

# **Modelos Lineares Generalizados e Extensões**

**Gauss Moutinho Cordeiro**

Departamento de Estatística e Informática, UFRPE,

Rua Dom Manoel de Medeiros, s/n

50171-900, Recife, PE

Email: [gauss@deinfo.ufrpe.br](mailto:gauss@deinfo.ufrpe.br)

**Clarice G.B. Demétrio**

Departamento de Ciências Exatas, ESALQ, USP

Caixa Postal 9

13418-900 Piracicaba, SP

Email: [clarice@esalq.usp.br](mailto:clarice@esalq.usp.br)

11 de março de 2008

## **Prefácio**

Este livro é resultante de vários anos de lecionamento de cursos e minicursos sobre modelos lineares generalizados e tem como objetivo apresentar noções gerais desses modelos, algumas de suas extensões e aplicações. Enumerar as pessoas a quem devemos agradecimentos é uma tarefa difícil, pois são muitos aqueles que contribuíram de forma direta ou indireta para a elaboração deste material. Agradecemos a Eduardo Bonilha, funcionário do Departamento de Ciências Exatas da ESALQ/USP, o auxílio na digitação, e a todos que nos ajudaram lendo versões anteriores, cuidadosamente, e dando sugestões muito proveitosas. Agradecemos, também, ao CNPq, à CAPES e à FAPESP por financiamentos de projetos que trouxeram contribuições importantes para a elaboração deste livro.

Finalmente, assumimos total responsabilidade pelas imperfeições e solicitamos aos leitores que nos apresentem críticas e sugestões para uma futura edição revisada.

Gauss Moutinho Cordeiro

Clarice Garcia Borges Demétrio

Piracicaba, fevereiro de 2008

# Sumário

<b>1</b>	<b>Família exponencial de distribuições</b>	<b>1</b>
1.1	Introdução . . . . .	1
1.2	Família exponencial uniparamétrica . . . . .	2
1.3	Componente aleatório . . . . .	4
1.4	Função geradora de momentos . . . . .	7
1.5	Estatística suficiente . . . . .	12
1.6	Família exponencial multiparamétrica . . . . .	13
1.7	Exercícios . . . . .	16
<b>2</b>	<b>Modelo Linear Generalizado</b>	<b>23</b>
2.1	Introdução . . . . .	23
2.2	Exemplos de motivação . . . . .	26
2.3	Definição . . . . .	40
2.4	Modelos especiais . . . . .	44
2.4.1	Modelo clássico de regressão . . . . .	44
2.4.2	Modelo de Poisson . . . . .	46
2.4.3	Modelo binomial . . . . .	48
2.4.3.1	Dados na forma de proporções . . . . .	49
2.4.3.2	Dados binários agrupados . . . . .	51
2.4.4	Modelo gama . . . . .	52
2.4.5	Modelo normal inverso . . . . .	54
2.4.6	Modelo binomial negativo . . . . .	55

2.4.7	Modelo secante hiperbólico generalizado . . . . .	56
2.4.8	Modelos definidos por transformações . . . . .	57
2.5	Metodologia . . . . .	58
2.5.1	Formulação de modelos . . . . .	58
2.5.2	Ajuste dos modelos . . . . .	63
2.5.3	Inferência . . . . .	64
2.6	Exercícios . . . . .	65
<b>3</b>	<b>Estimação</b>	<b>69</b>
3.1	Estatísticas suficientes . . . . .	69
3.2	O algoritmo de estimação . . . . .	71
3.3	Estimação em modelos especiais . . . . .	77
3.4	Resultados adicionais na estimação . . . . .	79
3.5	Seleção do modelo . . . . .	82
3.6	Considerações sobre a função de verossimilhança . . . . .	85
3.7	Exercícios . . . . .	89
<b>4</b>	<b>Métodos de Inferência</b>	<b>93</b>
4.1	Distribuição dos estimadores dos parâmetros . . . . .	93
4.2	Função desvio e estatística de Pearson generalizada . . . . .	99
4.3	Análise do desvio e seleção de modelos . . . . .	109
4.4	Estimação do parâmetro de dispersão . . . . .	113
4.5	Comparação dos três métodos de estimação do parâmetro de dispersão no modelo gama . . . . .	115
4.6	Testes de hipóteses . . . . .	116
4.6.1	Teste de uma hipótese nula simples . . . . .	117
4.6.2	Teste de uma hipótese nula composta . . . . .	119
4.7	Regiões de confiança . . . . .	122
4.8	Seleção de covariáveis . . . . .	123
4.9	Método das variáveis explanatórias adicionais . . . . .	125

4.10	Seleção da função de ligação . . . . .	127
4.11	Exercícios . . . . .	129
<b>5</b>	<b>Resíduos e Diagnósticos</b>	<b>135</b>
5.1	Introdução . . . . .	135
5.2	Técnicas para a verificação do ajuste de um modelo . . . . .	136
5.3	Análise de resíduos e diagnóstico para o modelo clássico de regressão	137
5.3.1	Tipos de resíduos . . . . .	138
5.3.2	Estatísticas para diagnósticos . . . . .	140
5.3.3	Tipos de gráficos . . . . .	144
5.4	Análise de resíduos e diagnóstico para modelos lineares generalizados	150
5.4.1	Tipos de resíduos . . . . .	151
5.4.2	Tipos de gráficos . . . . .	157
5.4.3	Resíduos de Pearson estudentizados . . . . .	159
5.5	Verificação da função de ligação . . . . .	161
5.6	Verificação da adequação da função de variância . . . . .	164
5.7	Verificação da adequação das escalas das variáveis explanatórias . . .	165
5.8	Verificação de anomalias no componente sistemático através da análise dos resíduos . . . . .	169
5.9	Exercícios . . . . .	172
<b>6</b>	<b>Aplicações a Dados Contínuos</b>	<b>179</b>
6.1	Dados de volume de árvores . . . . .	179
6.2	Dados de gordura no leite . . . . .	184
6.3	Dados de importação Brasileira . . . . .	185
6.4	Dados de tempos de sobrevivência de ratos . . . . .	190
6.5	Dados de assinaturas de TV a cabo . . . . .	192
6.6	Dados de demanda de energia elétrica . . . . .	204
6.7	Dados de tempo de funcionamento de um transformador . . . . .	210
6.8	Dados de malária . . . . .	218

<b>7</b>	<b>Aplicações a Dados Discretos</b>	<b>223</b>
7.1	Dados binários e proporções . . . . .	223
7.1.1	Estimação da dose efetiva e seu intervalo de confiança . . . . .	223
7.1.2	Paralelismo entre retas no modelo logístico linear . . . . .	226
7.2	Dados de contagem . . . . .	233
7.2.1	Modelo de Poisson . . . . .	233
7.2.2	Modelos log-lineares para tabelas $2 \times 2$ . . . . .	235
7.3	Introdução . . . . .	239
7.4	Metodologia . . . . .	240
7.5	Modelo Binomial Negativo . . . . .	242
7.6	Conclusões . . . . .	247
<b>8</b>	<b>Inferência Adicional</b>	<b>253</b>
8.1	Existência, finitude e unicidade dos estimadores dos parâmetros $\beta$ . . . . .	253
8.2	Uma família de funções de ligação . . . . .	254
8.3	Identificação de observações não explicadas pelo modelo . . . . .	257
8.4	Identificação de observações não explicadas pelo modelo . . . . .	258
8.4.1	Pontos influentes no modelo linear generalizado . . . . .	259
8.4.2	Diagnóstico local de um único ponto influente . . . . .	260
8.4.3	Diagnóstico global de um único ponto influente . . . . .	263
8.4.4	Diagnóstico local e global da influência de um conjunto de pontos . . . . .	265
8.5	Teste de hipóteses com restrições . . . . .	266
8.5.1	O teste através da razão de verossimilhanças . . . . .	267
8.5.2	O teste através da estatística escore . . . . .	267
8.5.3	O teste segundo a estatística de Wald . . . . .	268
8.5.4	Comparações entre as três estatísticas . . . . .	269
8.5.5	Aplicação do teste na prática . . . . .	269
8.5.6	Componente sistemático em termos de restrições . . . . .	271
8.6	Exercícios . . . . .	274

<b>9</b>	<b>Modelo Normal Não-Linear</b>	<b>275</b>
9.1	Introdução . . . . .	275
9.2	Estimação de Máxima Verossimilhança . . . . .	279
9.2.1	Resultados Assintóticos . . . . .	282
9.2.2	Exemplos . . . . .	284
9.3	Medidas de Não-Linearidade . . . . .	286
9.3.1	Medidas de Curvatura de Bates e Watts . . . . .	288
9.3.2	Viés de Ordem $n^{-1}$ de $\hat{\beta}$ . . . . .	291
9.3.3	Aperfeiçoamento da Estatística da Razão de Verossimilhanças . . . . .	293
9.3.4	Exemplos . . . . .	294
9.4	Técnicas de Diagnóstico . . . . .	296
9.4.1	Matriz de Projeção . . . . .	296
9.4.2	Resíduo Projetado . . . . .	297
9.4.3	Medidas de Influência . . . . .	299
9.4.4	Gráfico da Variável Adicionada . . . . .	300
9.4.5	Exemplos . . . . .	301
9.5	Exercícios . . . . .	304
<b>10</b>	<b>Outros Modelos de regressão</b>	<b>309</b>
10.1	Introdução . . . . .	309
10.2	Modelo de Box e Cox . . . . .	309
10.3	Modelos lineares generalizados com função de ligação composta . . . . .	312
10.4	Modelos semi-paramétricos . . . . .	313
10.5	Modelos aditivos generalizados . . . . .	313
10.6	Modelos de quase-verossimilhança . . . . .	315
10.7	Modelos para análise de dados de sobrevivência . . . . .	320
10.7.1	Modelos de riscos proporcionais . . . . .	320
10.7.2	Riscos proporcionais de Cox . . . . .	323
10.8	Modelos lineares generalizados com covariáveis de dispersão . . . . .	324

10.9 Modelos lineares generalizados com superdispersão . . . . .	327
10.10 Modelos com matriz de covariância não-escalar . . . . .	330
10.11 Modelo de regressão rígida . . . . .	333
10.12 Modelos heterocedásticos . . . . .	334
10.13 Modelos autocorrelacionados . . . . .	338
<b>11 Modelos não-lineares da família exponencial</b>	<b>341</b>
11.1 Introdução . . . . .	341
<b>12 Modelos simétricos não-lineares</b>	<b>343</b>
12.1 Introdução . . . . .	343



# Capítulo 1

## Família exponencial de distribuições

### 1.1 Introdução

Muitas das distribuições conhecidas podem ser reunidas em uma família denominada família exponencial de distribuições. Assim, por exemplo, pertencem a essa família as distribuições normal, binomial, binomial negativa, gama, Poisson, normal inversa, multinomial, beta, logarítmica, entre outras. Essa classe de distribuições foi proposta independentemente por Koopman, Pitman e Darrois através do estudo de propriedades de suficiência estatística. Posteriormente, muitos outros aspectos dessa família foram estudados e tornaram-se importantes na teoria moderna de Estatística. O conceito de família exponencial foi introduzido na Estatística por Fisher mas os modelos da família exponencial apareceram na Mecânica Estatística no final do século XIX e foram desenvolvidos por Maxwell, Boltzmann e Gibbs. A importância da família exponencial de distribuições teve maior destaque, na área dos modelos de regressão, a partir do trabalho pioneiro de Nelder e Wedderburn (1972) que definiram os modelos lineares generalizados. Na década de 80, esses modelos popularizaram-se, inicialmente, no Reino Unido, e, posteriormente, nos Estados Unidos e na Europa.

## 1.2 Família exponencial uniparamétrica

A família exponencial uniparamétrica é caracterizada por uma função (de probabilidade ou densidade) da forma

$$f(x; \theta) = h(x) \exp [\eta(\theta) t(x) - b(\theta)], \quad (1.1)$$

em que as funções  $\eta(\theta)$ ,  $b(\theta)$ ,  $t(x)$  e  $h(x)$  assumem valores em subconjuntos dos reais. As funções  $\eta(\theta)$ ,  $b(\theta)$  e  $t(x)$  não são únicas. Por exemplo,  $\eta(\theta)$  pode ser multiplicada por uma constante  $k$  e  $t(x)$  pode ser dividida pela mesma constante.

Várias distribuições importantes podem ser escritas na forma (1.1), tais como: Poisson, binomial, Rayleigh, normal, gama e normal inversa (as três últimas com a suposição de que um dos parâmetros é conhecido). No artigo de Cordeiro et al. (1995), são apresentadas 24 distribuições na forma (1.1). O suporte da família exponencial (1.1), isto é,  $\{x; f(x; \theta) > 0\}$ , não pode depender de  $\theta$ . Assim, a distribuição uniforme em  $(0, \theta)$  não é um modelo da família exponencial. Pelo teorema da fatoração de Neyman-Fisher, a estatística  $t(X)$  é suficiente para  $\theta$ .

É muito fácil verificar se uma distribuição pertence, ou não, à família exponencial (1.1), como é mostrado nos três exemplos que se seguem.

**Exemplo 1.1:** A distribuição de Poisson  $P(\theta)$  de parâmetro  $\theta > 0$ , usada para análise de dados na forma de contagens, tem função de probabilidade

$$f(x; \theta) = \frac{e^{-\theta} \theta^x}{x!} = \frac{1}{x!} \exp(x \log \theta - \theta)$$

e, portanto, é um membro da família exponencial (1.1) com  $\eta(\theta) = \log \theta$ ,  $b(\theta) = \theta$ ,  $t(x) = x$  e  $h(x) = 1/x!$ .

**Exemplo 1.2:** A distribuição binomial  $B(m, \theta)$ , com  $0 < \theta < 1$  e  $m$ , o número conhecido de ensaios independentes, é usada para análise de dados na forma de proporções e tem função de probabilidade

$$f(x; \theta) = \binom{m}{x} \theta^x (1 - \theta)^{m-x} = \binom{m}{x} \exp \left[ x \log \left( \frac{\theta}{1 - \theta} \right) + m \log(1 - \theta) \right]$$

com  $\eta(\theta) = \log[\theta/(1 - \theta)]$ ,  $b(\theta) = -m \log(1 - \theta)$ ,  $t(x) = x$  e  $h(x) = \binom{m}{x}$ , sendo, portanto, um membro da família exponencial (1.1).

**Exemplo 1.3:** A distribuição de Rayleigh, usada para análise de dados contínuos positivos, tem função densidade ( $x > 0, \theta > 0$ )

$$f(x; \theta) = \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right) = x \exp\left(-\frac{1}{2\theta^2}x^2 - \log \theta^2\right),$$

e, portanto, pertence à família exponencial (1.1) com  $\eta(\theta) = -1/(2\theta^2)$ ,  $b(\theta) = \log \theta^2$ ,  $t(x) = x^2$  e  $h(x) = x$ .

A família exponencial na forma canônica é definida a partir de (1.1), considerando que as funções  $\eta(\theta)$  e  $t(x)$  são iguais à função identidade, de forma que

$$f(x; \theta) = h(x) \exp[\theta x - b(\theta)]. \quad (1.2)$$

Na parametrização (1.2),  $\theta$  é chamado de **parâmetro canônico**. O logaritmo da função de verossimilhança correspondente a uma única observação no modelo (1.2) é dado por

$$\ell(\theta) = \theta x - b(\theta) + \log[h(x)]$$

e, portanto, a função escore  $U = U(\theta) = d\ell(\theta)/d\theta$  resulta em  $U = x - b'(\theta)$ .

É fácil verificar das propriedades da função escore,  $E(U) = 0$  e  $\text{Var}(U) = -E[d^2\ell(\theta)/d\theta^2]$  (esta última igualdade é a informação de Fisher), que

$$E(X) = b'(\theta) \quad \text{e} \quad \text{Var}(X) = b''(\theta). \quad (1.3)$$

O fato simples de se calcularem momentos da família exponencial (1.2) em termos de derivadas da função  $b(\theta)$  (denominada de função geradora de cumulantes) em relação ao parâmetro canônico  $\theta$  é muito importante na teoria dos modelos lineares generalizados, principalmente, no contexto assintótico.

Suponha que  $X_1, \dots, X_n$  sejam  $n$  variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), seguindo (1.1). A distribuição conjunta de  $X_1, \dots, X_n$

é dada por

$$f(x_1, \dots, x_n; \theta) = \left[ \prod_{i=1}^n h(x_i) \right] \exp \left[ \eta(\theta) \sum_{i=1}^n t(x_i) - nb(\theta) \right]. \quad (1.4)$$

A equação (1.4) implica que a distribuição conjunta de  $X_1, \dots, X_n$  é, também, um modelo da família exponencial. A estatística suficiente é  $\sum_{i=1}^n T(X_i)$  e tem dimensão um qualquer que seja  $n$ .

É, geralmente, verdade que a estatística suficiente de um modelo da família exponencial segue, também, a família exponencial. Por exemplo, se  $X_1, \dots, X_n$  são variáveis aleatórias i.i.d. com distribuição de Poisson  $P(\theta)$ , então a estatística suficiente  $\sum_{i=1}^n T(X_i)$  tem, também, distribuição de Poisson  $P(n\theta)$  e, assim, é um modelo exponencial uniparamétrico.

### 1.3 Componente aleatório

Como será visto, na Seção 2.3, o componente aleatório de um modelo linear generalizado é definido a partir da família exponencial uniparamétrica na forma canônica (1.2) com a introdução de um parâmetro  $\phi > 0$  de perturbação. Nelder e Wedderburn (1972) ao fazerem isso, conseguiram incorporar distribuições biparamétricas no componente aleatório do modelo. Tem-se,

$$f(y; \theta, \phi) = \exp \left\{ \phi^{-1} [y\theta - b(\theta)] + c(y, \phi) \right\}, \quad (1.5)$$

em que  $b(\cdot)$  e  $c(\cdot)$  são funções conhecidas. Quando  $\phi$  é conhecido, a família de distribuições (1.5) é idêntica à família exponencial na forma canônica (1.2). Na Seção 1.4, será demonstrado que o valor esperado e a variância de  $Y$  com distribuição na família (1.5) são

$$E(Y) = \mu = b'(\theta) \text{ e } \text{Var}(Y) = \phi b''(\theta).$$

Observa-se, então, que  $\phi$  é um parâmetro de **dispersão** do modelo e seu inverso  $\phi^{-1}$ , uma medida de **precisão**. A função que relaciona o parâmetro canônico

$\theta$  com a média  $\mu$  (inversa da função  $b'(\cdot)$ ) é denotada por  $\theta = q(\mu)$ . A função da média  $\mu$  na variância é representada por  $b''(\theta) = V(\mu)$ . Denomina-se  $V(\mu)$  de **função de variância**. Observe-se que o parâmetro canônico pode ser obtido de  $\theta = \int V^{-1}(\mu)d\mu$ , pois  $V(\mu) = d\mu/d\theta$ . A Tabela 1.1 apresenta várias distribuições importantes na família (1.5), caracterizando as funções  $b(\theta)$ ,  $c(y, \phi)$ , a média  $\mu$  em termos do parâmetro canônico  $\theta$  e a função de variância  $V(\mu)$ . Nessa tabela,  $\Gamma(\cdot)$  é a função gama, isto é,  $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ ,  $\alpha > 0$ . A família de distribuições (1.5) permite incorporar distribuições que exibem assimetria e de natureza discreta ou contínua e com suportes que são restritos a intervalos do conjunto dos reais, conforme bem exemplificam as distribuições dadas na Tabela 1.1. Essas distribuições são estudadas no Capítulo 2.

Convém salientar que se  $\phi$  não for conhecido, a família (1.5) pode, ou não, pertencer à família exponencial biparamétrica (Seção 1.6). Para (1.5) pertencer à família exponencial biparamétrica quando  $\phi$  é desconhecido, a função  $c(y, \phi)$  deve ser decomposta, segundo Cordeiro e McCullagh (1991), como  $c(y, \phi) = \phi^{-1}d(y) + d_1(y) + d_2(\phi)$ . Esse é o caso das distribuições normal, gama e normal inversa.

Morris (1982) demonstra que existem apenas seis distribuições na família (1.5) cuja função de variância é uma função, no máximo, quadrática da média. Essas distribuições são normal ( $V = 1$ ), gama ( $V = \mu^2$ ), binomial ( $V = \mu(1 - \mu)$ ), Poisson ( $V = \mu$ ), binomial negativa ( $V = \mu + \mu^2/k$ ) e a sexta, chamada secante hiperbólica generalizada ( $V = 1 + \mu^2$ ), cuja densidade é dada por

$$\pi(y; \theta) = \frac{1}{2} \exp[\theta y + \log(\cos \theta)] \cosh\left(\frac{\pi y}{2}\right), \quad y \in \mathbb{R}. \quad (1.6)$$

A distribuição secante hiperbólica generalizada (1.6) compete com a distribuição normal na análise de observações contínuas irrestritas. A seguir, são apresentadas duas distribuições que são membros da família (1.5).

**Exemplo 1.4:** A distribuição normal de média  $\mu \in \mathbb{R}$  e variância  $\sigma^2 > 0$ , tem função

Tabela 1.1: Algumas distribuições importantes na família exponencial (1.5).

Distribuição	$\phi$	$\theta$	$b(\theta)$	$c(y, \phi)$	$\mu(\theta)$	$V(\mu)$
Normal: $N(\mu, \sigma^2)$	$\sigma^2$	$\mu$	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left[ \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$	$\theta$	1
Poisson: $P(\mu)$	1	$\log \mu$	$e^\theta$	$-\log y!$	$e^\theta$	$\mu$
Binomial: $B(m, \pi)$	1	$\log \left( \frac{\mu}{m - \mu} \right)$	$m \log(1 + e^\theta)$	$\log \binom{m}{y}$	$\frac{me^\theta}{1 + e^\theta}$	$\frac{\mu}{m}(m - \mu)$
Binomial Negativa: $BN(\mu, k)$	1	$\log \left( \frac{\mu}{\mu + k} \right)$	$-k \log(1 - e^\theta)$	$\log \left[ \frac{\Gamma(k + y)}{\Gamma(k)y!} \right]$	$\frac{ke^\theta}{1 - e^\theta}$	$\mu \left( \frac{\mu}{k} + 1 \right)$
Gama: $G(\mu, \nu)$	$\nu^{-1}$	$\frac{1}{-\frac{1}{\mu}}$	$-\log(-\theta)$	$\nu \log(\nu y) - \log y - \log \Gamma(\nu)$	$\frac{1}{-\frac{1}{\theta}}$	$\mu^2$
Normal Inversa: $IG(\mu, \sigma^2)$	$\sigma^2$	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\frac{1}{2} \left[ \log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y} \right]$	$(-2\theta)^{-1/2}$	$\mu^3$

densidade dada por

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(y - \mu)^2}{2\sigma^2} \right].$$

Tem-se, então,

$$\begin{aligned} f(y; \mu, \sigma^2) &= \exp \left[ -\frac{(y - \mu)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right] \\ &= \exp \left[ \frac{1}{\sigma^2} \left( y\mu - \frac{\mu^2}{2} \right) - \frac{1}{2} \log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} \right], \end{aligned}$$

obtendo-se os elementos da primeira linha da Tabela 1.1, isto é,

$$\theta = \mu, \quad \phi = \sigma^2, \quad b(\theta) = \frac{\mu^2}{2} = \frac{\theta^2}{2} \quad \text{e} \quad c(y, \phi) = -\frac{1}{2} \left[ \frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right],$$

o que mostra que a distribuição  $N(\mu, \sigma^2)$  pertence à família exponencial (1.5).

**Exemplo 1.5:** A distribuição binomial tem função de probabilidade

$$f(y; \pi) = \binom{m}{y} \pi^y (1 - \pi)^{m-y}, \quad \pi \in [0, 1], \quad y = 0, 1, \dots, m.$$

Tem-se, então,

$$\begin{aligned} f(y; \pi) &= \exp \left[ \log \binom{m}{y} + y \log \pi + (m - y) \log(1 - \pi) \right] \\ &= \exp \left[ y \log \left( \frac{\pi}{1 - \pi} \right) + m \log(1 - \pi) + \log \binom{m}{y} \right], \end{aligned}$$

obtendo-se os elementos da terceira linha da Tabela 1.1, isto é,

$$\begin{aligned} \phi = 1, \quad \theta = \log \left( \frac{\pi}{1 - \pi} \right) &= \log \left( \frac{\mu}{m - \mu} \right), \quad \text{o que implica} \quad \mu = \frac{me^\theta}{(1 + e^\theta)}, \\ b(\theta) &= -m \log(1 - \pi) = m \log(1 + e^\theta) \quad \text{e} \quad c(y, \phi) = \log \binom{m}{y} \end{aligned}$$

e, portanto, a distribuição binomial pertence à família exponencial (1.5).

## 1.4 Função geradora de momentos

A função geradora de momentos (f.g.m.) da família (1.5) é dada por

$$M(t; \theta, \phi) = E(e^{tY}) = \exp \{ \phi^{-1} [b(\phi t + \theta) - b(\theta)] \}. \quad (1.7)$$

**Prova:** A prova será feita apenas para o caso de variáveis aleatórias contínuas. No caso discreto, basta substituir a integral pelo somatório. Lembrando-se que

$$\int f(y; \theta, \phi) dy = 1,$$

chega-se a

$$\int \exp \{ \phi^{-1} [\theta y - b(\theta)] + c(y, \phi) \} dy = 1,$$

obtendo-se

$$\int \exp [\phi^{-1} \theta y + c(y, \phi)] dy = \exp [\phi^{-1} b(\theta)]. \quad (1.8)$$

Logo,

$$\begin{aligned} M(t; \theta, \phi) &= E(e^{tY}) = \int \exp(ty) f(y) dy \\ &= \int \exp \{ \phi^{-1} [(\phi t + \theta)y - b(\theta)] + c(y, \phi) \} dy \\ &= \frac{1}{\exp [\phi^{-1} b(\theta)]} \int \exp [\phi^{-1} (\phi t + \theta)y + c(y, \phi)] dy \end{aligned}$$

e, usando-se (1.8), tem-se

$$M(t; \theta, \phi) = \frac{\exp [\phi^{-1} b(\phi t + \theta)]}{\exp [\phi^{-1} b(\theta)]}$$

ou ainda, demonstrando (1.7),

$$M(t; \theta, \phi) = \exp \{ \phi^{-1} [b(\phi t + \theta) - b(\theta)] \}.$$

A função geradora de cumulantes (f.g.c.) correspondente é, então,

$$\varphi(t; \theta, \phi) = \log[M(t; \theta, \phi)] = \phi^{-1} [b(\phi t + \theta) - b(\theta)]. \quad (1.9)$$

A f.g.c. desempenha um papel mais importante do que a f.g.m., na Estatística, pois uma grande parte da teoria assintótica depende de suas propriedades. Derivando-se (1.9), sucessivamente, em relação a  $t$ , vem

$$\varphi^{(r)}(t; \theta, \phi) = \phi^{r-1} b^{(r)}(\phi t + \theta),$$



em que  $b^{(r)}(\cdot)$  indica a derivada de  $r$ -ésima ordem de  $b(\cdot)$  em relação a  $t$ . Para  $t = 0$ , obtém-se o  $r$ -ésimo cumulante da família (1.5) como

$$\kappa_r = \phi^{r-1} b^{(r)}(\theta). \quad (1.10)$$

Como enfatizado anteriormente, podem-se agora deduzir, a partir de (1.10), o valor esperado  $\kappa_1 = \mu$  e a variância  $\kappa_2$  da família (1.5) para  $r = 1$  e  $2$ , respectivamente. Tem-se que  $\kappa_1 = \mu = b'(\theta)$  e  $\kappa_2 = \phi b''(\theta) = \phi d\mu/d\theta$ .

A equação (1.10) mostra que existe uma relação interessante de recorrência entre os cumulantes da família (1.5), isto é,  $\kappa_{r+1} = \phi d\kappa_r/d\theta$  para  $r = 1, 2, \dots$ . Isso é fundamental para obtenção de propriedades assintóticas dos estimadores de máxima verossimilhança nos modelos lineares generalizados.

Podem-se, alternativamente, deduzir essas expressões, usando-se as propriedades da função escore. Seja  $\ell = \ell(\theta, \phi) = \log f(y; \theta, \phi)$  o logaritmo da função de verossimilhança correspondente a uma única observação em (1.5). Tem-se

$$U = \frac{d\ell}{d\theta} = \phi^{-1}[y - b'(\theta)]$$

e

$$U' = \frac{d^2\ell}{d\theta^2} = -\phi^{-1}b''(\theta).$$

Logo,

$$E(U) = \phi^{-1} [E(Y) - b'(\theta)] = 0 \quad \text{que implica em} \quad E(Y) = b'(\theta)$$

e, assim,

$$\text{Var}(U) = -E(U') = \phi^{-1}b''(\theta) \quad \text{e} \quad \text{Var}(U) = E(U^2) = \phi^{-2}\text{Var}(Y).$$

Então,

$$\text{Var}(Y) = \phi b''(\theta).$$

**Exemplo 1.6:** Considere o Exemplo 1.4 da distribuição normal e obtenha  $\varphi(t)$  e  $M(t)$ , representadas, agora, sem os parâmetros  $\theta$  e  $\phi$ . Tem-se que  $\phi = \sigma^2$ ,  $\theta = \mu$  e  $b(\theta) = \theta^2/2$ . De (1.9) vem a f.g.c.

$$\begin{aligned}\varphi(t) &= \frac{1}{\sigma^2} \left[ \frac{(\sigma^2 t + \theta)^2}{2} - \frac{\theta^2}{2} \right] \\ &= \frac{1}{2} (\sigma^2 t^2 + 2t\theta) = t\mu + \frac{\sigma^2 t^2}{2}.\end{aligned}$$

Note que, derivando-se  $\varphi(t)$  e fazendo-se  $t = 0$ , tem-se que  $\kappa_1 = \mu$ ,  $\kappa_2 = \sigma^2$  e  $\kappa_r = 0$ ,  $r \geq 3$ . Assim, todos os cumulantes da distribuição normal de ordem maior do que dois são nulos.

Logo, a f.g.m. é dada por

$$M(t) = \exp \left( t\mu + \frac{\sigma^2 t^2}{2} \right).$$

**Exemplo 1.7:** Considere o Exemplo 1.5 da distribuição binomial e obtenha  $\varphi(t)$  e  $M(t)$ . Tem-se que

$$\phi = 1, \theta = \log[\mu/(m - \mu)] \text{ e } b(\theta) = -m \log(1 - \pi) = m \log(1 + e^\theta).$$

Logo, usando-se a f.g.c. (1.9), vem

$$\begin{aligned}\varphi(t) &= m [\log(1 + e^{t+\theta}) - \log(1 + e^\theta)] \\ &= \log \left( \frac{1 + e^{t+\theta}}{1 + e^\theta} \right)^m = \log \left( \frac{m - \mu}{m} + \frac{\mu}{m} e^t \right)^m.\end{aligned}$$

Assim, a f.g.m. é

$$M(t) = e^{\varphi(t)} = \left( \frac{m - \mu}{m} + \frac{\mu}{m} e^t \right)^m.$$

A Tabela 1.2 apresenta as funções geradoras de momentos para as distribuições dadas na Tabela 1.1.

Pode-se demonstrar, que especificando a forma da função  $\mu = q^{-1}(\theta)$ , a distribuição em (1.5) é univocamente determinada. Assim, uma relação funcional variância-média caracteriza a distribuição na família (1.5). Entretanto,

Tabela 1.2: Funções geradoras de momentos para algumas distribuições.

Distribuição	Função geradora de momentos $M(t; \theta, \phi)$
Normal: $N(\mu, \sigma^2)$	$\exp\left(t\mu + \frac{\sigma^2 t^2}{2}\right)$
Poisson: $P(\mu)$	$\exp[\mu(e^t - 1)]$
Binomial: $B(m, \pi)$	$\left(\frac{m - \mu}{m} + \frac{\mu}{m}e^t\right)^m$
Bin. Negativa: $BN(\mu, k)$	$\left[1 + \frac{\mu}{k}(1 - e^t)\right]^{-k}$
Gama: $G(\mu, \nu)$	$\left(1 - \frac{t\mu}{\nu}\right)^{-\nu}, \quad t < \frac{\nu}{\mu}$
Normal Inversa: $IG(\mu, \sigma^2)$	$\exp\left\{\frac{1}{\sigma^2}\left[\frac{1}{\mu} - \left(\frac{1}{\mu^2} - 2t\sigma^2\right)^{1/2}\right]\right\}, \quad t < \frac{1}{2\sigma^2\mu^2}$

essa relação não caracteriza a distribuição na família exponencial não-linear  $\pi(y; \theta, \phi) = \exp\{\phi^{-1}[t(y)\theta - b(\theta)] + c(y, \phi)\}$ . Esse fato é comprovado com os três exemplos que se seguem.

**Exemplo 1.8:** Se  $Y$  tem distribuição beta com parâmetros  $\phi^{-1}\mu$  e  $\phi^{-1}(1 - \mu)$  e f.d.p. dada por

$$f(y; \mu, \phi) = \frac{y^{\phi^{-1}\mu-1}(1-y)^{\phi^{-1}(1-\mu)-1}}{B[\phi^{-1}\mu, \phi^{-1}(1-\mu)]},$$

em que  $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$  é a função beta, tem-se que  $t(y) = \log[y/(1-y)]$ ,  $\theta = \mu$  e  $\text{Var}(Y) = \phi\mu(1-\mu)/(1+\phi)$ , obtendo-se uma função de variância do mesmo tipo que a do modelo binomial.

**Exemplo 1.9:** Se  $Y$  tem distribuição de Euler com média  $\mu$  e f.d.p.

$$f(y; \mu) = \exp\{\mu \log y - \mu - \log[\Gamma(\mu)]\},$$

tem-se que  $t(y) = \log y$ ,  $\theta = \mu$  e  $\text{Var}(Y) = \mu$  que é do mesmo tipo que a função de variância do modelo de Poisson.

**Exemplo 1.10:** Se  $Y$  tem distribuição log normal de parâmetros  $\alpha$  e  $\sigma^2$  e f.d.p. dada por

$$f(y; \alpha, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} \exp \left[ -\frac{(\log y - \alpha)^2}{2\sigma^2} \right],$$

então,  $E(Y) = \mu = \exp(\alpha + \sigma^2/2)$ ,  $t(y) = \log y$ ,  $\theta = \alpha/\sigma^2$  e  $\text{Var}(Y) = \mu^2[\exp(\sigma^2) - 1]$ , que é do mesmo tipo que a função de variância do modelo gama.

## 1.5 Estatística suficiente

Uma estatística  $T = T(\mathbf{Y})$  é suficiente para um parâmetro  $\theta$  (que pode ser um vetor) quando contém toda informação sobre esse parâmetro contida na amostra  $\mathbf{Y}$ . Se  $T$  é suficiente para  $\theta$ , então, a distribuição condicional de  $\mathbf{Y}$  dada a estatística  $T(\mathbf{Y})$  é independente de  $\theta$ , isto é,

$$P(\mathbf{Y} = \mathbf{y} | T = t, \theta) = P(\mathbf{Y} = \mathbf{y} | T = t).$$

O critério da fatoração é uma forma conveniente de caracterizar uma estatística suficiente. Uma condição necessária e suficiente para  $T$  ser suficiente para um parâmetro  $\theta$  é que a f.d.p.  $f_{\mathbf{Y}}(\mathbf{y}, \theta)$  possa ser decomposta como

$$f_{\mathbf{Y}}(\mathbf{y}, \theta) = h(\mathbf{y})g(t, \mathbf{y}),$$

em que  $t = T(\mathbf{y})$  e  $h(\mathbf{y})$  não dependem de  $\theta$ .

Seja  $Y_1, \dots, Y_n$  uma amostra aleatória (a.a.) de uma distribuição que pertence à família (1.5). A distribuição conjunta de  $Y_1, \dots, Y_n$  é dada por

$$\begin{aligned} f(y; \theta, \phi) &= \prod_{i=1}^n f(y_i; \theta, \phi) = \prod_{i=1}^n \exp \{ \phi^{-1} [y_i \theta - b(\theta)] + c(y_i, \phi) \} \\ &= \exp \left\{ \phi^{-1} \left[ \theta \sum_{i=1}^n y_i - n b(\theta) \right] \right\} \exp \left[ \sum_{i=1}^n c(y_i, \phi) \right]. \end{aligned}$$

Pelo teorema da fatoração de Neyman-Fisher e supondo  $\phi$  conhecido, tem-se que  $T = \sum_{i=1}^n Y_i$  é uma estatística suficiente para  $\theta$ , pois

$$f(y; \theta, \phi) = g(t, \theta) h(y_1, \dots, y_n),$$

sendo que  $g(t, \theta)$  depende de  $\theta$  e dos  $y$ 's apenas através de  $t$  e  $h(y_1, \dots, y_n)$  independe de  $\theta$ .

Isso mostra, que se uma distribuição pertence à família exponencial uniparamétrica, então, existe uma estatística suficiente. Na realidade, usando-se o Teorema de Lehmann-Scheffé (Mendenhall et al., 1981) mostra-se que  $T = \sum_{i=1}^n Y_i$  é uma estatística suficiente minimal.

## 1.6 Família exponencial multiparamétrica

A família exponencial multiparamétrica de dimensão  $k$  é caracterizada por uma função (de probabilidade ou densidade) da forma

$$f(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x}) \exp \left[ \sum_{i=1}^k \eta_i(\boldsymbol{\theta}) t_i(\mathbf{x}) - b(\boldsymbol{\theta}) \right], \quad (1.11)$$

em que  $\boldsymbol{\theta}$  é um vetor de parâmetros, usualmente, de dimensão  $k$ , e as funções  $\eta_i(\boldsymbol{\theta})$ ,  $b(\boldsymbol{\theta})$ ,  $t_i(\mathbf{x})$  e  $h(\mathbf{x})$  assumem valores em subconjuntos dos reais. Obviamente, a forma dada em (1.1) é um caso especial de (1.11). Pelo teorema da fatoração, o vetor  $\mathbf{T} = [T_1(\mathbf{X}), \dots, T_k(\mathbf{X})]^T$  é suficiente para o vetor de parâmetros  $\boldsymbol{\theta}$ . Quando  $\eta_i(\boldsymbol{\theta}) = \theta_i$ ,  $i = 1, \dots, k$ , obtém-se de (1.11) a família exponencial na forma canônica com parâmetros canônicos  $\theta_1, \dots, \theta_k$  e estatísticas canônicas  $T_1(\mathbf{X}), \dots, T_k(\mathbf{X})$ . Tem-se

$$f(\mathbf{x}; \boldsymbol{\theta}) = h(\mathbf{x}) \exp \left[ \sum_{i=1}^k \theta_i t_i(\mathbf{x}) - b(\boldsymbol{\theta}) \right]. \quad (1.12)$$

É fácil verificar (Exercício 12) que as distribuições normal, gama, normal inversa e beta pertencem à família exponencial biparamétrica canônica (1.12) com  $k = 2$ .

Gelfand e Dalal (1990) estudaram a família exponencial biparamétrica  $f(x; \theta, \tau) = h(x) \exp[\theta x + \tau t(x) - b(\theta, \tau)]$  que é um caso especial de (1.11) com  $k = 2$ . Essa família tem despertado interesse, recentemente, como o componente aleatório dos modelos lineares generalizados superdispersos (Dey et al., 1997). Dois casos especiais importantes dessa família são diretamente obtidos:

- a. a família exponencial canônica uniparamétrica (1.2) surge, naturalmente, quando  $\tau = 0$ ;
- b. o componente aleatório (1.5) dos modelos lineares generalizados é obtido incorporando o parâmetro de dispersão  $\phi$ .

**Exemplo 1.11:** Considere a distribuição multinomial com função de probabilidade

$$f(\mathbf{x}; \boldsymbol{\pi}) = \frac{n!}{x_1! \dots x_k!} \pi_1^{x_1} \dots \pi_k^{x_k},$$

em que  $\sum_{i=1}^k x_i = n$  e  $\sum_{i=1}^k \pi_i = 1$ . Essa distribuição pertence, obviamente, à família exponencial canônica (1.12) com parâmetro canônico  $\boldsymbol{\theta} = (\log \pi_1, \dots, \log \pi_k)^T$  e estatística canônica  $\mathbf{T} = (X_1, \dots, X_k)^T$ . Entretanto, devido à restrição  $\sum_{i=1}^k \pi_i = 1$ , a representação mínima da família exponencial é obtida considerando  $\boldsymbol{\theta} = [\log(\pi_1/\pi_k), \dots, \log(\pi_{k-1}/\pi_k)]^T$  e  $\mathbf{t} = (x_1, \dots, x_{k-1})^T$ , ambos vetores de dimensão  $k - 1$ , resultando na família exponencial multiparamétrica de dimensão  $k - 1$

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{n!}{x_1! \dots x_k!} \exp \left[ \sum_{i=1}^{k-1} \theta_i x_i - b(\boldsymbol{\theta}) \right], \quad (1.13)$$

com  $\theta_i = \log(\pi_i/\pi_k)$ ,  $i = 1, \dots, k - 1$ , e  $b(\boldsymbol{\theta}) = n \log \left( 1 + \sum_{i=1}^{k-1} e^{\theta_i} \right)$ .

Pode-se demonstrar que os dois primeiros momentos da estatística suficiente  $\mathbf{T} = [T_1(\mathbf{X}), \dots, T_k(\mathbf{X})]^T$  na família exponencial canônica (1.12) são dados por

$$E(\mathbf{T}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad \text{Cov}(\mathbf{T}) = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \quad (1.14)$$

As expressões (1.14) generalizam (1.3). Nas equações (1.14), o vetor  $\partial b(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$  de dimensão  $k$  tem um componente típico  $E[T_i(\mathbf{X})] = \partial b(\boldsymbol{\theta})/\partial \theta_i$  e a matriz  $\partial^2 b(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T$  de ordem  $k$  tem como elemento típico  $\text{Cov}(T_i(\mathbf{X}), T_j(\mathbf{X})) =$

$\partial^2 b(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j$ . Assim, os valores esperados e as covariâncias das estatísticas suficientes do modelo (1.12) são facilmente obtidos por simples diferenciação. A demonstração das equações (1.14) é deixada como Exercício 19.

Para o modelo multinomial (1.13), usando as equações (1.14), vem

$$\begin{aligned} E(X_i) &= n \frac{\partial}{\partial\theta_i} \log \left( 1 + \sum_{i=1}^{k-1} e^{\theta_i} \right) \\ &= \frac{ne^{\theta_i}}{1 + \sum_{i=1}^{k-1} e^{\theta_i}} = \frac{n \frac{\pi_i}{\pi_k}}{1 + \sum_{i=1}^{k-1} \frac{\pi_i}{\pi_k}} = n\pi_i \end{aligned}$$

e para  $i \neq j$

$$\begin{aligned} \text{Cov}(X_i, X_j) &= n \frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \left( 1 + \sum_{i=1}^{k-1} e^{\theta_i} \right) \\ &= \frac{-ne^{\theta_i}e^{\theta_j}}{\left( 1 + \sum_{i=1}^{k-1} e^{\theta_i} \right)^2} = -n\pi_i\pi_j \end{aligned}$$

e para  $i = j$

$$\begin{aligned} \text{Var}(X_i) &= n \frac{\partial^2}{\partial\theta_i^2} \log \left( 1 + \sum_{i=1}^{k-1} e^{\theta_i} \right) \\ &= n\pi_i(1 - \pi_i). \end{aligned}$$

Finalmente, apresenta-se mais uma distribuição na família exponencial canônica (1.12) com  $k = 2$ .

**Exemplo 1.12:** Considere a distribuição Gaussiana inversa reparametrizada por  $(\alpha, \beta > 0)$

$$f(x; \alpha, \beta) = \sqrt{\frac{\alpha}{2\pi}} e^{\sqrt{\alpha\beta}x - 3/2} \exp \left[ -\frac{1}{2}(\alpha x^{-1} + \beta x) \right], \quad x > 0.$$

Pode-se escrever essa f.d.p. na forma (1.12) com  $\mathbf{t} = \left( -\frac{1}{2}x^{-1}, -\frac{1}{2}x \right)^T$ ,  $\boldsymbol{\theta} = (\alpha, \beta)^T$  e  $b(\boldsymbol{\theta}) = -\frac{1}{2} \log \alpha - \sqrt{\alpha\beta}$ . Usando-se as equações (1.14), obtêm-se, por simples diferenciação,

$$E(X) = \sqrt{\frac{\alpha}{\beta}}, \quad E(X^{-1}) = \alpha^{-1} + \sqrt{\frac{\beta}{\alpha}}$$

e

$$\text{Cov}(X, X^{-1}) = \begin{pmatrix} \alpha^{1/2}\beta^{-3/2} & -(\alpha\beta)^{-1/2} \\ -(\alpha\beta)^{-1/2} & 2\alpha^{-1} + \alpha^{-3/2}\beta^{1/2} \end{pmatrix}.$$

## 1.7 Exercícios

1. Verifique se as distribuições que se seguem pertencem à família (1.5). Obtenha  $\varphi(t)$ ,  $M(t)$ ,  $E(Y)$ ,  $\text{Var}(Y)$  e  $V(\mu)$ .

a) Poisson:  $Y \sim P(\mu)$ ,  $\mu > 0$

$$f(y; \mu) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \dots$$

b) Binomial negativa ( $k$  fixo):  $Y \sim \text{BN}(\mu, k)$ ,  $k > 0$ ,  $\mu > 0$

$$f(y; \mu, k) = \frac{\Gamma(k+y)}{\Gamma(k)y!} \frac{\mu^y k^k}{(\mu+k)^{k+y}}, \quad y = 0, 1, \dots$$

c) Gama:  $Y \sim G(\mu, \nu)$ ,  $\nu > 0$ ,  $\mu > 0$

$$f(y; \mu, \nu) = \frac{\left(\frac{\nu}{\mu}\right)^\nu}{\Gamma(\nu)} y^{\nu-1} \exp\left(-\frac{y\nu}{\mu}\right), \quad y > 0$$

d) Normal inversa (ou inversa Gaussiana):  $Y \sim \text{IG}(\mu, \sigma^2)$ ,  $\sigma^2 > 0$ ,  $\mu > 0$

$$f(y; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2 y^3}\right)^{1/2} \exp\left[-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y}\right], \quad y > 0.$$

2. Seja  $X$  uma v.a. com distribuição gama  $G(\nu)$  de um parâmetro  $\nu > 0$ , com f.d.p.

$$f(x; \nu) = \frac{x^{\nu-1}e^{-x}}{\Gamma(\nu)}, \quad x > 0.$$

Dado que  $E(X) = \nu$ , mostre que usando-se a transformação  $Y = \frac{X}{\nu}\mu$ , obtém-se a f.d.p. usada no item c) do Exercício 1.

3. Seja  $Y$  uma v.a. com distribuição de Poisson truncada com parâmetro  $\lambda > 0$ , isto é, com função de probabilidade dada por

$$f(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!(1-e^{-\lambda})}, \quad y = 1, 2, \dots$$



Mostre que (Ridout e Demétrio, 1992):

a) essa distribuição é um membro da família exponencial na forma canônica;

$$\text{b) } E(Y) = \mu = \frac{\lambda}{1 - e^{-\lambda}};$$

$$\text{c) } \text{Var}(Y) = \frac{\lambda}{1 - e^{-\lambda}} \left( 1 - \frac{\lambda e^{-\lambda}}{1 - e^{-\lambda}} \right) = \mu(1 + \lambda - \mu);$$

$$\text{d) } M(t) = \frac{\exp(\lambda e^t) - 1}{e^\lambda - 1}.$$

4. Seja  $Y$  uma v.a. com distribuição binomial truncada com probabilidade de sucesso  $0 < \pi < 1$ , isto é, com função de probabilidade dada por

$$f(y; \pi) = \frac{\binom{m}{y} \pi^y (1 - \pi)^{(m-y)}}{1 - (1 - \pi)^m}, \quad y = 1, \dots, m.$$

Mostre que (Vieira et al., 2000):

a) essa distribuição é um membro da família exponencial na forma canônica;

$$\text{b) } E(Y) = \mu = \frac{m\pi}{1 - (1 - \pi)^m};$$

$$\text{c) } \text{Var}(Y) = \mu[1 + \pi(m - 1) - \mu];$$

$$\text{d) } M(t) = \frac{(1 - \pi + \pi e^t)^m - (1 - \pi)^m}{1 - (1 - \pi)^m}.$$

5. De acordo com Smyth (1989), uma distribuição pertence à família exponencial se sua f.d.p. puder ser colocada na forma

$$f(y; \theta, \phi) = \exp \left\{ \frac{w}{\phi} [y\theta - b(\theta)] + c(y, \phi) \right\},$$

sendo  $b(\cdot)$  e  $c(\cdot)$  funções conhecidas, e  $\phi > 0$ , chamado parâmetro de dispersão, e  $w$ , um peso *a priori*. Se a constante  $\phi$  é desconhecida, então, a expressão anterior define uma família exponencial com dois parâmetros apenas se

$$c(y, \phi) = -\frac{w}{\phi} g(y) - \frac{1}{2} s \left( -\frac{w}{\phi} \right) + t(y)$$

para  $g(\cdot)$ ,  $s(\cdot)$  e  $t(\cdot)$  conhecidas e, nesse caso,  $g'(\cdot)$  deve ser a inversa de  $b'(\cdot)$  tal que

$\theta = g'(\mu)$ . Mostre que isso ocorre para as distribuições normal, normal inversa e gama.

6. Seja  $Y | P \sim B(m, P)$  e  $P \sim \text{Beta}(\alpha, \beta)$ ,  $\alpha > 0$ ,  $\beta > 0$ ,  $0 < p < 1$ , isto é,

$$f(y | p) = \binom{m}{y} p^y (1-p)^{m-y} \quad \text{e} \quad f(p) = \frac{p^{\alpha-1} (1-p)^{\beta-1}}{B(\alpha, \beta)},$$

sendo  $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ . Mostre que (Hinde e Demétrio, 1998a):

a) incondicionalmente,  $Y$  tem distribuição beta-binomial com f.d.p. dada por

$$f(y) = \binom{m}{y} \frac{B(\alpha+y, m+\beta-y)}{B(\alpha, \beta)};$$

b)  $E(Y) = m \frac{\alpha}{\alpha+\beta} = m\pi$  e  $\text{Var}(Y) = m\pi(1-\pi)[1 + \rho(m-1)]$ , sendo  $\rho = \frac{1}{\alpha+\beta+1}$ ;

c) a distribuição beta-binomial não pertence à família (1.5).

7. Seja  $Y | Z = z \sim P(z)$ , isto é,

$$P(Y = y | Z = z) = \frac{e^{-z} z^y}{y!}, \quad y = 0, 1, 2, \dots$$

Então, se:

a)  $Z \sim G(k, \lambda)$ ,  $z > 0$ , isto é, com f.d.p. dada por

$$f(z; k, \lambda) = \frac{\left(\frac{\lambda}{k}\right)^\lambda}{\Gamma(\lambda)} z^{\lambda-1} \exp\left(-\frac{z\lambda}{k}\right),$$

mostre que para  $k$  fixo, incondicionalmente,  $Y$  tem distribuição binomial negativa, que pertence à família exponencial, com  $E(Y) = \frac{k}{\lambda} = \mu$  e  $\text{Var}(Y) = \mu + \frac{\mu^2}{k}$ ;

b)  $Z \sim G(r, \lambda)$ ,  $z > 0$ , isto é, com f.d.p. dada por

$$f(z; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} z^{r-1} e^{-\lambda z},$$

mostre que para  $\lambda$  fixo, incondicionalmente,  $Y$  tem distribuição binomial negativa, que não pertence à família exponencial, com  $E(Y) = \frac{r}{\lambda} = \mu$  e  $\text{Var}(Y) = \mu + \frac{\mu}{\lambda} = \phi\mu$ , sendo  $\phi = 1 + \frac{1}{\lambda}$ .

**8.** Uma forma geral para representar a função de probabilidade da distribuição binomial negativa é dada por

$$P(Y = y) = \frac{\Gamma\left(y + \frac{\mu^c}{\nu}\right)}{\Gamma\left(\frac{\mu^c}{\nu}\right) y!} \left(1 + \frac{\mu^{c-1}}{\nu}\right)^{-y} (1 + \nu\mu^{1-c})^{-\frac{\mu^c}{\nu}}, \quad y = 0, 1, 2, \dots$$

a) mostre que  $E(Y) = \mu$  e  $\text{Var}(Y) = \mu + \nu\mu^{2-c}$ . Obtenha  $E(Y)$  e  $\text{Var}(Y)$  para os casos mais comuns ( $c = 0$  e  $c = 1$ ) da distribuição binomial negativa;

b) mostre que  $P(Y = y)$  pertence à família (1.5) apenas para  $c = 0$ .

**9.** Uma distribuição para explicar o excesso de zeros em dados de contagem é a distribuição de Poisson inflacionada de zeros, com função de probabilidade dada por

$$P(Y = y) = \begin{cases} \omega + (1 - \omega)e^{-\lambda} & y = 0 \\ (1 - \omega)\frac{e^{-\lambda}\lambda^y}{y!} & y = 1, 2, \dots \end{cases}$$

Mostre que  $E(Y) = (1 - \omega)\lambda = \mu$  e  $\text{Var}(Y) = \mu + \left(\frac{\omega}{1 - \omega}\right)\mu^2$  (Ridout et al., 1998).

**10.** Uma distribuição alternativa para explicar o excesso de zeros em dados de contagem é a distribuição binomial negativa inflacionada de zeros, com função de probabilidade dada por (Ridout et al., 1998)

$$P(Y = y) = \begin{cases} \omega + (1 - \omega)(1 + \alpha\lambda^c)^{-\frac{\lambda^{1-c}}{\alpha}}, & y = 0 \\ (1 - \omega)\frac{\Gamma\left(y + \frac{\lambda^{1-c}}{\alpha}\right)}{y!\Gamma\left(\frac{\lambda^{1-c}}{\alpha}\right)}(1 + \alpha\lambda^c)^{-\frac{\lambda^{1-c}}{\alpha}}\left(1 + \frac{\lambda^{1-c}}{\alpha}\right)^{-y}, & y = 1, 2, \dots \end{cases}$$

Mostre que  $E(Y) = (1 - \omega)\lambda$  e  $\text{Var}(Y) = (1 - \omega)\lambda(1 + \omega\lambda + \alpha\lambda^c)$ .

**11.** Obtenha as funções geradoras de momentos e de cumulantes da distribuição

secante hiperbólica generalizada dada pela expressão (1.6).

**12.** Mostre que as distribuições normal, gama, normal inversa e beta pertencem à família exponencial canônica biparamétrica dada em (1.12) com  $k = 2$  e identifique  $t_1(\mathbf{x})$ ,  $t_2(\mathbf{x})$ ,  $h(\mathbf{x})$  e  $b(\boldsymbol{\theta})$ .

**13.** No Exercício 12, use as equações (1.14) para calcular  $E(\mathbf{T})$  e  $\text{Cov}(\mathbf{T})$ , sendo  $\mathbf{T} = [T_1(\mathbf{x}), T_2(\mathbf{x})]^T$ .

**14.** Usando as equações (1.14), obtenha  $E[T(\mathbf{X})]$  e  $\text{Var}[T(\mathbf{X})]$  para as 24 distribuições dadas em Cordeiro et al. (1995) na família exponencial uniparamétrica (1.1).

**15.** Demonstre as fórmulas de  $E(\mathbf{X})$ ,  $E(\mathbf{X}^{-1})$  e  $\text{Cov}(\mathbf{X}, \mathbf{X}^{-1})$  dadas no Exemplo 1.12.

**16.** Seja  $f(x; \theta) = h(x) \exp[g(x; \theta)]$  uma distribuição uniparamétrica arbitrária. Demonstre que uma condição necessária para ela não pertencer à família exponencial (1.1) é que, dados quatro pontos amostrais  $x_1$ ,  $x_2$ ,  $x_3$  e  $x_4$ , o quociente  $\frac{g(x_1, \theta) - g(x_2, \theta)}{g(x_3, \theta) - g(x_4, \theta)}$  seja uma função que depende de  $\theta$ .

**17.** Usando o Exercício 16, mostre que a distribuição de Cauchy  $f(x; \theta) = \frac{1}{\pi [1 + (x - \theta)^2]}$  não é um membro da família exponencial uniparamétrica (1.1).

**18.** Demonstre que para a família exponencial biparamétrica  $f(x; \theta, \tau) = h(x) \exp[\theta x + \tau t(x) - b(\theta, \tau)]$ , tem-se:  $E(X) = b^{(1,0)}$ ,  $\text{Var}(X) = b^{(2,0)}$ ,  $E[T(X)] = b^{(0,1)}$  e  $\text{Cov}[X, T(X)] = b^{(1,1)}$ , sendo que  $b^{(r,s)} = \frac{\partial^{(r+s)} b(\theta, \tau)}{\partial \theta^r \partial \tau^s}$ .

**19.** Considere a família exponencial multiparamétrica na forma canônica (1.12). Demonstre que os dois primeiros momentos do vetor  $\mathbf{T}$  de estatísticas suficientes são dados por (1.14).

**20.** Suponha que  $Y_1$  e  $Y_2$  têm distribuições de Poisson independentes com médias  $\mu$  e  $\rho\mu$ , respectivamente. Mostre que

- a)  $Y_+ = Y_1 + Y_2$  tem distribuição de Poisson com média  $\mu(1 + \rho)$ ;
- b)  $Y_1|Y_+ = m$  tem distribuição binomial  $B(m, (1 + \rho)^{-1})$ .

**21.** Seja  $X$  uma variável aleatória binomial  $B(m, \theta)$ .

- a) Se  $m \rightarrow \infty$  e  $\theta \rightarrow 0$  de modo que  $m\theta = \mu$  permanece constante, mostre que  $P(X = k) \rightarrow e^{-\mu} \mu^k / k!$ . Esse limite é a base da aproximação de Poisson para a distribuição binomial.
- b) Demonstre, ainda, que

$$P(X = k) \approx \frac{1}{\sqrt{2\pi m\theta(1-\theta)}} \exp \left[ -\frac{(k - m\theta)^2}{2m\theta(1-\theta)} \right].$$

**22.** Obtenha uma expressão geral para o momento central de ordem  $r$  da família de distribuições (1.5) a partir da expressão geral dos cumulantes (1.10).

**23.** Seja uma distribuição na família exponencial natural com f.d.p. ( $y > 0$ )

$$f(y; \theta) = c(y) \exp[\theta y - b(\theta)]$$

e média  $\mu = \tau(\theta)$ . Mostre que  $g(y; \theta) = yf(y; \theta)/\tau(\theta)$  é uma nova f.d.p. e calcule suas funções geratrizes de momentos e de cumulantes.

**24.** A distribuição logarítmica é definida pela função de probabilidade

$$f(y; \rho) = -\frac{\rho^y}{y \log(1 - \rho)}$$

para  $y = 1, 2, \dots$  e  $0 < \rho < 1$ . Mostre que essa distribuição pertence à família exponencial e que

$$E(Y) = \frac{\rho}{b(\rho)(1 - \rho)} \quad \text{e} \quad \text{Var}(Y) = \frac{\rho[1 - \frac{\rho}{b(\rho)}]}{b(\rho)(1 - \rho)^2},$$

em que  $b(\rho) = -\log(1 - \rho)$ .

**25.** Demonstrar as fórmulas de recorrência para os momentos ordinários  $(\mu'_r)$  e centrais  $(\mu_r)$  da distribuição binomial:

$$\mu_{r+1} = \mu(1 - \mu) \left[ mr\mu_{r-1} + \frac{d\mu_r}{d\mu} \right] \quad \text{e} \quad \mu'_{r+1} = \mu(1 - \mu) \left[ \frac{m\mu'_r}{(1 - \mu)} + \frac{d\mu'_r}{d\mu} \right].$$

**26.** Se  $X$  tem distribuição exponencial de média unitária, mostre que a função geratriz de momentos de  $Y$  é igual a  $M(t) = \Gamma(1 + t)$  e que a sua f.d.p. é  $f(y) = \exp(y - e^y)$ .

# Capítulo 2

## Modelo Linear Generalizado

### 2.1 Introdução

A seleção de modelos é uma parte importante de toda pesquisa em modelagem estatística e envolve a procura de um modelo que seja o mais simples possível e que descreva bem os dados observados que surgem em diversas áreas do conhecimento como agricultura, demografia, ecologia, economia, engenharia, geologia, medicina, ciência política, sociologia, zootecnia, entre outras.

Nelder e Wedderburn (1972) mostraram que uma série de técnicas estatísticas, comumente estudadas separadamente, podem ser formuladas, de uma maneira unificada, como uma classe de modelos de regressão. A essa teoria unificadora de modelagem estatística, uma extensão dos modelos clássicos de regressão, deram o nome de **modelos lineares generalizados** (MLG). Esses modelos envolvem uma variável resposta univariada, variáveis explanatórias e uma amostra aleatória de  $n$  observações independentes, sendo que

- i) a variável resposta, **componente aleatório** do modelo, tem uma distribuição pertencente à família de distribuições (1.5) que engloba as distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens;
- ii) as variáveis explanatórias entram na forma de uma estrutura linear, constituindo o **componente sistemático** do modelo;

- iii) a ligação entre os componentes aleatório e sistemático é feita através de uma função adequada como, por exemplo, logarítmica para os modelos log-lineares, chamada **função de ligação**.

O componente sistemático é estabelecido durante o planejamento (fundamental para a obtenção de conclusões confiáveis) do experimento, resultando em modelos de regressão (linear simples, múltipla, não linear etc.), de análise de variância (delineamentos inteiramente casualizados, blocos casualizados, quadrados latinos com estrutura de tratamentos fatorial, parcelas subdivididas etc.) e de análise de covariância. O componente aleatório é estabelecido assim que são definidas as medidas a serem feitas, que podem ser contínuas ou discretas, exigindo o ajuste de distribuições diferentes. A partir de um mesmo experimento podem ser obtidas medidas de diferentes tipos, como por exemplo, dados de altura de plantas, número de lesões por planta e proporção de plantas doentes.

No modelo clássico de regressão, tem-se

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon},$$

sendo  $\mathbf{Y}$  o vetor, de dimensões  $n \times 1$ , da variável resposta,  $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ , o componente sistemático,  $\mathbf{X}$  a matriz do modelo, de dimensões  $n \times p$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ , o vetor dos parâmetros,  $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ , o componente aleatório com  $\epsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, \dots, n$ . Nesse caso, tem-se que  $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$  e o vetor de médias  $\boldsymbol{\mu}$  da distribuição normal, que define o componente aleatório, é igual ao preditor linear que representa o componente sistemático. Essa é a forma mais simples de ligação entre esses dois componentes, sendo denominada função de ligação identidade.

Em muitos casos, porém, essa estrutura aditiva entre o componente sistemático e o erro aleatório não é satisfeita. Além disso, não há razão para se restringir à estrutura simples dada pela função de ligação identidade, nem à distribuição normal para o componente aleatório e à suposição de homogeneidade de variâncias.

Outros modelos foram surgindo e os desenvolvimentos que levaram a essa visão geral da modelagem estatística, remontam a quase dois séculos. Assim, um



MLG é definido por uma distribuição de probabilidade, membro da família (1.5) de distribuições, para a variável resposta, um conjunto de variáveis independentes descrevendo a estrutura linear do modelo e uma função de ligação entre a média da variável resposta e a estrutura linear. Entre os métodos estatísticos para a análise de dados univariados que são casos especiais dos MLG, citam-se:

- (a) modelo clássico de regressão múltipla (Legendre, Gauss, início do século XIX) e modelo de análise de variância para experimentos planejados (Fisher, 1920 a 1935) com o erro aleatório tendo distribuição normal;
- (b) modelo complemento log-log para ensaios de diluição, envolvendo a distribuição binomial (Fisher, 1922);
- (c) modelo probito (Bliss, 1935) para o estudo de proporções, envolvendo a distribuição binomial;
- (d) modelo logístico (Berkson, 1944; Dyke e Patterson, 1952; Rasch, 1960; Cox, 1970) para o estudo de proporções, envolvendo a distribuição binomial;
- (e) modelos log-lineares para análise de dados na forma de contagens em tabelas de contingência, envolvendo a distribuição de Poisson e a multinomial (Birch, 1963; Haberman, 1970);
- (f) modelo logístico para tabelas multidimensionais de proporções;
- (g) os modelos de testes de vida, envolvendo a distribuição exponencial (Feigl e Zelen, 1965; Zippin e Armitage, 1966; Gasser, 1967);
- (h) polinômios inversos para ensaios de adubação, envolvendo a distribuição normal na escala logarítmica e linearidade na escala inversa (Nelder, 1966);
- (i) modelo de análise de variância com efeitos aleatórios;
- (j) modelo estrutural para dados com distribuição gama;

(1) modelo de regressão não-simétrica

e outros modelos familiares.

Além dessas técnicas usuais, outros modelos podem ser definidos dentro do contexto dos MLG como, por exemplo, os modelos de Box e Cox (1964) e alguns modelos de séries temporais. Devido ao grande número de métodos estatísticos que engloba, a teoria dos MLG vem desempenhando um papel importante na Estatística moderna, tanto para especialistas, quanto para não-especialistas. Esses modelos podem ainda representar um meio unificado de ensino da Estatística, em qualquer curso de graduação ou pós-graduação.

Algumas referências para o estudo dos MLG e extensões são: Cordeiro (1986), McCullagh e Nelder (1989), Firth (1991), Francis et al. (1993), Fahrmeir e Tutz (1994), McCulloch e Searle (2000), Dobson (2001), Collet (2002), Paula (2004), Aitkin et al. (2005), Molenberghs e Verbeke (2005) e Lee et al. (2006).

## 2.2 Exemplos de motivação

A seguir, serão apresentados alguns dos modelos que apareceram na literatura, independentemente, e que, conforme será mostrado, podem ser agrupados de acordo com algumas propriedades comuns, o que permite um método unificado para a estimação dos parâmetros.

### a) Ensaios do tipo dose-resposta

Ensaios do tipo dose-resposta são aqueles em que uma determinada droga é administrada em  $k$  diferentes doses,  $d_1, \dots, d_k$ , respectivamente, a  $m_1, \dots, m_k$  indivíduos. Suponha que cada indivíduo responde, ou não, à droga, tal que a resposta é quantal (tudo ou nada, isto é, 1 ou 0), obtendo-se, após um período especificado,  $y_1, \dots, y_k$  indivíduos que mudam de estado (ocorrência de um sucesso). Por exemplo, quando um inseticida é aplicado a um determinado número de insetos, eles respondem (morrem), ou não (sobrevivem), à dose aplicada. Quando uma droga benéfica é administrada a um grupo de pacientes, eles podem melhorar (sucesso), ou

não (fracasso). Dados resultantes desse tipo de ensaio podem ser considerados como provenientes de uma distribuição binomial com probabilidade  $\pi_i$ , que é a probabilidade de ocorrência (sucesso) do evento sob estudo, ou seja, o número de sucessos  $Y_i$  tem distribuição binomial  $B(m_i, \pi_i)$ .

Os objetivos desse tipo de experimento são, em geral, modelar a probabilidade de sucesso  $\pi_i$  como função de variáveis explanatórias e, então, determinar doses efetivas ( $DL_p$ , doses que causam mudança de estado em 100p% dos indivíduos, por exemplo,  $DL_{50}, DL_{90}$ ), comparar potências de diferentes produtos etc.

**Exemplo 2.1:** Os dados da Tabela 2.1 referem-se a um ensaio de toxicidade de rotenone (Martin, 1942), no delineamento completamente casualizado, em que doses ( $d_i$ ) do inseticida foram aplicadas a  $m_i$  insetos (*Macrosiphoniella sanborni*, pulgão do crisântemo) e após um certo tempo foram observados os números ( $y_i$ ) de insetos mortos.

Tabela 2.1: Número de insetos mortos ( $y_i$ ) de ( $m_i$ ) insetos que receberam a dose  $d_i$  de rotenone.

Dose ( $d_i$ )	$m_i$	$y_i$	$p_i$
0,0	49	0	0,00
2,6	50	6	0,12
3,8	48	16	0,33
5,1	46	24	0,52
7,7	49	42	0,86
10,2	50	44	0,88

O interesse do pesquisador estava na determinação das doses letais que matam 50% ( $DL_{50}$ ) e 90% ( $DL_{90}$ ) dos insetos, para recomendação de aplicação do inseticida no campo. Pode-se observar que o gráfico (Figura 2.1) de dispersão das proporções ( $p_i = y_i/m_i$ ) de insetos mortos versus as doses ( $d_i$ ) tem um aspecto

sigmóide o que orienta a escolha do modelo para  $\pi_i$ .

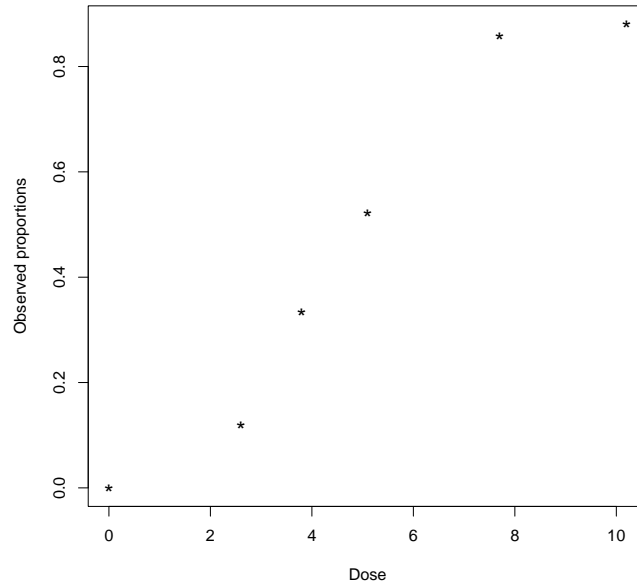


Figura 2.1: Gráfico de dispersão das proporções ( $p_i$ ) versus doses ( $d_i$ ) de rotenone, referentes à Tabela 2.1.

São dois aspectos, portanto, a serem considerados nos ensaios de dose-resposta. Um é a intensidade do estímulo que pode ser a dose de uma droga (inseticida, fungicida, herbicida, medicamento) e o outro que é o indivíduo (um inseto, um esporo, uma planta, um paciente). O estímulo é aplicado a uma intensidade especificada em unidades de concentração e como resultado uma resposta do indivíduo é obtida. Quando a resposta é binária (0 ou 1), sua ocorrência, ou não, dependerá da intensidade do estímulo aplicado. Para todo indivíduo haverá um certo nível de intensidade abaixo do qual a resposta não ocorre e acima do qual ela ocorre; na terminologia farmacológica e toxicológica, esse valor é chamado tolerância (Ashton, 1972). Essa tolerância varia de um indivíduo para outro da população e, então, há uma distribuição de tolerâncias à qual pode-se associar uma variável aleatória  $U$  com f.d.p. representada por curvas, simétricas ou assimétricas, dos tipos apresentados na Figura 2.2.

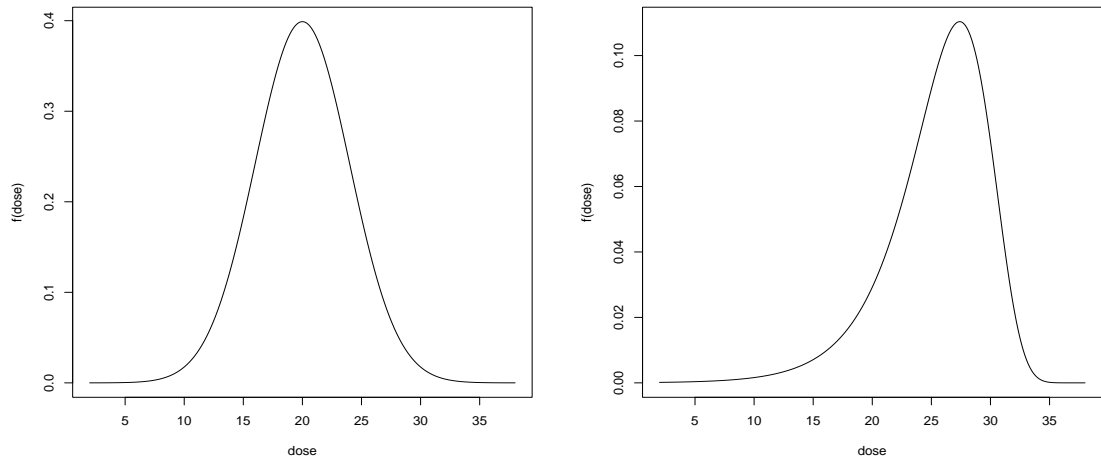


Figura 2.2: Dois tipos de curvas para distribuições de tolerância.

Se a dose  $d$  é dada para a população toda e  $f(u)$  é a função densidade para a distribuição de tolerâncias, todo indivíduo cuja tolerância é menor do que  $d$  responderá à droga, e a probabilidade de que um indivíduo escolhido ao acaso responda à dose, conforme a Figura 2.3, é dada por

$$\pi = P(U \leq d) = F(d) = \int_{-\infty}^d f(u) du. \quad (2.1)$$

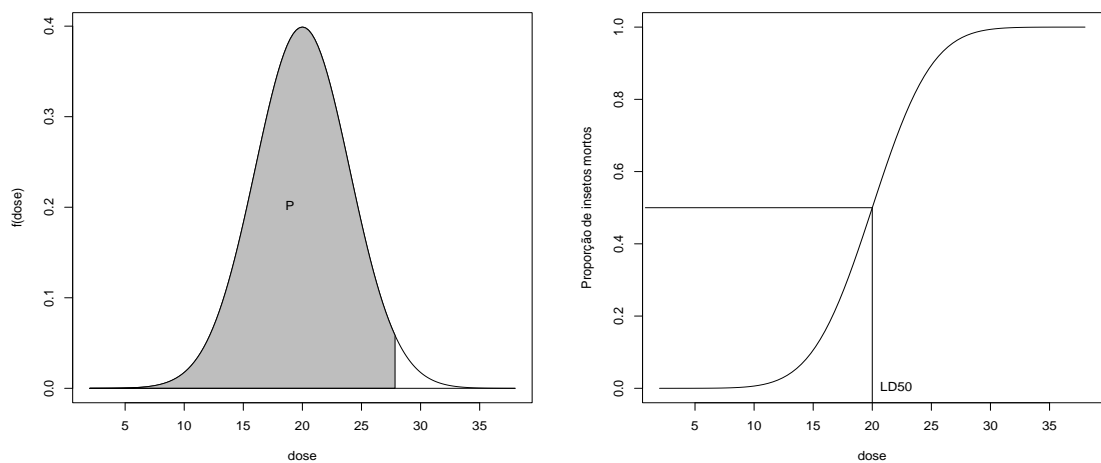


Figura 2.3: Área sob a curva de tolerância e correspondente distribuição acumulada.

A probabilidade de ocorrer uma resposta (sucesso) é tipicamente nula para valores pequenos de  $d$ , unitária para valores grandes de  $d$  (pois, então, um sucesso é certo) e é uma função estritamente crescente de  $d$ . Uma tal curva tem as propriedades matemáticas de uma função de distribuição contínua acumulada e tem a forma sigmóide típica como mostrada na Figura 2.3.

Observe-se que nenhum indivíduo responde se a dose é muito pequena e que todos os indivíduos respondem se a dose é muito grande. Essas suposições nem sempre são razoáveis. Pode haver indivíduos que respondem, naturalmente, sem a droga (morte natural) e outros que são imunes à droga, o que pode causar um excesso de zeros (Ridout et al., 1998) e uma variabilidade maior do que a esperada (superdispersão) (Hinde e Demétrio, 1998a,b).

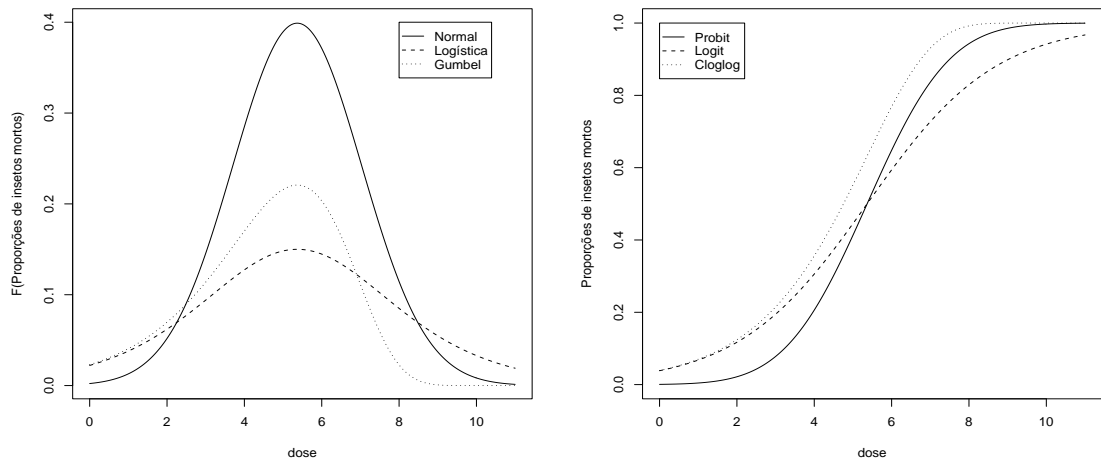


Figura 2.4: Curvas para distribuições de tolerância e correspondentes sigmóides.

O problema, então, está em se encontrar uma curva sigmóide que se ajuste bem aos dados e a partir dela obter  $DL_{50}$  e  $DL_{90}$ . O que ocorre, porém, é que são modelos não-lineares nos parâmetros e, então, a idéia é se fazer uma transformação tal que essa curva sigmóide se transforme em uma reta e, assim, procedimentos comuns de regressão possam ser usados para se estimarem os parâmetros. A Figura 2.4 mostra as distribuições, e suas correspondentes curvas sigmóides, mais

comumente usadas, cujas expressões e respectivas transformações lineares são apresentadas, a seguir.

i) **Modelo probito (“Probability unit”)**

Nesse caso, assume-se que  $U$  tem distribuição normal de média  $\mu \in \mathbb{R}$  e variância  $\sigma^2 > 0$ , isto é,

$$f_U(u; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(u - \mu)^2}{2\sigma^2} \right],$$

e, portanto, com  $Z = \frac{U - \mu}{\sigma} \sim N(0, 1)$ . Então,

$$\pi_i = P(U \leq d_i) = P \left( Z \leq -\frac{\mu}{\sigma} + \frac{1}{\sigma} d_i \right) = P(Z \leq \beta_1 + \beta_2 d_i)$$

para  $\beta_1 = -\mu/\sigma$  e  $\beta_2 = 1/\sigma$ . Logo,

$$\pi_i = \Phi(\beta_1 + \beta_2 d_i),$$

em que  $\Phi(\cdot)$  representa a função de distribuição normal padrão, é uma função não-linear em um conjunto linear de parâmetros. É linearizada por

$$\text{probit}(\pi_i) = \Phi^{-1}(\pi_i) = \beta_1 + \beta_2 d_i.$$

ii) **Modelo logístico (“Logistic unit”)**

Nesse caso, assume-se que  $U$  tem distribuição logística com parâmetros  $\mu \in \mathbb{R}$  e  $\tau > 0$ , que é similar à distribuição normal na forma, com caudas um pouco mais longas e tem f.d.p. dada por

$$f_U(u; \mu, \tau) = \frac{1}{\tau} \frac{\exp \left( \frac{u - \mu}{\tau} \right)}{\left[ 1 + \exp \left( \frac{u - \mu}{\tau} \right) \right]^2},$$

com média  $E(U) = \mu$  e variância  $\sigma^2 = \text{Var}(U) = \pi^2 \tau^2 / 3$ . Fazendo-se,  $\beta_1 = -\mu/\tau$  e  $\beta_2 = 1/\tau$ , tem-se

$$f_U(u; \beta_1, \beta_2) = \frac{\beta_2 e^{\beta_1 + \beta_2 u}}{(1 + e^{\beta_1 + \beta_2 u})^2}.$$

Logo,

$$\pi_i = P(U \leq d_i) = F(d_i) = \frac{e^{\beta_1 + \beta_2 d_i}}{1 + e^{\beta_1 + \beta_2 d_i}}$$

é uma função não-linear em um conjunto linear de parâmetros, sendo linearizada por

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_1 + \beta_2 d_i.$$

### iii) Modelo complemento log-log

Nesse caso, assume-se que  $U$  tem distribuição Gumbel de valor extremo com parâmetros  $\alpha$  e  $\tau$ , que é uma distribuição assimétrica ao contrário das duas anteriores que são simétricas, e tem f.d.p. dada por

$$f_U(u; \alpha, \tau) = \frac{1}{\tau} \exp\left(\frac{u - \alpha}{\tau}\right) \exp\left[-\exp\left(\frac{u - \alpha}{\tau}\right)\right], \quad \alpha \in \mathbb{R}, \quad \tau > 0,$$

com média  $E(U) = \alpha + \gamma\tau$  e variância  $\sigma^2 = \text{Var}(U) = \pi^2\tau^2/6$ , sendo  $\gamma \approx 0,577216$  o número de Euler que é definido por  $\gamma = -\psi(1) = \lim_{n \rightarrow \infty} (\sum_{i=1}^n i^{-1} - \log n)$ , em que  $\psi(p) = d \log \Gamma(p)/dp$  é a função digama. Fazendo-se,  $\beta_1 = -\alpha/\tau$  e  $\beta_2 = 1/\tau$ , tem-se

$$f_U(u; \beta_1, \beta_2) = \beta_2 \exp(\beta_1 + \beta_2 u - e^{\beta_1 + \beta_2 u}).$$

Logo,

$$\pi_i = P(U \leq d_i) = F(d_i) = 1 - \exp[-\exp(\beta_1 + \beta_2 d_i)]$$

é uma função não-linear em um conjunto linear de parâmetros e é linearizada por

$$\log[-\log(1 - \pi_i)] = \beta_1 + \beta_2 d_i.$$

Então, esses três exemplos têm em comum

i) a distribuição dos  $Y_i$  (binomial) é um membro da família exponencial, com

$$E(Y_i) = \mu_i = m_i \pi_i;$$

ii) as variáveis explanatórias entram na forma de uma soma linear de seus efeitos sistemáticos, ou seja,

$$\eta_i = \sum_{j=1}^2 x_{ij} \beta_j = \mathbf{x}_i^T \boldsymbol{\beta},$$

sendo  $\mathbf{x}_i^T = (1, d_i)$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$  e  $\eta_i$  o preditor linear.



iii) a média  $\mu_i$  é funcionalmente relacionada ao preditor linear, isto é,

$$\eta_i = g\left(\frac{\mu_i}{m_i}\right) = g(\pi_i)$$

que nos casos analisados foram:

modelo probito:  $\eta_i = g(\pi_i) = \Phi^{-1}(\pi_i)$ ;

modelo logístico:  $\eta_i = g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$ ;

modelo complemento log-log:  $\eta_i = g(\pi_i) = \log[-\log(1 - \pi_i)]$ .

Portanto, tem-se que esses modelos são baseados na família exponencial uniparamétrica (1.2) com médias que são não-lineares em um conjunto de parâmetros lineares, isto é,

modelo probito:  $\mu_i = m_i \Phi(\beta_1 + \beta_2 d_i)$ ;

modelo logístico:  $\mu_i = m_i \frac{e^{\beta_1 + \beta_2 d_i}}{1 + e^{\beta_1 + \beta_2 d_i}}$ ;

modelo complemento log-log:  $\mu_i = m_i \{1 - \exp[-\exp(\beta_1 + \beta_2 d_i)]\}$ .

## b) Ensaios de diluição

É prática comum, o uso dos ensaios de diluição para se estimar a concentração  $\lambda$  de um organismo (número por unidade de volume, de área, de peso etc.) em uma amostra. Quando a contagem direta não é possível, mas a presença ou ausência do organismo em sub-amostras pode ser detectada (Ridout e Fenlon, 1998) pode-se, também, estimar  $\lambda$ . Em geral, registrar a presença, ou ausência, fica mais econômico do que fazer a contagem. Por exemplo, pode-se detectar se uma determinada bactéria está presente, ou não, em um líquido por um teste de cor, ou se um fungo está presente, ou não, em uma amostra de solo, plantando-se uma planta susceptível nesse solo e verificando se a planta apresenta sintomas da doença. Esse método está baseado na suposição de que o número de indivíduos presentes segue

uma distribuição de Poisson, o que é uma suposição forte e é importante verificar se é verdadeira. Por exemplo, a distribuição espacial de um fungo no solo está longe de ser aleatória e pode ser que o número de indivíduos em diferentes amostras desse solo não siga a distribuição de Poisson.

Nos ensaios de diluição, a solução original é diluída progressivamente e na  $i$ -ésima diluição são feitas as contagens (Exemplo 2.2) ou, então, são testadas  $m_i$  sub-amostras das quais  $Y_i$  apresentam resultado positivo para a presença do organismo (Exemplo 2.3). Seja  $\nu_i$  o volume da amostra original que está presente em cada uma das sub-amostras na  $i$ -ésima diluição. Em geral, mas nem sempre, são usadas diluições iguais, tal que os  $\nu_i$ 's ficam em progressão geométrica.

**Exemplo 2.2:** A Tabela 2.2 mostra os dados referentes a contagens de partículas de vírus para cinco diluições diferentes, sendo que foram usadas quatro repetições para as quatro primeiras diluições e cinco repetições para a última diluição. O objetivo do experimento era estimar o número de partículas de vírus por unidade de volume.

Tabela 2.2: Números de partículas de vírus para cinco diluições diferentes.

Diluição	Contagens				
0,3162	13	14	17	22	
0,1778	9	14	6	14	
0,1000	4	4	3	5	
0,0562	3	2	1	3	
0,0316	2	1	3	2	2

Fonte: Ridout (1990), notas de aula

**Exemplo 2.3:** A Tabela 2.3 mostra os dados de um ensaio de diluição realizado para determinar o número de esporos de *Bacillus mesentericus* por grama (g) de farinha de batata (Fisher e Yates, 1970). Uma suspensão líquida foi preparada e sujeita a sucessivas diluições para que resultassem soluções com 4, 2, ..., 1/128g de farinha

por 100ml de solução. Para cada diluição foram tomadas cinco amostras de 1ml e foi contado o número de amostras com esporos.

Tabela 2.3: Números de amostras ( $Y$ ) que contêm esporos em cinco amostras, para diferentes quantias (g) de farinha de batata em cada diluição.

g/100 ml	4	2	1	1/2	1/4	1/8	1/16	1/32	1/64	1/128
y	5	5	5	5	4	3	2	2	0	0

O parâmetro de interesse é  $\lambda$ , a concentração de organismos por unidade de volume ( $\nu_i$ ). Se os organismos estão aleatoriamente distribuídos, o número de organismos em uma sub-amostra da  $i$ -ésima diluição segue a distribuição de Poisson com média  $\lambda\nu_i$ , isto é,

$$\mu_i = \lambda\nu_i.$$

Assim, se forem feitas contagens dos indivíduos após a diluição, tem-se que essa expressão, pode ser linearizada, usando-se a função logarítmica, ou seja,

$$\eta_i = \log(\mu_i) = \log(\lambda) + \log(\nu_i) = \beta_1 + \text{offset}, \quad (2.2)$$

em que  $\log(\nu_i)$  entra na regressão como variável *offset* que é um valor conhecido na equação.

Quando se observa o número de amostras em que o indivíduo está presente tem-se  $Y_i \sim B(m_i, \pi_i)$ , desde que as sub-amostras de cada diluição sejam independentes, sendo que a probabilidade  $\pi_i$  de que o organismo esteja presente na sub-amostra  $i$  é dada por

$$\pi_i = P(\text{pelo menos um organismo presente}) = 1 - \exp(-\lambda\nu_i).$$

Logo,

$$\eta_i = \log[-\log(1 - \pi_i)] = \log(\lambda) + \log(\nu_i) = \beta_1 + \text{offset}. \quad (2.3)$$

Tem-se, em (2.2) e (2.3), que  $\beta_1 = \log(\lambda)$  e  $\log(\nu_i)$  é a variável *offset*. Além disso, para (2.2) tem-se a função de ligação logarítmica para o modelo de Poisson enquanto que para (2.3) tem-se a função de ligação complemento log-log para o modelo binomial.

Esse método de diluição em série é muito utilizado em diversas áreas da Biologia. Podem ser tratados de forma semelhante os problemas de estimação de:

- a) proporção de sementes doentes em um lote de sementes, em que  $n$  é o tamanho da amostra de sementes,  $\theta$  é a probabilidade de uma semente infectada e

$$\pi = P(\text{pelo menos uma semente doente}) = 1 - (1 - \theta)^n = 1 - e^{n \log(1-\theta)};$$

- b) proporção de um determinado tipo de célula em uma população em estudos de imunologia;
- c) probabilidade de uma partícula de vírus matar um inseto, nos ensaios de controle biológico;
- d) taxa média de falha de um determinado componente quando os tempos de falha são distribuídos exponencialmente.

Nesse exemplo, verifica-se, novamente, que:

- i) a distribuição dos  $Y_i$  (Poisson ou binomial) é um membro da família exponencial uniparamétrica (1.2), com  $E(Y_i) = \mu_i$  (Poisson) ou  $E(Y_i) = \mu_i = m_i \pi_i$  (binomial);
- ii) as variáveis explanatórias entram na forma de uma soma linear de seus efeitos, ou seja,

$$\eta_i = \sum_{j=1}^2 x_{ij} \beta_j = \mathbf{x}_i^T \boldsymbol{\beta},$$

sendo  $\mathbf{x}_i = (1, d_i)^T$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$  e  $\eta_i$  o preditor linear.

iii) a média  $\mu_i$  é funcionalmente relacionada ao preditor linear, isto é,

$$\eta_i = g(\mu_i) \quad \text{ou} \quad \eta_i = g\left(\frac{\mu_i}{m_i}\right) = g(\pi_i)$$

que nos casos analisados foram:

modelo log-linear:  $\eta_i = g(\mu_i) = \log \mu_i$ ;

modelo complemento log-log:  $\eta_i = g(\pi_i) = \log[-\log(1 - \pi_i)]$ .

Portanto, esses modelos são baseados na família exponencial uniparamétrica (1.2), cujas médias são não-lineares em um conjunto de parâmetros lineares, isto é,

modelo log-linear:  $\mu_i = e^{\beta_1 + \text{offset}}$ ;

modelo complemento log-log:  $\mu_i = m_i\{1 - \exp[-\exp(\beta_1 + \text{offset})]\}$ ,

sendo  $\beta_2 = 1$  e  $\log(\nu_i) = \text{offset}$ .

### c) Tabelas de contingência

Dados na forma de contagens são oriundos da simples contagem de eventos (por exemplo, número de brotos por explante), ou então, da frequência de ocorrências em várias categorias e que dão origem às tabelas de contingência. Sejam os exemplos que se seguem.

**Exemplo 2.4:** Os dados da Tabela 2.4 referem-se a coletas de insetos em armadilhas adesivas de duas cores, em que os indivíduos coletados de uma determinada espécie foram sexados, tendo como objetivo verificar se havia influência da cor da armadilha sobre a atração de machos e fêmeas dessa espécie.

Tem-se que o número de insetos que chegam às armadilhas, seja do sexo feminino ou do sexo masculino, é um número aleatório, caracterizando uma observação de uma variável com distribuição de Poisson. A hipótese de interesse é a hipótese de independência, isto é, o sexo do inseto não afeta a escolha pela cor da armadilha.

Tabela 2.4: Números de insetos coletados em armadilhas adesivas e sexados.

Armadilha	Machos	Fêmeas	Totais
Alaranjada	246	17	263
Amarela	458	32	490
Totais	704	49	753

Fonte: Silveira Neto et al. (1976)

**Exemplo 2.5:** Os dados da Tabela 2.5 referem-se a um ensaio de controle de brocas do fruto do tomateiro através de quatro tratamentos. Tem-se aqui, também, um

Tabela 2.5: Números de frutos de tomateiro sadios e com broca.

Inseticidas	Frutos		Totais
	Sadios	Com broca	
Diazinon	1690	115	1805
Phosdrin	1578	73	1651
Sevin	2061	53	2114
Testemunha	1691	224	1915
Totais	7020	465	7485

Fonte: Silveira Neto et al. (1976)

caso em que o número total de frutos com broca é uma variável aleatória e, portanto, pode ser estudada pela distribuição de Poisson. A hipótese a ser testada é a da homogeneidade, isto é, a proporção de frutos sadios é a mesma para todos os inseticidas.

A distribuição de Poisson é especialmente útil na análise de tabelas de contingência em que as observações consistem de contagens ou frequências nas caselas pelo cruzamento das variáveis resposta e explanatórias.

Considerando-se uma tabela de contingência bidimensional e a hipótese de

independência, se  $y_{ij}$  representa o número de observações numa classificação cruzada de dois fatores  $i$  e  $j$  com  $I$  e  $J$  níveis, respectivamente, para  $i = 1, \dots, I$  e  $j = 1, \dots, J$ , então,

$$\mu_{ij} = E(Y_{ij}) = m\pi_{i+}\pi_{+j},$$

em que  $m = \sum_{i=1}^I \sum_{j=1}^J y_{ij}$  e  $\pi_{i+} = \sum_{j=1}^J \pi_{ij}$  e  $\pi_{+j} = \sum_{i=1}^I \pi_{ij}$  são as probabilidades marginais de uma observação pertencer às classes  $i$  e  $j$ , respectivamente. Pode-se, então, supor que  $Y_{ij}$  tem distribuição de Poisson com média  $\mu_{ij}$ .

Vê-se, então, que uma função logarítmica lineariza esse modelo, isto é,

$$\eta_{ij} = \log(\mu_{ij}) = \log(m) + \log(\pi_{i+}) + \log(\pi_{+j}) = \mu + \alpha_i + \beta_j.$$

Novamente, tem-se:

- i) a distribuição de  $Y_{ij}$  (Poisson) é um membro da família exponencial, com  $E(Y_{ij}) = \mu_{ij}$ ;
- ii) as variáveis explanatórias entram na forma de uma soma linear de seus efeitos, ou seja,

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

sendo  $\boldsymbol{\eta} = (\eta_{11}, \dots, \eta_{1J}, \dots, \eta_{I1}, \dots, \eta_{IJ})^T$  o preditor linear,  $\mathbf{X}$  uma matriz, de dimensões  $IJ \times (I + J + 1)$ , de variáveis “dummy” e  $\boldsymbol{\beta} = (\mu, \alpha_1, \dots, \alpha_I, \beta_1, \dots, \beta_J)^T$ ;

- iii) a média é funcionalmente relacionada ao preditor linear, isto é,

$$\eta_{ij} = g(\mu_{ij}) = \log \mu_{ij}.$$

Portanto, tem-se que esses modelos são baseados na família exponencial uniparamétrica (1.2), cujas médias são não-lineares em um conjunto de parâmetros lineares, ou seja,  $\boldsymbol{\mu} = \exp(\boldsymbol{\eta}) = \exp(\mathbf{X}^T \boldsymbol{\beta})$ .

De forma semelhante, pode ser verificado que, em geral, para dados colocados em tabelas de contingência, as hipóteses mais comuns podem ser expressas

como modelos multiplicativos para as frequências esperadas das caselas (McCullagh e Nelder, 1989; Agresti, 2002; Paulino e Singer, 2006). Verifica-se, então, que na análise de dados categorizados, de uma forma geral, a média  $\mu$  é obtida como um produto de outras médias marginais. Isso sugere que uma transformação logarítmica do valor esperado lineariza essa parte do modelo (daí vem o nome de modelo log-linear).

## 2.3 Definição

Os modelos lineares generalizados podem ser usados quando se tem uma única variável aleatória  $Y$  associada a um conjunto de variáveis explanatórias  $x_1, \dots, x_p$ . Para uma amostra de  $n$  observações  $(y_i, \mathbf{x}_i)$  em que  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  é o vetor coluna de variáveis explicativas, o MLG envolve os três componentes:

- i) **Componente aleatório:** representado por um conjunto de variáveis aleatórias independentes  $Y_1, \dots, Y_n$  provenientes de uma mesma distribuição que faz parte da família de distribuições (1.5) com médias  $\mu_1, \dots, \mu_n$ , ou seja,

$$E(Y_i) = \mu_i, \quad i = 1, \dots, n,$$

sendo  $\phi > 0$  um parâmetro de dispersão e o parâmetro  $\theta_i$  denominado parâmetro canônico. Então, a f.d.p. de  $Y_i$  é dada por

$$f(y_i; \theta_i, \phi) = \exp \left\{ \phi^{-1} [y_i \theta_i - b(\theta_i)] + c(y_i, \phi) \right\}, \quad (2.4)$$

sendo  $b(\cdot)$  e  $c(\cdot)$  funções conhecidas. Conforme foi visto na Seção 1.4

$$E(Y_i) = \mu_i = b'(\theta_i) \quad \text{e} \quad \text{Var}(Y_i) = \phi b''(\theta_i) = \phi V_i,$$

em que  $V_i = V(\mu_i) = d\mu_i/d\theta_i$  é denominada de função de variância e depende unicamente da média  $\mu_i$ . O parâmetro natural  $\theta_i$  pode ser expresso como

$$\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i), \quad (2.5)$$



sendo  $q(\mu_i)$  uma função conhecida da média  $\mu_i$ . Dada uma relação funcional para a função de variância  $V(\mu)$ , o parâmetro canônico é obtido da equação (2.5) e a distribuição fica determinada na família exponencial (2.4). A importância da família (2.4) na teoria dos MLG é que ela permite incorporar dados que exibem assimetria, dados de natureza discreta ou contínua e dados que são restritos a um intervalo do conjunto dos reais, como o intervalo (0,1).

- ii) **Componente sistemático:** as variáveis explicativas entram na forma de uma soma linear de seus efeitos

$$\eta_i = \sum_{r=1}^p x_{ir}\beta_r = \mathbf{x}_i^T \boldsymbol{\beta} \quad \text{ou} \quad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad (2.6)$$

sendo  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$  a matriz do modelo,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  o vetor de parâmetros e  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^T$  o preditor linear. Se um parâmetro tem valor conhecido, o termo correspondente na estrutura linear é chamado *offset*, como visto nos ensaios de diluição (Seção 2.2).

- iii) **Função de ligação:** uma função que **relaciona** o componente aleatório ao componente sistemático, ou seja, vincula a média ao preditor linear, isto é,

$$\eta_i = g(\mu_i), \quad (2.7)$$

sendo  $g(\cdot)$  uma função monótona e diferenciável.

Assim, vê-se que para a especificação do modelo, os parâmetros  $\theta_i$  da família de distribuições (2.4) não são de interesse direto (pois há um para cada observação) mas sim um conjunto menor de parâmetros  $\beta_1, \dots, \beta_p$  tais que uma combinação linear dos  $\beta$ 's seja igual a alguma função do valor esperado de  $Y_i$ . Como o parâmetro natural  $\theta_i$  é uma função unívoca da média  $\mu_i$ , pode-se expressar a função de ligação em termos desse parâmetro, isto é,  $\eta_i = g(q^{-1}(\theta_i))$ .

Portanto, uma decisão importante na escolha do MLG é definir os termos do trinômio: (i) distribuição da variável resposta; (ii) matriz do modelo e (iii) função

de ligação. Nesses termos, um MLG é definido por uma distribuição da família (2.4), uma estrutura linear (2.6) e uma função de ligação (2.7). Por exemplo, quando  $\theta = \mu$  e a função de ligação é linear, obtém-se o modelo clássico de regressão como um caso particular. Os modelos log-lineares são deduzidos supondo  $\theta = \log \mu$  com função de ligação logarítmica  $\log \mu = \eta$ . Torna-se clara, agora, a palavra “generalizado”, significando uma distribuição mais ampla do que a normal para a variável resposta, e uma função não-linear em um conjunto linear de parâmetros conectando a média dessa variável com a parte determinística do modelo.

Observe-se que na definição de um MLG por (2.4), (2.6) e (2.7) não existe, em geral, aditividade entre a média  $\mu$  e o erro aleatório  $\epsilon$ , como ocorre no modelo clássico de regressão descrito na Seção 2.1. Define-se no MLG uma distribuição para a variável resposta que representa as observações e não uma distribuição para o erro aleatório  $\epsilon$ .

A escolha da distribuição em (2.4) é, usualmente, feita pela natureza dos dados (discreta ou contínua) e pelo seu intervalo de variação (conjunto dos reais, reais positivos ou um intervalo como  $(0,1)$ ). Na escolha da matriz do modelo  $\mathbf{X} = \{x_{ir}\}$ , de dimensões  $n \times p$  e suposta de posto completo,  $x_{ir}$  pode representar a presença ou ausência de um nível de um fator classificado em categorias, ou pode ser o valor de uma covariável quantitativa. A forma da matriz do modelo representa matematicamente o desenho do experimento. A escolha da função de ligação depende do problema em particular e, pelo menos em teoria, cada observação pode ter uma função de ligação diferente.

As funções de ligação usuais são: potência  $\eta = \mu^\lambda$  em que  $\lambda$  é um número real, logística  $\eta = \log[\mu/(m - \mu)]$ , probito  $\eta = \Phi^{-1}(\mu/m)$  sendo  $\Phi(\cdot)$  a função de distribuição acumulada (f.d.a.) da distribuição normal padrão e a complemento log-log  $\eta = \log[-\log(1 - \mu/m)]$ , em que  $m$  é o número de ensaios independentes. As três últimas funções de ligação são apropriadas para o modelo binomial, pois transformam o intervalo  $(0,1)$  em  $(-\infty, +\infty)$  (Exercício 1.1). Casos importantes da função de ligação potência são identidade, recíproca, raiz quadrada e logarítmica,

correspondentes, a  $\lambda = 1, -1, 1/2$  e  $0$ , respectivamente.

Se a função de ligação é escolhida de tal forma que  $g(\mu_i) = \theta_i = \eta_i$ , o preditor linear modela diretamente o parâmetro canônico  $\theta_i$ , sendo denominada função de ligação canônica. Os modelos correspondentes são denominados **canônicos**. Isso resulta, freqüentemente, em uma escala adequada para a modelagem com interpretação prática para os parâmetros de regressão, além de vantagens teóricas em termos da existência de um conjunto de estatísticas suficientes para o vetor de parâmetros  $\beta$  e alguma simplificação no algoritmo de estimação. A estatística suficiente para  $\beta$  é  $\mathbf{T} = \mathbf{X}^T \mathbf{Y}$ , com componentes  $T_r = \sum_{i=1}^n x_{ir} Y_i$ ,  $r = 1, \dots, p$ . As funções de ligação canônicas para as principais distribuições estão apresentadas na Tabela 2.6.

Tabela 2.6: Funções de ligação canônicas.

Distribuição	Função de ligação canônica
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \log \mu$
Binomial	Logística: $\eta = \log\left(\frac{\pi}{1-\pi}\right) = \log\left(\frac{\mu}{m-\mu}\right)$
Gama	Recíproca: $\eta = \frac{1}{\mu}$
Normal Inversa	Recíproca do quadrado: $\eta = \frac{1}{\mu^2}$

Deve ser lembrado, porém, que embora as funções de ligação canônicas conduzam a propriedades estatísticas desejáveis para o modelo, principalmente, no caso de amostras pequenas, não há nenhuma razão a priori para que os efeitos sistemáticos do modelo devam ser aditivos na escala dada por tais funções. Para o modelo clássico de regressão, a função de ligação canônica é a identidade, pois o preditor linear é igual à média. Essa função de ligação é adequada no sentido em que ambos,  $\eta$  e  $\mu$ , podem assumir valores na reta real. Entretanto, certas restrições surgem quando se trabalha, por exemplo, com a distribuição de Poisson em que  $\mu > 0$  e, portanto, a função de ligação identidade não deve ser usada, pois  $\hat{\mu}$  poderá assumir valores

negativos, dependendo dos valores obtidos para  $\hat{\beta}$ . Além disso, dados de contagem dispostos em tabelas de contingência, sob a suposição de independência, conduzem, naturalmente, a efeitos multiplicativos cuja linearização pode ser obtida através da função de ligação logarítmica, isto é,  $\eta = \log \mu$  e, portanto,  $\mu = e^\eta$  (conforme visto nos ensaios de diluição descritos na Seção 2.2).

Aranda-Ordaz (1981) propôs a família de funções de ligação para análise de dados na forma de proporções dada por

$$\eta = \log \left[ \frac{(1 - \pi)^{-\lambda} - 1}{\lambda} \right],$$

sendo  $\lambda$  uma constante desconhecida e que tem como casos particulares as funções de ligação logística para  $\lambda = 1$  e complemento log-log quando  $\lambda \rightarrow 0$ .

Uma família importante de funções de ligação, principalmente para dados com média positiva, é a família potência (Exercício 2), especificada por

$$\begin{cases} \frac{\mu^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log \mu & \lambda = 0 \end{cases}$$

ou então,

$$\begin{cases} \mu^\lambda & \lambda \neq 0 \\ \log \mu & \lambda = 0 \end{cases}$$

sendo  $\lambda$  uma constante desconhecida.

## 2.4 Modelos especiais

### 2.4.1 Modelo clássico de regressão

A distribuição normal foi deduzida, por Laplace, em 1774, como uma aproximação para a distribuição hipergeométrica. Em 1778, Laplace tabulou a f.d.a.  $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^x e^{-t^2/2} dt$  da distribuição normal padronizada. Gauss, em dois artigos publicados em 1809 e 1816, estabeleceu técnicas baseadas na distribuição normal que se tornaram métodos corriqueiros durante o século XIX. No seu artigo de 1816, Gauss deduziu a distribuição normal como a distribuição limite da soma de

um número muito grande de erros independentes, podendo assim ser considerado um dos resultados mais antigos do teorema central do limite. A f.d.p. da distribuição normal está dada na Seção 1.3 (Exemplo 1.4).

A função geratriz de momentos da distribuição normal é  $M(t; \mu, \sigma^2) = \exp(\mu t + \sigma^2 t^2/2)$ , sendo, então, seus cumulantes  $\kappa_r = 0$ , para  $r > 2$ . Entre outras características, citam-se: média, moda e mediana iguais a  $\mu$ , coeficientes de assimetria e curtose iguais a 0 e 3, respectivamente,  $r$ -ésimo momento central igual a 0 se  $r$  é ímpar, e  $\frac{\sigma^r r!}{2^{r/2}} \left(\frac{r}{2}\right)!$ , se  $r$  é par.

Existem várias aproximações para calcular a f.d.a.  $\Phi(x)$  da distribuição normal padronizada que podem ser encontradas em Johnson et al. (2004).

As origens do modelo clássico de regressão estão nos trabalhos de astronomia de Gauss em 1809 e 1821. O método de mínimos quadrados foi desenvolvido por Legendre em 1805 e por Gauss em 1809 para determinar a órbita do asteroide Ceres. As idéias de obtenção da matriz modelo nos planejamentos dos experimentos surgiram na Estação Experimental de Rothamsted, Inglaterra, com Fisher (1920 a 1935).

O modelo normal  $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$  para o vetor  $\mathbf{Y}$  da variável resposta, em que  $\mathbf{I}$  é a matriz identidade, é usado na análise de variância com efeitos fixos, como modelo amostral e, mais comumente, como um modelo aproximado para uma distribuição desconhecida. É o caso mais simples do MLG correspondendo a  $\eta = \theta = \mu$ .

Embora a estimação por máxima verossimilhança seja estudada na Seção 3.2, convém salientar que no modelo clássico de regressão, o estimador de máxima verossimilhança de  $\boldsymbol{\beta}$ , que coincide com o de mínimos quadrados, é dado em forma explícita por  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ . A função de verossimilhança só depende dos dados através de  $\hat{\boldsymbol{\beta}}$  e da soma dos quadrados dos resíduos  $SQR = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ . Sabe-se que  $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$  e  $SQR \sim \sigma^2 \chi_{n-p}^2$ . Os testes para componentes de  $\boldsymbol{\beta}$  são realizados, exatamente, através das estatísticas com distribuições  $\chi^2$  e  $F$ .

### 2.4.2 Modelo de Poisson

Em 1837, Poisson publicou a distribuição que leva seu nome, obtendo-a como uma distribuição limite da distribuição binomial. Se a variável aleatória  $Y$  tem distribuição de Poisson,  $P(\mu)$ , com parâmetro  $\mu > 0$ , então sua função de probabilidade é expressa como

$$f(y; \mu) = \frac{e^{-\mu} \mu^y}{y!}, \quad \text{para } y = 0, 1, 2, \dots$$

A função geratriz de momentos é dada por  $M(t; \mu) = \exp\{\mu[\exp(t) - 1]\}$ , sendo todos os cumulantes iguais a  $\mu$  e o  $r$ -ésimo momento central  $\mu_r$  ( $r \geq 2$ ) pode ser obtido pela fórmula de recorrência  $\mu_r = \mu \sum_{i=0}^{r-1} \binom{r-1}{i} \mu_i$ , com  $\mu_0 = 1$ . A moda corresponde ao maior inteiro menor do que  $\mu$ , e para  $\mu$  inteiro, assume os valores  $\mu$  e  $\mu - 1$ . Os coeficientes de assimetria e curtose são iguais a  $\mu^{-1/2}$  e  $3 + \mu^{-1}$ , respectivamente. O  $r$ -ésimo momento fatorial é  $E[Y(Y-1)\dots(Y-r+1)] = \mu^r$ .

Quando  $\mu \rightarrow \infty$ , tem-se  $(Y - \mu)\mu^{-1/2} \sim N(0, 1) + O_p(\mu^{-1/2})$ . Em geral, para  $\mu > 9$ , a aproximação da distribuição de Poisson  $P(\mu)$  pela distribuição normal  $N(\mu, \mu)$  é satisfatória. Probabilidades individuais podem ser computadas pela expressão aproximada  $P(Y = y) = \Phi(y_2) - \Phi(y_1)$ , sendo que  $\Phi(\cdot)$  é a f.d.a. da distribuição normal padronizada,  $y_2 = (y - \mu + 0.5)\mu^{-1/2}$  e  $y_1 = (y - \mu - 0.5)\mu^{-1/2}$ . O resultado bastante conhecido  $P(Y \leq y) = P(\chi_{2(1+y)}^2 > 2\mu)$  é, muitas vezes, útil no cálculo da função de distribuição acumulada de Poisson.

Uma fórmula alternativa aproximada para calcular a distribuição acumulada da Poisson, baseada na distribuição acumulada da distribuição normal padrão, é  $P(Y \leq y) \approx \Phi[g(y - 0.5)]$ , em que

$$g(y) = \begin{cases} 3y^{1/2} - 3y^{1/6}\mu^{1/3} + \mu^{-1/2}/6, & y \neq 0; \\ -(2\mu)^{1/2} + \mu^{-1/2}/6, & y = 0. \end{cases}$$

O modelo de Poisson tem um importante papel na análise de dados em forma de contagens. Suas características principais são:

- a) proporciona, em geral, uma descrição satisfatória de dados experimentais cuja variância é proporcional à média;

- b) pode ser deduzido teoricamente de princípios elementares com um número mínimo de restrições;
- c) se eventos ocorrem independente e aleatoriamente no tempo, com taxa média de ocorrência constante, o modelo determina o número de eventos, em um intervalo de tempo especificado.

O modelo de regressão de Poisson desempenha na análise de dados categorizados, o mesmo papel do modelo normal, na análise de dados contínuos. A diferença fundamental é que a estrutura multiplicativa para as médias do modelo de Poisson é mais apropriada do que a estrutura aditiva das médias do modelo normal. Tem-se constatado, na análise de dados categorizados, que a média  $\mu$  é, geralmente, obtida como um produto de outras médias marginais que se tornam os parâmetros lineares do modelo. A estrutura linear adotada é dada, na escala do parâmetro canônico da distribuição, por  $\log \mu = \eta$ , com os parâmetros  $\beta$ 's medindo efeitos sobre a escala logarítmica das frequências esperadas. Por exemplo, independência de dois fatores numa tabela de contingência  $r \times s$  equivale ao modelo  $\mu_{ij} = \mu_{i+}\mu_{+j}/\mu_{++}$ , com a notação usual para a soma, e isso implica, que o logaritmo de  $\mu_{ij}$  é expresso como uma estrutura linear formada pelos efeitos principais dos fatores sem a interação.

O modelo log-linear é definido pela distribuição de Poisson,  $P(\mu)$ , com  $\log(\mu) = \eta = \mathbf{X}\beta$ , sendo um dos casos especiais de MLG de maior importância, pelo seu papel na análise de dados categorizados dispostos em tabelas de contingência. Pode-se supor que a tabela de contingência é proveniente de um modelo de Poisson, multinomial ou produto-multinomial, dependendo do planejamento feito. Para os dois últimos modelos, demonstra-se que isso equivale a um conjunto de distribuições condicionadas de Poisson com a condição do total das frequências observadas ser fixo (Seção 7.2.2).

Podem-se transformar  $Y$  na forma de contagens e definir modelos alternativos para os dados transformados. Geralmente, usa-se a transformação  $Y^{1/2}$  que estabiliza a variância supondo  $\mu$  grande, ou trata-se  $Y^{2/3}$  como, aproximadamente,

normal. Entretanto, ao se fazer isso, ignora-se a natureza discreta dos dados.

### 2.4.3 Modelo binomial

A distribuição binomial foi deduzida por James Bernoulli em 1713, embora tenha sido encontrada anteriormente em trabalhos de Pascal.

Suponha que  $Y = mP$  tenha distribuição binomial  $B(m, \pi)$ , com função de probabilidade dada no Exemplo 1.2, sendo que  $P$  representa a proporção de sucessos em  $m$  ensaios independentes com probabilidade de sucesso  $\pi$ . A função geratriz de momentos de  $Y$  é dada por  $M(t; \pi, m) = \{\pi[\exp(t) - 1] + 1\}^m$  e os seus momentos centrais,  $\mu_{2r}$  e  $\mu_{2r+1}$ , são  $O(m^r)$ , para  $r = 1, 2, \dots$ . O  $r$ -ésimo momento central de  $P$  é, simplesmente,  $m^{-r}\mu_r$ . Todos os cumulantes de  $Y$  são de ordem  $O(m)$  e, portanto,

$$\frac{Y - m\pi}{[m\pi(1 - \pi)]^{1/2}} \sim N(0, 1) + O_p(m^{-1/2}),$$

sendo a taxa de convergência dada pelo terceiro cumulante padronizado. A moda de  $Y$  pertence ao intervalo  $[(m + 1)\pi - 1, (m + 1)\pi]$ , e os seus coeficientes de assimetria e curtose são, respectivamente,

$$\frac{(1 - 2\pi)}{[m\pi(1 - \pi)]^{1/2}} \text{ e } 3 - \frac{6}{m} + \frac{1}{m\pi(1 - \pi)}.$$

Quando  $m\pi > 5$  e  $0,1 \leq \pi \leq 0,9$ , ou  $m\pi > 25$ , sendo  $\pi$  qualquer, o modelo binomial  $B(m, \pi)$  pode ser aproximado pelo modelo normal  $N(m\pi, m\pi(1 - \pi))$ . Uma melhor aproximação é dada por  $P(Y \leq y) = \Phi(y_1) + \Phi'(y_1)/\{2[m\pi(1 - \pi)]^{1/2}\}$ , em que  $y_1 = (y - m\pi)/[m\pi(1 - \pi)]$  e  $\Phi'(\cdot)$  é a f.d.p. da distribuição normal padrão, com erro inferior a  $(0,2 + 0,25 |1 - 2\pi|)/[m\pi(1 - \pi)] + \exp\{-1,5[m\pi(1 - \pi)]^{-1/2}\}$ , se  $m\pi(1 - \pi) \geq 25$ . A aproximação normal com correção de continuidade  $P(Y \leq y) = \Phi(y_2)$ , em que  $y_2 = (y + 0,5 - m\pi)/[m\pi(1 - \pi)]^{1/2}$ , tem erro menor do que  $0,140[m\pi(1 - \pi)]^{-1/2}$  (Cordeiro, 1986).

Se  $y = mp$  é inteiro, um número de aproximações para as probabilidades



binomiais são baseadas na equação

$$\begin{aligned} P(Y \geq y) &= \sum_{i=y}^m \binom{m}{i} \pi^i (1-\pi)^{m-i} \\ &= B(y, m-y+1)^{-1} \int_0^\pi t^{y-1} (1-t)^{m-y} dt = I_\pi(y, m-y+1), \end{aligned}$$

em que  $I_\pi(y, m-y+1)$  representa a função razão beta incompleta.

Pode-se ainda usar a aproximação da distribuição binomial pela distribuição de Poisson  $P(m\pi)$  quando  $\pi < 0, 1$ , o erro da aproximação sendo  $O(m^{-1})$ , ou, então, a fórmula  $P(Y \leq y) = 1 - P\{F[2(y+1), 2(m-y)] < \pi(m-y)/[(1+y)(1-\pi)]\}$ , em que  $F[2(y+1), 2(m-y)]$  representa a distribuição F de Snedecor com  $2(y+1)$  e  $2(m-y)$  graus de liberdade.

Para finalizar, sejam  $B(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$  e  $P(y) = \frac{e^{-\mu} \mu^y}{y!}$ , as probabilidades pontuais das distribuições binomial e Poisson, respectivamente. Considerando  $\mu = m\pi$  e supondo  $\mu$  fixo, pode-se mostrar, com base na aproximação de Stirling para o fatorial, que quando  $m-y \rightarrow \infty$ ,

$$\frac{B(y)}{P(y)} \approx \left( \frac{m}{m-y} \right)^{1/2}.$$

Esse resultado pode ser, também, facilmente, comprovado numericamente.

O modelo binomial é usado, principalmente, no estudo de dados na forma de proporções, como nos casos da análise probito (Finney, 1952), logística (ou “logit”) (Ashton, 1972) e complemento log-log (Fisher, 1922) (Seção 2.2), e na análise de dados binários, como na regressão logística linear (Cox, 1970).

#### 2.4.3.1 Dados na forma de proporções

Considera-se o modelo binomial para o estudo de dados na forma de proporções e que são aplicadas doses de uma droga a  $n$  conjuntos de indivíduos, sendo  $m_i$  o número de indivíduos testados no conjunto  $i, i = 1, \dots, n$ . Conforme visto na Seção 2.2, o sucesso de um teste é determinado por uma variável latente  $U$ , denominada *tolerância*, com distribuição de probabilidade acumulada  $F(\cdot)$ . Os indivíduos

do conjunto  $i$  recebem uma dose fixa  $x_i$  da droga e a probabilidade de sucesso correspondente é dada por  $\pi_i = P(U \leq x_i) = F(\alpha + \beta x_i)$  em que  $\alpha$  e  $\beta$  são parâmetros desconhecidos que dependem dos parâmetros da distribuição proposta para  $U$ .

Sejam  $P_1, \dots, P_n$  as proporções de sucessos, supostas independentes, nos conjuntos  $1, \dots, n$ . O modelo para o estudo dessas proporções, no contexto dos MLG, tem variável resposta  $Y_i = m_i P_i$  com distribuição binomial, função de ligação  $F^{-1}(\cdot)$  e estrutura linear  $\eta_i = \alpha + \beta x_i$ . Convém salientar, que é postulada uma relação linear entre alguma função de  $\mu$  e  $x$ , ao invés de uma função de  $P$  e  $x$ . A variância da variável resposta não é constante, como no modelo clássico de regressão, e depende do valor da média.

Vários casos particulares desse modelo binomial são obtidos através da definição da distribuição da tolerância conforme explicado na Seção 2.2. Se se supõe que a tolerância tem distribuição normal, o modelo correspondente  $\pi_i = \Phi(\alpha + \beta x_i)$  é denominado probito (Finney, 1952). Se se supõe que tem distribuição logística, o modelo  $\pi_i = \exp(\alpha + \beta x_i) / [1 + \exp(\alpha + \beta x_i)]$  é chamado logístico (Berkson, 1944), e quando tem distribuição de valor extremo, a função de ligação  $F^{-1}(\cdot)$  corresponde ao modelo complemento log-log. O modelo logístico, postulando uma regressão linear para  $\log[\pi/(1 - \pi)]$  (“log odds”), tem sido muito usado na área de Medicina, pois tem uma interpretação simples, enquanto que o probito é o mais usado na área de Entomologia, por influência do artigo de Bliss (1935).

Existe pouca diferença entre as distribuições normal e logística para a tolerância, e, quando essas são re-escaladas adequadamente, por exemplo, para terem as médias e os desvios-padrão iguais, tornam-se bastante similares no intervalo  $[0, 1; 0, 9]$ . Por essa razão, é, geralmente, difícil diferenciá-las com base no ajuste do modelo. As funções de ligação logística e probito são simétricas em relação ao ponto de inflexão, isto é,  $F^{-1}(\pi) = -F^{-1}(1 - \pi)$ , o que não ocorre com função de ligação complemento log-log. Essa última função de ligação é mais apropriada para análise de dados sobre incidência de doenças. Para valores de  $\mu$  próximos de 0, as funções de ligação complemento log-log e logística são equivalentes. A família de funções de

ligação de Aranda-Ordaz (1981) com um parâmetro  $g(\mu; \lambda) = \log\{[(1 - \mu)^{-\lambda} - 1]/\lambda\}$  contém a função de ligação logística ( $\lambda = 1$ ) e a complemento log-log ( $\lambda = 0$ ).

#### 2.4.3.2 Dados binários agrupados

Apresenta-se, agora, o estudo de variáveis binárias agrupadas. Sejam  $n$  variáveis binárias,  $R_1, \dots, R_n$ , tendo somente os valores 0 e 1, classificadas em  $t$  grupos, o grupo  $i$  com  $m_i$  variáveis independentes com probabilidade de sucesso (resposta igual a 1) associada  $\pi_i$ ,  $i = 1, \dots, t$ , sendo  $\sum_{i=1}^t m_i = n$ . Definem-se  $Y_i$  e  $P_i$  como o número e a proporção de sucessos no grupo  $i$ , respectivamente, em que  $Y_i = m_i P_i$  tem distribuição binomial  $B(m_i, \pi_i)$ ,  $i = 1, \dots, t$ . O modelo para experimentos com respostas binárias não-agrupadas corresponde ao caso especial  $m_i = 1$  e  $n = t$ .

O modelo para  $m_i P_i$  com distribuição binomial  $B(m_i, \pi_i)$  e função de ligação  $g(\pi_i) = g(\mu_i/m_i) = \eta_i = \sum_{r=1}^p x_{ir} \beta_r$  pertence à classe dos MLG devendo a função de ligação ser uma função do intervalo  $(0, 1)$  na reta real. O modelo logístico linear é obtido definindo  $g(\pi_i) = g(\mu_i/m_i) = \log[\pi_i/(1 - \pi_i)] = \log[\mu_i/(m_i - \mu_i)]$ .

Um modelo alternativo para análise de dados binários agrupados é formulado por variáveis aleatórias independentes  $Z_i = g(Y_i/m_i)$ ,  $i = 1, \dots, t$ . A variável  $Z_i$  tem, aproximadamente, distribuição normal de média  $g(\pi_i)$  e variância  $g'(\pi_i)^2 \pi_i (1 - \pi_i)/m_i$ , desde que  $m_i \rightarrow \infty$  e que  $\pi_i$  não seja próximo de 0 ou 1. Essa variância é, consistentemente, estimada por  $v_i = g'(p_i)^2 p_i (1 - p_i)/m_i$ , substituindo  $\pi_i$  pelo valor amostral  $p_i$  de  $P_i$ .

Considera-se  $\mathbf{z} = (z_1, \dots, z_t)^T$  em que  $z_i = g(p_i)$ , como realizações de variáveis aleatórias com médias  $E(\mathbf{Z}) = \mathbf{X}\boldsymbol{\beta}$  e estrutura de covariância aproximada  $\mathbf{V} = \text{diag}\{v_1, \dots, v_t\}$ , sendo  $\mathbf{X}$  a matriz modelo de dimensões  $t \times p$  e  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ . Se não ocorrerem proporções de sucessos iguais a 0 ou 1, o método de mínimos quadrados ponderados, que equivale a minimizar  $(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})$  em relação a  $\boldsymbol{\beta}$ , produzirá o estimador  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{z}$ . Esse estimador é

diferente do estimador de máxima verossimilhança de  $\beta$ . Nesse modelo alternativo, testes e regiões de confiança para os parâmetros são obtidos na forma do modelo clássico de regressão.

Escolhendo a função de ligação  $g(\cdot)$  como a logística, tem-se  $Z_i = \log[Y_i/(m_i - Y_i)]$ , denominada *transformação logística empírica* de  $Y_i/m_i$ , sendo  $\text{Var}(Z_i)$  estimada por  $m_i/[Y_i(m_i - Y_i)]$ . Uma transformação mais adequada é obtida acrescentando-se 0,5 ao numerador e denominador, implicando

$$Z_i = \log \left( \frac{Y_i + 0,5}{m_i - Y_i + 0,5} \right),$$

pois  $E(Z_i) = \log[\pi_i/(1 - \pi_i)] + O(m_i^{-2})$ , além de ser definida para proporções de sucessos iguais a zero e um. Um estimador não-tendencioso de  $\text{Var}(Z_i)$  é dado por

$$v_i = \frac{(m_i + 1)(m_i + 2)}{m_i(Y_i + 1)(m_i - Y_i + 1)}.$$

Escolhendo a função de ligação arco seno, tem-se  $Z_i = \arcsen(\sqrt{Y_i/m_i})$ , denominada “transformação angular empírica” que, aproximadamente, estabiliza a variância para  $m_i$  grande. A média e a variância de  $Z_i$  são, aproximadamente, iguais a  $\arcsen(\sqrt{\pi_i})$  e  $1/(4m_i)$ , respectivamente.

#### 2.4.4 Modelo gama

Seja  $Y$  com distribuição gama,  $G(\mu, \phi)$ , com parâmetros positivos  $\mu$  e  $\phi$ , isto é, com f.d.p. dada por

$$f(y; \mu, \phi) = \frac{\left(\frac{\phi}{\mu}\right)^\phi}{\Gamma(\phi)} y^{\phi-1} \exp\left(-\frac{\phi y}{\mu}\right), \quad y > 0,$$

sendo a média  $\mu$  e o coeficiente de variação igual a  $\sqrt{\phi}$ . Tem-se, então, a função geratriz de momentos  $M(t; \mu, \phi) = (1 - \mu\phi t)^{-\phi^{-1}}$ , se  $t < (\phi\mu)^{-1}$ ,  $r$ -ésimo momento central  $(\mu\phi)^r \prod_{j=0}^{r-1} (j + \phi^{-1})$ ,  $r$ -ésimo cumulante  $(r-1)! \mu^r \phi^{r-1}$ , coeficientes de assimetria e curtose iguais a  $2\sqrt{\phi}$  e  $3 + 6\phi$ , respectivamente. Logo, o modelo gama  $G(\mu, \phi)$  tem o modelo normal como limite quando o parâmetro de dispersão  $\phi \rightarrow 0$ . A moda

da distribuição é igual a  $\mu(1 - \phi)$  para  $\phi \leq 1$  e, se  $\phi > 1$ , a função densidade da distribuição gama decresce quando  $y$  cresce.

Se a variável aleatória  $Y$  tem distribuição gama  $G(\mu, \phi)$ , a sua f.d.a. pode ser calculada por

$$P(Y \leq x) = \frac{\Gamma_{\phi\mu^{-1}x}(\phi)}{\Gamma(\phi)},$$

em que a função gama incompleta é  $\Gamma_y(\phi) = \int_0^y t^{\phi-1} e^{-t} dt$ . A função  $\Gamma(\phi) = \int_0^\infty t^{\phi-1} e^{-t} dt$  é a função gama. Essas funções estão disponíveis nos principais software estatísticos e pode ser encontrada, também, em <http://mathworld.wolfram.com>.

O modelo gama é usado na análise de dados contínuos não-negativos que apresentam uma variância crescente com a média e mais, fundamentalmente, quando o coeficiente de variação dos dados for, aproximadamente, constante. É, também, aplicado na estimação de componentes de variância de modelos com efeitos aleatórios, e como uma distribuição aproximada de medições físicas, tempos de sobrevivência etc.

Uma aplicação do modelo gama é na análise de variância com efeitos aleatórios, em que as somas de quadrados, supondo que a variável resposta tem distribuição normal, são proporcionais a variáveis qui-quadrados. Sejam  $k$  somas de quadrados  $SQ_1, \dots, SQ_k$  independentes, tais que  $SQ_i \sim \eta_i \chi_{\nu_i}^2$ , em que  $\nu_i$  é o número de graus de liberdade associado a  $SQ_i$  e  $\eta_i = \sum_{j=1}^p x_{ij} \sigma_j^2$  é uma constante de proporcionalidade, dada como uma combinação linear de  $p$  variâncias desconhecidas  $\sigma_1^2, \dots, \sigma_p^2$ . Como os quadrados médios  $QM_i = SQ_i/\nu_i$  têm distribuição  $(\eta_i/\nu_i) \chi_{\nu_i}^2$ , pode-se considerar  $QM_i, i = 1, \dots, k$ , como sendo a variável resposta, no contexto dos MLG, seguindo o modelo  $G(\eta_i, (\nu_i/2)^{-1})$  com função de ligação identidade.

Suponha, agora, que a variável aleatória  $Y$  tem distribuição gama  $G(\mu, \phi)$  com coeficiente de variação  $\sqrt{\phi}$  bastante pequeno. Obtêm-se as aproximações

$$E(\log Y) \approx \log \mu - \frac{\phi}{2} \quad \text{e} \quad \text{Var}(\log Y) \approx \phi.$$

Assim, ao invés de analisar os dados  $y$  através do modelo gama  $G(\mu, \phi)$  com função de ligação  $g(\cdot)$ , pode-se construir um modelo normal alternativo de variância constante  $\phi$  e função de ligação  $g(\exp(\cdot))$  ajustado aos logaritmos dos dados. Além disso, a variância da variável transformada, isto é,  $\phi = \text{Var}(\log Y)$ , pode ser estimada, após o ajuste do modelo normal, por exemplo, pelo quadrado médio do resíduo.

Finalmente, pode-se demonstrar que o logaritmo da função de verossimilhança do modelo gama  $G(\mu, \phi)$  é, aproximadamente, quadrático, na escala  $\mu^{-1/3}$ , e que a diferença entre o seu máximo e o valor num ponto arbitrário  $\mu$ , é dada por  $9y^{2/3}(y^{-1/3} - \mu^{-1/3})^2/2$  (McCullagh e Nelder, 1989, Seção 7.2). Ainda, tem-se que a variável transformada  $3[(\frac{Y}{\mu})^{1/3} - 1]$  é, aproximadamente, normal.

### 2.4.5 Modelo normal inverso

A distribuição normal inversa (ou Gaussiana inversa) foi deduzida por Wald e Tweedie em dois artigos publicados independentemente em 1947. A f.d.p. da distribuição normal inversa  $IG(\mu, \phi)$  com média  $\mu > 0$  e parâmetro  $\phi > 0$ , representando uma medida de dispersão, é dada por

$$\pi(y; \mu, \phi) = (2\pi\phi y^3)^{-1/2} \exp \left[ \frac{-(y - \mu)^2}{2\mu^2\phi y} \right], y > 0.$$

O parâmetro  $\mu$  é, portanto, uma medida de locação e o parâmetro  $\phi$ , uma medida de dispersão, isto é,  $\phi$  é a razão entre a variância e o cubo da média.

As características da distribuição  $IG(\mu, \phi)$  são: função geratriz de momentos  $M(t; \mu, \phi) = \exp \{(\phi\mu)^{-1}[1 - (1 + 2\mu^2\phi t)^{1/2}]\}$ , cumulantes para  $r \geq 2$  obtidos de  $\kappa_r = 1.3.5 \dots (2r - 1)\mu^{2r-1}\phi^{r-1}$ , coeficientes de assimetria e curtose iguais a  $3\sqrt{\mu\phi}$  e  $3 + 15\mu\phi$ , respectivamente, e moda  $\mu[(1 + 9\mu^2\phi^2/4)^{1/2} - 3\mu\phi/2]$ . A distribuição é unimodal e sua forma depende apenas do valor do produto  $\phi\mu$ . Uma relação importante entre os momentos positivos e negativos é  $E(Y^{-r}) = E(Y^{r+1})/\mu^{2r+1}$ .

A f.d.a. da distribuição normal inversa  $IG(\mu, \phi)$  pode ser obtida a partir da distribuição  $N(0, 1)$  por  $P(Y \leq y) = \Phi(y_1) + \exp[2/(\phi\mu)]\Phi(y_2)$ , em que  $y_1 = (\phi y)^{-1/2}(-1 + y/\mu)$  e  $y_2 = -(\phi y)^{-1/2}(1 + y/\mu)$ .

A distribuição normal inversa tem distribuição assintótica normal, da mesma forma que a gama, a log normal e outras distribuições assimétricas. Quando  $\phi \rightarrow 0$ , a distribuição normal inversa  $IG(\mu, \phi)$  é, assintoticamente,  $N(\mu, \mu^3\phi)$ .

As aplicações do modelo normal inverso  $IG(\mu, \phi)$  concentram-se no estudo do movimento Browniano de partículas, análise de regressão com dados consideravelmente assimétricos, testes de confiabilidade, análise seqüencial e análogo de análise de variância para classificações encaixadas. Outras aplicações incluem modelagem de tempos, como: duração de greves, tempo de primeira passagem nos passeios aleatórios, tempos de sobrevivência, tempo gasto para injetar uma substância no sistema biológico etc.

Existem muitas analogias entre os modelos normal e normal inverso. Por exemplo, o dobro do termo do expoente com sinal negativo nas funções densidades normal e normal inversa, tem distribuição  $\chi_1^2$ . Um estudo completo do modelo normal inverso  $IG(\mu, \phi)$  é apresentado por Folks e Chhikara (1978).

#### 2.4.6 Modelo binomial negativo

A distribuição binomial negativa com parâmetros  $k > 0$  e  $0 < p < 1$  é definida por

$$P(Y = y) = \binom{k+y-1}{k-1} \left(\frac{p}{p+1}\right)^y \frac{1}{(p+1)^k}$$

para  $y = 0, 1, 2, \dots$ . O parâmetro  $\mu = kp$  é igual à média e pode ser usado no lugar de  $p$  (Tabela 1.1 e Exercício 1b do Capítulo 1). Quando  $k$  é inteiro, essa distribuição é, também, chamada distribuição de Pascal. Um caso especial importante é a distribuição geométrica quando  $k = 1$ . Formas especiais da distribuição binomial negativa surgiram com Pascal e Fermat, em 1679. Gosset (“Student”), em 1907, usou a distribuição binomial negativa como um modelo para contagens no lugar da distribuição de Poisson.

A função geratriz de momentos é  $M(t) = [1 + p(1 - e^t)]^{-k}$ . A sua variância é igual a  $\text{Var}(Y) = kp(1 + p)$  e os coeficientes de assimetria e curtose valem  $(2p +$

$1)/\sqrt{kp(p+1)}$  e  $3 + [1 + 6p(1+p)]/[kp(1+p)]$ , respectivamente. Observe-se que sua variância pode ser escrita em termos da média como  $\text{Var}(Y) = \mu(1 + \mu/k)$ , o que caracteriza o modelo binomial negativo como um dos modelos adequados para estudar **superdispersão**, isto é,  $\text{Var}(Y) > E(Y)$  (Hinde e Demétrio, 1998a,b).

Pode-se verificar que  $P(Y = y + 1) > P(Y = y)$  quando  $y < \mu(1 - k^{-1}) - 1$  e  $P(Y = y + 1) < P(Y = y)$  quando  $y > \mu(1 - k^{-1}) - 1$ .

A distribuição acumulada da binomial negativa  $P(Y \leq y)$  para  $y$  inteiro pode ser calculada a partir da distribuição acumulada da variável aleatória  $X$  tendo distribuição binomial com parâmetros  $k + y$  e  $(1 + p)^{-1}$  por  $P(Y \leq y) = P(X \geq k)$ . Alternativamente, a distribuição acumulada da binomial negativa pode ser calculada de forma aproximada por

$$P(Y \leq y) \approx e^{-\mu} \sum_{i=0}^y \frac{\mu^i}{i!} - \frac{(y - \mu) k e^{-\mu} \mu^y}{2 (\mu + k) y!}.$$

#### 2.4.7 Modelo secante hiperbólico generalizado

A distribuição secante hiperbólica generalizada (SHG) foi estudada por Morris (1982) no contexto da função de variância da família exponencial, sendo uma função quadrática da média. A sua f.d.p. é dada por ( $y \in \mathbb{R}$ )

$$f(y; \mu, \phi) = \exp \left\{ \frac{1}{\phi} [y \arctan \mu - \frac{1}{2} \log(1 + \mu^2)] + c(y, \phi) \right\},$$

sendo

$$c(y, \phi) = \log \left\{ \frac{2^{(1-2\phi)/\phi}}{\pi \phi \Gamma(\phi^{-1})} \right\} - \sum_{j=0}^{\infty} \log \left\{ 1 + \frac{y^2}{(1 + 2j\phi)^2} \right\}.$$

Em relação a outras distribuições na família exponencial, a forma de sua função  $c(y, \phi)$  é bastante complicada. Entretanto, a distribuição SHG pode ser adequada para análise de dados contínuos reais como distribuição alternativa à normal. A sua função de variância é obtida de  $\theta = \arctan \mu$  como  $V = d\mu/d\theta = 1 + \mu^2$ . Morris (1982) demonstrou que existem na família exponencial (2.4), exatamente, seis distribuições com função de variância quadrática  $V(\mu) = c_0 + c_1\mu + c_2\mu^2$ , a



saber: binomial ( $c_0 = 0, c_1 = 1, c_2 = -1$ ), Poisson ( $c_0 = c_2 = 0, c_1 = 1$ ), normal ( $c_0 = 1, c_1 = c_2 = 0$ ), gama ( $c_0 = c_1 = 0, c_2 = 1$ ), binomial negativa ( $c_0 = 0, c_1 = 1, c_2 > 0$ ) e SHG ( $c_0 = c_2 = 1, c_1 = 0$ ).

### 2.4.8 Modelos definidos por transformações

Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias tais que após uma transformação  $h(\cdot)$ , as variáveis resultantes  $Z_1, \dots, Z_n$ , em que  $Z_i = h(Y_i)$ , têm distribuições normais de médias  $\mu_1, \dots, \mu_n$  e variância constante  $\sigma^2$ , e que para uma outra transformação  $g(\cdot)$  produz linearidade dos efeitos sistemáticos, isto é,  $g[E(\mathbf{Y})] = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ . Usando-se expansão de Taylor em torno de  $\mu$  até primeira ordem, tem-se  $g(h^{-1}(\mu)) = \eta$  e, portanto, esses modelos pertencem, de forma aproximada, à classe dos MLG, com distribuição normal e função de ligação  $g(h^{-1}(\cdot))$ . A variância da variável resposta original pode ser obtida, aproximadamente, de  $\text{Var}(Y) = \sigma^2 / \{h'[g^{-1}(\eta)]^2\}$ .

Usando-se a transformação potência de Box e Cox (1964), pode-se definir uma subclasse de modelos por

$$Z = h(Y) = \phi^{-1}(Y^\phi - 1) \sim N(\mu, \sigma^2)$$

e

$$g[E(Y)] = \theta^{-1}\{[E(Y)]^\theta - 1\},$$

em que  $g[E(\mathbf{Y})] = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ . Aqui,  $\phi = 0$  e  $\theta = 0$  correspondem à transformação logarítmica. Logo, essa subclasse é representada, no contexto dos MLG, por uma distribuição normal com função de ligação  $\phi^{-1}[(1 + \phi\mu)^{\theta/\phi} - 1] = \eta$ , supondo  $\theta \neq 0$  e  $\phi \neq 0$ .

Quando  $\theta = \phi = 1$  tem-se o modelo clássico de regressão. Um caso importante, denominado *polinômios inversos* (Nelder, 1966), é definido por  $\phi = 0$  e  $\theta = -1$  e, portanto, admite erros normais na escala logarítmica e linearidade na escala inversa, sendo equivalente ao modelo normal  $N(\mu, \sigma^2)$  com função de ligação  $\eta = 1 - \exp(-\mu)$ .

## 2.5 Metodologia

O processo de ajuste dos MLG pode ser dividido em três etapas: (i) formulação dos modelos; (ii) ajuste dos modelos e (iii) inferência.

Os MLG formam um ferramental de grande utilidade prática, pois apresentam grande flexibilidade na etapa (i), computação simples em (ii) e critérios razoáveis em (iii). Essas etapas são realizadas sequencialmente. Na análise de dados complexos, após a realização da etapa de inferência, pode-se voltar à etapa (i) e escolher outros modelos, a partir de informações mais detalhadas obtidas do estudo feito em (iii).

Uma característica importante dos MLG é que se supõe independência das variáveis respostas (ou, pelo menos, não correlação) e, portanto, dados exibindo autoregressões como as séries temporais, em princípio, podem ser excluídos. Uma segunda característica é que a estrutura do erro é suposta única embora, usualmente, existam várias variáveis explanatórias (covariáveis) na estrutura linear desses modelos. Assim, outras técnicas estatísticas devem ser consideradas para analisar dados, que ocorrem em planejamentos de experimentos com mais de uma fonte de erro. Ainda, variáveis respostas com distribuições fora da família (2.4), como a distribuição de Cauchy, e estruturas não lineares do tipo  $\eta = \sum \beta_j \exp(\alpha_j x_j)$ , a menos que os  $\alpha_j$  sejam conhecidos, devem também ser excluídos.

Apresentam-se, agora, as características principais das etapas que formam a metodologia de trabalho com os MLG.

### 2.5.1 Formulação de modelos

A etapa de formulação dos modelos compreende a escolha de opções para a distribuição de probabilidade da variável resposta, covariáveis (matriz modelo) e função de ligação. Essas opções visam a descrever as características principais da variável resposta.

Para se escolher razoavelmente a distribuição em (2.4), devem-se exami-

nar cuidadosamente os dados, principalmente quanto aos seguintes pontos básicos: assimetria, natureza contínua ou discreta (por exemplo, contagens) e intervalo de variação.

As distribuições gama e normal inversa são associadas a dados contínuos assimétricos. Se os dados exibem simetria e o intervalo de variação é o conjunto dos reais, a distribuição normal deve ser escolhida. Entretanto, se os dados têm intervalo de variação em  $(0, \infty)$ , a suposição de normalidade pode ser mais apropriada para alguma transformação dos dados, por exemplo, a logarítmica. Alternativamente, podem-se supor as distribuições normal inversa e gama, cujos intervalos de variação são positivos. Quando os dados apresentam coeficientes de variação constante, o modelo gama deve ser o preferido.

A distribuição de Poisson aplica-se a observações na forma de contagens, mas pode também ser usada para análise de dados contínuos que apresentam variância aproximadamente igual à média. Quando a variância dos dados é maior do que a média (ao invés de igual), pode-se trabalhar com as distribuições gama, normal inversa e binomial negativa. Esse fenômeno é denominado **superdispersão** para distribuições discretas (Hinde e Demétrio, 1998a,b). A escolha entre essas três distribuições pode depender, exclusivamente, da dispersão dos dados. A variância da binomial negativa ( $V(\mu) = \mu + \mu^2/r$ ) pode ser aproximada, para um intervalo razoável de variação de  $\mu$ , por  $V(\mu) = \lambda\mu$ , em que a função de variância contém um parâmetro multiplicador  $\lambda > 1$ , desconhecido. Portanto, a distribuição de Poisson pode ser empregada para análise de dados que apresentam superdispersão, desde que seja obtida uma estimativa para  $\lambda$ . O fenômeno de subdispersão, em que a variância dos dados é menor do que a média, pode ser tratado através do modelo de Poisson com  $\lambda < 1$ , mas é muito incomum na prática. Nesse caso, a distribuição binomial pode ser a mais adequada.

A distribuição binomial serve para análise de dados na forma de proporções, podendo ainda ser útil na análise de dados contínuos ou discretos apresentando subdispersão. A superdispersão analisada através da distribuição binomial é possível,

com um parâmetro multiplicador na função de variância, porém não é freqüente na prática.

A escolha de uma função de ligação compatível com a distribuição proposta para os dados, deve resultar de considerações a priori, exame intensivo dos dados, facilidade de interpretação do modelo e, mais usualmente, uma mistura de tudo isso.

No modelo clássico de regressão a função de ligação é a identidade no sentido de que valores esperados e preditores lineares podem ter qualquer valor real. Entretanto, quando os dados são contagens e a distribuição é de Poisson, a função de ligação identidade, como visto anteriormente, é menos atrativa, pois não restringe os valores esperados ao intervalo  $(0, \infty)$ . Quando efeitos sistemáticos multiplicativos contribuem para as médias dos dados, uma função de ligação logarítmica torna os efeitos aditivos contribuindo para os preditores lineares e, portanto, pode ser a mais apropriada. Analogamente, as funções de ligação adequadas para os dados na forma de proporções, devem ser funções de  $(0, 1)$  no conjunto dos reais, como probito, logística, complemento log-log e arco seno. As funções de ligação compatíveis com os modelos gama, normal inverso e binomial negativo devem restringir as médias dos dados ao intervalo  $(0, \infty)$ .

A Tabela 2.7 apresenta a combinação distribuição da variável resposta/função de ligação para os casos especiais de MLG  $(a), (b), \dots, (l)$ , descritos na Seção 2.1.

Existem funções de ligação que produzem propriedades estatísticas desejáveis para o modelo, particularmente, em pequenas amostras. Essas funções são definidas visando aos seguintes efeitos de forma separada: constância da informação de Fisher e da curvatura do logaritmo da função de verossimilhança, estatísticas suficientes de dimensão mínima, normalização aproximada das estimativas de máxima verossimilhança dos parâmetros lineares e simetria do logaritmo da função de verossimilhança. Nenhuma função de ligação pode produzir todos estes efeitos desejados e, muitas vezes, se existe uma função de ligação superior, ela pode conduzir a dificuldades de interpretação.

Tabela 2.7: Combinação da distribuição da variável resposta e da função de ligação para os casos especiais de MLG descritos na Seção 2.1.

Função de ligação	Distribuição				
	Normal	Poisson	Binomial	Gama	Normal Inversa
Identidade	(a)	–	–	(i)	–
Logarítmica	–	(e)	–	–	–
Inversa	(h)	–	–	(g)(j)	–
Inversa do quadrado	–	–	–	–	(l)
Logística	–	–	(d)(f)	–	–
Probit	–	–	(c)	–	–
Complemento log-log	–	–	(b)	–	–

Observação: Para os casos (g), (j) e (l) foram escolhidas as funções de ligação mais usuais (canônicas) que correspondem a  $\theta = \eta$ .

A terceira escolha na formulação do modelo é a do conjunto de variáveis explicativas para representar a estrutura linear do MLG, ou seja, a formação da matriz modelo. Em geral, as covariáveis escolhidas devem ser não-correlacionadas. Os termos da estrutura linear podem ser *contínuos*, *qualitativos* e *mistos*.

Uma covariável contínua  $x$ , geralmente, corresponde a um único parâmetro  $\beta$ , contribuindo com o termo  $\beta x$  para o modelo, enquanto uma covariável qualitativa  $A$ , denominada freqüentemente de fator, inclui na estrutura linear um conjunto de parâmetros  $\alpha_i$ , em que  $i$  é o índice que representa os níveis do fator. Assim, na estrutura linear  $\eta_i = \alpha_i + \beta x$ , representando grupos distintos de um fator  $A$  mais uma covariável contínua  $x$ , a ordenada varia com o nível do fator, mas a declividade é a mesma. Entretanto, em alguns casos, a declividade deve variar com o nível do fator e, portanto, o termo  $\beta x$  deve ser substituído pelo mais geral  $\beta_i x$ , produzindo  $\eta = \alpha_i + \beta_i x$ . O termo  $\beta_i x$  é denominado misto, pois a declividade associada à covariável é suposta diferente para cada nível do fator.

Freqüentemente, as observações são classificadas por dois ou mais fatores

simultaneamente e, então, termos representando interações entre os fatores devem ser incluídos no modelo. Uma covariável contínua  $x$  pode ser transformada por uma função não-linear  $h(x)$ , sem prejudicar a linearidade do modelo, desde que  $h(\cdot)$  não contenha parâmetros desconhecidos. Assim, a estrutura linear do modelo pode conter polinômios em  $x$ . Transformações simples nas covariáveis podem implicar num grande aperfeiçoamento do componente sistemático do modelo. O caso de funções não-lineares das covariáveis com parâmetros desconhecidos, será discutido na Seção 5.7.

Em muitas aplicações, a combinação linear das covariáveis  $x_1, \dots, x_p$  depende, fortemente, das características do experimento e deve propiciar uma contribuição útil na explicação do comportamento da variável resposta associada aos dados  $y$ .

Um MLG é considerado como uma boa representação dos dados se conseguir explicar a relação variância/média satisfatoriamente, e se produzir efeitos aditivos na escala definida pela função de ligação. Um modelo parcimonioso é, também, uma exigência, no sentido de que o número de parâmetros seja tão pequeno quanto possível. Por exemplo, se os dados são classificados por dois ou mais fatores, um modelo parcimonioso deve minimizar o número de interações entre os fatores.

Um ponto fundamental no processo de escolha de um MLG, é que não se deve ficar restrito a um único modelo, achando que ele é o mais importante e excluir outros alternativos. É prudente considerar a escolha restrita a um conjunto amplo de modelos estabelecidos por princípios como: facilidade de interpretação, boas previsões anteriores e conhecimento profundo da estrutura dos dados. Algumas características nos dados podem não ser descobertas, mesmo por um modelo muito bom e, portanto, um conjunto razoável de modelos adequados aumenta a possibilidade de se detectarem essas características.

### 2.5.2 Ajuste dos modelos

A etapa de ajuste representa o processo de estimação dos parâmetros lineares dos modelos e de determinadas funções das estimativas desses parâmetros, que representam medidas de adequação dos valores estimados. Vários métodos podem ser usados para estimar os parâmetros dos MLG. Nesse ponto, convém recordar a citação “nada é tão fácil quanto inventar métodos de estimação” de Sir Ronald Fisher (1925). Como o método de máxima verossimilhança nos MLG conduz a um procedimento de estimação bastante simples, esse método é o mais usado.

O algoritmo para a solução das equações de máxima verossimilhança nos MLG foi desenvolvido por Nelder e Wedderburn (1972) e equivale ao cálculo repetido de uma regressão linear ponderada, como será descrito na Seção 3.2. O algoritmo é similar a um processo iterativo de Newton-Raphson, mas a característica principal é o uso da matriz de valores esperados das derivadas parciais de segunda ordem do logaritmo da função de verossimilhança (informação), em relação aos  $\beta$ 's, no lugar da matriz correspondente de valores observados. Essa característica foi, primeiramente, desenvolvida por Fisher (1935), para o caso da distribuição binomial com função de ligação “probit” e o processo é denominado “método score para estimação de parâmetros”.

O algoritmo de Nelder e Wedderburn (1972) tem como casos especiais os algoritmos de Finney (1952) para o cálculo de curvas ajustadas de resposta a um conjunto de doses de um medicamento, e de Haberman (1970) para o cálculo de estimativas nos modelos log-lineares.

Vários software estatísticos como R, SAS, S-PLUS e MATLAB fornecem para cada ajuste, as estimativas dos parâmetros  $\beta$ ,  $\eta$  e  $\mu$  do modelo, resíduos, estruturas de covariância e correlação entre as estimativas e outras funções de interesse.

O algoritmo de estimação nos MLG é bastante robusto, convergindo rapidamente. Entretanto, pode falhar em convergir de duas maneiras distintas:

- (a) os parâmetros tomam valores infinitos, embora o máximo do logaritmo da

função de verossimilhança esteja convergindo para o valor correto;

- (b) o logaritmo da função de verossimilhança, ao invés de sempre crescer no processo iterativo, começa a decrescer ou oscilar, constituindo uma divergência real.

Quando  $\theta = \eta$ , implicando um modelo com  $p$  estatísticas suficientes minimais, tem-se constatado que exemplos de divergência são muito raros.

Ao ocorrer falha do algoritmo, torna-se necessário repetir o procedimento de estimação, a partir dos valores ajustados correntes, usando um modelo diferente, pois a convergência pode ser alcançada (Cordeiro, 1986).

### 2.5.3 Inferência

A etapa de inferência tem como objetivo principal verificar a adequação do modelo como um todo e realizar um estudo detalhado quanto a discrepâncias locais. Essas discrepâncias, quando significantes, podem implicar na escolha de outro modelo, ou em aceitar a existência de dados aberrantes. Em qualquer caso, toda a metodologia de trabalho deverá ser repetida.

O analista deve, nessa etapa, verificar a precisão e a interdependência das estimativas, construir regiões de confiança e testes sobre os parâmetros de interesse, analisar estatisticamente os resíduos e realizar previsões.

A precisão das previsões depende basicamente do modelo selecionado e, portanto, um critério de adequação do ajuste é verificar se a precisão de uma previsão em particular é maximizada. Muitas vezes, é possível otimizar a precisão por simples alteração do componente sistemático do modelo.

Um gráfico dos resíduos padronizados versus valores ajustados, sem nenhuma tendência, é um indicativo de que a relação funcional variância/média proposta para os dados é satisfatória. Gráficos dos resíduos versus covariáveis que não estão no modelo são bastante úteis. Se nenhuma covariável adicional é necessária, então não se deve encontrar qualquer tendência nesses gráficos. Observações com erros



grosseiros podem ser detectadas como tendo resíduos grandes e *leverages* pequenos ou resíduos pequenos e *leverages* ( $h$ ) grandes, ou o modelo ajustado deve requerer mais covariáveis, por exemplo, interações de ordem superior. A inspeção gráfica é um meio poderoso de inferência nos MLG.

Para a verificação do ajuste do MLG, pode-se adotar o critério da razão da verossimilhanças em relação ao modelo saturado e a estatística de Pearson generalizada (Seção 4.2). Quase toda a parte de inferência nos MLG é baseada em resultados assintóticos, e pouco tem sido estudado sobre a validade desses resultados em amostras muito pequenas.

Um modelo mal ajustado aos dados pode apresentar uma ou mais das seguintes condições:

- (a) inclusão de um grande número de covariáveis no modelo, muitas das quais são redundantes e algumas explicando somente uma pequena fração de dados;
- (b) formulação de um modelo bastante pobre em covariáveis, que não revela e nem reflete as características do mecanismo gerador dos dados;
- (c) as observações mostram-se insuficientes para que falhas do modelo sejam detectadas.

A condição (a) representa uma superparametrização do modelo implicando numa imprecisão das estimativas e (b) é a situação oposta de (a): um subparametrização que acarreta previsões ruins. A terceira condição é um tipo de falha difícil de se detectar, e é devida a uma combinação inadequada distribuição/função de ligação, que nada tem a ver com as observações em questão.

## 2.6 Exercícios

1. Para o modelo binomial as funções de ligação mais comuns são: logística, probito e complemento log-log. Comparar os valores do preditor linear para essas funções de

ligação no intervalo  $(0, 1)$ .

2. Mostre que

$$\lim_{\lambda \rightarrow 0} \frac{\mu^\lambda - 1}{\lambda} = \log \mu.$$

3. Considere a família de funções de ligação dada por Aranda-Ordaz (1981)

$$\eta = \log \left[ \frac{(1 - \pi)^{-\lambda} - 1}{\lambda} \right], \quad 0 < \pi < 1 \text{ e } \lambda \text{ uma constante.}$$

Mostre que a função de ligação logística é obtida para  $\lambda = 1$  e que quando  $\lambda \rightarrow 0$ , tem-se a função de ligação complemento log-log.

4. Comparar os gráficos de  $\eta = \log \left[ \frac{(1 - \mu)^{-\lambda} - 1}{\lambda} \right]$  versus  $\mu$  para  $\lambda = -1, -0.5, 0, 0.5, 1$  e  $2$ .

5. Explicar como um modelo de Box-Cox poderia ser formulado no contexto dos MLG.

6. Demonstrar que se  $Y$  tem uma distribuição binomial  $B(m, \pi)$ , então para  $m$  grande  $\text{Var}(\arcsen \sqrt{Y/m})$  é, aproximadamente,  $1/(4m)$ , com o ângulo expresso em radianos. Em que situações, uma estrutura linear associada a essa transformação poderá ser adequada?

7. Suponha que  $Y$  tem distribuição binomial  $B(m, \pi)$  e que  $g(Y/m)$  é uma função arbitrária. Calcular o coeficiente de assimetria assintótico de  $g(Y/m)$ . Demonstrar que ele se anula quando  $g(\pi) = \int_0^\pi t^{-1/3}(1-t)^{-1/3}dt$  e, portanto, a variável aleatória definida por  $[g(Y/m) - g(\alpha)]/[\pi^{1/6}(1-\pi)^{1/6}m^{-1/2}]$ , em que  $\alpha = \pi - (1-2\pi)/(6m)$ , tem distribuição próxima da normal reduzida (Cox e Snell, 1968).

8. Sejam  $Y_1$  e  $Y_2$  variáveis aleatórias binomiais de parâmetros  $\pi_1$  e  $\pi_2$  em dois grupos de tamanhos  $m_1$  e  $m_2$ , respectivamente. O número de sucessos  $Y_1$  no primeiro grupo dado que o total de sucessos nos dois grupos é  $r$ , tem distribuição hipergeométrica

generalizada de parâmetros  $\pi_1$ ,  $\pi_2$ ,  $m_1$ ,  $m_2$  e  $r$ . Demonstrar que essa distribuição é um membro da família (2.4) com parâmetro  $\theta = \log\{\pi_1(1 - \pi_2)/[\pi_2(1 - \pi_1)]\}$ ,  $\phi = 1$  e  $\pi = D_1(\theta)/D_0(\theta)$ , em que  $D_i(\theta) = \sum_x x^i \binom{m_1}{x} \binom{m_2}{r-x} \exp(\theta x)$  para  $i = 0, 1$ . Calcular a expressão do  $r$ -ésimo cumulante dessa distribuição.

**9.** Se  $Y$  tem distribuição de Poisson  $P(\mu)$  demonstrar:

- (a) que o coeficiente de assimetria  $Y^{2/3}$  é de ordem  $\mu^{-1}$  enquanto que os de  $Y$  e  $Y^{1/2}$  são de ordem  $\mu^{-1/2}$ ;
- (b) que o logaritmo da função de verossimilhança para uma única observação é, aproximadamente, quadrático na escala  $\mu^{1/3}$ ;
- (c) a fórmula do  $r$ -ésimo momento fatorial  $E[Y(Y-1)\dots(Y-r+1)] = \mu^r$ ;
- (d) a fórmula de recorrência entre os momentos centrais  $\mu_{r+1} = r\mu\mu_{r-1} + \mu \frac{d\mu_r}{d\mu}$ ;
- (e) que  $2\sqrt{Y}$  é, aproximadamente, normal  $N(0, 1)$ .

**10.** Se  $Y$  tem distribuição gama  $G(\mu, \phi)$  demonstrar que:

- (a) quando  $\phi < 1$  a função densidade é zero na origem e tem uma única moda no ponto  $\mu(1 - \phi)$ ;
- (b) o logaritmo da função de verossimilhança para uma única observação é, aproximadamente, quadrático na escala  $\mu^{-1/3}$ ;
- (c) a variável transformada  $3[(Y/\mu)^{1/3} - 1]$  é, aproximadamente, normal.

**11.** Se  $Y$  tem distribuição binomial  $B(m, \pi)$ , demonstrar que a média e a variância de  $\log[(Y + 0,5)/(m - Y + 0,5)]$  são iguais a  $\log[\pi/(1 - \pi)] + O(m^{-2})$  e  $E\{(Y + 0,5)^{-1} + (m - Y + 0,5)^{-1}\} + O(m^{-3})$ , respectivamente.

**12.** Se  $Y$  tem distribuição de Poisson  $P(\mu)$ , obter uma expansão para  $\text{Var}\{(Y+c)^{1/2}\}$

em potências de  $\mu^{-1}$ , e mostrar que o coeficiente de  $\mu^{-1}$  é zero quando  $c = \frac{3}{8}$ . Achar uma expansão similar para  $\text{Var}\{Y^{1/2} + (Y + 1)^{1/2}\}$ .

**13.** Qual é a distribuição da tolerância correspondente à função de ligação  $\arcsen\sqrt{\cdot}$ ?

**14.** Se  $Y$  tem distribuição binomial  $B(m, \pi)$  demonstrar que os momentos da estatística  $Z = \pm\{2Y \log(Y/\mu) + 2(m - Y) \log[(m - Y)/(m - \mu)]\}^{1/2} + \{(1 - 2\pi)/[m\pi(1 - \pi)]\}^{1/2}/6$  diferem dos correspondentes da distribuição normal  $N(0, 1)$  com erro  $O(m^{-1})$ . Essa transformação induz simetria e estabiliza a variância simultaneamente (McCullagh e Nelder, 1989).

**15.** Se  $Y$  tem distribuição binomial  $B(m, \pi)$ , demonstrar a expressão aproximada  $P(Y \leq y) = \Phi(y_1)$ , em que  $y_1 = 2m^{1/2}\{\arcsen[(y+3/8)/(m+3/4)]^{1/2} - \arcsen(\pi^{1/2})\}$ .

**16.** Suponha que  $Y \sim B(m, \pi)$  em que  $\pi = e^\lambda(1 + e^\lambda)^{-1}$ . Mostre que  $m - Y$  tem distribuição binomial com parâmetro induzido correspondente  $\lambda' = -\lambda$ .

**17.** Demonstrar que para a variável aleatória  $Y$  com distribuição de Poisson, tem-se:

(a)  $E(Y^{1/2}) \approx \mu^{1/2}$  e  $\text{Var}(Y^{1/2}) \approx \frac{1}{4}$ ;

(b)  $E(Y^{1/2}) = \mu^{1/2} \left(1 - \frac{1}{8\mu}\right) + O(\mu^{-3/2})$  e  $\text{Var}(Y^{1/2}) = \frac{1}{4} \left(1 + \frac{3}{8\mu}\right) + O(\mu^{-3/2})$ ;

(c)  $E(Y^{2/3}) \approx \mu^{2/3} \left(1 - \frac{1}{9\mu}\right)$  e  $\text{Var}(Y^{2/3}) \approx \frac{4\mu^{1/3}}{9} \left(1 + \frac{1}{6\mu}\right)$ .

**18.** Se  $Y$  tem distribuição de Poisson com média  $\mu$ , mostre que:

(a)  $P(Y \leq y) = P(\chi_{2(y+1)}^2 > 2\mu)$ ;

(b)  $P(Y \leq y) = \Phi(z) - \phi(z) \left( \frac{z^2 - 1}{6\sqrt{\mu}} + \frac{z^5 - 7z^3 + 3z}{72\mu} \right) + O(\mu^{-3/2})$ , em que  $z = (y + 0.5 - \mu)\mu^{-1/2}$  e  $\Phi(\cdot)$  e  $\phi(\cdot)$  são, respectivamente, a f.d.a. e a f.d.p. da distribuição normal reduzida.

# Capítulo 3

## Estimação

### 3.1 Estatísticas suficientes

Seja um MLG definido pelas expressões (2.4), (2.6) e (2.7) e suponha que os dados a serem analisados sejam representados pelo vetor  $\mathbf{y} = (y_1, \dots, y_n)^T$ . O logaritmo da função de verossimilhança como função apenas de  $\boldsymbol{\beta}$  (considerando-se o parâmetro de dispersão  $\phi$  conhecido) dado o vetor  $\mathbf{y}$  é definido por  $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; \mathbf{y})$  e usando-se a expressão (2.4) tem-se

$$\ell(\boldsymbol{\beta}) = \phi^{-1} \sum_{i=1}^n [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^n c(y_i, \phi), \quad (3.1)$$

em que  $\theta_i = q(\mu_i)$ ,  $\mu_i = g^{-1}(\eta_i)$  e  $\eta_i = \sum_{r=1}^p x_{ir} \beta_r$ .

A estimação do parâmetro de dispersão  $\phi$  será discutida na Seção 4.4. Existem  $n$  parâmetros canônicos  $\theta_1, \dots, \theta_n$  e  $n$  médias  $\mu_1, \dots, \mu_n$  que são desconhecidos, mas que são funções de  $p$  parâmetros lineares  $\beta_1, \dots, \beta_p$  do modelo. Deve-se, primeiramente, estimar o vetor de parâmetros  $\boldsymbol{\beta}$  para depois calcular as estimativas do vetor das médias  $\boldsymbol{\mu}$  e do vetor dos parâmetros  $\boldsymbol{\theta}$  pelas relações funcionais  $\mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$  e  $\theta_i = q(\mu_i)$ .

Se o intervalo de variação dos dados não depender de parâmetros, pode-se demonstrar para os modelos contínuos (Cox e Hinkley, 1986, Capítulo 9), que todas as derivadas de  $\int \exp[\ell(\boldsymbol{\beta})] d\mathbf{y} = 1$  podem ser computadas dentro do sinal de integração e que o ponto  $\hat{\boldsymbol{\beta}}$  correspondente ao máximo do logaritmo da função

de verossimilhança (3.1) está próximo do vetor  $\boldsymbol{\beta}$  de parâmetros verdadeiros com probabilidade próxima de 1. Para os modelos discretos, a integração é substituída pelo somatório. Isso ocorre em problemas denominados **regulares**.

Um caso importante dos MLG surge quando o vetor de parâmetros  $\boldsymbol{\theta}$  da família (2.4) e o vetor de preditores lineares  $\boldsymbol{\eta}$  em (2.6) são iguais, conduzindo às funções de ligação canônicas. Tem-se,  $\theta_i = \eta_i = \sum_{r=1}^p x_{ir}\beta_r$  para  $i = 1, \dots, n$ . As estatísticas  $S_r = \sum_{i=1}^n x_{ir}Y_i$  para  $r = 1, \dots, p$  são suficientes para os parâmetros  $\beta_1, \dots, \beta_p$  e têm dimensão mínima  $p$ . Sejam  $s_r = \sum_{i=1}^n x_{ir}y_i$  as realizações de  $S_r$ ,  $r = 1, \dots, p$ . Então, (3.1) pode ser escrita na forma:

$$\phi^{-1}\left[\sum_{r=1}^p s_r\beta_r - \sum_{i=1}^n b(\theta_i)\right] + \sum_{i=1}^n c(y_i, \phi)$$

e, portanto,  $\ell(\boldsymbol{\beta})$  tem a seguinte decomposição

$$\ell(\boldsymbol{\beta}) = \ell_1(\mathbf{s}, \boldsymbol{\beta}) + \ell_2(\mathbf{y}),$$

em que  $\ell_1(\mathbf{s}, \boldsymbol{\beta}) = \phi^{-1} \sum_{r=1}^p s_r\beta_r - \phi^{-1} \sum_{i=1}^n b(\sum_{r=1}^p x_{ir}\beta_r)$  e  $\ell_2(\mathbf{y}) = \sum_{i=1}^n c(y_i, \phi)$ .

Pelo teorema da fatoração,  $\mathbf{S} = (S_1, \dots, S_p)^T$  é suficiente de dimensão mínima  $p$  para  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  e, portanto, ocorre uma redução na dimensão das estatísticas suficientes de  $n$  (o número de observações) para  $p$  (o número de parâmetros a serem estimados). As estatísticas  $S_1, \dots, S_p$  correspondem à maior redução que os dados podem alcançar, sem qualquer perda de informação relevante para se fazer inferência sobre o vetor de parâmetros desconhecidos  $\boldsymbol{\beta}$ .

Conforme visto na Seção 2.3, as funções de ligação que fornecem estatísticas suficientes de dimensão mínima  $p$ , para as diversas distribuições são denominadas *canônicas*. A Tabela 2.6 mostra que essas funções de ligação para os modelos normal, Poisson, binomial, gama e normal inverso são  $\eta = \mu$ ,  $\eta = \log \mu$ ,  $\eta = \log[\mu/(m - \mu)]$ ,  $\eta = \mu^{-1}$  e  $\eta = \mu^{-2}$ , respectivamente.

As funções de ligação canônicas produzem propriedades estatísticas de interesse para o modelo, tais como, suficiência, facilidade de cálculo, unicidade das estimativas de máxima verossimilhança e, em alguns casos, interpretação simples e,

a princípio, pode-se trabalhar com ela quando não existirem indicativos de outra preferível. Entretanto, não existe razão para se considerarem sempre os efeitos sistemáticos como aditivos na escala dada pela função de ligação canônica. A escolha da função de ligação será vista, com mais detalhes, na Seção 4.10.

## 3.2 O algoritmo de estimação

A decisão importante na aplicação dos MLG é a escolha do trinômio: distribuição da variável resposta  $\times$  matriz modelo  $\times$  função de ligação. A seleção pode resultar de simples exame dos dados ou de alguma experiência anterior. Inicialmente, considera-se esse trinômio fixo para se obter uma descrição adequada dos dados através das estimativas dos parâmetros do modelo. Muitos métodos podem ser usados para estimar os parâmetros  $\beta'$ s, inclusive o qui-quadrado mínimo, o Bayesiano e a estimação-M. O último inclui o método de máxima verossimilhança (MV) que tem muitas propriedades ótimas, tais como, consistência e eficiência assintótica.

Neste livro, considera-se apenas o método de MV para estimar os parâmetros lineares  $\beta_1, \dots, \beta_p$  do modelo. O vetor escore é formado pelas derivadas parciais de primeira ordem do logaritmo da função de verossimilhança. Da expressão (3.1) pode-se calcular, pela regra da cadeia, o vetor escore  $\mathbf{U}(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$  de dimensão  $p$ , com elemento típico  $U_r = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_r} = \sum_{i=1}^n \frac{d\ell_i}{d\theta_i} \frac{d\theta_i}{d\mu_i} \frac{d\mu_i}{d\eta_i} \frac{\partial \eta_i}{\partial \beta_r}$ , pois

$$\ell(\boldsymbol{\beta}) = f(\theta_1, \theta_2, \dots, \theta_i, \dots, \theta_n)$$

$$\downarrow$$

$$\theta_i = \int V_i^{-1} d\mu_i = q(\mu_i)$$

$$\downarrow$$

$$\mu_i = g^{-1}(\eta_i) = h(\eta_i)$$

$$\downarrow$$

$$\eta_i = \sum_{r=1}^p x_{ir} \beta_r$$

e, sabendo-se que  $\mu_i = b'(\theta_i)$  e  $d\mu_i/d\theta_i = V_i$ , tem-se

$$U_r = \phi^{-1} \sum_{i=1}^n (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ir} \quad (3.2)$$

para  $r = 1, \dots, p$ .

A estimativa de máxima verossimilhança (EMV)  $\hat{\beta}$  do vetor de parâmetros  $\beta$  é obtida igualando-se  $U_r$  a zero para  $r = 1, \dots, p$ . Em geral, as equações  $U_r = 0$ ,  $r = 1, \dots, p$ , não são lineares e têm que ser resolvidas numericamente por processos iterativos do tipo Newton-Raphson.

O método iterativo de Newton-Raphson para a solução de uma equação  $f(x) = 0$  é baseado na aproximação de Taylor para a função  $f(x)$  na vizinhança do ponto  $x_0$ , ou seja,

$$f(x) = f(x_0) + (x - x_0)f'(x_0) = 0,$$

obtendo-se

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

ou, de uma forma mais geral,

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})},$$

sendo  $x^{(m+1)}$  o valor de  $x$  no passo  $(m+1)$ ,  $x^{(m)}$  o valor de  $x$  no passo  $m$ ,  $f(x^{(m)})$  a função  $f(x)$  avaliada em  $x^{(m)}$  e  $f'(x^{(m)})$  a derivada da função  $f(x)$  avaliada em  $x^{(m)}$ .

Considerando-se que se deseja obter a solução do sistema de equações  $\mathbf{U} = \mathbf{U}(\beta) = \partial\ell(\beta)/\partial\beta = \mathbf{0}$  e, usando-se a versão multivariada do método de Newton-Raphson, tem-se

$$\beta^{(m+1)} = \beta^{(m)} + (\mathbf{J}^{(m)})^{-1}\mathbf{U}^{(m)},$$

sendo  $\beta^{(m)}$  e  $\beta^{(m+1)}$  os vetores de parâmetros estimados nos passos  $m$  e  $(m+1)$ , respectivamente,  $\mathbf{U}^{(m)}$  o vetor escore avaliado no passo  $m$ , e  $(\mathbf{J}^{(m)})^{-1}$  a inversa da negativa da matriz de derivadas parciais de segunda ordem de  $\ell(\beta)$ , com elementos  $-\partial^2\ell(\beta)/\partial\beta_r\partial\beta_s$ , avaliada no passo  $m$ .



Quando as derivadas parciais de segunda ordem são avaliadas facilmente, o método de Newton-Raphson é bastante útil. Acontece, porém, que isso nem sempre ocorre e no caso dos MLG usa-se o método escore de Fisher que, em geral, é mais simples (coincidindo com o método de Newton-Raphson no caso das funções de ligação canônicas). Esse método envolve a substituição da matriz de derivadas parciais de segunda ordem pela matriz de valores esperados das derivadas parciais, isto é, a substituição da matriz de informação observada,  $\mathbf{J}$ , pela matriz de informação esperada de Fisher,  $\mathbf{K}$ . Logo,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{K}^{(m)})^{-1} \mathbf{U}^{(m)}, \quad (3.3)$$

sendo que  $\mathbf{K}$  tem elementos típicos dados por

$$\kappa_{r,s} = -E \left[ \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s} \right] = E \left[ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_r} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_s} \right],$$

que é a matriz de covariâncias dos  $U'_r s$ .

Multiplicando-se ambos os membros de (3.3) por  $\mathbf{K}^{(m)}$ , tem-se

$$\mathbf{K}^{(m)} \boldsymbol{\beta}^{(m+1)} = \mathbf{K}^{(m)} \boldsymbol{\beta}^{(m)} + \mathbf{U}^{(m)}. \quad (3.4)$$

O elemento típico  $\kappa_{rs}$  de  $\mathbf{K}$  é obtido de (3.2) como

$$\kappa_{r,s} = E(U_r U_s) = \phi^{-2} \sum_{i=1}^n E(Y_i - \mu_i)^2 \frac{1}{V_i^2} \left( \frac{d\mu_i}{d\eta_i} \right)^2 x_{ir} x_{is}$$

ou

$$\kappa_{r,s} = \phi^{-1} \sum_{i=1}^n w_i x_{ir} x_{is},$$

sendo  $w_i = V_i^{-1} (d\mu_i/d\eta_i)^2$  denominado função peso. Logo, a matriz de informação de Fisher para  $\boldsymbol{\beta}$  tem a forma

$$\mathbf{K} = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X},$$

sendo  $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$  uma matriz diagonal de pesos que capta a informação sobre a distribuição e a função de ligação usadas e poderá incluir também um termo

para peso a priori. No caso das funções de ligação canônicas tem-se  $w_i = V_i$ , pois  $V_i = V(\mu_i) = d\mu_i/d\eta_i$ . Note-se que a informação é inversamente proporcional ao parâmetro de dispersão.

O vetor escore  $\mathbf{U} = \mathbf{U}(\boldsymbol{\beta})$  com componentes em (3.2) pode, então, ser escrito na forma

$$\mathbf{U} = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{G} (\mathbf{y} - \boldsymbol{\mu}),$$

com  $\mathbf{G} = \text{diag} \{d\eta_1/d\mu_1, \dots, d\eta_n/d\mu_n\} = \text{diag} \{g'(\mu_1), \dots, g'(\mu_n)\}$ . Assim, a matriz diagonal  $\mathbf{G}$  é formada pelas derivadas de primeira ordem da função de ligação.

Substituindo  $\mathbf{K}$  e  $\mathbf{U}$  em (3.4) e eliminando  $\phi$ , tem-se

$$\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \boldsymbol{\beta}^{(m)} + \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{G}^{(m)} (\mathbf{y} - \boldsymbol{\mu}^{(m)}),$$

ou, ainda,

$$\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T \mathbf{W}^{(m)} [\boldsymbol{\eta}^{(m)} + \mathbf{G}^{(m)} (\mathbf{y} - \boldsymbol{\mu}^{(m)})].$$

Define-se a variável dependente ajustada  $\mathbf{z} = \boldsymbol{\eta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$ . Logo,

$$\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)}$$

ou

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)}. \quad (3.5)$$

A equação matricial (3.5) é válida para qualquer MLG e mostra que a solução das equações de MV equivale a calcular repetidamente uma regressão linear ponderada de uma variável dependente ajustada  $\mathbf{z}$  sobre a matriz  $\mathbf{X}$  usando uma função de peso  $\mathbf{W}$  que se modifica no processo iterativo. As funções de variância e de ligação entram no processo iterativo através de  $\mathbf{W}$  e  $\mathbf{z}$ . Note-se que  $\text{Cov}(\mathbf{z}) = \mathbf{G} \text{Cov}(\mathbf{Y}) \mathbf{G} = \phi \mathbf{W}^{-1}$ , isto é, os  $z_i$  não são correlacionados. É importante enfatizar que a equação iterativa (3.5) não depende do parâmetro de dispersão  $\phi$ .

A demonstração de (3.5), em generalidade, foi dada por Nelder e Wedderburn (1972). Eles generalizaram procedimentos iterativos obtidos para casos especiais dos MLG: probit (Fisher, 1935), log-lineares (Haberman, 1970) e logístico-lineares (Cox, 1972).

A variável dependente ajustada depende da derivada de primeira ordem da função de ligação. Quando a função de ligação é linear ( $\boldsymbol{\eta} = \boldsymbol{\mu}$ ), isto é, a identidade, tem-se  $\mathbf{W} = \mathbf{V}^{-1}$  em que  $\mathbf{V} = \text{diag}\{V_1, \dots, V_n\}$ ,  $\mathbf{G} = \mathbf{I}$  e  $\mathbf{z} = \mathbf{y}$ , ou seja, a variável dependente ajustada reduz-se ao vetor de observações. Para o modelo normal linear ( $\mathbf{V} = \mathbf{I}, \boldsymbol{\mu} = \boldsymbol{\eta}$ ), tornando  $\mathbf{W}$  igual à matriz identidade de dimensão  $n$ ,  $\mathbf{z} = \mathbf{y}$  e de (3.5) obtém-se que a estimativa  $\hat{\boldsymbol{\beta}}$  reduz-se à fórmula esperada  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ . Esse é o único caso em que  $\hat{\boldsymbol{\beta}}$  é calculado de forma exata sem ser necessário um procedimento iterativo.

O método usual para iniciar o processo iterativo é especificar uma estimativa inicial e sucessivamente alterá-la até que a convergência seja obtida e, portanto,  $\boldsymbol{\beta}^{(m+1)}$  aproxime-se de  $\hat{\boldsymbol{\beta}}$  quando  $m$  cresce. Note, contudo, que cada observação pode ser considerada como uma estimativa do seu valor médio, isto é,  $\mu_i^{(1)} = y_i$  e, portanto, calcula-se

$$\eta_i^{(1)} = g(\mu_i^{(1)}) = g(y_i) \quad \text{e} \quad w_i^{(1)} = \frac{1}{V(y_i)[g'(y_i)]^2}.$$

Usando-se  $\boldsymbol{\eta}^{(1)}$  como variável resposta,  $\mathbf{X}$ , a matriz do modelo, e  $\mathbf{W}^{(1)}$ , a matriz diagonal de pesos com elementos  $w_i^{(1)}$ , obtém-se o vetor

$$\boldsymbol{\beta}^{(2)} = (\mathbf{X}^T \mathbf{W}^{(1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(1)} \boldsymbol{\eta}^{(1)}.$$

O algoritmo de estimação, para  $m = 2, \dots, k$ , sendo  $k-1$  o número necessário de iterações para convergência, pode ser resumido nos seguintes passos:

(1) obter as estimativas

$$\eta_i^{(m)} = \sum_{r=1}^p x_{ir} \beta_r^{(m)} \quad \text{e} \quad \mu_i^{(m)} = g^{-1}(\eta_i^{(m)});$$

(2) obter a variável dependente ajustada

$$z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})g'(\mu_i^{(m)})$$

e os pesos

$$w_i^{(m)} = \frac{1}{V(\mu_i^{(m)})[g'(\mu_i^{(m)})]^2};$$

3) calcular

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)},$$

voltar ao passo (1) com  $\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m+1)}$  e repetir o processo até obter a convergência, definindo-se, então,  $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$ .

Dentre os muitos existentes, um critério para verificar a convergência poderia ser

$$\sum_{r=1}^p \left( \frac{\beta_r^{(m+1)} - \beta_r^{(m)}}{\beta_r^{(m)}} \right)^2 < \xi,$$

tomando-se para  $\xi$  um valor suficientemente pequeno. Em geral, esse algoritmo é robusto e converge rapidamente (menos de 10 iterações são suficientes). Entretanto, o critério do desvio é o mais usado e consiste em verificar se  $|\text{desvio}^{(m+1)} - \text{desvio}^{(m)}| < \xi$ , sendo desvio definido na Seção 4.2.

Deve-se tomar cuidado se a função  $g(\cdot)$  não é definida para alguns valores  $y_i$ . Por exemplo, se a função de ligação for dada por

$$\eta = g(\mu) = \log \mu$$

e forem observados valores  $y_i = 0$ , o processo não pode ser iniciado. Um método geral para contornar esse problema é substituir  $y$  por  $y + c$  tal que  $E[g(y + c)]$  seja o mais próxima possível de  $g(\mu)$ . Para o modelo de Poisson com função de ligação logarítmica usa-se  $c = 1/2$ . Para o modelo logístico usa-se  $c = (1 - 2\pi)/2$  e  $\pi = \mu/m$ , sendo  $m$  o índice da distribuição binomial. De uma forma geral, usando-se a expansão de Taylor até segunda ordem para  $g(y + c)$  em relação a  $g(\mu)$ , tem-se

$$g(y + c) \approx g(\mu) + (y + c - \mu)g'(\mu) + (y + c - \mu)^2 \frac{g''(\mu)}{2}$$

com valor esperado dado por

$$E[g(Y + c)] \approx g(\mu) + cg'(\mu) + \text{Var}(Y) \frac{g''(\mu)}{2}$$

que implica

$$c \approx -\frac{1}{2} \text{Var}(Y) \frac{g''(\mu)}{g'(\mu)}.$$

Para pequenas amostras, a equação (3.5) pode divergir. O número de iterações até convergência depende inteiramente do valor inicial arbitrado para  $\hat{\beta}$ , embora, geralmente, o algoritmo convirja rapidamente. A desvantagem do método tradicional de Newton-Raphson com o uso da matriz observada de derivadas de segunda ordem é que, normalmente, não converge para determinados valores iniciais.

Vários software estatísticos utilizam o algoritmo iterativo (3.5) para obter as EMV  $\hat{\beta}_1, \dots, \hat{\beta}_p$  dos parâmetros lineares do MLG, entre os quais, R, S-PLUS, SAS, GENSTAT e MATLAB.

### 3.3 Estimação em modelos especiais

Para as funções de ligação canônicas  $w = V = d\mu/d\eta$  que produzem os modelos denominados canônicos, as equações de MV têm a seguinte forma, facilmente deduzidas de (3.2),

$$\sum_{i=1}^n x_{ir} y_i = \sum_{i=1}^n x_{ir} \hat{\mu}_i$$

para  $r = 1, \dots, p$ . Em notação matricial, tem-se

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \hat{\boldsymbol{\mu}}. \quad (3.6)$$

Nesse caso, as estimativas de MV dos  $\beta$ 's são únicas. Sendo  $\mathbf{S} = (S_1, \dots, S_p)^T$  o vetor de estatísticas suficientes definidas por  $S_r = \sum_{i=1}^n x_{ir} Y_i$ , conforme descrito na Seção 3.1, e  $\mathbf{s} = (s_1, \dots, s_p)^T$  os seus valores amostrais, as equações (3.6) podem ser expressas por

$$E(\mathbf{S}; \hat{\boldsymbol{\mu}}) = \mathbf{s},$$

significando que as estimativas de MV das médias  $\mu_1, \dots, \mu_n$  nos modelos canônicos são obtidas igualando-se as estatísticas suficientes minimais aos seus valores esperados.

Se a matriz modelo corresponde a uma estrutura fatorial, consistindo somente de zeros e uns, o modelo pode ser especificado pelas margens que são as estatísticas minimais, cujos valores esperados devem igualar aos totais marginais.

As equações (3.6) são válidas para os seguintes modelos canônicos: modelo clássico de regressão, modelo log-linear, modelo logístico linear, modelo gama com função de ligação recíproca e modelo normal inverso com função de ligação recíproca ao quadrado. Para os modelos canônicos, o ajuste é feito pelo algoritmo (3.5) com  $\mathbf{W} = \text{diag}\{V_i\}$ ,  $\mathbf{G} = \text{diag}\{V_i^{-1}\}$  e variável dependente ajustada com componente típica expressa por  $z_i = \eta_i + (y_i - \mu_i)/V_i$ .

Nos modelos com respostas binárias, a variável resposta tem distribuição binomial  $B(m_i, \pi_i)$ , e o logaritmo da função de verossimilhança em (3.1) é expresso como

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left[ y_i \log \left( \frac{\mu_i}{m_i - \mu_i} \right) + m_i \log \left( \frac{m_i - \mu_i}{m_i} \right) \right] + \sum_{i=1}^n \log \binom{m_i}{y_i},$$

em que  $\mu_i = m_i \pi_i$ . É importante notar que se  $y_i = 0$ , tem-se como componente típico dessa função  $\ell_i(\boldsymbol{\beta}) = m_i \log[(m_i - \mu_i)/m_i]$  e se  $y_i = m_i$ ,  $\ell_i(\boldsymbol{\beta}) = m_i \log(\mu_i/m_i)$ .

Para o modelo logístico linear, obtém-se  $\eta_i = g(\mu_i) = \log[\mu_i/(m_i - \mu_i)]$ . As iterações em (3.5) são realizadas com matriz de pesos  $\mathbf{W} = \text{diag}\{\mu_i(m_i - \mu_i)/m_i\}$ ,  $\mathbf{G} = \text{diag}\{m_i/[\mu_i(m_i - \mu_i)]\}$  e variável dependente ajustada com componentes iguais a  $z_i = \eta_i + [m_i(y_i - \mu_i)]/[\mu_i(m_i - \mu_i)]$ . O algoritmo (3.5), em geral, converge, exceto quando ocorrem médias ajustadas próximas a zero ou ao índice  $m_i$ .

Nos modelos log-lineares para análise de observações na forma de contagens, a variável resposta tem distribuição de Poisson  $P(\mu_i)$  com função de ligação logarítmica e, portanto,  $\eta_i = \log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ ,  $i = 1, \dots, n$ . Nesse caso, as iterações em (3.5) são realizadas com matriz de pesos  $\mathbf{W} = \text{diag}\{\mu_i\}$ ,  $\mathbf{G} = \text{diag}\{\mu_i^{-1}\}$  e variável dependente ajustada com componentes iguais a  $z_i = \eta_i + (y_i - \mu_i)/\mu_i$ . Esse caso

especial do algoritmo (3.5) foi apresentado, primeiramente, por Haberman (1978).

Para análise de dados contínuos, três modelos são, usualmente, adotados com função de variância potência  $V(\mu) = \mu^\delta$  para  $\delta = 0$  (normal),  $\delta = 2$  (gama) e  $\delta = 3$  (normal inversa). Para a função de variância potência, a matriz  $\mathbf{W}$  entra no algoritmo (3.5) com a expressão típica  $\mathbf{W} = \text{diag} \{ \mu_i^{-\delta} (d\mu_i/d\eta_i)^2 \}$  sendo  $\delta$  qualquer real especificado. Outras funções de variância podem ser adotadas no algoritmo (3.5) como aquelas dos modelos de quase-verossimilhança que serão estudados na Seção 10.6. Por exemplo,  $V(\mu) = \mu^2(1 - \mu)^2$ ,  $V(\mu) = \mu + \delta\mu^2$  (binomial negativo) ou  $V(\mu) = 1 + \mu^2$  (secante hiperbólica generalizada, Seção 1.3).

O algoritmo (3.5) pode ser usado para ajustar inúmeros outros modelos, como aqueles baseados na família exponencial (1.1) que estão descritos por Cordeiro et al. (1995), bastando identificar as funções de variância e de ligação.

### 3.4 Resultados adicionais na estimação

A partir da obtenção da EMV  $\hat{\beta}$  em (3.5), podem-se calcular as estimativas de MV dos preditores lineares  $\hat{\eta} = \mathbf{X}\hat{\beta}$  e das médias  $\hat{\mu} = g^{-1}(\hat{\eta})$ . A EMV do vetor de parâmetros canônicos  $\theta$  é, simplesmente, igual a  $\hat{\theta} = q(\hat{\mu})$ .

A inversa da matriz de informação estimada em  $\hat{\beta}$  representa a estrutura de covariância assintótica de  $\hat{\beta}$ , isto é, a matriz de covariância de  $\hat{\beta}$  quando  $n \rightarrow \infty$ . Logo, a matriz de covariância de  $\hat{\beta}$  é estimada por

$$\widehat{\text{Cov}}(\hat{\beta}) = \phi(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1}, \quad (3.7)$$

em que  $\widehat{\mathbf{W}}$  é o valor da matriz de pesos  $\mathbf{W}$  avaliada em  $\hat{\beta}$ .

Intervalos de confiança assintóticos para os parâmetros  $\beta'$ s podem ser deduzidos da aproximação (3.7). Observa-se que o parâmetro de dispersão  $\phi$  é um fator multiplicativo na matriz de covariância assintótica de  $\hat{\beta}$ . Assim, se  $\widehat{\text{Var}}(\hat{\beta}_r)$  é o elemento  $(r, r)$  da matriz  $\phi(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1}$ , um intervalo de 95% de confiança para  $\beta_r$  pode ser obtido dos limites (inferior corresponde a - e superior a +)

$$\hat{\beta}_r \mp 1,96 \widehat{\text{Var}}(\hat{\beta}_r)^{1/2}.$$

Na prática, uma estimativa consistente de  $\phi$  deve ser usada nesse intervalo.

A estrutura da covariância assintótica das estimativas de MV dos preditores lineares em  $\hat{\boldsymbol{\eta}}$  é obtida diretamente de  $\text{Cov}(\hat{\boldsymbol{\eta}}) = \mathbf{X}\text{Cov}(\hat{\boldsymbol{\beta}})\mathbf{X}^T$ . Logo,

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\eta}}) = \phi \mathbf{X}(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T. \quad (3.8)$$

A matriz  $\mathbf{Z} = \{z_{ij}\} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T$  que aparece em (3.8) desempenha um papel importante na teoria assintótica dos MLG (Cordeiro, 1983; Cordeiro e McCullagh, 1991). Essa matriz surge no valor esperado da função desvio (Seção 4.2) até termos de ordem  $O(n^{-1})$  e no valor esperado da estimativa  $\hat{\boldsymbol{\beta}}$  até essa ordem.

A estrutura de covariância assintótica das estimativas de MV das médias em  $\hat{\boldsymbol{\mu}}$  pode ser calculada expandindo  $\hat{\boldsymbol{\mu}} = g^{-1}(\hat{\boldsymbol{\eta}})$  em série de Taylor. Tem-se,

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\eta} + \frac{dg^{-1}(\boldsymbol{\eta})}{d\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$$

e, portanto,

$$\text{Cov}(\hat{\boldsymbol{\mu}}) = \mathbf{G}^{-1} \text{Cov}(\hat{\boldsymbol{\eta}}) \mathbf{G}^{-1}, \quad (3.9)$$

lembrando que a matriz diagonal  $\mathbf{G} = \text{diag}\{d\eta_i/d\mu_i\}$  foi introduzida na Seção 3.2. Essa matriz é estimada por

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\mu}}) = \phi \widehat{\mathbf{G}}^{-1} \mathbf{X}(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{G}}^{-1}.$$

As matrizes  $\text{Cov}(\hat{\boldsymbol{\eta}})$  e  $\text{Cov}(\hat{\boldsymbol{\mu}})$  em (3.8) e (3.9) são de ordem  $n^{-1}$ .

Os erros-padrão  $\hat{z}_{ii}^{1/2}$  de  $\hat{\eta}_i$  e os coeficientes de correlação estimados

$$\widehat{\text{Corr}}(\hat{\eta}_i, \hat{\eta}_j) = \frac{\hat{z}_{ij}}{(\hat{z}_{ii}\hat{z}_{jj})^{1/2}},$$

dos preditores lineares estimados,  $\hat{\eta}_1, \dots, \hat{\eta}_n$ , são resultados aproximados que dependem fortemente do tamanho da amostra. Entretanto, são guias úteis de informação sobre a confiabilidade e a interdependência das estimativas dos preditores lineares, e podem, também, ser usados para obtenção de intervalos de confiança aproximados para esses parâmetros. Para alguns MLG, é possível achar uma forma fechada



para a inversa da matriz de informação e, conseqüentemente, para as estruturas de covariância assintótica das estimativas de  $\hat{\boldsymbol{\beta}}$ ,  $\hat{\boldsymbol{\eta}}$  e  $\hat{\boldsymbol{\mu}}$ .

Freqüentemente, nos modelos de análise de variância, admite-se que os dados são originados de populações com variâncias iguais. Em termos de MLG, isso implica no uso de uma função de ligação  $g(\cdot)$ , tal que  $\mathbf{W}$ , não depende da média  $\boldsymbol{\mu}$  e, portanto, que a matriz de informação seja constante. Nesse caso, pelo menos, assintoticamente, a matriz de covariância das estimativas dos parâmetros lineares é estabilizada.

Essa função de ligação é denominada *estabilizadora* e implica na constância da matriz de pesos do algoritmo de estimação. A função de ligação estabilizadora será vista (como o caso  $\delta = 1/2$ ) na Seção 8.2, mas pode ser obtida como solução da equação diferencial  $d\mu/d\eta = kd\eta/d\theta$ , sendo que  $k$  é uma constante arbitrária. Por exemplo, para os modelos gama e Poisson, as soluções dessa equação são o logaritmo e a raiz quadrada, respectivamente. Para as funções de ligação estabilizadoras, é mais fácil obter uma forma fechada para a matriz de informação, que depende inteiramente da matriz modelo, isto é, do desenho do experimento.

Em muitas situações, os parâmetros de interesse não são aqueles básicos dos MLG. Seja  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^T$  um vetor de parâmetros, em que  $\gamma_i = h_i(\boldsymbol{\beta})$ , sendo as funções  $h_i(\cdot)$ ,  $i = 1, \dots, q$ , conhecidas. Supõe-se que essas funções, em geral, não-lineares, são suficientemente bem comportadas. Seja a matriz  $q \times p$  de derivadas  $\mathbf{D} = \{\partial h_i / \partial \beta_j\}$ . As estimativas  $\hat{\gamma}_1, \dots, \hat{\gamma}_q$  podem ser calculadas diretamente de  $\hat{\gamma}_i = h_i(\hat{\boldsymbol{\beta}})$ , para  $i = 1, \dots, q$ . A matriz de covariância assintótica de  $\hat{\boldsymbol{\gamma}}$  é igual a  $\phi \mathbf{D}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{D}^T$  e deve ser estimada no ponto  $\hat{\boldsymbol{\beta}}$ . Uma aplicação será vista na Seção 10.3.

Considere, por exemplo, que após o ajuste de um MLG, tenha-se interesse em estudar as estimativas dos parâmetros  $\gamma$ 's definidos por um modelo de regressão assintótico em três parâmetros  $\beta_0$ ,  $\beta_1$  e  $\beta_2$

$$\gamma_r = \beta_0 - \beta_1 \beta_2^{z_r}, \quad r = 1, \dots, q.$$

A matriz  $D$  de dimensões  $q \times 3$  é igual, portanto, a

$$D = \begin{bmatrix} 1 & -\beta_2^{z_1} & -\beta_1\beta_2^{z_1} \log \beta_2 \\ \dots & \dots & \dots \\ 1 & -\beta_2^{z_q} & -\beta_1\beta_2^{z_q} \log \beta_2 \end{bmatrix}.$$

### 3.5 Seleção do modelo

É difícil propor uma estratégia geral para o processo de escolha de um MLG a ser ajustado às observações que se dispõe. Isso está intimamente relacionado ao problema fundamental da estatística que, segundo Fisher, é “o que se deve fazer com os dados?”.

Em geral, o algoritmo de ajuste deve ser aplicado não a um MLG isolado, mas a vários modelos de um conjunto bem amplo que deve ser, realmente, relevante para o tipo de observações que se pretende analisar. Se o processo é aplicado a um único modelo, não levando em conta possíveis modelos alternativos, existe o risco de não se obter um dos modelos mais adequados aos dados. Esse conjunto de modelos pode ser formulado de várias maneiras:

- (a) definindo uma família de funções de ligação;
- (b) considerando diferentes opções para a escala de medição;
- (c) adicionando (ou retirando) vetores colunas independentes a partir de uma matriz básica original.

Pode-se propor um conjunto de modelos para dados estritamente positivos, usando-se a família potência de funções de ligação  $\eta = g(\mu; \lambda) = (\mu^\lambda - 1)\lambda^{-1}$ , em que  $\lambda$  é um parâmetro que indexa o conjunto. Para dados reais positivos ou negativos, outras famílias podem ser definidas como  $g(\mu; \lambda) = [\exp(\lambda\mu) - 1]\lambda^{-1}$ . A estimativa de MV de  $\lambda$ , em geral, define um modelo bastante adequado, porém, muitas vezes, de difícil interpretação.

Devem-se analisar não somente os dados brutos mas procurar modelos alternativos aplicados aos dados transformados  $z = h(y)$ . O problema crucial é a escolha

da função de escala  $h(\cdot)$ . No modelo clássico de regressão, essa escolha visa a combinar, aproximadamente, normalidade e constância da variância do erro aleatório, bem como, aditividade dos efeitos sistemáticos. Entretanto, não existe nenhuma garantia que tal escala  $h(\cdot)$  exista, nem mesmo que produza algumas das propriedades desejadas.

Para dar uma ilustração, suponha que as observações  $y$  representam contagens, com estrutura de Poisson de média  $\mu$  e que os efeitos sistemáticos dos fatores que classificam os dados sejam multiplicativos. A transformação  $\sqrt{y}$  produz, para valores grandes de  $\mu$ ,  $E(\sqrt{Y}) \doteq \sqrt{\mu}$  e  $\text{Var}(\sqrt{Y}) \doteq 1/4$ , sendo os erros de ordem  $\mu^{-1/2}$ . Portanto, a escala raiz quadrada implica na constância da variância dos dados transformados. Entretanto, se o objetivo é obter uma normalidade aproximada, uma escala preferida deve ser  $h(y) = \sqrt[3]{y^2}$ , pois o coeficiente de assimetria padronizado de  $Y^{2/3}$  é de ordem  $\mu^{-1}$ , ao invés de  $\mu^{-1/2}$  para  $Y$  ou  $Y^{1/2}$ . Ainda a escala  $h(y) = \log y$  é bem melhor para obtenção da aditividade dos efeitos sistemáticos.

Não existe nenhuma escala que produza os três efeitos desejados, embora a escala definida por  $h(y) = (3y^{1/2} - 3y^{1/6}\mu^{1/3} + \mu^{1/2})/6$ , se  $y \neq 0$  e  $h(y) = [-(2\mu)^{1/2} + \mu^{-1/2}]/6$ , se  $y = 0$ , conduza a simetria e constância da variância (McCullagh e Nelder, 1989, Capítulo 6). As probabilidades nas extremidades da distribuição de Poisson podem ser calculadas por  $P(Y \geq y) \doteq 1 - \Phi[h(y - 1/2)]$ , com erro de ordem  $\mu^{-1}$ , em que  $\Phi(\cdot)$  é a f.d.a. da normal reduzida.

Nos MLG, o fator escala não é tão crucial como no modelo clássico de regressão, pois constância da variância e normalidade não são essenciais para a distribuição da variável resposta e, ainda, pode-se achar uma estrutura aditiva aproximada de termos para representar a média da distribuição, usando uma função de ligação apropriada, diferente da escala de medição dos dados. Entretanto, não são raros os casos em que os dados devem ser primeiramente transformados para se obter um MLG com um bom ajuste.

A terceira maneira de selecionar o modelo é através da definição do conjunto de variáveis independentes a serem incluídas na estrutura linear. Considere um certo

número de possíveis covariáveis  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}$ , em que cada vetor coluna  $\mathbf{x}^{(r)}$  é de dimensão  $n$ , definindo um conjunto amplo de  $2^m$  modelos. O objetivo é selecionar um modelo de  $p \leq m$  covariáveis, cujos valores ajustados expliquem adequadamente os dados. Se  $m$  for muito grande, torna-se impraticável o exame de todos esses  $2^m$  modelos, mesmo considerando os avanços da tecnologia computacional.

Um processo simples de seleção é de natureza seqüencial, adicionando (ou eliminando) covariáveis (uma de cada vez) a partir de um modelo original até se obterem modelos adequados. Esse método seqüencial tem várias desvantagens, tais como:

- (a) modelos potencialmente úteis podem não ser descobertos, se o procedimento é finalizado numa etapa anterior, para o qual nenhuma covariável isolada mostrou-se razoável para ser explorada;
- (b) modelos similares (ou mesmo melhores) baseados em subconjuntos de covariáveis, distantes das covariáveis em exame, podem não ser considerados.

Devido aos avanços recentes da estatística computacional, os métodos seqüenciais (“stepwise methods”) foram substituídos por procedimentos ótimos de busca de modelos. O procedimento de busca examina, sistematicamente, somente os modelos mais promissores de determinada dimensão  $k$  e, baseado em algum critério, exibe os resultados de ajuste dos melhores modelos de  $k$  variáveis exploratórias, com  $k$  variando no processo de 1 até o tamanho  $p$  do subconjunto final de modelos considerados bons.

O analista deve sempre tentar eliminar a priori modelos medíocres, observando a estrutura dos dados, por meio de análises exploratórias gráficas. Na seleção do modelo, sempre será feito um balanço entre o grau de complexidade e a qualidade de ajuste do modelo.

### 3.6 Considerações sobre a função de verossimilhança

Expandindo a função suporte  $\ell = \ell(\boldsymbol{\beta})$ , dada na Seção 3.2, em série multivariada de Taylor ao redor de  $\hat{\boldsymbol{\beta}}$  e notando que  $\mathbf{U}(\hat{\boldsymbol{\beta}}) = 0$ , obtém-se, aproximadamente,

$$\hat{\ell} - \ell \doteq \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\mathbf{J}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}), \quad (3.10)$$

em que  $\hat{\ell} = \ell(\hat{\boldsymbol{\beta}})$  e  $\hat{\mathbf{J}}$  é a informação observada (Seção 3.2) em  $\hat{\boldsymbol{\beta}}$ . Essa equação aproximada revela que a diferença entre o suporte máximo e o suporte num ponto arbitrário, que pode ser vista como a quantidade de informação dos dados sobre  $\boldsymbol{\beta}$ , é proporcional a  $\hat{\mathbf{J}}$  (isto é, à informação observada no ponto  $\hat{\boldsymbol{\beta}}$ ). O determinante de  $\hat{\mathbf{J}}$  ( $|\hat{\mathbf{J}}|$ ) pode ser interpretado geometricamente como a curvatura esférica da superfície suporte no seu máximo. A forma quadrática do lado direito de (3.10) aproxima a superfície suporte por um parabolóide, passando pelo seu ponto de máximo, com a mesma curvatura esférica da superfície nesse ponto. O recíproco de  $|\hat{\mathbf{J}}|$  mede a variabilidade de  $\boldsymbol{\beta}$  ao redor da EMV  $\hat{\boldsymbol{\beta}}$ . E, como esperado, quanto maior a informação sobre  $\boldsymbol{\beta}$  menor será a dispersão de  $\boldsymbol{\beta}$  ao redor de  $\hat{\boldsymbol{\beta}}$ .

A interpretação geométrica desses conceitos é melhor compreendida no caso uniparamétrico, pois (3.10) reduz-se à equação de uma parábola  $\ell \doteq \hat{\ell} - \frac{1}{2}(\beta - \hat{\beta})^2 \hat{\mathbf{J}}$ . Uma inspeção gráfica mostrará que essa parábola aproxima a curva suporte, coincidindo no ponto máximo e tendo a mesma curvatura dessa curva em  $\hat{\beta}$ , revelando ainda que quanto maior a curvatura, menor a variação de  $\beta$  em torno de  $\hat{\beta}$ .

A equação (3.10) implica que a função de verossimilhança  $L = L(\boldsymbol{\beta})$  num ponto qualquer  $\boldsymbol{\beta}$  segue, aproximadamente, a expressão

$$L \doteq \hat{L} \exp \left[ -\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \hat{\mathbf{J}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right], \quad (3.11)$$

em que  $\hat{L}$  é a função de verossimilhança avaliada em  $\hat{\boldsymbol{\beta}}$ , que representa a forma da curva normal multivariada com média  $\hat{\boldsymbol{\beta}}$  e estrutura de covariância igual a  $\hat{\mathbf{J}}^{-1}$ .

Através dessa aproximação, pode-se, então, tratar o vetor de parâmetros como se fosse um vetor de variáveis aleatórias tendo distribuição normal multivariada com média igual à EMV  $\hat{\beta}$  e estrutura de covariância  $\hat{\mathbf{J}}^{-1}$ . Quando a função suporte for quadrática, a verossimilhança terá a forma normal. A forma de  $L$  se aproximará mais da normal quando  $n$  tender para infinito.

O lado direito de (3.11) é bem interpretado no contexto Bayesiano. Considere qualquer função densidade a priori não-nula para  $\beta$ , por exemplo,  $\pi(\beta)$ . Usando o teorema de Bayes, pode-se escrever a função densidade a posteriori de  $\beta$  como proporcional a  $L\pi(\beta)$ . Quando  $n \rightarrow \infty$ , pois  $\pi(\beta)$  não depende de  $n$ , a função densidade a posteriori de  $\beta$  segue de (3.11) com uma constante de proporcionalidade adequada, e, então, converge para a distribuição normal multivariada  $N(\hat{\beta}, \hat{\mathbf{J}}^{-1})$ . Uma demonstração matemática dessa convergência está fora dos objetivos desse texto. No caso uniparamétrico, a variabilidade de  $\beta$  fica restrita ao intervalo  $|\beta - \hat{\beta}| \leq 3\hat{J}^{-1/2}$  com probabilidade próxima de um.

A fórmula (3.11) mostra a decomposição da função de verossimilhança, pelo menos para  $n$  grande, estabelecendo, pelo teorema da fatoração, a suficiência assintótica da EMV. Conclui-se que, embora as estimativas de MV não sejam necessariamente suficientes para os parâmetros do modelo, essa suficiência será alcançada quando a dimensão do vetor de observações tender para infinito.

Citam-se, aqui, algumas propriedades da matriz de informação. Seja  $\mathbf{K}_y(\beta)$  a informação sobre um vetor paramétrico  $\beta$  contida nos dados  $\mathbf{y}$  obtidos de certo experimento. A informação é aditiva para amostras  $\mathbf{y}$  e  $\mathbf{z}$  independentes, isto é,  $\mathbf{K}_{y+z}(\beta) = \mathbf{K}_y(\beta) + \mathbf{K}_z(\beta)$ . Como  $\hat{\mathbf{U}} = \mathbf{U}(\hat{\beta}) = 0$ , segue-se a relação aproximada (por expansão multivariada de Taylor):

$$\hat{\beta} - \beta \doteq \mathbf{J}^{-1}\mathbf{U} \quad (3.12)$$

entre a EMV e a função escore  $\mathbf{U} = \mathbf{U}(\beta)$  e a informação observada  $\mathbf{J} = \mathbf{J}(\beta)$  avaliadas no ponto  $\beta$  próximo de  $\hat{\beta}$ .

O método de Newton-Raphson, introduzido na Seção 3.2, de cálculo da EMV

consiste em usar a equação (3.12) iterativamente. Obtém-se uma nova estimativa  $\beta^{(m+1)}$  a partir de uma anterior  $\beta^{(m)}$  por meio de

$$\beta^{(m+1)} = \beta^{(m)} + \mathbf{J}^{(m)-1} \mathbf{U}^{(m)}, \quad (3.13)$$

em que quantidades avaliadas na  $m$ -ésima iteração do procedimento iterativo são indicadas com o superescrito  $(m)$ . O processo é, então, repetido a partir de  $\beta^{(m+1)}$  até a distância entre  $\beta^{(m+1)}$  e  $\beta^{(m)}$  se tornar desprezível ou menor do que uma quantidade pequena especificada. Geometricamente, uma iteração do método equivale a ajustar um parabolóide à superfície suporte em  $\beta^{(m)}$ , tendo o mesmo gradiente e curvatura da superfície nesse ponto, e, então, obter o ponto máximo do parabolóide que corresponderá à estimativa atualizada  $\beta^{(m+1)}$ . Quando  $\beta$  é um escalar, a equação (3.13) reduz-se a  $\beta^{(m+1)} = \beta^{(m)} - U^{(m)} / U'^{(m)}$ , sendo  $U' = dU/d\beta$ , que representa o método das tangentes bastante usado para calcular a solução de uma equação não-linear  $\hat{U} = 0$ .

A seqüência  $\{\beta^{(m)}; m \geq 1\}$  gerada depende fundamentalmente do vetor inicial  $\beta^{(1)}$ , dos valores amostrais e do modelo estatístico e, em determinadas situações, em que  $n$  é pequeno, pode revelar irregularidades específicas aos valores amostrais obtidos do experimento e, portanto, pode não convergir e mesmo divergir da EMV  $\hat{\beta}$ . Mesmo quando há convergência, se a função de verossimilhança tem raízes múltiplas, não se tem garantia de que o procedimento converge para a raiz correspondente ao maior valor absoluto da verossimilhança. No caso uniparamétrico, se a estimativa inicial  $\beta^{(1)}$  for escolhida próxima de  $\hat{\beta}$  e se  $J^{(m)}$  para  $m \geq 1$  for limitada por um número real positivo, existirá uma chance apreciável que essa seqüência convirja para  $\hat{\beta}$ .

A expressão (3.12) tem uma forma alternativa equivalente, assintoticamente, pois pela lei dos grandes números  $\mathbf{J}$  deve convergir para  $\mathbf{K}$  quando  $n \rightarrow \infty$ . Assim, substituindo a informação observada em (3.12) pela esperada, obtém-se a aproximação

$$\hat{\beta} - \beta \doteq \mathbf{K}^{-1} \mathbf{U}. \quad (3.14)$$

O procedimento iterativo baseado em (3.14) é denominado *método escore* de Fisher para parâmetros, isto é,  $\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + \mathbf{K}^{(m)^{-1}} \mathbf{U}^{(m)}$ , como foi explicitado na equação (3.3). O aspecto mais trabalhoso dos dois esquemas iterativos é a inversão das matrizes  $\mathbf{J}$  e  $\mathbf{K}$ . Ambos os procedimentos são muito sensíveis em relação à estimativa inicial  $\boldsymbol{\beta}^{(1)}$ . Se o vetor  $\boldsymbol{\beta}^{(1)}$  for uma estimativa consistente, ambos os métodos convergirão em apenas um passo para uma estimativa eficiente, assintoticamente.

Existe evidência empírica que o método de Fisher é melhor em termos de convergência do que o método de Newton-Raphson. Ainda, tem a vantagem de usufruir (através da matriz de informação) de características específicas ao modelo estatístico. Ademais, em muitas situações, é mais fácil determinar a inversa de  $\mathbf{K}$  em forma fechada do que a inversa de  $\mathbf{J}$ , sendo a primeira menos sensível às variações de  $\boldsymbol{\beta}$  do que a segunda. Nesse sentido,  $\mathbf{K}$  pode ser considerada, aproximadamente, constante em todo o processo iterativo, requerendo que a inversão seja feita apenas uma vez. Uma vantagem adicional do método escore é que  $\mathbf{K}^{-1}$  é usada para obter aproximações de primeira ordem para as variâncias e covariâncias das estimativas  $\hat{\beta}_1, \dots, \hat{\beta}_p$ .

Os procedimentos iterativos descritos são casos especiais de uma classe de algoritmos iterativos para maximizar o logaritmo da função de verossimilhança  $\ell(\boldsymbol{\beta})$ . Essa classe tem a forma

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - s^{(m)} \mathbf{Q}^{(m)} \mathbf{U}^{(m)}, \quad (3.15)$$

em que  $s^{(m)}$  é um escalar,  $\mathbf{Q}^{(m)}$  é uma matriz quadrada que determina a direção da mudança de  $\boldsymbol{\beta}^{(m)}$  para  $\boldsymbol{\beta}^{(m+1)}$  e  $\mathbf{U}^{(m)}$  é o vetor gradiente do logaritmo da função de verossimilhança  $\ell(\boldsymbol{\beta})$ , com todas essas quantidades variando no processo iterativo. Os algoritmos iniciam num ponto  $\boldsymbol{\beta}^{(1)}$  e procedem, por meio da expressão (3.15), para calcular aproximações sucessivas para a EMV  $\hat{\boldsymbol{\beta}}$ . Vários algoritmos nessa classe são discutidos por Judge et al. (1985). Nos procedimentos iterativos de Newton-Raphson e escore de Fisher,  $s^{(m)}$  é igual a um, e a matriz de direção  $\mathbf{Q}^{(m)}$  é igual à inversa da matriz Hessiana e à inversa do valor esperado dessa matriz, respectivamente. Esses



dois procedimentos devem ser iniciados a partir de uma estimativa consistente a fim de se garantir convergência para  $\hat{\beta}$ . A escolha do melhor algoritmo em (3.15) é função da geometria do modelo em consideração e, em geral, não existe um algoritmo superior aos demais em qualquer espectro amplo de problemas de estimação.

### 3.7 Exercícios

1. Definir o algoritmo de estimação dado em (3.5) para os modelos canônicos relativos às distribuições estudadas na Seção 1.3 (Tabela 1.1), calculando  $\mathbf{W}$ ,  $\mathbf{G}$  e  $\mathbf{z}$ .
2. Definir o algoritmo de estimação dado em (3.5), calculando  $\mathbf{W}$ ,  $\mathbf{G}$  e  $\mathbf{z}$  para os modelos normal, gama, normal inverso e Poisson com função de ligação potência  $\eta = \mu^\lambda$ ,  $\lambda$  conhecido (Cordeiro, 1986). Para o modelo normal, considere, ainda, o caso da função de ligação logarítmica, isto é,  $\eta = \log \mu$ .
3. Definir o algoritmo (3.5), calculando  $\mathbf{W}$ ,  $\mathbf{G}$  e  $\mathbf{z}$ , para o modelo binomial com função de ligação  $\eta = \log\{[(1 - \mu)^{-\lambda} - 1]\lambda^{-1}\}$ ,  $\lambda$  conhecido. Obter ainda as formas do algoritmo para os modelos (c) e (d), definidos na Tabela 2.7 da Seção 2.5.1.
4. Considere a estrutura linear  $\eta_i = \beta x_i$ ,  $i = 1, \dots, n$ , com um único parâmetro  $\beta$  desconhecido e função de ligação  $\eta = (\mu^\lambda - 1)\lambda^{-1}$ ,  $\lambda$  conhecido. Calcular a estimativa de MV de  $\beta$  para os modelos normal, Poisson, gama, normal inverso e binomial negativo. Fazer o mesmo para o modelo binomial com função de ligação dada no Exercício 3. Obter ainda as estimativas no caso de  $x_1 = x_2 = \dots = x_n$ .
5. Para os modelos e funções de ligação citados no Exercício 4, calcular as estimativas de MV de  $\alpha$  e  $\beta$ , considerando a estrutura linear  $\eta_i = \alpha + \beta x_i$ ,  $i = 1, \dots, n$ . Obter ainda a estrutura de covariância aproximada dessas estimativas.
6. Caracterizar as distribuições log-normal e log-gama no contexto dos MLG, definindo o algoritmo de ajuste desses modelos com a função de ligação potência

$\eta = \mu^\lambda$ ,  $\lambda$  conhecido.

**7.** Formular o procedimento iterativo de cálculo das estimativas de mínimos quadrados dos parâmetros  $\beta$ 's nos MLG, que equivale a minimizar  $(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ , em que  $\mathbf{V} = \text{diag}\{V_1, \dots, V_n\}$ , com relação a  $\boldsymbol{\beta}$ . Como aplicação, obter essas estimativas nos Exercícios 4 e 5.

**8.** Calcular a forma da matriz de informação para o modelo log-linear associado a uma tabela de contingência com dois fatores sem interação, sendo uma observação por cela. Fazer o mesmo para o modelo de Poisson com função de ligação raiz quadrada. Qual a grande vantagem desse último modelo?

**9.** Calcular a forma da matriz de informação para os parâmetros  $\beta$ 's no modelo de classificação de um fator  $A$  com  $p$  níveis  $g(\mu_i) = \eta_i = \beta + \beta_i^A$ , com  $\beta_+^A = 0$ , considerando a variável resposta como normal, gama, normal inversa e Poisson. Determinar as matrizes de covariância assintótica das estimativas  $\hat{\boldsymbol{\beta}}$ ,  $\hat{\boldsymbol{\eta}}$  e  $\hat{\boldsymbol{\mu}}$ . Obter as expressões para o cálculo dessas estimativas.

**10.** Como o modelo binomial do Exercício 3 poderia ser ajustado se  $\lambda$  fosse desconhecido? E os modelos do Exercício 4, ainda  $\lambda$  desconhecido?

**11.** Sejam variáveis aleatórias  $Y_i$  com distribuições de Poisson  $P(\mu_i)$ ,  $i = 1, \dots, n$ , supostas independentes. Define-se  $f(\cdot)$  como uma função diferenciável tal que  $[f(\mu + x\mu^{1/2}) - f(\mu)]/\mu^{1/2}f'(\mu) = x + O(\mu^{-1/2})$ , para todo  $x$  com  $\mu \rightarrow \infty$ . Demonstrar que a variável aleatória  $[f(Y_i) - f(\mu_i)]/[\mu_i^{1/2}f'(\mu_i)]$  converge em distribuição para a distribuição normal  $N(0, 1)$  quando  $\mu_i \rightarrow \infty$ . Provar, ainda, que a parte do logaritmo da função de verossimilhança que só depende dos  $\mu_i$ 's tende, assintoticamente, para  $-2^{-1} \sum_{i=1}^n [f(y_i) - f(\mu_i)]^2 / [y_i f'(y_i)]^2$  quando  $\mu_i \rightarrow \infty$ ,  $i = 1, \dots, n$ , em que  $y_1, \dots, y_n$  são as realizações dos  $Y$ 's.

**12.** A probabilidade de sucesso  $\pi = \mu/m$  de uma distribuição binomial  $B(m, \pi)$

depende de uma variável  $x$  de acordo com a relação  $\pi = F(\alpha + \beta x)$ , em que  $F(\cdot)$  é uma função de distribuição acumulada especificada. Admite-se que para os valores  $x_1, \dots, x_n$  de  $x$ ,  $m_1, \dots, m_n$  ensaios independentes foram realizados, sendo obtidas proporções de sucessos  $p_1, \dots, p_n$ , respectivamente. Comparar as estimativas  $\hat{\alpha}$  e  $\hat{\beta}$  para as escolhas de  $F(\cdot)$ : probito, logística, arcsen  $\sqrt{\cdot}$  e complemento log-log.

**13.** Considere a f.d.p.  $f(y) = \exp(-\sum_{i=1}^r \alpha_i y^i)$  com parâmetros  $\alpha_1, \dots, \alpha_r$  desconhecidos. Demonstrar que as estimativas de MV e dos momentos desses parâmetros coincidem.

**14.** Considere um modelo log-gama com componente sistemático  $\log \mu_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta}$  e parâmetro de dispersão  $\phi$ . Mostre que

$$E[\log(Y_i)] = \alpha^* + \mathbf{x}_i^T \boldsymbol{\beta}$$

e

$$\text{Var}[\log(Y_i)] = \psi'(\phi^{-1}),$$

em que  $\alpha^* = \alpha + \psi(\phi^{-1}) + \log \phi$ . Seja  $\tilde{\boldsymbol{\beta}}$  o estimador de mínimos quadrados de  $\boldsymbol{\beta}$  obtido ajustando-se um modelo de regressão linear aos dados transformados  $\log y_i$ ,  $i = 1, \dots, n$ . Mostre que  $\tilde{\boldsymbol{\beta}}$  é um estimador consistente para  $\boldsymbol{\beta}$ .



# Capítulo 4

## Métodos de Inferência

### 4.1 Distribuição dos estimadores dos parâmetros

No modelo clássico de regressão em que a variável resposta tem distribuição normal e a função de ligação é a identidade, as distribuições dos estimadores dos parâmetros e das estatísticas usadas para verificar o ajuste do modelo aos dados podem ser determinadas exatamente. Em geral, porém, a obtenção de distribuições exatas nos MLG é muito complicada e resultados assintóticos são, rotineiramente, usados. Esses resultados, porém, dependem de algumas condições de regularidade e do número de observações independentes mas, em particular, para os MLG essas condições são satisfeitas (Fahrmeir e Kaufmann, 1985).

A idéia básica é que se  $\hat{\theta}$  é um estimador consistente para um parâmetro  $\theta$  e  $\text{Var}(\hat{\theta})$  é a variância desse estimador, então, para amostras grandes, tem-se:

i)  $\hat{\theta}$  é assintoticamente imparcial;

ii) a estatística

$$Z_n = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \rightarrow Z \text{ quando } n \rightarrow \infty, \text{ sendo que } Z \sim N(0, 1)$$

ou, de forma equivalente,

$$Z_n^2 = \frac{(\hat{\theta} - \theta)^2}{\text{Var}(\hat{\theta})} \rightarrow Z^2 \text{ quando } n \rightarrow \infty, \text{ sendo que } Z^2 \sim \chi_1^2.$$

Se  $\hat{\boldsymbol{\theta}}$  é um estimador consistente de um vetor  $\boldsymbol{\theta}$  de  $p$  parâmetros, tem-se, assintoticamente, que

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \mathbf{V}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \chi_p^2,$$

sendo  $\mathbf{V}$  a matriz de variâncias e covariâncias de  $\hat{\boldsymbol{\theta}}$ , suposta não-singular. Se  $\mathbf{V}$  é singular, usa-se uma matriz inversa generalizada ou, então, uma reparametrização de forma a se obter uma nova matriz de variâncias e covariâncias não-singular.

Considere-se um MLG definido por uma distribuição em (2.4), uma estrutura linear (2.6) e uma função de ligação (2.7). As propriedades do estimador  $\hat{\boldsymbol{\beta}}$  dos parâmetros lineares do modelo em relação à existência, finitude e unicidade serão apresentadas na Seção 8.1. Em geral, não é possível obter distribuições exatas para os estimadores de MV e para as estatísticas de testes usadas nos MLG e trabalha-se com resultados assintóticos. As condições de regularidade que garantem esses resultados são satisfeitas para os MLG. É fato conhecido que os estimadores de MV têm poucas propriedades que são satisfeitas para todos os tamanhos de amostras, como, por exemplo, suficiência e invariância. As propriedades assintóticas de segunda-ordem de  $\hat{\boldsymbol{\beta}}$ , como o viés de ordem  $O(n^{-1})$  e a sua matriz de covariância de ordem  $O(n^{-2})$ , foram estudadas por Cordeiro e McCullagh (1991) e Cordeiro (2004a,b,c), respectivamente.

Seja o vetor escore  $\mathbf{U}(\boldsymbol{\beta}) = \partial \ell(\boldsymbol{\beta}) / \partial \boldsymbol{\beta}$  como definido na Seção 3.2. Como o vetor escore tem valor esperado zero e estrutura de covariância igual à matriz de informação  $\mathbf{K}$  em problemas regulares (Cox e Hinkley, 1986, Capítulo 9), tem-se de (3.2) que  $E\{\mathbf{U}(\boldsymbol{\beta})\} = \mathbf{0}$  e

$$\text{Cov}[\mathbf{U}(\boldsymbol{\beta})] = E[\mathbf{U}(\boldsymbol{\beta})\mathbf{U}(\boldsymbol{\beta})^T] = E\left[\frac{-\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}\right] = \mathbf{K}. \quad (4.1)$$

Conforme demonstrado na Seção 3.2, a matriz de informação para  $\boldsymbol{\beta}$  nos MLG é dada por  $\mathbf{K} = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X}$ .

O teorema central do limite aplicado a  $\mathbf{U}(\boldsymbol{\beta})$  (que equivale a uma soma de variáveis aleatórias independentes) implica que a distribuição assintótica de  $\mathbf{U}(\boldsymbol{\beta})$

é normal  $p$ -variada, isto é,  $N_p(\mathbf{0}, \mathbf{K})$ . Para amostras grandes, a estatística escore definida pela forma quadrática  $S_R = \mathbf{U}(\boldsymbol{\beta})^T \mathbf{K}^{-1} \mathbf{U}(\boldsymbol{\beta})$  tem, aproximadamente, distribuição  $\chi_p^2$  supondo o modelo, com o vetor de parâmetros  $\boldsymbol{\beta}$  especificado, verdadeiro.

De uma forma resumida têm-se, a seguir, algumas propriedades para o estimador  $\hat{\boldsymbol{\beta}}$ :

i) O estimador  $\hat{\boldsymbol{\beta}}$  é assintoticamente não viesado, isto é, para amostras grandes  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ . Suponha que o logaritmo da função de verossimilhança tem um único máximo em  $\hat{\boldsymbol{\beta}}$  que está próximo do verdadeiro valor de  $\boldsymbol{\beta}$ . A expansão em série multivariada de Taylor para o vetor escore  $\mathbf{U}(\hat{\boldsymbol{\beta}})$  em relação a  $\boldsymbol{\beta}$ , até termos de primeira ordem, substituindo-se a matriz de derivadas parciais de segunda ordem por  $-\mathbf{K}$ , é dada por

$$\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{U}(\boldsymbol{\beta}) - \mathbf{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{0},$$

pois  $\hat{\boldsymbol{\beta}}$  é a solução do sistema de equações  $\mathbf{U}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ . As variáveis aleatórias  $\mathbf{U}(\boldsymbol{\beta})$  e  $\mathbf{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$  diferem por quantidades estocásticas de ordem  $O_p(1)$ . Portanto, tem-se até ordem  $n^{-1/2}$  em probabilidade

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{K}^{-1} \mathbf{U}(\boldsymbol{\beta}), \quad (4.2)$$

desde que  $\mathbf{K}$  seja não-singular.

A expressão aproximada (4.2) é de grande importância para a determinação de propriedades do estimador de MV  $\hat{\boldsymbol{\beta}}$ . As variáveis aleatórias  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$  e  $\mathbf{K}^{-1} \mathbf{U}(\boldsymbol{\beta})$  diferem por variáveis aleatórias de ordem  $n^{-1}$  em probabilidade. Tem-se, então, que

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{K}^{-1} E[\mathbf{U}(\boldsymbol{\beta})] = \mathbf{0} \Rightarrow E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta},$$

pois  $E[\mathbf{U}(\boldsymbol{\beta})] = \mathbf{0}$  e, portanto,  $\hat{\boldsymbol{\beta}}$  é um estimador imparcial para  $\boldsymbol{\beta}$  (pelo menos assintoticamente). Na realidade,  $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + O(n^{-1})$ , sendo que o termo de ordem  $O(n^{-1})$  foi obtido por Cordeiro e McCullagh (1991). Mais recentemente, Cordeiro e Barroso (2007) obtiveram o termo de ordem  $O(n^{-2})$  da expansão de  $E(\hat{\boldsymbol{\beta}})$ .

ii) Denotando-se  $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{U}$  e usando (4.2) e (4.1) tem-se que a matriz de variâncias e covariâncias de  $\hat{\boldsymbol{\beta}}$ , para amostras grandes, é dada por

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = \mathbf{K}^{-1}E(\mathbf{U}\mathbf{U}^T)\mathbf{K}^{-1T} = \mathbf{K}^{-1}\mathbf{K}\mathbf{K}^{-1} = \mathbf{K}^{-1},$$

pois  $\mathbf{K}^{-1}$  é simétrica. Na realidade,  $\text{Cov}(\hat{\boldsymbol{\beta}}) = \mathbf{K}^{-1} + O(n^{-2})$ , sendo o termo matricial de ordem  $O(n^{-2})$  dado em Cordeiro (2004a).

iii) Para amostras grandes, tem-se a aproximação

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{K} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_p^2 \quad (4.3)$$

ou, de forma equivalente,

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \mathbf{K}^{-1}), \quad (4.4)$$

ou seja,  $\hat{\boldsymbol{\beta}}$  tem distribuição assintótica normal  $p$ -variada, que é a base para a construção de testes e intervalos de confiança para os parâmetros lineares de um MLG. Para modelos lineares com variáveis respostas com distribuição normal, (4.3) e (4.4) são resultados exatos. Fahrmeir e Kaufmann (1985), em um artigo bastante matemático, desenvolvem condições gerais que garantem a consistência e a normalidade assintótica do estimador de MV  $\hat{\boldsymbol{\beta}}$  nos MLG.

Para amostras pequenas, como citado em i), o estimador  $\hat{\boldsymbol{\beta}}$  é viesado e torna-se necessário computar o viés de ordem  $n^{-1}$  que pode ser apreciável. Também, para  $n$  não muito grande, como visto em ii), a estrutura de covariância das estimativas de MV dos parâmetros lineares difere de  $\mathbf{K}^{-1}$ . Uma demonstração rigorosa dos resultados assintóticos (4.3) e (4.4) exige argumentos do teorema central do limite adaptado ao vetor escore  $\mathbf{U}(\boldsymbol{\beta})$  e da lei fraca dos grandes números aplicada à matriz de informação  $\mathbf{K}$ . Pode-se, então, demonstrar, com mais rigor, a normalidade assintótica de  $\hat{\boldsymbol{\beta}}$ , com média igual ao parâmetro verdadeiro  $\boldsymbol{\beta}$  desconhecido, e com matriz de covariância consistentemente estimada por  $\hat{\mathbf{K}}^{-1} = \phi(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1}$  em que  $\widehat{\mathbf{W}}$  é a matriz de pesos  $\mathbf{W}$  avaliada em  $\hat{\boldsymbol{\beta}}$ .

Para as distribuições binomial e de Poisson  $\phi = 1$ . Se o parâmetro de dispersão  $\phi$  for constante para todas as observações e desconhecido afetará a matriz



de covariância assintótica  $\hat{\mathbf{K}}^{-1}$  de  $\hat{\boldsymbol{\beta}}$  mas não o valor de  $\hat{\boldsymbol{\beta}}$ . Na prática, se  $\phi$  for desconhecido, deverá ser substituído por alguma estimativa consistente (Seção 4.4).

A distribuição assintótica normal  $p$ -variada  $N_p(\boldsymbol{\beta}, \mathbf{K}^{-1})$  de  $\hat{\boldsymbol{\beta}}$  é a base da construção de testes e intervalos de confiança, em amostras grandes, para os parâmetros lineares dos MLG. O erro dessa aproximação para a distribuição de  $\hat{\boldsymbol{\beta}}$  é de ordem  $n^{-1}$  em probabilidade, significando que os cálculos de probabilidade baseados na função de distribuição acumulada da distribuição normal assintótica  $N_p(\boldsymbol{\beta}, \mathbf{K}^{-1})$ , apresentam erros de ordem de magnitude  $n^{-1}$ .

A distribuição assintótica normal  $p$ -variada  $N_p(\boldsymbol{\beta}, \mathbf{K}^{-1})$  será uma boa aproximação para a distribuição de  $\hat{\boldsymbol{\beta}}$ , se o logaritmo da função de verossimilhança for razoavelmente uma função quadrática. Pelo menos, assintoticamente, todos os logaritmos das funções de verossimilhança têm essa forma. Para amostras pequenas, isso pode não ocorrer para  $\boldsymbol{\beta}$ , embora possa existir uma reparametrização  $\boldsymbol{\gamma} = h(\boldsymbol{\beta})$ , que conduza o logaritmo da função de verossimilhança a uma função, aproximadamente, quadrática. Assim, testes e regiões de confiança mais precisos poderão ser baseados na distribuição assintótica de  $\hat{\boldsymbol{\gamma}} = h(\hat{\boldsymbol{\beta}})$ .

Anscombe (1964), no caso de um único parâmetro  $\beta$ , obtém uma parametrização geral que elimina a assimetria do logaritmo da função de verossimilhança. A solução geral é da forma

$$\gamma = h(\beta) = \int \exp \left[ \frac{1}{3} \int v(\beta) d\beta \right] d\beta,$$

em que  $v(\beta) = \frac{d^3 \ell(\beta)}{d\beta^3} \left[ \frac{d^2 \ell(\beta)}{d\beta^2} \right]^{-1}$ . Essa transformação tem a propriedade de anular a derivada de terceira ordem do logaritmo da função de verossimilhança, em relação a  $\gamma$ , e, portanto, eliminar a principal contribuição da assimetria.

Para os MLG, a assimetria do logaritmo da função de verossimilhança é eliminada usando uma função de ligação apropriada, como será explicado na Seção 8.2 (caso  $\delta = 1/3$ ). Usando-se a expressão de Anscombe (1964), obtém-se, diretamente,  $\eta = \int \exp \left\{ \int b'''(\theta) / [3b''(\theta)] d\theta \right\} d\theta = \int b''(\theta)^{1/3} d\theta$ , a função de ligação que simetriza  $\ell(\boldsymbol{\beta})$ . Quando a função de ligação é diferente desse caso, e se  $\boldsymbol{\beta}$ ,

tem dimensão maior do que 1, em geral, não é possível anular a assimetria. Em particular, parametrizações componente a componente  $\gamma_i = h(\beta_i)$ ,  $i = 1, \dots, p$ , não apresentam um bom aperfeiçoamento na forma do logaritmo da função de verossimilhança, a menos que as variáveis explanatórias sejam, mutuamente, ortogonais (Pregibon, 1979).

**Exemplo 4.1:** Seja  $Y_1, \dots, Y_n$  uma amostra aleatória de uma distribuição normal  $N(\mu_i, \sigma^2)$ , sendo que  $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ . Considerando a função de ligação identidade, isto é,  $\eta_i = \mu_i$ , tem-se que  $g'(\mu_i) = 1$ . Além disso,  $V_i = 1$  e, portanto,  $w_i = 1$ . Logo, a matriz de informação é dada por

$$\mathbf{K} = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} = \sigma^{-2} \mathbf{X}^T \mathbf{X}$$

e a variável dependente ajustada fica sendo  $z_i = y_i$ .

Portanto, o algoritmo de estimação (3.5) reduz-se a

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

e, desde que  $\mathbf{X}^T \mathbf{X}$  tenha inversa,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \quad (4.5)$$

que é a solução usual de mínimos quadrados para o modelo clássico de regressão. Tem-se, então,

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}$$

e

$$\begin{aligned} \text{Cov}(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T] \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}, \end{aligned}$$

pois  $E[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T] = \sigma^2 \mathbf{I}$ .

Como  $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$  e o vetor  $\hat{\boldsymbol{\beta}}$  dos estimadores de MV é uma transformação linear do vetor  $\mathbf{y}$  em (4.5), o vetor  $\hat{\boldsymbol{\beta}}$  tem distribuição normal  $p$ -variada  $N_p(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$  exatamente. Logo, tem-se, exatamente, que

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^T)\mathbf{K}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_p^2,$$

sendo  $\mathbf{K} = \sigma^{-2}\mathbf{X}^T\mathbf{X}$  a matriz de informação.

Os erros-padrão dos estimadores de MV  $\hat{\beta}_1, \dots, \hat{\beta}_p$  são iguais às raízes quadradas dos elementos da diagonal de  $\hat{\mathbf{K}}^{-1}$  e podem fornecer informações valiosas sobre a exatidão desses estimadores. Usa-se aqui a notação  $\mathbf{K}^{-1} = \{\kappa^{r,s}\}$  para a inversa da matriz de informação em que, aproximadamente,  $\text{Cov}(\hat{\beta}_r, \hat{\beta}_s) = \kappa^{r,s}$ . Então, com nível de confiança de 95%, intervalos de confiança para os parâmetros  $\beta'_r$ s podem ser deduzidos de

$$\hat{\beta}_r \mp 1,96\sqrt{\hat{\kappa}^{r,r}},$$

em que  $\hat{\kappa}^{r,r} = \widehat{\text{Var}}(\hat{\beta}_r)$  é o valor de  $\kappa^{r,r}$  em  $\hat{\boldsymbol{\beta}}$ . Nas Seções 4.6 e 4.7 serão apresentados testes e regiões de confiança construídos com a função desvio.

A correlação estimada  $\hat{\rho}_{rs}$  entre as estimativas  $\hat{\beta}_r$  e  $\hat{\beta}_s$  segue como

$$\hat{\rho}_{rs} = \widehat{\text{Corr}}(\hat{\beta}_r, \hat{\beta}_s) = \frac{\hat{\kappa}^{r,s}}{\sqrt{\hat{\kappa}^{r,r}\hat{\kappa}^{s,s}}},$$

sendo obtida diretamente da inversa da matriz de informação  $\mathbf{K}$  avaliada em  $\hat{\boldsymbol{\beta}}$ . Essas correlações permitem verificar, pelo menos aproximadamente, a interdependência dos  $\hat{\beta}'_r$ s.

## 4.2 Função desvio e estatística de Pearson generalizada

O ajuste de um modelo a um conjunto de observações  $\mathbf{y}$  pode ser tratado como uma maneira de se substituir  $\mathbf{y}$  por um conjunto de valores estimados  $\hat{\boldsymbol{\mu}}$  para um modelo com um número, relativamente pequeno, de parâmetros. Logicamente, os

$\hat{\mu}'s$  não serão exatamente iguais aos  $y's$ , e a questão, então, que aparece é em quanto eles diferem. Isso porque, uma discrepância pequena pode ser tolerável enquanto que uma discrepância grande, não.

Assim, admitindo-se uma combinação satisfatória da distribuição da variável resposta e da função de ligação, o objetivo é determinar quantos termos são necessários na estrutura linear para uma descrição razoável dos dados. Um número grande de variáveis explanatórias pode conduzir a um modelo que explique bem os dados mas com um aumento de complexidade na interpretação. Por outro lado, um número pequeno de variáveis explanatórias pode conduzir a um modelo de interpretação fácil, porém, que se ajuste pobremente aos dados. O que se deseja na realidade é um modelo intermediário, entre um modelo muito complicado e um modelo pobre em ajuste.

Considerando  $n$  observações, a elas podem ser ajustados modelos contendo até  $n$  parâmetros. O modelo mais simples é o **modelo nulo** que tem um único parâmetro, representado por um valor  $\mu$  comum a todos os dados. A matriz do modelo, então, reduz-se a um vetor coluna, formado de 1's. Esse modelo atribui toda a variação entre os  $y's$  ao componente aleatório. No modelo nulo, o valor comum para todas as médias dos dados é igual à média amostral, isto é,  $\bar{y} = \sum_{i=1}^n y_i/n$ , mas não representa a estrutura dos dados. No outro extremo, está o **modelo saturado** ou **completo** que tem  $n$  parâmetros especificados pelas médias  $\mu_1, \dots, \mu_n$  linearmente independentes, ou seja, correspondendo a uma matriz modelo igual à matriz identidade de ordem  $n$ . O modelo saturado tem, então,  $n$  parâmetros, um para cada observação, e as estimativas de MV das médias são  $\tilde{\mu}_i = y_i$ , para  $i = 1, \dots, n$ . O til é colocado para diferir das estimativas de MV do MLG com matriz modelo  $\mathbf{X}$ , de dimensões  $n \times p$ , com  $p < n$ . O modelo saturado atribui toda a variação dos dados ao componente sistemático e, assim, ajusta-se perfeitamente, reproduzindo os próprios dados.

Na prática, o modelo nulo é muito simples e o saturado é não-informativo, pois não sumariza os dados, mas, simplesmente, os repete. Existem dois outros mo-

delos, não tão extremos, quanto os modelos nulo e saturado: o **modelo minimal** que contém o menor número de termos necessário para o ajuste, e o **modelo maximal** que inclui o maior número de termos, que pode ser considerado. Os termos desses modelos extremos são, geralmente, obtidos por interpretações a priori da estrutura dos dados. Em geral, trabalha-se com modelos encaixados, e o conjunto de matrizes dos modelos pode, então, ser formado pela inclusão sucessiva de termos ao modelo minimal até se chegar ao modelo maximal. Qualquer modelo com  $p$  parâmetros linearmente independentes, situado entre os modelos minimal e maximal, é chamado de *modelo sob pesquisa* ou *modelo corrente*.

Determinados parâmetros têm que estar no modelo como é o caso, por exemplo, de efeitos de blocos em planejamento de experimentos ou então, totais marginais fixados em tabelas de contingência para análise de observações na forma de contagens. Assim, considerando-se um experimento em blocos casualizados, com tratamentos no esquema fatorial com 2 fatores, têm-se os modelos:

$$\text{nulo: } \eta_i = \mu$$

$$\text{minimal: } \eta_i = \mu + \beta_\ell$$

$$\text{maximal: } \eta_i = \mu + \beta_\ell + \alpha_j + \gamma_k + (\alpha\gamma)_{jk}$$

$$\text{saturado: } \eta_i = \mu + \beta_\ell + \alpha_j + \gamma_k + (\alpha\gamma)_{jk} + (\beta\alpha)_{\ell j} + (\beta\gamma)_{\ell k} + (\beta\alpha\gamma)_{\ell jk},$$

sendo  $\mu$  o efeito associado à média geral;  $\beta_\ell$  o efeito associado ao bloco  $\ell$ ,  $\ell = 1, \dots, b$ ;  $\alpha_j$  o efeito associado ao  $j$ -ésimo nível do fator A;  $\gamma_k$  o efeito associado ao  $k$ -ésimo nível do fator B;  $(\alpha\gamma)_{jk}$ ,  $(\beta\alpha)_{\ell j}$ ,  $(\beta\gamma)_{\ell k}$ ,  $(\beta\alpha\gamma)_{\ell jk}$  os efeitos associados às interações. O modelo saturado inclui, nesse caso, todas as interações com blocos que não são de interesse prático.

Em geral, trabalha-se com modelos encaixados e o conjunto de matrizes dos modelos pode, então, ser formado pela adição sucessiva de termos ao modelo minimal até se chegar ao modelo maximal. O problema é determinar a utilidade de um parâmetro extra no modelo corrente (sob pesquisa) ou, então, verificar a falta de ajuste induzida pela omissão dele. A fim de discriminar entre modelos, medidas de discrepância devem ser introduzidas para medir o ajuste de um modelo.

Nelder e Wedderburn (1972) propuseram, como medida de discrepância, a “deviance” (traduzida como **desvio** por Cordeiro (1986)), com expressão dada por

$$S_p = 2(\hat{\ell}_n - \hat{\ell}_p),$$

sendo  $\hat{\ell}_n$  e  $\hat{\ell}_p$  os máximos do logaritmo da função de verossimilhança para os modelos saturado e corrente (sob pesquisa), respectivamente. Vê-se que o modelo saturado é usado como base de medida do ajuste de um modelo sob pesquisa (modelo corrente). Do logaritmo da função de verossimilhança (3.1) obtém-se:

$$\hat{\ell}_n = \phi^{-1} \sum_{i=1}^n [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)] + \phi^{-1} \sum_{i=1}^n c(y_i, \phi)$$

e

$$\hat{\ell}_p = \phi^{-1} \sum_{i=1}^n [y_i \hat{\theta}_i - b(\hat{\theta}_i)] + \phi^{-1} \sum_{i=1}^n c(y_i, \phi),$$

sendo  $\tilde{\theta}_i = q(y_i)$  e  $\hat{\theta}_i = q(\hat{\mu}_i)$  as estimativas de MV do parâmetro canônico sob os modelos saturado e corrente, respectivamente.

Então, tem-se,

$$S_p = \phi^{-1} D_p = 2\phi^{-1} \sum_{i=1}^n [y_i(\tilde{\theta}_i - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\tilde{\theta}_i)], \quad (4.6)$$

em que  $S_p$  e  $D_p$  são denominados de desvio escalonado e desvio, respectivamente. O desvio  $D_p$  é função apenas dos dados  $y$  e das médias ajustadas  $\hat{\mu}$ . O desvio escalonado  $S_p$  depende de  $D_p$  e do parâmetro de dispersão  $\phi$ . Pode-se, ainda, escrever

$$S_p = \phi^{-1} \sum_{i=1}^n d_i^2,$$

sendo que  $d_i^2$  mede a diferença dos logaritmos das funções de verossimilhança observada e ajustada, para a observação  $i$  correspondente, e é chamado **componente do desvio**. A soma deles mede a discrepância total entre as duas funções de verossimilhança na escala logarítmica. É portanto, uma medida da distância dos valores ajustados  $\hat{\mu}'s$  em relação às observações  $y's$ , ou de forma equivalente, do modelo

corrente em relação ao modelo saturado. Verifica-se que o desvio equivale a uma constante menos duas vezes o máximo do logaritmo da função de verossimilhança para o modelo corrente, isto é,

$$S_p = 2\hat{\ell}_n - 2\hat{\ell}_p = \text{constante} - 2\hat{\ell}_p.$$

Assim, um modelo bem (mal) ajustado aos dados, com uma verossimilhança máxima grande (pequena), tem um pequeno (grande) desvio. Entretanto, um grande número de variáveis explanatórias, visando reduzir o desvio, significa um grau de complexidade na interpretação do modelo. Procuram-se, na prática, modelos simples com desvios moderados, situados entre os modelos mais complicados e os que se ajustam mal aos dados.

O desvio é computado facilmente para qualquer MLG a partir da estimativa de MV de  $\boldsymbol{\mu}$  dada por  $\hat{\boldsymbol{\mu}} = g^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}})$ . O desvio é sempre maior do que ou igual a zero, e à medida que variáveis explanatórias entram no componente sistemático, o desvio decresce até se tornar zero para o modelo saturado. Para o teste, define-se o número de graus de liberdade do desvio do modelo por  $\nu = n - p$ , isto é, como o número de observações menos o posto da matriz do modelo sob pesquisa. Em alguns casos especiais, como nos modelos normal e log-linear, o desvio torna-se igual a estatísticas comumente usadas nos testes de ajuste.

**Exemplo 4.2:** Seja  $Y_1, \dots, Y_n$  uma amostra aleatória de uma distribuição normal  $N(\mu_i, \sigma^2)$ , sendo que  $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ . Tem-se,  $\phi = \sigma^2$ ,  $\theta_i = \mu_i$  e  $b(\theta_i) = \frac{\theta_i^2}{2} = \frac{\mu_i^2}{2}$ . Logo,

$$\begin{aligned} S_p &= \frac{1}{\sigma^2} \sum_{i=1}^n 2 \left[ y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right] = \frac{1}{\sigma^2} \sum_{i=1}^n (2y_i^2 - 2\hat{\mu}_i y_i - y_i^2 + \hat{\mu}_i^2) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 = \frac{SQRes}{\sigma^2} \end{aligned}$$

que coincide com a estatística clássica SQRes com  $(n - p)$  graus de liberdade dividida por  $\sigma^2$ .

**Exemplo 4.3:** Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias representando contagens de sucessos em amostras independentes de tamanhos  $m_i$ . Suponha que  $Y_i \sim B(m_i, \pi_i)$ ,  $\phi = 1$ ,  $\theta_i = \log\left(\frac{\mu_i}{m_i - \mu_i}\right)$  e  $b(\theta_i) = m_i \log(1 + e^{\theta_i}) = -m_i \log\left(\frac{m_i - \mu_i}{m_i}\right)$ .

Logo,

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

ou ainda,

$$S_p = 2 \sum_{i=1}^n \left[ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right) \right].$$

Essa expressão é válida para  $0 < y_i < m_i$ . Se  $y_i = 0$  ou  $y_i = m_i$ , o  $i$ -ésimo termo de  $S_p$  deve ser substituído por  $2m_i \log[m_i/(m_i - \hat{\mu}_i)]$  ou  $2m_i \log(m_i/\hat{\mu}_i)$ , respectivamente (Paula, 2004). Se  $m_i = 1$ , isto é,  $Y_i \sim \text{Bernoulli}(\pi_i)$  e a função de ligação considerada é a logística, a função desvio é apenas uma função das observações e, portanto, não é informativa com relação ao ajuste do modelo aos dados. O mesmo é válido para as funções de ligação probit e complemento log-log.

Para o modelo de Poisson, o desvio tem a forma

$$S_p = 2 \left[ \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \sum_{i=1}^n \frac{\hat{\mu}_i}{y_i} \right]$$

e, em particular, para os modelos log-lineares a segunda soma é igual a zero, desde que a matriz  $\mathbf{X}$  tenha uma coluna de 1's (Exercício 5 da Seção 4.11). Nesse caso, o desvio é igual à razão de verossimilhanças (chamada de  $G^2$  ou  $Y^2$ ), que é, geralmente, usada nos testes de hipóteses em tabelas de contingência.

Para o modelo gama ( $\theta = -\mu^{-1}$ ) com média  $\mu$  e parâmetro de dispersão  $\phi$  ( $= \text{Var}(Y)/E(Y)^2$ ), o desvio tem a forma

$$S_p = 2\phi^{-1} \sum_{i=1}^n \left[ \log\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} \right],$$



que pode ainda ser simplificada em alguns casos especiais (Exercício 6 da Seção 4.11). Se algum componente é igual a zero, segundo Paula (2004), pode-se substituir  $D_p$  por

$$D_p = 2c(y) + 2 \sum_{i=1}^n \left( \log \hat{\mu}_i + \frac{y_i}{\hat{\mu}_i} \right),$$

sendo  $c(y)$  uma função arbitrária, porém limitada. Pode ser usada, por exemplo, a expressão  $c(y) = \sum_{i=1}^n \frac{y_i}{1 + y_i}$ .

Na Tabela 4.1 apresentam-se as funções desvios para os principais modelos.

Tabela 4.1: Funções desvios para alguns modelos.

Modelo	Desvio
Normal	$D_p = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2$
Binomial	$D_p = 2 \sum_{i=1}^n \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \log \left( \frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right]$
Poisson	$D_p = 2 \sum_{i=1}^n \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]$
Binomial negativo	$D_p = 2 \sum_{i=1}^n \left[ y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) + (y_i + k) \log \left( \frac{\hat{\mu}_i + k}{y_i + k} \right) \right]$
Gama	$D_p = 2 \sum_{i=1}^n \left[ \log \left( \frac{\hat{\mu}_i}{y_i} \right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]$
Normal inverso	$D_p = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$

Quanto melhor for o ajuste do MLG aos dados tanto menor será o valor do desvio  $D_p$ . Assim, um modelo bem ajustado aos dados, terá uma métrica  $\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|$  pequena, sendo essa métrica definida na escala da função desvio.

Uma maneira de se conseguir a diminuição do desvio é aumentar o número de parâmetros, o que, porém, significa um aumento do grau de complexidade na interpretação do modelo. Na prática, procuram-se modelos simples com desvios

moderados, situados entre os modelos mais complicados e os que se ajustam mal às observações. Para testar a adequação de um MLG, o valor calculado do desvio com  $n - p$  graus de liberdade, sendo  $p$  o posto da matriz do modelo, deve ser comparado com o percentil de alguma distribuição de probabilidade de referência. Para o modelo normal com função de ligação identidade, assumindo-se que o modelo usado é verdadeiro e que  $\sigma^2$  é conhecido, tem-se o resultado exato

$$S_p = \frac{D_p}{\sigma^2} \sim \chi_{n-p}^2.$$

Entretanto, para modelos normais com outras funções de ligação, esse resultado é apenas uma aproximação. Em alguns casos especiais da matriz modelo, com delineamentos experimentais simples, considerando-se as distribuições exponencial (caso especial da gama) e normal inversa, também, podem ser obtidos resultados exatos. No geral, porém, apenas alguns resultados assintóticos estão disponíveis e, em alguns casos, o desvio, não tem distribuição  $\chi_{n-p}^2$ , nem mesmo assintoticamente. O desvio corrigido por uma correção de Bartlett proposta para os MLG por Cordeiro (1983, 1987, 1995) tem sido usado para melhorar a sua aproximação pela distribuição  $\chi_{n-p}^2$  de referência. Com efeito, o desvio modificado  $S_p = (n - p)S_p / \hat{E}(S_p)$ , em que a correção de Bartlett é dada por  $(n - p) / \hat{E}(S_p)$  quando  $E(S_p)$  é determinada até termos de ordem  $O(n^{-1})$ , sendo  $\hat{E}(S_p)$  o valor de  $E(S_p)$  avaliada em  $\hat{\mu}$ , é melhor aproximado pela distribuição  $\chi_{n-p}^2$  de referência do que o desvio  $S_p$ , conforme comprovam os estudos de simulação de Cordeiro (1993).

Assumindo-se que o modelo corrente é verdadeiro, para o modelo binomial, quando  $n$  é fixo e  $m_i \rightarrow \infty$ ,  $\forall i$  (não vale quando  $m_i \pi_i (1 - \pi_i)$  permanece limitado) e para o modelo de Poisson, quando  $\mu_i \rightarrow \infty$ ,  $\forall i$ , tem-se que (lembre-se que  $\phi = 1$ ):  $S_p = D_p$  é, aproximadamente, distribuído como  $\chi_{n-p}^2$ .

Nos casos em que  $S_p$  depende do parâmetro de dispersão  $\phi$ , Jørgensen (1987) mostra que  $S_p \xrightarrow{D} \chi_{n-p}^2$  quando  $\phi \rightarrow 0$ , isto é, quando o parâmetro de dispersão é pequeno, a aproximação  $\chi_{n-p}^2$  para  $S_p$  é satisfatória. Para o modelo gama, a aproximação da distribuição de  $S_p$  por  $\chi_{n-p}^2$  será tanto melhor quanto mais próximo

de um estiver o parâmetro de dispersão. Em geral, porém, não se conhece  $\phi$ , que precisa ser substituído por uma estimativa consistente (Seções 4.4 e 4.5).

Na prática, contenta-se em testar um MLG comparando-se o valor de  $S_p$  com os percentis da distribuição  $\chi^2_{n-p}$ . Assim, quando

$$S_p = \phi^{-1} D_p \leq \chi^2_{n-p;\alpha},$$

ou seja,  $S_p$  é inferior ao valor crítico  $\chi^2_{n-p;\alpha}$  da distribuição  $\chi^2_{n-p}$ , pode-se considerar que existem evidências, a um nível aproximado de  $100\alpha\%$  de significância, que o modelo proposto está bem ajustado aos dados. Ou ainda, se o valor de  $D_p$  for próximo do valor esperado  $n - p$  de uma distribuição  $\chi^2_{n-p}$ , pode ser um indicativo de que o modelo ajustado aos dados é adequado.

O desvio  $D_p$  pode funcionar como um critério de parada do algoritmo de ajuste descrito em (3.5) e, após a convergência, o seu valor com o correspondente número de graus de liberdade podem ser computados.

Uma outra medida da discrepância do ajuste de um modelo a um conjunto de dados é a estatística de Pearson generalizada  $X_p^2$  cuja expressão é dada por

$$X_p^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \quad (4.7)$$

sendo  $V(\hat{\mu}_i)$  a função de variância estimada sob o modelo que está sendo ajustado aos dados.

Para respostas com distribuição normal,  $\frac{X_p^2}{\sigma^2} = \frac{SQRes}{\sigma^2}$  e, então,

$$X_p^2 \sim \sigma^2 \chi^2_{n-p},$$

sendo esse resultado exato somente se a função de ligação for a identidade e  $\sigma^2$  conhecido.

Para dados provenientes das distribuições binomial e de Poisson, em que  $\phi = 1$ ,  $X_p^2$  é a estatística original de Pearson, comumente usada na análise dos modelos logístico e log-linear em tabelas multidimensionais, e que tem a forma

$$X_p^2 = \sum_{i=1}^n \frac{(o_i - e_i)^2}{e_i},$$

sendo  $o_i$  a frequência observada e  $e_i$  a frequência esperada.

Para as distribuições não-normais, têm-se apenas resultados assintóticos para  $X_p^2$ , isto é, a distribuição  $\chi_{n-p}^2$  pode ser usada, somente, como uma aproximação para a distribuição de  $X_p^2$ , que em muitos casos pode ser inadequada. Além disso,  $X_p^2$  tem como desvantagem o fato de tratar as observações simetricamente. Note-se que para o modelo normal  $X_p^2 = D_p$ .

**Exemplo 4.4:** Considere os dados do Exemplo 2.1 da Seção 2.2. A variável resposta tem distribuição binomial, isto é,  $Y_i \sim B(m_i, \pi_i)$ . Adotando-se a função de ligação logística (canônica) e o preditor linear dado por uma regressão linear simples, isto é,

$$\eta_i = \log \left( \frac{\mu_i}{m_i - \mu_i} \right) = \beta_0 + \beta_1 d_i,$$

tem-se  $S_p = D_p = 10,26$  e  $X_p^2 = 9,70$  com 4 graus de liberdade. Da tabela da distribuição  $\chi^2$ , tem-se  $\chi_{4;0,05}^2 = 9,49$  e  $\chi_{4;0,01}^2 = 13,29$ , indicando que existem evidências, a um nível de significância entre 5% e 1% de probabilidade, que o modelo logístico linear ajusta-se razoavelmente a esse conjunto de dados. Necessita-se, porém, adicionalmente, do teste da hipótese  $H_0 : \beta_1 = 0$ , uma análise de resíduos e de medidas de diagnóstico.

A expressão (4.7) da estatística de Pearson generalizada tem uma forma equivalente dada em termos da variável dependente ajustada do algoritmo (3.5) para o modelo canônico. Tem-se,

$$X_p^2 = (\mathbf{z} - \hat{\boldsymbol{\eta}})^T \widehat{\mathbf{W}} (\mathbf{z} - \hat{\boldsymbol{\eta}}).$$

O desvio  $S_p$  tem a grande vantagem como medida de discrepância por ser aditivo para um conjunto de modelos encaixados, enquanto  $X_p^2$ , em geral, não tem essa propriedade, apesar de ser preferido em relação ao desvio, em muitos casos, por facilidade de interpretação.

### 4.3 Análise do desvio e seleção de modelos

A análise do desvio (“Analysis of the Deviance” - ANODEV) é uma generalização da análise de variância para os MLG, visando obter, a partir de uma seqüência de modelos encaixados, cada um incluindo mais termos do que os anteriores, os efeitos de fatores, covariáveis e suas interações. Utiliza-se o desvio como uma medida de discrepância do modelo e forma-se uma tabela de diferenças de desvios.

Seja  $M_{p_1}, M_{p_2}, \dots, M_{p_r}$  uma seqüência de modelos encaixados de dimensões respectivas  $p_1 < p_2 < \dots < p_r$ , matrizes dos modelos  $\mathbf{X}_{p_1}, \mathbf{X}_{p_2}, \dots, \mathbf{X}_{p_r}$  e desvios  $D_{p_1}, D_{p_2}, \dots, D_{p_r}$ , tendo os modelos a mesma distribuição e a mesma função de ligação. Essas desigualdades entre os desvios, em geral, não se verificam para a estatística de Pearson  $X_p^2$  generalizada e, por essa razão, a comparação de modelos encaixados é feita, principalmente, usando-se a função desvio. Assim, para o caso de um ensaio inteiramente casualizado, com  $r$  repetições e tratamentos no esquema fatorial, com  $a$  níveis para o fator  $A$  e  $b$  níveis para o fator  $B$ , obtêm-se os resultados mostrados na Tabela 4.2.

Tabela 4.2: Um exemplo de construção de uma tabela de Análise de Desvio.

Modelo	gl	desvio	Dif. de desvios	Dif. de gl	Significado
Nulo	$rab - 1$	$D_1$			
			$D_1 - D_A$	a-1	$A$ ignorando $B$
A	$a(rb - 1)$	$D_A$			
			$D_A - D_{A+B}$	b-1	$B$ incluído $A$
A+B	$a(rb - 1) - (b - 1)$	$D_{A+B}$			
			$D_{A+B} - D_{A*B}$	(a-1)(b-1)	Interação $AB$
					incluídos $A$ e $B$
A+B+A.B	$ab(r - 1)$	$D_{A*B}$			
			$D_{A*B}$		Resíduo
Saturado	0	0			

Dois termos  $A$  e  $B$  são ortogonais se a redução que  $A$  (ou  $B$ ) causa no desvio  $D_p$  é a mesma, esteja  $B$  (ou  $A$ ) incluído, ou não em  $M_p$ . Em geral, para os MLG ocorre a não-ortogonalidade dos termos e a interpretação da tabela ANODEV é mais complicada do que a ANOVA usual.

Sejam os modelos encaixados  $M_q$  e  $M_p$  ( $M_q \subset M_p$ ,  $q < p$ ), com  $q$  e  $p$  parâmetros, respectivamente. A estatística  $D_q - D_p$  com  $(p - q)$  graus de liberdade é interpretada como uma medida de variação dos dados, explicada pelos termos que estão em  $M_p$  e não estão em  $M_q$ , incluídos os efeitos dos termos em  $M_q$  e ignorando quaisquer efeitos dos termos que não estão em  $M_p$ . Tem-se, assintoticamente, para  $\phi$  conhecido

$$S_q - S_p = \phi^{-1}(D_q - D_p) \sim \chi_{p-q}^2,$$

que é simplesmente a estatística da razão de verossimilhanças (Seção 4.6).

Nesses termos, quando o modelo com menor número de parâmetros ( $q$ ) é verdadeiro,  $S_q - S_p$  tem distribuição assintótica  $\chi_{p-q}^2$ . Entretanto, cada desvio isolado não é distribuído, assintoticamente, como qui-quadrado. O teorema de Wilks (1937) exige que os espaços de parâmetros, segundo os modelos nulo e alternativo, sejam de dimensão fixa, enquanto  $n$  cresce e, portanto, não se aplica ao desvio isolado, cujo modelo alternativo é o saturado de dimensão  $n$ .

Se  $\phi$  é desconhecido, deve-se obter uma estimativa  $\hat{\phi}$  consistente, de preferência baseada no modelo maximal (com  $m$  parâmetros), e a inferência pode ser baseada na estatística  $F$ , dada por

$$F = \frac{(D_q - D_p)/(p - q)}{\hat{\phi}} \sim F_{p-q, n-m}.$$

Para a distribuição normal, tem-se

$$\frac{(SQRes_q - SQRes_p)/(p - q)}{SQRes_m/(n - m)} \sim F_{p-q, n-m},$$

sendo a distribuição  $F$  exata para o modelo normal linear.

**Exemplo 4.5:** Considere os dados do Exemplo 2.1 da Seção 2.2. A variável resposta tem distribuição binomial, isto é,  $Y_i \sim B(m_i, \pi_i)$ . Adotando-se a função de ligação logística (canônica) e o preditor linear dado por uma regressão linear simples, isto é,

$$\eta_i = \log \left( \frac{\mu_i}{m_i - \mu_i} \right) = \beta_0 + \beta_1 d_i,$$

dois modelos encaixados podem ser propostos para a análise desses dados, a saber:

- a) o modelo nulo:  $\eta_i = \beta_0$  e
- b) o modelo de regressão linear:  $\eta_i = \beta_0 + \beta_1 d_i$ .

Tabela 4.3: Desvios e  $X^2$  residuais obtidos para dois modelos encaixados ajustados aos dados da Tabela 2.1.

Modelo	g.l.	Desvios	$X^2$
$\eta_i = \beta_0$	5	163,74	135,70
$\eta_i = \beta_0 + \beta_1 d_i$	4	10,26	9,70

A Tabela 4.3 apresenta os desvios e os valores da estatística de Pearson generalizada e seus respectivos números de graus de liberdade (g.l.), e a Tabela 4.4, a análise do desvio correspondente.

Tabela 4.4: Análise do Desvio, considerando o modelo logístico linear ajustado aos dados da Tabela 2.1.

Causa de Variação	g.l.	Desvios	Valor $p$
Regressão linear	1	153,48	$< 0,0001$
Resíduo	4	10,26	
Total	5	163,74	

$\chi^2_{1;0,05} = 3,84$ ;  $\chi^2_{1;0,01} = 6,64$

O exame da Tabela 4.3, confirmando o que já foi visto no Exemplo 4.4, mostra que existem evidências, a um nível de significância entre 5% e 1% de probabilidade, que o modelo logístico linear ajusta-se razoavelmente a esse conjunto de dados, mas rejeita-se o modelo nulo. Pelo exame da Tabela 4.4, rejeita-se a hipótese nula  $H_0 : \beta_1 = 0$ , confirmando a adequação do modelo logístico linear. Necessita-se, porém, adicionalmente, de uma análise de resíduos e de diagnósticos.

Tem-se, ainda, que  $\hat{\beta}_0 = -3,226[s(\hat{\beta}_0) = 0,3699]$  e  $\hat{\beta}_1 = 0,6051[s(\hat{\beta}_1) = 0,0678]$ . O número esperado de insetos mortos  $\hat{\mu}_i$  para a dose  $d_i$  é dado por

$$\hat{\mu}_i = m_i \frac{\exp(-3,226 + 0,6051d_i)}{1 + \exp(-3,226 + 0,6051d_i)}.$$

Na Figura 4.1 estão representados a curva do modelo ajustado e os valores observados. Um programa simples em linguagem R (R Development Core Team, 2006) para a obtenção desses resultados é dado no Apêndice.

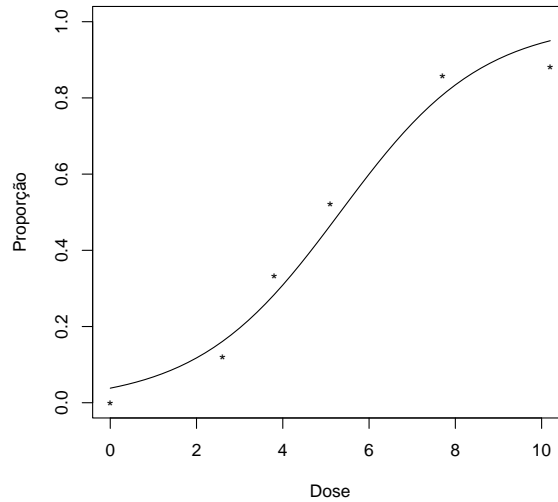


Figura 4.1: Valores observados e curva ajustada pelo modelo logístico linear aos dados da Tabela 2.1.



## 4.4 Estimação do parâmetro de dispersão

Para as distribuições binomial e Poisson tem-se que o parâmetro de dispersão  $\phi = 1$ . Quando  $\phi$  é desconhecido (distribuições normal, normal inversa e gama), admite-se que seja o mesmo para todas as observações, isto é, constante. Necessária se faz sua estimação para a obtenção (conforme visto na Seção 3.4) dos erros-padrão dos  $\hat{\beta}'s$ , intervalos de confiança e testes de hipóteses para os  $\beta's$  etc. Os métodos mais usados para a estimação de  $\phi$  são: método do desvio, método de Pearson e método de máxima verossimilhança.

O método do desvio é baseado na aproximação  $\chi^2_{n-p}$  para o desvio escalonado dado pela equação (4.6). Para um modelo bem ajustado às observações, espera-se, portanto, que o desvio escalonado  $S_p$  tenha valor esperado igual a  $n - p$ . Assim, obtém-se a estimativa de  $\phi$

$$\hat{\phi}_d = \frac{D_p}{n - p}, \quad (4.8)$$

em que o desvio  $D_p$  é obtido de (4.6) como função das observações  $y$  e dos valores ajustados  $\hat{\mu}$ . O estimador  $\hat{\phi}_d$  é, aproximadamente, não viesado para os modelos normal e normal inverso. Para o modelo normal linear  $\hat{\phi}_d = \sum (y_i - \hat{\mu}_i)^2 / (n - p)$  é o estimador usual não-viesado de  $\sigma^2$ . Para os modelos gama e normal inverso, as expressões correspondentes dos desvios  $D_p$  estão na Tabela 4.1, permitindo obter  $\hat{\phi}_d$  de (4.8).

O método de Pearson é baseado na aproximação da distribuição da estatística de Pearson  $X_p^2$  generalizada, dada em (4.7), dividida por  $\phi$ , pela distribuição  $\chi^2_{n-p}$ . Obtém-se, assim, a estimativa de  $\phi$

$$\hat{\phi}_P = \frac{1}{n - p} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \quad (4.9)$$

Para o modelo normal  $\hat{\phi}_d = \hat{\phi}_P$ . Para os demais modelos contínuos, esses estimadores diferem em valor. Os estimadores  $\hat{\phi}_P$  para os modelos gama e normal inverso são obtidos de (4.9) fazendo-se  $V(\mu) = \mu^2$  e  $V(\mu) = \mu^3$ , respectivamente.

O método de máxima verossimilhança é sempre possível em teoria, mas pode tornar-se intratável computacionalmente quando não existir solução explícita para a EMV. Se  $\phi$  é o mesmo para todas as observações, a EMV de  $\beta$  independe de  $\phi$ . Entretanto, a matriz de variâncias e covariâncias dos  $\hat{\beta}'$ s envolve esse parâmetro. Interpretando o logaritmo da função de verossimilhança  $\ell(\beta, \phi)$  como função de  $\beta$  e de  $\phi$ , dado  $\mathbf{y}$ , pode-se escrever de (3.1)

$$\ell(\beta, \phi) = \phi^{-1} \sum_{i=1}^n [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^n c(y_i, \phi). \quad (4.10)$$

A função escore relativa ao parâmetro  $\phi$  é dada por

$$U_\phi = \frac{\partial \ell(\beta, \phi)}{\partial \phi} = -\phi^{-2} \sum_{i=1}^n [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^n \frac{dc(y_i, \phi)}{d\phi}.$$

Observe-se que  $U_\phi$  é função de  $\beta$  através de  $\theta$  (ou  $\mu$ ) e de  $\phi$ , supondo  $\mathbf{y}$  conhecido. A EMV  $\hat{\phi}$  de  $\phi$  é obtida igualando-se  $\partial \ell(\hat{\beta}, \phi)/\partial \phi$  a zero. Claro que a EMV  $\hat{\phi}$  é função das médias ajustadas  $\hat{\mu}$  e dos dados  $\mathbf{y}$ . Da forma da função  $c(y, \phi)$  dada na Seção 1.3 (Tabela 1.1), verifica-se facilmente que  $\hat{\phi} = D_p/n$  para os modelos normal e normal inverso. Para o modelo gama, obtém-se a EMV  $\hat{\phi}$  de  $\phi$  como solução da equação não-linear

$$\log(\hat{\phi}^{-1}) - \psi(\hat{\phi}^{-1}) = \frac{D_p}{2n}, \quad (4.11)$$

em que o desvio  $D_p$  é dado na Tabela 4.1 e  $\psi(r) = d \log \Gamma(r)/dr$  é a função digama (função psi). Uma aproximação para  $\hat{\phi}$  obtida de (4.11) foi dada por Cordeiro e McCullagh (1991) para valores pequenos de  $\phi$

$$\hat{\phi} \approx \frac{2D_p}{n \left[ 1 + \left( 1 + \frac{2D_p}{3n} \right)^{1/2} \right]}.$$

Derivando  $U_\phi$  em relação a  $\beta_r$  vem

$$U_{\phi r} = \frac{\partial^2 \ell(\beta, \phi)}{\partial \phi \partial \beta_r} = -\phi^{-2} \sum_{i=1}^n \frac{(y_i - \mu_i)}{V_i} \frac{d\mu_i}{d\eta_i} x_{ir}.$$

Logo,  $E(U_{\phi r}) = 0$ , o que mostra que os parâmetros  $\phi$  e  $\beta$  são ortogonais. Isso implica que os EMV de  $\beta$  e  $\phi$  são, assintoticamente, independentes.

Como  $U_{\phi}$  é função de  $\phi$  e  $\mu$ , escreve-se  $U_{\phi} = U_{\phi}(\phi, \mu)$ . Pode-se mostrar que  $2U_{\phi}(\hat{\phi}, \mathbf{y}) = D_p$ , isto é, duas vezes a função escore relativa a  $\phi$  avaliada no ponto  $(\hat{\phi}, \mathbf{y})$  é igual ao desvio do modelo.

## 4.5 Comparação dos três métodos de estimação do parâmetro de dispersão no modelo gama

Nesta seção, comparam-se as três estimativas  $\hat{\phi}_d$ ,  $\hat{\phi}_P$  e  $\hat{\phi}$  de  $\phi$  no modelo gama. Cordeiro e McCullagh (1991) usaram a desigualdade  $\frac{1}{2x} < \log x - \psi(x) < \frac{1}{x}$  para mostrar que

$$\frac{D_p}{2n} < \hat{\phi} < \frac{D_p}{n}$$

e, portanto,

$$\frac{\hat{\phi}_d(n-p)}{2n} < \hat{\phi} < \frac{\hat{\phi}_d(n-p)}{n}.$$

Logo, para  $n$  grande, a EMV de  $\phi$  deve ficar entre  $\hat{\phi}_d/2$  e  $\hat{\phi}_d$ , ou seja, será menor do que  $\hat{\phi}_d$ .

Para comparar  $\hat{\phi}_d$  e  $\hat{\phi}_P$ , admite-se que a matriz modelo  $\mathbf{X}$  tenha uma coluna de uns relativa ao intercepto. Nesse caso, o desvio  $D_p$  reduz-se a  $D_p = 2 \sum_{i=1}^n \log(\hat{\mu}_i/y_i)$ , pois  $\sum_{i=1}^n (y_i - \hat{\mu}_i)/\hat{\mu}_i = 0$ . Considere a expansão em série de Taylor

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \dots$$

e a função  $f(y_i) = \log(\hat{\mu}_i/y_i)$  com  $x = y_i$  e  $a = \hat{\mu}_i$ . Então,  $f'(y_i) = -y_i^{-1}$ ,  $f''(y_i) = y_i^{-2}$  e  $f'''(y_i) = -2y_i^{-3}$  e

$$f(y_i) = \log\left(\frac{\hat{\mu}_i}{y_i}\right) \approx -\frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} + \frac{(y_i - \hat{\mu}_i)^2}{2\hat{\mu}_i^2} - \frac{(y_i - \hat{\mu}_i)^3}{3\hat{\mu}_i^3}.$$

Logo,

$$D_p = 2 \sum_{i=1}^n \log \left( \frac{\hat{\mu}_i}{y_i} \right) \approx -2 \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i} + \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i^2} - \frac{2}{3} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^3}{\hat{\mu}_i^3}. \quad (4.12)$$

O primeiro termo dessa expansão é nulo, pois o MLG tem por hipótese uma coluna de uns. Dividindo a equação (4.12) por  $n - p$  e usando (4.8) e (4.9), tem-se

$$\hat{\phi}_d \approx \hat{\phi}_P - \frac{2}{3(n-p)} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^3}{\hat{\mu}_i^3}.$$

Como a última soma pode ser positiva ou negativa, conclui-se que  $\hat{\phi}_d$  pode ser maior do que, menor do que ou igual a  $\hat{\phi}_P$ . Se o MLG tiver um bom ajuste, as médias ajustadas e as observações serão próximas e, assim,  $\hat{\phi}_d \doteq \hat{\phi}_P$ .

## 4.6 Testes de hipóteses

Os métodos de inferência nos MLG baseiam-se, fundamentalmente, na teoria de máxima verossimilhança. De acordo com essa teoria, existem três estatísticas usualmente utilizadas para testar hipóteses relativas aos parâmetros  $\beta's$ , que são deduzidas de distribuições assintóticas de funções adequadas dos estimadores de MV dos  $\beta's$ . São elas: i) razão de verossimilhanças, ii) Wald e iii) escore, que são assintoticamente equivalentes.

Sob a hipótese nula  $H_0$  e supondo que o parâmetro de dispersão  $\phi$  é conhecido, as três estatísticas convergem para uma variável aleatória com distribuição  $\chi_p^2$ , sendo, porém, a razão de verossimilhanças, o critério que define um teste uniformemente mais poderoso. Um estudo comparativo dessas estatísticas pode ser encontrado em Buse (1982) para o caso de hipóteses simples. Dentre outras, referências importantes são Silvey (1975), Cordeiro (1986), Dobson (2001), McCulloch e Searle (2000) e Paula (2004).

A razão de verossimilhanças para testar componentes do vetor  $\beta$  pode ser obtida como uma diferença de desvios entre modelos encaixados. A estatística de Wald (1943), também chamada de “máxima verossimilhança” por alguns autores, é

baseada na distribuição normal assintótica de  $\hat{\beta}$ . A estatística escore (Rao, 1973, Seção 6e) é obtida da função escore introduzida na Seção 3.2.

Dependendo da hipótese a ser testada, em particular, qualquer uma dessas três estatísticas pode ser a mais apropriada. Para hipóteses relativas a um único coeficiente  $\beta_r$ , a estatística de Wald é a mais usada. Para hipóteses relativas a vários coeficientes, a razão de verossimilhanças é, geralmente, preferida. A estatística escore tem sido usada na Bioestatística, com a finalidade de realizar testes como os do tipo de Mantel e Haenszel (1959).

#### 4.6.1 Teste de uma hipótese nula simples

Considere o teste da hipótese nula simples  $H_0 : \beta = \beta_0$  em um MLG supondo  $\phi$  conhecido, em que  $\beta_0$  é um vetor especificado para os parâmetros desconhecidos  $\beta$ , versus a hipótese alternativa  $H : \beta \neq \beta_0$ . Esse teste não é muito usado, pois, na prática, o interesse é especificar um subconjunto de componentes de  $\beta$ . As três estatísticas para testar  $H_0$  têm as seguintes formas

$$\text{razão de verossimilhanças: } w = 2\{\ell(\hat{\beta}) - \ell(\beta_0)\},$$

$$\text{estatística de Wald: } W = (\hat{\beta} - \beta_0)^T \hat{\mathbf{K}} (\hat{\beta} - \beta_0), \text{ e}$$

$$\text{estatística escore: } S_R = \mathbf{U}(\beta_0)^T \mathbf{K}_0^{-1} \mathbf{U}(\beta_0),$$

em que  $\ell(\hat{\beta})$  e  $\ell(\beta_0)$  são os valores do logaritmo da função de verossimilhança (3.1) em  $\hat{\beta}$  e  $\beta_0$ , respectivamente,  $\mathbf{U}(\beta_0)$  e  $\mathbf{K}_0$  são o vetor escore e a matriz de informação avaliadas em  $\beta_0$ , e  $\hat{\mathbf{K}}$  a matriz de informação avaliada na EMV  $\hat{\beta}$ . Na estatística de Wald,  $\hat{\mathbf{K}}$  pode ser substituída por  $\mathbf{K}_0$  para definir uma estatística de Wald modificada assintoticamente equivalente. Uma vantagem da estatística escore é que não é necessário calcular a EMV de  $\beta$  segundo  $H$ , embora na prática essa estatística seja importante.

As três estatísticas descritas são assintoticamente equivalentes e, segundo  $H_0$ , convergem em distribuição para a variável  $\chi_p^2$ . Entretanto, a razão de verossimilhanças, nesse caso, é geralmente preferida, pois, se existe um teste uniformemente mais poderoso, esse critério o define. Se o modelo tem um único parâmetro, usando-

se as estatísticas  $\sqrt{S_R}$  e  $\sqrt{W}$ , com um sinal adequado, no lugar de  $S_R$  e  $W$ , obtêm-se testes de mais fácil interpretação.

A estatística escore é definida pela forma quadrática  $S_R = \mathbf{U}^T \mathbf{K}^{-1} \mathbf{U}$  e pode ser deduzida da maneira que se segue. O vetor escore  $\mathbf{U}$  tem as seguintes propriedades descritas na Seção 4.1:  $E(\mathbf{U}) = \mathbf{0}$  e  $\text{Cov}(\mathbf{U}) = E(\mathbf{U}\mathbf{U}^T) = \mathbf{K}$ . Supondo observações independentes, o vetor escore é definido por uma soma de variáveis aleatórias independentes que, pelo teorema central do limite, tem distribuição assintótica normal  $p$ -dimensional  $N_p(\mathbf{0}, \mathbf{K})$ . Logo, para amostras grandes, a estatística escore  $S_R = \mathbf{U}^T \mathbf{K}^{-1} \mathbf{U}$  converge, assintoticamente, para uma distribuição  $\chi_p^2$  supondo que o modelo com os parâmetros especificados seja verdadeiro.

**Exemplo 4.6:** Seja  $Y_1, \dots, Y_n$  uma amostra aleatória de uma distribuição normal  $N(\mu, \sigma^2)$  com  $\mu$  desconhecido e  $\sigma^2$  conhecido. Visto como MLG, tem-se:

- i) somente um parâmetro de interesse,  $\mu$ ;
- ii) não há variáveis explanatórias e,
- iii) a função de ligação é a identidade, isto é,  $\eta = \mu$ .

O logaritmo da função de verossimilhança é

$$\ell = \ell(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 - \frac{n}{2} \log(2\pi\sigma^2),$$

a partir do qual se obtêm:

$$U = \frac{d\ell}{d\mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu) = \frac{n}{\sigma^2} (\bar{y} - \mu),$$

$$E(U) = \frac{n}{\sigma^2} [E(\bar{Y}) - \mu] = 0$$

e

$$K = \text{Var}(U) = \frac{n^2}{\sigma^4} \text{Var}(\bar{Y}) = \frac{n}{\sigma^2}.$$

Portanto,

$$S_R = U^T K^{-1} U = \frac{n^2(\bar{Y} - \mu)^2 \sigma^2}{\sigma^4 n} = \frac{(\bar{Y} - \mu)^2}{\frac{\sigma^2}{n}} \sim \chi_1^2,$$

resultado que pode ser usado para a obtenção de intervalos de confiança para  $\mu$ .

**Exemplo 4.7:** Suponha que  $Y$  tem distribuição binomial  $B(m, \pi)$ . Então, conforme foi visto,

$$\ell(\pi) = \log \binom{m}{y} + y \log \pi + (m - y) \log(1 - \pi)$$

e, portanto,

$$U = \frac{d\ell(\pi)}{d\pi} = \frac{y}{\pi} - \frac{(m - y)}{1 - \pi} = \frac{y - m\pi}{\pi(1 - \pi)}.$$

Mas,  $E(Y) = \mu = m\pi$  e  $\text{Var}(Y) = m\pi(1 - \pi) = \frac{\mu}{m}(m - \mu)$ . Logo,

$$E(U) = 0 \quad \text{e} \quad K = \text{Var}(U) = \frac{\text{Var}(Y)}{\pi^2(1 - \pi)^2} = \frac{m}{\pi(1 - \pi)}.$$

Assim,

$$S_R = U^T K^{-1} U = \frac{(Y - m\pi)^2 \pi(1 - \pi)}{\pi^2(1 - \pi)^2 m} = \frac{(Y - m\pi)^2}{m\pi(1 - \pi)} = \frac{[Y - E(Y)]^2}{\text{Var}(Y)}$$

que, pelo teorema central do limite, tem distribuição  $\chi_1^2$ , ou, equivalentemente,

$$\frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}} = \frac{\sqrt{m}(Y - \mu)}{\sqrt{\mu(m - \mu)}} \rightarrow N(0, 1),$$

resultado que pode ser usado para se fazer inferência sobre  $\mu$ .

#### 4.6.2 Teste de uma hipótese nula composta

Quando se tem um vetor de parâmetros  $\beta$  em um MLG, muitas vezes há interesse no teste de hipóteses de apenas um subconjunto deles. Supõe-se que o parâmetro de dispersão  $\phi$  é conhecido. Seja, então, uma partição do vetor de parâmetros  $\beta$  dada por  $\beta = (\beta_1^T \quad \beta_2^T)^T$ , em que  $\beta_1$ , de dimensão  $q$ , é o vetor

de interesse e  $\beta_2$ , de dimensão  $(p - q)$ , o vetor de parâmetros de perturbação. De forma semelhante, tem-se a partição da matriz modelo  $\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2)$ , do vetor escore  $\mathbf{U} = \phi^{-1}\mathbf{X}^T\mathbf{W}\mathbf{H}(\mathbf{y} - \boldsymbol{\mu}) = (\mathbf{U}_1^T \ \mathbf{U}_2^T)^T$  com  $\mathbf{U}_1 = \phi^{-1}\mathbf{X}_1^T\mathbf{W}\mathbf{H}(\mathbf{y} - \boldsymbol{\mu})$  e  $\mathbf{U}_2 = \phi^{-1}\mathbf{X}_2^T\mathbf{W}\mathbf{H}(\mathbf{y} - \boldsymbol{\mu})$  e da matriz de informação de Fisher para  $\beta$

$$\mathbf{K} = \phi^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X} = \begin{bmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix},$$

sendo que  $\mathbf{K}_{12} = \mathbf{K}_{21}^T$ .

Usando-se resultados conhecidos de álgebra linear, envolvendo partição de matrizes (Searle, 1982), tem-se, para amostras grandes, a variância assintótica de  $\hat{\beta}_1$ :

$$\text{Var}(\hat{\beta}_1) = (\mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{21})^{-1} = \phi[\mathbf{X}_1^T\mathbf{W}^{1/2}(\mathbf{I} - \mathbf{P}_2)\mathbf{W}^{1/2}\mathbf{X}_1]^{-1},$$

sendo  $\mathbf{P}_2 = \mathbf{W}^{1/2}\mathbf{X}_2(\mathbf{X}_2^T\mathbf{W}\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{W}^{1/2}$  a matriz projeção segundo o modelo com matriz  $\mathbf{X}_2$ .

Sejam as hipóteses

$$H_0 : \beta_1 = \beta_{1,0} \text{ versus } H : \beta_1 \neq \beta_{1,0},$$

sendo  $\beta_{1,0}$  um vetor especificado para  $\beta_1$ . Seja  $\hat{\beta} = (\hat{\beta}_1^T \ \hat{\beta}_2^T)^T$  o EMV de  $\beta$  sem restrição e  $\tilde{\beta} = (\beta_{1,0}^T \ \tilde{\beta}_2^T)^T$  o EMV restrito de  $\beta$  em que  $\tilde{\beta}_2$  é o EMV de  $\beta_2$  sob  $H_0$ . A seguir, são definidos os três testes mais usados para testar a hipótese  $H_0$ .

#### (a) Teste da razão de verossimilhanças

Envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada sem restrição ( $\ell(\hat{\beta}_1, \hat{\beta}_2)$ ) e sob  $H_0$  ( $\ell(\beta_{1,0}, \tilde{\beta}_2)$ ), ou, em termos do desvio, a comparação de  $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$  e  $D(\mathbf{y}; \tilde{\boldsymbol{\mu}})$  em que  $\tilde{\boldsymbol{\mu}} = g^{-1}(\tilde{\boldsymbol{\eta}})$  e  $\tilde{\boldsymbol{\eta}} = \mathbf{X}\tilde{\beta}$ . Esse teste é, geralmente, preferido no caso de hipóteses relativas a vários coeficientes  $\beta's$ . Se as diferenças são grandes, então,  $H_0$  é rejeitada. A estatística para esse teste é dada como uma diferença de desvios por:

$$w = 2[\ell(\hat{\beta}_1, \hat{\beta}_2) - \ell(\beta_{1,0}, \tilde{\beta}_2)] = \phi^{-1}[D(\mathbf{y}; \tilde{\boldsymbol{\mu}}) - D(\mathbf{y}; \hat{\boldsymbol{\mu}})]. \quad (4.13)$$



Para amostras grandes, rejeita-se  $H_0$ , a um nível de  $100\alpha\%$  de significância, se  $w > \chi_{q,1-\alpha}^2$ .

(b) **Teste de Wald**

É baseado na distribuição normal assintótica de  $\hat{\beta}$  e é uma generalização da estatística t de Student (Wald, 1943). É, geralmente, o mais usado no caso de hipóteses relativas a um único coeficiente  $\beta_r$ . Tem como vantagem, em relação ao teste da razão de verossimilhanças, o fato de não haver necessidade de se calcular a EMV restrita  $\tilde{\beta}_2$ . Como visto antes, assintoticamente,  $\hat{\beta} \sim N_p(\beta, \mathbf{K}^{-1})$ . Assim, a estatística para esse teste é

$$W = (\hat{\beta}_1 - \beta_{1,0})^T \widehat{\text{Cov}}(\hat{\beta}_1)^{-1} (\hat{\beta}_1 - \beta_{1,0}), \quad (4.14)$$

sendo  $\widehat{\text{Cov}}(\hat{\beta}_1)$  a matriz  $\text{Cov}(\hat{\beta}_1)$  avaliada em  $\hat{\beta} = (\hat{\beta}_1^T \ \hat{\beta}_2^T)^T$ . Para amostras grandes, rejeita-se  $H_0$ , a um nível de  $100\alpha\%$  de significância, se  $W > \chi_{q,1-\alpha}^2$ .

(c) **Teste escore**

Obtido a partir da função escore, a estatística de teste é dada por

$$S_R = \mathbf{U}_1^T(\tilde{\beta}) \widetilde{\text{Cov}}(\hat{\beta}_1) \mathbf{U}_1(\tilde{\beta}), \quad (4.15)$$

sendo  $\widetilde{\text{Cov}}(\hat{\beta}_1)$  a matriz  $\text{Cov}(\hat{\beta}_1)$  avaliada em  $\tilde{\beta} = (\beta_{1,0}^T \ \tilde{\beta}_2^T)^T$ . Para amostras grandes, rejeita-se  $H_0$ , a um nível de  $100\alpha\%$  de significância, se  $S_R > \chi_{q,1-\alpha}^2$ .

As três estatísticas (4.13), (4.14) e (4.15) diferem por termos de ordem  $O_p(n^{-1})$ . As expansões assintóticas das distribuições dessas três estatísticas são descritas no livro de Cordeiro (1999).

Para o cálculo das estatísticas Wald e escore, deve-se obter  $\text{Var}(\hat{\beta}_1)$  da inversa da matriz de informação subdividida como  $\mathbf{K}$ , ou seja,

$$\text{Cov}(\hat{\beta}) = \mathbf{K}^{-1} = \phi(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} = \begin{pmatrix} \mathbf{K}^{11} & \mathbf{K}^{12} \\ \mathbf{K}^{21} & \mathbf{K}^{22} \end{pmatrix},$$

sendo que  $\mathbf{K}^{12} = \mathbf{K}^{21^T}$ ,  $\text{Var}(\hat{\beta}_1) = \mathbf{K}^{11}$ ,  $\text{Var}(\hat{\beta}_2) = \mathbf{K}^{22}$  e  $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \mathbf{K}^{12}$ .

## 4.7 Regiões de confiança

Regiões de confiança assintóticas para  $\beta_1$  podem ser construídas usando-se qualquer uma das três estatísticas de teste. A partir da estatística da razão de verossimilhanças, uma região de confiança para  $\beta_1$ , com um coeficiente de confiança de  $100(1 - \alpha)\%$ , inclui todos os valores de  $\beta_1$  tais que:

$$2[\ell(\hat{\beta}_1, \hat{\beta}_2) - \ell(\beta_1, \tilde{\beta}_2)] < \chi_{q,1-\alpha}^2,$$

em que  $\tilde{\beta}_2$  é a EMV de  $\beta_2$  para cada valor de  $\beta_1$  que é testado ser pertencente, ou não, ao intervalo, e  $\chi_{q,1-\alpha}^2$  é o percentil da distribuição  $\chi^2$  com  $q$  graus de liberdade, correspondente a um nível de significância igual a  $100\alpha\%$ .

Usando-se a estatística de Wald, uma região de confiança para  $\beta_1$ , com um coeficiente de confiança de  $100(1 - \alpha)\%$ , inclui todos os valores de  $\beta_1$  tais que:

$$(\hat{\beta}_1 - \beta_1)^T \widehat{\text{Var}}(\hat{\beta}_1)^{-1} (\hat{\beta}_1 - \beta_1) < \chi_{q,1-\alpha}^2.$$

Alternativamente, regiões de confiança para os parâmetros lineares  $\beta_1, \dots, \beta_p$  de um MLG podem ser construídos através da função desvio. Deseja-se uma região de confiança aproximada para um conjunto particular de parâmetros  $\beta_1, \dots, \beta_q$  de interesse. Sejam  $S_p$  o desvio do modelo  $M_p$  com todos os  $p$  parâmetros e  $S_{p-q}$  o desvio do modelo  $M_{p-q}$  com  $p - q$  parâmetros linearmente independentes, e os  $q$  parâmetros de interesse tendo valores fixados:  $\beta_r = \beta_r^*$ ,  $r = 1, \dots, q$ . No ajuste do modelo  $M_{p-q}$ , a quantidade  $\sum_{r=1}^q \beta_r^* \mathbf{x}^{(r)}$  funciona como “offset” (isto é, uma parte conhecida na estrutura linear do modelo), sendo  $\mathbf{x}^{(r)}$  a  $r$ -ésima coluna da matriz modelo  $\mathbf{X}$  correspondente a  $\beta_r$ .

Uma região aproximada de  $100(1 - \alpha)\%$  de confiança para  $\beta_1, \dots, \beta_q$  é dada pelo conjunto de pontos  $\beta_r^*$ ,  $r = 1, \dots, q$ , não rejeitados pela estatística  $S_{p-q} - S_p$ , isto é, por

$$\{\beta_r^*, r = 1, \dots, q; S_{p-q} - S_p < \chi_{q,1-\alpha}^2\}. \quad (4.16)$$

Embora, na prática, o cálculo dessas regiões de confiança apresente um trabalho considerável, os software R, S-Plus, SAS e MATLAB têm as facilidades necessárias incluindo o uso de gráficos.

No caso do intervalo de confiança para um único parâmetro  $\beta_r$ , tem-se

$$\{\beta_r^*; S_{p-1} - S_p < \chi_{1,1-\alpha}^2\}, \quad (4.17)$$

em que  $S_{p-1}$  é o desvio do modelo com os parâmetros  $\beta_1, \dots, \beta_{r-1}, \beta_{r+1}, \dots, \beta_p$  e “offset”  $\beta_r^* \mathbf{x}^{(r)}$ . Um outro intervalo aproximado para  $\beta_r$ , simétrico e assintoticamente equivalente a (4.17), pode ser obtido de

$$[\hat{\beta}_r - a_{\alpha/2}(-\hat{\kappa}^{rr})^{1/2}, \hat{\beta}_r + a_{\alpha/2}(-\hat{\kappa}^{rr})^{1/2}], \quad (4.18)$$

em que  $-\hat{\kappa}^{rr}$  é o elemento  $(r, r)$  de  $\hat{\mathbf{K}}^{-1}$  e  $\Phi(-a_{\alpha/2}) = \alpha/2$ , sendo  $\Phi(\cdot)$  a f.d.a. da normal  $N(0, 1)$ .

A construção de (4.18) é muito mais simples do que (4.17), pois é necessário apenas o ajuste do modelo  $M_p$ . A grande vantagem do uso de (4.17), ao invés de (4.18), é de ser independente da parametrização adotada. Por exemplo, com uma parametrização diferente para o parâmetro de interesse  $\gamma_r = h(\beta_r)$ , o intervalo baseado na distribuição normal assintótica de  $\hat{\gamma}_r$  não corresponde exatamente a (4.18). Entretanto, usando (4.17), o intervalo para  $\gamma_r$  pode ser obtido por simples transformação  $\{h(\beta_r^*); S_{p-1} - S_p < \chi_{1,1-\alpha}^2\}$ .

## 4.8 Seleção de covariáveis

Na prática, é difícil selecionar um conjunto de covariáveis para formar um modelo parcimonioso, devido aos problemas de ordem combinatória e estatística. O problema de cunho combinatório é selecionar todas as combinações possíveis de covariáveis que deverão ser testadas para inclusão no modelo. O problema estatístico é definir, com a inclusão de um novo termo no preditor linear, o balanço entre o efeito de reduzir a discrepância entre  $\hat{\mu}$  e  $y$  e o fato de se ter um modelo mais complexo.

Outras estatísticas que servem como medidas de comparação de qualidade de ajuste do modelo e o seu grau de complexidade são os critérios de informação de Akaike (1974)  $AIC_p = -2\hat{\ell}_p + 2p$  e de Bayes  $BIC_p = -2\hat{\ell}_p + p \log n$  (Schwarz, 1978) que para os MLG podem ser expressos como

$$AIC_p = S_p + 2p - 2\hat{\ell}_n. \quad (4.19)$$

e

$$BIC_p = S_p + p \log n - 2\hat{\ell}_n. \quad (4.20)$$

Se o modelo envolver um parâmetro de dispersão  $\phi$ , esse deve ser estimado, como descrito na Seção 4.4, para calcular um valor numérico em (4.19) e (4.20).

O critério de Akaike foi desenvolvido para estender o método de máxima verossimilhança para a situação de ajustes de muitos modelos com diferentes números de parâmetros e para decidir quando parar o ajuste. A estatística (4.19) pode ajudar na seleção de modelos complexos e tem demonstrado produzir soluções razoáveis para muitos problemas de seleção de modelos que não podem ser tratados pela teoria convencional de máxima verossimilhança. Um valor baixo para  $AIC_p$  é considerado como representativo de um melhor ajuste e os modelos são selecionados visando a obter um mínimo  $AIC_p$ . De forma semelhante interpreta-se  $BIC_p$ .

Uma outra medida de comparação equivalente ao critério de Akaike é

$$C_p^* = S_p + 2p - n = AIC_p + 2\hat{\ell}_n - n. \quad (4.21)$$

Para um MLG isolado é, usualmente, mais simples trabalhar com  $C_p^*$  do que  $AIC_p$ . Para o modelo normal-linear com variância constante  $\sigma^2$ ,  $C_p^*$  reduz-se à estatística  $C_p = SQR_p/\tilde{\sigma}^2 + 2p - n$  (Mallows, 1966), em que  $SQR_p = \sum_{\ell=1}^n (y_\ell - \hat{\mu}_\ell)^2$  e  $\tilde{\sigma}^2 = SQR_m/(n - m)$  é, a menos de um coeficiente multiplicador, o resíduo quadrático médio baseado no modelo maximal com  $m$  parâmetros. Nesse caso,  $AIC_p = SQR_p/\tilde{\sigma}^2 + 2p + n \log(2\pi\tilde{\sigma}^2)$ . Note-se que  $C_m = m$ .

Em geral,  $E(C_p^*) \neq p$ . Para o modelo normal-linear com variância conhecida tem-se  $E(C_p^*) = p$ , supondo que o modelo é verdadeiro. Se a variância for desconhecida, o valor esperado de  $C_p^*(= C_p)$  será muito maior do que  $p$ , quando o modelo não se ajustar bem aos dados. Um gráfico de  $C_p^*$  (ou  $AIC_p$ ) versus  $p$  fornece uma boa indicação para comparação de modelos alternativos. Considerando dois modelos encaixados  $M_q \subset M_p$ ,  $p > q$ , vem  $AIC_p - AIC_q = C_p^* - C_q^* = S_p - S_q + 2(p - q)$  e, portanto, supondo  $M_q$  verdadeiro,  $E(AIC_p - AIC_q) = p - q + O(n^{-1})$ .

Na comparação de modelos, sucessivamente, mais ricos, a declividade esperada do segmento de reta unindo  $AIC_p$  com  $AIC_q$  (ou  $C_p^*$  com  $C_q^*$ ) deve ser próxima de um, supondo o modelo mais pobre  $M_q$  verdadeiro. Pares de modelos com declividade observada maior do que um, indicam que o modelo maior ( $M_p$ ) não é, significativamente, melhor do que o menor ( $M_q$ ).

Uma outra tentativa para seleção de covariáveis é minimizar a expressão (Atkinson, 1981)

$$A_p = D_p + \frac{p\alpha}{\phi}, \quad (4.22)$$

em que  $D_p$  é o desvio do modelo  $M_p$  sem o parâmetro de dispersão  $\phi$  e  $\alpha$  é uma constante ou função de  $n$ . Para o cálculo de (4.22),  $\phi$  é estimado como descrito na Seção 4.4. Tem-se  $A_p = [C_p^* + p(\alpha - 2) + n]/p$  e para  $\alpha = 2$ ,  $A_p$  é equivalente a  $C_p^*$  (ou  $AIC_p$ ).

## 4.9 Método das variáveis explanatórias adicionais

O método das variáveis explanatórias adicionais (Pregibon, 1979, Capítulo 3), consiste em aumentar a estrutura linear do modelo através de variáveis explanatórias bastante adequadas para representar anomalias específicas no MLG usual. A forma mais comum do método tem origem no trabalho de Box e Tidwell (1962), que consideraram uma regressão com parâmetros não-lineares nas variáveis explanatórias. No preditor linear aparecendo uma função  $h(x; \gamma)$ , em que  $\gamma$  é não-linear em  $x$ , expande-se a função em série de Taylor ao redor de um valor próximo

conhecido  $\gamma^{(o)}$  tornando  $\gamma$  um parâmetro linear na variável explanatória adicional  $\frac{\partial h(x; \gamma)}{\partial \gamma} \Big|_{\gamma=\gamma^{(o)}}$ .

No método, a estrutura linear do modelo aumentado é do tipo

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}, \quad (4.23)$$

em que  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_q)$ , sendo  $\mathbf{z}_r$  um vetor coluna de dimensão  $n$  conhecido e  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_q)^T$ . Em casos especiais, as colunas  $\mathbf{z}_r$  podem ser funções do ajuste do modelo usual, isto é,  $\mathbf{Z} = \mathbf{Z}(\mathbf{X}\hat{\boldsymbol{\beta}})$ , ou funções específicas das variáveis explanatórias originais  $\mathbf{z}_r = \mathbf{z}_r(\mathbf{x}^{(r)})$ .

A importância das variáveis explanatórias adicionais é dada pela diferença dos desvios dos modelos  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$  e (4.23). Se a adição das variáveis explanatórias  $\mathbf{Z}\boldsymbol{\gamma}$  altera substancialmente o ajuste, as anomalias em questão afetam, seriamente, o modelo original. Em geral, quando isso ocorre, as formas das variáveis explanatórias adicionais produzem uma ação corretiva.

Um bom exemplo do uso de uma variável explanatória adicional está no teste de Tukey (1949) de um grau de liberdade para verificar a não-aditividade de um modelo. Em termos de MLG, considera-se  $(\mathbf{X}\hat{\boldsymbol{\beta}}) \otimes (\mathbf{X}\hat{\boldsymbol{\beta}})$ , em que  $\otimes$  é o produto direto, como uma variável explanatória adicional e, se no ajuste do modelo aumentado, o coeficiente dessa variável explanatória for significantemente diferente de zero, aceita-se a não-aditividade no modelo original. Uma transformação do tipo potência da variável resposta, pode ser uma medida corretiva para eliminar a não-aditividade.

Para verificar se a escala de uma variável explanatória isolada  $\mathbf{x}^{(r)}$  está correta, o teste de Tukey trata  $\hat{\beta}_r^2(\mathbf{x}^{(r)} \otimes \mathbf{x}^{(r)})$ , em que  $\hat{\beta}_r$  é o coeficiente estimado de  $\mathbf{x}^{(r)}$ , como uma variável explanatória adicional. Quando o coeficiente associado a essa variável explanatória, no ajuste do modelo aumentado, for estatisticamente zero, aceita-se a linearidade de  $\boldsymbol{\eta}$  em  $\mathbf{x}^{(r)}$ .

Pregibon (1979) recomenda um método gráfico, alternativo, baseado na estatística  $\mathbf{v}_r = \hat{\beta}_r \mathbf{x}^{(r)} + \hat{\mathbf{z}} - \hat{\boldsymbol{\eta}} = \hat{\beta}_r \mathbf{x}^{(r)} + \hat{\mathbf{H}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ , que representa uma medida da linearidade da variável explanatória  $\mathbf{x}^{(r)}$ . A estatística  $\mathbf{v}_r$  é, simplesmente, um

resíduo parcial generalizado para a variável explanatória  $\mathbf{x}^{(r)}$ , expresso na escala da variável dependente modificada  $\mathbf{z}$ . A escala de  $\mathbf{x}^{(r)}$  é considerada correta, se o gráfico de  $\mathbf{v}_r$  versus  $\mathbf{x}^{(r)}$  é, aproximadamente, linear. Caso contrário, a forma do gráfico deve sugerir a ação corretiva.

Inferência sobre  $\boldsymbol{\gamma}$  pode ser realizada a partir da redução do desvio do modelo com a inclusão de  $\mathbf{Z}\boldsymbol{\gamma}$ , ou através da distribuição normal assintótica de  $\hat{\boldsymbol{\gamma}}$ , de média igual ao parâmetro verdadeiro  $\boldsymbol{\gamma}$  e matriz de covariância dada por

$$(\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} + \mathbf{L}(\mathbf{X}^T \mathbf{W} \mathbf{X} - \mathbf{X}^T \mathbf{W} \mathbf{Z} \mathbf{L})^{-1} \mathbf{L}^T,$$

em que  $\mathbf{L} = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{X}$ . O método das variáveis explanatórias adicionais é bastante usado para estimar a função de ligação e para identificar observações não explicadas pelo modelo.

## 4.10 Seleção da função de ligação

Na Seção 8.2, serão apresentadas várias funções de ligação com objetivos diferentes. Muitas vezes, para um conjunto particular de observações, pode ser difícil decidir qual a melhor função de ligação e, ainda, essa pode não pertencer à uma família especificada.

Uma estratégia freqüente para verificar se uma função de ligação é adequada, seria computar a redução no desvio após a inclusão da variável explanatória  $\hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$ . Se isso causar uma redução significativa no desvio, a função de ligação não é satisfatória. Um método alternativo é traçar o gráfico da variável dependente modificada estimada  $\hat{\mathbf{z}} = \hat{\boldsymbol{\eta}} + \hat{\mathbf{H}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$  versus  $\hat{\boldsymbol{\eta}}$ . Se o gráfico for, aproximadamente, linear, a função de ligação estará correta.

Apresenta-se, agora, um método de estimação da função de ligação, desenvolvido por Pregibon (1980), usando variáveis explanatórias adicionais, obtidas de uma linearização da função de ligação. Seja  $g(\boldsymbol{\mu}; \boldsymbol{\lambda}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$  a função de ligação dependendo de um conjunto de parâmetros  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_r)^T$ , supostos desconhe-

cidos. Uma família de funções de ligação com um único parâmetro é a potência  $g(\mu; \lambda) = (\mu^\lambda - 1)/\lambda$  ou  $\mu^\lambda$ .

Um teste aproximado da hipótese nula composta  $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}^{(0)}$ , em que  $\boldsymbol{\lambda}^{(0)}$  é um valor especificado para  $\boldsymbol{\lambda}$ , versus  $H : \boldsymbol{\lambda} \neq \boldsymbol{\lambda}^{(0)}$ , é obtido expandindo  $g(\mu; \lambda) = \boldsymbol{\eta}$  em série de Taylor ao redor de  $\boldsymbol{\lambda}^{(0)}$  até primeira ordem. Tem-se,

$$g(\mu; \lambda) = g(\mu; \lambda^{(0)}) + \mathbf{D}(\mu; \lambda^{(0)})(\lambda - \lambda^{(0)}), \quad (4.24)$$

em que  $\mathbf{D}(\mu; \lambda) = \frac{\partial g(\mu; \lambda)}{\partial \lambda}$  é uma matriz de dimensões  $n \times r$  que depende de  $\beta$  e  $\lambda$ . Seja  $\hat{\beta}_0$  a EMV de  $\beta$  obtida do ajuste do modelo  $g(\mu; \lambda^{(0)}) = \mathbf{X}\beta$  e  $\hat{\mu}_0 = g^{-1}(\mathbf{X}\hat{\beta}_0; \lambda^{(0)})$ . Estima-se  $\mathbf{D}(\mu; \lambda^{(0)})$  por  $\hat{\mathbf{D}}^{(0)} = \mathbf{D}(\hat{\mu}_0; \lambda^{(0)})$ .

Se a expansão (4.24) for adequada, pode-se considerar a estrutura linear

$$g(\mu; \lambda^{(0)}) = \begin{pmatrix} \mathbf{X} & -\hat{\mathbf{D}}^{(0)} \end{pmatrix} \begin{pmatrix} \beta \\ \lambda \end{pmatrix} + \hat{\mathbf{D}}^{(0)} \lambda^{(0)} \quad (4.25)$$

com uma aproximação de  $g(\mu; \lambda) = \mathbf{X}\beta$  com  $\lambda$  desconhecido.

Na estrutura (4.25), o vetor de parâmetros  $\lambda$  aparece como linear nas variáveis explanatórias adicionais  $-\hat{\mathbf{D}}^{(0)}$  e o preditor linear envolve  $\hat{\mathbf{D}}^{(0)}\lambda^{(0)}$  como *offset*. Essas variáveis explanatórias adicionais representam uma medida da distância da função de ligação  $\lambda^{(0)}$  à função de ligação verdadeira. A inferência sobre  $\lambda$  pode ser realizada de maneira análoga a  $\beta$ , como descrito na Seção 4.6.2.

Logo, testar  $H_0 : \lambda = \lambda^{(0)}$  contra  $H : \lambda \neq \lambda^{(0)}$  corresponde, aproximadamente, a comparar os modelos  $\mathbf{X}$  e  $(\mathbf{X} \quad -\hat{\mathbf{D}}^{(0)})$ , ambos tendo a mesma função de ligação  $g(\mu; \lambda^{(0)}) = \boldsymbol{\eta}$ . Se a diferença de desvios entre esses modelos é for maior do que  $\chi_r^2(\alpha)$ , rejeita-se  $H_0$ .

A aproximação do teste depende fortemente da linearização (4.24). Quando o  $\lambda$  verdadeiro estiver distante de  $\lambda^{(0)}$ , não existirá garantia de convergência no ajuste de (4.25) e, mesmo convergindo, a estimativa de  $\lambda$  obtida pode diferir substancialmente do valor correto de sua EMV. Para calcular uma melhor aproximação dessa estimativa, o processo (4.25) deverá ser repetido com as variáveis explanatórias adicionais sendo reestimadas a cada etapa, a partir das estimativas correspondentes de  $\beta$  e  $\lambda$ .



Um processo alternativo para obter uma boa estimativa de  $\boldsymbol{\lambda}$ , é considerar  $\boldsymbol{\lambda}$  fixado e pertencendo a um conjunto amplo de valores arbitrários e, então, computar o desvio  $S_p(\boldsymbol{\lambda})$  como função de  $\boldsymbol{\lambda}$ . Traça-se o gráfico da superfície  $S_p(\boldsymbol{\lambda})$  versus  $\boldsymbol{\lambda}$ , escolhendo a estimativa  $\tilde{\boldsymbol{\lambda}}$  correspondente ao valor mínimo de  $S_p(\boldsymbol{\lambda})$  nesse conjunto. Se  $\boldsymbol{\lambda}$  é unidimensional, o processo é bastante simples, caso contrário, pode ser impraticável. Uma região de  $100(1 - \alpha)\%$  de confiança para  $\boldsymbol{\lambda}$  é determinada no gráfico por  $\{\boldsymbol{\lambda}; S_p(\boldsymbol{\lambda}) - S_p(\tilde{\boldsymbol{\lambda}}) \leq \chi_r^2(\alpha)\}$ , sendo independente da parametrização adotada. Um teste de  $H_o : \boldsymbol{\lambda} = \boldsymbol{\lambda}^{(0)}$  pode ser baseado nessa região. Pode-se calcular, numericamente, a EMV de  $\boldsymbol{\lambda}$ , embora com uma maior complexidade computacional.

## 4.11 Exercícios

1. Para os modelos normal, gama, normal inverso e Poisson com componentes sistemáticos  $\eta_i = \mu_i^\lambda = \beta_0 + \beta_1 x_i$ , e para o modelo binomial com  $\eta_i = \log\{[(1 - \mu_i)^{-\lambda} - 1]\lambda^{-1}\} = \beta_0 + \beta_1 x_i$ , sendo  $\lambda$  conhecido, calcular: a) as estruturas de covariância assintótica de  $\hat{\boldsymbol{\beta}}$  e  $\hat{\boldsymbol{\mu}}$ ; b) as estatísticas escore, de Wald e da razão de verossimilhanças nos testes:  $H_1 : \beta_1 = 0$  versus  $H'_1 : \beta_1 \neq 0$  e  $H_2 : \beta_0 = 0$  versus  $H'_2 : \beta_0 \neq 0$ ; c) intervalos de confiança para os parâmetros  $\beta_0$  e  $\beta_1$ .
2. Sejam  $Y_1, \dots, Y_n$  variáveis binárias independentes e identicamente distribuídas com  $P(Y_i = 1) = 1 - P(Y_i = 0) = \mu$ ,  $0 < \mu < 1$ . A distribuição de  $Y_i$  pertence à família (1.5) com parâmetro natural  $\theta$ . Demonstrar que a estatística de Wald para testar  $H_0 : \theta = 0$  versus  $H : \theta \neq 0$  é  $W = [n\hat{\theta}^2 \exp(\hat{\theta})]/[1 + \exp(\hat{\theta})]^2$ , sendo os valores possíveis de  $\hat{\theta}$  iguais a  $\log[t/(n-t)]$ ,  $t = 1, \dots, n-1$ . Quais as formas das estatísticas escore e da razão de verossimilhanças?
3. Obtenha as expressões das estatísticas desvio  $D_p$  e  $X_p^2$  de Pearson para as distribuições estudadas no Capítulo 1.
4. a) Mostre que para os modelos log-lineares com matriz modelo tendo uma coluna

de 1's, o desvio reduz-se a  $S_p = 2 \sum_{i=1}^n y_i \log(y_i/\hat{\mu}_i)$ ; b) Mostre que para o modelo gama com índice  $\nu$  e função de ligação potência  $\eta = \mu^\lambda$  ou  $\eta = \log \mu$ , nesse último caso a matriz  $\mathbf{X}$  tendo uma coluna de 1's, o desvio reduz-se a  $S_p = 2\nu \sum_{i=1}^n \log(\hat{\mu}_i/y_i)$ .

5. Os dois exercícios em 4. são casos particulares do resultado mais geral  $\sum_{i=1}^n (y_i - \hat{\mu}_i) \hat{\mu}_i V^{-1}(\hat{\mu}_i) = 0$  quando o modelo tem função de ligação  $\eta = \mu^\lambda (\lambda \neq 0)$  ou  $\eta = \log \mu$ , nesse último caso,  $\mathbf{X}$  com uma coluna de 1's. Mostre esse resultado.

6. a) Mostre que para o modelo gama simples com índice  $\nu$ , em que todas as médias são iguais, o desvio reduz-se à estatística clássica  $S_1 = 2n\nu \log(\bar{y}/\tilde{y})$ , em que  $\bar{y}$  e  $\tilde{y}$  são as médias aritmética e geométrica dos dados, respectivamente. b) Mostre que, para um MLG, sendo  $\ell$  o logaritmo da função de verossimilhança total,  $E(\partial^2 \ell / \partial \phi \partial \beta_j) = 0$  e, portanto, os parâmetros  $\phi$  e  $\beta$  são ortogonais.

7. Mostre que a estimativa de máxima verossimilhança do parâmetro de dispersão  $\phi$  é dada por

a)  $\hat{\phi} = D_p/n$  (modelos normal e normal inverso);

b)  $\hat{\phi} = \frac{D_p}{n} \left( 1 + \sqrt{1 + \frac{2D_p}{3n}} \right)^{-1}$  (modelo gama, expressão aproximada para  $\phi$  pequeno) (Cordeiro e McCullagh, 1991).

8. Considere uma única resposta  $Y \sim B(m, \pi)$ .

a) obtenha a expressão para a estatística de Wald  $W = (\hat{\pi} - \pi)^T \hat{K} (\hat{\pi} - \pi)$ , em que  $\hat{\pi}$  é a estimativa de máxima verossimilhança de  $\pi$  e  $\hat{K}$  é a informação de Fisher estimada em  $\hat{\pi}$ ;

b) obtenha a expressão para a estatística escore  $S_R = \tilde{U}^T \tilde{K}^{-1} \tilde{U}$  e verifique que é igual à estatística de Wald;

c) obtenha a expressão para a estatística da razão de verossimilhanças  $\Lambda = 2[\ell(\hat{\mu}) - \ell(\mu)]$ ;

d) para amostras grandes, as estatísticas escore, de Wald e da razão de verossimilhanças têm distribuição assintótica  $\chi_1^2$ . Sejam  $m = 10$  e  $y = 3$ . Compare essas

estatísticas usando  $\pi = 0, 1$ ;  $\pi = 0, 3$  e  $\pi = 0, 5$ . Quais as conclusões obtidas?

**9.** Seja  $Y_1, \dots, Y_n$  uma amostra aleatória de uma distribuição exponencial de média  $\mu$ . Sejam as hipóteses  $H_0 : \mu = \mu_0$  versus  $H : \mu \neq \mu_0$ . Mostre que:

- a)  $w = 2n \left[ \log \left( \frac{\mu_0}{\bar{y}} \right) + \frac{\bar{y} - \mu_0}{\mu_0} \right]$  (teste da razão de verossimilhanças);
- b)  $W = \frac{n(\bar{y} - \mu_0)^2}{\bar{y}^2}$  (teste de Wald);
- c)  $S_R = \frac{n(\bar{y} - \mu_0)^2}{\mu_0^2}$  (teste escore).

**10.** Sejam  $Y_1, \dots, Y_n$  variáveis independentes com distribuição de Poisson com média  $\mu_i = \mu \rho^{i-1}$  ( $i = 1, \dots, n$ ). Obter as estatísticas escore, de Wald e da razão de verossimilhanças para o teste das hipóteses que se seguem:

- a)  $H_0 : \mu = \mu_0$  versus  $H : \mu \neq \mu_0$ , quando  $\rho$  é conhecido;
- b)  $H_0 : \rho = \rho_0$  versus  $H : \rho \neq \rho_0$ , quando  $\mu$  é conhecido.

**11.** Considere a estrutura linear  $\eta_i = \beta x_i, i = 1, \dots, n$ , com um único parâmetro  $\beta$  desconhecido e função de ligação  $\eta = (\mu^\lambda - 1)\lambda^{-1}$ ,  $\lambda$  conhecido. Obter o estimador de máxima verossimilhança de  $\beta$ , considerando-se os modelos normal, Poisson, gama, normal inverso e binomial negativo. Fazer o mesmo para o modelo binomial com função de ligação  $\eta = \log\{[(1 - \mu)^{-\lambda} - 1]\lambda^{-1}\}$ ,  $\lambda$  conhecido. Obter, ainda, as estimativas no caso de  $x_1 = x_2 = \dots = x_n$ .

**12.** No exercício anterior, considere o teste de  $H_0 : \beta = \beta_0$  versus  $H : \beta \neq \beta_0$ , sendo  $\beta_0$  um valor especificado para o parâmetro desconhecido. Obtenha: a) a variância assintótica para  $\hat{\beta}$ ; b) as estatísticas para os testes da razão de verossimilhanças, Wald e escore; c) um intervalo de confiança, com um coeficiente de confiança de  $100(1 - \alpha)\%$ , para  $\beta$ ; d) um intervalo de confiança, com um coeficiente de confiança de  $100(1 - \alpha)\%$ , para uma função  $g(\beta)$  com  $g(\cdot)$  conhecido.

**13.** Seja  $Y_1, \dots, Y_n$  uma amostra aleatória de uma distribuição gama  $G(\mu, \phi)$  com parâmetro de dispersão  $\phi$ . Demonstrar que: a) a estimativa de MV de  $\phi$  satisfaz

$\log \hat{\phi} + \psi(\hat{\phi}^{-1}) = \log(\tilde{y}/\bar{y})$ , sendo  $\bar{y}$  e  $\tilde{y}$  as médias aritmética e geométrica dos dados, respectivamente, e  $\psi(\cdot)$  é a função digama; b) uma solução aproximada é dada por  $\hat{\phi} = 2(\bar{y} - \tilde{y})/\bar{y}$ .

**14.** Sejam  $Y_i \sim N(\mu_i, \sigma^2)$  e os  $Y_i$ 's independentes,  $i = 1, \dots, n$ , com variância constante desconhecida e  $\mu_i = \exp(\beta x_i)$ . Calcular: a) a matriz de informação para  $\beta$  e  $\sigma^2$ ; b) as estatísticas escore, de Wald e da razão de verossimilhanças nos seguintes testes:  $H_1 : \beta = \beta^{(0)}$  versus  $H'_1 : \beta \neq \beta^{(0)}$  e  $H_2 : \sigma^2 = \sigma^{(0)2}$  versus  $H'_2 : \sigma^2 \neq \sigma^{(0)2}$ ; c) intervalos de confiança para  $\beta$  e  $\sigma^2$ .

**15.** Sejam  $Y_i \sim P(\mu_i)$  com  $\mu_i = \mu \rho^{i-1}$ ,  $i = 1, \dots, n$ . Calcular as estatísticas escore, de Wald e da razão de verossimilhanças nos seguintes testes: a) de  $H_0 : \mu = \mu^{(0)}$  versus  $H : \mu \neq \mu^{(0)}$  para os casos de  $\rho$  conhecido e desconhecido; b) de  $H_0 : \rho = \rho^{(0)}$  versus  $H : \rho \neq \rho^{(0)}$  para os casos de  $\mu$  conhecido e desconhecido.

**16.** Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias independentes com distribuição gama  $G(\mu_i, \phi)$ , sendo  $\phi$  o parâmetro de dispersão, com  $\mu_i = \phi^{-1} \exp(-\alpha - \beta x_i)$ , em que  $\phi$ ,  $\alpha$  e  $\beta$  são parâmetros desconhecidos, e os  $x_i$ 's são valores especificados,  $i = 1, \dots, n$ . Calcular estatísticas adequadas para os seguintes testes: a)  $H_1 : \beta = 0$  versus  $H'_1 : \beta \neq 0$ ; b)  $H_2 : \phi = \phi^{(0)}$  versus  $H'_2 : \phi \neq \phi^{(0)}$ ; c)  $H_3 : \alpha = 0$  versus  $H'_3 : \alpha \neq 0$ .

**17.** Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias normais  $N(\mu_i, \sigma^2)$ , com  $\sigma^2$  conhecido, e  $\mu_i = \alpha + \beta \exp(-\gamma x_i)$ ,  $i = 1, \dots, n$ , em que  $\alpha$ ,  $\beta$  e  $\gamma$  são parâmetros desconhecidos. a) Obter intervalos de confiança para  $\alpha$ ,  $\beta$  e  $\gamma$ ; b) Testar  $H_0 : \gamma = 0$  versus  $H : \gamma \neq 0$  por meio das estatísticas escore, de Wald e da razão de verossimilhanças; c) Como proceder em a) e b) se  $\sigma^2$  for desconhecido?

**18.** Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias independentes e identicamente distribuídas como normal  $N(\mu, \sigma^2)$ . Define-se  $Z_i = |Y_i|$ ,  $i = 1, \dots, n$ . Demonstrar que a razão de verossimilhanças no teste de  $H_0 : \mu = 0$  versus  $H : \mu \neq 0$ ,  $\sigma^2$  desconhecido,

é, assintoticamente, equivalente ao teste baseado em valores grandes da estatística

$$T = \sum_{i=1}^n Z_i^4 / \sum_{i=1}^n Z_i^2, \text{ que é uma estimativa do coeficiente de curtose de } Z.$$

**19.** Considere o teste do exercício anterior. Demonstrar as expressões das estatísticas escore  $S_R = n\bar{y}^2/\tilde{\sigma}^2$  e de Wald  $W = n\bar{y}^2/\hat{\sigma}^2$  em que  $\bar{y}$  é a média dos  $y_i$ s,  $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \bar{y})^2/n$  e  $\tilde{\sigma}^2 = \hat{\sigma}^2 + \bar{y}^2$  são as estimativas de  $\sigma^2$ , segundo  $H$  e  $H_0$ , respectivamente, e que, segundo  $H_0$ ,  $E(W) = (n-1)/(n-3)$  e  $E(S_R) = 1 + O(n^{-2})$ .

**20.** Sejam  $k$  amostras independentes de tamanhos  $n_i$  ( $i = 1, \dots, k$ ;  $n_i \geq 2$ ) retiradas de populações normais diferentes de médias  $\mu_i$  e variâncias  $\sigma_i^2$ ,  $i = 1, \dots, k$ . Formular o critério da razão de verossimilhanças para o teste de homogeneidade de variâncias,  $H_0 : \sigma_1^2 = \dots = \sigma_k^2$  *versus*  $H : \sigma_i^2$  não é constante. Como realizar esse teste na prática?

**21.** Seja um MLG com estrutura linear  $\eta_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$  e função de ligação  $g(\cdot)$  conhecida. Determinar as estatísticas para os testes da razão de verossimilhanças, de Wald e escore, para as hipóteses: a)  $H_0 : \beta_2 = \beta_3 = 0$  *versus*  $H : H_0$  é falsa; b)  $H_0 : \beta_2 = 0$  *versus*  $H : \beta_2 \neq 0$ ; c)  $H_0 : \beta_3 = 0$  *versus*  $H : \beta_3 \neq 0$ .

**22.** Considere uma tabela de contingência  $r \times s$ , em que  $Y_{ij}$  tem distribuição de Poisson  $P(\mu_{ij})$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, s$ . Para o teste da hipótese de independência linha-coluna *versus* uma alternativa geral, calcular a forma das estatísticas escore, de Wald e da razão de verossimilhanças.

**23.** Considere o problema de testar  $H_0 : \mu = 0$  *versus*  $H : \mu \neq 0$  numa distribuição normal  $N(\mu, \sigma^2)$  com  $\sigma^2$  desconhecido. Comparar as estatísticas dos testes escore, de Wald e da razão de verossimilhanças entre si e com a distribuição  $\chi^2$  assintótica.

**24.** Seja uma distribuição multinomial com probabilidades  $\pi_1, \dots, \pi_m$  dependendo de um parâmetro  $\theta$  desconhecido. Considere uma amostra de tamanho  $n$ . Calcular a forma das estatísticas dos testes da razão de verossimilhanças, escore e de Wald

para testar as hipóteses  $H_0 : \theta = \theta^{(0)}$  versus  $H : \theta \neq \theta^{(0)}$ , sendo  $\theta^{(0)}$  um valor especificado.

**25.** A estatística escore pode ser usada para escolha de um entre dois modelos separados. Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias independentes com  $Y_i$  tendo distribuição normal  $N(\mu_i, \sigma^2)$ , sendo  $\mu_i = \beta x_i$  ou  $\mu_i = \gamma z_i$ ,  $i = 1, \dots, n$ , sendo todos os parâmetros desconhecidos e os  $x_i$ 's e os  $z_i$ 's conhecidos. Propor um teste baseado na estatística escore para a escolha entre uma dessas estruturas.

**26.** Sejam  $Y_1, \dots, Y_n$  variáveis aleatórias independentes sendo que  $Y_i$ ,  $i = 1, \dots, n$ , tem distribuição binomial negativa inflacionada de zeros (BNIZ) com  $P(Y_i = y_i)$  dada no Exercício 10 do Capítulo 1 em que  $\log(\lambda) = X\beta$ ,  $\log[(\omega/(1 - \omega))] = Z\gamma$ ,  $X$  e  $Z$  são matrizes de variáveis explanatórias e  $\beta$  e  $\gamma$  vetores de parâmetros.

a) Mostre que um teste escore para a hipótese  $H_0$ : PIZ versus  $H$  : BNIZ, isto é,  $H_0 : \alpha = 0$  versus  $H_1 : \alpha > 0$  é dado por  $T = S\sqrt{\hat{J}^{\alpha\alpha}}$ , em que  $S = \frac{1}{2} \sum_i \hat{\lambda}_i^{c-1} \left\{ [(y_i - \hat{\lambda}_i)^2 - y_i] - I_{(y_i=0)} \hat{\lambda}_i^2 \hat{\omega}_i / \hat{p}_{0,i} \right\}$ ,  $\hat{J}^{\alpha\alpha}$  é o elemento superior esquerdo da inversa da matriz de informação de Fisher

$$J = \begin{bmatrix} J_{\alpha\alpha} & J_{\alpha\beta} & J_{\alpha\gamma} \\ J_{\alpha\beta} & J_{\beta\beta} & J_{\beta\gamma} \\ J_{\alpha\gamma} & J_{\beta\gamma} & J_{\gamma\gamma} \end{bmatrix}$$

avaliada na estimativa de MV sob  $H_0$ . Note que  $J_{\alpha\alpha}$  é um escalar, e que os outros elementos são, em geral, matrizes com dimensões determinadas pelas dimensões dos vetores de parâmetros  $\beta$  e  $\gamma$ . No limite, quando  $\alpha \rightarrow \infty$ , os elementos típicos da matriz de informação são dados em Ridout et al. (2001).

b) Mostre que o caso particular em que não há variáveis explanatórias para  $\lambda$  e  $\omega$ , o teste escore simplifica-se para

$$T = \frac{\sum_i [(y_i - \hat{\lambda})^2 - y_i] - n\hat{\lambda}^2\hat{\omega}}{\hat{\lambda} \sqrt{n(1 - \hat{\omega}) \left( 2 - \frac{\hat{\lambda}^2}{e^{\hat{\lambda}} - 1 - \hat{\lambda}} \right)}}.$$

# Capítulo 5

## Resíduos e Diagnósticos

### 5.1 Introdução

A escolha de um MLG envolve três passos: i) definição da distribuição (que determina a função de variância); ii) definição da função de ligação; iii) definição da matriz do modelo.

Na prática, porém, pode acontecer que após uma escolha cuidadosa de um modelo e subsequente ajuste a um conjunto de observações o resultado obtido seja insatisfatório. Isso pode ocorrer em função de algum desvio sistemático entre as observações e os valores ajustados ou, então, porque uma ou mais observações são discrepantes em relação às demais.

Desvios sistemáticos podem ocorrer pela escolha inadequada da função de variância, da função de ligação e da matriz do modelo, ou ainda pela definição errada da escala da variável dependente ou das variáveis explanatórias. Discrepâncias isoladas podem ocorrer ou porque os pontos estão nos extremos da amplitude de validade da variável explanatória, ou porque eles estão realmente errados como resultado de uma leitura errada ou uma transcrição mal feita, ou ainda porque algum fator não controlado influenciou a sua obtenção.

Na prática, em geral, há uma combinação dos diferentes tipos de falhas. Assim, por exemplo, a detecção de uma escolha errada da função de ligação pode ocorrer porque ela está realmente errada ou porque uma ou mais variáveis explanatórias estão na escala errada ou devido à presença de alguns pontos discrepantes. Isso faz

com que a verificação da adequação de um modelo para um determinado conjunto de dados seja um processo realmente difícil.

Maiores detalhes podem ser vistos em Atkinson (1985), Atkinson et al. (1989), Cordeiro (1986), McCullagh e Nelder (1989), Francis et al. (1993) e Paula (2004).

## 5.2 Técnicas para a verificação do ajuste de um modelo

As técnicas usadas para esse fim podem ser formais ou informais. As informais baseiam-se em exames visuais de gráficos para a detecção de padrões, ou então, de pontos discrepantes. As formais envolvem colocar o modelo sob pesquisa em uma classe mais ampla pela inclusão de um parâmetro (ou vetor de parâmetros) extra  $\gamma$ . As mais usadas são baseadas nos testes da razão de verossimilhanças e escore. Parâmetros extras podem aparecer devido a:

- inclusão de uma variável explanatória adicional;
- inclusão de uma variável explanatória  $x$  em uma família  $h(x, \gamma)$  indexada por um parâmetro  $\gamma$ , sendo um exemplo a família de Box-Cox;
- inclusão de uma função de ligação  $g(\mu)$  em uma família mais ampla  $g(\mu, \gamma)$ , sendo um exemplo a família de Aranda-Ordaz (1981), dada no Exercício 3 do Capítulo 2;
- inclusão de uma variável construída, por exemplo  $\hat{\eta}^2$ , a partir do ajuste original, para o teste de adequação da função de ligação;
- inclusão de uma variável “dummy” assumindo o valor 1 (um) para a unidade discrepante e 0 (zero) para as demais. Isso é equivalente a eliminar essa observação do conjunto de dados, fazer a análise com a observação discrepante e



sem ela e verificar, então, se a mudança no valor do desvio é significativa, ou não. Ambos, porém, dependem da localização do(s) ponto(s) discrepante(s).

### 5.3 Análise de resíduos e diagnóstico para o modelo clássico de regressão

No modelo clássico de regressão  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$  os elementos  $\epsilon_i$  do vetor  $\boldsymbol{\epsilon}$  são as diferenças entre os valores observados  $y_i$ 's e aqueles esperados pelo modelo ( $\mu_i$ 's). Esses elementos são chamados de erros aleatórios (ou resíduos brancos) e admite-se que os  $\epsilon_i$ 's são independentes e, além disso, que  $\epsilon_i$  tem distribuição normal  $N(0, \sigma^2)$ . Esses termos representam a variação natural dos dados, mas podem, também, ser interpretados como o efeito cumulativo de fatores que não foram considerados no modelo. Se as pressuposições do modelo são violadas, a análise resultante pode conduzir a resultados duvidosos. Esse tipo de violação do modelo dá origem às chamadas falhas sistemáticas (não linearidade, não-normalidade, heterocedasticidade, não-independência, etc). Outro fato bastante comum é a presença de pontos atípicos (falhas isoladas), que podem influenciar, ou não, no ajuste do modelo. Eles podem surgir de várias maneiras. Algumas possibilidades são:

- devido a erros grosseiros na variável resposta ou nas variáveis explanatórias, por medidas erradas ou registro da observação, ou ainda, erros de transcrição;
- observação proveniente de uma condição distinta das demais;
- modelo mal especificado (falta de uma ou mais variáveis explanatórias, modelo inadequado, etc);
- escala usada errada, talvez os dados sejam melhor descritos após uma transformação, do tipo logarítmica ou raiz quadrada;
- a parte sistemática do modelo e a escala estão corretas, mas a distribuição da resposta tem uma cauda mais longa do que a distribuição normal.

Dado um conjunto de observações e ajustando-se um determinado modelo com  $p$  parâmetros linearmente independentes, para a verificação das pressuposições devem ser considerados como elementos básicos:

- os valores estimados (ou ajustados)  $\hat{\mu}_i$ ;
- os resíduos ordinários  $r_i = y_i - \hat{\mu}_i$ ;
- a variância residual estimada (ou quadrado médio residual),  $\hat{\sigma}^2 = s^2 = \text{QMRes} = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / (n - p)$ ;
- os elementos da diagonal (“leverage”) da matriz de projeção  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ , isto é,

$$h_{ii} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i,$$

sendo  $\mathbf{x}_i^T = (x_{i1}, \dots, x_{ip})$ .

Uma idéia importante, também, é a da deleção (“deletion”), isto é, a comparação do ajuste do modelo escolhido, considerando-se todos os pontos, com o ajuste do mesmo modelo sem os pontos atípicos. As estatísticas obtidas pela omissão de um certo ponto  $i$  são denotadas com um índice entre parênteses. Assim, por exemplo,  $s_{(i)}^2$  representa a variância residual estimada para o modelo ajustado, excluído o ponto  $i$ .

### 5.3.1 Tipos de resíduos

Vale destacar que os resíduos têm papel fundamental na verificação do ajuste de um modelo. Vários tipos de resíduos foram propostos na literatura (Cook e Weisberg, 1982; Atkinson, 1985).

#### a) Resíduos ordinários

Os resíduos do processo de ajuste por mínimos quadrados são dados por:

$$r_i = y_i - \hat{\mu}_i.$$

Enquanto os erros  $\epsilon_i$ 's são independentes e têm a mesma variância, o mesmo não ocorre com os resíduos obtidos a partir do ajuste do modelo através de mínimos quadrados. Tem-se,

$$\text{Var}(\mathbf{r}) = \text{Var}[(\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{I} - \mathbf{H})\sigma^2.$$

Em particular, a variância do  $i$ -ésimo resíduo é dada por  $\text{Var}(r_i) = (1 - h_{ii})\sigma^2$ , e a covariância dos resíduos relativos às observações  $i$  e  $j$  é igual a  $\text{Cov}(r_i, r_j) = -h_{ij}$ .

Assim, o uso dos resíduos ordinários pode não ser adequado devido à heterogeneidade de variâncias. Foram, então, propostas diferentes padronizações para minimizar esse problema.

**b) Resíduos estudentizados internamente (“Studentized residuals”)**

Considerando-se  $s^2 = \text{QMRes}$  como a estimativa de  $\sigma^2$ , tem-se que um estimador não tendencioso para  $\text{Var}(r_i)$  é dado por

$$\widehat{\text{Var}}(r_i) = (1 - h_{ii})s^2 = (1 - h_{ii})\text{QMRes}$$

e como  $E(r_i) = E(Y_i - \hat{\mu}_i) = 0$ , então, o resíduo estudentizado internamente é dado por

$$rsi_i = \frac{r_i}{s\sqrt{(1 - h_{ii})}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})\text{QMRes}}}.$$

Esses resíduos são mais sensíveis do que os anteriores por considerarem variâncias distintas. Entretanto, um valor discrepante pode alterar profundamente a variância residual dependendo do modo como se afasta do grupo maior das observações. Além disso, o numerador e o denominador dessa expressão são variáveis dependentes, isto é,  $\text{Cov}(r, \text{QMRes}) \neq 0$ .

**c) Resíduos estudentizados externamente (“jackknifed residuals”, “deletion residuals”, “externally studentized residuals”, “RStudent”)**

Para garantir a independência do numerador e denominador na padronização dos resíduos, define-se o resíduo estudentizado externamente, como

$$rse_{(i)} = \frac{r_i}{s_{(i)}\sqrt{(1 - h_{ii})}},$$

sendo  $s_{(i)}^2$  o quadrado médio residual livre da influência da observação  $i$ , ou seja, a estimativa de  $\sigma^2$ , omitindo-se a observação  $i$ . Prova-se que

$$rse_{(i)} = rsi_i \sqrt{\frac{n - p - 1}{n - p - rsi_i^2}},$$

sendo  $p$  o número de parâmetros independentes do modelo e  $rsi_i$  definido no item b).

A vantagem de usar o resíduo  $rse_{(i)}$  é que, sob normalidade, tem distribuição  $t$  de Student com  $(n - p - 1)$  graus de liberdade. Embora não seja recomendada a prática de testes de significância na análise de resíduos, sugere-se que a  $i$ -ésima observação seja merecedora de atenção especial se  $|rse_{(i)}|$  for maior do que o  $100[1 - \alpha/(2n)]$ -ésimo percentil da distribuição  $t$  com  $(n - p - 1)$  graus de liberdade, sendo que o nível de significância  $\alpha$  é dividido por  $n$  por ser esse o número de pontos sob análise.

### 5.3.2 Estatísticas para diagnósticos

Discrepâncias isoladas (pontos atípicos) podem ser caracterizadas por terem  $h_{ii}$  e/ou resíduos grandes, serem inconsistentes e/ou serem influentes (McCullagh e Nelder, 1989, p. 404). Uma observação inconsistente é aquela que destoa da tendência geral das demais. Quando uma observação está distante das outras em termos das variáveis explanatórias ela pode ser, ou não, influente. Uma observação influente é aquela cuja omissão do conjunto de dados resulta em mudanças substanciais em certos aspectos do modelo. Essa observação pode ser um “outlier” (observação aberrante), ou não. Uma observação pode ser influente de diversas maneiras, isto é,

- no ajuste geral do modelo;
- no conjunto de estimativas dos parâmetros;

- na estimativa de um determinado parâmetro;
- na escolha de uma transformação da variável resposta ou de uma variável explanatória.

As estatísticas mais utilizadas para verificar de pontos atípicos são:

- Medida de “leverage”:  $h_{ii}$ ;
- Medida de inconsistência:  $rse_{(i)}$ ;
- Medida de influência sobre o parâmetro  $\beta_j$ :  $DFBetaS_{(i)}$  para  $\beta_j$ ;
- Medidas de influência geral:  $DFFitS_{(i)}$ ,  $D_{(i)}$  ou  $C_{(i)}$ .

De uma forma geral, pode-se classificar uma observação como:

- Ponto inconsistente: ponto com  $rse_{(i)}$  grande, isto é, tal que  $|rse_{(i)}| \geq t_{\gamma/(2n);n-p-1}$ , com nível de significância igual a  $100\gamma\%$ ;
- Ponto de alavanca: ponto com  $h_{ii}$  grande, isto é, tal que  $h_{ii} \geq 2p/n$ . Pode ser classificado como bom, quando consistente, ou ruim, quando inconsistente;
- “Outlier”: ponto inconsistente com *leverage* pequeno, ou seja, com  $rse_{(i)}$  grande e  $h_{ii}$  pequeno;
- Ponto influente: ponto com  $DFFitS_{(i)}$ ,  $C_{(i)}$ ,  $D_{(i)}$  ou  $DFBetaS_{(i)}$  grande.

Essa medida é considerada grande se  $DFFitS_{(i)} \geq 2\sqrt{p/n}$ .

No pacote estatístico R, a  $i$ -ésima observação é considerada influente se  $|DFBetaS_{(i)}| > 1$ , se  $|DFFitS_{(i)}| > 3\sqrt{p/(n-p)}$ , se  $|1 - COVRATIO| > 3p/(n-p)$ , se  $D_{(i)} > F_{0,5;p;n-p}$ , ou se  $h_{ii} > 3p/n$ , em que  $COVRATIO = [E_{(i)}/E]^2$  com  $E \propto [s^{2p}/|\mathbf{X}^T\mathbf{X}|]^{1/2}$ ,  $E_{(i)} \propto [s_{(i)}^{2p}/|\mathbf{X}^T\mathbf{X}|]^{1/2}$  e  $s^2 = QMRes$ .

A seguir são descritas as estatísticas citadas.

#### a) Elementos da diagonal da matriz de projeção H ( $h_{ii}$ , “leverage”)

A distância de uma observação em relação às demais é medida por  $h_{ii}$  (medida de “leverage”). No caso particular da regressão linear simples, usando-se a variável centrada  $x_i = X_i - \bar{X}$ , tem-se:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum_{i=1}^n x_i^2} \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix}$$

e, portanto,

$$h_{ii} = \frac{1}{n} + \frac{x_i^2}{\sum_{i=1}^n x_i^2} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^n x_i^2}, \text{ elementos da diagonal de } \mathbf{H} \text{ e}$$

$$h_{ij} = \frac{1}{n} + \frac{x_i x_j}{\sum_{i=1}^n x_i^2} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n x_i^2}, \text{ elementos fora da diagonal de } \mathbf{H},$$

o que mostra que à medida que  $X$  se afasta de  $\bar{X}$  o valor de  $h_{ii}$  aumenta e que seu valor mínimo é  $1/n$ . Esse valor mínimo ocorre para todos os modelos que incluem uma constante. No caso em que o modelo de regressão passa pela origem, o valor mínimo de  $h_{ii}$  é 0 para uma observação  $X_i = 0$ . O valor máximo de  $h_{ii}$  é 1, ocorrendo quando o modelo ajustado é irrelevante para a predição em  $X_i$  e o resíduo é igual a 0. Sendo  $\mathbf{H}$  uma matriz de projeção, tem-se  $\mathbf{H} = \mathbf{H}^2$  e, portanto,

$$h_{ii} = \sum_{j=1}^n h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

concluindo-se que  $0 \leq h_{ii} \leq 1$  e  $\sum_{j=1}^n h_{ij} = 1$ . Além disso,

$$r(\mathbf{H}) = \text{tr}[\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] = \text{tr}[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}] = \text{tr}(\mathbf{I}_p) = \sum_{i=1}^n h_{ii} = p,$$

e, então, o valor médio de  $h_{ii}$  é  $p/n$ .

No processo de ajuste, como  $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$ , tem-se

$$\hat{\mu}_i = \sum_{j=1}^n h_{ij} y_j = h_{i1} y_1 + h_{i2} y_2 + \dots + h_{ii} y_i + \dots + h_{in} y_n \text{ com } 1 \leq i \leq n.$$

Vê-se, portanto, que o valor ajustado  $\hat{\mu}_i$  é a média ponderada dos valores observados e que o peso de ponderação é o valor de  $h_{ij}$ . Assim, o elemento da diagonal de  $\mathbf{H}$  é o peso com que a observação  $y_i$  participa do processo de obtenção do valor ajustado  $\hat{\mu}_i$ . Valores de  $h_{ii} \geq 2p/n$  indicam observações que merecem uma

análise mais apurada (Belsley et al., 1980, p. 17).

### b) **DFBeta** e **DFBetaS**

Essas estatísticas são importantes quando o coeficiente de regressão tem um significado prático.  $\text{DFBeta}_{(i)}$  mede a alteração no vetor estimado  $\hat{\beta}$  ao se retirar o  $i$ -ésimo ponto da análise, isto é,

$$\text{DFBeta}_{(i)} = \hat{\beta} - \hat{\beta}_{(i)} = \frac{r_i}{(1 - h_{ii})} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i.$$

ou ainda, considerando que  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{C} \mathbf{Y}$  em que  $\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ , tem-se:

$$\text{DFBeta}_{(i)} = \frac{r_i}{(1 - h_{ii})} \mathbf{c}_i^T, \quad i = 1, \dots, n,$$

sendo  $\mathbf{c}_i^T$  a  $i$ -ésima linha de  $\mathbf{C}$ . Então,

$$\text{DFBeta}_{j(i)} = \frac{r_i}{(1 - h_{ii})} c_{ji}, \quad i = 1, \dots, n, \quad j = 0, \dots, p - 1.$$

Cook e Weisberg (1982) propuseram curvas empíricas para o estudo dessa medida. Dado que  $\text{Cov}(\hat{\beta}) = \mathbf{C} \text{Var}(\mathbf{Y}) \mathbf{C}^T$ , a versão estudentizada de  $\text{DFBeta}_{j(i)}$  reduz-se a

$$\text{DFBetaS}_{j(i)} = \frac{c_{ji}}{(\sum c_{ji}^2) s_{(i)}} \frac{r_i}{(1 - h_{ii})}.$$

### c) **DFFit** e **DFFitS**

A estatística DFFit e sua versão estudentizada DFFitS medem a alteração provocada no valor ajustado pela retirada da observação  $i$ . São dadas por

$$\text{DFFit}_{(i)} = \mathbf{x}_i^T (\hat{\beta} - \hat{\beta}_{(i)}) = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}$$

e

$$\text{DFFitS}_{(i)} = \frac{\text{DFFit}_{(i)}}{\sqrt{h_{ii} s_{(i)}^2}} = \frac{\mathbf{x}_i^T (\hat{\beta} - \hat{\beta}_{(i)})}{\sqrt{h_{ii} s_{(i)}^2}} = \frac{\hat{\mu}_i - \hat{\mu}_{(i)}}{\sqrt{h_{ii} s_{(i)}^2}} \frac{r_i}{(1 - h_{ii})} \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$$

ou, ainda,

$$\text{DFFitS}_{(i)} = \left( \frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}} \frac{r_i}{s_{(i)} (1 - h_{ii})^{\frac{1}{2}}} = \left( \frac{h_{ii}}{1 - h_{ii}} \right)^{\frac{1}{2}} rse_i,$$

sendo o quociente  $\frac{h_{ii}}{1 - h_{ii}}$ , chamado potencial de influência, uma medida da distância do ponto  $x_i$  em relação às demais observações. Nota-se que DFFitS pode ser grande porque  $h_{ii}$  é grande ou porque o resíduo estudentizado externamente é grande. Valores absolutos excedendo  $2\sqrt{p/n}$  podem identificar observações influentes (Belsley et al., 1980, p. 28).

#### d) Distância de Cook

É também uma medida de afastamento do vetor de estimativas provocado pela retirada da observação  $i$ . É uma expressão muito semelhante ao DFFitS mas que usa como estimativa da variância residual aquela obtida com todas as  $n$  observações, ou ainda, considera o resíduo estudentizado internamente. É dada por

$$D_{(i)} = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T (\mathbf{X}^T \mathbf{X}) (\hat{\beta} - \hat{\beta}_{(i)})}{ps^2} = \frac{h_{ii}}{(1 - h_{ii})^2} \frac{r_i^2}{ps^2} = \left[ \frac{r_i}{(1 - h_{ii})^{\frac{1}{2}} s} \right]^2 \frac{h_{ii}}{p(1 - h_{ii})}$$

ou, ainda,

$$D_{(i)} = \frac{h_{ii} r s i_i^2}{p (1 - h_{ii})}.$$

#### e) Distância de Cook modificada

Atkinson (1981, p.25) sugere uma modificação para a distância de Cook

$$C_{(i)} = \left[ \frac{(n - p)}{p} \frac{h_{ii}}{(1 - h_{ii})} \right]^{\frac{1}{2}} |rse_{(i)}| = \left( \frac{n - p}{p} \right)^{\frac{1}{2}} \text{DFFitS}_{(i)}.$$

### 5.3.3 Tipos de gráficos

#### a) Valores observados ( $y$ ) versus variáveis explanatórias ( $x_j$ )

Esse tipo de gráfico indica a estrutura que pode existir entre a variável dependente e as diversas covariáveis. Pode indicar, também, a presença de heterocedasticidade. Pode, porém, levar a uma idéia falsa no caso de muitas covariáveis (a não ser que haja ortogonalidade entre todas).



**b) Variável explanatória  $x_j$  versus variável explanatória  $x_{j'}$** 

Esse tipo de gráfico pode indicar a estrutura que pode existir entre duas variáveis explanatórias. Pode indicar, também, a presença de heterocedasticidade. Pode, porém, levar a uma idéia falsa no caso de muitas variáveis explanatórias (a não ser que haja ortogonalidade entre todas).

**c) Resíduos versus variáveis explanatórias não incluídas ( $x_{fora}$ )**

Pode mostrar se existe uma relação entre os resíduos do modelo ajustado e uma variável ainda não incluída no modelo. Pode conduzir, também, a evidência de heterocedasticidade. Pode levar, porém, ao mesmo tipo de problema apontado nos itens a) e b). Uma alternativa melhor para esse tipo de gráfico é o gráfico da variável adicionada (“Added variable plot”).

**d) Resíduos versus variáveis explanatórias incluídas ( $x_{dentro}$ )**

Pode mostrar se ainda existe uma relação sistemática entre os resíduos e a variável  $x_j$  já incluída no modelo, isto é, por exemplo se  $x_{dentro}^2$  deve ser incluída. Esse tipo de gráfico apresenta o mesmo tipo de problema que o citado nos itens a), b) e c). Uma alternativa melhor para isso é o gráfico dos resíduos parciais (“partial residual plot”). O padrão para esse tipo de gráfico é uma distribuição aleatória de média 0 e amplitude constante. Desvios sistemáticos podem indicar:

- escolha errada da variável explanatória,
- falta de termo quadrático (ou de ordem superior),
- escala errada da variável explanatória.

**e) Resíduos versus valores ajustados**

O padrão para esse tipo de gráfico é uma distribuição aleatória de média 0 e amplitude constante. Pode mostrar heterogeneidade de variâncias.

**f) Gráficos de índices**

Servem para localizar observações com resíduos,  $h_{ii}$  (“leverage”), distância de Cook modificada etc, grandes.

**g) Gráfico da variável adicionada ou da regressão parcial (“Added variable plot”)**

Embora os gráficos dos resíduos *versus* variáveis não incluídas no modelo possam indicar a necessidade de variáveis extras no modelo, a interpretação exata deles não é clara. A dificuldade está em que, a menos que a variável explanatória, considerada para inclusão, seja ortogonal a todas as variáveis que já estão no modelo, o coeficiente angular do gráfico de resíduos não é o mesmo que o coeficiente angular no modelo ajustado, incluindo a variável em questão.

Esse tipo de gráfico pode ser usado para detectar a relação de  $\mathbf{y}$  com uma variável explanatória  $\mathbf{u}$ , ainda não incluída no modelo, livre do efeito de outras variáveis, e como isto é influenciado por observações individuais. Note que  $\mathbf{u}$  pode ser, também, uma variável construída para verificar a necessidade de uma transformação para a variável resposta, e/ou para as variáveis explanatórias. No caso do modelo linear geral, tem-se

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{u},$$

sendo  $\mathbf{u}$  uma variável a ser adicionada e  $\gamma$ , o parâmetro escalar adicional. O interesse está em se saber se  $\gamma = 0$ , isto é, se não há necessidade de se incluir a variável  $\mathbf{u}$  no modelo. A partir do sistema de equações normais, tem-se

$$\begin{bmatrix} \mathbf{X}^T \\ \mathbf{u}^T \end{bmatrix} \begin{bmatrix} \mathbf{X} & \mathbf{u} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \mathbf{y} \\ \mathbf{u}^T \mathbf{y} \end{bmatrix} \Rightarrow \begin{cases} \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{X}^T \mathbf{u} \hat{\gamma} = \mathbf{X}^T \mathbf{y} \\ \mathbf{u}^T \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{u}^T \mathbf{u} \hat{\gamma} = \mathbf{u}^T \mathbf{y} \end{cases}$$

e, portanto,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{u} \hat{\gamma})$$

e

$$\hat{\gamma} = \frac{\mathbf{u}^T (\mathbf{I} - \mathbf{H}) \mathbf{y}}{\mathbf{u}^T (\mathbf{I} - \mathbf{H}) \mathbf{u}} = \frac{\mathbf{u}^T (\mathbf{I} - \mathbf{H}) (\mathbf{I} - \mathbf{H}) \mathbf{y}}{\mathbf{u}^T (\mathbf{I} - \mathbf{H}) (\mathbf{I} - \mathbf{H}) \mathbf{u}} = \frac{\mathbf{u}^{*T} \mathbf{r}}{\mathbf{u}^{*T} \mathbf{u}}$$

que é o coeficiente angular de uma reta que passa pela origem sendo  $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$  os resíduos de  $\mathbf{y}$  ajustado para  $\mathbf{X}$  e  $\mathbf{u}^* = (\mathbf{I} - \mathbf{H})\mathbf{u}$  os resíduos de  $\mathbf{u}$  ajustado para  $\mathbf{X}$ .

O gráfico da variável adicionada (“added variable plot”) de  $\mathbf{r}$  versus  $\mathbf{u}^*$ , portanto, tem coeficiente angular  $\hat{\gamma}$  (diferente do gráfico de  $\mathbf{r}$  versus  $\mathbf{u}$ ) e é obtido a partir dos resíduos ordinários da regressão de  $\mathbf{y}$  como função de todas as variáveis explanatórias, exceto  $\mathbf{u} = x_j$ , versus os resíduos ordinários da regressão de  $\mathbf{u} = x_j$  como função das mesmas variáveis explanatórias usadas para modelar  $\mathbf{y}$ . Assim, por exemplo, para um modelo com 3 variáveis explanatórias, o gráfico da variável adicionada para  $x_3$  é obtido a partir de

$$\hat{\boldsymbol{\mu}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \Rightarrow \mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}}$$

e

$$\hat{x}_3 = \hat{\beta}'_0 + \hat{\beta}'_1 x_1 + \hat{\beta}'_2 x_2 \Rightarrow \mathbf{u}^* = x_3 - \hat{x}_3.$$

O padrão nulo do gráfico de  $\mathbf{r}$  versus  $\mathbf{u}^*$  indicará a não necessidade de inclusão da variável  $\mathbf{u}$ .

#### h) Gráfico de resíduos parciais ou gráfico de resíduos mais componente (“Partial residual plot”)

Se o interesse está em se detectar uma estrutura omitida, tal como uma forma diferente de dependência em  $\mathbf{u}$ , um gráfico usando  $\mathbf{u}$  pode ser de mais utilidade. Esse gráfico, também, tem coeficiente angular  $\hat{\gamma}$ . Consiste em se plotarem os resíduos do modelo  $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{u}$  mais  $\hat{\gamma}\mathbf{u}$  versus  $\mathbf{u}$ , isto é, no gráfico dos resíduos aumentados  $\tilde{\mathbf{r}} = \mathbf{r} + \hat{\gamma}\mathbf{u}$  versus  $\mathbf{u}$ . Por isso, também, esse gráfico é chamado de gráfico do resíduo mais componente.

#### i) Gráficos normal e semi-normal de probabilidades (“Normal plots” e “Half normal plots”)

O gráfico normal de probabilidades destaca-se por dois aspectos (Weisberg, 2005):

- identificação da distribuição originária dos dados e
- identificação de valores que se destacam no conjunto.

Seja uma amostra aleatória de tamanho  $n$ . As estatísticas de ordem correspondentes aos resíduos padronizados obtidos a partir do ajuste de um determinado modelo são  $d_{(1)}, \dots, d_{(i)}, \dots, d_{(n)}$ . O fundamento geral para a construção do gráfico normal de probabilidades é que se os valores de uma dada amostra provêm de uma distribuição normal, então os valores das estatísticas de ordem e os  $z_i$  correspondentes, obtidos da distribuição normal padrão são linearmente relacionados. Portanto, o gráfico de  $d_{(i)}$  versus  $z_i$  deve ser, aproximadamente, uma reta. Formatos aproximados comuns que indicam ausência de normalidade são:

**S (Esse)** - indica distribuições com caudas muito curtas, isto é, distribuições cujos valores estão muito próximos da média;

**S invertido (Esse invertido)** - indica distribuições com caudas muito longas e, portanto, presença de muitos valores extremos;

**J e J invertido** - indicam distribuições assimétricas, positivas e negativas, respectivamente.

Esses gráficos, na realidade são muito dependentes do número de observações, atingindo a estabilidade quando o número de observações é grande (em torno de 300). Para a construção desse gráfico seguem-se os passos:

a) ajuste um determinado modelo a um conjunto de dados e obtenha  $d_{(i)}$ , os valores ordenados de uma certa estatística de diagnóstico (resíduos, distância de Cook,  $h$  etc);

b) dada a estatística de ordem na posição  $(i)$ , calcule a respectiva probabilidade acumulada  $p_i$  e o respectivo quantil, ou seja, o inverso da função de distribuição normal  $\Phi(\cdot)$ , no ponto  $p_i$ . Essa probabilidade  $p_i$  é, em geral, aproximada por

$$p_i = \frac{i - c}{n - 2c + 1}$$

sendo  $0 < c < 1$ . Diversos valores têm sido propostos para a constante  $c$ . Vários autores recomendam a utilização de  $c = 3/8$ , ficando, então,

$$z_i = \Phi^{-1} \left( \frac{i - 0,375}{n + 0,25} \right), \text{ para } i = 1, 2, \dots, n.$$

c) coloque, em um gráfico,  $d_{(i)}$  versus  $z_i$ .

Esse gráfico tem, também, o nome de “Q-Q plot”, por relacionar os valores de um quantil amostral ( $d_{(i)}$ ) versus os valores do quantil correspondente da distribuição normal ( $z_i$ ).

A construção do gráfico semi-normal de probabilidades (“half normal plot”) é o resultado do conjunto de pontos obtidos pelo gráfico dos valores  $|d_{(i)}|$  versus  $z_i$  em que  $z_i = \Phi^{-1}(i + n - 0,125)/(2n + 0,5)$ .

McCullagh e Nelder (1989) sugerem o uso do gráfico normal de probabilidades para os resíduos e o gráfico semi-normal de probabilidades (“half normal plot”) para medidas positivas como é o caso de  $h_{ii}$  (medida de “leverage”) e da distância de Cook modificada. No caso do gráfico normal de probabilidades para os resíduos, espera-se que na ausência de pontos discrepantes, o aspecto seja linear, mas não há razão para se esperar que o mesmo aconteça quando são usados  $h_{ii}$  ou a distância de Cook modificada. Os valores extremos aparecerão nos extremos do gráfico, possivelmente com valores que desviam da tendência indicada pelos demais.

Para auxiliar na interpretação do gráfico semi-normal de probabilidades (“half normal plot”), Atkinson (1985) propôs a adição de um envelope simulado. Esse gráfico é obtido, seguindo-se os passos:

a) ajuste um determinado modelo a um conjunto de dados e obtenha  $d_{(i)}$ , os valores absolutos ordenados de uma certa estatística de diagnóstico (resíduos, distância de Cook,  $h_{ii}$  (“leverage”) etc);

b) simule 19 amostras da variável resposta, usando as estimativas obtidas após um determinado modelo ser ajustado aos dados e os mesmos valores para as variáveis explanatórias;

- c) ajuste o mesmo modelo a cada uma das 19 amostras e calcule os valores absolutos ordenados da estatística de diagnóstico de interesse,  $d_{j(i)}^*$ ,  $j = 1, \dots, 19$ ,  $i = 1, \dots, n$ ;
- d) para cada  $i$ , calcule a média, o mínimo e o máximo dos  $d_{j(i)}^*$ ;
- e) coloque em um gráfico as quantidades obtidas no item anterior e  $d_{(i)}$  versus  $z_i$ .

Esse envelope é tal que sob o modelo correto as estatísticas (resíduos, “leverage”, distância de Cook etc) obtidas a partir das observações caem dentro do envelope. Demétrio e Hinde (1997) apresentam um conjunto de macros que permitem fazer esses gráficos para uma grande variedade de modelos, usando o software GLIM.

#### j) Valores observados ( $y$ ) ou Resíduos *versus* tempo

Mesmo que o tempo não seja uma variável incluída no modelo, gráficos de respostas ( $y$ ) ou de resíduos versus tempo devem ser feitos sempre que possível. Esse tipo de gráfico pode conduzir à detecção de padrões não suspeitados, devido ao tempo ou, então, a alguma variável altamente correlacionada com o tempo.

## 5.4 Análise de resíduos e diagnóstico para modelos lineares generalizados

As técnicas usadas para análise de resíduos e diagnóstico para MLG são semelhantes àsquelas usadas para o modelo clássico de regressão, com algumas adaptações. Assim, por exemplo, na verificação da pressuposição de linearidade para o modelo clássico de regressão usam-se os vetores  $\mathbf{y}$  e  $\hat{\boldsymbol{\mu}}$  enquanto que para o MLG devem ser usados  $\hat{\mathbf{z}}$ , a variável dependente ajustada estimada, e  $\hat{\boldsymbol{\eta}}$ , o preditor linear. A variância residual  $s^2$  é substituída por uma estimativa consistente do

parâmetro de dispersão  $\phi$  e a matriz de projeção  $\mathbf{H}$  é definida por

$$\mathbf{H} = \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}, \quad (5.1)$$

o que é equivalente a substituir  $\mathbf{X}$  por  $\mathbf{W}^{1/2} \mathbf{X}$ . Note-se que, agora  $\mathbf{H}$  depende das variáveis explanatórias, da função de ligação e da função de variância, tornando mais difícil a interpretação da medida de “leverage”. Demonstra-se que

$$\mathbf{V}^{-1/2}(\hat{\boldsymbol{\mu}} - \mathbf{u}) \cong \mathbf{H} \mathbf{V}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu}),$$

sendo  $\mathbf{V} = \text{diag}\{V(\mu_i)\}$ . Isso mostra que  $\mathbf{H}$  mede a influência em unidades estudentizadas de  $\mathbf{y}$  sobre  $\hat{\boldsymbol{\mu}}$ .

#### 5.4.1 Tipos de resíduos

Os resíduos são importantes para detectar a presença de observações aberrantes que devem ser estudadas detalhadamente. O resíduo  $R_i$  deve expressar uma discrepância (distância) entre a observação  $y_i$  e o seu valor ajustado  $\hat{\mu}_i$

$$R_i = h_i(y_i, \hat{\mu}_i), \quad (5.2)$$

em que  $h_i$  é uma função adequada de fácil interpretação, usualmente escolhida para estabilizar a variância ou induzir simetria na distribuição amostral de  $R_i$ . A definição (5.2) foi dada por Cox e Snell (1968). A mesma função  $h_i(\cdot) = h(\cdot)$  pode ser usada para as diversas observações. A matriz  $\mathbf{H}$  em (5.1) desempenha um papel importante na análise dos resíduos em MLG e tem as propriedades  $\text{tr}(\mathbf{H}) = p$  e  $0 \leq h_{ii} \leq 1$ , descritas na Seção 5.3.2 no contexto do modelo clássico de regressão. Outra matriz importante de projeção é definida como  $\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{W}^{1/2} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}$ . As escolhas mais comuns de  $h_i$  são  $R_i = (y_i - \hat{\mu}_i)/[\text{Var}(y_i)]^{1/2}$  e  $R_i = (y_i - \hat{\mu}_i)/[\text{Var}(y_i - \hat{\mu}_i)]^{1/2}$ , a primeira forma sendo a mais usual, sendo que as expressões da variância nos denominadores são estimadas segundo o modelo sob pesquisa. Em algumas aplicações esses resíduos não são apropriados para detectar anomalias no ajuste do modelo estatístico.

Em geral, a definição da função  $h_i$  depende, basicamente, do tipo de anomalia que se deseja detectar no modelo. Entre as anomalias mais freqüentes, citam-se:

- i) uma falsa distribuição populacional para a variável resposta;
- ii) uma ou mais observações não pertencendo à distribuição proposta para a variável resposta;
- iii) algumas observações que se mostram dependentes ou exibindo alguma forma de correlação serial;
- iv) um parâmetro importante que está sendo omitido no modelo.

Escolhendo  $h_i$ , adequadamente, estas anomalias podem ser descobertas através dos gráficos respectivos:

i') resíduos ordenados  $R_{(i)}$  *versus* pontos percentuais de alguma distribuição de probabilidade de referência  $F(\cdot)$ ; esses pontos podem ser definidos por  $F^{-1}[(i - \alpha)/(n - 2\alpha + 1)]$  para  $0 \leq \alpha \leq 0,5$ ;

ii')  $R_i$  *versus*  $\hat{\mu}_i$ ;

iii')  $R_i$  *versus*  $i$ ;

iv')  $R_i$  *versus* os níveis da variável ou fator correspondente ao parâmetro omitido.

Geralmente, esses gráficos representam o método mais importante de análise dos resíduos. A definição dos resíduos através de (5.2) deve satisfazer, aproximadamente, propriedades de segunda ordem, tais como,  $E(R_i) = 0$ ,  $\text{Var}(R_i) = \text{constante}$  e  $\text{Cov}(R_i, R_j) = 0$ ,  $i \neq j$ , pois, em muitos casos, essas condições são suficientes para especificar a forma da distribuição de  $R_i$ .

O **resíduo verdadeiro** é definido por  $\epsilon_i = h_i(y_i, \mu_i)$ . A quantidade de observações para estimar os parâmetros  $\beta'$ s do modelo e fornecer informações sobre a distribuição de probabilidade dos resíduos deve ser grande. Frequentemente,  $p$  é pequeno comparado com  $n$  e as combinações de parâmetros são estimadas com erro padrão de ordem  $n^{-1/2}$ . Nesse caso, o resíduo  $R_i$  difere de  $\epsilon_i$  de uma quantidade de ordem  $n^{-1/2}$  em probabilidade, e muitas propriedades estatísticas dos  $R_i$ 's são equivalentes às propriedades respectivas dos  $\epsilon_i$ 's.



Em geral, a distribuição exata de  $R_i$  não é conhecida e trabalha-se com resultados assintóticos, tais como, valor esperado  $E(R_i)$  e variância  $\text{Var}(R_i)$  até ordem  $n^{-1}$ . Theil (1965) sugere usar uma combinação linear dos resíduos  $\mathbf{R}^* = \mathbf{C}\mathbf{R}$ , no lugar de  $\mathbf{R} = (R_1, \dots, R_n)^T$ , para testes e gráficos, em que  $\mathbf{C}$  é uma matriz  $(n - p) \times n$ , escolhida de tal forma que  $\mathbf{R}^*$  tenha, aproximadamente, uma distribuição normal multivariada  $N(\mathbf{0}, \sigma^2 \mathbf{I})$ . Apresentam-se, a seguir, os tipos de resíduos mais comuns nos MLG.

#### a) Resíduos ordinários

$$r_i = y_i - \hat{\mu}_i.$$

Esses resíduos não têm maior interesse nos MLG.

#### b) Resíduos de Pearson

O resíduo mais simples é o de Pearson, definido por

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\hat{V}_i^{1/2}}. \quad (5.3)$$

Esta quantidade é um componente da estatística de Pearson generalizada  $X_p^2 = \sum_{i=1}^n r_i^{P^2}$  dada em (4.3). Para os modelos log-lineares tem-se que  $r_i^P = (y_i - \hat{\mu}_i) \hat{\mu}_i^{-1/2}$  e Haberman (1974) sugere o ajuste  $r_i^{P*} = r_i^P / (1 - \hat{\mu}_i \hat{z}_{ii})^{1/2}$ , em que  $\mathbf{Z} = \{z_{ij}\} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T$  com  $\mathbf{W} = \text{diag}\{\mu_i\}$ , para tornar a distribuição do resíduo  $r_i^{P*}$ , aproximadamente, normal  $N(0, 1)$ . A variância média dos  $r_i^{P*}$  é  $1 - (1 + p)n^{-1}$ . Podem-se incorporar pesos a priori na fórmula (5.3). A desvantagem do resíduo de Pearson é que sua distribuição é, geralmente, bastante assimétrica para modelos não-normais. Cordeiro (2004a) apresenta expressões para a média e a variância de  $r_i^P$  corretas até ordem  $n^{-1}$ .

#### c) Resíduos de Anscombe

Anscombe (1953) apresenta uma definição geral de resíduo, através de uma transformação  $N(y_i)$  da observação  $y_i$ , escolhida visando tornar a sua distribuição o

mais próxima possível da distribuição normal. Barndorff-Nielsen (1978) demonstra que, para os MLG,  $N(\cdot)$  é dada por  $N(\mu) = \int V^{-1/3} d\mu$ . Como  $N'(\mu)(V/\phi)^{1/2}$  é a aproximação de primeira ordem do desvio padrão de  $N(y)$ , o resíduo de Anscombe, visando à normalização e à estabilização da variância, é expresso por

$$A_i = \frac{N(y_i) - N(\hat{\mu}_i)}{N'(\hat{\mu}_i)\hat{V}_i^{1/2}}. \quad (5.4)$$

Da definição do resíduo de Anscombe, conclui-se que a transformação aplicada aos dados para normalizar os resíduos é a mesma que aplicada às médias dos dados normaliza a distribuição de  $\hat{\beta}$  (caso  $\delta = 2/3$ , em (8.1)).

Para os modelos de Poisson, gama e normal inverso, os resíduos de Anscombe são facilmente obtidos de (5.4) como  $3(y^{2/3} - \hat{\mu}^{2/3})/(2\hat{\mu}^{1/6})$ ,  $3(y^{1/3} - \hat{\mu}^{1/3})/\hat{\mu}^{1/3}$  e  $(\log y - \log \hat{\mu})/\hat{\mu}^{1/2}$ , respectivamente. Para o modelo binomial  $B(m, \mu)$ , (5.4) reduz-se a  $A_i = m_i^{1/2}[N(y_i) - N(\hat{\mu}_i)]/[\hat{\mu}_i(1 - \hat{\mu}_i)]^{1/6}$ , em que  $N(\mu) = \int [\mu(1 - \mu)]^{-1/3} d\mu$ . Cox e Snell (1968) calculam esse resíduo através da função beta incompleta.

#### d) Resíduos de Pearson estudentizados

$$r_i^{P'} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - h_{ii})}}, \quad (5.5)$$

sendo  $h_{ii}$  o  $i$ -ésimo elemento da diagonal da matriz definida em (5.1). Os resíduos estudentizados (5.5) têm, aproximadamente, variância igual a um quando o parâmetro de dispersão  $\phi \rightarrow 0$ .

#### e) Componentes do desvio

Os resíduos podem, também, ser definidos como iguais às raízes quadradas dos componentes do desvio com sinal dado pelo sinal de  $y_i - \hat{\mu}_i$ . Tem-se,

$$r_i^D = \text{sin}(\text{al})(y_i - \hat{\mu}_i)\sqrt{2}[v(y_i) - v(\hat{\mu}_i) + q(\hat{\mu}_i)(\hat{\mu}_i - y_i)]^{1/2}, \quad (5.6)$$

em que a função  $v(x) = xq(x) - b(q(x))$  é definida em termos das funções  $b(\cdot)$  e  $q(\cdot)$  dadas na Seção 1.3.

O resíduo  $r_i^D$  representa uma distância da observação  $y_i$  ao seu valor ajustado  $\hat{\mu}_i$ , medida na escala do logaritmo da função de verossimilhança. Tem-se  $D_p = \sum_{i=1}^n r_i^{D^2}$ . Um valor grande para  $r_i^D$  indica que a  $i$ -ésima observação é mal ajustada pelo modelo. Pregibon (1979) demonstra que, se existe uma transformação  $h_i$  que normaliza a distribuição do resíduo  $R_i = h_i(y_i, \hat{\mu}_i)$ , então as raízes quadradas dos componentes do desvio são resíduos que exibem as mesmas propriedades induzidas por essa transformação. Assim, os resíduos  $r_i^D$  podem ser tratados, aproximadamente, como variáveis aleatórias normais reduzidas e, conseqüentemente,  $r_i^{D^2}$  como tendo, aproximadamente, distribuição  $\chi_1^2$ .

Para os modelos de Poisson, gama, binomial e normal inverso, os resíduos definidos como as raízes quadradas dos componentes do desvio, têm as formas respectivas:  $\delta \{2[y \log(y/\hat{\mu}) + \hat{\mu} - y]\}^{1/2}$ ,  $\delta \{2[\log(\hat{\mu}/y) + (y - \hat{\mu})/\hat{\mu}]\}^{1/2}$ ,  $\delta (2m\{y \log(y/\hat{\mu}) + (1 - y) \log[(1 - y)/(1 - \hat{\mu})]\})^{1/2}$  e  $(y - \hat{\mu})/(y^{1/2}\hat{\mu})$ , em que  $\delta$  representa o sinal de  $(y - \hat{\mu})$ .

As vantagens do resíduo (5.6) são: a) não requer o conhecimento da função normalizadora; b) computação simples após o ajuste do MLG; c) é definido para toda observação e, mesmo para observações censuradas, desde que essas forneçam uma contribuição para o logaritmo da função de verossimilhança.

#### f) Componentes do desvio estudentizados

$$r_i^{D'} = \frac{r_i^D}{\sqrt{1 - \hat{h}_{ii}}}.$$

Os resíduos  $r_i^{D'}$  são, então, definidos a partir de (5.6). Os resíduos de Pearson, de Anscombe e componentes do desvio dados em (5.3), (5.4) e (5.6), respectivamente, são os mais importantes nas aplicações dos MLG.

No modelo normal, nenhuma distinção é feita entre esses três tipos de resíduos. Para modelos bem ajustados, as diferenças entre  $r_i^D$  e  $r_i^P$  devem ser pequenas. Entretanto, para os modelos mal-ajustados e/ou para observações aberrantes, podem ocorrer diferenças consideráveis entre esses resíduos. Embora os

resíduos, definidos por (5.4) e (5.6), apresentem formas bem diferentes para modelos não-normais, os seus valores, dados  $y$  e  $\hat{\mu}$ , são similares. Admite-se que  $\hat{\mu} = cy$ , em que  $c$  é um real qualquer. Seja A/D o quociente entre os resíduos de Anscombe (A) e aquele, definido como a raiz quadrada do componente do desvio (D). Para os modelos de Poisson, gama e normal inverso esse quociente é igual a  $3\delta(1 - c^{2/3})/(2\sqrt{2})c^{1/6}(c - 1 - \log c)^{1/2}$ ;  $3\delta(1 - c^{1/3})c^{1/6}/\sqrt{2}(c \log c + 1 - c)^{1/2}$  e  $c^{1/2} \log c/(c - 1)$ , respectivamente, em que  $\delta = +1(-1)$  quando  $c < 1(> 1)$ .

A Tabela 5.1 apresenta valores do quociente A/D para esses três modelos. Dessa tabela, conclui-se que esses dois resíduos são, aproximadamente, equivalentes. Essa equivalência poderia ainda ser determinada por expansões em série de Taylor. McCullagh e Nelder (1989) comparam os resíduos de Pearson, de Anscombe e como componentes do desvio para o modelo de Poisson.

Tabela 5.1: Relação A/D entre o resíduo de Anscombe e o definido como a raiz quadrada do componente do desvio, para três modelos.

c	Poisson	gama	normal inverso
0,1	1,0314	0,9462	0,8090
0,2	1,0145	0,9741	0,8997
0,4	1,0043	0,9918	0,9658
0,6	1,0014	0,9977	0,9892
0,8	1,0010	0,9994	0,9979
2,0	1,0019	0,9958	0,9802
3,0	1,0048	0,9896	0,9514
5,0	1,0093	0,9790	0,8997
10,0	1,0169	0,9598	0,8090

Definindo-se uma distribuição teórica conveniente para os resíduos, podem-se aplicar as diversas técnicas analíticas e gráficas para detectar desvios do modelo sob pesquisa.

### 5.4.2 Tipos de gráficos

São basicamente os mesmos gráficos já estudados na Seção 5.3.3 com algumas modificações e com interpretações semelhantes.

#### a) Resíduos versus alguma função dos valores ajustados

É recomendado o gráfico de algum tipo de resíduo estudentizado ( $r_i^{P'}$  ou  $r_i^{D'}$ ) *versus*  $\hat{\eta}_i$ , ou então, *versus* os valores ajustados transformados de tal forma a se ter variância constante para a distribuição em uso. Assim, usar, no eixo das abscissas,  $\hat{\mu}_i$  para a distribuição normal,  $2\sqrt{\hat{\mu}_i}$  para a Poisson,  $2\arcsen(\hat{\mu}_i/m_i)$  para a binomial,  $\log \hat{\mu}_i$  para a gama e  $-2\hat{\mu}_i^{-1/2}$  para a normal inversa. O padrão nulo desse gráfico é uma distribuição dos resíduos em torno de zero com amplitude constante. Desvios sistemáticos podem ter algum tipo de curvatura ou, então, mudança sistemática da amplitude com o valor ajustado.

#### b) Resíduos *versus* variáveis explanatórias não incluídas

Esse gráfico pode mostrar se existe uma relação entre os resíduos do modelo ajustado e uma variável ainda não incluída no modelo. Uma alternativa melhor para esse tipo de gráfico é o gráfico da variável adicionada (“added variable plot”). O padrão nulo desse gráfico é uma distribuição dos resíduos em torno de zero com amplitude constante.

#### c) Resíduos *versus* variáveis explanatórias já incluídas

Esse gráfico pode mostrar se ainda existe uma relação sistemática entre os resíduos e uma variável que está incluída no modelo. Alternativa melhor para isso é o gráfico de resíduos parciais (“partial residual plot”). O padrão nulo para esse tipo de gráfico é uma distribuição aleatória de média 0 e amplitude constante.

#### d) Gráfico da variável adicionada ou da regressão parcial (“added variable plot”)

Inicialmente, ajusta-se o modelo com preditor linear  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ . Em seguida,

faz-se o gráfico de  $\widehat{\mathbf{W}}^{-1/2}\mathbf{s}$  *versus*  $(\mathbf{I} - \widehat{\mathbf{H}})\widehat{\mathbf{W}}^{1/2}\mathbf{u}$ , sendo  $\mathbf{s}$  o vetor com elementos estimados por

$$s_i = \frac{(y_i - \hat{\mu}_i) \widehat{d\mu}_i}{V(\hat{\mu}_i) d\eta_i}$$

e  $\mathbf{u}$  o vetor com os valores da variável a ser adicionada (Wang, 1985). Aqui  $\widehat{\mathbf{W}}^{-1/2}\mathbf{s}$  representa o vetor de elementos  $(y_i - \hat{\mu}_i)V(\hat{\mu}_i)^{-1/2}$  (resíduo de Pearson generalizado da regressão ponderada de  $\mathbf{y}$  em relação a  $\mathbf{X}$  com matriz de pesos estimada  $\widehat{\mathbf{W}}$ ) e  $(\mathbf{I} - \widehat{\mathbf{H}})\widehat{\mathbf{W}}^{1/2}\mathbf{u}$  representa os resíduos da regressão ponderada de  $\mathbf{u}$  em relação a  $\mathbf{X}$  com matriz de pesos estimada  $\widehat{\mathbf{W}}$ .

**e) Gráfico de resíduos parciais ou gráfico de resíduos mais componente (“partial residual plot”)**

Inicialmente, ajusta-se o MLG com preditor linear  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{u}$ , obtendo-se  $\widehat{\mathbf{W}}^{-1}\mathbf{s}$  e  $\hat{\gamma}$ . Em seguida, faz-se o gráfico de  $\widehat{\mathbf{W}}^{-1}\mathbf{s} + \hat{\gamma}\mathbf{u}$  *versus*  $\mathbf{u}$  (Wang, 1987). O padrão nulo desse gráfico é linear com coeficiente angular  $\hat{\gamma}$  se a escala da variável  $\mathbf{u}$  está adequada. A forma desse gráfico pode sugerir uma escala alternativa para  $\mathbf{u}$ .

**f) Gráficos de índices**

Servem para localizar observações com resíduo, “leverage” ( $h_{ii}$ ), distância de Cook modificada etc, grandes.

**g) Gráfico normal e semi-normal de probabilidades (“normal plots” e “half normal plots”)**

São construídos da mesma forma que para os modelos clássicos de regressão, usando-se, porém, a distribuição pertinente.

**h) Valores observados ou resíduos versus tempo**

Mesmo que o tempo não seja uma variável incluída no modelo, gráficos de respostas ( $y$ ) ou de resíduos *versus* tempo devem ser construídos sempre que possível.

Esse tipo de gráfico pode conduzir à detecção de padrões concebidos a priori, devido ao tempo ou, então, a alguma variável altamente correlacionada com o tempo.

### 5.4.3 Resíduos de Pearson estudentizados

Na expressão geral dos resíduos  $R_i = h_i(y_i, \hat{\mu}_i)$ , dada em (5.2), a função  $h_i$  deve ser escolhida visando a satisfazer as propriedades de segunda ordem:  $E(R_i) = 0$  e  $\text{Var}(R_i) = \text{constante}$ . Cox e Snell (1968) apresentam fórmulas gerais para  $E(R_i)$ ,  $\text{Cov}(R_i, R_j)$  e  $\text{Var}(R_i)$  até termos de ordem  $n^{-1}$ , válidas para qualquer função  $h_i$  especificada. Essas fórmulas possibilitam calcular resíduos modificados cujas distribuições são melhor aproximadas pelas distribuições de probabilidade de referência.

Cordeiro (2004a) usa os resultados de Cox e Snell para obter expressões matriciais aproximadas, até ordem  $n^{-1}$ , para os valores esperados, variâncias e covariâncias dos resíduos de Pearson, válidas para qualquer MLG. Essas expressões dependem das funções de ligação e de variância e de suas duas primeiras derivadas.

Demonstra-se, a seguir, que os resíduos de Pearson têm estrutura de covariância dada, aproximadamente, pela matriz de projeção do MLG, vista na Seção 5.4.1, isto é, por  $\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{Z}\mathbf{W}^{1/2}$ , em que  $\mathbf{Z} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$  é a covariância assintótica de  $\hat{\boldsymbol{\eta}}$ . Essa aproximação não está correta até termos de ordem  $n^{-1}$  (Cordeiro, 2004a).

O algoritmo (3.5) de ajuste do MLG avaliado na EMV  $\hat{\boldsymbol{\beta}}$  implica

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{W}}\hat{\mathbf{z}},$$

em que  $\hat{\mathbf{z}} = \hat{\boldsymbol{\eta}} + \widehat{\mathbf{H}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ . Logo, da definição da matriz  $\mathbf{Z}$ , tem-se

$$\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}} = (\mathbf{I} - \widehat{\mathbf{Z}}\widehat{\mathbf{W}})\hat{\mathbf{z}}.$$

Supondo que  $\mathbf{Z}$  e  $\mathbf{W}$  são tais que, aproximadamente, pelo menos o produto  $\widehat{\mathbf{Z}}\widehat{\mathbf{W}}$  é constante, pode-se escrever

$$\text{Cov}(\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}}) \approx (\mathbf{I} - \mathbf{Z}\mathbf{W})\text{Cov}(\hat{\mathbf{z}})(\mathbf{I} - \mathbf{Z}\mathbf{W})^T$$

e como  $\text{Cov}(\hat{\mathbf{z}}) = \mathbf{W}^{-1}$ , tem-se

$$\text{Cov}(\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}}) \approx \mathbf{W}^{-1/2}(\mathbf{I} - \mathbf{H})\mathbf{W}^{-1/2},$$

em que  $\mathbf{H}_2 = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{Z}\mathbf{W}^{1/2}$ . Logo,

$$\text{Cov}[\widehat{\mathbf{W}}^{1/2}(\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}})] \approx \mathbf{H}_2;$$

A expressão  $\widehat{\mathbf{W}}^{1/2}(\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}})$  é igual a  $\widehat{\mathbf{V}}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ , em que  $\mathbf{V} = \text{diag}\{V_1, \dots, V_n\}$ , e representa o vetor cujos componentes são iguais aos resíduos de Pearson (5.3) e, então, a demonstração está concluída. Convém enfatizar que os resultados de Cordeiro (2004a) mostram que a expressão de  $\text{Cov}[\widehat{\mathbf{W}}^{1/2}(\hat{\mathbf{z}} - \hat{\boldsymbol{\eta}})]$  está correta até ordem  $n^{-1}$  para os elementos fora da diagonal, mas não para os elementos da diagonal.

A conclusão prática importante é que para uma análise mais cuidadosa dos gráficos i'), i''), i''') e iv'), descritos na Seção 5.4.1, devem-se usar os resíduos de Pearson estudentizados  $r_i^{P'}$  definidos em (5.5), em que o denominador é  $[V(\hat{\mu}_i)(1 - \hat{h}_{ii})]^{1/2}$  ao invés de  $V(\hat{\mu}_i)^{1/2}$ .

Nos casos em que as variáveis explanatórias apresentam configurações irregulares, o uso dos resíduos  $r_i^{P'}$  é fundamental. Essa correção, também, será importante para identificar observações aberrantes. Em geral, no gráfico de  $r_i^{P'}$  versus  $1 - h_{ii}$  observam-se, facilmente, dados com grandes resíduos e/ou variâncias pequenas e, portanto, esse gráfico pode ajudar na identificação de pontos aberrantes.

Os resíduos  $r_i^{P'}$  apresentam propriedades razoáveis de segunda ordem mas podem ter distribuições bem diferentes da normal. Por essa razão, os resíduos definidos em (5.5) como as raízes quadradas dos componentes do desvio podem ser preferidos nos gráficos de resíduos versus pontos percentuais da distribuição normal padrão. Pregibon (1979) propõe, também, o uso do mesmo fator de correção  $(1 - \hat{h}_{ii})^{1/2}$  para os resíduos  $r_i^D$ , isto é, sugere trabalhar com as raízes quadradas dos componentes do desvio divididos por  $(1 - \hat{h}_{ii})^{1/2}$ , ou seja, com os componentes do desvio estudentizados definidos na Seção 5.4.1 (item e). Entretanto, a adequação da



distribuição normal para aproximar a distribuição de  $r_i^{D'}$  ainda é um problema a ser pesquisado.

## 5.5 Verificação da função de ligação

Um método informal para a verificação da adequação da função de ligação usada é o gráfico da variável dependente ajustada estimada  $\hat{\mathbf{z}}$  *versus* o preditor linear estimado  $\hat{\boldsymbol{\eta}}$ . O padrão nulo é uma reta. O gráfico da variável adicionada também pode ser usado, considerando-se  $\mathbf{u} = \hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$ , sendo que o padrão nulo indicará que a função de ligação usada é adequada.

Para funções de ligação na família potência, uma curvatura para cima no gráfico indica que deve ser usada uma função de ligação com expoente maior enquanto que uma curvatura para baixo indica um expoente menor. Esse tipo de gráfico não serve para dados binários.

Existem dois métodos formais para a verificar a adequacidade da função de ligação utilizada:

i) o mais simples consiste em se adicionar  $\mathbf{u} = \hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$  como uma variável explanatória extra e examinar a mudança ocorrida no desvio, o que equivale ao teste da razão de verossimilhanças. Se ocorrer uma diminuição drástica há evidência de que a função de ligação é insatisfatória. Pode-se usar, também, o teste escore;

ii) outro método formal consiste em indexar a família de funções de ligação por um parâmetro  $\lambda$  e fazer um teste da hipótese  $H_0 : \lambda = \lambda_0$ , usando-se os testes da razão de verossimilhanças e escore. Incerteza sobre a função de ligação é mais comum com dados contínuos que têm distribuição gama e com proporções cujo número de sucessos segue a distribuição binomial. Assim, por exemplo, para dados com distribuição gama, pode-se usar a família de funções de ligação  $\eta = \mu^\lambda$ . Para dados com distribuição binomial, pode-se usar a família de funções de ligação  $\eta = \log[(1 - \pi)^{-\lambda} - 1]/\lambda$  de Aranda-Ordaz (1981) que tem como casos especiais a função de ligação logística para  $\lambda = 1$  e a complemento log-log quando  $\lambda \rightarrow 0$ . Em

geral, usa-se o método do logaritmo da função de verossimilhança perfilada para se estimar  $\lambda$ . Para o modelo clássico de regressão, esse teste equivale ao teste proposto por Tukey (1949) para não-aditividade.

A verificação da adequação da função de ligação é, inevitavelmente, afetada pela falha em estabelecer escalas corretas para as variáveis explanatórias no preditor linear. Em particular, se o teste formal construído pela adição de  $\hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$  ao preditor linear produz uma redução significativa no desvio do modelo, isso pode indicar uma função de ligação errada ou escalas erradas para as variáveis explanatórias ou ambas. Pontos atípicos, também, podem afetar a escolha da função de ligação.

**Exemplo 5.1:** Seja a função de ligação  $g_0(\boldsymbol{\mu}) = g(\boldsymbol{\mu}, \lambda_0) = X\boldsymbol{\beta}$ , incluída em uma família paramétrica  $g(\boldsymbol{\mu}, \lambda)$ , indexada pelo parâmetro escalar  $\lambda$ , por exemplo,

$$g(\boldsymbol{\mu}, \lambda) = \begin{cases} \frac{\boldsymbol{\mu}^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log \boldsymbol{\mu} & \lambda = 0 \end{cases} \quad (5.7)$$

que inclui as funções de ligação identidade, logarítmica etc, ou então, a família de Aranda-Ordaz,

$$\boldsymbol{\mu} = \begin{cases} 1 - (1 + \lambda e^{\boldsymbol{\eta}})^{-\frac{1}{\lambda}} & \lambda e^{\boldsymbol{\eta}} > -1 \\ 1 & c.c. \end{cases}$$

que inclui as funções de ligação logística, complemento log-log etc.

A expansão de Taylor para  $g(\boldsymbol{\mu}, \lambda)$  em torno de um valor conhecido  $\lambda_0$ , produz

$$g(\boldsymbol{\mu}, \lambda) \simeq g(\boldsymbol{\mu}, \lambda_0) + (\lambda - \lambda_0)\mathbf{u}(\lambda_0) = X\boldsymbol{\beta} + \gamma\mathbf{u}(\lambda_0),$$

em que  $\mathbf{u}(\lambda_0) = \left. \frac{\partial g(\boldsymbol{\mu}, \lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0}$ .

De uma forma geral usa-se  $\mathbf{u} = \hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$ . A seguir, justifica-se o uso de  $\hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$  como variável adicionada ao preditor linear do modelo para o teste de adequação da função de ligação de um MLG.

Suponha que a função de ligação usada foi  $\boldsymbol{\eta} = g(\boldsymbol{\mu})$  e que a função de ligação verdadeira seja  $g^*(\boldsymbol{\mu})$ . Então,

$$g(\boldsymbol{\mu}) = g[g^{*-1}(\boldsymbol{\eta})] = h(\boldsymbol{\eta}).$$

A hipótese nula é  $H_0 : h(\boldsymbol{\eta}) = \boldsymbol{\eta}$  e a alternativa é  $H : h(\boldsymbol{\eta}) = \text{não linear}$ .

Fazendo-se a expansão de  $g(\boldsymbol{\mu})$  em série de Taylor, tem-se:

$$g(\boldsymbol{\mu}) \simeq h(0) + h'(0)\boldsymbol{\eta} + \frac{h''(0)}{2}\boldsymbol{\eta} \otimes \boldsymbol{\eta}$$

e, então, a variável adicionada é  $\hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}$ , desde que o modelo tenha termos para o qual a média geral seja marginal.

**Exemplo 5.2:** Considere os dados do Exemplo 2.1. A variável resposta tem distribuição binomial, isto é,  $Y_i \sim B(m_i, \pi_i)$ . Adotando-se a função de ligação logística (canônica) e os preditores lineares dados por

$$\eta_i = \log \left( \frac{\mu_i}{m_i - \mu_i} \right) = \beta_1 + \beta_2 d_i,$$

e

$$\eta_i = \log \left( \frac{\mu_i}{m_i - \mu_i} \right) = \beta_1 + \beta_2 d_i + \gamma u_i,$$

sendo  $u_i = \eta_i^2$ , usa-se a diferença de desvios para testar a adequação da função de ligação, obtendo-se os resultados da Tabela 5.2. Verifica-se que se rejeita a hipótese nula  $H_0 : \gamma = 0$ , ao nível de 5% de significância, indicando que a função de ligação logística não é adequada. A estimativa para  $\gamma$  é  $\hat{\gamma} = -0,2087$  com erro padrão 0,0757.

Fazendo-se uma análise de resíduos verifica-se que a primeira observação é discrepante. Eliminando-a e refazendo-se o teste para a função de ligação, a hipótese nula  $H_0 : \gamma = 0$  não é rejeitada, indicando a adequação da função de ligação logística. Tem-se, então,  $\hat{\gamma} = 0,0757$  com erro padrão 0,086 e,

$$\eta_i = \log \left( \frac{\hat{\mu}_i}{m_i - \hat{\mu}_i} \right) = -3,5823 + 0,7506 d_i.$$

Tabela 5.2: Análise de desvio e teste da função de ligação para os dados do Exemplo 2.1.

Causa de variação	g.l.	Desvio	Valor de p
Regressão linear	1	153,480	$< 0,0001$
Função de ligação	1	9,185	0,0024
Novo Resíduo	3	1,073	
(Resíduo)	4	10,260	0,0527
Total	5	163,740	

## 5.6 Verificação da adequação da função de variância

Um método informal para a verificação da adequação da função de variância (que é definida ao se escolher uma determinada distribuição) é o gráfico dos resíduos absolutos *versus* os valores ajustados transformados em uma escala com variância constante (item a) da Seção 5.4.2. O padrão nulo para esse tipo de gráfico é uma distribuição aleatória de média 0 (zero) e amplitude constante. A escolha errada da função de variância mostrará uma tendência na média. Em geral, a não adequação da função de variância será tratada como superdispersão (Hinde e Demétrio, 1998a,b).

Um método formal para a verificação da adequação da função de variância consiste em indexar essa função por um parâmetro  $\lambda$  e fazer um teste de hipótese  $H_0 : \lambda = \lambda_0$ , usando-se os testes da razão de verossimilhanças e escore. Assim, por exemplo, pode-se usar  $V(\mu) = \mu^\lambda$  e observar como o ajuste varia em função da variação de  $\lambda$ . Em geral, usa-se o logaritmo da função de verossimilhança perfilada para se estimar  $\lambda$ .

Para se compararem ajustes de modelos com diferentes funções de variância, o desvio não pode ser usado, há necessidade de se usar a teoria de quase verossimilhança estendida, que será discutida na Seção 10.6.

A verificação da adequação da função de variância é, inevitavelmente, afetada pela falha em estabelecer escalas corretas para as variáveis explanatórias no preditor linear, escolha errada da função de ligação e pontos atípicos.

## 5.7 Verificação da adequação das escalas das variáveis explanatórias

O gráfico de resíduos parciais é uma ferramenta importante para saber se um termo  $\beta x$  no preditor linear pode ser melhor expresso como  $\beta h(x; \lambda)$  para alguma função monótona  $h(\cdot; \lambda)$ . Em MLG, o vetor de resíduos parciais (ou resíduos + componente) é dado por

$$\tilde{r} = \hat{\mathbf{z}} - \hat{\boldsymbol{\eta}} + \hat{\gamma} \mathbf{x},$$

sendo  $\hat{\mathbf{z}}$  a variável dependente ajustada estimada,  $\hat{\boldsymbol{\eta}}$  o preditor linear estimado e  $\hat{\gamma}$  a estimativa do parâmetro para a variável explanatória  $\mathbf{x}$ .

O gráfico de  $\tilde{r}$  *versus*  $\mathbf{x}$  conduz a um método informal. Se a escala de  $\mathbf{x}$  é satisfatória, o gráfico deve ser, aproximadamente, linear. Se não, sua forma pode sugerir um modelo alternativo. Poderão, entretanto, ocorrer distorções se as escalas das outras variáveis explanatórias estiverem erradas e, então, pode ser necessário analisar gráficos de resíduos parciais para diversos  $x$ 's.

Um método formal consiste em colocar  $x$  em uma família  $h(\cdot; \lambda)$  indexada por  $\lambda$ ; calcular, então, o desvio para um conjunto de valores de  $\lambda$  e determinar  $\hat{\lambda}$  como aquele valor que conduz a um desvio mínimo (método da verossimilhança perfilada). O ajuste para  $\hat{\lambda}$  será, então, comparado com o ajuste para a escolha inicial  $\lambda_0$  que, em geral, é 1. Esse procedimento pode ser usado para vários  $x$ 's simultaneamente e é, particularmente, útil quando se têm as mesmas dimensões físicas, tal que seja necessária uma transformação comum. A família mais comum de transformações é a família de Box e Cox (1964) dada por  $h(x; \lambda) = (x^\lambda - 1)/\lambda$ , se  $\lambda \neq 0$ , e  $h(x; \lambda) = \log x$ , se  $\lambda = 0$ .

Um método informal para o estudo de uma única variável explanatória implica na forma  $\mathbf{u}(\lambda_0) = \left. \frac{dh(\boldsymbol{\mu}, \lambda)}{d\lambda} \right|_{\lambda=\lambda_0}$  que é, então, usada como variável adicional para o teste de adequação da escala usada para a variável explanatória de interesse. Pode-se, então, fazer o gráfico de resíduos parciais, como já discutido na Seção 5.4.2, item e). Essa mesma variável  $u$  construída pode ser usada como uma variável adicional no modelo para o teste da hipótese  $H_0 : \lambda = \lambda_0$  (o que equivale ao teste de  $H_0 : \gamma = 0$ ) que, se não rejeitada, indicará a adequação da escala escolhida para a variável explanatória de interesse.

### Exemplo 5.3: Transformação para a variável dependente

Seja a família de transformações de Box-Cox normalizada

$$\mathbf{z}(\lambda) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{cases} \frac{\mathbf{y}^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} + \boldsymbol{\epsilon} & \lambda \neq 0 \\ \dot{y} \log \mathbf{y} + \boldsymbol{\epsilon} & \lambda = 0, \end{cases}$$

sendo  $\dot{y}$  a média geométrica das observações. A expansão de  $\mathbf{z}(\lambda)$  em série de Taylor em relação a  $\lambda_0$ , suposto conhecido, é

$$\mathbf{z}(\lambda) \approx \mathbf{z}(\lambda_0) + (\lambda - \lambda_0)\mathbf{u}(\lambda_0),$$

sendo  $\mathbf{u}(\lambda_0) = \left. \frac{d\mathbf{z}(\lambda)}{d\lambda} \right|_{\lambda=\lambda_0}$ . Então, o modelo linear aproximado é

$$\mathbf{z}(\lambda_0) = \mathbf{z}(\lambda) - (\lambda - \lambda_0)\mathbf{u}(\lambda_0) + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{u} + \boldsymbol{\epsilon}.$$

Mas  $\mathbf{z}(\lambda) = \frac{\mathbf{y}^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} + \boldsymbol{\epsilon}$  e, portanto,

$$\mathbf{u}(\lambda) = \frac{d\mathbf{z}(\lambda)}{d\lambda} = \frac{\mathbf{y}^\lambda \log \mathbf{y} - (\mathbf{y}^\lambda - 1)(\lambda^{-1} + \log \dot{y})}{\lambda \dot{y}^{\lambda-1}}.$$

O interesse, em geral, está em testar alguns valores de  $\lambda$ , tais como  $\lambda_0 = 1$  (sem transformação) e  $\lambda_0 = 0$  (transformação logarítmica). Desde que são necessários apenas os resíduos de  $\mathbf{u}(\lambda)$ , então, constantes podem ser ignoradas se  $\boldsymbol{\beta}$  contém uma constante. Então,

$$\mathbf{u}(1) = \mathbf{y} \left[ \log \left( \frac{\mathbf{y}}{\dot{\mathbf{y}}} \right) - 1 \right], \text{ variável construída para testar se } \lambda_0 = 1$$

e

$$\mathbf{u}(0) = \dot{\mathbf{y}} \log \mathbf{y} \left[ \frac{\log \mathbf{y}}{2} - \log \dot{\mathbf{y}} \right], \text{ variável construída para testar se } \lambda_0 = 0.$$

Como  $-\gamma = \lambda - \lambda_0$ , tem-se que uma estimativa para  $\lambda$  pode ser obtida como  $\hat{\lambda} = \lambda_0 - \hat{\gamma}$ . Usa-se, em geral, um valor para  $\lambda$  próximo de  $\hat{\lambda}$  que tenha uma interpretação prática.

#### Exemplo 5.4: Transformação para as variáveis explanatórias

Se em lugar de transformar  $y$  houver necessidade de transformar  $x_k$ , tem-se que o componente sistemático mais amplo é dado por

$$E(\mathbf{Y}) = \beta_0 + \sum_{j \neq k} \beta_j \mathbf{x}_j + \beta_k \mathbf{x}_k^\lambda.$$

A expansão de  $\mathbf{z}(\lambda)$  em série de Taylor em relação a  $\lambda_0$ , suposto conhecido, é

$$\mathbf{z}(\lambda) \approx \mathbf{z}(\lambda_0) + (\lambda - \lambda_0) \frac{d\mathbf{z}(\lambda)}{d\lambda} \Big|_{\lambda=\lambda_0}.$$

Então,

$$\mathbf{z}(\lambda) \approx \beta_0 + \sum_{j \neq k} \beta_j \mathbf{x}_j + \beta_k \mathbf{x}_k^{\lambda_0} + \beta_k (\lambda - \lambda_0) \mathbf{x}_k^{\lambda_0} \log \mathbf{x}_k = \beta_0 + \sum_{j \neq k} \beta_j \mathbf{x}_j + \beta_k \mathbf{x}_k^{\lambda_0} + \gamma \mathbf{u}(\lambda_0),$$

pois  $\frac{d\mathbf{z}(\lambda)}{d\lambda} = \beta_k \mathbf{x}_k^\lambda \log x_k$ . Portanto, testar  $\lambda = \lambda_0$  é equivalente a testar  $\gamma = 0$  para a regressão com a variável construída  $\mathbf{u}_x(\lambda_0) = \mathbf{x}_k^{\lambda_0} \log \mathbf{x}_k$  com  $\mathbf{x}_k^{\lambda_0}$  incluída no modelo.

Para  $\lambda_0 = 1$ , tem-se

$$E(\mathbf{Y}) = \beta_0 + \sum_{j \neq k} \beta_j \mathbf{x}_j + \beta_k \mathbf{x}_k + \beta_k (\lambda - 1) \mathbf{x}_k \log \mathbf{x}_k = \mathbf{X}\boldsymbol{\beta} + \gamma \mathbf{u}_x(\lambda_0 = 1),$$

sendo  $\mathbf{u}_x = \mathbf{x}_k \log \mathbf{x}_k$  e  $\gamma = \beta_k (\lambda - 1)$ . Portanto, faz-se a regressão de  $\mathbf{Y}$  em relação a todos os  $\mathbf{x}_j$ ,  $j = 1, \dots, p-1$ , e a  $\mathbf{u}_x = \mathbf{x}_k \log \mathbf{x}_k$ . Rejeita-se  $H_0 : \lambda = 1$  se  $t = \hat{\gamma}/se(\hat{\gamma}) > t_{n-p-1, \gamma/2}$  e, nesse caso, mostrando que a escala das observações não está adequada. A contribuição de observações individuais pode ser examinada

através do gráfico da variável adicionada.

**Exemplo 5.5: Transformação simultânea para as variáveis resposta e explanatórias**

Para a transformação simultânea das variável resposta e  $p - 1$  variáveis explanatórias (exceto a constante  $1^\lambda = 1$ ), o modelo para um vetor de  $p$  parâmetros  $\boldsymbol{\lambda} = (\lambda_1 \dots \lambda_p)^T$  é

$$\mathbf{z}(\lambda_p) = \beta_0 + \sum_{j=1}^{p-1} \beta_j \mathbf{x}_j^{\lambda_j} + \boldsymbol{\epsilon}. \quad (5.8)$$

Em (5.8) cada variável, incluindo a variável resposta, pode ter um parâmetro de transformação diferente. De forma semelhante aos Exemplos 5.3 e 5.4, a expansão de Taylor desse modelo em torno de um  $\lambda_0$  comum, suposto conhecido, é

$$\mathbf{z}(\lambda_0) = \beta_0 - (\lambda_p - \lambda_0) \mathbf{u}(\lambda_0) + \sum_{j=1}^{p-1} \beta_j \mathbf{x}_j^{\lambda_0} + \sum_{j=1}^{p-1} (\lambda_j - \lambda_0) \beta_j \mathbf{x}_j^{\lambda_0} \log \mathbf{x}_j + \boldsymbol{\epsilon}.$$

Para o caso de  $\lambda_0 = 1$ , tem-se

$$\mathbf{z}(\lambda_0) = \beta_0 + \sum_{j=1}^{p-1} \beta_j \mathbf{x}_j + \gamma \left[ \sum_{j=1}^{p-1} \beta_j \mathbf{x}_j \log \mathbf{x}_j - \mathbf{u}(1) \right] + \boldsymbol{\epsilon},$$

sendo  $\gamma = \lambda - 1$  e definindo a variável construída por

$$\mathbf{u}_{xy}(1) = \beta_0 + \sum_{j=1}^{p-1} \hat{\beta}_j \mathbf{x}_j \log \mathbf{x}_j - \mathbf{y} \left[ \log \left( \frac{\mathbf{y}}{\hat{\mathbf{y}}} \right) - 1 \right]$$

usando-se no lugar de  $\beta_j$ ,  $\hat{\beta}_j$ , as estimativas de mínimos quadrados do modelo sem transformação. Rejeita-se  $H_0 : \boldsymbol{\lambda} = 1$  se  $t = \hat{\gamma}/se(\hat{\gamma}) > t_{n-p-1, \gamma/2}$  e, nesse caso, mostrando que a escala das observações e das variáveis explanatórias não está adequada. A contribuição de observações individuais pode ser examinada através do gráfico da variável adicionada.



## 5.8 Verificação de anomalias no componente sistemático através da análise dos resíduos

Considera-se um MLG com distribuição na família (1.5) e componente sistemático  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ . As possíveis anomalias no componente aleatório do modelo podem ser descobertas pelos gráficos i'), ii') e iii') descritos na Seção 5.4.1, desde que os resíduos sejam definidos apropriadamente. Nesta seção, apresenta-se uma técnica geral para verificar anomalias no componente sistemático do modelo definido pelas equações (2.5) e (2.6).

Admite-se que o componente sistemático correto contém uma variável explanatória  $\mathbf{z}$  adicional (Seção 4.9) e um parâmetro escalar  $\gamma$ , isto é,

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + h(\mathbf{z}; \gamma), \quad (5.9)$$

em que  $h(\mathbf{z}; \gamma)$  pode representar:

- a) um termo adicional em uma ou mais variáveis explanatórias originais, por exemplo:  $h(\mathbf{z}; \gamma) = \gamma x_j^2$  ou  $h(\mathbf{z}; \gamma) = \gamma x_j x_k$ ;
- b) uma contribuição linear ou não-linear de alguma variável explanatória omitida, por exemplo:  $g(\mathbf{z}; \gamma) = \gamma \mathbf{z}$  ou  $g(\mathbf{z}; \gamma) = \mathbf{z}^\gamma$ .

O objetivo é definir resíduos modificados  $\tilde{\mathbf{R}}$  para o modelo ajustado  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$  tal que  $E(\tilde{\mathbf{R}}) = h(\mathbf{z}; \gamma)$ . Se isso acontecer, um gráfico de  $\tilde{\mathbf{R}}$  versus  $\mathbf{z}$ , desprezando a variação aleatória, exibirá a função  $h(\mathbf{z}; \gamma)$ .

Para fixar idéias considere o modelo normal linear e os resíduos usuais:  $\mathbf{R} = \mathbf{y} - \hat{\boldsymbol{\mu}} = [\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T] \mathbf{y} = (\mathbf{I} - \mathbf{H}) \mathbf{y}$ . Supondo que o componente sistemático correto é (5.9), tem-se  $\mathbf{R} = (\mathbf{I} - \mathbf{H})[\mathbf{X}\boldsymbol{\beta} + h(\mathbf{z}; \gamma) + \boldsymbol{\varepsilon}]$ , em que  $\boldsymbol{\varepsilon}$  é um ruído branco. Como  $\mathbf{X}$  é ortogonal a  $\mathbf{I} - \mathbf{H}$ , tem-se  $\mathbf{R} = (\mathbf{I} - \mathbf{H})h(\mathbf{z}; \gamma) + \boldsymbol{\varepsilon}$  e, portanto,  $E(\mathbf{R}) = (\mathbf{I} - \mathbf{H})h(\mathbf{z}; \gamma)$ . Assim, um gráfico de  $\mathbf{R}$  versus  $\mathbf{z}$  não apresentará nenhuma semelhança com  $h(\mathbf{z}; \gamma)$ . Entretanto, se  $h(\mathbf{z}; \gamma)$  for aproximadamente linear, um gráfico de  $\mathbf{R}$  versus  $(\mathbf{I} - \mathbf{H})\mathbf{z}$  poderá ser usado. A declividade da reta de mínimos quadrados

ajustada aos pontos desse gráfico proporcionará uma estimativa de  $\gamma$  no modelo (5.9). Se a declividade for próxima de zero, o modelo  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$  poderá ser aceito ao invés de (5.9).

Para o modelo normal linear, supondo  $h(\mathbf{z}; \gamma)$  aproximadamente linear, Larsen e McCleary (1972) definem *resíduos parciais* por

$$\tilde{\mathbf{R}} = \mathbf{y} - \hat{\boldsymbol{\mu}} + \hat{\gamma}\mathbf{H}\mathbf{z} = (\mathbf{I} - \mathbf{H})\mathbf{y} + \hat{\gamma}\mathbf{H}\mathbf{z}, \quad (5.10)$$

em que  $\hat{\gamma}$  é a estimativa de mínimos quadrados de  $\gamma$  baseada na regressão de  $\mathbf{y} - \hat{\boldsymbol{\mu}}$  sobre a matriz  $(\mathbf{I} - \mathbf{H})\mathbf{z}$ , isto é,  $\hat{\gamma} = [\mathbf{z}^T(\mathbf{I} - \mathbf{H})\mathbf{z}]^{-1}\mathbf{z}^T(\mathbf{I} - \mathbf{H})(\mathbf{y} - \hat{\boldsymbol{\mu}})$ , com  $\mathbf{z} = (z_1, \dots, z_n)^T$ .

Pode-se demonstrar que os resíduos parciais (5.10) podem ser expressos como combinações lineares dos resíduos  $\mathbf{y} - \hat{\boldsymbol{\mu}}$  e, também, como combinações lineares das observações  $\mathbf{y}$ .

Ainda no modelo normal linear, a noção de resíduos parciais pode ser estendida para determinar se variáveis explanatórias, com contribuições não-lineares, estão omissas no componente sistemático do modelo. Suponha, agora, que  $\boldsymbol{\gamma}$  seja um vetor de parâmetros. Isso é possível, desde que a função  $h(\mathbf{z}; \boldsymbol{\gamma})$  possa ser aproximada por um polinômio de grau baixo, isto é,  $h(\mathbf{z}; \boldsymbol{\gamma}) \approx \mathbf{T}\boldsymbol{\gamma}$ , em que  $\mathbf{T} = \mathbf{T}(\mathbf{z}) = (\mathbf{z}, \mathbf{z}^{(2)}, \mathbf{z}^{(3)} \dots)$  com  $\mathbf{z}^{(i)} = (z_1^i, \dots, z_n^i)^T$ .

Com essa aproximação, definem-se os *resíduos aumentados* de Andrews e Pregibon (1978), por uma expressão análoga a (5.10),

$$\tilde{\mathbf{R}} = \mathbf{y} - \hat{\boldsymbol{\mu}} + \mathbf{H}\mathbf{T}\hat{\boldsymbol{\gamma}} = (\mathbf{I} - \mathbf{H})\mathbf{y} + \mathbf{H}\mathbf{T}\hat{\boldsymbol{\gamma}}, \quad (5.11)$$

em que  $\hat{\boldsymbol{\gamma}}$  é a estimativa de mínimos quadrados de  $\boldsymbol{\gamma}$  na regressão linear de  $\mathbf{y} - \hat{\boldsymbol{\mu}}$  sobre  $(\mathbf{I} - \mathbf{H})\mathbf{T}$ , isto é,  $\hat{\boldsymbol{\gamma}} = [\mathbf{T}^T(\mathbf{I} - \mathbf{H})\mathbf{T}]^{-1}\mathbf{T}^T(\mathbf{I} - \mathbf{H})(\mathbf{y} - \hat{\boldsymbol{\mu}})$ .

Tem-se  $E(\tilde{\mathbf{R}}) = \mathbf{T}\boldsymbol{\gamma} \approx h(\mathbf{z}; \boldsymbol{\gamma})$  e, portanto, exceto por variações aleatórias, um gráfico de  $\tilde{\mathbf{R}}$  versus  $\mathbf{z}$  poderá exibir a forma da função  $h(\mathbf{z}; \boldsymbol{\gamma})$ .

Para os MLG os resíduos aumentados podem ser definidos a partir de resíduos medidos na escala linear

$$\mathbf{R} = \hat{\mathbf{z}} - \hat{\boldsymbol{\eta}} = (\mathbf{I} - \widehat{\mathbf{Z}\mathbf{W}})\hat{\mathbf{z}}. \quad (5.12)$$

Essa expressão já foi vista na Seção 5.4.3. Aqui, estima-se  $\boldsymbol{\gamma}$  ajustando o modelo aumentado  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{T}\boldsymbol{\gamma}$  aos dados. Isso determinará opções de aperfeiçoamento da estrutura linear do modelo. O ajuste de polinômios de graus elevados é, numericamente, bastante instável, sendo melhor considerar no máximo  $\mathbf{T} = (\mathbf{z}, \mathbf{z}^{(2)}, \mathbf{z}^{(3)})$ .

Tem-se  $\mathbf{R} = (\mathbf{I} - \widehat{\mathbf{Z}\mathbf{W}})(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{T}\hat{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}) = (\mathbf{I} - \widehat{\mathbf{Z}\mathbf{W}})(\mathbf{T}\hat{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon})$  e, portanto, os resíduos aumentados nos MLG são expressos por

$$\tilde{\mathbf{R}} = \mathbf{R} + \widehat{\mathbf{Z}\mathbf{W}}\mathbf{T}\hat{\boldsymbol{\gamma}} \quad (5.13)$$

e têm valores esperados próximos de  $h(\mathbf{z}; \boldsymbol{\gamma})$ . Em (5.13) as estimativas de  $\mathbf{Z}$  e  $\mathbf{W}$  são segundo o modelo reduzido  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ .

A expressão (5.11) é um caso especial de (5.13) quando  $\mathbf{W}$  é igual à matriz identidade. Um gráfico de  $\tilde{\mathbf{R}}$  versus  $\mathbf{z}$  poderá indicar se essa variável explanatória deve estar incluída no modelo e, se isso acontecer, poderá ainda sugerir a forma de inclusão. Não se devem comparar os resíduos aumentados em (5.13) com os resíduos ordinários  $\mathbf{R}$ , pois os primeiros são baseados no ajuste do modelo aumentado.

A análise gráfica dos resíduos aumentados pode ser bastante útil nos estágios preliminares de seleção de variáveis explanatórias, quando se têm muitas variáveis explanatórias que devem ser consideradas. A formação do componente sistemático pode ser feita, passo a passo, com a introdução de uma única variável explanatória, a cada passo, pelo método descrito.

Para determinar a contribuição de uma variável explanatória  $\mathbf{x}_i = (x_{i1}, \dots, x_{in})^T$  da própria matrix  $\mathbf{X}$  no ajuste do modelo reduzido  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$  aos dados, pode-se trabalhar com os resíduos parciais generalizados

$$v_i = \hat{\mathbf{z}} - \hat{\boldsymbol{\eta}} + \hat{\boldsymbol{\beta}}_i x_i. \quad (5.14)$$

Os resíduos (5.14), vistos na Seção 5.7, são muito mais simples de serem computados do que os resíduos aumentados definidos em (5.13).

## 5.9 Exercícios

1. Comparar os resíduos de Anscombe, Pearson e como raiz quadrada do componente do desvio, para o modelo de Poisson. Como sugestão supor  $\hat{\mu} = cy$  e variar  $c$ , por exemplo, 0(0.2)2(0.5)10. Fazer o mesmo para os modelos binomial, gama e normal inverso.

2. Definir os resíduos de Anscombe, Pearson e como raiz quadrada do componente do desvio para o modelo binomial negativo, fazendo uma comparação entre os três resíduos.

3. Seja um MLG com estrutura linear  $\eta_i = \alpha + \beta x_i + x_i^\gamma$  e função de ligação  $g(\cdot)$  conhecida.

(a) Formular, por meio da função desvio, critérios para os seguintes testes:  $H_1 : \gamma = \gamma^{(0)}$  versus  $H'_1 : \gamma \neq \gamma^{(0)}$ ;  $H_2 : \beta = \beta^{(0)}$ ,  $\gamma = \gamma^{(0)}$  versus  $H'_2 : \beta \neq \beta^{(0)}$ ,  $\gamma = \gamma^{(0)}$  e versus  $H''_2 : \beta \neq \beta^{(0)}$ ,  $\gamma \neq \gamma^{(0)}$ ;  $H_3 : \beta = \beta^{(0)}$  versus  $H_3 : \beta \neq \beta^{(0)}$ ;

(b) como obter um intervalo de confiança para  $\gamma$  usando a função desvio?

(c) se a função de ligação dependesse de um parâmetro  $\lambda$  desconhecido, como determinar critérios para os testes citados?

4. Os dados da Tabela 5.3 referem-se a medidas de diâmetro a 4,5 pés acima do solo ( $D$ , polegadas) e altura ( $H$ , pés) de 21 cerejeiras (“black cherry”) em pé e de volume ( $V$ , pés cúbicos) de árvores derrubadas. O objetivo desse tipo de experimento é verificar de que forma essas variáveis estão relacionadas para, através de medidas nas árvores em pé, poder prever o volume de madeira em uma área de floresta (*Allegheny National Forest*). Pede-se:

- fazer os gráficos de variáveis adicionadas para  $H$  e  $D$ ;
- fazer os gráficos de resíduos parciais para  $H$  e  $D$ ;

Tabela 5.3: Medidas de diâmetro a 4,5 pés acima do solo ( $D$ , polegadas) e altura ( $H$ , pés) de 21 cerejeiras (“black cherry”) em pé e de volume ( $V$ , pés cúbicos) de árvores derrubadas

Amostra	$D$	$H$	$V$	Amostra	$D$	$H$	$V$
1	8,3	70	10,3	17	12,9	85	33,8
2	8,6	65	10,3	18	13,3	86	27,4
3	8,8	63	10,2	19	13,7	71	25,7
4	10,5	72	16,4	20	13,8	64	24,9
5	10,7	81	18,8	21	14,0	78	34,5
6	10,8	83	19,7	22	14,2	80	31,7
7	11,0	66	15,6	23	14,5	74	36,3
8	11,0	75	18,2	24	16,0	72	38,3
9	11,1	80	22,6	25	16,3	77	42,6
10	11,2	75	19,9	26	17,3	81	55,4
11	11,3	79	24,2	27	17,5	82	55,7
12	11,4	76	21,0	28	17,9	80	58,3
13	11,4	76	21,4	29	18,0	80	51,5
14	11,7	69	21,3	30	18,0	80	51,0
15	12,0	75	19,1	31	20,6	87	77,0
16	12,9	74	22,2				

**Fonte:** Ryan et al. (1976, p. 329).

c) fazer as transformações  $LV = \log(V)$ ,  $LH = \log(H)$  e  $LD = \log(D)$  e repetir os gráficos dos itens (a) e (b);

d) verificar se existem pontos discrepantes em ambas as escalas;

e) usando

$$\mathbf{u}(1) = \sum_{j=2}^p \hat{\beta}_j \mathbf{x}_j \log \mathbf{x}_j - \mathbf{y} \left[ \log \left( \frac{\mathbf{y}}{\dot{\mathbf{y}}} \right) - 1 \right],$$

obtido como no Exemplo 5.5 da Seção 5.7, como variável adicionada, verifique que há necessidade da transformação simultânea de  $V$ ,  $H$  e  $D$ .

**5.** Os dados da Tabela 5.4 referem-se à mortalidade de escaravelhos após 5 h de exposição a diferentes doses de bissulfeto de carbono ( $CS^2$ ). Pede-se:

Tabela 5.4: Número de insetos mortos ( $y_i$ ) de ( $m_i$ ) insetos após 5 h de exposição a diferentes doses de  $CS^2$

$\log(\text{Dose}) (d_i)$	$(m_i)$	$y_i$
1,6907	59	6
1,7242	60	13
1,7552	62	18
1,7842	56	28
1,8113	63	52
1,8369	59	53
1,8610	62	61
1,8839	60	60

a) ajuste o modelo logístico linear e faça o teste para a função de ligação;

b) ajuste o modelo complemento log-log e faça o teste para a função de ligação;

c) faça o gráfico da variável adicionada para os itens a) e b);

d) verifique se há necessidade de transformação para a variável dose usando

o gráfico de resíduos parciais.

6. Os dados da Tabela 5.5 são provenientes de um experimento em delineamento casualizado em blocos em que foram usadas como tratamentos 8 doses de um inseticida fosforado e foram contadas quantas ( $y$ ) cenouras estavam danificadas de totais de  $m$  cenouras.

Tabela 5.5: Número de cenouras danificadas ( $y_i$ ) de ( $m_i$ ) cenouras

log(Dose)	Bloco I		Bloco II		Bloco III	
$d_i$	$m_i$	$y_i$	$m_i$	$y_i$	$m_i$	$y_i$
1,52	10	35	17	38	10	34
1,64	16	42	10	40	10	38
1,76	8	50	8	33	5	36
1,88	6	42	8	39	3	35
2,00	9	35	5	47	2	49
2,12	9	42	17	42	1	40
2,24	1	32	6	35	3	22
2,36	2	28	4	35	2	31

**Fonte:** Phelps (1982).

- a) ajuste o modelo logístico linear e faça o teste para a função de ligação;
- b) ajuste o modelo complemento log-log e faça o teste para a função de ligação;
- c) faça o gráfico da variável adicionada para os itens (a) e (b);
- d) usando a família de funções de ligação de Aranda-Ordaz, obtenha a variável construída e estime  $\lambda$ ;
- e) ajuste o modelo logístico com preditor linear quadrático e faça o teste para a função de ligação.

7. Considere a família de funções de ligação dada por (5.7). Mostre que a variável

construída para o teste da hipótese  $H_0 : \lambda = 0$  é dada por  $\mathbf{u}(\lambda_0) = \frac{dh(\boldsymbol{\mu}, \lambda)}{d\lambda} \Big|_{\lambda=0} = -\frac{\log \hat{\boldsymbol{\mu}} \otimes \log \hat{\boldsymbol{\mu}}}{2} = -\frac{\hat{\boldsymbol{\eta}} \otimes \hat{\boldsymbol{\eta}}}{2}$  (Atkinson, 1985, p. 238) em que  $\odot$  representa o produto termo a termo.

8. Seja  $Y_i \sim B(m_i, \mu_i)$  com a notação usual  $\boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ , etc.

Demonstrar que os resíduos podem ser definidos por

$$\frac{\left[ G\left(\frac{Y_i}{m_i}\right) - G'(\hat{\mu}_i) \right]}{G'(\hat{\mu}_i)} \left[ \frac{\hat{\mu}_i(1 - \hat{\mu}_i)}{m_i} \right]^{1/2}.$$

Quais as vantagens das escolhas  $G(\mu) = \mu$ ,  $G(\mu) = \log[\mu/(1 - \mu)]$  e  $g(\mu) = \int_0^\mu x^{-1/3}(1 - x)^{-1/3}dx$ ?

9. No modelo normal linear com estrutura para a média dada por  $\boldsymbol{\mu} = E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + g(\mathbf{z}; \boldsymbol{\gamma})$ , sendo a função  $g(\mathbf{z}; \boldsymbol{\gamma})$  aproximadamente linear, demonstrar que os resíduos parciais  $\hat{\mathbf{R}} = (\mathbf{I} - \mathbf{H})\mathbf{y} + \mathbf{H}\mathbf{z}\hat{\boldsymbol{\gamma}}$ , em que  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$  é a matriz de projeção, podem ser expressos como combinações lineares dos resíduos ordinários  $\mathbf{y} - \hat{\boldsymbol{\mu}}$  e, também, como combinações lineares dos dados  $\mathbf{y}$ .

10. Demonstrar as fórmulas aproximadas (8.15) e (8.16) para se fazer o diagnóstico global de influência de uma única observação sobre o ajuste do MLG.

11. Demonstrar as expressões (8.17) e (8.19) para se fazer o diagnóstico local da influência de um conjunto de pontos sobre as estimativas dos parâmetros lineares do modelo.

12. Demonstrar as expressões (8.20) e (8.22) para se diagnosticar a influência de um conjunto de pontos sobre o ajuste global do modelo.

13. Quais as vantagens e desvantagens das estatísticas (8.21) e (8.22) ((8.15) e (8.16))?

14. Os resíduos  $r_i^{P'}$  definidos em (5.5) são, também, denominados resíduos de Stu-



dent (W.S. Gosset). Calcular expressões para  $a_0^{(1)}$ ,  $b_i$  e  $c_i$  em função desses resíduos.

**15.** Seja um modelo normal, ou gama ou normal inverso com componente usual  $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$  e admita-se que o parâmetro  $\phi$  seja constante para todas as observações, embora desconhecido. Determinar, através da função desvio, critérios para os seguintes testes:

(a)  $\phi = \phi^{(0)}$  versus  $\phi \neq \phi^{(0)}$ ;

(b)  $\beta = \beta^{(0)}$  versus  $\beta \neq \beta^{(0)}$  (Cordeiro, 1986).



# Capítulo 6

## Aplicações a Dados Contínuos

Neste Capítulo, serão apresentadas análises para vários conjuntos de dados contínuos.

**(Precisa ser melhorada)**

Os programas para as análises foram feitos em R e encontram-se no Apêndice.

### 6.1 Dados de volume de árvores

Os dados da Tabela 5.3 referem-se a medidas de diâmetro a 4,5 pés acima do solo ( $D$ , polegadas) e altura ( $H$ , pés) de 21 cerejeiras (“black cherry”) em pé e de volume ( $V$ , pés cúbicos) de árvores derrubadas (Ryan et al., 1976). O objetivo desse tipo de experimento é verificar de que forma essas variáveis estão relacionadas para, através de medidas nas árvores em pé, poder prever o volume de madeira em uma área de floresta (*Allegheny National Forest*).

A Figura 6.1 mostra os gráficos de dispersão das variáveis duas a duas para os dados observados sem transformação e com transformação logarítmica. Pode-se verificar que existe uma relação mais forte entre volume e diâmetro à altura do peito, do que entre volume e altura. Além disso, que a variável altura tem variabilidade maior do que a variável diâmetro à altura do peito.

Como um primeiro modelo ( $M_1$ ), supõe-se que (i) a variável resposta  $Y = \mu + \varepsilon_1$  em que  $Y = V$  e  $\varepsilon_1 \sim N(0, \sigma_1^2)$  e, portanto,  $Y \sim N(\mu, \sigma_1^2)$ , (ii) a função de

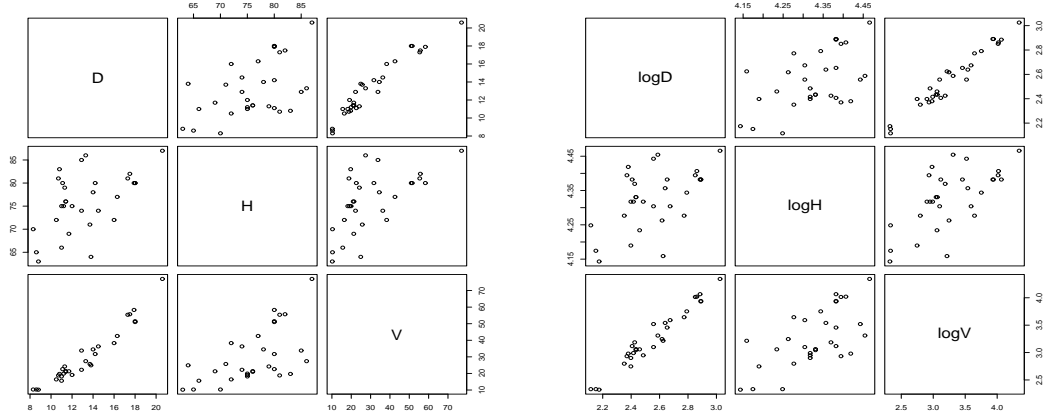


Figura 6.1: Gráfico de dispersão - valores observados e transformados na escala logarítmica

ligação é a identidade, isto é,  $\eta = \mu$  e (iii) o preditor linear é dado por (6.1)

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (6.1)$$

em que  $x_1 = D$  e  $x_2 = H$ .

Um segundo modelo ( $M_2$ ) baseia-se no fato que o volume é proporcional ao produto de diâmetro à altura do peito pela altura, isto é,  $V \approx \gamma_0 D^{\beta_1} H^{\beta_2}$  e, portanto,  $\log(V) \approx \beta_0 + \beta_1 \log(D) + \beta_2 \log(H)$ . Então, pode-se supor que (i) a variável resposta transformada  $Y = \mu + \varepsilon_2$ , em que  $Y = \log(V)$  e  $\varepsilon_2 \sim N(0, \sigma_2^2)$  e, portanto,  $Y \sim N(\mu, \sigma_2^2)$ , (ii) a função de ligação é a identidade, isto é,  $\eta = \mu$  e (iii) o preditor linear é dado por (6.1), em que  $x_1 = \log(D)$  e  $x_2 = \log(H)$ .

Como um terceiro modelo ( $M_3$ ), supõe-se que (i) a variável resposta  $Y = \mu + \varepsilon_3$  em que  $Y = V$ ,  $\mu = \gamma_0 D^{\beta_1} H^{\beta_2}$  e  $\varepsilon_3 \sim N(0, \sigma_1^3)$  e, portanto,  $Y \sim N(\mu, \sigma_3^2)$ , (ii) a função de ligação é a logarítmica, isto é,  $\eta = \log(\mu)$  e (iii) o preditor linear é dado por (6.1) em que  $x_1 = \log(D)$  e  $x_2 = \log(H)$ .

A Tabelas 6.1 e 6.2 mostram os resultados obtidos, considerando-se diversos submodelos para o preditor linear, para a análise dos dados sem transformação ( $M_1$ ) e com transformação logarítmica ( $M_2$ ). Verifica-se que existem evidências, ao nível de 1% de significância, que os efeitos tanto de diâmetro à altura do peito como de altura são significativos, sendo que o efeito de diâmetro à altura do peito é maior do

que o de altura, tanto para o caso de dados não transformados como para transformados. É importante, lembrar, também, que o teste para o modelo com ambas as variáveis (regressão parcial) simultaneamente tem um nível de significância conjunto, enquanto que na análise seqüencial não se sabe o nível conjunto de significância dos testes. Verifica-se que existem evidências que ambas as variáveis explanatórias altura e diâmetro são necessárias para explicar o volume e que o melhor ajuste é obtido com os dados transformados. Testes  $t$  (equivalentes aos testes  $F$ ) e intervalos de confiança para os parâmetros e intervalos de previsão para  $Y$  podem ser obtidos. Há necessidade, porém, de um estudo mais detalhado, fazendo-se análise de resíduos e diagnósticos, para a escolha do modelo final.

Conforme, pode-se ver na Figura ?? há indicação de que o modelo 1 não se ajusta bem às observações.

Tabela 6.1: Análise de variância, teste  $F$  e estimativas - Dados sem transformação ( $M_1$ ).

$\eta = \beta_0 + \beta_1 D$				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	7.581,8	7.581,8	419,4 **
Resíduo	29	524,3	18,1	
Total	30	8.106,1		
$\hat{V} = -36,94 + 5,066D$ $R^2 = 0,935$ $\bar{R}^2 = 0,933$ $s(\hat{\beta}_0) = 3,36$ e $s(\hat{\beta}_1) = 0,247$				
$\eta = \beta_0 + \beta_2 H$				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	2.901,2	2.901,2	16,2 **
Resíduo	29	5.204,9	179,5	
Total	30	8.106,1		
$\hat{V} = -87,12 + 1,543H$ $R^2 = 0,358$ $\bar{R}^2 = 0,336$ $s(\hat{\beta}_0) = 29,27$ e $s(\hat{\beta}_2) = 0,384$				
$\eta = \beta_0 + \beta_1 D + \beta_2 H$ - Parcial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP e Altura	2	7.684,4	3.842,2	255,0 **
Resíduo	28	421,9	15,1	
Total	30	8.106,1		
$\hat{V} = -57,99 + 4,708D + 0,339H$ $R^2 = 0,948$ $\bar{R}^2 = 0,944$ $s(\hat{\beta}_0) = 8,64$ , $s(\hat{\beta}_1) = 0,264$ e $s(\hat{\beta}_2) = 0,130$				
$\eta = \beta_0 + \beta_1 D + \beta_2 H$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	7.581,8	7.581,8	503,1 **
Altura DAP	1	102,4	102,4	6,8*
Resíduo	28	421,9	15,1	
Total	30	8.106,1		
$\eta = \beta_0 + \beta_1 D + \beta_2 H$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	2.901,2	2.901,2	192,5 **
DAP Altura	1	4.783,0	4.783,0	317,4 **
Resíduo	28	421,9	15,1	
Total	30	8.106,1		
$F_{1,29;0,05} = 4,18$ , $F_{2,28;0,05} = 3,34$ e $F_{1,28;0,05} = 4,20$ $F_{1,29;0,01} = 7,60$ , $F_{2,28;0,01} = 5,45$ e $F_{1,28;0,01} = 7,64$				

Tabela 6.2: Análise de variância, teste  $F$  e estimativas - Dados transformados ( $M_2$ ).

$\eta = \beta_0 + \beta_1 \log(D)$				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	7,9254	7,9254	599,7 **
Resíduo	29	0,3832	0,0132	
Total	30	8,3087		
$\widehat{\log(V)} = -2,353 + 2,2 \log(D) \quad R^2 = 0,954 \quad \bar{R}^2 = 0,952$ $s(\hat{\beta}_0) = 0,231$ e $s(\hat{\beta}_1) = 0,089$				
$\eta = \beta_0 + \beta_2 \log(H)$				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	3,496	3,496	21,06 **
Resíduo	29	4,8130	0,166	
Total	30	8,3087		
$\widehat{\log(V)} = -13,96 + 3,982 \log(H) \quad R^2 = 0,421 \quad \bar{R}^2 = 0,401$ $s(\hat{\beta}_0) = 3,76$ e $s(\hat{\beta}_2) = 0,868$				
$\eta = \beta_0 + \beta_1 \log(D) + \beta_2 \log(H)$ - Parcial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP e Altura	2	8,1228	4,0614	615,36 **
Resíduo	28	0,1855	0,0066	
Total	30	8,3087		
$\widehat{\log(V)} = -6,632 + 1,983 \log(D) + 1,117 \log(H) \quad R^2 = 0,978$ $\bar{R}^2 = 0,976 \quad s(\hat{\beta}_0) = 0,799, s(\hat{\beta}_1) = 0,0075$ e $s(\hat{\beta}_2) = 0,204$				
$\eta = \beta_0 + \beta_1 \log(D) + \beta_2 \log(H)$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
DAP	1	7,9254	7,9254	1196,5 **
Altura DAP	1	0,1978	0,1978	29,9 **
Resíduo	28	0,1855	0,0066	
Total	30	8,3087		
$\eta = \beta_0 + \beta_1 \log(D) + \beta_2 \log(H)$ - Sequencial				
Causas de variação	G.L.	S.Q.	Q.M.	F
Altura	1	3,4957	3,4957	527,8 **
DAP Altura	1	4,6275	4,6275	698,6 **
Resíduo	28	0,1855	0,0066	
Total	30	8,3087		
$F_{1,29;0,05} = 4,18, F_{2,28;0,05} = 3,34$ e $F_{1,28;0,05} = 4,20$ $F_{1,29;0,01} = 7,60, F_{2,28;0,01} = 5,45$ e $F_{1,28;0,01} = 7,64$				

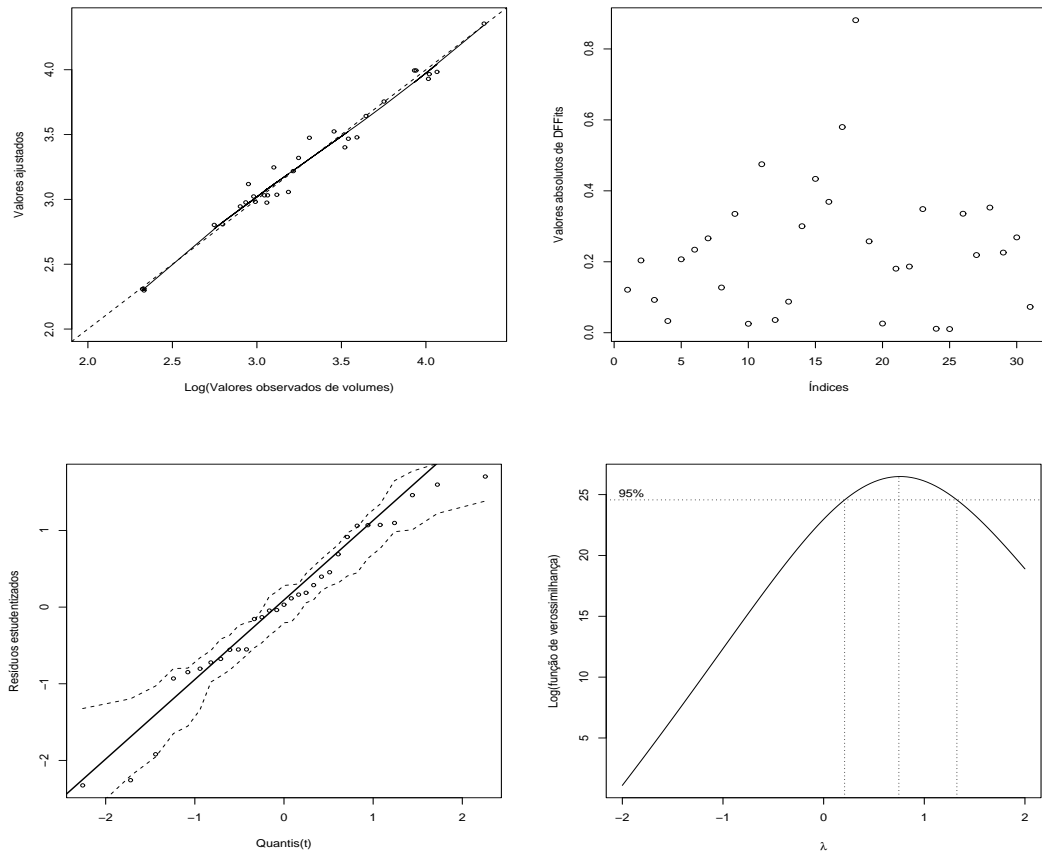


Figura 6.2: Gráficos de valores ajustados (modelo 2) *versus* log(valores observados), valores absolutos de DFFits *versus* índices, qqplot e Box-Cox

## 6.2 Dados de gordura no leite

A Tabela 6.3 refere-se a produções médias diárias de gordura (kg/dia) no leite de uma única vaca durante 35 semanas (McCulloch, 2001).

É comum supor que a produção média de gordura  $Y_i$  tem distribuição com média

$$\mu_i \propto \alpha t_i^\beta e^{\gamma t_i},$$

em que  $t$  representa a semana,  $\alpha$ ,  $\beta$  e  $\gamma$  são parâmetros.

Portanto,

$$\log \mu_i = \log \alpha + \beta \log(t_i) + \gamma t_i,$$

o que mostra a necessidade do uso de função de ligação logarítmica. Pode-se supor ainda



Tabela 6.3: Produções médias diárias de gordura (kg/dia) do leite de uma vaca.

0.31	0.39	0.50	0.58	0.59	0.64	0.68
0.66	0.67	0.70	0.72	0.68	0.65	0.64
0.57	0.48	0.46	0.45	0.31	0.33	0.36
0.30	0.26	0.34	0.29	0.31	0.29	0.20
0.15	0.18	0.11	0.07	0.06	0.01	0.01

que  $Y_i \sim N(\mu_i, \tau^2)$ , isto é,

$$Y_i = \mu_i + \delta_i = \alpha t_i^\beta e^{\gamma t_i} + \delta_i,$$

em que  $\delta_i \sim N(0, \tau^2)$ .

Entretanto, na prática é comum supor que  $\log(Y_i) \sim N(\log \mu_i, \sigma^2)$ , isto é,

$$\log(Y_i) = \log \mu_i + \epsilon_i = \log \alpha + \beta \log(t_i) + \gamma t_i + \epsilon_i,$$

em que  $\epsilon_i \sim N(0, \sigma^2)$ .

A Figura 6.3 mostra que usar a distribuição normal com função de ligação logarítmica produz um melhor ajuste do que fazer transformação logarítmica dos dados e usar distribuição normal com função de ligação identidade. Necessária se faz a complementação com análise de resíduos e diagnósticos. O programa em R encontra-se no Apêndice.

## 6.3 Dados de importação Brasileira

Os dados da Tabela 6.4 referem-se a importações brasileiras (IM) em milhões de dólares, taxa de câmbio (TCI) e o Produto Interno Bruto representando a renda nacional (RN). Os dados são trimestrais das contas externas do Brasil no período de 1980 a 1998 (Banco Central). A taxa de câmbio representa a relação entre reais e dólar, isto é, quantos reais são gastos para comprar um dólar americano e, por fim, a renda nacional em número índice (dez90=100).

A Figura 6.4 mostra os gráficos de dispersão das variáveis duas a duas para os dados observados e considerando o logaritmo e o inverso da variável resposta. Verifica-se

Tabela 6.4: Importações brasileiras (IM) em milhões de dólares, taxa de câmbio (TCI) e o Produto Interno Bruto representando a renda nacional (RN), no período de 1980 a 1998.

IM	TCI	RN	IM	TCI	RN
5482	1.629	82.17	4046	1.423	109.40
5749	1.517	88.80	5495	1.356	111.36
6043	1.331	87.94	5173	1.244	105.50
5679	1.181	85.28	4576	1.046	97.60
5605	1.315	82.06	4265	1.091	96.39
5565	1.217	86.49	5474	1.091	106.01
5610	1.177	82.62	6345	1.300	100.01
5309	1.135	78.30	4330	1.380	91.70
4804	1.434	78.34	5034	1.354	104.02
4872	1.306	87.11	5614	1.314	108.26
5071	1.209	85.77	6015	1.452	101.05
4646	1.156	80.91	4630	1.499	97.02
3824	1.740	75.88	4725	1.626	101.71
3651	2.004	83.65	5221	1.467	103.80
3907	1.957	82.80	5976	1.441	101.30
4044	1.959	80.10	5230	1.421	99.90
3155	1.971	79.10	6007	1.388	106.90
3406	2.015	87.59	7328	1.340	108.92
3730	2.024	87.19	6914	1.305	106.01
3623	2.027	85.94	6049	1.283	104.01
3094	2.036	84.55	7087	1.279	109.66
3016	2.219	92.47	8023	1.075	115.30
3132	2.201	95.23	11814	0.957	116.45
3925	2.131	94.44	12065	0.942	113.92
3352	2.013	90.69	13651	0.955	116.09
2760	2.023	99.48	11917	0.951	115.67
3661	1.991	102.87	12030	0.970	114.93
4270	1.924	101.15	10738	0.980	111.63
3565	1.832	97.65	12478	0.995	118.06
3610	1.792	106.21	14235	1.012	122.90
3987	1.914	103.45	15837	1.030	120.69
3888	1.789	101.10	13150	1.049	116.90
3516	1.692	97.72	15405	1.067	123.85
3349	1.657	105.78	16930	1.086	126.37
3776	1.643	105.84	15873	1.106	122.55
3963	1.607	98.87	13415	1.126	118.11
3548	1.557	95.01	14591	1.147	125.74

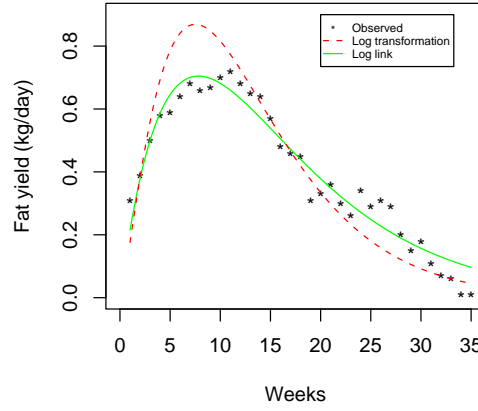


Figura 6.3: Valores observados e curvas ajustadas

que existe uma relação não-linear entre IM e as variáveis explanatórias TCI e RN. Essa relação não-linear diminui, porém, não desaparece quando se usa o logaritmo da variável resposta. Entretanto, a transformação  $1/IM$  parece linearizar a relação com as variáveis explanatórias. Nota-se, contudo, a presença de heterogeneidade de variâncias em todos os casos, principalmente relacionada à variável RN. A relação entre TCI e RN não parece muito forte e mostra uma variabilidade grande. Portanto, espera-se de antemão uma falta de ajuste do modelo com distribuição normal.

Inicialmente, ajusta-se o modelo de regressão normal linear, supondo-se que

$$IM_i = \beta_0 + \beta_1 TCI_i + \beta_2 RN_i + \epsilon_i, \quad i = 1, 2, \dots, 74,$$

em que os  $\epsilon_i$ 's são erros aleatórios com as suposições usuais de um modelo de regressão normal linear, isto é,  $\epsilon_i \sim N(0, \sigma_1^2)$ . Como esperado, a Figura 6.5 mostra a falta de ajuste desse modelo, por meio dos gráficos dos valores observados *versus* valores ajustados, resíduos studentizados *versus* valores ajustados e resíduos studentizados *versus* TCI e RN. Vê-se que as suposições de homocedasticidade e independência dos erros aleatórios são violadas.

Resultados semelhantes para os gráficos de verificação de ajuste foram obtidos, supondo-se

$$\log(IM_i) = \beta_0 + \beta_1 TCI_i + \beta_2 RN_i + \epsilon_i, \quad i = 1, 2, \dots, 74,$$

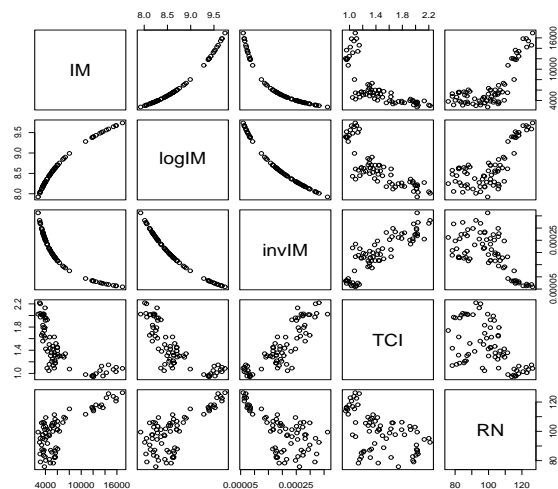


Figura 6.4: Gráficos de dispersão - valores observados e transformados e variáveis explanatórias

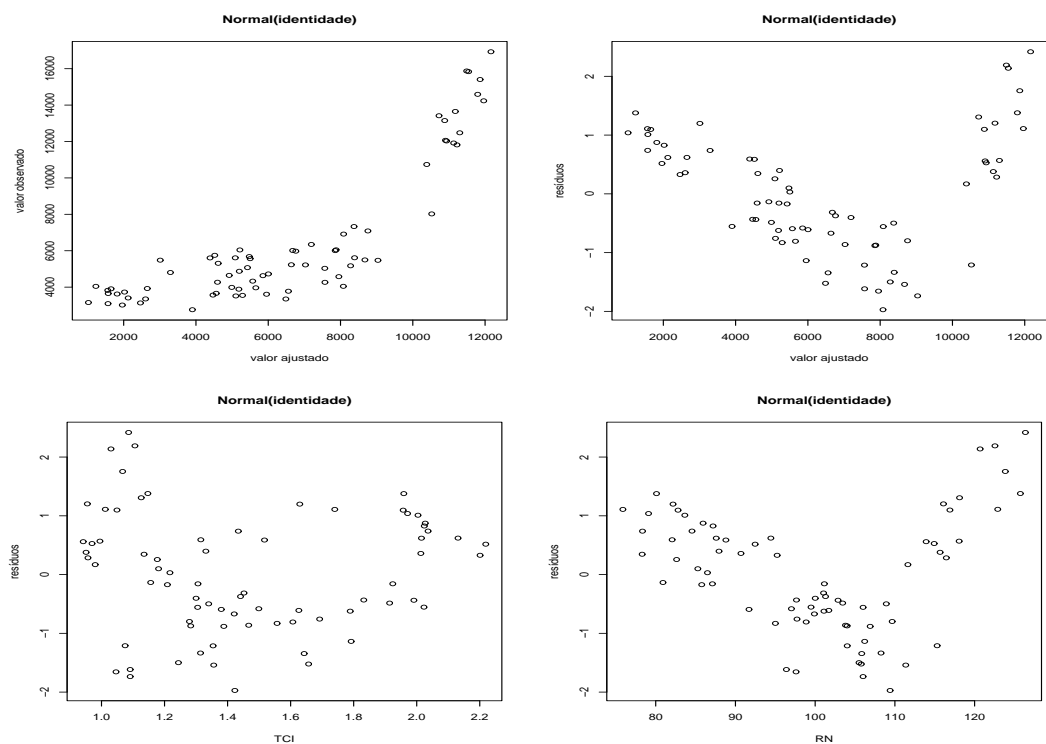


Figura 6.5: Importação - Gráficos para verificação de ajuste, distribuição normal

em que  $\epsilon_i \sim N(0, \sigma_2^2)$  e, também, para o modelo em que  $IM_i \sim N(\mu_i, \sigma_3^2)$  com  $\mu_i = \beta_0 + \beta_1 TCI_i + \beta_2 RN_i$ . A Figura 6.6 confirma a falta de ajuste adequado desses três modelos fazendo-se suposição de distribuição normal para a variável resposta com funções de ligação identidade e logarítmica ou suposição de distribuição normal para o logaritmo da variável resposta com função de ligação identidade.

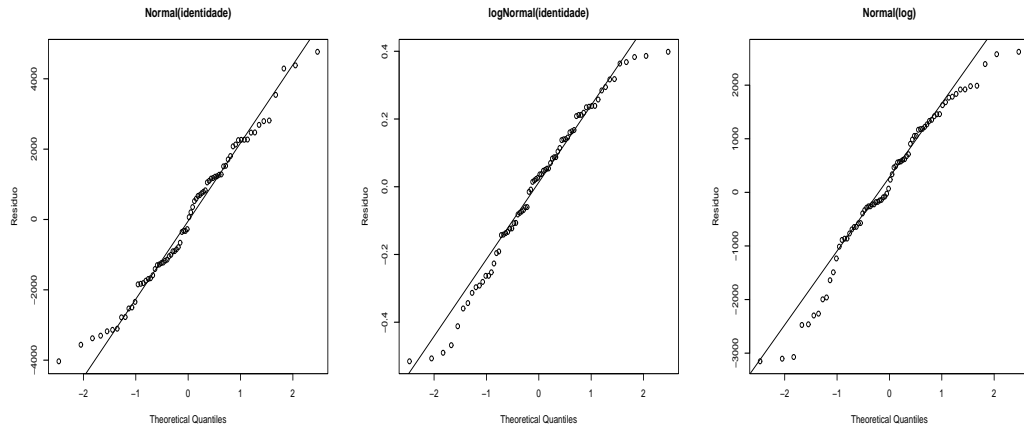


Figura 6.6: Importação - Gráficos normais de probabilidade, supondo-se distribuição normal

É importante notar que o desvio para modelos com distribuições contínuas depende da escala dos dados como mostra a Tabela 6.5, não servindo, portanto, como estatística para verificação de ajuste. No caso da distribuição normal, os desvios residuais são chamados de Somas de Quadrados Residuais. Verifica-se, também, que a estatística AIC depende da escala da variável resposta (IM ou  $\log(IM)$ ).

Tabela 6.5: Resumo do ajuste do modelo linear generalizado normal para diferentes funções de ligação.

Variável resposta	F. de Ligação	Desvio Residual	$\hat{\sigma}$	AIC
IM	$\mu$	315765900	2109	1347,7
$\log(IM)$	$\mu$	4,0	0,2382	2,6
IM	$\log \mu$	146543227	1436,7	1290,9

Modelos alternativos podem ser usados, supondo-se que  $IM_i \sim G(\mu_i, \phi)$  com as funções de ligação canônica (inversa), logarítmica e identidade. A Tabela 6.7 mostra um

Tabela 6.6: Resumo do ajuste do modelo linear generalizado gama para diferentes funções de ligação.

F. de Ligação	$\hat{\phi}$	Desvio	AIC
$1/\mu$	0,0307	2,294	1240,9
$\log \mu$	0,0524	3,908	1280,6
$\mu$	0,0892	6,191	1315,0

resumo, considerando o ajuste para esses três casos. Vê-se que o modelo com menor AIC é aquele com distribuição gama e função de ligação canônica (inversa). Entretanto, seria interessante completar com estudos de simulação e testes "bootstrap" para a escolha do melhor modelo.

A Figura 6.7 mostra os gráficos dos valores ajustados *versus* valores observados e os gráficos normais de probabilidade desses três modelos fazendo-se suposição de distribuição gama para a variável resposta com funções de ligação inversa, identidade e logarítmica, confirmando o modelo escolhido.

Outros estudos referentes a pontos discrepantes e/ou influentes são necessários. Um resumo das estatísticas para o modelo escolhido encontra-se na Tabela 6.7.

Tabela 6.7: Resumo do ajuste do mlg gama com função de ligação inversa  $\log(1/\mu)$ .

Parâmetro	Estimativa	e.p.	t	$\Pr(> t )$
(Intercepto)	2.587e-04	4.372e-05	5.917	1.05e-07 ***
TCI	1.413e-04	1.320e-05	10.699	< 2e-16 ***
RN	-2.729e-06	2.891e-07	-9.439	3.64e-14 ***

## 6.4 Dados de tempos de sobrevivência de ratos

Os dados da Tabela 6.8 referem-se a tempos de sobrevivência de ratos após envenenamento, sob 4 diferentes tratamentos e 4 tipos de venenos (Box e Cox, 1964).

ANOVA for Time

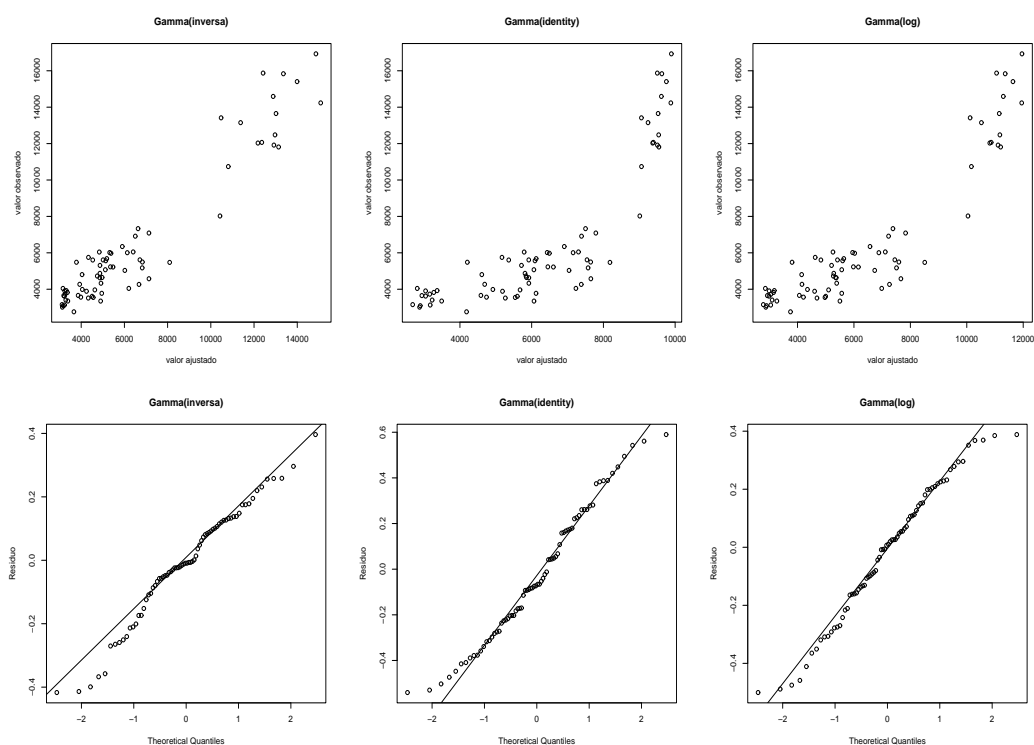


Figura 6.7: Importação - Gráficos dos valores ajustados *versus* valores observados, supondo-se distribuição gama

Source	SS	df	MS	F-ratio
Type	1.0330	2	0.5165	23.27
Treat	0.9212	3	0.3071	16.71
Interaction	0.2501	6	0.0417	1.88
Residual	0.8007	36	0.0222	

ANOVA for Rate

Source	SS	df	MS	F-ratio
Type	34.877	2	17.439	72.46
Treat	20.414	3	6.805	28.35
Interaction	1.571	6	0.262	1.09
Residual	8.643	36	0.240	

$$F_{0.05; 6,36} = 2.364 \quad F_{0.10; 6,36} = 1.945$$

$$F_{0.05; 2,36} = 3.259 \quad F_{0.05; 3,36} = 2.866$$

Tabela 6.8: Tempos de sobrevivência de ratos após envenenamento.

Tempo	Tipo	Trat.	Tempo	Tipo	Trat.	Tempo	Tipo	Trat.	Tempo	Tipo	Trat.
0,31	1	1	0,45	1	1	0,46	1	1	0,43	1	1
0,82	1	2	1,10	1	2	0,88	1	2	0,72	1	2
0,43	1	3	0,45	1	3	0,63	1	3	0,76	1	3
0,45	1	4	0,71	1	4	0,66	1	4	0,62	1	4
0,36	2	1	0,29	2	1	0,4	2	1	0,23	2	1
0,92	2	2	0,61	2	2	0,49	2	2	1,24	2	2
0,44	2	3	0,35	2	3	0,31	2	3	0,40	2	3
0,56	2	4	1,02	2	4	0,71	2	4	0,38	2	4
0,22	3	1	0,21	3	1	0,18	3	1	0,23	3	1
0,30	3	2	0,37	3	2	0,38	3	2	0,29	3	2
0,23	3	3	0,25	3	3	0,24	3	3	0,22	3	3
0,30	3	4	0,36	3	4	0,31	3	4	0,33	3	4

## 6.5 Dados de assinaturas de TV a cabo

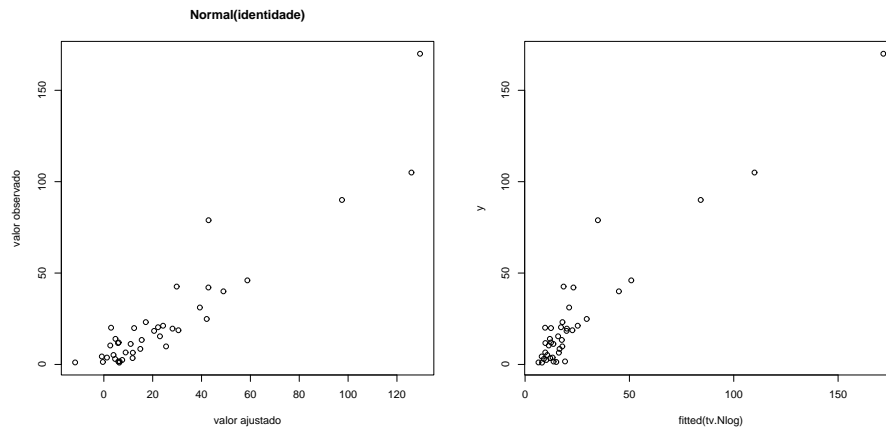
Os dados da Tabela 6.9 referem-se a número de assinantes (em milhares) de TV a Cabo  $Y$  em 40 áreas metropolitanas (Ramanathan, 1993), número de domicílios (em milhares) na área ( $x_1$ ), renda per capita (em US\$) por domicílio com TV a cabo ( $x_2$ ), taxa de instalação ( $x_3$ ), custo médio mensal de manutenção ( $x_4$ ), número de canais a cabo disponíveis na área ( $x_5$ ) e número de canais não pagos com sinal de boa qualidade disponíveis na área ( $x_6$ ). O interesse está em modelar o número de assinantes (variável resposta) como função das demais variáveis (variáveis explanatórias).

Inicialmente, será suposto que a variável resposta tem distribuição normal, isto é,  $Y \sim N(\mu, \sigma^2)$ . Como funções de ligação serão usadas a identidade,  $\eta = \mu$  e a logarítmica,  $\eta = \log \mu$ , e como preditor linear  $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$ .



Tabela 6.9: Número de assinantes (em milhares) de TV a Cabo  $Y$  em 40 áreas metropolitanas, número de domicílios (em milhares) na área ( $x_1$ ), renda per capita (em US\$) por domicílio com TV a cabo ( $x_2$ ), taxa de instalação ( $x_3$ ), custo médio mensal de manutenção ( $x_4$ ), número de canais a cabo disponíveis na área ( $x_5$ ) e número de canais não pagos com sinal de boa qualidade disponíveis na área ( $x_6$ ), (Ramanathan, 1993)

$Y$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
105,000	350,000	9839	14,95	10,00	16	13
90,000	255,631	10606	15,00	7,50	15	11
14,000	31,000	10455	15,00	7,00	11	9
11,700	34,840	8958	10,00	7,00	22	10
46,000	153,434	11741	25,00	10,00	20	12
11,217	26,621	9378	15,00	7,66	18	8
12,000	18,000	10433	15,00	7,50	12	8
6,428	9,324	10167	15,00	7,00	17	7
20,100	32,000	9218	10,00	5,60	10	8
8,500	28,000	10519	15,00	6,50	6	6
1,600	8,000	10025	17,50	7,50	8	6
1,100	5,000	9714	15,00	8,95	9	9
4,355	15,204	9294	10,00	7,00	7	7
78,910	97,889	9784	24,95	9,49	12	7
19,600	93,000	8173	20,00	7,50	9	7
1,000	3,000	8967	9,95	10,00	13	6
1,650	2,600	10133	25,00	7,55	6	5
13,400	18,284	9361	15,50	6,30	11	5
18,708	55,000	9085	15,00	7,00	16	6
1,352	1,700	10067	20,00	5,60	6	6
170,000	270,000	8908	15,00	8,75	15	5
15,388	46,540	9632	15,00	8,73	9	6
6,555	20,417	8995	5,95	5,95	10	6
40,000	120,000	7787	25,00	6,50	10	5
19,900	46,390	8890	15,00	7,50	9	7
2,450	14,500	8041	9,95	6,25	6	4
3,762	9,500	8605	20,00	6,50	6	5
24,882	81,980	8639	18,00	7,50	8	4
21,187	39,700	8781	20,00	6,00	9	4
3,487	4,113	8551	10,00	6,85	11	4
3,000	8,000	9306	10,00	7,95	9	6
42,100	99,750	8346	9,95	5,73	8	5
20,350	33,379	8803	15,00	7,50	8	4
23,150	35,500	8942	17,50	6,50	8	5
9,866	34,775	8591	15,00	8,25	11	4
42,608	64,840	9163	10,00	6,00	11	6
10,371	30,556	7683	20,00	7,50	8	6
5,164	16,500	7924	14,95	6,95	8	5
31,150	70,515	8454	9,95	7,00	10	4
18,350	42,040	8429	20,00	7,00	6	4



```
glm(formula = ASSIN ~ DOMIC + RENDA + TAXA + CUSTO + CADI +
CANAIS,
     family = gaussian(identity), data = tv)
```

Residual deviance: 5791.4 on 33 degrees of freedom

```
glm(formula = ASSIN ~ DOMIC + RENDA + TAXA + CUSTO + CADI +
CANAIS,
     family = gaussian(log), data = tv)
```

Residual deviance: 4632.4 on 33 degrees of freedom

Os modelos não são aceitos pelo valor tabelado da distribuição qui-quadrado com 33 graus de liberdade ao nível de 5% (47.40). Com isso, será utilizado um modelo com erro gama e as ligações identidade e logarítmica para tentar obter um melhor ajuste.

```
glm(formula = ASSIN ~ DOMIC + RENDA + TAXA + CUSTO + CADI +
CANAIS,
     family = Gamma(identity), data = tv)
```

Residual deviance: 4.3142 on 33 degrees of freedom

```
glm(formula = ASSIN ~ DOMIC + RENDA + TAXA + CUSTO + CADI +
```

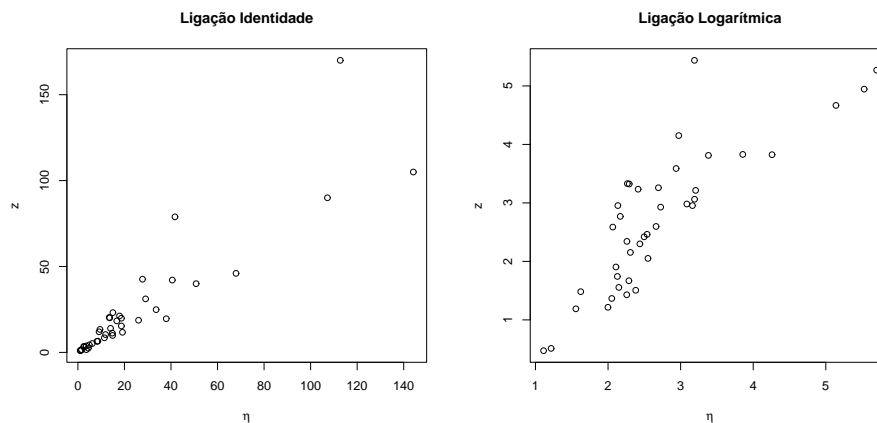
CANAIS,

```
family = Gamma(log), data = tv)
```

Residual deviance: 18.549 on 33 degrees of freedom

Os modelos são aceitos, pois possuem desvio inferior ao ponto crítico da distribuição qui-quadrado com 33 graus de liberdade. Porém o escolhido é o modelo gama com ligação identidade, pois o seu desvio de 4.3142 é menor que o desvio do modelo gama com ligação logarítmica. Ainda a ligação identidade apresenta gráfico da variável dependente modificada estimada  $z$  versus  $\hat{\eta}$  mais próximo da 1ª bissetriz, conforme Figura 1.1.

Figura 1.1: Gráficos de  $z$  versus  $\hat{\eta}$  segundo um modelo gama para ASSIN com as ligações identidade e logarítmica.



As características do modelo ajustado escolhido.

```
Call: glm(formula = ASSIN ~ DOMIC + RENDA + TAXA + CUSTO + CADI +
CANAIS,
          family = Gamma(identity), data = tv)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.71642	-0.26677	-0.07681	0.16052	0.71305

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-5.4897012	5.7682598	-0.952	0.3482	
DOMIC	0.4092640	0.0330192	12.395	5.8e-14 ***	
RENDA	0.0007127	0.749	0.4589	0.1159527	
TAXA	1.224	0.2298	CUSTO	-0.5436267	0.2527754
CADI	* CADI	0.4667268	0.1748683	2.669	0.0117 *
CANAIS	-0.2027451	0.1865639	-1.087	0.2850	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

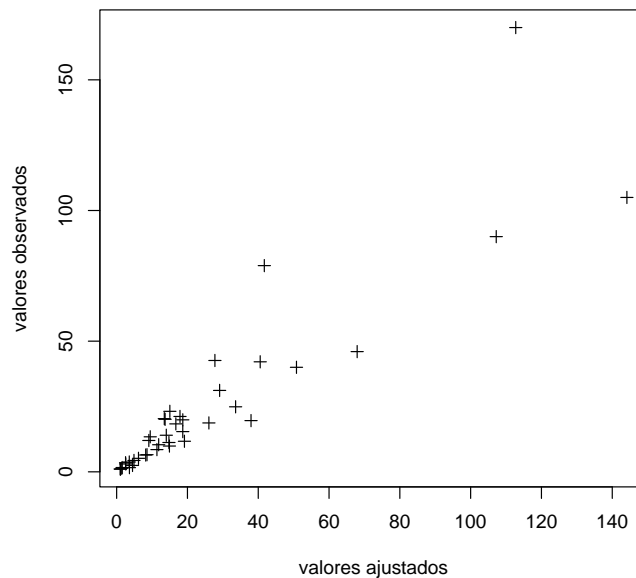
(Dispersion parameter for Gamma family taken to be 0.1322429)

Null deviance: 57.4761 on 39 degrees of freedom  
 Residual deviance: 4.3142 on 33 degrees of freedom AIC: 239.6

Number of Fisher Scoring iterations: 8

A Figura 1.2, mostra que os dados foram bem ajustados pelo modelo gama com ligação identidade.

Figura 1.2: *Valores observados versus valores ajustados.*

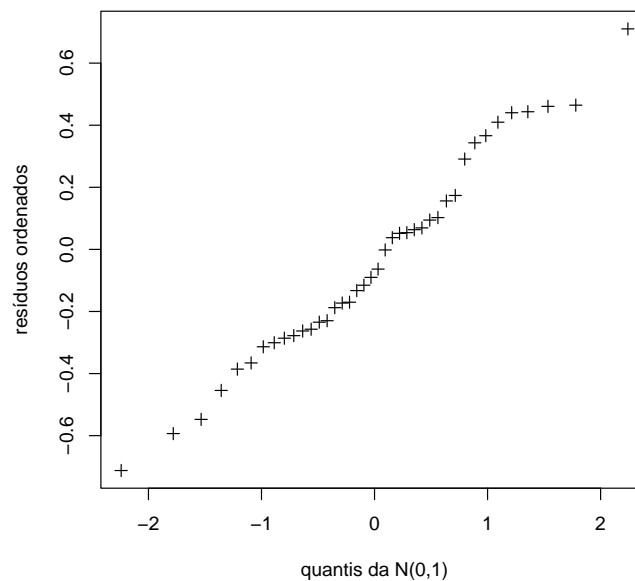


Para verificar se a função de ligação é adequada, foi acrescentado ao modelo original a covariável  $u = \hat{\eta} * \hat{\eta}$ .

```
glm(formula = ASSIN ~ DOMIC + RENDA + TAXA + CUSTO + CADI + CANAIS
+
      u, family = Gamma(identity), data = tv)
```

Residual deviance: 4.312 on 32 degrees of freedom

A redução no desvio (acima), provocada pela inclusão da variável  $u$ , não é significativa, indicando que a ligação identidade está correta. Na figura 1.3 observamos um comportamento próximo da reta  $y = x$  (1ª bissetriz), mostrando que a distribuição gama para o erro está adequada.

Figura 1.3: *Resíduos ordenados de Anscombe versus quantis da normal  $N(0,1)$ .*

As covariáveis *REND*A, *TAXA* e *CANAIS* não são significativas, com isso foi ajustado um novo modelo retirando as covariáveis *REND*A e *CANAIS*, mas supondo o mesmo erro e a mesma ligação.

Considera-se agora um novo modelo, retirando as covariáveis *REND*A e *CANAIS*, que não são significativas.

```
Call: glm(formula = ASSIN ~ DOMIC + TAXA + CUSTO + CADI, family =
Gamma(identity),
data = tv)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.65697	-0.27661	-0.08197	0.20657	0.71901

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-2.19631	2.14116	-1.026	0.31204	DOMIC
0.40066	0.03078	13.018	5.63e-15 ***		TAXA
0.06428	2.782	0.00864	**		CUSTO
				-0.69322	0.21777

```

-3.183  0.00305 ** CADI          0.55072    0.16206    3.398
0.00171 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for Gamma family taken to be 0.1302632)

```

Null deviance: 57.4761  on 39  degrees of freedom
Residual deviance:  4.4585  on 35  degrees of freedom AIC: 236.94

```

Number of Fisher Scoring iterations: 7

Apesar desse novo modelo ter um desvio um pouco maior do que o desvio do modelo anterior, o mesmo também é aceito pelo teste aproximado da distribuição qui-quadrado. Todas as covariáveis são significativas, mas o sinal da covariável *TAXA* não é o esperado, pois se a taxa de instalação é acrescida de US\$ 1 o número esperado de assinantes cresce, diferentemente do que se esperaria. Neste caso, a taxa teria que ser negativa para que tivéssemos um decréscimo no número esperado de assinantes. Com isso foi também retirada do modelo a covariável *TAXA*, pois o valor da taxa de instalação cobrado pelas empresas de TV a cabo é irrelevante para o nível de renda americano.

```

Call: glm(formula = ASSIN ~ DOMIC + CUSTO + CADI, family =
Gamma(identity),
      data = tv)

```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-0.8397	-0.2887	-0.1366	0.2242	0.8025

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.12892	1.37626	2.273	0.0291 *
DOMIC	0.39798	0.03323	11.975	4.08e-14 ***
CUSTO	-0.52672			
CADI	0.23631	-2.229	0.0321 *	0.14850
	0.10958			1.355
	0.1838			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.1493105)

Null deviance: 57.4761 on 39 degrees of freedom

Residual deviance: 5.2983 on 36 degrees of freedom AIC: 241.98

Number of Fisher Scoring iterations: 16

Esse novo modelo também é aceito pelo teste qui-quadrado ao nível de 5%, sendo que a covariável *CADI* não é significativa, mas os sinais das três covariáveis estão corretos. Na Figura 1.4 os valores observados versus valores ajustados revelam uma boa adequação do modelo.

Figura 1.4: *Valores observados versus valores ajustados.*

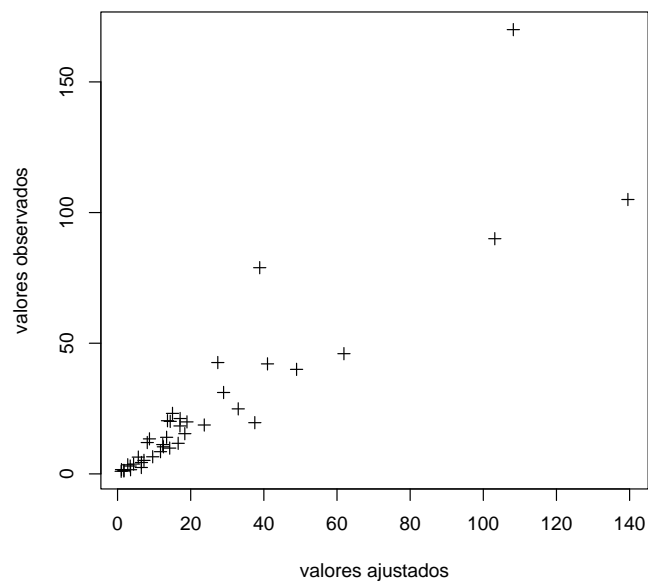
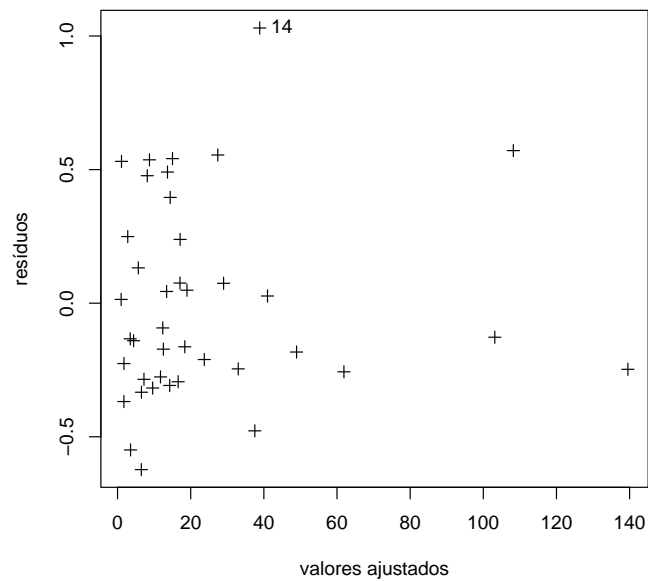


Figura 1.5: *Resíduos de Pearson versus valores ajustados.*





Os resíduos (Figura 1.5) apresentam-se de forma aleatória, o que mostra que a variância dos resíduos é constante e, também, como o resíduo da observação 14 se diferencia dos demais. Sendo o sinal da covariável *TAXA* diferente do esperado, será definida uma nova covariável, com o objetivo de obter o sinal desejado para a mesma.

A nova covariável adicionada ao modelo com as covariáveis *DOMI*, *CUSTO* e *CADI* foi *TAXA2* que é a covariável *TAXA* ao quadrado.

```
Call: glm(formula = ASSIN ~ DOMIC + CUSTO + CADI + TAXA + TAXA2,
  family = Gamma(identity), data = tv)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-0.60465	-0.27494	-0.04147	0.17900	0.71180

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.591161	3.415993	0.173	0.86363	DOMIC
0.403714	0.030642	13.175	6.49e-15 ***		CUSTO
0.203558	-3.387	0.00180	**		CADI
				0.503504	0.162519

```

3.098  0.00389 ** TAXA      -0.122873  0.298862 -0.411
0.68355 TAXA2      0.008371  0.008321  1.006  0.32150

```

```

---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

(Dispersion parameter for Gamma family taken to be 0.1302548)

```

```

Null deviance: 57.4761  on 39  degrees of freedom

```

```

Residual deviance:  4.3325  on 34  degrees of freedom AIC: 237.77

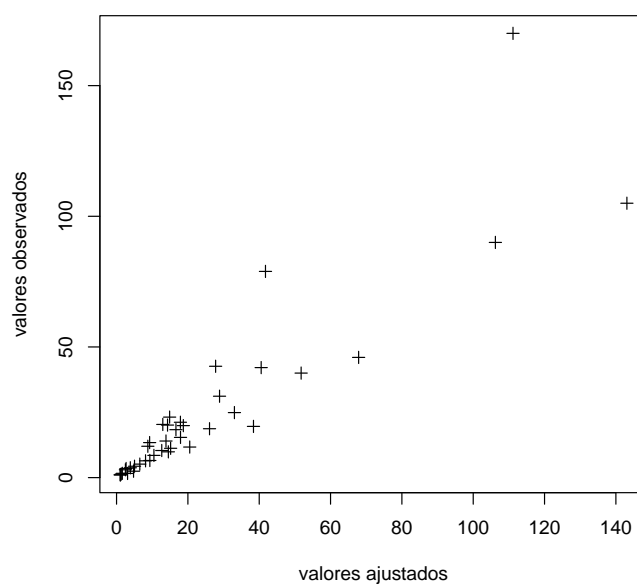
```

```

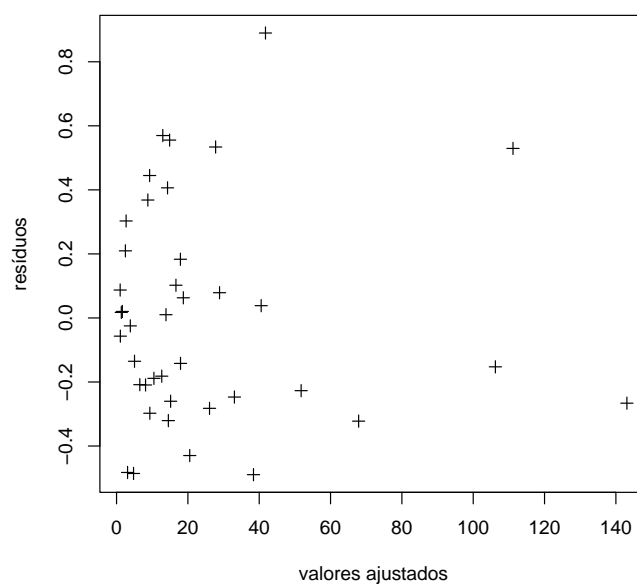
Number of Fisher Scoring iterations: 8

```

O modelo é aceito pelo teste qui-quadrado ao nível de 5%. Temos que as co-variáveis *TAXA* e *TAXA2* não são significativas mas o sinal da covariável *TAXA* agora apresenta-se correto às custas da não-linearidade do modelo.

Figura 1.6: *Valores observados versus valores ajustados.*

Na Figura 1.6 os pontos apresentam-se de forma linear, indicando que os dados foram bem ajustados.

Figura 1.7: *Resíduos de Pearson versus valores ajustados.*

Os pontos da Figura 1.7 apresentam-se de forma aleatória satisfazendo à hipótese de variância constante.

A partir das análises e dos resultados apresentados anteriormente, observa-se que aumentando o número de domicílios e o número de canais disponíveis na área teremos um aumento no número de assinantes; e, aumentando-se o custo de manutenção, tem-se um decréscimo no número de assinantes, isto é, os sinais obtidos pela regressão são os esperados. Assim, o melhor modelo para explicar os dados acima é dado por:

$$ASSIN = 3.1289 + 0.3980DOMIC - 0.5267CUSTO + 0.1485CADI.$$

Com este modelo pode-se concluir que: para cada aumento de uma unidade no número de domicílios, corresponderá um aumento de 0.3980 unidades no número de assinantes, mantidas constantes as demais covariáveis. Entretanto, para cada aumento de uma unidade no custo de manutenção, corresponderá uma redução de 0.5267 unidades no número de assinantes.

## 6.6 Dados de demanda de energia elétrica

O segundo modelo tem como variável resposta a demanda de eletricidade agregada per capita para o setor residencial (ELAR), e como variáveis explicativas o preço médio da eletricidade para o setor residencial (PER), o preço do gás natural para o setor residencial (PGR) e a renda per capita (RECA). Ainda, D1, D2, D3 e D4 são variáveis binárias e foram incluídas no modelo pois os dados são trimestrais. T representa o trimestre e os dados foram coletados no primeiro trimestre de 1961 até o quarto trimestre de 1983, com o total de 92 observações. Abaixo estão apresentados o número de observações e todas as variáveis do modelo.

ANO	T	ELAR	PER	PGR	RECA	D1	D2	D3	D4
1	1	0.30800536	7.64518690	2.77420998	0.00914456	1	0	0	0
1	2	0.26834363	7.95841503	3.10906148	0.00923471	0	1	0	0
1	3	0.27840772	7.92997503	4.04409552	0.00932230	0	0	1	0
1	4	0.28370830	7.82164145	3.05730581	0.00950548	0	0	0	1
2	1	0.33067492	7.35322905	2.71285081	0.00960076	1	0	0	0
2	2	0.28388155	7.71690655	3.14473939	0.00966927	0	1	0	0
2	3	0.30097651	7.64894676	3.47958493	0.00972013	0	0	1	0
2	4	0.29878822	7.53726721	3.01232100	0.00964969	0	0	0	1
3	1	0.35450837	7.04945183	2.66247821	0.00974009	1	0	0	0
3	2	0.29236847	7.52932024	3.09602141	0.00984403	0	1	0	0
3	3	0.32083428	7.37974453	3.95054865	0.00998568	0	0	1	0
3	4	0.30998397	7.31903124	3.03680444	0.01003013	0	0	0	1
4	1	0.36952662	6.81957054	2.62996173	0.01020502	1	0	0	0
4	2	0.31365973	7.20112085	3.01820755	0.01028083	0	1	0	0
4	3	0.35007703	7.02109432	3.96968317	0.01034642	0	0	1	0
4	4	0.33276981	7.02124262	2.90021181	0.01034942	0	0	0	1
5	1	0.38749585	6.54028463	2.74633431	0.01053808	1	0	0	0
5	2	0.33387709	6.86014271	3.09525871	0.01066791	0	1	0	0
5	3	0.36804986	6.66966391	3.92323565	0.01077701	0	0	1	0
5	4	0.35709164	6.63340855	3.02050757	0.01099775	0	0	0	1
6	1	0.41694346	6.15353727	2.66674948	0.01118029	1	0	0	0
6	2	0.35326710	6.51159859	3.01723003	0.01119937	0	1	0	0
6	3	0.40777826	6.27930784	3.81770802	0.01126028	0	0	1	0
6	4	0.38217804	6.20854807	2.84517026	0.01128659	0	0	0	1
7	1	0.44221917	5.87383795	2.57694674	0.01131980	1	0	0	0
7	2	0.38583204	6.20719862	2.94127989	0.01137994	0	1	0	0
7	3	0.42855132	6.06665373	3.66671538	0.01149168	0	0	1	0
7	4	0.41222385	5.98085690	2.74726343	0.01152810	0	0	0	1
8	1	0.49082169	5.49876261	2.47987032	0.01163357	1	0	0	0
8	2	0.40941107	5.83722544	2.79997373	0.01180093	0	1	0	0
8	3	0.48547110	5.61731529	3.45636535	0.01186746	0	0	1	0
8	4	0.44673607	5.56372929	2.64927459	0.01182800	0	0	0	1
9	1	0.53332543	5.13844633	2.35906005	0.01195509	1	0	0	0
9	2	0.44059545	5.48616648	2.68346119	0.01195672	0	1	0	0
9	3	0.54803473	5.21186781	3.31664300	0.01198937	0	0	1	0
9	4	0.49101120	5.22422218	2.56152606	0.01190421	0	0	0	1
10	1	0.57242423	4.84008980	2.32434344	0.01180006	1	0	0	0
10	2	0.48410484	5.13360834	2.64912558	0.01176797	0	1	0	0
10	3	0.60302770	4.98096657	3.27019763	0.01186475	0	0	1	0
10	4	0.52503026	5.08426189	2.55258965	0.01171888	0	0	0	1
11	1	0.60602528	4.76719999	2.32727671	0.01198772	1	0	0	0
11	2	0.51891249	5.01803827	2.62444520	0.01194521	0	1	0	0
11	3	0.62209785	4.94619703	3.33343983	0.01198712	0	0	1	0
11	4	0.56083840	4.99554968	2.58277440	0.01193268	0	0	0	1
12	1	0.62708759	4.79266357	2.37980080	0.01218264	1	0	0	0
12	2	0.54876824	5.09319210	2.68980694	0.01239293	0	1	0	0
12	3	0.65694511	4.95712137	3.23334769	0.01247493	0	0	1	0
12	4	0.60439968	4.91112804	2.51575303	0.01268085	0	0	0	1

### Análise por GLM

O ajuste do modelo será iniciado usando erro normal e as ligações identidade e logarítmica, respectivamente. Abaixo temos o resultados do valor do desvio e graus de liberdade do modelo. Observa-se através da Tabela 1 que o modelo que melhor ajusta os dados é o com erro normal e ligação logarítmica (desvio = 0.17169).

*Tabela 1 - Desvio e graus de liberdade dos modelos ajustados erro normal.*

ANO	T	ELAR	PER	PGR	RECA	D1	D2	D3	D4
13	1	0.68328059	4.67283297	2.33333063	0.01294289	1	0	0	0
13	2	0.57989609	4.94276857	2.67354584	0.01295302	0	1	0	0
13	3	0.72811598	4.79395962	3.13997459	0.01291298	0	0	1	0
13	4	0.62451297	4.83387899	2.55854464	0.01298187	0	0	0	1
14	1	0.66959435	4.83421087	2.40839648	0.01289692	1	0	0	0
14	2	0.59413171	5.32074070	2.75469518	0.01289350	0	1	0	0
14	3	0.70640928	5.39235258	3.19338322	0.01269503	0	0	1	0
14	4	0.62540507	5.39791536	2.73541474	0.01255311	0	0	0	1
15	1	0.70960039	5.22349358	2.61702061	0.01228601	1	0	0	0
15	2	0.62260377	5.44529819	2.95232224	0.01237817	0	1	0	0
15	3	0.74306965	5.50917530	3.47252870	0.01256718	0	0	1	0
15	4	0.63985091	5.46223164	3.01631594	0.01269196	0	0	0	1
16	1	0.74697447	5.23494911	2.91738129	0.01291349	1	0	0	0
16	2	0.61285406	5.55359745	3.27993631	0.01294898	0	1	0	0
16	3	0.75429350	5.64516401	3.91158652	0.01297108	0	0	1	0
16	4	0.69813275	5.46667147	4.27899122	0.01306254	0	0	0	1
17	1	0.81564754	5.30334044	3.27748561	0.01319841	1	0	0	0
17	2	0.63987577	5.68160534	3.70696568	0.01338583	0	1	0	0
17	3	0.81182355	5.90110493	4.23934031	0.01361182	0	0	1	0
17	4	0.69549668	5.62990713	3.48335361	0.01353800	0	0	0	1
18	1	0.84910756	5.35183573	3.37630939	0.01362886	1	0	0	0
18	2	0.66610706	5.73035097	3.68710351	0.01401979	0	1	0	0
18	3	0.82361311	5.77223778	4.21130323	0.01409499	0	0	1	0
18	4	0.71349722	5.51756096	3.52143955	0.01423942	0	0	0	1
19	1	0.87685442	5.17210197	4.39531507	0.01419568	1	0	0	0
19	2	0.67969620	5.58356667	3.75331378	0.01415907	0	1	0	0
19	3	0.81007040	5.78466034	4.43317604	0.01423306	0	0	1	0
19	4	0.71948880	5.53953552	3.98764658	0.01415617	0	0	0	1
20	1	0.84437078	5.37417889	3.97319126	0.01426184	1	0	0	0
20	2	0.68406653	5.80723810	4.34946060	0.01389695	0	1	0	0
20	3	0.89883024	6.06001234	5.06670094	0.01386312	0	0	1	0
20	4	0.73912853	5.74602461	4.36355448	0.01399696	0	0	0	1
21	1	0.85256535	5.66703844	4.19112778	0.01423567	1	0	0	0
21	2	0.69459844	6.27355528	4.63667440	0.01415394	0	1	0	0
21	3	0.88925880	6.57580376	5.15262365	0.01417765	0	0	1	0
21	4	0.73861104	6.19287395	4.57044888	0.01394008	0	0	0	1
22	1	0.86724007	6.18621683	4.59979963	0.01368745	1	0	0	0
22	2	0.69785839	6.52221394	5.05689907	0.01369381	0	1	0	0
22	3	0.84755844	6.66881037	5.81978750	0.01355230	0	0	1	0
22	4	0.73958969	6.39538670	5.41910744	0.01353536	0	0	0	1
23	1	0.82811236	6.25222349	5.49710894	0.01362200	1	0	0	0
23	2	0.68105930	6.60154247	5.79531860	0.01390618	0	1	0	0
23	3	0.94196534	6.87017965	6.52311754	0.01406361	0	0	1	0
23	4	0.74517667	6.52699089	5.60170937	0.01427785	0	0	0	1

Ligação	Desvio	DF
Identidade	0.21417	85
Logarítmica	0.17169	85

Abaixo temos as estimaticas do modelo com erro normal e ligação logarítmica.

Observe que todas as variáveis foram significativas ao nível de 5% de significância.

```
>fit2<-glm(elar~per+pgr+reca+DD1+DD2+DD3,
family = gaussian(link = "log"))
> summary(fit2)
```

```
Call: glm(formula = elar ~ per + pgr + reca + DD1 + DD2 + DD3,
family = gaussian(link = "log"))
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-8.904e-02	-3.255e-02	8.283e-05	2.854e-02	1.037e-01

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-2.22778	0.23949	-9.302	1.32e-14 ***	per
	-0.11247	0.02396	-4.694	1.02e-05 ***	pgr
	0.02012	3.628	0.000486	***	reca
	11.517	< 2e-16 ***	DD1	0.12624	0.02217
	1.74e-07 ***	DD2	-0.04949	0.02409	-2.054
	DD3	0.11021	0.02369	4.652	1.20e-05 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.002019894)

Null deviance: 3.20412 on 91 degrees of freedom  
 Residual deviance: 0.17169 on 85 degrees of freedom AIC: -301.03

Number of Fisher Scoring iterations: 4

Para verificar se o modelo ajustado é razoável, fazemos o gráfico de dispersão dos valores observados versus os valores ajustados. Observa-se através da Figura 1 que é razoável o ajuste aplicado. Além disso, ainda na Figura 1, temos que a distribuição normal para os erros ordenados de Anscombe é aceita; os resíduos de Pearson apresenta uma distribuição aleatória quando feita a sua dispersão versus os valores ajustados, indicando assim, que os resíduos são não correlacionados, ou seja, a hipótese de independência e variância constante para os resíduos são aceitas.

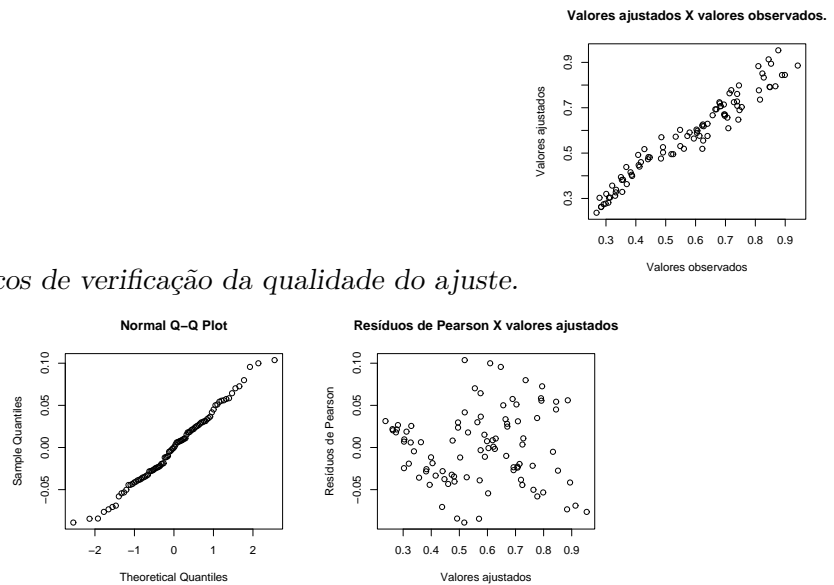


Figura 1: Gráficos de verificação da qualidade do ajuste.

Assim, concluímos que o modelo para ajustar a demanda de energia elétrica é dada por:

$$\log(elar) = -2.228 - 0.1125per + 0.073pgr + 163reca + 0.1262DD1 - 0.04949DD2 + 0.1102DD3$$



## Algoritmo

```
dados <- read.table("dados - Demanda de energia.txt", header=TRUE)
ano<-dados[,1] t<-dados[,2] elar<-dados[,3] per<-dados[,4]
pgr<-dados[,5] reca<-dados[,6] D1<-dados[,7] D2<-dados[,8]
D3<-dados[,9] D4<-dados[,10] DD1<-factor(D1) DD2<-factor(D2)
DD3<-factor(D3) DD4<-factor(D4)

##### Modelo Normal com função de ligação identidade
fit1<-glm(elar~per+pgr+reca+DD1+DD2+DD3,
          family = gaussian(link = "identity"))
summary(fit1)

##### Modelo Normal com função de ligação log
fit2<-glm(elar~per+pgr+reca+DD1+DD2+DD3,
          family = gaussian(link = "log"))
summary(fit2)

plot(elar,fit2$fitted, ylab="Valores ajustados",
     xlab="Valores observados",
     main="Valores ajustados X valores observados.")

##### calculo do resíduo de Anscomb ra<-
(elar-fit2$fitted) ra<-sort(ra) qqnorm(ra)

##### calculo do resíduo de pearson rp<-(elar-fit2$fitted)
plot(fit2$fitted,rp, ylab="Resíduos de Pearson",
     xlab="Valores ajustados",
     main="Resíduos de Pearson X valores ajustados")
```

## 6.7 Dados de tempo de funcionamento de um transformador

O tempo de funcionamento de qualquer equipamento do sistema elétrico definido como o tempo de funcionamento até o primeiro acidente (ou falha) é modelado através de um modelo exponencial log-linear de regressão. Esse modelo considera que a distribuição do tempo segue a distribuição exponencial cujo logaritmo da média é dado por uma estrutura linear contendo as variáveis que explicam o comportamento desse tempo e parâmetros desconhecidos. Nessa seção, nos restringiremos aos transformadores do sistema CHESF que totalizam XXX equipamentos.

A Tabela 7.1 apresenta os dados do tempo de funcionamento em meses de 61 transformadores e as suas respectivas características físicas.

O objetivo é estimar uma equação de regressão para modelar a média do tempo de funcionamento (*tempo*) dos transformadores, considerando como variáveis explicativas a espécie (*especie*) do equipamento, o número de enrolamentos (*enrolam*), o tipo do transformador (*tipo*) e a sua potência (em MVA) (*pot*).

Tabela 7.1: *Dados de Tempos de Funcionamento de Transformadores.*

COMPONENTE	TEMPO DE FUNCIONAMENTO	ESPÉCIE	ENROLAMENTO	TIPO	POTÊNCIA (MVA)
04T1-BGI	178	Trifásico	2	Transformador	100
04T3-MRD	135	Trifásico	2	Transformador	100
04T3-MTT	187	Trifásico	2	Transformador	100
04T3-RLD	29	Trifásico	2	Transformador	100
04T2-SNB	302	Trifásico	2	Transformador	33.34
01T3-C-ULG	36	Monofásico	3	Trafo Elevador	92.5
01T4-C-USD	288	Monofásico	2	Trafo Elevador	30
01T3-A-USD	314	Monofásico	2	Trafo Elevador	25
01T4-B-USQ	135	Monofásico	2	Trafo Elevador	150
01T6-B-USQ	121	Monofásico	2	Trafo Elevador	150
04T6-FNL	314	Trifásico	2	Transformador	100
05E2-B-OLD	141	Monofásico	1	Reator	50
04T1-PRI	230	Trifásico	3	Transformador	33.34
02A1-SMD	127	Trifásico	2	Reator	15.25
04T1-TSA	85	Trifásico	3	Transformador	33.34
01T3-C-ULG	51	Monofásico	3	Trafo Elevador	92.5
01T2-C-USQ	86	Monofásico	2	Trafo Elevador	150
01T1-UTC	162	Trifásico	2	Trafo Elevador	80
02T8-A-BGI	342	Trifásico	2	Transformador	5
04T2-BJS	247	Trifásico	3	Transformador	39.9
04T3-BJS	144	Trifásico	3	Transformador	39
04T2-GNN	102	Trifásico	3	Transformador	100
04T3-MRR	163	Trifásico	3	Transformador	100
04T1-MTT	31	Trifásico	2	Transformador	100
04T1-TAC	97	Trifásico	2	Transformador	100
04E1-C-TSA	277	Monofásico	1	Reator	3.33
01T3-B-USQ	142	Monofásico	2	Trafo Elevador	150
01T3-B-USQ	141	Monofásico	2	Trafo Elevador	150
04T2-ACD	54	Trifásico	1	Transformador	55
02T8-A-BGI	353	Trifásico	2	Transformador	5
04T6-FNL	344	Trifásico	2	Transformador	100
04T2-JCR	128	Trifásico	3	Transformador	100
04T3-MRD	6	Trifásico	2	Transformador	100
04T2-MRR	191	Trifásico	3	Transformador	100
04E1-B-TSA	291	Monofásico	1	Reator	3.33
04T1-TSA	115	Trifásico	3	Transformador	33.34
04T1-TSA	117	Trifásico	3	Transformador	33.34
06T2-A-UFL	359	Monofásico	2	Trafo Elevador	4.8
01T2-C-ULG	72	Monofásico	3	Trafo Elevador	92.5
01T1-C-USQ	104	Monofásico	2	Trafo Elevador	150
01T5-A-USQ	131	Monofásico	2	Trafo Elevador	150
01T3-C-USQ	142	Monofásico	2	Trafo Elevador	150
01T4-B-UST	268	Monofásico	2	Trafo Elevador	80
03T1-CRD	234	Trifásico	3	Transformador	38.6
02T1-A-CRD	87	Monofásico	2	Transformador	5
04E1-B-FTZ	123	Monofásico	1	Reator	3.33
04T1-GNN	96	Trifásico	3	Transformador	100
04T2-IRE	280	Trifásico	3	Transformador	39.9
05T2-B-MSI	4	Monofásico	2	Transformador	200
02T5-NTD	81	Trifásico	2	Transformador	1
04T1-OLD	160	Trifásico	2	Transformador	40
01T1-B-ULG	83	Monofásico	3	Trafo Elevador	92.5
04T4-ACD	350	Trifásico	3	Transformador	39
04T2-BJS	294	Trifásico	3	Transformador	39.9
04T2-GNN	137	Trifásico	3	Transformador	100
04T1-RIB	26	Trifásico	2	Transformador	100
05E1-C-SJI	187	Monofásico	1	Reator	33.3
06T1-C-UFL	272	Monofásico	2	Trafo Elevador	4.8
01T1-B-ULG	93	Monofásico	3	Trafo Elevador	92.5
01T1-B-ULG	102	Monofásico	3	Trafo Elevador	92.5
01T4-C-UXG	12	Monofásico	2	Trafo Elevador	185

As variáveis *tempo* e *pot* são quantitativas e *especie*, *enrolam* e *tipo* são variáveis qualitativas, também, denominadas fatores. O fator *especie* tem dois níveis, monofásico (M) e trifásico (T); o fator *enrolam* tem três níveis, 1 (A), 2 (B) e 3 (C) correspondentes a um, dois e três, rolamentos, respectivamente; e, finalmente, o fator *tipo* também com três níveis, transformador (A), reator (B) e trafo elevador (C). A potência (*pot*) é uma covariável contínua medida em *MVA* e o tempo (*tempo*) representa o número de meses ocorridos entre a energização e a falha detectada em um equipamento.

Inicialmente será realizada uma análise descritiva, a fim de melhor conhecer o comportamento das variáveis em questão.

A Figura 7.1 apresenta o box plot para a variável resposta *tempo*. Notou-se que ela apresenta uma grande variabilidade e assimetria positiva.

Figura 7.1: *Box plot para a variável resposta tempo.*

Na Figura 7.2 são apresentados os box plots da variável resposta *tempo* por espécie, número de enrolamentos e tipo do equipamento.

Figura 7.2: *Box plots da variável resposta tempo pelos fatores especie, enrolam e tipo.*

A Figura 7.3 apresenta o box plot para a variável explicativa *pot*. Nota-se que ela apresenta um valor extremo, com assimetria negativa.

Figura 7.3: *Box Plot da possível variável explicativa pot.*

Supondo que o tempo de funcionamento segue a distribuição exponencial e considerando uma função de ligação logarítmica para a média dessa distribuição, um modelo que melhor se ajusta aos dados é dado pela seguinte equação obtida através do software *R*:

$$\begin{aligned} \log \hat{\mu} = & 5.4623 - 0.9969\text{especieT} - 1.4239\text{enrolamB} - 2.3420\text{enrolamC} + \\ & 1.9692\text{tipoC} - 0.0087\text{pot} + 2.6591\text{especieT} * \text{enrolamB} + \\ & 3.5913\text{especieT} * \text{enrolamC} - 1.8892\text{especieT} * \text{tipoC}. \end{aligned}$$

Essa equação pode, também, ser escrita como:

$$\begin{aligned} \hat{\mu} = & 235.6357 \times 0.3690^{\text{especieT}} \times 0.2408^{\text{enrolamB}} \times 0.0961^{\text{enrolamC}} \times \\ & 7.1650^{\text{tipoC}} \times 0.9914^{\text{pot}} \times 14.2840^{\text{especieT} * \text{enrolamB}} \times \\ & 36.2803^{\text{especieT} * \text{enrolamC}} \times 0.1512^{\text{especieT} * \text{tipoC}}. \end{aligned}$$

Call:

```
glm(formula = tempo ~ especieT + enrolamB + enrolamC + tipoC +
    pot + especieT * enrolamB + especieT * enrolamC + especieT *
    tipoC, family = Gamma(log))
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	5.462287	0.450489	12.125	< 2e-16 ***

```

especieT      -0.996875   1.100525  -0.906   0.36503
enrolamB      -1.423926   0.871417  -1.634   0.10225
enrolamC      -2.342058   0.987979  -2.371   0.01776 *
tipoC         1.969210   0.756184   2.604   0.00921 **
pot           -0.008662   0.002906  -2.981   0.00288 **
especieT:enrolamB  2.659138   1.343139   1.980   0.04773 *
especieT:enrolamC  3.591276   1.425664   2.519   0.01177 *
especieT:tipoC   -1.889252   1.278320  -1.478   0.13943
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

(Dispersion parameter for Gamma family taken to be 1)

```

Null deviance: 35.016 on 60 degrees of freedom
Residual deviance: 20.441 on 52 degrees of freedom
AIC: 715.74

```

Number of Fisher Scoring iterations: 8

Este modelo foi aceito pelo teste do desvio, pois possui desvio (20.441) inferior ao ponto crítico (ou seja valor tabelado) da distribuição qui-quadrado com 52 graus de liberdade ao nível de 5% (69.83). As principais estimativas dos parâmetros são significativas, com exceção dos parâmetros *especieT*, *enrolamB*, *especieT : tipoC*.

Adota-se aqui os resíduos de Anscombe deduzidos para o modelo exponencial, com expressão

$$a_i = 3(\text{tempo}_i^{1/3} - \hat{\mu}_i^{1/3})/\hat{\mu}_i^{1/3},$$

pois esses resíduos podem ser considerados, aproximadamente, normais.

O gráfico de  $a_i$  versus  $\hat{\mu}_i$  mostrado na Figura 7.4 não revela dados aberrantes e os pontos se apresentam aleatoriamente distribuídos, sem nenhum padrão definido, concluindo-se, então, que esses tempos são variáveis independentes.

Figura 7.4: *Resíduos de Anscombe versus médias ajustadas segundo um modelo exponencial com a ligação logarítmica.*

Apesar de que alguns dados não estão bem ajustados, a Figura 7.5, não revela nenhum resíduo fora do intervalo  $(-2, 2)$ .

Figura 7.5: *Valores observados versus valores ajustados.*

A Figura 7.6 apresenta o gráfico da variável dependente modificada estimada  $z$  versus o preditor linear estimado  $\hat{\eta}$  indicando que a função de ligação está correta.

Figura 7.6: *Gráfico de  $z$  versus  $\hat{\eta}$  segundo um modelo exponencial para tempo com a função de ligação logarítmica.*



Os resíduos ordenados versus os quantis da  $N(0, 1)$  mostrados na Figura 7.7 suportam a distribuição exponencial para *tempo*.

Figura 7.7: *Resíduos de Anscombe ordenados versus os quantis da  $N(0, 1)$  segundo um modelo exponencial com a ligação logarítmica.*

A Figura 7.8 apresenta o gráfico da estatística de Cook versus o índice das observações. Nota-se que as observações #45 e #49 são pontos influentes.

Figura 7.8: *Estatística de Cook versus o índice das observações segundo um modelo exponencial para tempo.*

As observações #45 e #49 que foram detectadas como pontos influentes no modelo se referem a componentes do tipo transformador monofásico com dois enrolamentos, ou seja, referentes aos componentes 02T1-A-CRD e 05T2-B-MSI, respectivamente. Como existem apenas estas duas observações com estas características elas foram mantidas no modelo.

A chance do transformador só sofrer acidente após o tempo  $t$  (sobrevida ao tempo  $t$ ) foi calculada por:

$$S(t) = \exp\{-t \exp(-x^T \hat{\beta})\} = \exp(-t/\hat{\mu}).$$

As Tabelas 7.2 e 7.3 apresentam as sobrevidas a acidentes dos transformadores para as espécies monofásicos e trifásicos, respectivamente. As curvas de sobrevidas dessas duas espécies de transformadores por tipo de equipamento estão dadas nas Figuras 7.9 (monofásicos) e 7.10 (trifásicos). Pode-se concluir que para os equipamentos monofásicos a sobrevida é maior para o tipo transformador, depois reator e bem menor para o trafo-elevador. Entretanto, para os equipamentos trifásicos, a sobrevida é superior para o tipo reator, depois trafo-elevador e inferior quando o equipamento é um transformador.

Tabela 7.2: *Sobrevida a acidentes dos transformadores monofásicos.*

Meses	Transformador 02T1-A-CRD	Reator 05E2-B-OLD	Trafo Elevador 01T3-A-USD
12	0.8018	0.9245	0.9640
24	0.6429	0.8546	0.9293
36	0.5155	0.7901	0.8958
48	0.4133	0.7304	0.8636
60	0.3314	0.6753	0.8325
120	0.1098	0.4560	0.6931
180	0.0364	0.3079	0.5770
240	0.0121	0.2079	0.4804

Figura 7.9: *Curva de sobrevida a acidentes para equipamentos monofásicos.*

Tabela 7.3: *Sobrevida a acidentes dos transformadores trifásicos.*

Meses	Transformador 04T2-GNN	Reator 02A1-SMD	Trafo Elevador 01T1-UTC
12	0.9102	0.9552	0.9286
24	0.8285	0.9125	0.8623
36	0.7541	0.8716	0.8007
48	0.6863	0.8326	0.7436
60	0.6247	0.7953	0.6905
120	0.3903	0.6326	0.4768
180	0.2438	0.5031	0.3292
240	0.1523	0.4002	0.2273

Figura 7.10: *Curva de sobrevida a acidentes para equipamentos trifásicos.*

## 6.8 Dados de malária

Tabela A: Distribuição do número de casos notificados de malária e variáveis sócio-econômicas por municípios das regiões Norte, Sudeste e Sul do Brasil no ano de 2004. Base de dados: IBGE, IPEA e DATASUS.

Município	<i>casos_malaria</i>	<i>renda</i>	<i>pop_total</i>	<i>saude_san</i>	<i>gini</i>	<i>idh_ed</i>	<i>tx_analf</i>	<i>medicos</i>	<i>hab_urb</i>	<i>trans_fed</i>	<i>trans_est</i>
Boca do Acre	69	92.58	26959	1734185	0.623	0.637	53.7	0	858741	28818	0
Manaus	667	262.40	1405835	68841915	0.639	0.909	8.6	1	99996899	6563705	1940033
Lábrea	45	66.46	28956	1062929	0.690	0.633	58.5	0	93419	28818	0
Apuí	72	130.08	13864	920610	0.762	0.778	32.5	0	296952	613	0
Manacapuru	18	101.23	73695	0	0.590	0.761	34.3	0	0	0	0
Careiro	38	72.41	27554	1218307	0.660	0.709	51.7	0	1270967	28818	0
Borba	9	60.61	28619	1029544	0.668	0.734	38.9	0	1157283	69791	0
Tefé	27	117.55	64457	0	0.593	0.750	30.9	0	0	0	0
São Gabriel	36	106.61	29947	2658711	0.806	0.740	29.2	0	1413561	128079	338000
Uruará	55	132.73	18372	1633061	0.766	0.800	18.8	0	1005426	53642	0
Barcelos	7	68.85	24197	1407991	0.655	0.537	41.0	0	1083369	1057454	1137594
Tabatinga	25	142.08	37919	2813418	0.641	0.780	28.6	0	2153573	2072889	458431
Novo Airão	9	93.27	9651	807289	0.584	0.771	33.5	2	555089	90255	0
Iranduba	28	113.85	32303	2127159	0.559	0.766	27.8	0	484284	272518	0
Codajás	7	70.90	17507	757799	0.532	0.679	48.8	0	679546	108775	0
Altamira	130	205.00	62285	9165028	0.592	0.797	30.2	1	8016004	35393	0
Belém	38	313.93	1272354	200851724	0.651	0.928	6.9	2	28018810	873377	0
Marabá	50	188.59	134373	17859298	0.632	0.799	23.7	0	7192799	833855	760028
Santarém	59	139.90	186297	19701070	0.618	0.884	17.2	0	5181294	3963112	0
Itaituba	405	162.60	64486	3978979	0.649	0.797	33.9	0	15828	92240	0
Jacareacanga	66	140.92	5670	0	0.644	0.615	0.0	0	0	0	0
Tucuruí	43	180.99	60918	7881357	0.592	0.867	26.0	1	2248312	245105	0
Anajás	13	89.82	4613	0	0.556	0.529	64.9	0	0	0	0
Santana	257	121.08	17326	0	0.618	0.751	34.3	0	0	0	0
Castanhal	14	162.25	121249	9440432	0.573	0.854	17.5	0	4713655	372188	0
Dom Eliseu	4	145.70	23801	2544074	0.650	0.726	44.6	0	1263420	294750	60000
Paragominas	5	166.04	58240	0	0.606	0.766	38.0	0	0	0	0
Redenção	69	200.72	59613	7042532	0.605	0.836	25.5	0	2286022	620692	454291
Itupiranga	61	85.71	14754	2481718	0.602	0.671	46.8	0	1406690	669217	0
Monte Alegre	14	87.87	20921	68028	0.590	0.784	24.9	0	354449	1381626	0

Município	casos_malaria	renda	pop_total	saude_san	gini	idh_ed	tx_analf	medicos	hab_urb	trans-fed	trans_est
São Paulo	46	610.04	10434252	990610375	0.618	0.919	4.9	3	995987427	919686	2153193
Campinas	11	614.86	969396	188598384	0.584	0.925	5.0	3	54282636	0	233418
Ribeirão Preto	10	539.84	504923	178576664	0.564	0.918	4.4	5	13992434	1588622	318750
Botucatu	2	426.18	108306	4700134	0.539	0.909	5.6	4	6337118	0	1421360
Presidente Prudente	2	482.62	189186	13514842	0.592	0.924	6.2	3	7870710	162497	329927
S. José do Rio Preto	3	512.01	358523	67113072	0.557	0.916	5.4	4	10440369	3296099	419711
Mará	2	274.82	363392	46289950	0.490	0.909	6.6	0	30853650	0	445102
Osasco	1	390.45	652593	100746145	0.522	0.913	5.8	1	3021103	544659	39170
Guarulá	2	308.41	264812	33891820	0.525	0.885	8.4	1	39854124	0	209985
Mogi Guaçu	1	453.33	124228	31535951	0.519	0.886	7.1	1	9455674	536000	576000
Jacaré	1	353.95	191291	46644843	0.528	0.913	6.0	1	20987098	430620	88977
Araçatuba	1	503.17	169254	56088438	0.634	0.909	6.3	2	12576699	5150000	21840
Presidente Venceslau	1	364.66	37347	3685724	0.597	0.893	9.7	1	1911612	790287	995358
Bauru	1	500.27	316064	43704771	0.585	0.908	5.2	1	24287115	172107	180000
Fernandópolis	1	366.50	61647	4260736	0.541	0.891	9.2	2	2241073	288000	82090
Rio de Janeiro	18	596.65	5857904	828097433	0.616	0.933	4.4	4	385126048	450000	0
Paraty	1	312.07	29544	4624791	0.594	0.827	12.8	0	1089385	0	0
Niterói	4	809.18	459451	82892564	0.587	0.960	3.6	7	74000493	0	0
Angra dos Reis	1	275.66	119247	35583431	0.553	0.870	8.9	1	4200714	2764114	92000
S. João do Meriti	1	233.12	449476	32713472	0.475	0.895	5.7	0	12017744	2898377	0
Resende	0	365.45	104549	21094469	0.565	0.918	6.9	2	5826621	1233001	158931
Rio Bonito	0	276.19	49691	2726191	0.574	0.833	12.8	2	2897896	1047412	60718
Duque de Caxias	0	226.14	775456	87343897	0.528	0.873	8.0	0	29444592	220000	50074
Nova Iguaçu	0	237.50	920599	73949921	0.526	0.884	7.2	1	29738159	7758792	39973800
Itaboraí	0	202.29	187479	21249420	0.521	0.884	10.8	0	3783274	0	0
Macaé	0	392.94	132461	17637142	0.562	0.889	7.9	1	3133151	0	0
Nilópolis	0	298.30	153712	7702620	0.479	0.933	5.7	0	2751281	2753797	0
São Gonçalo	0	268.79	891119	58871177	0.489	0.896	5.8	0	21752118	0	0
Saquarana	0	266.64	52461	10272548	0.551	0.848	12.0	1	1482040	2615055	6316115
Maricá	0	321.41	76737	5893003	0.549	0.881	8.7	1	3966433	90000	0

Município	casos_malaria	renda	pop_total	saude_san	gini	idh_ed	tx_analf	medicos	hab_urb	trans_fed	trans_est
Curitiba	7	619.82	1587315	107701036	0.594	0.946	3.4	3	191897099	1708000	2154373
Foz do Iguaçu	12	326.19	258543	26119487	0.582	0.905	7.5	1	17906123	12546	58700
Pranchita	5	254.47	6260	698298	0.584	0.855	11.4	1	44409	70000	103264
Londrina	1	439.35	447065	110208024	0.577	0.910	7.1	2	6284351	95912	1069051
Apucarana	1	288.47	107827	1888614	0.513	0.877	9.6	1	5052465	196600	10009
Umuarama	2	313.76	90690	15110164	0.537	0.898	9.8	1	3463195	390088	386547
Mal. Cândido Rondon	1	341.71	41007	4777760	0.567	0.932	4.3	2	5930630	35063	0
Medianeira	1	277.50	37827	1912808	0.513	0.904	7.2	1	1188248	45051	345647
Arapongas	1	304.07	85428	4527118	0.504	0.883	8.4	1	6570150	677169	204322
Francisco Beltrão	1	307.97	67132	10442384	0.577	0.918	8.4	1	3587493	840827	122662
Paranaguá	1	305.36	127339	8861291	0.563	0.897	5.9	1	3571266	8492	60295
Mandaguari	1	238.95	31395	1531968	0.502	0.888	11.8	0	160495	0	0
Telêmaco Borba	1	275.60	61238	3612343	0.580	0.865	10.5	1	3611809	651000	76809
Paranavaí	1	312.97	75750	4438090	0.567	0.886	9.8	2	1135285	330720	329757
Missal	1	237.91	10433	1659870	0.549	0.898	7.8	0	936972	255000	0
Porto Alegre	4	709.88	1360590	690428164	0.606	0.951	3.6	6	104207038	496871	41989
Caxias do Sul	2	490.65	360419	71900922	0.511	0.945	3.7	2	112657	177059	510162
Passo Fundo	2	405.65	168458	11830539	0.585	0.912	5.6	3	6102389	15000	236165
Frederico Westphalen	2	297.09	26759	2198041	0.551	0.937	7.6	1	652048	482003	287544
Estrela	1	332.79	27401	1577862	0.469	0.929	3.6	1	696319	198283	220206
Maximiliano de Almeida	0	192.34	5651	579431	0.531	0.883	10.8	1	256593	149641	25000
Torres	2	317.74	30880	2239220	0.531	0.911	6.3	2	1780991	0	342
Ronda Alta	1	245.54	10051	1314708	0.575	0.867	9.4	0	195813	602930	39488
Alegria	2	160.89	5367	930560	0.563	0.854	12.3	0	90857	20000	51000
Bento Gonçalves	1	466.84	91486	6011847	0.473	0.938	3.9	1	2197841	275457	433468
Lajeado	1	395.34	64133	4631319	0.506	0.923	4.1	1	2235892	216630	79871
Ijuí	1	332.05	78461	4752438	0.572	0.926	5.7	1	1603367	710196	327550
Taquara	1	345.44	52825	4285	0.524	0.896	6.8	1	158914	37644	18092
Sapucaia do Sul	0	271.38	122751	6385902	0.464	0.900	5.5	0	2868537	50213	0
São Leopoldo	0	370.06	193547	40757529	0.551	0.922	4.8	1	6275356	400000	967529



# Capítulo 7

## Aplicações a Dados Discretos

Neste Capítulo serão apresentadas análises para três conjuntos de dados.

### 7.1 Dados binários e proporções

#### 7.1.1 Estimação da dose efetiva e seu intervalo de confiança

Como já foi visto no Capítulo 2, ensaios do tipo dose-resposta são muito usados na área de toxicologia. Em geral, os dados resultantes são proporções e os modelos mais usados são logístico, probit e complemento log-log. Tais modelos, ajustados a conjuntos de dados, no caso em que o preditor linear é uma regressão linear simples, podem ser usados para sumarizá-los através do par de estimativas  $(\hat{\beta}_0, \hat{\beta}_1)$  dos parâmetros e formam a base para comparação de diferentes conjuntos de dados (Morgan, 1992). Assim, por exemplo, podem ser usados para a comparação de potência de diferentes produtos (inseticidas, fungicidas, herbicidas etc).

Em muitos casos, porém, o interesse está na determinação de estimativas de doses efetivas,  $\theta_p$  ( $DE_{100p}$ ), que são doses, as quais sob o modelo ajustado causam uma mudança de estado em 100p% dos indivíduos. Um exemplo muito comum é a determinação da  $DL_{50}$  (também chamada dose mediana) que é a dose que causa 50% de mortalidade dos indivíduos. Assim, para os modelos citados tem-se:

$$\text{logit}(p) = \log \frac{p}{1-p} = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} (\log \frac{p}{1-p} - \hat{\beta}_0), \text{ logístico};$$

$$\text{probit}(p) = \Phi^{-1}(p) = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} [\Phi^{-1}(p) - \hat{\beta}_0], \text{ probit};$$

$$\log[-\log(1-p)] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} \{\log[-\log(1-p)] - \hat{\beta}_0\}, \text{ clog-log e}$$

$$\log \left[ \frac{1 - (1-p)^\lambda}{\lambda(1-p)^\lambda} \right] = \hat{\beta}_0 + \hat{\beta}_1 \hat{\theta}_p \Rightarrow \hat{\theta}_p = \frac{1}{\hat{\beta}_1} \left\{ \log \left[ \frac{1 - (1-p)^\lambda}{\lambda(1-p)^\lambda} \right] - \hat{\beta}_0 \right\}, \text{ Aranda-Ordaz (1981)}.$$

De uma forma geral, tem-se

$$\hat{\theta}_p = \frac{F^{-1}(p) - \hat{\beta}_0}{\hat{\beta}_1} = g(\hat{\beta}_0, \hat{\beta}_1), \quad (7.1)$$

sendo  $F(\cdot)$  uma f.d.a. de interesse.

Se  $p = 0,50$ , verifica-se que, para qualquer modelo simétrico, portanto, incluindo logístico e probit, a dose efetiva é obtida por

$$\hat{\theta}_{50} = -\frac{\hat{\beta}_0}{\hat{\beta}_1}$$

enquanto que para o modelo complemento log-log é dada por

$$\hat{\theta}_{50} = \frac{\log(\log 2) - \hat{\beta}_0}{\hat{\beta}_1}$$

e para o modelo de Aranda-Ordaz, por

$$\hat{\theta}_{50} = \frac{1}{\hat{\beta}_1} \left[ \log \left( \frac{2^\lambda - 1}{\lambda} \right) - \hat{\beta}_0 \right].$$

É importante notar que se o modelo está como função do logaritmo, em uma base  $b$  qualquer, da dose, então,  $\hat{\theta}_p = \log_b \hat{d}_p$  e, portanto a dose efetiva é obtida fazendo-se  $\hat{d}_p = b^{\hat{\theta}_p}$ .

Lembrando que  $\hat{\beta}_0 \sim N(\beta_0, \text{Var}(\hat{\beta}_0))$ ,  $\hat{\beta}_1 \sim N(\beta_1, \text{Var}(\hat{\beta}_1))$  e  $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \neq 0$ , isto é,  $\hat{\beta} \sim N(\beta, \mathbf{V})$  em que  $\mathbf{V} = \text{Cov}(\hat{\beta})$  é a matriz de variâncias e covariâncias dos estimadores dos parâmetros (inversa da matriz de informação de Fisher), os métodos mais comumente usados para a construção de intervalos de confiança para doses efetivas são: o método Delta, o de Fieller e o da razão de verossimilhanças (perfil de verossimilhanças) (Collet, 2002; Morgan, 1992).



**Método Delta**

Fazendo-se uma expansão de Taylor de primeira ordem para a expressão (7.1) de  $g(\hat{\beta}_0, \hat{\beta}_1)$  em torno de  $(\beta_0, \beta_1)$  tem-se:

$$\hat{\theta}_p = g(\hat{\beta}_0, \hat{\beta}_1) \approx g(\beta_0, \beta_1) + (\hat{\beta}_0 - \beta_0) \frac{\partial g(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} \Big|_{(\beta_0, \beta_1)} + (\hat{\beta}_1 - \beta_1) \frac{\partial g(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} \Big|_{(\beta_0, \beta_1)},$$

e, portanto,

$$\text{Var}(\hat{\theta}_p) = \hat{\gamma}^T \hat{\mathbf{V}} \hat{\gamma} = \frac{1}{\hat{\beta}_1^2} \{ \widehat{\text{Var}}(\hat{\beta}_0) + \hat{\theta}_p^2 \widehat{\text{Var}}(\hat{\beta}_1) + 2\hat{\theta}_p \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) \}.$$

$$\text{em que } \boldsymbol{\gamma}^T = \left( \frac{\partial g(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} \Big|_{(\beta_0, \beta_1)}, \frac{\partial g(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} \Big|_{(\beta_0, \beta_1)} \right) = \left( -\frac{1}{\beta_1}, -\frac{F^{-1}(p) - \beta_0}{\beta_1^2} \right).$$

Logo,

$$\widehat{\text{Var}}(\hat{\theta}_p) = \hat{\gamma}^T \hat{\mathbf{V}} \hat{\gamma} = \frac{1}{\hat{\beta}_1^2} \{ \widehat{\text{Var}}(\hat{\beta}_0) + \hat{\theta}_p^2 \widehat{\text{Var}}(\hat{\beta}_1) + 2\hat{\theta}_p \widehat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1) \}.$$

Assintoticamente,  $g(\hat{\boldsymbol{\beta}}) \sim N(g(\boldsymbol{\theta}), \boldsymbol{\gamma}^T \mathbf{V} \boldsymbol{\gamma})$  e, portanto, um intervalo de confiança para a dose efetiva  $\theta_p$  é dado por:

$$IC(\theta_p) : \hat{\theta}_p \mp z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\theta}_p)}.$$

Uma desvantagem desse intervalo de confiança é ser sempre simétrico, e, além disso, estar baseado na distribuição normal assintótica de  $g(\hat{\boldsymbol{\beta}})$ .

**Método baseado no teorema de Fieller**

O teorema de Fieller é um resultado geral que permite a obtenção de intervalos de confiança para razões de duas variáveis aleatórias normalmente distribuídas.

Suponha que  $\theta_p = \frac{F^{-1}(p) - \beta_0}{\beta_1}$  e considere a função  $\psi = \hat{\beta}_0 - F^{-1}(p) + \hat{\beta}_1 \theta_p$ . Então,  $E(\psi) = \beta_0 - F^{-1}(p) + \beta_1 \theta_p = 0$  e  $\text{Var}(\psi) = \text{Var}(\hat{\beta}_0) + 2\theta_p \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \text{Var}(\hat{\beta}_1)$ . Portanto,  $\hat{\beta}_1 \theta_p + \hat{\beta}_0 - F^{-1}(p) \sim N(0, \text{Var}(\psi))$  e

$$\frac{[\hat{\beta}_1 \theta_p + \hat{\beta}_0 - F^{-1}(p)]^2}{\text{Var}(\hat{\beta}_0) + 2\theta_p \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \theta_p^2 \text{Var}(\hat{\beta}_1)} \sim N(0, 1).$$

Logo, um intervalo de confiança para  $\theta_p$  pode ser obtido como a solução da inequação

$$\frac{[\hat{\beta}_1\theta_p + \hat{\beta}_0 - F^{-1}(p)]^2}{\text{Var}(\hat{\beta}_0) + 2\theta_p\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \theta_p^2\text{Var}(\hat{\beta}_1)} < z_{\alpha/2}^2,$$

sendo que os limites do intervalo de confiança serão dados pelas raízes da correspondente equação de segundo grau. No caso de raízes complexas, o intervalo não existirá. Em geral, os resultados são semelhantes aos obtidos pelo método delta. Uma outra alternativa é o método baseado na razão de verossimilhanças, que não será tratado aqui.

### Método baseado na razão de verossimilhanças

**Exemplo 7.1:** Usando-se os dados do **Exemplo 4.5**, e calculando-se a dose letal que mata 50% dos insetos e os intervalos de confiança com um coeficiente de confiança de 90% de probabilidade, pelos 3 métodos obtiveram-se os resultados:

i) dose letal:  $\hat{\theta}_{50} = \frac{3,226}{0,6051} = 5,33;$

ii) intervalos de confiança:

Fieller:  $4,8 < \theta_{50} < 5,9,$

Delta:  $4,8 < \theta_{50} < 5,9,$

Perfil de verossimilhança:  $5,0 < \theta_{50} < 5,7.$

### 7.1.2 Paralelismo entre retas no modelo logístico linear

Na área de toxicologia é muito comum o interesse na comparação da eficiência de produtos (fungicidas, inseticidas, herbicidas, medicamentos etc) ou tratamentos. Considerando-se o modelo logístico linear com uma variável quantitativa  $x$  (dose ou  $\log(\text{dose})$ ) e  $k$  produtos a serem testados, os preditores lineares a serem considerados são:

$$\text{logit}(p_{ij}) = \alpha_j + \beta_j \log(\text{dose}_i) - \text{retas concorrentes}$$

$$\text{logit}(p_{ij}) = \alpha_j + \beta \log(\text{dose}_i) - \text{retas paralelas}$$

$$\text{logit}(p_{ij}) = \alpha + \beta_j \log(\text{dose}_i) - \text{retas com intercepto comum}$$

$$\text{logit}(p_{ij}) = \alpha + \beta \log(\text{dose}_i) - \text{retas coincidentes.}$$

para  $j = 1, \dots, k$ . O ajuste desses modelos aos dados é testado através das diferenças dos

desvios residuais. No caso em que existem evidências de que o modelo de retas paralelas ajusta-se bem aos dados, tem-se, então, que a dose efetiva ( $\hat{\theta}_j^{(p)}$ ) para 100p% dos indivíduos é obtida a partir de:

$$\text{logit}(p) = \log \frac{p}{1-p} = \hat{\alpha}_j + \hat{\beta} \hat{\theta}_j^{(p)}, \quad j = 1, \dots, k.$$

Portanto, para  $j \neq j'$ , tem-se

$$\frac{\hat{\alpha}_j - \hat{\alpha}_{j'}}{\hat{\beta}} = \hat{\theta}_{j'}^{(p)} - \hat{\theta}_j^{(p)}.$$

Se  $x = \log(d)$ , então,

$$\frac{\hat{\alpha}_j - \hat{\alpha}_{j'}}{\hat{\beta}} = \log \frac{\hat{d}_j^{(p)}}{\hat{d}_{j'}^{(p)}} = \log \hat{\rho}_{jj'} \Rightarrow \hat{\rho}_{jj'} = \frac{\hat{\alpha}_j - \hat{\alpha}_{j'}}{\hat{\beta}} = \frac{DE_{j'}^{(50)}}{DE_j^{(50)}}$$

sendo  $\hat{\rho}_{jj'}$  a estimativa da eficiência relativa  $\rho_{jj'}$  do produto  $j$  em relação ao  $j'$  e  $\log[\hat{d}_j^{(p)}] - \log[\hat{d}_{j'}^{(p)}]$ , medindo a diferença horizontal entre as duas retas paralelas. Portanto,  $\rho_{jj'}$  é a razão de duas doses igualmente efetivas. Intervalos de confiança para  $\rho_{jj'}$  podem ser obtidos pelos métodos Delta, de Fieller e da razão de verossimilhanças (perfil de verossimilhanças) (Morgan, 1992; Collet, 2002).

### Exemplo 7.2: Resistência a cypermethrin

Amostras de 20 insetos, *Heliothis virescens* (praga do algodão), resistentes a cypermethrin, foram expostas a doses crescentes do inseticida, dois dias depois da emergência da pupa (Collet, 2002). Após 72h foram contados os números de insetos mortos e os resultados obtidos estão na Tabela 7.1.

### Considerações

Variável resposta:  $Y_i$  – número de insetos mortos em amostras de tamanho  $m_i = 20$

Distribuição: Binomial

Parte sistemática: completamente casualizado, modelos de regressão.

**Objetivo:** determinação de doses letais.

A Tabela 7.2 apresenta os desvios residuais, estatísticas  $X^2$  e seus respectivos números de graus de liberdade (g.l.) e a Tabela 7.3, a Análise de desvios.

Tabela 7.1: Números de insetos machos mortos em amostras de 20 insetos machos e fêmeas expostos a doses ( $d_i$ ) crescentes de cypermethrin

Doses ( $d_i$ )	Número de insetos mortos	
	Machos	Fêmeas
1,0	1	0
2,0	4	2
4,0	9	6
8,0	13	10
16,0	18	12
32,0	20	16

Tabela 7.2: Desvios residuais

Modelo	g.l.	Desvios	Valor de $p$	$X^2$	Valor de $p$
Constante	11	124,9	$< 0,0001$	101,4	$< 0,0001$
Sexo	10	118,8	$< 0,0001$	97,4	$< 0,0001$
Dose	6	15,2	0,0191	12,9	0,0446
Sexo + Dose	5	5,0	0,4146	3,7	0,5933

Vê-se que existem evidências de que o modelo com preditor linear com dois fatores, sexo (com dois níveis,  $j = 1, 2$ ) e dose (com 6 níveis,  $k = 1, \dots, 6$ , em princípio sem levar em consideração o fato de serem quantitativos), ajusta-se bem aos dados, enquanto que os modelos mais simples, não.

Tabela 7.3: Análise de Desvios

Causas de Variação	g.l.	Desvios	Valor de $p$
Sexo	1	6,1	0,0144
Sexo Dose	1	10,1	0,0002
Dose	5	109,7	$< 0,0001$
Dose Sexo	5	113,8	$< 0,0001$
Resíduo	5	5,0	0,5841
Total	11	124,9	

Pela Tabela 7.3 verifica-se que há evidências para efeito significativo de sexo e de dose. Note-se, ainda, que os desvios para sexo ignorando dose e, para sexo ajustado para dose, são diferentes devido à não ortogonalidade por se estar considerando a distribuição binomial. O mesmo ocorre para dose ignorando sexo e, para dose ajustada para sexo.

Pode-se, ainda, tentar uma simplificação desse modelo, considerando que dose é um fator quantitativo. Se for usado como preditor linear um polinômio com  $x = dose$ , verifica-se que há necessidade de grau 3. Como, porém, as doses estão em progressão geométrica é conveniente usar como variável regressora  $x = \log_2(dose)$ , considerando-se os modelos de retas concorrentes, paralelas, com intercepto comum e coincidentes. Os resultados para o desvio e a estatística  $X^2$  residuais estão apresentados na Tabela 7.4.

Tabela 7.4: Desvios residuais

Modelo	g.l.	Desvios	Valor de $p$	$X^2$	Valor de $p$
Constante	11	124,9	< 0,0001	101,4	< 0,0001
Sexo + Sexo.log(dose)	8	4,99	0,7586	3,51	0,8991
Sexo + log(dose)	9	6,75	0,6621	5,31	0,8074
Const. + Sexo.log(Dose)	9	5,04	0,8308	3,50	0,9411
Const. + log(Dose)	10	16,98	0,0748	14,76	0,1395

Pela Tabela 7.4, vê-se que existem evidências que os modelos com retas concorrentes, paralelas e com intercepto comum ajustam-se bem aos dados. Tem-se, ainda, que as diferenças de deviances entre os modelos com retas paralelas e retas concorrentes ( $6,76 - 4,99 = 1,77$ ) e entre os modelos com intercepto comum e retas concorrentes ( $5,04 - 4,99 = 0,05$ ), ambas com 1 grau de liberdade, não são estatisticamente significativas. Utilizando de parcimônia e facilidade de interpretação opta-se pelo modelo de retas paralelas. A Tabela 7.5 traz a análise de desvios para o modelo escolhido.

Tabela 7.5: Análise de Desvios

Causas de Variação	g.l.	Desvios	Valor de $p$
Sexo	1	6,1	0,0144
Regressão Linear	1	112,0	< 0,0001
Resíduo	9	6,8	0,7473
Total	11	124,9	

A partir do modelo escolhido obtêm-se, então, respectivamente, para machos e fêmeas, as equações:

$$\log \frac{\hat{p}_i}{1 - \hat{p}_i} = -2,372 + 1,535 \log_2(dose_i) - \text{machos},$$

$$\log \frac{\hat{p}_i}{1 - \hat{p}_i} = -3,473 + 1,535 \log_2(dose_i) - \text{fêmeas};$$

e as doses que matam 50% dos insetos

$$\log_2(\hat{DL}_{50}) = \frac{2,372}{1,535} = 1,54 \Rightarrow DL_{50} = 4,68 - \text{machos},$$

$$\log_2(\hat{DL}_{50}) = \frac{3,473}{1,535} = 2,26 \Rightarrow DL_{50} = 9,61 - \text{fêmeas}.$$

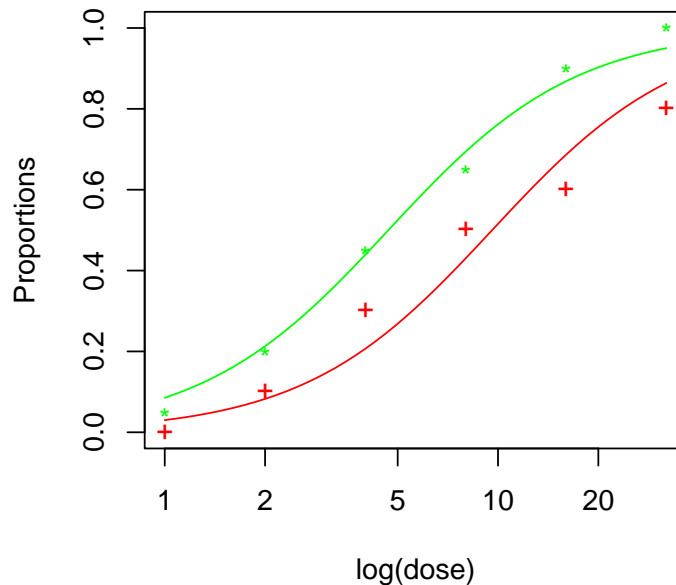


Figura 7.1: Cypermectrin - Proporções observadas e curvas ajustadas

Verifica-se que as fêmeas são mais resistentes, pois para matar 100p% das fêmeas há necessidade de uma dose 2 vezes maior do que para matar 100p% dos machos. Pode-se verificar que a dose letal para  $p = 0,9$  para as fêmeas está fora do intervalo estudado o que é perigoso pois acima da dose 32 não se sabe se o comportamento será o mesmo. Se o interesse estiver na estimação dessa dose há necessidade de se aumentar a amplitude de doses para fêmeas em um novo experimento. Necessária se faz ainda uma análise de resíduos e diagnósticos. A Figura 7.1 mostra o gráfico das curvas ajustadas e os valores observados. O programa em R encontra-se no Apêndice.

### Exemplo 7.3: Potência relativa - Mortalidade do besouro da farinha

Grupos de insetos (*Tribolium castaneum*, praga da farinha) foram expostos a doses (mg/l) crescentes de DDT,  $\gamma$ -BHC e mistura dos dois. Depois de 6 dias foram contados

os números de insetos mortos e os resultados obtidos estão na Tabela 7.6 (Collet, 2002).

Tabela 7.6: Proporções de insetos mortos quando expostos a doses crescentes de DDT,  $\gamma$ -BHC e mistura dos dois.

Inseticida	log(Doses)					
	2,00	2,64	3,48	4,59	6,06	8,00
DDT	3/50	5/49	19/47	19/50	24/49	35/50
$\gamma$ -BHC	2/50	14/49	20/50	27/50	41/50	40/50
DDT + $\gamma$ -BHC	28/50	37/50	46/50	48/50	48/50	50/50

### Considerações

Variável resposta:  $Y_i$  – número de insetos mortos em amostras de tamanho  $m_i$

Distribuição: Binomial

Parte sistemática: completamente casualizado, modelos de regressão.

**Objetivo:** determinação de doses letais e comparação de inseticidas.

A Tabela 7.7 apresenta os desvios e as estatísticas  $X^2$  residuais e seus respectivos números de graus de liberdade (g.l.) e a Tabela 7.8, a Análise de desvios, considerando-se o modelo logístico.

Tabela 7.7: Desvios residuais

Modelo	d.f.	Desvios	Valor de $p$	$X^2$	Valor de $p$
Constante	17	413,6	$< 0,0001$	347,1	$< 0,0001$
Inseticida	15	234,7	$< 0,0001$	215,0	$< 0,0001$
Dose	12	242,6	$< 0,0001$	218,9	$< 0,0001$
Inseticida + Dose	10	12,8	0,2316	11,8	0,2989

Vê-se que existem evidências de que o modelo com preditor linear com dois fatores, inseticida (com três níveis,  $j = 1, 2, 3$ ) e dose (com 6 níveis,  $k = 1, \dots, 6$ , em princípio sem levar em consideração o fato de serem quantitativos), ajusta-se bem aos dados, enquanto que os modelos mais simples, não.

Pela Tabela 7.8 verifica-se que há evidências para efeito significativo de inseticida e de dose. Note-se, ainda, que os desvios para inseticida ignorando dose e, para inseticida

Tabela 7.8: Análise de Desvios

Causas de variação	g.l.	Desvios	Valor de $p$
Inseticida	2	178,9	$< 0,0001$
Inseticida Dose	2	229,8	$< 0,0001$
Dose	5	171,0	$< 0,0001$
Dose Inseticida	5	221,8	$< 0,0001$
Resíduo	10	12,8	0,2316
Total	17	413,6	

ajustado para dose, são diferentes devido à não ortogonalidade por se estar considerando a distribuição binomial. O mesmo ocorre para dose ignorando inseticida e, para dose ajustada para inseticida. Pode-se, ainda, tentar uma simplificação desse modelo, considerando que dose é um fator quantitativo. Se for usado como preditor linear um polinômio com  $x = \log(\text{dose})$ , considerando-se os modelos de retas concorrentes, paralelas, com intercepto comum e coincidentes. Os resultados para o desvio e a estatística  $X^2$  residuais estão apresentados na Tabela 7.9.

Tabela 7.9: Desvios residuais

Modelo	g.l.	Desvios	Valor de $p$	$X^2$	Valor de $p$
Constante	17	413,6	$< 0,0001$	347,1	$< 0,0001$
Inseticida + Inseticida $\log(\text{dose})$	12	17,9	0,1191	17,6	0,1280
Inseticida + $\log(\text{dose})$	14	21,2	0,0946	20,3	0,1203
Const. + Inseticida $\log(\text{dose})$	14	24,7	0,0375	28,0	0,0141
Const. + $\log(\text{dose})$	16	246,8	$< 0,0001$	219,8	$< 0,0001$

Pela Tabela 7.9, vê-se que existem evidências que os modelos com retas concorrentes e paralelas ajustam-se bem aos dados. Tem-se, ainda, que a diferença de desvios entre os modelos com retas paralelas e retas concorrentes com 2 graus de liberdade, não é estatisticamente significativa. Utilizando de parcimônia e facilidade de interpretação opta-se pelo modelo de retas paralelas cuja análise de desvios está na Tabela 7.10.

A partir do modelo escolhido obtêm-se, então, as equações:

$$\text{DDT: } \log \frac{\hat{p}_i}{1 - \hat{p}_i} = -3,8425 + 2,6958 \log(\text{dose}_i)$$

$$\gamma\text{-BHC: } \log \frac{\hat{p}_i}{1 - \hat{p}_i} = -4,5553 + 2,6958 \log(\text{dose}_i)$$



Tabela 7.10: Análise de Desvios

Causas de variação	d.f.	Desvios	Valor de $p$
Inseticida	2	178,9	$< 0,0001$
Regressão Linear	1	213,4	$< 0,0001$
Resíduo	14	21,2	0,0946
Total	17	413,6	

$$\text{DDT} + \gamma\text{-BHC: } \log \frac{\hat{p}_i}{1 - \hat{p}_i} = -1,4248 + 2,6958 \log(\text{dose}_i)$$

as doses que matam 50% dos insetos

$$\text{DDT: } \log(\hat{LD}_{50}) = \frac{3,8425}{2,6958} = 1,42 \Rightarrow LD_{50} = 4,16$$

$$\gamma\text{-BHC: } \log(\hat{LD}_{50}) = \frac{4,5553}{2,6958} = 1,69 \Rightarrow LD_{50} = 5,42$$

$$\text{DDT} + \gamma\text{-BHC: } \log(\hat{LD}_{50}) = \frac{1,4248}{2,6958} = 0,53 \Rightarrow LD_{50} = 1,70$$

e as potências relativas

$$\text{da mistura em relação ao DDT: } \frac{4,16}{1,696} = 2,45$$

$$\text{da mistura em relação ao } \gamma\text{-BHC: } \frac{5,417}{1,696} = 3,19,$$

mostrando evidência de **sinergismo**, isto é, a mistura dos inseticidas potencializa o efeito. Necessária se faz ainda uma análise de resíduos e diagnósticos. A Figura 7.2 mostra o gráfico das curvas ajustadas e os valores observados. O programa em R encontra-se no Apêndice.

## 7.2 Dados de contagem

### 7.2.1 Modelo de Poisson

#### Exemplo 7.4: Armazenamento de microorganismos

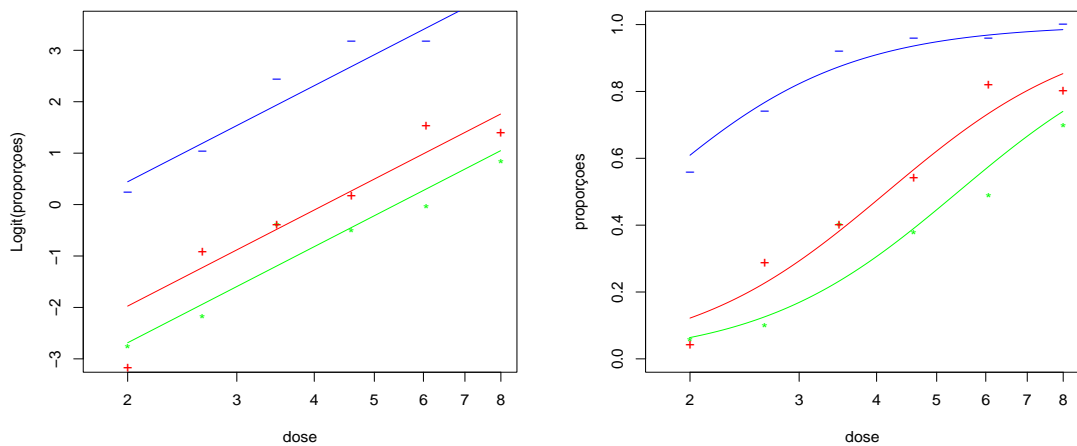


Figura 7.2: Tribolium - Proporções observadas e curvas ajustadas

A Tabela 7.11 mostra concentrações de bactérias (contagens por área fixa) feitas no congelamento inicial ( $-70^{\circ}\text{C}$ ) e após 1, 2, 6 e 12 meses (Francis et al., 1993).

Tabela 7.11: Concentrações de bactérias por área fixa

Tempo	0	1	2	6	12
Contagem	31	26	19	15	20

Pode-se supor, inicialmente, que  $Y$ , o número de bactérias por área fixa, segue distribuição Poisson com média  $\mu$ , isto é,  $Y \sim P(\mu)$ . Além disso, em geral, espera-se que a contagem média decresça com o tempo, isto é,

$$\mu_i \propto \frac{1}{(\text{tempo})^\gamma}$$

e, portanto,

$$\log \mu_i = \beta_0 + \beta_1 \log(\text{tempo}_i + 0,1),$$

sendo a constante 0,1 adicionada para evitar problemas com o tempo 0. A Tabela 7.7 apresenta os desvios e as estatísticas  $X^2$  residuais e seus respectivos números de graus de liberdade (g.l.) e a Tabela 7.8, a Análise de desvios, considerando-se o modelo logístico.

Pela Tabela 7.12 vê-se que existem evidências de que o modelo com preditor linear com  $\log(\text{tempo})$ , ajusta-se bem aos dados, enquanto que o modelo nulo, não. Pela Tabela

Tabela 7.12: Desvios residuais

Modelo	g.l.	Desvios	$X^2$
Constante	4	7,0672	7,1532
log(Tempo)	3	1,8338	1,8203

Tabela 7.13: Análise de Desvios

Causas de variação	g.l.	Desvios	Valor de $p$
Linear Regression	1	5,2334	0,0222
Error	3	1,8338	
Total	4	7,0672	

7.13 verifica-se que há evidências para efeito significativo de regressão linear. A equação da curva ajustada é dada por

$$\log(\mu_i) = 3,149 - 0,1261 \log(\text{tempo}_i)$$

que pode ser observada na Figura 7.2.1 juntamente com os valores observados. Necessária se faz ainda uma análise de resíduos e diagnósticos. O programa em R encontra-se no Apêndice.

## 7.2.2 Modelos log-lineares para tabelas $2 \times 2$

Considere a tabela de contingência  $2 \times 2$  que se segue, em que  $m_{ij}$  são contagens associadas aos fatores  $A$  e  $B$ .

	B		
A	1	2	
1	$y_{11}$	$y_{12}$	$y_{1.}$
2	$y_{21}$	$y_{22}$	$y_{2.}$
	$m_{.1}$	$y_{.2}$	$y_{..}$

Uma forma de se medir a associação entre os fatores  $A$  e  $B$  é por meio da razão das chances, dada por:

$$\text{Razão de chances observada} = \hat{\psi} = \frac{y_{11} \times y_{22}}{y_{12} \times y_{21}}$$

O interesse, em geral, está em se saber se o valor obtido não difere, estatisticamente, de 1, isto é, no teste da hipótese  $H_0 : \psi = 1$ . Isso corresponde, ao teste de independência para tabelas de contingência, como será mostrado a seguir.

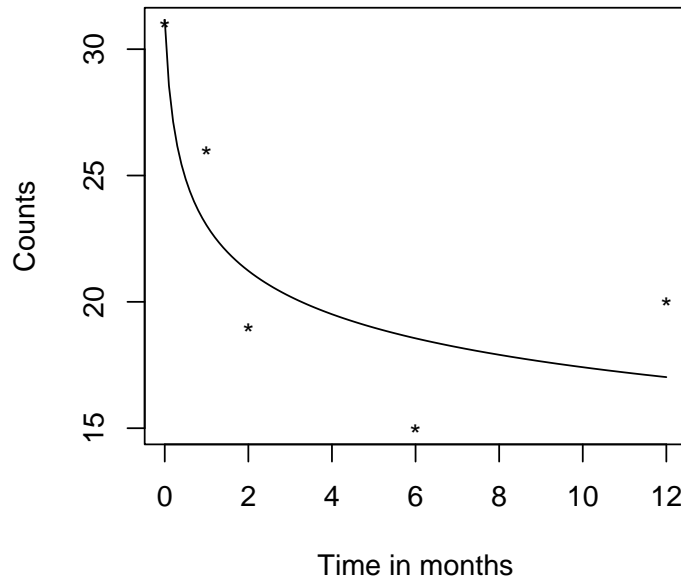


Figura 7.3: Concentrações de bactérias por área fixa: valores observados e curva ajustada

Pode-se supor, inicialmente, que  $Y_{ij}$  são variáveis aleatórias com distribuição Poisson de média  $\mu_{ij}$ . Em geral, as distribuições marginais de A e B não são de interesse. Os modelos de interesse, portanto, são: o modelo de independência e o modelo saturado.

(i) **Modelo de independência:**  $A + B$

Como visto no Capítulo 2, o modelo sob independência dos fatores A e B pode ser expresso por

$$\log \mu_{ij} = \mu + \alpha_i + \beta_j$$

ou ainda,

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B \quad i, j = 1, 2$$

com  $\lambda_1^A = \lambda_1^B = 0$ , isto é, com preditor linear  $\log(\mu_{ij})$  conforme o quadro que se segue.

A	B	
	1	2
1	$\lambda$	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B$

Verifica-se que o logaritmo da razão das chances é dado por

$$\log \psi = \log(\mu_{11}) + \log(\mu_{22}) - \log(\mu_{12}) - \log(\mu_{21}) = (\lambda + \lambda_2^A + \lambda_2^B) + \lambda - (\lambda + \lambda_2^B) - (\lambda + \lambda_2^A) = 0,$$

isto é, a razão das chances  $\psi = 1$ .

Além disso, pode-se mostrar que

$$\hat{\lambda} = \log\left(\frac{y_{1.}y_{.1}}{y_{..}}\right), \quad \hat{\lambda}_2^A = \log\left(\frac{y_{2.}}{y_{1.}}\right) \quad \text{e} \quad \hat{\lambda}_2^B = \log\left(\frac{y_{.2}}{y_{.1}}\right).$$

Mas,

$$\begin{aligned} y_{..} &= e^{\hat{\lambda}} + e^{\hat{\lambda} + \hat{\lambda}_2^A} + e^{\hat{\lambda} + \hat{\lambda}_2^B} + e^{\hat{\lambda} + \hat{\lambda}_2^A + \hat{\lambda}_2^B} \\ &= e^{\hat{\lambda}}(1 + e^{\hat{\lambda}_2^A})(1 + e^{\hat{\lambda}_2^B}) \end{aligned}$$

e, portanto,

$$\begin{aligned} \log y_{..} &= \hat{\lambda} + \log\left(1 + \frac{y_{2.}}{y_{1.}}\right) + \log\left(1 + \frac{y_{.2}}{y_{.1}}\right) \\ &= \hat{\lambda} + \log\left(\frac{y_{..}}{y_{1.}}\right) + \log\left(\frac{y_{..}}{y_{.1}}\right) \end{aligned}$$

implicando em

$$\hat{\lambda} = \log\left(y_{..} \frac{y_{1.}}{y_{..}} \frac{y_{.1}}{y_{..}}\right) = \log\left(\frac{y_{1.}y_{.1}}{y_{..}}\right),$$

isto é,

$$e^{\hat{\lambda}} = \hat{\mu}_{11} = y_{..} \frac{y_{1.}}{y_{..}} \frac{y_{.1}}{y_{..}} = y_{..} \hat{\pi}_{1.} \hat{\pi}_{.1} \text{ (independência)}$$

De forma semelhante, obtém-se

$$\begin{aligned} \hat{\mu}_{12} &= e^{\hat{\lambda} + \hat{\lambda}_2^B} = e^{\hat{\lambda}} e^{\hat{\lambda}_2^B} = y_{..} \frac{y_{1.}}{y_{..}} \frac{y_{.1}}{y_{..}} \frac{y_{.2}}{y_{.1}} \\ &= y_{..} \frac{y_{1.}}{y_{..}} \frac{y_{.2}}{y_{..}} = y_{..} \hat{\pi}_{1.} \hat{\pi}_{.2} \end{aligned}$$

(ii) **Modelo saturado ou de interação:**  $A*B \equiv A+B+A.B$

O preditor linear nesse caso é dado por

$$\log \mu_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad i, j = 1, 2$$

com  $\lambda_1^A = \lambda_1^B = \lambda_{1j}^{AB} = \lambda_{i1}^{AB} = 0$ , isto é, preditor linear  $\log(\mu_{ij})$  conforme quadro que se segue

A	B	
	1	2
1	$\lambda$	$\lambda + \lambda_2^B$
2	$\lambda + \lambda_2^A$	$\lambda + \lambda_2^A + \lambda_2^B + \lambda_{22}^{AB}$

Pode-se mostrar que

$$\hat{\lambda} = \log(m_{11}), \quad \hat{\lambda}_2^A = \log\left(\frac{y_{21}}{y_{11}}\right), \quad \hat{\lambda}_2^B = \log\left(\frac{y_{12}}{y_{11}}\right) \quad \text{e} \quad \hat{\lambda}_{22}^{AB} = \log\left(\frac{y_{22}y_{11}}{y_{12}y_{21}}\right) = \log(\hat{\psi}).$$

Tem-se, portanto, que o logaritmo da razão de chances corresponde ao parâmetro de interação e testar a hipótese  $H_0 : \psi = 1 \Rightarrow \log(\psi) = 0$  é o mesmo que testar o efeito de interação no modelo Poisson log-linear.

### Exemplo 7.5: Coletas de insetos em armadilhas adesivas

Considere os dados descritos no **Exemplo 2.4** em que os insetos de uma determinada espécie coletados em armadilhas adesivas de duas cores foram sexados, tendo como objetivo verificar se havia influência da cor da armadilha sobre a atração de machos e fêmeas. Tem-se

$$\text{Razão de chances observada} = \hat{\psi} = \frac{246 \times 32}{458 \times 17} = 1,01$$

A Tabela 7.14 apresenta os desvios e as estatísticas  $X^2$  residuais e seus respectivos números de graus de liberdade (g.l.), considerando-se o modelo Poisson log-linear.

Tabela 7.14: Análise de Desvios

Model	d.f.	Desvios	$X^2$
Cor da armadilha + sexo	1	0,001254	0,001252
Cor da armadilha * sexo	0	0	

Vê-se que existem evidências que o modelo de independência ajusta-se bem aos dados. Como esperado o modelo de interação (saturado) tem desvio e estatística  $X^2$  iguais a zero. As estimativas dos parâmetros do modelo saturado são:

	estimativa	e.p.	parâmetro
1	5,505	0,0638	1
2	0,622	0,0790	armcor(2)
3	-2,672	0,2508	sexo(2)
4	0,011	0,3104	armcor(2).sexo(2)

É importante notar que o modelo saturado reproduz os dados e que o logaritmo da razão de chances ajustada é  $\hat{\psi} = 0,01098$  resultando em  $\hat{\psi} = \exp(0,01098) = 1,01$ .

As estimativas para o modelo de efeitos principais (independência) são:

	estimativa	e.p.	parâmetro
1	5,505	0,0624	1
2	0,622	0,0764	armcor(2)
3	-2,665	0,1478	sexo(2)

Nota-se que agora o logaritmo da razão de chances é zero. Pela Tabela 7.14 tem-se que a diferença de desvios é 0,00125 ( $p = 0,9117$ ), não significativa, isto é, existem evidências para não se rejeitar a hipótese que a razão de chances é igual a 1, isto é, não há associação entre sexo do inseto e preferência por cor de armadilha adesiva.

## 7.3 Introdução

A malária é uma doença tropical endêmica em muitos países, vitimando anualmente milhões de pessoas em todo o mundo. Cerca de 40% da população mundial vive em áreas com risco de transmissão de malária, resultando em não menos que 300 milhões de pessoas infectadas no mundo a cada ano, mais de 90% em países africanos, com um número anual de mortes entre 1 e 1,5 milhão. A transmissão ocorre em mais de 100 países da América do Norte (México), América Central, América do Sul (principalmente na Bacia Amazônica), Caribe (República Dominicana e Haiti), África, Ásia (Subcontinente Indiano, Sudeste Asiático e Oriente Médio), Europa Oriental e Oceania.

No Brasil, a malária é endêmica na Região Norte, que responde por cerca de 80% dos casos em todo território nacional. Por este motivo, além do estudo da malária, abordando todos os seus aspectos epidemiológicos, fez-se, neste trabalho, o delineamento de modelos sócio-econômicos para representá-la no Brasil e nas regiões que apresentam fortes desigualdades sociais. Portanto, foram coletados, em 2004, dados sócio-econômicos e de incidência desta doença no país considerando as bases de dados do Instituto Brasileiro de Geografia e Estatística (IBGE), Instituto de Pesquisas Econômicas Aplicadas (IPEA) e DATASUS, em 90 municípios das regiões Norte, Sul e Sudeste. Ao final, é feita uma análise dessa infecção no país e nestas regiões através dos modelos normal linear e binomial

negativo log-linear com apresentação de resultados e sugestões para trabalhos posteriores de modelagem.

## 7.4 Metodologia

Em nosso estudo foram escolhidos, aleatoriamente, dois estados por região (Norte, Sul e Sudeste), sendo cada um deles representado por 15 dos seus respectivos municípios, igualmente selecionados, os quais encontram-se citados na Tabela 7.15.

Tabela 7.15: Regiões, estados e municípios estudados.

Região	Estado	Municípios		
Norte	Pará	Monte Alegre	Belém	Dom Eliseu
		Jacareacanga	Tucuruí	Anajás
		Santarém	Itaituba	Castanhal
		Paragominas	Marabá	Redenção
		Santana	Itupiranga	Altamira
	Amazonas	Boca do Acre	Manaus	Barcelos
		Manacapuru	Borba	Tabatinga
		Urucará	Lábrea	São Gabriel
		Novo Airão	Tefé	Iranduba
		Apuí	Careiro	Codajás
Sul	Rio Grande do Sul	Porto Alegre	Estrela	Taquara
		Bento Gonçalves	Torres	Caxias do Sul
		Sapucaia do Sul	Ijuí	Ronda Alta
		Maximiliano de Almeida	Lajeado	São Leopoldo
		Frederico Westphalen	Alegria	Passo Fundo
	Paraná	Foz do Iguaçu	Curitiba	Missal
		Mal. Cândido Rondon	Pranchita	Paranavaí
		Arapongas	Umuarama	Medianeira
		Telêmaco Borba	Londrina	Mandaguari
		Francisco Beltrão	Paranaguá	Apucarana
Sudeste	São Paulo	Ribeirão Preto	Campinas	São Paulo
		Presidente Prudente	Botucatu	Mauá
		S. José do Rio Preto	Osasco	Guarujá
		Fernandópolis	Jacareí	Araçatuba
		Presidente Venceslau	Bauru	Mogi Guaçu
	Rio de Janeiro	Rio de Janeiro	Paraty	Niterói
		Angra dos Reis	Itaboraí	Resende
		Duque de Caxias	Rio Bonito	Nova Iguaçu
		S. João do Mereti	Macaé	Nilópolis
		São Gonçalo	Saquarema	Maricá

A variável resposta, *malaria*, é o número de casos de malária registrados no ano de



2004, pelo SUS, em cada município, abrangendo os principais tipos de malária que assolam a região (Vivax, Malariae e Falciparum). Com base em informações fornecidas pelo IBGE e pelo IPEA, para cada um dos 90 (noventa) municípios selecionados, fez-se o levantamento de 10 (dez) possíveis variáveis explicativas, descritas abaixo. Vale ressaltar a grande dificuldade encontrada para a montagem da matriz com todas as variáveis independentes.

- ▷ *renda*: Renda per capita em reais, razão entre o somatório da renda de todos os membros da família e o número de seus membros;
- ▷ *pop\_total*: População total, levantamento censitário da população realizado pelo IBGE;
- ▷ *saude\_san*: Gastos em reais com saúde e saneamento básico, recursos municipais aplicados na área de habitação;
- ▷ *gini*: Índice de Gini, mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita;
- ▷ *idh\_ed*: Índice de Desenvolvimento Humano - Educação, índice obtido a partir da taxa de alfabetização e da taxa bruta de frequência à escola;
- ▷ *tx\_analf*: Taxa de analfabetismo, percentual de pessoas de 25 anos ou mais analfabetos;
- ▷ *medicos*: Número de médicos residente no município por 1000 habitantes;
- ▷ *hab\_urb*: Gastos em reais por função de habitação e urbanização, levantadas pelo IBGE;
- ▷ *trans\_fed*: Transferência em reais de recursos do governo federal aplicados à saúde para cada município;
- ▷ *trans\_est*: Transferência em reais de recursos do governo estadual aplicados à saúde para cada município.

## 7.5 Modelo Binomial Negativo

A variável dependente representando o número de casos de malária será modelada através do modelo de regressão binomial negativo log-linear, que considera a distribuição binomial negativa para a resposta cujo logaritmo da média é dado por uma estrutura linear. Os modelos são ajustados para os números de casos de malária para cada uma das três regiões, Norte, Sul e Sudeste, e para todas estas regiões conjuntamente.

Supondo que o número de casos de malária para o modelo nacional segue a distribuição binomial negativa e considerando uma função de ligação logarítmica para a média da distribuição, um modelo que melhor se ajusta aos dados é representado pela seguinte equação ( $p$ -valores abaixo das estimativas, entre parênteses) obtida com o software R:

$$\log \hat{\mu} = -12,85 + 24,31gini + 0,04758tx\_analf + 2,837 \times 10^{-7}trans\_fed.$$

$$(< 2 \times 10^{-16}) \quad (< 2 \times 10^{-16}) \quad (6,31 \times 10^{-6}) \quad (0,006)$$

Este modelo foi aceito pelo teste do desvio, pois possui desvio (95,213) inferior ao ponto crítico (ou seja, valor tabelado) da distribuição qui-quadrado com 86 graus de liberdade ao nível de 5% (108,65). Todas as estimativas dos parâmetros são altamente significativas e a estimativa do parâmetro  $\theta$  na função de variância do modelo  $V(\mu) = \mu + \mu^2/\theta$  é  $\hat{\theta} = 0,6455$  (E.P. 0,0990).

Os resíduos de Anscombe do modelo binomial negativo só podem ser calculados numericamente e, portanto, trabalha-se com os resíduos definidos como componentes do desvio que têm, aproximadamente, distribuição normal  $N(0, 1)$ .

A Figura 7.4 apresenta os gráficos dos resíduos do tipo componentes do desvio e dos elementos da diagonal da matriz de projeção do modelo para detectar os pontos aberrantes e os pontos de alavanca, respectivamente, para o modelo nacional. Os resíduos estão distribuídos aleatoriamente e apenas dois municípios, Altamira e São Gabriel correspondentes aos números de casos de malária 130 e 36, respectivamente, ficaram fora do intervalo  $(-2, 2)$ , porém estão muito próximos dos limites do intervalo, descartando a possibilidade de serem pontos aberrantes. Foram encontrados pontos de alavanca, isto é,

municípios que excederam duas vezes o valor da média dos  $h_{ii}$ . Os municípios foram Nova Iguaçu, Manaus, São Gabriel, Anajás, Urucará, Apuí, Lábrea, Codajás e Araçatuba, a grande maioria municípios da região Norte. Além de ser um ponto de alavanca, Manaus, também, foi identificada como um município influente através da estatística de Cook.

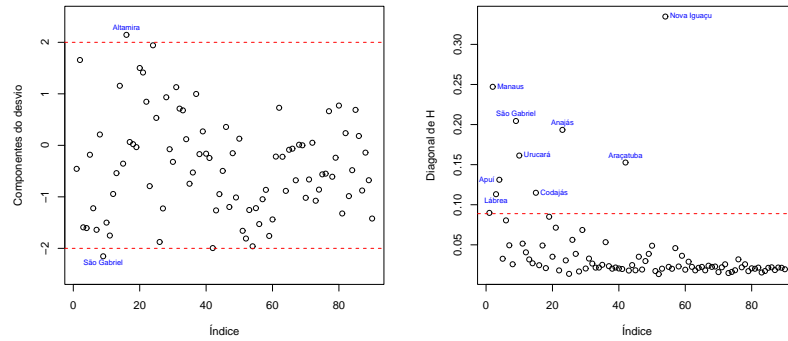


Figura 7.4: Resíduos do tipo componentes do desvio e elementos da diagonal da matriz de projeção versus índice das observações do modelo nacional.

A partir das análises e dos resultados apresentados anteriormente, observa-se que aumentando o índice de Gini, a taxa de analfabetismo e a transferência de recurso federais teremos um aumento no número esperado de casos de malária. Pode-se fazer, também, uma análise de sensibilidade com o objetivo de medir os impactos de cada variável explicativa no número esperado de casos de malária nos 90 municípios. Então, uma redução de 0,10 no índice de Gini e de 10% na taxa de analfabetismo, implica que o número esperado de casos de malária fica reduzido (marginalmente) em cerca de 91,21% e 37,86%, respectivamente.

A equação da regressão para a região Norte é

$$\begin{aligned} \log \hat{\mu} = & 0,01379 \text{ renda} + 4,241 \times 10^{-6} \text{pop\_total} - 4,205 \times 10^{-8} \text{saude\_san} \\ & + 3,643 \text{ gini} - 1,329 \times 10^{-6} \text{trans\_est}, \end{aligned}$$

sendo o  $p$ -valor da variável *renda* 0,000544, da *pop\_total* 0,005326, da *saude\_san* 0,000219, do *gini*  $3,37 \times 10^{-6}$  e da *trans\_est* 0,030731.

Este modelo foi aceito pelo teste do desvio, pois possui desvio (32,871) inferior ao ponto crítico (ou seja, valor tabelado) da distribuição qui-quadrado com 25 graus de liberdade ao nível de 5% (37,65). As estimativas dos parâmetros são altamente significativas, exceto aquela da variável *trans\_est* que é significativa ao nível de 5% de significância e  $\hat{\theta} = 1,375$  (E.P. 0,334).

A Figura 7.5 apresenta os gráficos dos resíduos do tipo componentes do desvio e dos elementos da diagonal da matriz de projeção do modelo para detectar os pontos aberrantes e os pontos de alavanca, respectivamente, para o modelo da região Norte. Os resíduos estão distribuídos aleatoriamente e três municípios, Santana, Dom Eliseu e Paragominas, correspondentes aos números de casos de malária 257, 4 e 5, respectivamente, ficaram fora do intervalo  $(-2, 2)$ , sendo, portanto, pontos aberrantes. Os municípios Manaus, Belém e Barcelos aparecem como pontos de alavanca. E Manaus, também, foi identificada como um município influente através da estatística de Cook.

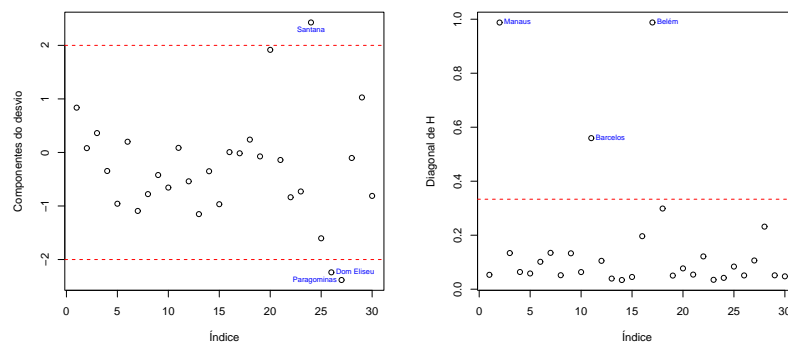


Figura 7.5: Resíduos do tipo componentes do desvio e elementos da diagonal da matriz de projeção versus índice das observações do modelo da região Norte.

Da equação acima, observa-se que aumentando a renda per capita, a população total e o índice de Gini e diminuindo os gastos com a saúde e saneamento básico e a transferência de recursos estaduais, teremos um aumento no número esperado de casos de malária. Uma análise econômica marginal mostra que o aumento da renda per capita em R\$ 1,00/dia na região Norte, conduzirá a um acréscimo de 51,24% no número esperado de casos de malária, enquanto uma redução de 0,10 no índice de Gini, implicará numa

diminuição de, aproximadamente, 30,53% no número esperado de casos de malária.

A equação da regressão para a região Sudeste é

$$\log \hat{\mu} = -0,6535 + 0,4711medicos + 3,304 \times 10^{-9}hab\_urb.$$

$$(0,0422) \quad (1,36 \times 10^{-6}) \quad (5,59 \times 10^{-7})$$

Este modelo foi aceito pelo teste do desvio, pois possui desvio (27,894) inferior ao ponto crítico (ou seja, valor tabelado) da distribuição qui-quadrado com 27 graus de liberdade ao nível de 5% (40,11). Todas as estimativas dos parâmetros são significativas e  $\hat{\theta} = 2,67$  (E.P. 1,52).

A Figura 7.6 apresenta os gráficos dos resíduos do tipo componentes do desvio e dos elementos da diagonal da matriz de projeção do modelo para detectar os pontos aberrantes e os pontos de alavanca, respectivamente, para o modelo da região Sudeste. Os resíduos estão distribuídos aleatoriamente e o município de Campinas com 11 casos de malária ficou fora do intervalo  $(-2, 2)$ , sendo, assim, um ponto aberrante. Os municípios São Paulo e Niterói aparecem como pontos de alavanca e, também, foram identificados como municípios influentes através da estatística de Cook.

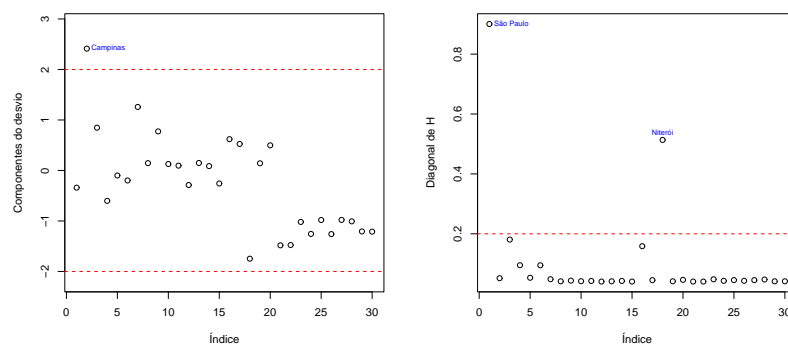


Figura 7.6: Resíduos do tipo componentes do desvio e elementos da diagonal da matriz de projeção versus índice das observações do modelo da região Sudeste.

Da equação anterior, observa-se que aumentando o número de médicos por mil habitantes e os gastos com habitação e urbanização teremos um acréscimo no número esperado de casos de malária. Em termos quantitativos, uma redução no número de médicos por mil habitantes em uma unidade, implica uma redução marginal do número esperado de casos de malária em cerca de 37,57%.

A equação da regressão para a região Sul é

$$\log \hat{\mu} = 1,296gini + 1,805 \times 10^{-8}hab\_urb - 1,399 \times 10^{-6}trans\_fed.$$

(0,00054)

(0,00106)

(0,03711)

Este modelo foi aceito pelo teste do desvio, pois possui desvio (23,917) inferior ao ponto crítico (ou seja, valor tabelado) da distribuição qui-quadrado com 27 graus de liberdade ao nível de 5% (40,11). Todas as estimativas dos parâmetros são significativas e  $\hat{\theta} = 4,93$  (E.P. 3,63).

A Figura 7.7 apresenta os gráficos dos resíduos do tipo componentes do desvio e dos elementos da diagonal da matriz de projeção do modelo para detectar os pontos aberrantes e os pontos de alavanca, respectivamente, para o modelo da região Sul. Os resíduos estão distribuídos aleatoriamente e o município de Foz do Iguaçu com 12 casos de malária ficou fora do intervalo  $(-2, 2)$ , sendo, então, um ponto aberrante. Os municípios Curitiba e Porto Alegre aparecem como pontos de alavanca. Ainda, os municípios Foz do Iguaçu, Curitiba e Porto Alegre foram identificados como influentes através da estatística de Cook.

Da equação acima, observa-se que aumentando o índice de Gini e os gastos com habitação e urbanização e diminuindo a transferência de recursos federais teremos um aumento no número de casos de malária. Logo, uma redução de 0,10 no índice de Gini conduz a uma redução marginal do número esperado de casos de malária de, aproximadamente, 12,16%.

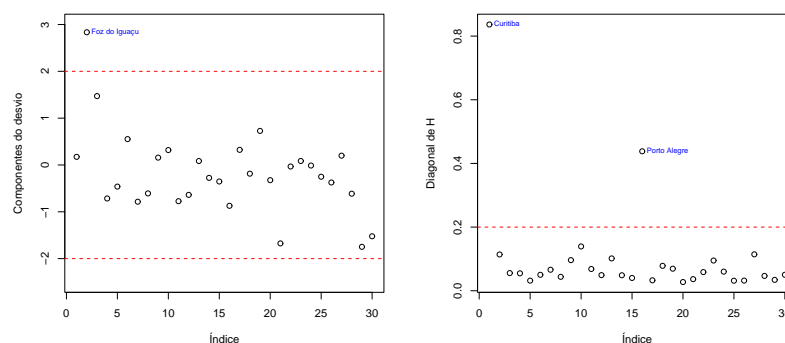


Figura 7.7: Resíduos do tipo componentes do desvio e elementos da diagonal da matriz de projeção versus índice das observações do modelo da região Sul.

## 7.6 Conclusões

No modelo binomial negativo log-linear as variáveis significativas foram Gini, taxa de analfabetismo e transferência de recursos do Governo Federal. Assim, com o crescimento da desigualdade de renda, da taxa de analfabetismo e da transferência de recursos federais cresce, também, o número de casos de malária. Para os modelos regionais, a variável Gini (habitação/urbanização) foi significativa para explicar o número de casos de malária na região Norte e Sul (Sudeste e Sul). Em ambas situações, a significância da estimativa ocorre quando existe maior diversidade do efeito da variável explicativa.

Tabela A: Distribuição do número de casos notificados de malária e variáveis sócio-econômicas por municípios das regiões Norte, Sudeste e Sul do Brasil no ano de 2004. Base de dados: IBGE, IPEA e DATASUS.

Município	<i>casos_malaria</i>	<i>renda</i>	<i>pop_total</i>	<i>saude_san</i>	<i>gini</i>	<i>idh_ed</i>	<i>tx_analf</i>	<i>medicos</i>	<i>hab_urb</i>	<i>trans_fed</i>	<i>trans_est</i>
Boca do Acre	69	92.58	26959	1734185	0.623	0.637	53.7	0	858741	28818	0
Manaus	667	262.40	1405835	68841915	0.639	0.909	8.6	1	99996899	6563705	1940033
Lábrea	45	66.46	28956	1062929	0.690	0.633	58.5	0	93419	28818	0
Apuí	72	130.08	13864	920610	0.762	0.778	32.5	0	296952	613	0
Manacapuru	18	101.23	73695	0	0.590	0.761	34.3	0	0	0	0
Careiro	38	72.41	27554	1218307	0.660	0.709	51.7	0	1270967	28818	0
Borba	9	60.61	28619	1029544	0.668	0.734	38.9	0	1157283	69791	0
Tefé	27	117.55	64457	0	0.593	0.750	30.9	0	0	0	0
São Gabriel	36	106.61	29947	2658711	0.806	0.740	29.2	0	1413561	128079	338000
Urucará	55	132.73	18372	1633061	0.766	0.800	18.8	0	1005426	53642	0
Barcelos	7	68.85	24197	1407991	0.655	0.537	41.0	0	1083369	1057454	1137594
Tabatinga	25	142.08	37919	2813418	0.641	0.780	28.6	0	2153573	2072889	458431
Novo Airão	9	93.27	9651	807289	0.584	0.771	33.5	2	555089	90255	0
Iranduba	28	113.85	32303	2127159	0.559	0.766	27.8	0	484284	272518	0
Codajás	7	70.90	17507	757799	0.532	0.679	48.8	0	679546	108775	0
Altamira	130	205.00	62285	9165028	0.592	0.797	30.2	1	8016004	35393	0
Belém	38	313.93	1272354	200851724	0.651	0.928	6.9	2	28018810	873377	0
Marabá	50	188.59	134373	17859298	0.632	0.799	23.7	0	7192799	833855	760028
Santarém	59	139.90	186297	19701070	0.618	0.884	17.2	0	5181294	3963112	0
Itaituba	405	162.60	64486	3978979	0.649	0.797	33.9	0	15828	92240	0
Jacareacanga	66	140.92	5670	0	0.644	0.615	0.0	0	0	0	0
Tucuruí	43	180.99	60918	7881357	0.592	0.867	26.0	1	2248312	245105	0
Anajás	13	89.82	4613	0	0.556	0.529	64.9	0	0	0	0
Santana	257	121.08	17326	0	0.618	0.751	34.3	0	0	0	0
Castanhal	14	162.25	121249	9440432	0.573	0.854	17.5	0	4713655	372188	0
Dom Eliseu	4	145.70	23801	2544074	0.650	0.726	44.6	0	1263420	294750	60000
Paragominas	5	166.04	58240	0	0.606	0.766	38.0	0	0	0	0
Redenção	69	200.72	59613	7042532	0.605	0.836	25.5	0	2286022	620692	454291
Itupiranga	61	85.71	14754	2481718	0.602	0.671	46.8	0	1406690	669217	0
Monte Alegre	14	87.87	20921	68028	0.590	0.784	24.9	0	354449	1381626	0



Município	casos_malaria	renda	pop_total	saude_san	gini	idh_ed	tx_analf	medicos	hab_urb	trans_fed	trans_est
São Paulo	46	610.04	10434252	990610375	0.618	0.919	4.9	3	995987427	919686	2153193
Campinas	11	614.86	969396	188598384	0.584	0.925	5.0	3	54282636	0	233418
Ribeirão Preto	10	539.84	504923	178576664	0.564	0.918	4.4	5	13992434	1588622	318750
Botucatu	2	426.18	108306	4700134	0.539	0.909	5.6	4	6337118	0	1421360
Presidente Prudente	2	482.62	189186	13514842	0.592	0.924	6.2	3	7870710	162497	329927
S. José do Rio Preto	3	512.01	358523	67113072	0.557	0.916	5.4	4	10440369	3296099	419711
Mauá	2	274.82	363392	46289950	0.490	0.909	6.6	0	30853650	0	445102
Osasco	1	390.45	652593	100746145	0.522	0.913	5.8	1	3021103	544659	39170
Guarujá	2	308.41	264812	33891820	0.525	0.885	8.4	1	39854124	0	209985
Mogi Guaçu	1	453.33	124228	31535951	0.519	0.886	7.1	1	9455674	536000	576000
Jacaré	1	353.95	191291	46644843	0.528	0.913	6.0	1	20987098	430620	88977
Araçatuba	1	503.17	169254	56088438	0.634	0.909	6.3	2	12576699	5150000	21840
Presidente Venceslau	1	364.66	37347	3685724	0.597	0.893	9.7	1	1911612	790287	995358
Bauru	1	500.27	316064	43704771	0.585	0.908	5.2	1	24287115	172107	180000
Fernandópolis	1	366.50	61647	4260736	0.541	0.891	9.2	2	2241073	288000	82090
Rio de Janeiro	18	596.65	5857904	828097433	0.616	0.933	4.4	4	385126048	450000	0
Paraty	1	312.07	29544	4624791	0.594	0.827	12.8	0	1089385	0	0
Niterói	4	809.18	459451	82892564	0.587	0.960	3.6	7	74000493	0	0
Angra dos Reis	1	275.66	119247	35583431	0.553	0.870	8.9	1	4200714	2764114	92000
S. João do Meretú	1	233.12	449476	32713472	0.475	0.895	5.7	0	12017744	2898377	0
Resende	0	365.45	104549	21094469	0.565	0.918	6.9	2	5826621	1233001	158931
Rio Bonito	0	276.19	49691	2726191	0.574	0.833	12.8	2	2897896	1047412	60718
Duque de Caxias	0	226.14	775456	87343897	0.528	0.873	8.0	0	29444592	220000	50074
Nova Iguaçu	0	237.50	920599	73949921	0.526	0.884	7.2	1	29738159	7758792	39973800
Itaboraí	0	202.29	187479	21249420	0.521	0.884	10.8	0	3783274	0	0
Macaé	0	392.94	132461	17637142	0.562	0.889	7.9	1	31331151	0	0
Nilópolis	0	298.30	153712	7702620	0.479	0.933	5.7	0	2751281	2753797	0
São Gonçalo	0	268.79	891119	58871177	0.489	0.896	5.8	0	21752118	0	0
Saquarema	0	266.64	52461	10272548	0.551	0.848	12.0	1	1482040	2615055	6316115
Maricá	0	321.41	76737	5893003	0.549	0.881	8.7	1	3966433	90000	0

Município	<i>casos_malaria</i>	<i>renda</i>	<i>pop_total</i>	<i>saude_san</i>	<i>gini</i>	<i>idh_ed</i>	<i>tx_analf</i>	<i>medicos</i>	<i>hab_urb</i>	<i>trans_fed</i>	<i>trans_est</i>
Curitiba	7	619.82	1587315	107701036	0.594	0.946	3.4	3	191897099	1708000	2154373
Foz do Iguaçu	12	326.19	258543	26119487	0.582	0.905	7.5	1	17906123	12546	58700
Pranchita	5	254.47	6260	698298	0.584	0.855	11.4	1	44409	70000	103294
Londrina	1	439.35	447065	110208024	0.577	0.910	7.1	2	6284351	95912	1069051
Apucarana	1	288.47	107827	1888614	0.513	0.877	9.6	1	5052465	196600	10000
Umuarama	2	313.76	90690	15110164	0.537	0.898	9.8	1	3463195	390088	386517
Mal. Cândido Rondon	1	341.71	41007	4777760	0.567	0.932	4.3	2	5930630	35063	0
Medianeira	1	277.50	37827	1912808	0.513	0.904	7.2	1	1188248	45051	34564
Arapongas	1	304.07	85428	4527118	0.504	0.883	8.4	1	6570150	677169	204382
Francisco Beltrão	1	307.97	67132	10442384	0.577	0.918	8.4	1	3587493	840827	122642
Paranaгуá	1	305.36	127339	8861291	0.563	0.897	5.9	1	3571266	8492	60295
Mandaguari	1	238.95	31395	1531968	0.502	0.888	11.8	0	160495	0	0
Telêmaco Borba	1	275.60	61238	3612343	0.580	0.865	10.5	1	3611809	651000	76809
Paranaуá	1	312.97	75750	4438090	0.567	0.886	9.8	2	1135285	330720	329757
Missal	1	237.91	10433	1659870	0.549	0.898	7.8	0	936972	255000	0
Porto Alegre	4	709.88	1360590	690428164	0.606	0.951	3.6	6	104207038	496871	41989
Caxias do Sul	2	490.65	360419	71900922	0.511	0.945	3.7	2	112657	177059	510162
Passo Fundo	2	405.65	168458	11830539	0.585	0.912	5.6	3	6102389	15000	236165
Frederico Westphalen	2	297.09	26759	2198041	0.551	0.937	7.6	1	652048	482003	287544
Estrela	1	332.79	27401	1577862	0.469	0.929	3.6	1	696319	198283	220206
Maximiliano de Almeida	0	192.34	5651	579431	0.531	0.883	10.8	1	256593	149641	25000
Torres	2	317.74	30880	2239220	0.531	0.911	6.3	2	1780991	0	342
Ronda Alta	1	245.54	10051	1314708	0.575	0.867	9.4	0	195813	602930	39488
Alegria	2	160.89	5367	930560	0.563	0.854	12.3	0	90857	20000	51000
Bento Gonçalves	1	466.84	91486	6011847	0.473	0.938	3.9	1	2197841	275457	433468
Lajeado	1	395.34	64133	4631319	0.506	0.923	4.1	1	2235892	216630	79871
Ijuí	1	332.05	78461	4752438	0.572	0.926	5.7	1	1603367	710196	327550
Taquara	1	345.44	52825	4285	0.524	0.896	6.8	1	158914	37644	18092
Sapucaia do Sul	0	271.38	122751	6385902	0.464	0.900	5.5	0	2868537	50213	0
São Leopoldo	0	370.06	193547	40757529	0.551	0.922	4.8	1	6275356	400000	967529

```
# Lendo os dados dados=
read.table(file="C:/UFRPE/Malária/dados.txt",header=T)
attach(dados)

## Binomial Negativa

library(MASS)

# Modelo - Nacional malaria.NB<-glm.nb(malaria ~ gini + tx_analf +
trans_fed, link = log) summary(malaria.NB)

qchisq(0.95,malaria.NB$df.residual)

# Valor ajustado mi <- fitted(malaria.NB)

# Resíduos - Componentes do desvio
rd<-residuals(malaria.NB,type="deviance")

# Gráfico do resíduo componente do desvio versus índice das
observações plot(rd, xlab="Índice", ylab="Componentes do desvio",
ylim=c(-2.5,2.5)) abline(-2,0,lty=2, col="red") abline(2,0,lty=2,
col="red") identify(rd, col="blue", label=municipio, cex=0.7)

# Distância de Cook di <- cooks.distance(malaria.NB) plot(di,
xlab="Índice", ylab="Distâncias de Cook", main="") identify(di,
col="blue", label=municipio, cex=0.7)

# Elementos da diagonal da matriz de projeção
p=length(malaria.NB$coef) n=length(malaria)
```

```
h <- lm.influence(malaria.NB)$hat plot(h, xlab="Índice",  
ylab="Diagonal de H", main="") abline((2*p)/n,0,lty=2, col="red")  
identify(h, col="blue", label=municipio, cex=0.7)
```

# Capítulo 8

## Inferência Adicional

### 8.1 Existência, finitude e unicidade dos estimadores dos parâmetros $\beta$

Se a matriz do modelo,  $\mathbf{X}$ , não é de posto completo, as equações (3.5) não admitem solução única. Existem várias soluções dependendo da inversa generalizada da informação  $\mathbf{K}$  mas, para fins estatísticos, a solução particular escolhida é irrelevante, pois as funções estimáveis  $\hat{\eta}$  e  $\hat{\mu}$  são únicas, qualquer que seja a inversa adotada.

Mesmo quando  $\mathbf{X}$  é de posto completo, o sistema (3.5) pode falhar em convergir para alguns MLG e, mesmo quando converge, se o logaritmo da função de verossimilhança  $\ell(\beta)$  tem várias raízes, é importante verificar que o processo converge para um  $\hat{\beta}$  produzindo um máximo global. Se  $\ell(\beta)$  é estritamente côncava como função de  $\beta$ , o máximo *único* está garantido, quando existir pelo menos um.

Outro ponto importante é verificar se existe um  $\hat{\beta}$  *finito* que maximize  $\ell(\beta)$  e se esse se encontra no *interior* de um espaço de parâmetros especificado.

As condições necessárias de unicidade, finitude e existência de um  $\hat{\beta}$  no interior de um espaço especificado são discutidas em Wedderburn (1976) para várias distribuições e funções de ligação. Wedderburn mostra que para alguns MLG a estimativa  $\hat{\beta}$  sempre existe e, para outros, só existe quando não ocorrer degenerescência nos dados. A Tabela 8.1 apresenta seus resultados para os modelos normal, Poisson e gama com função de ligação potência  $\eta = \mu^\lambda$  ( $\lambda = 0$  interpretado como logaritmo) e para o modelo binomial com função de ligação identidade, arcsen  $\sqrt{\cdot}$ , probito, logística e complemento log log. O modelo normal

Tabela 8.1: Propriedades das estimativas de MV dos  $\beta$ 's para várias distribuições e funções de ligação: F - finitude; I - existência no interior do espaço de parâmetros especificado; U - Unicidade.

(a) Modelos com $\mu > 0$ , se $\eta$ não for definido e finito quando $\mu = 0$			
Função de ligação potência $\eta = \mu^\lambda$	Distribuição		
	Normal	Gama	Poisson
$\lambda > -1$	I	FI	FI
$-1 \leq \lambda < 0$	I	FIU	FI
$\lambda = 0$	I	FIU	FIU
$0 < \lambda < 1$	F	FI	FIU
$\lambda = 1$	FU	FI	FIU
$\lambda > 1$	F	FI	FI
(b) Modelo normal com $\mu$ irrestrito			
$\eta = \mu$	FIU		
$\eta = \exp(-\mu)$	FI		
(b) Modelo binomial			
$\eta = \mu, \eta = \arcsen \sqrt{\mu/m}$	IU		
$\eta = \Phi^{-1}(\mu), \eta = \log[\mu/(m - \mu)]$	FIU		
$\eta = \log[-\log(1 - \mu/m)]$	FIU		

com função de ligação  $\eta = \exp(-\mu)$  foi também incluído, pois este ocorre como polinômios inversos (Seção 2.1).

Da Tabela 8.1 conclui-se que no modelo gama com função de ligação potência, excluindo dados iguais a zero, para qualquer  $\lambda$ , os  $\hat{\beta}$ 's são sempre finitos e pertencem ao interior do espaço dos  $\beta$ 's tal que  $g^{-1}(\mathbf{X}\beta) \geq 0$ , e são únicos para  $-1 \leq \lambda \leq 0$ .

## 8.2 Uma família de funções de ligação

Define-se a família de funções de ligação

$$\eta = \int b''(\theta)^\delta d\theta, \quad (8.1)$$

em que  $\delta$  é um número real e  $b''(\theta) > 0$ . Convém lembrar que  $b''(\theta) = \phi^{-1}\text{Var}(Y)$ . Por exemplo, quando  $Y$  tem distribuição de Poisson,  $b(\theta) = \exp(\theta)$ ,  $\theta = \log \mu$  e (8.1) reduz-se à família potência de funções de ligação  $\eta = \mu^\delta$ .

Um estudo completo de (8.1) pode ser encontrado em Vaeth (1985). A Tabela 8.2 apresenta casos importantes dessa família para várias distribuições. A expressão (8.1) tem duas formas equivalentes  $\eta = \int V^\delta d\theta$ , em que  $V = dq^{-1}(\theta)/d\theta$  e  $\eta = \int V^{\delta-1} d\mu$ , sendo

a última a mais usada. A sua grande utilidade é que engloba várias funções de ligação importantes. Para  $\delta = 0$  e  $\delta = 1$ , obtêm-se as funções de ligação canônica e identidade, respectivamente.

Para  $\delta = 1/2$ , tem-se  $(d\eta/d\theta)^2 = V$  que implica  $w = 1$ , sendo a matriz de informação  $\mathbf{K}$  constante e independente de parâmetros desconhecidos. Nesse caso, a matriz de covariância das estimativas de MV dos parâmetros lineares é estabilizada.

Considere o logaritmo da função de verossimilhança obtido de (3.1) e relativo a uma única observação:

$$\log f(y; \theta, \phi) = \phi^{-1}t(y, \theta) + c(y, \phi),$$

com  $t(y, \theta) = y\theta - b(\theta)$  e seja  $\eta = g[q^{-1}(\theta)]$  a transformação unívoca do parâmetro  $\theta$  representando o preditor linear.

Por simples diferenciação, tem-se

$$\frac{\partial^2 t(y, \theta)}{\partial \eta^2} = \frac{\partial^2 t(y, \theta)}{\partial \theta^2} \left( \frac{d\eta}{d\theta} \right)^{-2} \quad (8.2)$$

e

$$\frac{\partial^3 t(y, \theta)}{\partial \eta^3} = \left[ \frac{\partial^3 t(y, \theta)}{\partial \theta^3} - 3 \frac{\partial^2 t(y, \theta)}{\partial \theta^2} \frac{d^2 \eta}{d\theta^2} \frac{d\theta}{d\eta} \right] \left( \frac{d\eta}{d\theta} \right)^{-3}. \quad (8.3)$$

Usando (8.2), verifica-se que a função de ligação  $\delta = 1/2$ , além de estabilizar a covariância de  $\hat{\eta}$ , implica em  $\partial^2 t(y, \theta)/\partial \eta^2$  ser constante, isto é, a curvatura do logaritmo da função de verossimilhança  $\ell(\beta)$  é, também, estabilizada. Para as distribuições de Poisson, gama, binomial e normal inversa, as funções de ligação que produzem informação constante são:  $\eta = \sqrt{\mu}$ ,  $\eta = \log \mu$ ,  $\eta = \arcsen \sqrt{\mu/m}$  e  $\eta = \mu^{-1/2}$ , respectivamente. As transformações  $\sqrt{Y}$  para dados de Poisson e  $\arcsen \sqrt{Y/m}$  para dados binomiais foram estudadas por Box e Tiao (1973). Bartlett (1947) introduziu o conceito de estabilização de variância e mostrou as vantagens de realizar uma análise de variância com dados transformados satisfazendo essa propriedade.

Apresenta-se agora a função de ligação correspondente a  $\delta = 1/3$ . Tem-se  $t''(y, \theta) = -V$ ,  $d(d\eta/d\theta)/d\eta = d^2 \eta/d\theta^2 d\theta/d\eta$  e  $(3V)^{-1} dV/d\theta = d(d\eta/d\theta)/d\eta$ . Logo, a derivada de terceira ordem do logaritmo da função de verossimilhança  $\partial^3 t(y, \theta)/\partial \eta^3$  em (8.3) é igual a zero. Essa derivada representa o termo de primeira ordem do terceiro

Tabela 8.2: Membros importantes da família  $\eta = \int b''(\theta)^\delta d\theta$ .

Distribuição	$\delta$			
	2/3	1/2	1/3	0
Binomial	$\int [\mu(1 - \frac{\mu}{m})]^{-1/3} d\mu$	$\arcsen \sqrt{\frac{\mu}{m}}$	$\int [\mu(1 - \frac{\mu}{m})]^{-2/3} d\mu$	$\log[\mu/(1 - \frac{\mu}{m})]$
Poisson	$\mu^{2/3}$	$\sqrt{\mu}$	$\mu^{1/3}$	$\log \mu$
Gama	$\mu^{1/3}$	$\log \mu$	$\mu^{-1/3}$	$\mu^{-1}$
Normal	$\mu$	$\mu$	$\mu$	$\mu$
Normal inversa	$\log \mu$	$\mu^{-1/2}$	$\mu^{-1}$	$\mu^{-2}$
Observação: $\delta = 2/3$ - normaliza a distribuição de $\hat{\beta}$ ,				
$\delta = 1/2$ - estabiliza a informação $\mathbf{K}$ e a curvatura de $\ell(\beta)$ ,				
$\delta = 1/3$ - simetriza $\ell(\beta)$ no ponto $\hat{\beta}$ ,				
$\delta = 0$ - função de ligação canônica				

momento central de  $t(y, \theta)$ , parametrizado em termos de  $\eta$  e, portanto, essa função de ligação torna o coeficiente de assimetria do logaritmo da função de verossimilhança  $\ell(\beta)$ , aproximadamente, igual a zero.

Isso é facilmente visto da expansão de  $\ell(\beta)$  em série de Taylor ao redor de  $\hat{\beta}$ , pois  $\ell(\beta)$  só será uma função quadrática em  $\beta$  se todas as suas derivadas parciais, em relação a  $\beta$ , de ordem superior à segunda, forem identicamente iguais a zero. Como

$$\frac{\partial^3 \ell(\beta)}{\partial \beta_r \partial \beta_s \partial \beta_t} = \frac{1}{\phi} \sum_{i=1}^n \frac{\partial^3 t(y_i, \theta_i)}{\partial \eta_i^3} x_{ir} x_{is} x_{it}$$

é a primeira contribuição para a não-quadraticidade, espera-se que  $\eta = \int V^{1/3} d\theta$  simetrize  $\ell(\beta)$  localmente no seu ponto máximo  $\hat{\beta}$ . Para os modelos de Poisson, gama, binomial e normal inverso, essas funções de ligação são  $\eta = \mu^{1/3}$ ,  $\eta = \mu^{-1/3}$ ,  $\int [\mu(1 - \mu/m)]^{-2/3} d\mu$  e  $\eta = \mu^{-1}$ , respectivamente. Nos casos dos modelos de Poisson e gama, as transformações foram propostas por Anscombe (1964). Para a distribuição binomial, o inconveniente é que a função de ligação não pode ser dada em forma explícita.

Finalmente, pode-se demonstrar que a função de ligação  $\int b''(\theta)^{2/3} d\theta$  normaliza a distribuição de  $\hat{\beta}$ , tornando o seu coeficiente de assimetria, aproximadamente, nulo



(Barndorff-Nielsen, 1978, Seção 9.8). Para as distribuições de Poisson, gama, binomial e normal inversa, as funções de ligação com  $\delta = 2/3$  correspondem a  $\eta = \mu^{2/3}$ ,  $\eta = \mu^{1/3}$ ,  $\eta = \int [\mu(1 - \mu/m)]^{-1/3} d\mu$  e  $\eta = \log \mu$ , respectivamente. O caso binomial foi primeiramente proposto por Cox e Snell (1968).

### 8.3 Identificação de observações não explicadas pelo modelo

Considere que a análise dos resíduos mostrou que um modelo usual  $g(\mu) = \mathbf{X}\beta$  não se ajustou bem às observações  $\mathbf{y}_q = (y_{i_1}, \dots, y_{i_q})^T$ . Seja a partição do vetor de observações  $\mathbf{y} = (\mathbf{y}_q^T \ \mathbf{y}_{n-q}^T)^T$  com correspondentes partições da matriz do modelo  $\mathbf{X} = [\mathbf{X}_q \ \mathbf{X}_{n-q}]$ , do vetor de médias  $\boldsymbol{\mu} = (\boldsymbol{\mu}_q^T \ \boldsymbol{\mu}_{n-q}^T)^T$  em que  $\boldsymbol{\mu}_q = E(\mathbf{Y}_q)$  e  $\boldsymbol{\mu}_{n-q} = E(\mathbf{Y}_{n-q})$  e da matriz  $\mathbf{W} = (\mathbf{W}_q \ \mathbf{W}_{n-q})$ . As equações de máxima verossimilhança baseadas nos dados  $\mathbf{y}_{n-q}$  podem ser escritas como

$$\mathbf{X}_{n-q}^T \widetilde{\mathbf{W}}_{n-q} (\mathbf{y}_{n-q} - \tilde{\boldsymbol{\mu}}_{n-q}) = \mathbf{0}. \quad (8.4)$$

Seja  $g(\boldsymbol{\mu}) = \mathbf{X}\beta + \mathbf{Z}\gamma$  um modelo aumentado com todas as observações  $\mathbf{y}$ , sendo  $\mathbf{Z} = (\mathbf{z}_{i_1}, \dots, \mathbf{z}_{i_q})$  uma matriz de covariáveis adicionais e  $\mathbf{z}_{i_j}$  um vetor de ordem  $n$  com elementos iguais a zero exceto no componente correspondente à observação  $y_{i_j}$  mal ajustada pelo modelo para  $j = 1, \dots, q$ . As equações de máxima verossimilhança para o modelo aumentado são dadas por:

$$\begin{bmatrix} \mathbf{X}^T \\ \mathbf{Z}^T \end{bmatrix} \widehat{\mathbf{W}} (\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}. \quad (8.5)$$

Considerando as últimas  $q$  equações tem-se de (8.5)  $\hat{\mu}_{i_j} = y_{i_j}$ ,  $j = 1, \dots, q$ , isto é, a EMV da média da observação  $y_{i_j}$  mal ajustada pelo modelo é igual à própria observação. Substituindo nas  $p$  primeiras equações, constata-se que as estimativas  $\tilde{\boldsymbol{\mu}}_{n-q}$  satisfazem equações idênticas a (8.4). Como  $\mathbf{Z}\gamma$  não contribui para os preditores lineares em  $\boldsymbol{\eta}_{n-q}$ , as estimativas  $\tilde{\boldsymbol{\mu}}_{n-q}$  e  $\hat{\boldsymbol{\mu}}_{n-q}$  são iguais.

Assim, a eliminação de um conjunto de pontos mal ajustados pelo modelo, equivale a aumentar a matriz modelo com uma matriz adequada. Tem-se  $\hat{\boldsymbol{\eta}}_{n-q} = \mathbf{X}_{n-q}\hat{\boldsymbol{\beta}}$ ,  $\hat{\boldsymbol{\mu}}_{n-q} = g^{-1}(\hat{\boldsymbol{\eta}}_{n-q})$ ,  $\hat{\boldsymbol{\mu}}_q = \mathbf{y}_q$ ,  $\hat{\boldsymbol{\eta}}_q = g(\mathbf{y}_q)$  e  $\hat{\boldsymbol{\gamma}} = g(\mathbf{y}_q) - \mathbf{X}_q\hat{\boldsymbol{\beta}}$ .

A contribuição no logaritmo da função de verossimilhança de cada observação  $y_{i_j}$  mal ajustada pelo modelo, pode ser computada como o aumento no desvio resultante da eliminação da covariável  $z_{i_j}$  no modelo ampliado. Os vetores  $\mathbf{z}_{i_j}$ , que produzem aumentos significativos no desvio, correspondem às observações aberrantes do modelo original.

## 8.4 Identificação de observações não explicadas pelo modelo

Esta Ssecção é baseada em Pregibon (1979) que apresenta um estudo completo da verificação do modelo por meio da análise detalhada das observações. Contudo, algumas demonstrações foram estendidas.

Na Seção 4.9 foi visto o método das covariáveis adicionais, que proporciona um diagnóstico do componente sistemático do modelo. Esse método pode, ainda, fornecer uma ação corretiva das formas das covariáveis omitidas ou da função de ligação do modelo.

Em situações complexas, em que um gráfico da variável resposta *versus* as covariáveis no modelo, não pode ser construído, a diferença entre os desvios do modelo original e do modelo aumentado é o único guia para testar a validade do componente sistemático. Entretanto, isso não representa sempre um bom teste, pois um pequeno número  $q$  ( $q \ll n$ ) de observações pode apresentar grandes desvios e direcionar a uma alteração substancial no componente sistemático, enquanto as outras  $n - q$  observações podem suportar, adequadamente, o modelo corrente. Por exemplo, numa tabela de contingência de duas entradas, a interação entre linhas e colunas pode ser devida a umas poucas celas isoladas e, sem as quais, a aditividade é apropriada. Em muitos casos, a identificação das celas da tabela, para as quais a aditividade não é satisfeita, torna-se mais importante do que o próprio ajuste do modelo.

Conforme visto na Seção 5.3, no modelo normal linear, os elementos da matriz de projeção  $\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$  e os resíduos  $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$  são as quantidades básicas para se detectarem as observações *influentes* e as observações *aberrantes*, respectivamente.

O vetor  $\mathbf{r}$  é importante para identificar pontos aberrantes, mas não é útil para determinar as observações que influenciam, indevidamente, o ajuste do modelo. Em geral, resíduos pequenos passam despercebidos, embora possam influenciar bastante o ajuste do modelo.

Os elementos da diagonal da matriz  $\mathbf{I} - \mathbf{H}$ ,  $1 - h_{ii}$ ,  $i = 1, \dots, n$ , são usados para detectar os pontos influentes. Esses correspondem aos menores valores dos  $1 - h_{ii}$ 's, por exemplo, valores inferiores à média  $\text{tr}(\mathbf{I} - \mathbf{H})/n = 1 - p/n$ . Na prática, aceita-se a  $i$ -ésima observação como influente, se  $h_{ii} > 2p/n$ . A determinação dos pontos influentes pode ser feita nos gráficos de  $1 - h_{ii}$  versus  $i$  ou versus os valores ajustados  $\hat{\mu}_i$ . Os elementos fora da diagonal de  $\mathbf{I} - \mathbf{H}$  são usados quando se consideram efeitos conjuntos de várias observações sobre o ajuste do modelo.

### 8.4.1 Pontos influentes no modelo linear generalizado

A matriz de projeção para o modelo linear generalizado, já vista na Seção 5.4.1, é dada por  $\mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{Z}\mathbf{W}^{1/2}$ , com  $\mathbf{Z} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$ ;  $\mathbf{I} - \mathbf{H}$  é simétrica e idempotente. Usa-se a notação  $\mathbf{r}^P = (r_1^P, \dots, r_n^P)^T$  para representar os valores observados dos resíduos de Pearson em (5.3).

Demonstra-se, agora, que esses resíduos estão situados no subespaço gerado pelas colunas da matriz  $\mathbf{I} - \mathbf{H}$ , isto é,  $\mathbf{r}_P = (\mathbf{I} - \hat{\mathbf{H}})\mathbf{r}^P$ , em que  $\mathbf{I} - \hat{\mathbf{H}}$  representa a EMV de  $\mathbf{I} - \mathbf{H}$ . Os elementos de  $\mathbf{I} - \mathbf{H}$  são funções das médias e da matriz modelo e, portanto, devem ser estimados em  $\hat{\boldsymbol{\beta}}$ . Somente no caso do modelo normal linear é que  $\mathbf{I} - \mathbf{H}$  não depende de  $\boldsymbol{\mu}$ . A estimativa de MV de  $\boldsymbol{\beta}$  na convergência é  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}\hat{\mathbf{z}}$ , em que  $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$ . Pré-multiplicando por  $\mathbf{X}$  e subtraindo de  $\mathbf{z}$  vem  $\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \hat{\mathbf{Z}}\hat{\mathbf{W}})\hat{\mathbf{z}}$ . Logo,  $\hat{\mathbf{W}}^{-1/2}(\mathbf{I} - \hat{\mathbf{H}})\hat{\mathbf{W}}^{1/2}\hat{\mathbf{z}} = \hat{\mathbf{W}}^{-1/2}(\mathbf{I} - \hat{\mathbf{H}})\mathbf{r}^P = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{V}}^{-1/2}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ , pois  $\mathbf{G} = \mathbf{W}^{-1/2}\mathbf{V}^{-1/2}$ , sendo  $\mathbf{V} = \text{diag}\{V_1, \dots, V_n\}$ . Pré-multiplicando por  $\mathbf{W}^{1/2}$  obtém-se o resultado desejado, isto é,  $\mathbf{r}_P = (\mathbf{I} - \hat{\mathbf{H}})\mathbf{r}^P$ .

Analogamente ao modelo normal linear, os valores das estimativas dos elementos da diagonal de  $\mathbf{I} - \mathbf{H}$  menores do que  $1 - 2p/n$ , indicam quais os pontos influentes do MLG. Os  $\hat{h}_{ii}$ 's dependem fortemente da matriz modelo e, em geral, muito pouco da estimativa de  $\boldsymbol{\mu}$ . Os gráficos de  $\hat{h}_{ii}$ 's versus  $i$  ou  $\hat{\mu}_i$  são, geralmente, usados para visualizar os pontos influentes.

### 8.4.2 Diagnóstico local de um único ponto influente

As quantidades  $r_i^P$  e  $\hat{h}_{ii}$ ,  $i = 1, \dots, n$ , da seção anterior, embora possibilitem verificar quais as observações que não são bem explicadas pelo modelo e aquelas que dominam, em grande parte, o ajuste, não podem medir o efeito sobre todos os componentes do modelo ajustado. Nesta seção, trata-se de determinar, matematicamente, o efeito local de cada observação sobre as estimativas dos parâmetros. Utiliza-se um método simples de perturbação do modelo corrente através de um escalar, que possibilita o estudo dos efeitos de pontos individuais sobre o ajuste do modelo. Na Seção 8.4.4, esse método é generalizado para determinar os efeitos de várias observações, simultaneamente, sobre o ajuste do modelo.

Para cada observação define-se a função  $t_l : t_l = t (0 \leq t \leq 1)$  se  $l = i$  e  $t_l = 1$  se  $l \neq i$  que conduz a um MLG perturbado. Aqui, planeja-se estudar o efeito local da  $i$ -ésima observação sobre as estimativas dos parâmetros  $\beta$ 's do MLG. Para  $t = 1$  e  $t = 0$ , são definidos o modelo usual e o modelo com a omissão da  $i$ -ésima observação, respectivamente. Seja a matriz diagonal  $\mathbf{T} = \text{diag}\{t_1, \dots, t_n\}$ . No contexto dos MLG, essa matriz  $\mathbf{T}$  pode ser especificada como formada por pesos a priori.

O logaritmo da função de verossimilhança para  $\beta$  dado  $\mathbf{t} = (t_1, t_2, \dots, t_n)^T$  é  $l_t(\beta) = \phi^{-1} \sum_{i=1}^n t_i [y_i \theta_i - b(\theta_i)] + \sum_{i=1}^n c(y_i, \phi)$ . Tem-se a função escore  $\mathbf{U}_t = \partial l_t(\beta) / \partial \beta = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{T} \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$  e a matriz de informação  $\mathbf{K}_t = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{T} \mathbf{X}$ . O processo para calcular a EMV  $\hat{\beta}_t$  de  $\beta$  é

$$\beta_t^{(m+1)} = (\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{T} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{T} \mathbf{z}^{(m)}. \quad (8.6)$$

Para  $t = 1$  o processo (8.6) equivale a calcular a EMV usual  $\hat{\beta}$ , isto é,  $\hat{\beta}_1 = \hat{\beta}$ . A diferença  $\hat{\beta}_t - \hat{\beta}$  proporciona uma medida de efeito local da  $i$ -ésima observação, como função de  $\mathbf{t}$ , sobre as estimativas dos  $\beta$ 's. Se essa diferença se tornar bastante reduzida quanto  $\mathbf{t}$  decrescer, então a  $i$ -ésima observação exercerá pouca influência sobre as estimativas dos parâmetros e, portanto, sobre o próprio ajuste do modelo. Se por outro lado, pequenas alterações de  $\mathbf{t}$  implicam em grandes diferenças de  $\hat{\beta}_t - \hat{\beta}$ , o  $i$ -ésimo ponto será influente e, pode-se então, determinar quais os componentes de  $\hat{\beta}$  que são mais instáveis em relação a esse ponto.

Na prática, essa diferença pode ser calculada em qualquer “software” de ajuste

dos MLG sem muitas dificuldades, para um conjunto de pontos que possam ser considerados potencialmente influentes. A única tarefa é ajustar  $r + 1$  modelos, caso haja  $r$  pontos a se pesquisar. Apresenta-se, agora, um método alternativo, aproximado, para computar a diferença  $\hat{\beta}_t - \beta$ , a partir do ajuste de um único modelo: o modelo corrente.

O processo (8.6) é obtido da aproximação

$$\beta_t^{(m+1)} - \beta_t^{(m)} = \mathbf{K}_t^{-1(m)} \mathbf{U}_t^{(m)}. \quad (8.7)$$

Admite-se a aproximação de um passo  $\beta_t^{(1)}$  para  $\hat{\beta}_t$ , iniciando o processo (8.7) na EMV usual  $\hat{\beta} = \hat{\beta}_1$ . A função escore  $\mathbf{U}_t$  no ponto  $\hat{\beta}$  é dada por  $\phi^{-1}(t-1)w_i g_i(y_i - \mu_i)\mathbf{x}_i$ , em que  $\mathbf{x}_i^T$  é um vetor linha de  $\mathbf{X}$  e  $w_i$  e  $g_i$  são elementos de  $\mathbf{W}$  e  $\mathbf{G}$ , todos correspondentes à  $i$ -ésima observação. Tem-se

$$(\mathbf{X}^T \mathbf{W} \mathbf{T} \mathbf{X})^{-1} = [\mathbf{X}^T \mathbf{W} \mathbf{X} - (1-t)w_i \mathbf{x}_i \mathbf{x}_i^T]^{-1}.$$

Para calcular a inversa de  $\mathbf{K}_t$ , usa-se o seguinte teorema da álgebra linear: se  $\mathbf{M}$  é uma matriz simétrica  $p \times p$  e  $\mathbf{A}$  e  $\mathbf{B}$  são matrizes  $q \times p$  então:

$$(\mathbf{M} + \mathbf{A}^T \mathbf{B})^{-1} = \mathbf{M}^{-1} - \mathbf{M}^{-1} \mathbf{A}^T (\mathbf{I} + \mathbf{B} \mathbf{M}^{-1} \mathbf{A}^T) \mathbf{B} \mathbf{M}^{-1}.$$

Assim, obtém-se, facilmente, sem o multiplicador  $\phi^{-1}$ , a matriz de informação do MLG perturbado

$$\mathbf{K}_t^{-1} = \mathbf{K}^{-1} + \frac{(1-t)w_i}{1 + (t-1)w_i z_{ii}} \mathbf{K}^{-1} \mathbf{x}_i \mathbf{x}_i^T \mathbf{K}^{-1}, \quad (8.8)$$

em que  $\mathbf{K} = \mathbf{K}_1$  é a matriz de informação definida na Seção 3.2 sem o multiplicador  $\phi^{-1}$ , e  $z_{ii}$  é o  $i$ -ésimo elemento da diagonal de  $\mathbf{Z}$ .

Logo, levando a função escore  $\mathbf{U}_t$  e (8.8), ambos avaliados em  $\hat{\beta}$ , na expressão (8.7), obtém-se, com alguma álgebra,

$$\beta_t^{(1)} - \hat{\beta} = \frac{(t-1)\hat{w}_i \hat{h}_{ii}(y_i - \hat{\mu}_i)}{1 + (t-1)\hat{w}_i \hat{z}_{ii}} \hat{\mathbf{K}}^{-1} \mathbf{x}_i, \quad (8.9)$$

em que todas as quantidades estão estimadas em  $\hat{\beta}$ .

A grande vantagem dessa expressão aproximada é que pode ser calculada inteiramente a partir do ajuste do modelo corrente ajustado. A influência do  $i$ -ésimo ponto nas estimativas dos parâmetros é estudada computando o lado direito de (8.9) para alguns

valores de  $t$  em  $[0, 1]$ . O gráfico de cada componente de  $\beta_t^{(1)}$  versus  $t$  dará uma idéia da taxa de variação dessa estimativa como função de  $t$ . De  $w_i z_{ii} = 1 - h_{ii}$ ,  $g_i = w_i^{-1/2} V_i^{-1/2}$ ,  $V_i^{-1/2}(y_i - \mu_i) = r_i^P$ , obtém-se uma expressão mais simples do que (8.9)

$$\beta_t^{(1)} - \hat{\beta} = \frac{(t-1)\hat{w}_i^{1/2} r_i^P}{1 + (t-1)(1 - \hat{h}_{ii})} \hat{\mathbf{K}}^{-1} \mathbf{x}_i, \quad (8.10)$$

A taxa de variação de  $\beta_t^{(1)}$  com  $t$ ,  $\frac{\partial \beta_t^{(1)}}{\partial t} = \dot{\beta}_t^{(1)}$ , é dada por

$$\dot{\beta}_t^{(1)} = \frac{\hat{w}_i^{1/2} r_i^P}{[1 + (t-1)(1 - \hat{h}_{ii})]^2} \hat{\mathbf{K}}^{-1} \mathbf{x}_i, \quad (8.11)$$

O valor de  $\hat{\beta}_t^{(1)}$  no ponto  $t = 1$  descreve a alteração local que  $\beta_t^{(1)}$  sofre na estimativa  $\hat{\beta}$  e, no ponto  $t = 0$ , mede a variação de  $\beta_t^{(1)}$  na EMV de  $\beta$ , calculada sem a  $i$ -ésima observação. O valor de  $\dot{\beta}_t^{(1)}$  no ponto  $t = 0$  é bastante dependente de pontos aberrantes e de covariáveis extremas do modelo. Observar que o denominador de  $\dot{\beta}_t^{(1)}$  no ponto  $t = 0$  é  $\hat{h}_{ii}^2$  e, portanto, valores pequenos de  $\hat{h}_{ii}^2$  indicarão pontos bastante influentes, como visto na Seção 8.4.1.

A obtenção dos gráficos de cada componente de  $\beta_t^{(1)}$  versus  $t$ , no intervalo unitário, para todo ponto de interesse, é um processo trabalhoso, e pode-se usar a diferença

$$\Delta \beta^{(1)} = \beta_1^{(1)} - \beta_0^{(1)} = \frac{\hat{w}_i^{1/2} r_i^P}{\hat{h}_{ii}} \hat{\mathbf{K}}^{-1} \mathbf{x}_i. \quad (8.12)$$

A fórmula (8.12) é computacionalmente mais simples, como uma medida de diagnóstico da influência local da observação  $i$  sobre as diversas estimativas dos parâmetros  $\beta'$ s. Essa diferença mede uma discrepância na EMV de  $\beta$  devido à omissão da  $i$ -ésima observação.

Seja  $\hat{\beta}_r$  uma estimativa de interesse e  $\rho_r^T \hat{\mathbf{K}}^{-1}$  a linha correspondente da matriz  $\hat{\mathbf{K}}^{-1}$ , em que  $\rho_r^T$  é um vetor  $1 \times p$  de zeros com 1 no  $r$ -ésimo componente. O gráfico de

$$\frac{\Delta \beta_r^{(1)}}{\text{Var}(\hat{\beta}_r)^{1/2}} = - \frac{\hat{w}_i^{1/2} r_i^P}{\hat{h}_{ii} \hat{\kappa}^{rr/2}} \rho_r^T \hat{\mathbf{K}}^{-1} \mathbf{x}_i$$

versus  $i$ , denominado *curva empírica de influência* sobre a estimativa  $\hat{\beta}_r$ , é o mais usado para detectar as observações que causam instabilidade local nas estimativas de interesse, sendo  $\text{Var}(\hat{\beta}_r) = -\kappa^{rr}$ . Geralmente, esses gráficos discriminam que os mesmos dados

causam as maiores instabilidades em todas as estimativas. Entretanto, as alterações nos valores das estimativas  $\hat{\beta}'s$  podem se compensar de tal maneira que os valores ajustados variam muito pouco. Nesse caso, outros diagnósticos mais globais devem ser usados e os mesmos são baseados na razão de verossimilhanças, no desvio ou na estatística de Pearson generalizada.

### 8.4.3 Diagnóstico global de um único ponto influente

A diferença entre os máximos do logaritmo da função de verossimilhança para  $t$  especificado e para  $t = 1$ ,  $a_t = 2\{l_t(\hat{\beta}) - l(\hat{\beta})\}$ , representa uma *medida de influência global* da  $i$ -ésima observação sobre o ajuste do modelo. Embora, o cálculo de  $a_t$  não apresente dificuldades, obtém-se uma aproximação expandindo  $a_t$  em série de Taylor e substituindo  $\hat{\beta}_t$  por  $\beta_t^{(1)}$ , dado em (8.10),

$$a_t^{(1)} = \frac{(t-1)^2(1-\hat{h}_{ii})r_i^{P^2}}{[1+(t-1)(1-\hat{h}_{ii})]^2}. \quad (8.13)$$

Deve-se notar que  $a_t^{(1)}$  está expresso em termos dos elementos da diagonal da matriz de projeção  $\mathbf{I} - \mathbf{H}$  e dos resíduos de Pearson  $r_i^P$  e pode ser computado do ajuste usual.

A diferença  $a_0^{(1)} - a_1^{(1)} = (1-\hat{h}_{ii})r_i^{P^2}/\hat{h}_{ii}^2$  mede uma influência global sobre o ajuste do modelo, ao eliminar o  $i$ -ésimo ponto. Exceto por um coeficiente multiplicador,  $a_0^{(1)}$  é a medida de influência, relativa ao modelo linear, introduzida por Cook (1977). O gráfico de  $(1-\hat{h}_{ii})r_i^{P^2}/\hat{h}_{ii}^2$  versus  $i$  é muito simples de ser obtido e determina, quase sempre, as observações que influenciam, indevidamente, o ajuste global do modelo.

Diferenciando  $a_t^{(1)}$  em relação a  $t$  vem

$$a_t^{(1)} = \frac{-2(t-1)r_i^{P^2}}{[1+(1-t)(1-\hat{h}_{ii})]^3}. \quad (8.14)$$

Observe-se que  $a_t^{(1)}$  é uma função decrescente de  $t$  e isto, coincide, com a interpretação intuitiva de  $a_t$ .

Sejam  $S_p(t)$  e  $X_p^2(t)$  o desvio e a estatística de Pearson generalizada, supondo que a observação  $i$  está multiplicada pelo escalar  $t$ . Um diagnóstico global, alternativo, da influência da observação  $i$  sobre o ajuste do modelo, pode ser determinado calculando a

variação nas estatísticas  $S_p(t)$  e  $X_p^2(t)$  em função de  $t$ . A eliminação de pontos que sejam bastante influentes no modelo, provavelmente, alterará muito os valores dessas estatísticas. Com a exclusão de observações, o desvio sempre decresce embora isso não aconteça com a estatística de Pearson generalizada. Para ver isso, obtém-se da fórmula do desvio (4.6), sem o parâmetro de dispersão  $\phi$ ,

$$\frac{\partial S_p(t)}{\partial t} = 2 \left( y_i \{q(y_i) - q[g^{-1}(x_i \beta_t)]\} + b \{q[g^{-1}(x_i \hat{\beta}_t)]\} - b[q(y_i)] \right).$$

Assim,  $\partial S_p(t)/\partial t$  é sempre não-negativo indicando que o desvio  $S_p(t)$  cresce quando  $t$  cresce de 0 a 1.

As estatísticas  $b_i = S_p(1) - S_p(0)$  e  $c_i = X_p^2(1) - X_p^2(0)$  medem as variações respectivas no desvio e na estatística de Pearson generalizada, ao se eliminar a  $i$ -ésima observação do modelo. A quantidade  $b_i(c_i)$  representa a declividade da reta ligando  $S_p(0)(X_p^2(0))$  a  $S_p(1)(X_p^2(1))$ .

As estatísticas  $b_i$  e  $c_i$  são as mais usadas para diagnosticar, globalmente, a influência do  $i$ -ésimo ponto sobre o ajuste do modelo. Expressões aproximadas para essas quantidades são deduzidas de expansões em séries de Taylor (Pregibon, 1979) e podem ser escritas como

$$b_i = r_i^{D^2} + r_i^{P^2} \frac{(1 - \hat{h}_{ii})}{\hat{h}_{ii}} \quad (8.15)$$

e

$$c_i = \frac{r_i^{P^2}}{\hat{h}_{ii}}, \quad (8.16)$$

em que  $r_i^D$  é o resíduo definido a partir do componente do desvio em (5.6). Assim,  $b_i$  é uma média ponderada dos quadrados dos resíduos  $r_i^D$  e  $r_i^P$ .

Os gráficos de  $b_i$  e  $c_i$  versus  $i$  indicarão as observações influentes sobre o ajuste como um todo. Assim, todas as medidas para um diagnóstico global da influência dos dados sobre o ajuste do modelo são funções dos resíduos e das estimativas dos elementos da diagonal da matriz de projeção.



### 8.4.4 Diagnóstico local e global da influência de um conjunto de pontos

Aqui, os conceitos descritos nas Seções 8.4.2 e 8.4.3 são estendidos para diagnosticar a influência de um conjunto de pontos especificados, sobre as estimativas e o ajuste do modelo. Pode ocorrer que vários pontos controlem, simultaneamente, um aspecto importante do ajuste do modelo, embora os pontos isoladamente, não apresentem influências significativas, locais ou globais, no ajuste. As fórmulas apresentadas nesta seção podem ser obtidas de maneira equivalente às deduções feitas nas duas seções anteriores e são deixadas para o leitor.

O interesse é verificar a influência simultânea de um conjunto  $I$  de observações sobre o ajuste. Define-se a função  $T_I$  por:  $t_I = t$  se  $l \in I$  e  $t_I = 1$  se  $l \notin I$ . Para simplificar a notação, qualquer quantidade com o índice  $I$  representa a parte dessa quantidade correspondente aos dados em  $I$ .

As fórmulas (8.10), (8.11) e (8.12) são generalizadas para diagnosticar localmente a influência das observações em  $I$  sobre as estimativas dos parâmetros:

$$\beta_t^{(1)} - \hat{\beta} = (t - 1)\hat{\mathbf{K}}^{-1}\mathbf{X}_I^T\hat{\mathbf{W}}_I^{1/2}[\mathbf{I} + (t - 1)(\mathbf{I} - \hat{\mathbf{H}}_I)]^{-1}\mathbf{r}_I^P, \quad (8.17)$$

$$\Delta\beta_t^{(1)} = \hat{\mathbf{K}}^{-1}\mathbf{X}_I^T\hat{\mathbf{W}}_I^{1/2}\hat{\mathbf{H}}_I^{-1}\mathbf{r}_I^P, \quad (8.18)$$

$$\dot{\beta}_t^{(1)} = \hat{\mathbf{K}}^{-1}\mathbf{X}_I^T\hat{\mathbf{W}}_I^{1/2}[\mathbf{I} + (t - 1)(\mathbf{I} - \hat{\mathbf{H}}_I)]^{-2}\mathbf{r}_I^P. \quad (8.19)$$

A curva empírica de influência dos pontos em  $I$  sobre a estimativa  $\hat{\beta}_r$  é obtida de (8.18).

As estatísticas usadas para um diagnóstico global da influência dos pontos em  $I$  sobre o ajuste do modelo são dadas por

$$a_0^{(1)} = \mathbf{r}_i^{P^T}\hat{\mathbf{H}}_I^{-1}(\mathbf{I} - \hat{\mathbf{H}}_I)\hat{\mathbf{H}}_I^{-1}\mathbf{r}_I^P, \quad (8.20)$$

$$b_I = \mathbf{r}_I^{D^T}\mathbf{r}_I^D + \mathbf{r}_I^P\hat{\mathbf{H}}_I^{-1}(\mathbf{I} - \hat{\mathbf{H}}_I)\mathbf{r}_I^P \quad (8.21)$$

e

$$c_I = \mathbf{r}_I^{P^T}\hat{\mathbf{H}}_I^{-1}\mathbf{r}_I^P. \quad (8.22)$$

Aqui,  $\mathbf{r}_I^P$  e  $\mathbf{r}_I^D$  são os vetores dos resíduos de Pearson e como componentes do desvio, definidos em (5.3) e (5.6), correspondentes às observações em  $I$ .

Todas estas estatísticas medem a influência sobre o ajuste ao eliminar os pontos em  $I$ . Se  $I$  é um conjunto unitário obtém-se de (8.20) a (8.22), como casos especiais, as fórmulas correspondentes, citadas nas Seções 8.4.2 e 8.4.3. A estatística  $b_I$  é mais sensível do que  $c_I$  em relação às alterações do ajuste do modelo e, portanto, deve ser preferida. Se o conjunto de pontos  $I$  é um subconjunto de  $I^*$ , então  $b_I \leq b_{I^*}$ .

É importante notar as analogias entre os diagnósticos de um único ponto influente e de um conjunto de pontos influentes, sobre o ajuste do modelo. Claro está que a dimensão das matrizes envolvidas para o diagnóstico da influência de um conjunto de pontos é igual à cardinalidade desse conjunto. O número de conjuntos, possivelmente, influentes pode ser bastante grande e, na prática, o diagnóstico, fatalmente, depende de um programa de computador, que calcule as estatísticas (8.20) a (8.22), somente, para os conjuntos mais promissores à influência. Em geral, o interesse é identificar aqueles conjuntos que implicam em valores grandes dessas estatísticas, para o número de pontos em  $I$  variando de 1 a um valor máximo estipulado. Um algoritmo do tipo “branch and bound” é, geralmente, usado para definir os conjuntos de pontos influentes por meio de uma regra de decisão, por exemplo: um subconjunto  $I$  de um conjunto  $I^*$  será considerado não-influente quando  $b_I$  for estritamente menor do que uma fração especificada de  $b_{I^*}$ , em que  $|J| = |I^*|$ . Para considerações adicionais consultar Pregibon (1979).

## 8.5 Teste de hipóteses com restrições

Numa situação freqüente, a hipótese nula em um MLG define  $q$  combinações lineares dos  $\beta$ 's iguais a zero,  $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ , em que  $\mathbf{A}$  é uma matriz de dimensões  $q \times p$  de posto  $q < p$ . A determinação de  $\mathbf{A}$  pode ser feita pela estrutura geral dos dados ou por objetivos próprios. Seja  $\tilde{\boldsymbol{\beta}}$  a EMV de  $\boldsymbol{\beta}$  segundo  $H_0$ . O algoritmo de estimação dos MLG pode ser usado para calcular a EMV  $\tilde{\boldsymbol{\beta}}$  restrita a  $H_0$ , expressando a hipótese  $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$  na forma da estrutura linear usual de um MLG.

É sempre possível achar uma matriz  $\mathbf{M}$ , de dimensões  $p \times (p - q)$ , tal que o espaço de colunas de  $\mathbf{M}$  é o complemento ortogonal do espaço de colunas de  $\mathbf{A}^T$ , no espaço

Euclidiano de dimensão  $p$ . A determinação de  $\mathbf{M}$  pode ser realizada através do processo de decomposição de Gram-Schmidt, embora, muitas vezes, seja difícil e trabalhosa.

O posto de  $(\mathbf{A}^T \mathbf{M})$  é  $p$  e escreve-se  $\boldsymbol{\beta} = \mathbf{A}^T \boldsymbol{\delta}_q + \mathbf{M} \boldsymbol{\delta}_{p-q}$ , sendo  $\boldsymbol{\delta}_q$  e  $\boldsymbol{\delta}_{p-q}$  vetores colunas de dimensões  $q$  e  $p - q$ , respectivamente. De  $\mathbf{A} \mathbf{M} = \mathbf{0}$  tem-se  $\mathbf{A} \boldsymbol{\beta} = \mathbf{A} \mathbf{A}^T \boldsymbol{\delta}_q$  e, portanto, as hipóteses  $H_0 : \mathbf{A} \boldsymbol{\beta} = \mathbf{0}$  e  $H : \mathbf{A} \boldsymbol{\beta} \neq \mathbf{0}$  são equivalentes a  $H_0 : \boldsymbol{\delta}_q = \mathbf{0}$  e  $H : \boldsymbol{\delta}_q \neq \mathbf{0}$ . A estrutura linear do modelo alternativo fica igual a  $\boldsymbol{\eta} = \mathbf{X} \mathbf{A}^T \boldsymbol{\delta}_q + \mathbf{X} \mathbf{M} \boldsymbol{\delta}_{p-q}$ .

O algoritmo de estimação aplicado à matriz  $\mathbf{X} \mathbf{M}$  determinará a estimativa  $\tilde{\boldsymbol{\delta}}_{p-q}$  de  $\boldsymbol{\delta}_{p-q}$ , segundo a hipótese nula e, conseqüentemente,  $\tilde{\boldsymbol{\beta}} = \mathbf{M} \tilde{\boldsymbol{\delta}}_{p-q}$ . Sejam  $\hat{\boldsymbol{\delta}}_q$  e  $\hat{\boldsymbol{\delta}}_{p-q}$ , as estimativas dos parâmetros lineares segundo o modelo  $(\mathbf{X} \mathbf{A}^T, \mathbf{X} \mathbf{M})$ . Tem-se,  $\hat{\boldsymbol{\beta}} = \mathbf{A}^T \hat{\boldsymbol{\delta}}_q + \mathbf{M} \hat{\boldsymbol{\delta}}_{p-q}$ .

### 8.5.1 O teste através da razão de verossimilhanças

A razão de verossimilhanças para o teste de  $H_0$  é igual a duas vezes à diferença entre os máximos dos logaritmos da função de verossimilhança segundo os modelos  $(\mathbf{X} \mathbf{A}^T, \mathbf{X} \mathbf{M})$  e  $\mathbf{X} \mathbf{M}$ . Tem-se,

$$w = 2[\ell(\hat{\boldsymbol{\delta}}_q, \hat{\boldsymbol{\delta}}_{p-q}) - \ell(\mathbf{0}, \tilde{\boldsymbol{\delta}}_{p-q})], \quad (8.23)$$

em que  $\tilde{\boldsymbol{\delta}}_{p-q}$  é a EMV de  $\boldsymbol{\delta}_{p-q}$  segundo  $H_0$ .

Para calcular (8.23) basta obter os desvios  $S_{p-q}$  e  $S_p$  dos modelos encaixados  $\mathbf{X} \mathbf{M}$  e  $(\mathbf{X} \mathbf{A}^T, \mathbf{X} \mathbf{M})$ , respectivamente, pois a diferença  $S_{p-q} - S_p$  é igual à expressão (8.23). Assim, a diferença de desvios pode ser usada para comparação de modelos, pois é idêntica à razão de verossimilhanças. A estatística  $S_{p-q} - S_p$  representa uma perda no valor do ajuste ao estimar de um espaço de dimensão  $p - q$ , ao invés de um espaço de dimensão  $p$ , sendo disponível diretamente nos principais software (R, S-PLUS, SAS, GENSTAT, MATLAB) após os ajustes dos dois modelos.

### 8.5.2 O teste através da estatística escore

Seja  $\mathbf{U}(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q}) = [\mathbf{U}_q(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q})^T, \mathbf{U}_{p-q}(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q})^T]^T$ , a função escore subdividida de acordo com a partição  $(\boldsymbol{\delta}_q^T, \boldsymbol{\delta}_{p-q}^T)^T$ . Usando a fórmula matricial para a função escore dada na Seção 3.2, vem  $\mathbf{U}_q(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q}) = \mathbf{A} \mathbf{X}^T \mathbf{W} \mathbf{H}(\mathbf{y} - \boldsymbol{\mu})$  e  $\mathbf{U}_{p-q}(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q}) = \mathbf{M}^T \mathbf{X}^T \mathbf{W} \mathbf{H}(\mathbf{y} - \boldsymbol{\mu})$ . A matriz de informação para os parâmetros  $\boldsymbol{\delta}'s$ , subdividida da

mesma maneira, é obtida conforme descrito na Seção 3.2

$$\mathbf{K}(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q}) = \begin{bmatrix} \mathbf{K}_{qq} & \mathbf{K}_{q(p-q)} \\ \mathbf{K}_{(p-q)q} & \mathbf{K}_{(p-q)(p-q)} \end{bmatrix},$$

em que as submatrizes são definidas por  $\mathbf{K}_{qq} = \phi^{-1} \mathbf{A} \mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{A}^T$ ,  $\mathbf{K}_{q(p-q)} = \mathbf{K}_{(p-q)q}^T = \phi^{-1} \mathbf{A} \mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{M}$  e  $\mathbf{K}_{(p-q)(p-q)} = \phi^{-1} \mathbf{M}^T \mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{M}$ . A função escore e todas as submatrizes da informação dependem das médias  $\mu_1, \dots, \mu_n$ , através de funções das matrizes modelos  $\mathbf{X} \mathbf{A}^T$  e  $\mathbf{X} \mathbf{M}$  e dos parâmetros  $\boldsymbol{\delta}_{p-q}$  e  $\boldsymbol{\delta}_q$ .

A estatística escore para testar  $H_0 : \boldsymbol{\delta}_q = \mathbf{0}$  é deduzida da distribuição normal assintótica da estimativa do componente da função escore relativa à hipótese nula, suposta verdadeira, isto é, de  $\mathbf{U}_q(\mathbf{0}, \tilde{\boldsymbol{\delta}}_{p-q})$ . (Cox e Hinkley, 1986, capítulo 9) apresentam a dedução dessa estatística que tem a forma

$$E = \mathbf{U}_q^T(\mathbf{0}, \tilde{\boldsymbol{\delta}}_{p-q}) \mathbf{C}(\mathbf{0}, \tilde{\boldsymbol{\delta}}_{p-q}) \mathbf{U}_q(\mathbf{0}, \tilde{\boldsymbol{\delta}}_{p-q}), \quad (8.24)$$

em que  $\mathbf{C}(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q}) = \{\mathbf{K}_{qq} - \mathbf{K}_{q(p-q)} \mathbf{K}_{(p-q)(p-q)}^{-1} \mathbf{K}_{q(p-q)}^T\}^{-1}$  representa a matriz de covariância da distribuição marginal assintótica normal de  $\hat{\boldsymbol{\delta}}_q$ , e a expressão, que é subtraída de  $\mathbf{K}_{qq}$  tem o significado de uma matriz de regressão de  $\mathbf{U}_q(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q})$  sobre  $\mathbf{U}_{p-q}(\boldsymbol{\delta}_q, \boldsymbol{\delta}_{p-q})$ .

Se os parâmetros  $\boldsymbol{\delta}_q$  e  $\boldsymbol{\delta}_{p-q}$  resultarem ortogonais, isto é,  $\mathbf{K}_{q(p-q)} = \mathbf{0}$ , a estatística escore reduz-se a

$$E = \phi^{-1} (\mathbf{y} - \tilde{\boldsymbol{\mu}})^T \tilde{\mathbf{H}} \tilde{\mathbf{W}} \mathbf{X} \mathbf{A}^T (\mathbf{A} \mathbf{X}^T \tilde{\mathbf{W}} \mathbf{X} \mathbf{A}^T)^{-1} \mathbf{A} \mathbf{X}^T \tilde{\mathbf{W}} \tilde{\mathbf{H}} (\mathbf{y} - \tilde{\boldsymbol{\mu}}),$$

em que o til indica o valor da estimativa no ponto  $(\mathbf{0}, \tilde{\boldsymbol{\delta}}_{p-q}^T)^T$ .

Pode-se demonstrar que a estatística escore (8.24) para o teste de  $H_0 : \mathbf{A} \boldsymbol{\beta} = \mathbf{0}$  pode ser calculada de uma maneira simples como o incremento da estatística de Pearson generalizada quando se retira da matriz modelo  $(\mathbf{X} \mathbf{A}^T, \mathbf{X} \mathbf{M})$  a submatriz modelo  $\mathbf{X} \mathbf{A}^T$  (Pregibon, 1980). A demonstração é bastante complicada e não será dada aqui.

### 8.5.3 O teste segundo a estatística de Wald

Desenvolve-se, agora, a estatística de Wald  $W$  para o teste de  $H_0 : \boldsymbol{\delta}_q = \mathbf{0}$ . Essa estatística é baseada na distribuição marginal assintótica normal de  $\hat{\boldsymbol{\delta}}_q$  ou de alguma transformação linear conveniente desse estimador. No caso, adota-se  $\mathbf{A} \mathbf{A}^T \hat{\boldsymbol{\delta}}_q = \mathbf{A} \hat{\boldsymbol{\beta}}$ , que tem valor esperado e covariância dados por  $E(\mathbf{A} \hat{\boldsymbol{\beta}}) = \mathbf{A} \boldsymbol{\beta} + O(n^{-1})$  e  $\text{Cov}(\mathbf{A} \hat{\boldsymbol{\beta}}) =$

$\mathbf{A}(\mathbf{X}^T \mathbf{W} \Phi \mathbf{X})^{-1} \mathbf{A}^T + O(n^{-2})$ . Como  $\mathbf{A}\hat{\boldsymbol{\beta}}$  tem distribuição  $N_p(\mathbf{A}\boldsymbol{\beta}, \phi \mathbf{A}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{A}^T)$ , assintoticamente, vem

$$W = \phi^{-1} \hat{\boldsymbol{\beta}}^T \mathbf{A}^T [\mathbf{A}(\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{A}^T]^{-1} \mathbf{A} \hat{\boldsymbol{\beta}}, \quad (8.25)$$

sendo a matriz de pesos estimada em  $\hat{\boldsymbol{\beta}}$ .

### 8.5.4 Comparações entre as três estatísticas

As três estatísticas descritas anteriormente são equivalentes em grandes amostras, convergindo segundo a hipótese nula, para uma variável com distribuição  $\chi_q^2$ . Entretanto, na prática, a estatística (8.25) é mais usada do que as estatísticas (8.23) e (8.24), pois é bem mais simples e não envolve a matriz  $\mathbf{M}$ . A estatística  $w$  depende das estimativas segundo os dois modelos, enquanto  $E$  requer apenas aquelas relativas ao modelo nulo. A região crítica do teste de tamanho  $\alpha$  corresponde ao conjunto  $\{\mathbf{y}; \text{estatística} \geq \chi_q^2(\alpha)\}$ , em que  $\chi_q^2(\alpha)$  é o valor tabelado da distribuição  $\chi^2$  com  $q$  graus de liberdade e a um nível de significância igual a  $100\alpha\%$ .

Se a hipótese nula composta especifica restrições, não homogêneas, do tipo  $\mathbf{A}\boldsymbol{\beta} = \mathbf{d}$ , o mesmo procedimento é usado, escolhendo antes qualquer vetor  $\boldsymbol{\beta}^*$  tal que  $\mathbf{A}\boldsymbol{\beta}^* = \mathbf{d}$ , e substituindo os parâmetros lineares por  $\boldsymbol{\beta} - \boldsymbol{\beta}^*$ .

### 8.5.5 Aplicação do teste na prática

Uma aplicação usual do teste, descrito na seção anterior, é na seleção de co-variáveis, em que  $\mathbf{A}\boldsymbol{\beta}$  corresponde à hipótese que  $q$  componentes do vetor  $\boldsymbol{\beta}$  são identicamente iguais a zero, isto é, escolhendo  $\mathbf{A}$  para produzir  $\mathbf{A}\boldsymbol{\beta} = \boldsymbol{\beta}_q = \mathbf{0}$ . Como um caso simples especial tem-se o teste do  $r$ -ésimo componente de  $\boldsymbol{\beta}$  ser igual a zero, bastando fazer  $\mathbf{A} = \mathbf{l}_r^T$ , em que  $\mathbf{l}_r$  é um vetor coluna de zeros com 1 no  $r$ -ésimo componente.

É fácil observar no teste de  $H_0 : \boldsymbol{\beta}_q = \mathbf{0}$  versus  $H : \boldsymbol{\beta}_q \neq \mathbf{0}$ , que  $\mathbf{A}\mathbf{A}^T = \mathbf{I}$ , em que  $\mathbf{I}$  é a matriz identidade de dimensão  $q$ ,  $\boldsymbol{\delta} = \boldsymbol{\beta}$  e as matrizes  $\mathbf{X}\mathbf{A}^T$  e  $\mathbf{X}\mathbf{M}$  são iguais, respectivamente, às submatrizes  $\mathbf{X}_q$  e  $\mathbf{X}_{p-q}$  da partição de  $\mathbf{X}$  correspondentes à partição de  $\boldsymbol{\beta} = (\boldsymbol{\beta}_q, \boldsymbol{\beta}_{p-q})^T$ .

De maneira análoga à matriz de projeção do modelo clássico de regressão, define-se uma matriz de projeção para o MLG por  $\mathbf{P} = \mathbf{I} - \mathbf{W}^{1/2} \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{1/2}$ , que é

simétrica e idempotente de ordem  $n$ . A covariância assintótica de  $\hat{\beta}_q$  reduz-se a

$$\begin{aligned} \mathbf{C}(\beta_q, \beta_{p-q}) &= \phi[\mathbf{X}_q^T \mathbf{W} \mathbf{X}_q - \mathbf{X}_q^T \mathbf{W} \mathbf{X}_{p-q} (\mathbf{X}_{p-q}^T \mathbf{W} \mathbf{X}_{p-q})^{-1} \mathbf{X}_{p-q}^T \mathbf{W} \mathbf{X}_q]^{-1} \\ &= \phi(\mathbf{X}_q^T \mathbf{W}^{1/2} \mathbf{P}_{p-q} \mathbf{W}^{1/2} \mathbf{X}_q)^{-1}, \end{aligned}$$

em que  $\mathbf{P}_{p-q}$  é a matriz de projeção segundo o modelo  $\boldsymbol{\eta} = \mathbf{X}_{p-q} \beta_{p-q}$ .

As três estatísticas para o teste  $H_0 : \beta_q = \mathbf{0}$  simplificam-se para

$$w = 2[\ell(\hat{\beta}_q, \hat{\beta}_{p-q}) - \ell(\mathbf{0}, \tilde{\beta}_{p-q})], \quad (8.26)$$

$$E = \phi^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}})^T \tilde{\mathbf{H}} \tilde{\mathbf{W}} \mathbf{X}_q (\mathbf{X}_q^T \tilde{\mathbf{W}}^{1/2} \tilde{\mathbf{P}}_{p-q} \tilde{\mathbf{W}}^{1/2} \mathbf{X}_q)^{-1} \mathbf{X}_q^T \tilde{\mathbf{W}} \tilde{\mathbf{H}} (\mathbf{y} - \tilde{\boldsymbol{\mu}}), \quad (8.27)$$

$$W = \phi^{-1} \hat{\beta}_q^T \mathbf{X}_q^T \widehat{\mathbf{W}}^{1/2} \widehat{\mathbf{P}}_{p-q} \widehat{\mathbf{W}}^{1/2} \mathbf{X}_q \hat{\beta}_q, \quad (8.28)$$

em que, como antes, circunflexos e til correspondem às estimativas segundo os modelos  $\boldsymbol{\eta} = \mathbf{X}_{p-q} \beta_{p-q} + \mathbf{X}_q \beta_q$  e  $\boldsymbol{\eta} = \mathbf{X}_{p-q} \beta_{p-q}$ , respectivamente. O parâmetro de dispersão  $\phi$  funciona como divisor nas equações (8.26)-(8.28). Assim, quanto maior for o parâmetro de dispersão, menores serão os valores das estatísticas  $w$ ,  $E$  e  $W$ .

Se os parâmetros  $\beta_{p-q}$  e  $\beta_q$  resultarem ortogonais,  $\mathbf{C}(\beta_q, \beta_{p-q}) = \phi(\mathbf{X}_q^T \mathbf{W} \mathbf{X}_q)^{-1}$  e, então,

$$E = \phi^{-1}(\mathbf{y} - \tilde{\boldsymbol{\mu}})^T \tilde{\mathbf{H}} \tilde{\mathbf{W}} \mathbf{X}_q (\mathbf{X}_q^T \tilde{\mathbf{W}} \mathbf{X}_q)^{-1} \mathbf{X}_q^T \tilde{\mathbf{W}} \tilde{\mathbf{H}} (\mathbf{y} - \tilde{\boldsymbol{\mu}})$$

e

$$W = \phi^{-1} \hat{\beta}_q^T \mathbf{X}_q^T \widehat{\mathbf{W}} \mathbf{X}_q \hat{\beta}_q.$$

Planejando-se testar a hipótese que envolve  $q$  componentes do vetor  $\beta$  serem iguais aos valores especificados  $\beta_q^{(0)}$ ,  $H_0 : \beta_q = \beta_q^{(0)}$ , os resultados (8.26), (8.27) e (8.28) são aplicáveis, substituindo, simplesmente, na estrutura linear  $\beta_q$  por  $\beta'_q = \beta_q - \beta_q^{(0)}$ . Usualmente, considera-se o caso especial, em que a hipótese nula envolve um único parâmetro  $H_0 : \beta_r = 0$  (versus  $H : \beta_r \neq 0$ ). A estatística de Wald é obtida de (8.25) fazendo  $\mathbf{A} = \mathbf{I}_r^T$ , implicando em  $W = -\hat{\beta}_r^2 / \hat{\kappa}^{rr}$ , em que  $-\hat{\kappa}^{rr}$  é a variância estimada de  $\hat{\beta}_r$ .

Hauck e Donner (1977) estudaram o comportamento dessa estatística no modelo binomial logístico:  $Y_1, \dots, Y_n$  variáveis binárias independentes e  $P(Y_i = 1) = [1 +$

$\exp(-\sum_{r=1}^p x_{ir}\beta_r)]^{-1}$ ,  $i = 1, \dots, n$ . Demonstraram que, fixados  $\beta_1, \dots, \beta_{r-1}, \beta_{r+1}, \dots, \beta_p$  e  $\eta$ , o parâmetro de não-centralidade  $-\beta_r^2/\kappa^{rr}$  da distribuição  $\chi^2$  não-central assintótica de  $W$ , segundo a hipótese alternativa  $H : \beta_r \neq 0$ , tende a zero, quando  $|\beta_r|$  cresce ou, o que é equivalente,  $W \rightarrow 0$  quando  $|\hat{\beta}| \rightarrow \infty$ .

Esse comportamento aberrante de  $W$  é causado pelo fato de que a variância estimada  $-\hat{\kappa}^{rr}$  cresce mais rapidamente do que  $\hat{\beta}_r^2$  quando  $|\hat{\beta}_r| \rightarrow \infty$ . Vaeth (1985) generaliza os resultados de Hauck e Donner, mostrando que  $W$  se comporta igualmente de maneira aberrante nos MLG com parâmetro canônico discreto, como por exemplo, os modelos log-lineares.

### 8.5.6 Componente sistemático em termos de restrições

Trata-se, inicialmente, da situação descrita no final da Seção 8.5.4, ou seja, um MLG com componente sistemático  $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$  e a hipótese nula composta  $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}$ , em que  $\mathbf{A}$  é uma matriz de dimensões  $q \times p$  de posto  $q$  que representa a hipótese a ser testada versus  $H : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{d}$ . Depois considera-se o teste, supondo restrições de desigualdades não-homogêneas, isto é,  $\mathbf{A}\boldsymbol{\beta} \leq \mathbf{d}$ .

#### (a) Restrições de Igualdade

Como foi visto no final da Seção 8.5.4, sem perda de generalidade, o teste da hipótese  $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}$  poderia ser tratado como  $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ . Se  $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}$  é verdadeira, pode-se expressar  $\boldsymbol{\beta}$  como

$$\boldsymbol{\beta} = (\mathbf{I} - \mathbf{A}^- \mathbf{A})\boldsymbol{\gamma} + \mathbf{A}^- \mathbf{d},$$

em que  $\boldsymbol{\gamma}$  é um vetor de dimensão  $p$  arbitrário e  $\mathbf{A}^-$  é uma inversa generalizada de  $\mathbf{A}$ . A estrutura linear  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$  segundo a hipótese nula é igual a  $\boldsymbol{\eta} = \mathbf{X}^*\boldsymbol{\gamma} + \mathbf{X}\mathbf{A}^- \mathbf{d}$ , em que  $\mathbf{X}^* = \mathbf{X}(\mathbf{I} - \mathbf{A}^- \mathbf{A})$  representa uma matriz modelo transformada e  $\mathbf{X}\mathbf{A}^- \mathbf{d}$  é um termo constante, que é especificado na definição do modelo como “offset”.

A estatística-teste de  $H_0$  é igual à diferença entre os logaritmos das funções de verossimilhanças máximas segundo os modelos  $\mathbf{X}^*$  e  $\mathbf{X}$ . A simplicidade do teste depende da determinação de  $\mathbf{A}^-$ . Alguns casos especiais da estrutura de  $\mathbf{A}$  apresentam uma sim-

plificação considerável.

Um primeiro caso corresponde às linhas de  $\mathbf{A}$  serem ortogonais, isto é,  $\mathbf{A}\mathbf{A}^T = \mathbf{I}$ . Logo,  $\mathbf{A}^- = \mathbf{A}^T$  e  $\mathbf{X}^* = \mathbf{X}(\mathbf{I} - \mathbf{A}^T\mathbf{A})$  com “offset”  $\mathbf{X}\mathbf{A}^T\mathbf{d}$ . Essa situação ocorre, freqüentemente, com contrastes lineares ortogonais  $\lambda_1, \dots, \lambda_q$  definidos por  $\boldsymbol{\lambda} = \mathbf{A}\boldsymbol{\beta}$  a serem testados individualmente. Então,  $\boldsymbol{\beta} = \mathbf{A}^T\boldsymbol{\lambda}$  e  $\boldsymbol{\eta} = \mathbf{X}\mathbf{A}^T\boldsymbol{\lambda}$ . O teste de  $H_0 : \lambda_r = 0$  é realizado, como descrito na Seção 8.5.5, com uma matriz modelo transformada  $\mathbf{X}^* = \mathbf{X}\mathbf{A}^T$ .

Um segundo caso especial corresponde a um arranjo nas linhas de  $\mathbf{A}$  e na ordem dos parâmetros em  $\boldsymbol{\beta}$ , produzindo  $\mathbf{A} = (\mathbf{B} \ \mathbf{I})$ , em que  $\mathbf{B}$  é uma matriz de dimensões  $q \times (p - q)$  e  $\mathbf{I}$  é a matriz identidade de ordem  $q$ . Então,

$$\mathbf{A}^- = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} \quad \text{e} \quad \mathbf{X}^* = \mathbf{X} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{B} & \mathbf{0} \end{bmatrix}.$$

As  $q$  colunas zeros em  $\mathbf{X}^*$  podem ser eliminadas resultando

$$\mathbf{X}^* = \mathbf{X} \begin{bmatrix} \mathbf{I} \\ -\mathbf{B} \end{bmatrix} \quad \text{e} \quad \boldsymbol{\beta} = \begin{bmatrix} \mathbf{I} \\ -\mathbf{B} \end{bmatrix} \boldsymbol{\gamma},$$

em que  $\boldsymbol{\gamma}$  tem os  $p - q$  primeiros componentes iguais aos correspondentes de  $\boldsymbol{\beta}$ .

### (b) Restrições de Desigualdade

Considere agora o teste de  $H_0 : \mathbf{A}\boldsymbol{\beta} \leq \mathbf{d}$  versus  $H : \boldsymbol{\beta}$  arbitrário. Sem perda de generalidade, a hipótese nula pode ser reparametrizada, com as restrições correspondendo aos componentes dos parâmetros transformados  $\boldsymbol{\gamma}$ 's serem não-negativos, isto é,  $H_0 : \boldsymbol{\gamma} \geq \mathbf{0}$ . O teste de  $H_0$  versus  $H$  é baseado na razão de verossimilhanças. A dificuldade é encontrar a EMV de  $\boldsymbol{\gamma}$  segundo a hipótese nula. A maximização do logaritmo da função de verossimilhança segundo restrições de igualdade é muito mais simples do que a maximização segundo restrições de desigualdade.

Quatro métodos de cálculo da EMV segundo  $H_0 : \boldsymbol{\gamma} \geq \mathbf{0}$  serão discutidos a seguir. Todos esses métodos envolvem, simplesmente, maximização segundo restrições de igualdade, que poderão ser realizados pelo processo descrito no item (a). Estes métodos estão apresentados, minuciosamente, em McDonald e Diamond (1983).

Um problema de maximização do logaritmo da função de verossimilhança de um MLG com restrições do tipo  $\boldsymbol{\gamma} > \mathbf{0}$  é denotado pela sigla PCR (problema com restrições). A maximização sem restrições corresponde à sigla PSR (problema sem restrições). Um



subproblema sem restrições (SPSR) é deduzido do PSR pela eliminação de algumas covariáveis  $x's$  e, portanto, equivale a supor que os parâmetros  $\gamma's$  correspondentes a esses  $x's$  são iguais a zero.

Waterman (1977) demonstra que, se na resolução do PSR, os componentes estimados  $\hat{\gamma}_{r_1}, \dots, \hat{\gamma}_{r_k}$  são negativos, então, na solução do PCR haverá pelo menos um  $\gamma_{r_j}$ ,  $j = 1, \dots, k$ , igual a zero, isto é, pelo menos uma covariável será eliminada do modelo.

No primeiro método, formulam-se  $2^k - 1$  SPSR com parâmetros  $\gamma_{r_j} = 0$ ,  $j$  variando de tal maneira a gerar todos os subconjuntos não-vazios de  $\{1, \dots, k\}$  e os demais parâmetros livres. A estimativa de  $\gamma$  segundo  $H_0$  é selecionada como aquela que maximiza o logaritmo da função de verossimilhança, entre todas as estimativas desses SPSR que verificam  $H_0$ .

Se  $k$  tiver um valor grande, claro que a resolução de todos os SPSR apresentará dificuldades de ordem computacional. Pode-se obter uma árvore de busca dos SPSR com alguns dos  $\gamma's$  iguais a zero, partindo de  $k$  SPSR em que exatamente um dos parâmetros  $\gamma_{r_1}, \dots, \gamma_{r_k}$  é igual a zero; o processo será repetido enquanto forem encontradas estimativas de parâmetros negativas, devendo-se tomar algum cuidado para evitar duplicações.

Um segundo método de busca é baseado em condições necessárias e suficientes que caracterizam a otimalidade da solução do SPSR (Theil e Van de Panne, 1960). Suponha que na resolução do SPSR com  $k$  parâmetros iguais a zero, isto é,  $\gamma_{r_1} = \gamma_{r_2} = \dots = \gamma_{r_k} = 0$ , nenhum  $\hat{\gamma}_{r_j}$  resultou menor do que zero. Então, a solução do PCR terá sido encontrada, se e somente se, os  $k$  SPSR com  $k - 1$  parâmetros zeros, isto é,  $\gamma_{r_j}$  livre para  $j = 1, \dots, k$  e  $\gamma_{r_t} = 0$  para  $t = 1, \dots, j - 1, j + 1, \dots, k$ , apresentarem a estimativa do parâmetro livre menor do que zero.

Assim, para verificar se um SPSR com  $k$  parâmetros nulos corresponde à solução ótima, devem-se resolver  $k$  SPSR com  $k - 1$  parâmetros nulos. Caso o ótimo não tenha sido encontrado, deve-se repetir o processo para os casos que resultarem na estimativa do parâmetro livre ser maior do que ou igual a zero, tendo-se agora que checar a otimalidade do SPSR com  $k - 1$  parâmetros nulos.

Um terceiro método (McDonald e Diamond, 1982) usa um critério de parada baseado nas condições de Kuhn-Tucker, que determinam se a solução de um SPSR é, também, solução da maximização do logaritmo da função de verossimilhança do MLG

segundo  $H_0 : \gamma \geq 0$ . Considere um MLG com  $k$  parâmetros nulos  $\gamma_{r_1} = \gamma_{r_2} = \dots = \gamma_{r_k} = 0$  e os demais parâmetros livres, constituindo, portanto, um SPSR. Admite-se que as estimativas dos parâmetros livres, com o ajuste desse modelo, foram todas maiores do que ou iguais a zero. As condições de Kuhn-Tucker são formuladas pelo produto interno ponderado (matriz de pesos igual a  $\widehat{\mathbf{W}}\widehat{\mathbf{G}}$ ) de todas as colunas  $\mathbf{x}'s$  pelo vetor de resíduos  $\mathbf{y} - \hat{\boldsymbol{\mu}}$ , isto é, por  $\mathbf{x}^T \widehat{\mathbf{W}}\widehat{\mathbf{G}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ , em que os circunflexos indicam valores estimados, segundo o modelo ajustado com parâmetros  $\gamma_{r_1} = \gamma_{r_2} = \dots = \gamma_{r_k} = 0$ .

Se o produto interno ponderado for zero, quando  $\mathbf{x}$  corresponde a um parâmetro cuja estimativa é positiva, e menor do que ou igual a zero, quando essa estimativa é igual a zero, a solução desse SPSR será também solução do PCR. Essas condições são necessárias e suficientes para um máximo local do MLG, segundo  $H_0$ , quando o logaritmo da função de verossimilhança é uma função estritamente côncava. Essas condições são fáceis de serem implementadas na prática.

O quarto e último método é baseado num algoritmo do tipo “branch-and-bound”, em que uma árvore de SPSR é gerada; cada nó da árvore corresponde a um modelo com todo parâmetro zero ou livre. A eficiência do algoritmo é dada pela determinação implícita de que certos ramos da árvore não poderão produzir a solução ótima, e no cálculo de um limite inferior para o máximo de um SPSR que indicará uma melhor solução em cada etapa do algoritmo. A descrição do método pode ser encontrada em Armstrong e Frome (1976) e Gentle e Kennedy (1979).

## 8.6 Exercícios

1. Demonstrar que a função de ligação  $\eta = \int b''(\theta)^{2/3} d\theta$  normaliza a distribuição de  $\hat{\beta}$ , tornando o seu coeficiente de assimetria, aproximadamente, igual a zero.
2. Supondo que  $Y$  tem distribuição de Poisson  $P(\mu)$ , comparar os valores exatos de  $P(Y \geq y)$  com os valores aproximados usando a transformação da distribuição de Poisson, que simetriza a distribuição e estabiliza a variância (Seção 8.2), supondo  $\mu = 1, 3$  e  $5$  e  $y = \mu + 1, \mu + 2, \dots, \mu + 8$ .

# Capítulo 9

## Modelo Normal Não-Linear

### 9.1 Introdução

Até o início da década de 70 as principais técnicas desenvolvidas para os modelos de regressão não-linear se restringiam à suposição de normalidade para a variável de resposta. Em 1972, Nelder e Wedderburn ampliaram a distribuição da variável de resposta para a família exponencial de distribuições, definindo os Modelos Lineares Generalizados (vide Capítulo 1). Mesmo assim, os modelos normais não-lineares continuaram recebendo um tratamento especial, surgindo diversos trabalhos na década de 70 e nas décadas posteriores. Particularmente, destaca-se o livro de Ratkowsky (1983), onde vários modelos normais não-lineares são discutidos segundo diversos aspectos.

A principal característica desses modelos é que os mesmos em geral são deduzidos a partir de suposições teóricas (quase sempre equações diferenciais) e os parâmetros resultantes são interpretáveis. Assim, aproximá-los para os modelos normais lineares, mesmo que sejam alcançados ajustes satisfatórios, prejudicaria bastante a obtenção de estimativas mais realistas dos parâmetros de interesse.

Nem sempre os modelos normais não-lineares são expressos numa forma paramétrica adequada, que facilite a convergência rápida dos processos iterativos utilizados na estimação dos parâmetros, sendo necessário procurar, em muitos casos, uma parametrização mais apropriada.

Embora as técnicas de diagnóstico da regressão normal não-linear sejam simples extensões das técnicas da regressão linear, as interpretações não são diretamente aplicadas, particularmente em virtude dos resíduos ordinários não terem mais distribuição aproxi-

madamente normal. Isso levou ao desenvolvimento de técnicas específicas de diagnóstico para os modelos normais não-lineares (vide Cook e Tsai, 1985). Similarmente, as propriedades das somas de quadrados contidas nas tabelas clássicas de análise da variância (ANOVA), não são estendidas diretamente para o caso não-linear. Entretanto, alguns pesquisadores continuam construindo tais tabelas após o ajuste de modelos não-lineares e utilizam apenas descritivamente os valores obtidos para a estatística  $F$ .

A forma clássica do modelo normal não-linear é dada por

$$y_i = f_i(\beta; x) + \varepsilon_i = \mu_i(\beta) + \varepsilon_i, \quad i = 1, \dots, n, \quad (9.1)$$

em que os  $\varepsilon_i$ 's são distribuídos normalmente com média zero e variância constante  $\sigma^2$ , as  $f_i$ 's são funções diferenciáveis,  $\beta = (\beta_1, \dots, \beta_p)^T$  contém os parâmetros desconhecidos a serem estimados e  $x = (x_1, \dots, x_q)^T$  representa os valores de  $q$  covariáveis.

Esses modelos são aplicados nas mais diversas áreas, tais como econometria, agricultura, farmacologia, biologia, etc. A seguir são apresentados alguns modelos especiais e a(s) respectiva(s) área(s) em que cada um é mais utilizado.

#### 1. Modelo para avaliar a mistura de duas drogas

Esse modelo é, geralmente, aplicado na área farmacológica, sendo dado por

$$y = \alpha + \delta \log\{x_1 + \rho x_2 + k(\rho x_1 x_2)^{1/2}\} + \varepsilon,$$

em que  $x_1$  e  $x_2$  representam, respectivamente, os logaritmos das doses de duas drogas  $A$  e  $B$ ,  $\delta$  é a inclinação comum da relação log-dose-resposta,  $\rho$  é a potência da droga  $B$  em relação à droga  $A$  e  $k$  representa a interação entre as drogas, sendo interpretado da seguinte maneira:  $k = 0$  significa que há ação similar entre as duas drogas,  $k > 0$  representa sinergismo e  $k < 0$  significa antagonismo.

#### 2. Modelo de Von-Bertalanffy

Frequentemente aplicado na área ecológica para explicar o comprimento de um peixe pela sua idade. A forma mais conhecida desse modelo é dada por

$$y = \alpha[1 - \exp\{-\delta(x - \gamma)\}] + \varepsilon,$$

em que  $x$  representa a idade do peixe,  $\alpha$  é comprimento máximo esperado para a espécie,  $\delta$  é a taxa média de crescimento e  $\gamma$  é um valor nominal em que o comprimento do peixe é zero.

### 3. Modelos sigmoidais

Fenômenos produzindo curvas sigmoidais na forma de  $S$  são freqüentemente encontrados na agricultura, em biologia, ecologia, engenharia e economia. Essas curvas começam em algum ponto fixo e crescem monotonicamente até um ponto de inflexão, a partir daí a taxa de crescimento começa a diminuir até a curva se aproximar de um valor final, chamado de assíntota. Na Tabela 9.1 são relacionados alguns modelos usuais com essa forma.

Tabela 9.1: Alguns modelos do tipo sigmoidal

Modelo	Componente sistemático
Gompertz	$\alpha \exp\{-\exp(\beta - \gamma x)\}$
Logtístico	$\alpha / \{1 + \exp(\beta - \gamma x)\}$
Richards	$\alpha / \{1 + \exp(\beta - \gamma x)\}^{1/\delta}$
Morgan-Mercer-Flodin (MMF)	$(\beta\gamma + \alpha x^\delta) / (\gamma + x^\delta)$
Weibull	$\alpha - \beta \exp(-\gamma x^\delta)$

Fonte: Ratkowsky (1983)

Nesses modelos o parâmetro  $\alpha$  é o valor máximo esperado para a resposta, ou assíntota. O parâmetro  $\beta$  está relacionado com o intercepto, isto é, o valor de  $\mu = E(Y)$  correspondente a  $x = 0$ . Para todos os modelos da Tabela 9.1 esse parâmetro pelo menos determina o intercepto. O parâmetro  $\gamma$  está relacionado com a taxa média de crescimento da curva, e finalmente o parâmetro  $\delta$ , que aparece em alguns modelos, é utilizado para aumentar a flexibilidade dos mesmos no ajuste dos dados.

### 4. Modelos parcialmente não-lineares

Os modelos normais parcialmente não-lineares aparecem muito frequentemente na prática e por terem uma estrutura simples diversas técnicas usuais são simplificadas.

Esses modelos são expressos na forma

$$y = x^T \alpha + \delta f(\gamma) + \varepsilon,$$

em que  $\alpha = (\alpha_1, \dots, \alpha_{p-2})^T$ ,  $x$  é um vetor que contém os valores de  $p - 2$  covariáveis,  $\delta$  e  $\gamma$  são escalares e as  $f_i$ 's são funções diferenciáveis. Alguns exemplos são apresentados a seguir.

Gallant (1975) aplica o modelo  $\mu = \alpha_1 x_1 + \alpha_2 x_2 + \delta \exp(\gamma z)$  em um delineamento com um fator, em que  $x_1$  e  $x_2$  representam dois tratamentos e  $z$  o tempo, que afeta exponencialmente a resposta. Darby e Ellis (1976) utilizam o modelo  $\mu = \alpha + \delta \log(z_1 + \gamma z_2)$  para avaliar a atividade conjunta de duas drogas. Stone (1980) sugere um modelo parecido dado por  $\mu = \alpha + \delta \log\{z_1/(\gamma + z_2)\}$ , em que  $z_1$  e  $z_2$  representam, respectivamente, as concentrações de uma droga ativa e de um reagente. Outro exemplo é o modelo assintótico de regressão  $\mu = \alpha - \delta \gamma^z$  (Ratkowsky, 1983) que tem sido intensivamente aplicado na agricultura, assim como, na biologia, engenharia e, particularmente, na ecologia para explicar o comprimento  $y$  de um peixe pela idade  $z$  do mesmo. Um modelo parcialmente não-linear utilizado para explicar a resistência  $y$  de um termostato pela temperatura  $z$  é dado por  $\mu = -\alpha + \delta/(\gamma + z)$ .

Os seguintes modelos com estrutura não-linear mais geral são encontrados em Ratkowsky: (i)  $\mu = \alpha \exp\{-\beta/(\gamma + x)\}$ , para explicar a idade de um certo tipo de coelho selvagem pelo peso  $x$  dos olhos e (ii)  $\mu = \theta_1 \theta_3 x_1 / (1 + \theta_1 x_1 + \theta_2 x_2)$ , para explicar, numa determinada reação química, a razão da reação denotada por  $y$  pelas concentrações  $x_1$  e  $x_2$  de dois reagentes. Ratkowsky também apresenta diversas formas alternativas de reparametrização para a maioria dos modelos mencionados acima, com o intuito de diminuir o viés das estimativas e facilitar a convergência do processo iterativo utilizado.

Na Seção 9.2 é apresentado o processo iterativo de Newton-Raphson para a obtenção da estimativa de mínimos quadrados de  $\beta$  assim como alguns resultados assintóticos. Medidas de não-linearidade e algumas técnicas relacionadas com esse assunto são discutidas na Seção 9.3. As principais técnicas de diagnóstico utilizadas na regressão normal não-linear são apresentadas na Seção 9.4, seguidas de algumas ilustrações.

## 9.2 Estimação de Máxima Verossimilhança

Sejam  $y_1, \dots, y_n$  variáveis aleatórias independentes com a estrutura dada em (9.1). Será apresentado a seguir o algoritmo de Newton-Raphson para a obtenção da estimativa de mínimos quadrados de  $\beta$ , que coincide com a estimativa de máxima verossimilhança. Essa estimativa é obtida minimizando a função quadrática

$$S(\beta) = \sum_{i=1}^n \{y_i - \mu_i(\beta)\}^2.$$

Expandindo  $\mu(\beta)$  por série de Taylor em torno de um valor  $\beta^{(0)}$  até segunda ordem, chega-se ao seguinte processo iterativo para obter  $\hat{\beta}$ :

$$\beta^{(m+1)} = \beta^{(m)} + \{\tilde{X}^{(m)T} \tilde{X}^{(m)}\}^{-1} \tilde{X}^{(m)T} \{y - \mu(\beta^{(m)})\}, \quad (9.2)$$

$m = 0, 1, \dots$ , em que  $\tilde{X}$  é a matriz Jacobiana da transformação de  $\mu(\beta)$  em  $\beta$ . Esse processo iterativo, também conhecido como algoritmo de Newton-Raphson para o modelo normal não-linear, deve continuar até que  $|(\beta^{(m+1)} - \beta^{(m)})/\beta^{(m)}| < \varepsilon$  em que  $\varepsilon$  é um valor arbitrário.

A convergência de (9.10) em geral depende dos valores iniciais para os parâmetros do vetor  $\beta$ . Isso pode evitar que problemas relacionados com a estrutura paramétrica do modelo, tais como não-linearidade acentuada (Seção 9.3) e/ou mal condicionamento da matriz  $\tilde{X}$ , prejudiquem a convergência do processo iterativo. Em Souza (1986) há uma discussão detalhada do método de Newton-Raphson e de outros métodos iterativos usuais em regressão normal não-linear. Ratkowsky (1983) sugere algumas técnicas para se obter valores iniciais para os parâmetros de  $\beta$ , as quais serão aplicadas a seguir para alguns dos modelos descritos na Seção 9.1

### 1. Modelo para avaliar a mistura de duas drogas

Como  $\alpha$  e  $\delta$  representam, respectivamente, o intercepto e a inclinação quando somente a droga  $A$  é considerada, pode-se utilizar como bons valores iniciais as estimativas obtidas para esses parâmetros em pesquisas que envolveram apenas a droga  $A$ . Denotando tais estimativas por  $\alpha_0$  e  $\delta_0$ , os valores iniciais para os demais parâmetros podem ser obtidos através das estimativas de mínimos quadrados do modelo linear simples

$$z_0 = \rho x_2 + \theta t + \varepsilon,$$

em que  $z_0 = \exp\{(y - \alpha_0)/\delta_0\} - x_1$ ,  $\theta = k\rho^{1/2}$  e  $t = (x_1x_2)^{1/2}$ .

Uma maneira alternativa, quando não for possível conhecer  $\alpha_0$  e  $\delta_0$  pela forma acima, é através da fixação de estimativas para  $\rho$  e  $k$ , com os demais valores iniciais sendo dados pelas estimativas de mínimos quadrados do modelo

$$y = \alpha + \delta t + \varepsilon,$$

em que  $t = \log\{x_1 + \rho_0x_2 + k_0(\rho_0x_1x_2)^{1/2}\}$ . Se os valores obtidos não conduzirem (9.10) à convergência deve-se tentar novas estimativas para  $\rho$  e  $k$  e repetir o procedimento.

## 2. Modelo de Von-Bertalanffy

O primeiro passo nesse caso é obter um valor inicial para  $\alpha$ . Como esse parâmetro representa a assíntota, ou o tamanho máximo esperado para a espécie, um valor inicial razoável para  $\alpha$  pode ser  $\alpha_0 = y_{max}$ . Conhecendo  $\alpha_0$  e substituindo o mesmo na parte sistemática do modelo, obtém-se a seguinte relação:

$z_0 = \theta - \delta x$ , em que  $\theta = \gamma\delta$  e  $z_0 = \log\{1 - (\mu/\alpha_0)\}$ . Logo, valores iniciais para  $\gamma$  e  $\delta$  podem ser obtidos da regressão linear simples de  $\log\{1 - (y/\alpha_0)\}$  sobre  $x$ . Se as estimativas  $\alpha_0$ ,  $\gamma_0$  e  $\delta_0$  não levarem (9.10) à convergência, devem-se tentar novas estimativas para esses parâmetros e repetir o procedimento.

## 3. Modelos sigmoidais

Para os modelos de Gompertz e logístico, os valores iniciais são obtidos de maneira muito parecida. Para ambos, deve-se inicialmente atribuir um valor inicial para  $\alpha$ , por exemplo, observando o gráfico de  $y$  versus  $x$  ou tomando  $\alpha_0 = y_{max}$ , já que  $\alpha$  representa a assíntota. Conhecendo-se  $\alpha_0$  tem-se, respectivamente, as equações lineares em  $\beta$  e  $\gamma$

$$\log\{-\log(\mu/\alpha_0)\} = \beta - \gamma x$$

e

$$\log\{(\alpha_0/\mu) - 1\} = \beta - \gamma x.$$

Logo, os valores iniciais  $\beta_0$  e  $\gamma_0$  saem respectivamente, das regressões lineares de  $\log\{-\log(y/\alpha_0)\}$  e de  $\log\{(\alpha_0/y) - 1\}$  sobre  $x$ .



Para os demais modelos, Richards, MMF e Weibull, a estimativa inicial  $\alpha_0$  pode ser obtida da mesma forma acima. Entretanto, aparece agora o parâmetro adicional  $\delta$ . Em particular, para o modelo de Richards tem-se a relação linear

$$\log \left\{ \left( \frac{\alpha_0}{\mu} \right)^\delta - 1 \right\} = \beta - \gamma x;$$

logo, conhecendo-se uma estimativa para  $\delta$ , os valores iniciais  $\beta_0$  e  $\gamma_0$  serão obtidos da regressão linear de  $\log\{(\frac{\alpha_0}{\mu})^\delta - 1\}$  sobre  $x$ . Ratkowsky (1983) sugere que  $\delta_0$  seja obtido através do ponto de inflexão da curva, isto é, do ponto  $(x_F, \mu_F)$  tal que  $d^2\mu/dx^2$  seja igual a zero.

Diferenciando a parte sistemática do modelo de Richards duas vezes em relação a  $x$  e igualando a zero, obtém-se

$$x_F = (\beta - \log \delta)/\gamma \quad \text{e} \quad \mu_F = \alpha(1 + \delta)^{-1/\delta}.$$

Portanto, obtendo-se uma estimativa para  $\mu_F$ , por exemplo através do gráfico de  $y$  versus  $x$ , extrai-se  $\delta_0$  da equação para  $\mu_F$ .

No modelo MMF, o parâmetro  $\beta$  pode ser estimado inicialmente pelo gráfico de  $y$  versus  $x$ , por exemplo atribuindo a  $\beta$  o valor de  $\mu$  quando  $x = 0$ . Para esse modelo,  $\delta_0$  pode ser obtido através das equações para o ponto de inflexão  $(x_F, \mu_F)$ , em que

$$x_F = \left\{ \frac{\gamma(\delta - 1)}{\delta + 1} \right\}^{1/\delta} \quad \text{e} \quad \mu_F = \{\beta(\delta + 1) + \alpha(\delta - 1)\}/2\delta.$$

Logo, conhecendo-se as estimativas para  $\mu_F$  e  $x_F$ , os valores iniciais  $\delta_0$  e  $\gamma_0$  saem, respectivamente, das equações para  $\mu_F$  e  $x_F$ .

Para o modelo de Weibull,  $\beta_0$  pode ser obtido analogamente ao modelo MMF. Denotando por  $y_{INI}$  a estimativa para  $\mu$  tal que  $x = 0$ , obtém-se

$$\beta_0 = \alpha_0 - y_{INI}.$$

Substituindo  $\beta_0$  e  $\alpha_0$  no componente sistemático do modelo chega-se à seguinte relação linear:

$$z_0 = \log \gamma + \delta \log x,$$

em que  $z_0 = \log\{-\log(\frac{\alpha_0 - \mu}{\beta_0})\}$ , sugerindo que  $\gamma_0$  e  $\delta_0$  sejam obtidos da regressão linear simples de  $\log\{-\log(\frac{\alpha_0 - y}{\beta_0})\}$  sobre  $\log x$ .

## 4. Modelo assintótico de regressão

Para o modelo assintótico, dado por

$$y = \alpha - \beta\gamma^x + \varepsilon,$$

os valores iniciais são facilmente obtidos. Inicialmente deve-se estimar o parâmetro  $\alpha$ , a assíntota, graficamente ou por  $y_{max}$ . Substituindo  $\alpha_0$  na parte sistemática do modelo obtém-se a relação linear

$$z_0 = \log \beta + x \log \gamma,$$

em que  $z_0 = \log(\alpha_0 - \mu)$ . Logo,  $\beta_0$  e  $\gamma_0$  saem da regressão linear simples de  $\log(\alpha_0 - y)$  sobre  $x$ .

Pelos exemplos acima pode-se notar a importância da interpretabilidade dos parâmetros do componente sistemático de um modelo normal não-linear, na estimação desses mesmos parâmetros.

### 9.2.1 Resultados Assintóticos

Nesta seção serão apresentados os resultados assintóticos mais relevantes relacionados com a estimação e testes de hipóteses para o parâmetro  $\beta = (\beta_1, \dots, \beta_p)^T$  do modelo normal não-linear.

O logaritmo da função de verossimilhança do modelo (9.1), como função de  $\beta$ , é expresso na forma

$$L(\beta) = (2\pi\sigma^2)^{-n/2} \exp\{-S(\beta)/2\sigma^2\}.$$

A estimativa de máxima verossimilhança  $\hat{\beta}$ , é obtida pelo processo iterativo dado em (9.10), sendo consistente e tendo assintoticamente distribuição normal  $p$  variada de média  $\beta$  e estrutura de variância-covariância  $K^{-1} = \sigma^2(\tilde{X}^T \tilde{X})^{-1}$  (vide Jennrich, 1969). Analogamente à regressão linear, a estimativa mais usual para  $\sigma^2$  é dada por  $s^2 = S(\hat{\beta})/(n - p)$ , em que  $S(\hat{\beta})$  é a soma de quadrados dos resíduos do modelo ajustado. Logo, um intervalo de  $100(1 - \alpha)\%$  para  $\beta_j$ , será formado pelos limites

$$\hat{\beta}_j \pm t_{\alpha/2} \times (-\hat{k}^{jj})^{1/2},$$

em que  $t_{\alpha/2}$  é o quantil  $(1 - \alpha/2)$  de uma  $t$  de Student com  $(n - p)$  graus de liberdade e  $-\hat{k}^{jj}$  é a estimativa do elemento  $(j, j)$  de  $K^{-1}$ .

Uma região de aproximadamente  $100(1 - \alpha)\%$  de confiança para  $\beta$  foi proposta por Beale (1960), e é formada pelos contornos de  $S(\beta)$  tais que

$$S(\beta) = S(\hat{\beta}) \left\{ 1 + \frac{p}{n - p} F_{p, (n-p)}(\alpha) \right\}.$$

Em particular, se  $L(\beta)$  for aproximadamente quadrática, a região de confiança acima é bem aproximada por

$$(\hat{\beta} - \beta)^T (\hat{X}^T \hat{X}) (\hat{\beta} - \beta) \leq s^2 p F_{p, (n-p)}(\alpha),$$

em que  $F_{p, (n-p)}(\alpha)$  é o quantil  $(1 - \alpha)$  de uma distribuição  $F$  e a matriz  $\tilde{X}$  é avaliada em  $\hat{\beta}$ . Essa última expressão é uma adaptação da região de confiança da regressão normal-linear.

Para testar a hipótese  $H : \beta \in B$ , em que  $B$  é um subconjunto do espaço paramétrico, pode-se utilizar, usualmente, a estatística da razão de verossimilhanças, dada por

$$-2 \log \lambda = n \log \{ S(\tilde{\beta}) - S(\hat{\beta}) \},$$

em que  $S(\tilde{\beta})$  é a soma de quadrados de resíduos para o modelo ajustado em  $H$ . Sob essa hipótese, a estatística acima tem, assintoticamente, distribuição  $\chi^2$  com  $(p - m)$  graus de liberdade, em que  $m = \dim(B)$ . Johansen (1983) mostra que a estatística  $-2 \log \lambda$  é assintoticamente equivalente à estatística

$$n \sum_{i=1}^n \{ \mu_i(\tilde{\beta}) - \mu_i(\hat{\beta}) \}^2 / S(\hat{\beta}),$$

que é mais fácil de ser calculada.

Uma estatística alternativa para testar  $H$  é dada por

$$F = \frac{(n - p)}{(p - m)} \frac{S(\tilde{\beta}) - S(\hat{\beta})}{S(\hat{\beta})},$$

que sob essa hipótese tem assintoticamente distribuição  $F$  com  $(p - m)$  e  $(n - p)$  graus de liberdade. Logo, deve-se rejeitar  $H$ , para um nível de significância  $\alpha$ , se  $F \geq F_{(p-m), (n-p)}(\alpha)$ . Esse resultado vale também quando a variável de resposta não é normal, havendo contudo algumas condições adicionais de regularidade.

### 9.2.2 Exemplos

**Exemplo 9.1:** Modelo de Gompertz para explicar o comprimento de um certo tipo de feijoeiro

O modelo de Gompertz, dado por  $y = \alpha \exp\{-\exp(\beta - \gamma x)\}$ , em que  $\alpha$  representa a assíntota, é freqüentemente utilizado para explicar o comprimento de diversos tipos de feijoeiros ( $y$ ) pela quantidade de água na raiz dos mesmos ( $\ell$ ). Para ilustrar, será utilizado o conjunto de dados da Tabela 9.2. Usando-se  $x = 0.5 + \ell$ ,  $\ell = 0, 1, \dots, 14$  e iniciando o

Tabela 9.2: Observações de comprimento ( $y$ ) de um certo tipo de feijoeiro como função quantidade de água na raiz ( $\ell$ ).

$\ell$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$y$	1.3	1.3	1.9	3.4	5.3	7.1	10.6	16.0	16.4	18.3	20.9	20.5	21.3	21.2	20.9

processo iterativo (algoritmo de Newton-Raphson para o modelo normal não-linear) com os valores  $\alpha_0 = 3.0$ ,  $\beta_0 = 2.1$  e  $\gamma_0 = 0.4$  são necessárias 10 iterações para chegar a convergência, porém para este modelo utilizar o valor inicial de  $\alpha$  igual ao valor máximo de  $y$  diminui o número de iterações, pois  $\alpha$  representa a assíntota. Então, utilizando agora os valores  $\alpha_0 = 21.3$ ,  $\beta_0 = 2.1$  e  $\gamma_0 = 0.4$  chega-se à convergência após 7 iterações com as estimativas (erros padrão entre parênteses)  $\hat{\alpha} = 22.507(0.837)$ ,  $\hat{\beta} = 2.106(0.235)$  e  $\hat{\gamma} = 0.388(0.046)$ , as quais indicam que os parâmetros são bem determinados. Em particular, o tamanho máximo esperado para esse tipo de feijoeiro será aproximadamente  $\hat{\alpha} = 22.51\text{cm}$  com desvio padrão de 0.837.

A Figura 9.1 exibe o gráfico dos valores observados e ajustados versus a quantidade de água na raiz da planta, mostrando que o modelo é adequado para ajustar esse conjunto de dados.

#### Exemplo 9.2 Testando a Interação entre Duas Drogas

Considere o conjunto de dados apresentado na Tabela 9.3 e o modelo descrito na Seção 9.1, para avaliar a atividade conjunta de duas drogas.

Após o ajuste desse modelo, usualmente testa-se a significância da interação, isto é, a

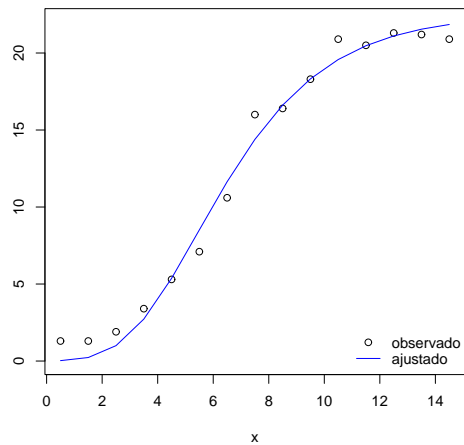


Figura 9.1: Valores observados e ajustados versus a covariável  $x$ .

hipótese  $H : k = 0$ .

As estimativas obtidas sob a hipótese alternativa são apresentadas na Tabela 1.2 e a soma de quadrados de resíduos correspondente vale 220.18. Sob  $H$ , inicializando o

Tabela 9.3: ???.

Tratamento	A	B	y	n
B	0	30.00	26	30
B	0	15.00	19	30
B	0	7.50	7	30
B	0	3.75	5	30
A25:B75	6.50	19.50	23	30
A25:B75	3.25	9.75	11	30
A25:B75	1.625	4.875	3	30
A25:B75	0.812	2.438	0	30
A50:B50	13.00	13.00	15	30
A50:B50	6.50	6.50	5	30
A50:B50	3.25	3.25	4	29
A50:B50	1.625	1.625	0	29
A75:B25	19.50	6.50	20	30
A75:B25	9.75	3.25	13	30
A75:B25	4.875	1.625	6	29
A75:B25	2.438	2.438	0	30
A	30.00	0	23	30
A	15.00	0	21	30
A	7.50	0	13	30
A	3.75	0	5	30

processo iterativo (1.2) com os valores  $\alpha_0 = 20.0$ ,  $\delta_0 = 14.0$  e  $\rho_0 = 0.10$  obtém-se na convergência  $\hat{\alpha} = -18.30(2.83)$ ,  $\hat{\delta} = 10.56(0.83)$ ,  $\hat{\rho} = 0.046(0.0035)$  e soma de quadrados de resíduos dada por 243.97.

Em ambos os ajustes, a comparação das estimativas obtidas com os respectivos desvios padrões indica que os parâmetros são bem determinados. Em particular, a significância da interação  $k$  pode ser avaliada através da estatística  $F$ , obtendo-se o valor

$$F = \frac{(56 - 4)}{(4 - 3)} \frac{243.97 - 220.18}{220.18} \cong 5.62.$$

que é significativo a 5%. Como  $\hat{k} < 0$ , há uma indicação de antagonismo entre as duas drogas, isto é, que a mistura de ambas produz um efeito menor que a soma dos efeitos individuais das duas drogas.

Será mostrado na Seção ??, através de algumas técnicas de diagnóstico, que a significância de  $k$  deve-se essencialmente a algumas “misturas” extremas que incluem apenas uma das drogas. Esse resultado mostra a necessidade de uma análise de diagnóstico após o ajuste do modelo. O valor de  $\hat{\rho} = 0,046$  indica que a insulina padrão é aproximadamente 22 vezes mais eficaz que a insulina na forma suberoyl A1-B29.

### 9.3 Medidas de Não-Linearidade

O principal objetivo das medidas de não-linearidade é verificar se o grau de não-linearidade de um problema de estimação não-linear é suficientemente pequeno para que as técnicas usuais de estimação, desenvolvidas para a regressão linear, sejam utilizadas como uma boa aproximação para o caso não-linear.

A primeira tentativa relevante no sentido de desenvolver uma medida de não-linearidade foi de Beale (1960). Contudo, Guttman e Meeter (1965) mostraram que a medida proposta por ele tende a subestimar o verdadeiro grau de não-linearidade do modelo. Uma outra contribuição importante foi a de Box (1971), que obteve a aproximação de ordem  $n^{-l}$  para o viés do estimador de máxima verossimilhança do vetor  $\beta$  de um modelo normal não-linear. Entretanto, foi somente no início da década de 80 que surgiu o trabalho mais relevante nesta área. Bates e Watts (1980) utilizando alguns conceitos de geometria diferencial, desenvolveram duas medidas de curvatura para os modelos normais

não-lineares. Essas medidas indicam, respectivamente, o grau de não-linearidade intrínseca de um modelo e o grau de não-linearidade aparente ou devida à parametrização utilizada.

Ratkowsky (1983) comparou algumas formas paramétricas para diversos modelos normais não-lineares através de simulações e utilizou as medidas de Box e de Bates e Watts.

Para que o leitor tenha uma idéia mais clara dos conceitos de não-linearidade intrínseca e de não-linearidade aparente, serão comparados a seguir um modelo linear e um modelo não-linear para o caso de  $n = 2$  e  $p = 1$ .

Considere, inicialmente, o modelo linear simples dado por  $y_i = \beta x_i + \varepsilon$ ,  $i = 1, 2$ , em que  $x$  denota uma covariável qualquer e  $\beta$  um parâmetro desconhecido. Nesse caso, o espaço de estimação (espaço de todos os valores ajustados possíveis) tem dimensão igual a um e é formado pelos pontos

$$X\beta = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \beta, \quad \beta \in \mathbf{R},$$

ou seja, é uma reta no  $\mathbf{R}^2$ . Além disso, para qualquer conjunto de soluções  $\beta^{(1)}, \beta^{(2)}, \dots$ , tais que  $\beta^{(i+1)} - \beta^{(i)} = \Delta$ , em que  $\Delta$  é uma constante arbitrária, as soluções possíveis para  $X\beta$  serão tais que

$$X\beta^{(i+1)} - X\beta^{(i)} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \Delta, \quad i = 1, 2, \dots,$$

ou seja, se as soluções para  $\beta$  forem igualmente espaçadas, então os valores ajustados correspondentes, também, serão igualmente espaçados.

Considere agora o modelo normal não-linear  $y_i = x_i^\beta + \varepsilon$ ,  $i = 1, 2$  e os dados apresentados em Ratkowsky (1983, pág. 7)

$$y = (2.5 \quad 10)^T \quad \text{e} \quad X = (2 \quad 3)^T.$$

Nesse caso o espaço de estimação não é mais uma reta, e sim uma curva em torno da estimativa de máxima verossimilhança  $\hat{\beta} = 2.05$ . A curva correspondente aos pontos  $(2^\beta \quad 3^\beta)^T$  com  $\beta$  variando em espaçamentos iguais a 0.5 é exibida pela Figura 1.2. Note que os pontos do espaço de estimação não são igualmente espaçados como ocorreu no caso linear.

Figura 1.2: Representação da curva  $(2^\beta \quad 3^\beta)^T$  com  $\beta$  variando em espaçamentos iguais Ratkowsky (1983).

Assim, quanto mais essa curva se afasta da reta tangente em  $\hat{\beta}$  maior será o que Bates e Watts (1980) chamam de “não-linearidade intrínseca” do modelo, e quanto mais desiguais forem os espaçamentos entre os pontos do espaço de estimação maior será o que ambos chamam de “não-linearidade aparente causada pela parametrização do modelo”.

Portanto, a não-linearidade de um modelo pode ser devida a duas causas. A primeira é a curvatura real do modelo ou intrínseca como definem Bates e Watts, que é invariante com qualquer tipo de reparametrização. A segunda é a curvatura devida à forma como os parâmetros aparecem no modelo. Essa última pode ser eliminada ou pelo menos reduzida através de reparametrizações. Para ilustrar isso, considere o modelo normal não-linear descrito anteriormente com a seguinte reparametrização:

$$y_i = x_i^{\log \phi} + \varepsilon_i, \quad i = 1, 2,$$

em que  $\phi = \exp(\beta)$ . A Figura 1.3 exibe os pontos da curva  $(2^{\log \phi} \ 3^{\log \phi})^T$  com espaçamentos iguais a 1.0 para  $\phi$ . Nota-se que os espaçamentos entre os pontos correspondentes são praticamente iguais, indicando que o grau de não-linearidade aparente foi substancialmente reduzido com essa reparametrização. Entretanto, a curvatura do espaço de estimação continua com a mesma forma anterior, como era de se esperar.

Figura 1.3: Representação da curva  $(2^{\log \phi} \ 3^{\log \phi})^T$  com  $\phi$  variando em espaçamentos iguais a 1.0 Ratkowsky (1983).

### 9.3.1 Medidas de Curvatura de Bates e Watts

Considere o modelo de regressão normal não-linear definido na Seção 9.1. Uma reta no espaço paramétrico passando por  $\hat{\beta}$ , pode ser expressa, usando um parâmetro  $b$  por

$$\beta(b) = \hat{\beta} + bh,$$

em que  $h = (h_1, \dots, h_p)^T$  é um vetor de valores não-nulos. Essa reta gera uma curva sobre o espaço de estimação, definida por

$$\mu_h(b) = \mu(\hat{\beta} + bh).$$



A tangente a essa curva no ponto  $b = 0$  é expressa na forma

$$\dot{\mu}_h = \tilde{X}h, \quad (9.3)$$

em que  $\tilde{X}$  é aqui a matriz Jacobiana da transformação de  $\mu(\beta)$  em  $\beta = \hat{\beta}$ . O conjunto de todas as combinações lineares da forma (1.3) é também chamado de plano tangente em  $\mu(\hat{\beta})$ .

A aceleração da curva  $\mu_h$  é definida por

$$\ddot{\mu}_h = h^T \hat{W}h,$$

em que  $W$  é um “vetor” de dimensão  $n \times p \times p$  com  $i$ -ésima face dada por  $W_i = (\partial^2 \mu_i / \partial \beta_r \partial \beta_s)$ ,  $i = 1, \dots, n$  e  $r, s = 1, \dots, p$ . Portanto, cada elemento do  $n \times 1$  é da forma  $h^T \hat{W}h$ ,  $i = 1, \dots, n$ .

O vetor de aceleração  $\ddot{\mu}_h$  pode ser decomposto em três componentes. O primeiro componente  $\ddot{\mu}_h^{IN}$  determina a variação na direção do vetor de velocidade instantânea  $\dot{\mu}_h$  normal ao plano tangente, enquanto o segundo e o terceiro componentes, cuja norma será denotada  $\ddot{\mu}^{PE}$ , determinam, respectivamente, a variação na direção de  $\ddot{\mu}_h$  paralela ao plano tangente e a variação na velocidade do ponto móvel. Esses componentes foram transformados por Bates e Watts (1980) nas seguintes curvaturas:

A - Curvatura intrínseca definida por

$$K_h^{IN} = \|\ddot{\mu}^{IN}\| / \|\dot{\mu}_h\|^2.$$

B - Curvatura devida à parametrização definida por

$$K_h^{PE} = \|\ddot{\mu}^{PE}\| / \|\dot{\mu}_h\|^2.$$

Essas curvaturas podem ser padronizadas de tal modo que fiquem invariantes com mudanças de escala. Isso é obtido multiplicando  $K_h^{IN}$  e  $K_h^{PE}$  por  $s\sqrt{p}$  com  $s = \{S(\hat{\beta})/(n - p)\}^{1/2}$ . Tem-se, portanto, as curvaturas padronizadas

$$\gamma_h^{IN} = s\sqrt{p} K_h^{IN} \quad \text{e} \quad \gamma_h^{PE} = s\sqrt{p} K_h^{PE}.$$

As medidas relativas acima podem ser usadas não somente para compara diferentes parametrizações de um determinado problema, mas também diferentes conjuntos de dados para o mesmo modelo ou para modelos diferentes.

As medidas de não-linearidade de Bates e Watts (1980) são definidas como sendo as curvaturas máximas

$$\gamma^{IN} = \max_h \{K_h^{IN}\} \quad \text{e} \quad \gamma^{PE} = \max_h \{K_h^{PE}\}.$$

Bates e Watts sugerem o critério

$$\gamma^{IN} \geq 2F^{-1/2} \quad \text{e} \quad \gamma^{PE} \geq 2F^{-1/2}$$

como guia para indicar se o modelo ajustado tem, respectivamente, curvatura intrínseca e curvatura aparente acentuada, em que  $F$  é o quantil  $(l - \alpha)$  de uma distribuição  $F$  com  $p$  e  $(n - p)$  graus de liberdade.

Para o cálculo dessas medidas é preciso, inicialmente, decompor a matriz  $\tilde{X}$  num produto de duas matrizes  $Q$  e  $R$ , isto é,  $\tilde{X} = QR$ , sendo  $Q$  matriz  $n \times n$  ortogonal e  $R$  uma matriz  $n \times p$  definida por

$$R = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix},$$

$\tilde{R}$ , sendo uma matriz  $p \times p$  triangular superior e inversível. As matrizes  $Q$  e  $R$  podem ser obtidas a partir da decomposição de Businger e Golub (1965).

A seguir, deve-se obter o “vetor”  $U = L^T \hat{W} L$ , sendo  $L = \tilde{R}^{-1}$ . Os elementos de  $U$  são vetores  $n \times 1$  denotados por  $U_{kj}$ ,  $k, j = 1, \dots, p$ . Define-se então o que Bates e Watts chamam de “vetor” de aceleração  $A = Q^T U$  de dimensão  $n \times p \times p$ . O  $(k, j)$ -ésimo elemento desse “vetor” é um vetor  $n \times 1$  expresso na forma  $Q^T U_{kj}$ . O “vetor”  $A$  é portanto dado por

$$Q = \begin{bmatrix} A^T U_{11} & \cdots & A^T U_{1p} \\ Q^T U_{p1} & \cdots & A^T U_{pp} \end{bmatrix},$$

em que  $Q^T U_{kj1} = (a_{kji}, \dots, a_{kjp})^T$ . A  $i$ -ésima face de  $A$  é expressa na forma

$$Q = \begin{bmatrix} a_{11i} & \cdots & a_{1pi} \\ \vdots & & \\ a_{p1i} & \cdots & a_{ppi} \end{bmatrix}, \quad i = 1, \dots, n.$$

Sejam  $A^{IN}$  o “vetor” constituído das  $p$  primeiras faces de  $A$  e  $A^{PE}$  o “vetor” constituído das últimas  $(n - p)$  faces de  $A$ . Então, as medidas de não-linearidade serão dadas por

$$\gamma^{IN} = \max_h \|h^T A^{IN} h\| \quad \text{e} \quad \gamma^{PE} = \max_h \|h^T A^{PE} h\|,$$

sendo  $\|h\| = 1$ . Para efetuar os cálculos acima não há, em geral, fórmulas explícitas, sendo necessário recorrer a algum processo iterativo. Souza discute a obtenção de  $\gamma^{IN}$  e  $\gamma^{PE}$  através de um processo iterativo proposto por Bates e Watts (1980).

### 9.3.2 Viés de Ordem $n^{-1}$ de $\hat{\beta}$

Cox e Snell (1968) deduziram uma aproximação de ordem  $n^{-1}$  para o viés do estimador de máxima verossimilhança do parâmetro  $\beta$  em uma classe geral de modelos que inclui o modelo normal não-linear como um caso particular. Box (1971) utilizando esse trabalho, obteve uma aproximação  $b \cong E\{\hat{\beta} - \beta\}$  em forma matricial, dada por

$$b = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T d, \quad (9.4)$$

em que  $d$  é um vetor  $n \times 1$  com elementos  $d_i = -\frac{1}{2}\sigma^2 \text{tr}\{(\tilde{X}^T \tilde{X})^{-1} W_i\}$ ,  $i = 1, \dots, n$  e, como antes,  $W_i = (\partial^2 \mu_i / \partial \beta_r \partial \beta_s)$  é uma matriz quadrada de ordem  $n$ . Portanto, o viés é simplesmente a estimativa de mínimos quadrados para o conjunto de coeficientes da regressão normal linear de  $d$  sobre as colunas de  $\tilde{X}$ . Aqui  $\tilde{X}$  é avaliada no parâmetro verdadeiro  $\beta$ .

Cook et al. (1986) mostraram que  $d$  é essencialmente a diferença entre os valores esperados das aproximações linear e quadrática para  $\mu(\beta)$ . Logo, o viés será pequeno se todos os elementos de  $d$  forem suficientemente próximos de zero, o que indica que o modelo é essencialmente linear, ou se  $d$  é ortogonal às colunas de  $\tilde{X}$ .

Bates e Watts (1980) mostraram que o viés de Box está relacionado com a medida de não-linearidade  $\gamma^{PE}$ . Portanto, o viés pode ser reduzido através de reparametrizações no modelo e a expressão (9.4) pode indicar quais parâmetros são os maiores responsáveis por um valor alto de não-linearidade.

Em particular, para os modelos normais parcialmente não-lineares termos de (9.4) tem-se

$$\tilde{X} = [x, f(\gamma), \delta f'(\gamma)]$$

e de (9.4) obtém-se

$$d_i = 2f'_i(\gamma)Cov(\hat{\delta}, \hat{\gamma}) + \delta f''_i(\gamma)Var(\hat{\gamma}), \quad i = 1, \dots, n.$$

Logo, o viés fica expresso por

$$b = -\delta^{-1}Cov(\hat{\delta}, \hat{\gamma})I_p - \frac{1}{2}\delta Var(\hat{\gamma})(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T f''(\gamma),$$

em que é um vetor  $p \times 1$  de zeros com o valor um na última posição e  $(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T f''(\gamma)$  são as estimativas dos coeficientes da regressão normal linear de  $f''(\gamma)$  sobre  $\tilde{X}$ . Note que  $Cov(\hat{\delta}, \hat{\gamma})$  contribui somente para o viés de  $\hat{\gamma}$ .

Box (1971) também desenvolveu uma fórmula para avaliar o viés dos estimadores de uma nova reparametrização, mostrando que o novo viés pode ser obtido através do viés da parametrização anterior.

Considere a reparametrização

$$\phi = g(\beta),$$

sendo  $\phi$  um escalar,  $g(\cdot)$  é uma função diferenciável e  $\beta = (\beta_1, \dots, \beta_p)^T$ . Seja  $b_{\hat{\gamma}}$  o viés de ordem  $n^{-1}$  de  $\phi$ . Box mostrou que

$$b_{\hat{\phi}} = G^T b + \frac{1}{2}tr\{M \text{ Var}(\hat{\beta})\},$$

em que  $G$  é um vetor  $p \times 1$  com as derivadas de  $g(\beta)$  em relação a  $\beta$  e  $M$  é uma matriz  $p \times p$  de derivadas  $\partial^2 g(\beta) / \partial \beta_r \partial \beta_s$ ,  $r, s = 1, \dots, p$ . Ambos,  $G$  e  $M$ , são avaliados em  $\hat{\beta}$ .

A variância de  $\hat{\phi}$  pode, também, ser expressa em função da variância de  $\hat{\beta}$  por

$$\text{Var}(\hat{\phi}) = tr\{(GG^T) \text{ Var}(\hat{\beta})\}.$$

Em particular para  $p = 1$

$$b_{\hat{\phi}} = b \frac{dg(\beta)}{d\beta} + \frac{1}{2} \text{Var}(\hat{\beta}) \frac{d^2 g(\beta)}{d\beta^2}$$

e

$$\text{Var}(\hat{\phi}) = \text{Var}(\hat{\beta}) \left\{ \frac{d^2 g(\beta)}{d\beta^2} \right\}^2,$$

com todas as derivadas sendo avaliadas em  $\hat{\beta}$ .

### 9.3.3 Aperfeiçoamento da Estatística da Razão de Verossimilhanças

Cordeiro e Paula (1989b) utilizando a metodologia de correção de Bartlett (1937) e as expansões de Lawley (1956) corrigiram a estatística da razão de verossimilhanças  $-2 \log \lambda$ , até ordem  $n^{-1}$ , para a classe dos modelos exponenciais não-lineares (Seção 7.1), que engloba como caso particular os modelos normais não-lineares. Esse fator de correção, denotado por  $c$ , faz com que a estatística corrigida  $-2c^{-1} \log \lambda$  se aproxime melhor da qui-quadrado de referência do que a estatística usual  $-2 \log \lambda$ .

Para ilustrar, suponha a partição  $\beta = (\beta_q^T \beta_{p-q}^T)^T$ ,  $p > q$  e a hipótese nula  $H : \beta_{p-q} = 0$ . Nesse caso, a estatística da razão de verossimilhanças é, simplesmente, dada por  $-2 \log \lambda = 2(\hat{\ell}_p - \hat{\ell}_q)$ , em que  $\hat{\ell}_p$  e  $\hat{\ell}_q$  são, respectivamente, o logaritmo da função de verossimilhança maximizada sob  $H$  e sob o modelo em pesquisa. A correção de Bartlett é dada por

$$c = 1 + (\varepsilon_p - \varepsilon_q)/(p - q), \quad (9.5)$$

em que  $\varepsilon_p$  é um termo bastante complicado envolvendo valores esperados de produtos de derivadas do logaritmo da função de verossimilhança. Particularmente, para os modelos normais não-lineares, tem-se Cordeiro e Paula (1989a)

$$\varepsilon_p = \frac{\sigma^2}{4} \psi = \frac{\sigma^2}{4} \{2 \text{tr}(B_d - BZ) - 1^T D M D I\}, \quad (9.6)$$

em que  $Z = \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}$ ,  $B = \{b_{ij}\}$ ,  $b_{ij} = \text{tr}\{W_i(\tilde{X}^T \tilde{X})^{-1} W_j(\tilde{X}^T \tilde{X})^{-1}\}$ ,  $D = \text{diag}\{d_1, \dots, d_n\}$ ,  $1$  é um vetor  $n \times 1$  de 1's,  $M = I - Z$  é o operador de projeção ortogonal de vetores do  $R^n$  no subespaço gerado pelas colunas da matriz  $\tilde{X}$  e  $B_d$  é uma matriz diagonal de ordem  $n$  com os elementos da diagonal de  $B$ .

Mostra-se, utilizando (9.6), que

$$E\left(\frac{SQRes}{\sigma^2}\right) \cong n - p - \frac{\sigma^2}{4} \psi,$$

isto é, o valor esperado da soma de quadrados de resíduos dividida por  $\sigma^2$  é, aproximadamente, igual a  $(n - p)$ , que é o valor esperado no caso linear, mais uma contribuição devida à não-linearidade em  $\mu(\beta)$ , multiplicada por  $-\sigma^2/4$ . ? relacionou  $\psi$  com a medida de não-linearidade de Beale (1960).

Restringindo-se à subclasse dos modelos parcialmente não-lineares,  $\varepsilon_p$  reduz-se para

$$\varepsilon_p = \frac{\sigma^2}{4} \text{Var}^2(\hat{\gamma}) 1^T \Gamma M \Gamma 1,$$

em que  $\sigma^2 \text{Var}(\hat{\sigma})$  é a variância assintótica de  $\hat{\sigma}$ ,  $\Gamma = \delta \text{diag}\{f_1''(\gamma), \dots, f_n''(\gamma)\}$  e  $1^T \Gamma M \Gamma 1$  é a soma de quadrados de resíduos após a regressão linear de  $\Gamma 1$  sobre as colunas de  $\tilde{X}$ .

Na prática o fator  $c$  deve ser estimado sob o menor modelo, isto é, tanto  $\varepsilon_p$  quanto  $\varepsilon_q$  em (9.5) devem ser computados sob  $H : \beta_{p-q} = 0$ . Para ilustrar, suponha que num modelo parcialmente não-linear o interesse é testar  $H : \gamma = \gamma^{(0)}$ . Logo, o fator de correção fica dado por

$$c = 1 + \frac{\sigma^2}{4} \text{Var}^2(\hat{\gamma}) 1^T \Gamma M \Gamma 1,$$

em que as quantidades  $\text{Var}(\hat{\gamma})$ ,  $\Gamma$  e  $M$  devem ser computadas sob  $H$ . Aqui, em particular,  $\text{Var}(\hat{\gamma})$  é o elemento que ocupa a posição  $(p, p)$  da matriz  $(\tilde{X}^T \tilde{X})^{-1}$ .

### 9.3.4 Exemplos

1 - Reparametrização do Modelo MMF através das Medidas de Bates e Watts e de Box.

Considere o modelo MMF, dado por

$$\mu = (\beta\gamma + \alpha x^\delta) / (\gamma + x^\delta),$$

e o conjunto de dados da Tabela 9.6 (Ratkowsky, 1983, pág. 88)

Tabela 9.4:  $\hat{\mu}(y)$  de um  $\hat{\mu}$  ( $\ell$ ).

x	9	14	21	28	42	57	63	70	79
y	8.93	10.80	18.59	22.33	39.35	56.11	61.73	64.62	67.08

em que  $y$  representa a produção e  $x$  o tempo de cultivo de uma certa cultura.

Na convergência, obtêm-se as estimativas  $\hat{\alpha} = 80.96$ ,  $\hat{\beta} = 8.895$ ,  $\hat{\gamma} = 49.577$  e  $\hat{\delta} = 2.828$ . As medidas de não-linearidade são estimadas por  $\hat{\gamma}^{IN} = 0.180$  e  $\hat{\gamma}^{PE} = 90.970$  e o valor crítico para um nível de significância de 5% vale  $1/2\sqrt{F} = 0.229$ , sendo  $F$  o quantil de 95% de uma distribuição  $F$  com 3 e 6 graus de liberdade. Portanto, a não-linearidade

devida à parametrização do modelo é altamente significativa, enquanto a não-linearidade intrínseca é não-significativa.

Para determinar quais os parâmetros que possivelmente estão causando essa não-linearidade acentuada, utiliza-se a medida de Box para o viés. É usual expressar o viés como uma porcentagem da estimativa correspondente. Para as estimativas acima essas porcentagens valem, respectivamente, 1.525%, -1.643%, 119.478% e 0.921%. Nota-se, portanto, um valor muito elevado para o viés de  $\hat{\gamma}$ , indicando que possivelmente uma reparametrização nesse parâmetro possa reduzir o viés.

Ratkowsky sugere a simulação das distribuições das estimativas com viés acentuado, para se ter uma idéia da reparametrização a ser aplicada. No exemplo acima, a transformação  $\phi = g(\gamma) = \log \gamma$  é a mais recomendada, obtendo-se  $\hat{\phi} = \log \hat{\gamma} = 10.81$  e  $b_{\hat{\phi}} = 0.1087$ . Logo, a porcentagem do viés vale agora  $0.1087 \times 100/10.81 \cong 1.00\%$ , uma redução substancial em relação à porcentagem inicial.

## 2 - Modelo para Explicar a Resistência de um Termostato

O modelo  $\mu = -\alpha + \delta/(\gamma + x)$  é, freqüentemente, utilizado para explicar a resistência  $y$  de um termostato pela temperatura  $x$ . Na versão acima do modelo, a variável resposta é expressa na forma  $\log y$ .

Esse modelo será utilizado para ajustar o conjunto de dados abaixo (Ratkowsky, 1983, pág. 120)

Tabela 9.5:  $??(y)$  de um  $??$  ( $\ell$ ).

34,780	28,610	23,650	19,630	16,370	13,720	11,540	9,744
8,261	7,030	6,005	5,147	4,427	3,820	3,307	2,872

e

$$x = 50 + 5 \times \ell, \quad \ell = 0, 1, \dots, 15.$$

Iniciando o processo iterativo (algoritmo de Newton-Raphson para o modelo normal não-linear) com os valores  $\alpha_0 = 1.0$ ,  $\delta_0 = 3.0$  e  $\gamma_0 = 2.0$  chega-se à convergência após 10 iterações. As estimativas dos parâmetros (erros padrão entre parênteses) são dadas por  $\hat{\alpha} = 12.053(0.017)$ ,  $\hat{\delta} = 6.14 \times 10^3(14.550)$  e  $\hat{\gamma} = 3.44 \times 10^2(0.507)$ .

A Figura 9.2 exibe o gráfico dos valores observados e ajustados versus a temperatura, mostrando que o modelo é adequado para ajustar esse conjunto de dados.

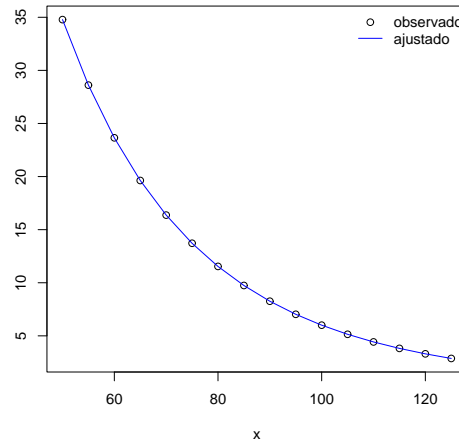


Figura 9.2: Valores observados e ajustados *versus* a covariável  $x$ .

As porcentagens do viés de Box são desprezíveis para cada estimativa (menores que 0.001%) e as simulações não indicam afastamentos da normalidade. Esse resultado, um tanto contraditório, pode ser explicado pelo fato das medidas de não-linearidade de Bates e Watts serem globais enquanto a medida de Box é individual, assim como as simulações, que indicam as distribuições marginais das estimativas. Logo, pode ocorrer que a medida de curvatura  $\gamma_{PE}$  seja significativa e nenhum parâmetro esteja influenciando de forma acentuada na não-linearidade do modelo.

## 9.4 Técnicas de Diagnóstico

Exceto com relação aos resíduos, as técnicas mais usuais de diagnóstico em regressão normal não-linear são simples adaptações da regressão linear. Algumas dessas técnicas serão apresentadas nesta seção.

### 9.4.1 Matriz de Projeção

No caso normal não-linear utiliza-se na detecção de pontos mais afastados dos demais, possivelmente pontos influentes, a matriz de projeção ortogonal no subespaço tangente a



$\eta(\hat{\beta})$ , dada por

$$\hat{H} = \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$$

em que  $\tilde{X}$  é avaliada em  $\hat{\beta}$ . Ao contrário da regressão linear, essa é uma matriz de projeção local, pois depende de  $\hat{\beta}$ . Mesmo assim, o critério  $\hat{h}_{ii} \leq 2p/n$  continua sendo adotado como guia para detectar pontos suspeitos de serem influentes.

### 9.4.2 Resíduo Projetado

Os resíduos ordinários no caso normal não-linear são definidos por  $r_i = y_i - \mu_i(\hat{\beta})$ ,  $i = 1, \dots, n$ . A distribuição desses resíduos agora é instratável, principalmente para pequenas amostras. Além disso, os mesmos em geral têm esperança diferente de zero e distribuição dependendo fortemente dos valores ajustados, o que pode levá-los a não refletirem exatamente a distribuição dos erros. Logo, nesses casos, os critérios de diagnóstico da regressão normal-linear podem falhar. Por exemplo, um resíduo muito diferente de zero, que segundo os critérios da regressão linear seria um ponto aberrante, pode agora não ser, caso o valor esperado desse seja também substancialmente diferente de zero.

Será definido a seguir um novo resíduo, que apesar de algebricamente ser mais complexo, tem propriedades mais próximas das do resíduo ordinário da regressão normal-linear.

Ao desenvolverem  $\mu'(\hat{\beta})$  e  $\mu(\beta)$  por série de Taylor em torno de  $\beta$  até primeira e segunda ordem, respectivamente, Cook e Tsai (1985) encontraram a seguinte aproximação para  $r$ :

$$r \cong (I - H)r - \tilde{X} \sum_{i=1}^n r_i W_i \Delta - \frac{1}{2}(I - H)\Delta^T W \Delta, \quad (9.7)$$

em que  $H$  é o projetor ortogonal em  $C(\tilde{X})$  (subespaço gerado pelas colunas de  $\tilde{X}$ ) e  $\Delta = \hat{\beta} - \beta$ .

Uma aproximação quadrática para  $r$  é obtida substituindo a primeira aproximação linear para  $r$  e  $\Delta$ , respectivamente, na expressão (9.7) mostrando que

$$E(r) \cong (I - H)f$$

e

$$\text{Cov}(r, \eta(\hat{\beta})) \cong NN^T \sigma^2 - \text{Var}(r),$$

em que  $f$  é um vetor  $n \times 1$  de elementos  $f_i = -\frac{1}{2}\sigma^2 \text{tr}(W_i)$   $i = 1, \dots, n$ ,  $N$  é uma matriz  $n \times n$  cujas colunas formam uma base ortonormal em  $C^*(\tilde{X})$  (subespaço gerado pelas colunas ortogonais a  $\tilde{X}$ ) e  $\text{Var}(r) = NN^T \sigma^2 +$  parte positiva. Logo, a covariância entre  $r$  e  $\mu(\hat{\beta})$  tende a ser negativa, o que pode dificultar a interpretação dos gráficos padrões, baseados em  $r$ .

Mostra-se que o segundo termo de (9.7) está em  $C(\tilde{X})$ , enquanto o terceiro termo está em  $C(W^*)$ , sendo  $W^*$  um “vetor”  $n \times p \times p$  cuja  $(k, j)$ -ésima coluna é a projeção de  $\tilde{X}_{ij} = (\partial^2 \eta_1 / \partial \beta_k \partial \beta_j, \dots, \partial^2 \eta_n / \partial \beta_k \partial \beta_j)^T$  em  $C^*(\tilde{X})$ , isto é,  $(I - H)\tilde{X}_{kj}$ .

Logo, as contribuições desses dois termos, que possivelmente explicam os problemas encontrados nas análises de diagnóstico baseadas em  $r$ , podem ser removidas projetando-se  $r$  em  $C^*(\tilde{X}, W^*)$ .

Sejam  $H_2$  e  $H_1$  os operadores de projeção ortogonal em  $C(\tilde{X}, W^*)$  e  $C(W^*)$ , respectivamente. Utilizando (9.7), Cook e Tsai (1985) definiram o resíduo projetado

$$(I - H_2)r = (I - H)\varepsilon - (I - H_1)\varepsilon. \quad (9.8)$$

O primeiro termo de (9.8) é a aproximação linear para o resíduo ordinário  $r$ , enquanto o segundo termo reflete a perda de informação necessária para se removerem as componentes não-lineares de (9.7). Se  $q = \text{posto}(H_1)$  for pequeno em relação a  $(n - p)$ , então essa perda, também, será pequena. Independente disso, se a medida de não-linearidade intrínseca  $\gamma^{IN}$  for significativa, isto é,  $\gamma^{IN} > 2F^{-1/2}$ , o ganho será substancial.

De (9.8) mostra-se, facilmente, que

$$E\{(I - H_2)r\} = 0, \quad \text{Var}\{(I - H_2)r\} = \sigma^2(I - H_2)$$

e

$$E\{r^T(I - H_2)r\} = \sigma^2 \text{tr}(I - H_2).$$

Logo, uma estimativa alternativa para  $\sigma^2$  é dada por

$$\tilde{\sigma}^2 = \frac{r^T(I - \hat{H}_2)r}{\text{tr}(H_2)}.$$

Os resíduos projetados superam os resíduos ordinários em diversos aspectos e muitas das técnicas de diagnóstico utilizadas na regressão linear são também aplicáveis aos

mesmos. Por exemplo, os gráficos de  $(I - \hat{H}_2)r$  contra covariáveis não incluídas no modelo podem revelar como esses termos aparecem no componente sistemático.

É importante lembrar que os operadores utilizados acima dependem de  $\beta$ , portanto na prática é preciso substituir essas quantidades pelas respectivas estimativas. Claramente  $r$  está em  $C^*(\tilde{X})$ , quando  $\tilde{X}$  é avaliado em  $\hat{\beta}$ ; logo,  $(I - \hat{H}_2)r = (I - \hat{H}_1 - \hat{H})r = (I - \hat{H}_1)r$  sendo  $\hat{H}_1 r$  os valores ajustados da regressão linear de  $r$  sobre  $(I - \hat{H})\tilde{X}_{kj}$ ,  $k, j = 1, \dots, p$ .

Na regressão linear, como foi visto no Capítulo 1, mesmo para erros não-correlacionados e de variância constante, os resíduos são correlacionados e com variância diferentes. São definidos então os resíduos studentizados que mesmo correlacionados, apresentam média zero e variância constante e igual a 1.

Similarmente, define-se agora  $s = s\{(I - \hat{H}_1)r\}$  como sendo o vetor de resíduos projetados estudentizados, cuja  $i$ -ésima componente será dada por

$$s_i = \frac{\{(I - \hat{H}_1)r\}_i}{\tilde{\sigma}\{(I - \hat{H}_2)r\}_{ii}^{1/2}}, \quad i = 1, \dots, n. \quad (9.9)$$

Cook e Tsai (1985) exibem para um exemplo particular o gráfico de  $(t_i - s_i)$  contra os valores de uma única covariável e mostram os diferentes diagnósticos que são obtidos se os critérios utilizados para  $s_i$  forem também adotados para os resíduos ordinários studentizados  $t_i$ ,  $i = 1, \dots, n$ . Paula (1987) mostra como obter os  $s_i$ 's pelo sistema GLIM.

Para avaliar se os erros  $\varepsilon_i$ 's têm distribuição aproximadamente normal, assim como para detectar se há pontos aberrantes e/ou influentes, o gráfico de probabilidades dos resíduos projetados ordenados  $s_{(i)}$  versus  $\Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$  pode ser útil, sendo  $\Phi(\cdot)$  a função acumulativa da distribuição normal padrão. A análise dos resíduos em (9.9) procede-se similarmente ao modelo normal linear.

### 9.4.3 Medidas de Influência

As medidas de influência para o modelo normal não-linear são baseadas na regressão linear. A única diferença, que pode ser relevante, é a substituição da estimativa  $\hat{\beta}(i)$  pela estimativa correspondente  $\hat{\beta}_{(i)}^1$ , que é obtida inicializando o processo iterativo (1.2) em  $\hat{\beta}$  sem a  $i$ -ésima observação e tomando a estimativa de um passo. Como o método de Newton-Raphson utiliza em cada passo uma aproximação quadrática para  $L(\beta)$ , a estimativa  $\hat{\beta}_{(i)}^1$  pode não estar muito próxima de  $\hat{\beta}_{(i)}$ , se  $L(\beta)$  não for localmente quadrática. Entretanto,

vários estudos de simulação têm mostrado que essa aproximação é suficiente para chamar a atenção dos pontos influentes.

Mostra-se que essa estimativa de um passo é dada por

$$\hat{\beta}_{(i)}^1 = \hat{\beta} - \frac{(\tilde{X}^T \tilde{X})^{-1}}{(1 - \hat{h}_{ii})} \tilde{x}_i r_i, \quad (9.10)$$

em que  $\tilde{x}_i$  é a  $i$ -ésima linha de  $\tilde{X}$  e  $\tilde{X}$  e  $\tilde{x}_i$  são avaliados em  $\hat{\beta}$ . Logo,  $\hat{\beta}_{(i)}^1$  depende de quantidades correspondentes ao  $i$ -ésimo ponto e de quantidades conhecidas que envolvem todas as observações.

A distância de Cook é agora dada por

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})^T (\tilde{X}^T \tilde{X}) (\hat{\beta}_{(i)} - \hat{\beta}) / ps^2, \quad (9.11)$$

em que  $s^2$  foi definido anteriormente. Usando (9.10) na expressão acima, obtém-se a forma aproximada

$$D_i^1 = \frac{\hat{t}_i^2}{p} \frac{\hat{h}_{ii}}{(1 - \hat{h}_{ii})},$$

sendo  $\hat{t}_i = r_i / s(1 - \hat{h}_{ii})^{1/2}$  o  $i$ -ésimo resíduo ordinário studentizado,  $i = 1, \dots, n$ . Os critérios de calibração para a regressão normal linear podem ser estendidos para o caso não-linear desde que os contornos de  $S(\beta) = \Sigma\{y_i - \mu_i(\beta)\}^2$  sejam aproximadamente elípticos. Isso porque em muitos problemas de regressão normal não-linear as regiões de confiança usuais para  $\beta$  podem ser seriamente viesadas Beale (1960), e o viés pode depender da parametrização escolhida Bates e Watts (1980). Logo, escolher uma parametrização adequada pode ser importante na detecção de pontos influentes.

O gráfico de  $D_i^1$  versus a ordem das observações é usual, devendo-se dar atenção àqueles pontos com o  $D_i^1$  correspondente mais afastado dos demais. Se o interesse é detectar pontos influentes nas estimativas individuais  $\hat{\beta}_j$ ,  $j = 1, \dots, p$ , sugere-se o gráfico de  $\Delta_i \hat{\beta}_j = (\hat{\beta}_j - \hat{\beta}_{(i)j}) / DP(\hat{\beta}_j)$  versus a ordem das observações.

#### 9.4.4 Gráfico da Variável Adicionada

O gráfico da variável adicionada pode revelar como as observações conjuntamente estão influenciando na estimativa do parâmetro que está sendo incluído no modelo. Giltinan et al. (1988) mostraram que esse gráfico pode ser estendido para a classe dos modelos

normais não-lineares, entretanto, de uma forma um pouco diferente. No modelo normal não-linear faz mais sentido incluir um novo parâmetro na parte sistemática, que em muitos casos pode significar uma interação, do que uma nova covariável.

Suponha, então, a média não-linear  $\mu(\beta)$  para o modelo reduzido e o preditor não-linear  $\mu(\beta, \gamma)$  com um parâmetro  $\gamma$  a ser incluído no modelo. Seja  $\tilde{x}_\gamma$  um vetor  $n \times 1$  com as derivadas parciais de  $\mu(\beta, \gamma)$  em relação a  $\gamma$ . Giltinan et al. (1988) sugerem o gráfico de  $r = y - \mu(\hat{\beta})$  versus  $(I - \hat{H})\tilde{x}_{\hat{\gamma}}$ , em que  $\hat{H}$  é a matriz de projeção correspondente ao modelo reduzido e  $\tilde{x}_{\hat{\gamma}}$  é o vetor  $\tilde{x}_\gamma$  computado sob a hipótese  $H : \gamma = 0$ . A estimativa  $\hat{\gamma}$  corresponde à estimativa do parâmetro da regressão linear simples, passando pela origem, de  $y - \mu(\hat{\beta})$  sobre  $(I - H)\tilde{x}_{\hat{\gamma}}$ . Logo, o gráfico proposto pode revelar como as observações estão contribuindo nessa relação e como estão se afastando dela.

### 9.4.5 Exemplos

#### 1 - Obtenção do Resíduo Projetado

Suponha um modelo normal não-linear com ligação entre  $\mu$  e  $\eta$  dada por

$$\log \mu = \eta, \quad \text{em que } \eta = \beta_1 + \beta_2 x_2.$$

Portanto tem-se

$$\tilde{X} = \left\{ \frac{d\mu}{d\eta} \frac{d\mu}{d\eta} x_2 \right\} = \{\exp(\eta) \ x_2 \exp(\eta)\}$$

e

$$W = \left\{ \frac{d^2\mu}{d\eta^2} \frac{d^2\mu}{d\eta^2} x_2 \frac{d^2\mu}{d\eta^2} x_2^2 \right\} = \{\exp(\eta) \ x_2 \exp(\eta) \ x_2^2 \exp(\eta)\}.$$

Serão utilizados para ajustar esse modelo os dados abaixo (Draper e Smith, 1981), em que  $y$  é a fração média de cloro disponível num produto manufaturado e  $x_2$  o tempo de fabricação do mesmo (em semanas):

Tabela 9.6:  $\mu(y)$  de um  $\mu$  ( $\ell$ ).

$y$ :	0.490	0.475	0.450	0.437	0.433	0.455	0.423	0.407	0.407
	0.407	0.405	0.393	0.405	0.400	0.395	0.400	0.390	0.390

e

$x_2 = 2(3 + \ell)$ ,  $\ell = 1, 2, \dots, 18$ . A Figura 1.4 exibe o gráfico de  $y$  contra  $x_2$ .

Figura 1.4: .

Utilizando o algoritmo de Newton-Raphson para o modelo normal não-linear chega-se à convergência após 1 iteração, com as estimativas (erros padrão entre parênteses)  $\hat{\beta}_1 = -0.723(0.0189)$  e  $\hat{\beta}_2 = -0.006(0.0007)$ .

A partir desse ajuste foram obtidos os resíduos ordinários  $r_i = y_i - \eta_i(\hat{\beta})$ ,  $i = 1, \dots, 18$  e a matriz de projeção  $\hat{H} = \tilde{X}(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T$ . Note que as duas primeiras colunas de  $W$  pertencem a  $C(\tilde{X})$ . Logo,  $(I - \hat{H})\tilde{X}_{11} = (I - \hat{H})\tilde{X}_{12} = 0$  e o vetor  $\hat{H}_1 r$  corresponderá aos valores ajustados da regressão linear de  $r$  sobre  $(I - \hat{H})\tilde{X}_{22}$ , em que  $\tilde{X}_{22}$  é a terceira coluna da matriz  $W$ . O vetor de resíduos projetados será então por  $r - \hat{H}_1 r$ . Como  $q = \text{posto}(H_2) = 1$ , a perda de informação quando se passa do subespaço dos resíduos ordinários para o subespaço dos resíduos projetados, será pequena. Os resíduos projetados studentizados são obtidos diretamente de (1.9).

A Figura 1.5 e 1.6 exibem, respectivamente, os gráficos de  $t_i$  versus  $x_{i2}$  e  $s_i$  versus  $x_{i2}$ ,  $i = 1, \dots, n$ .

Figura 1.5: *Gráfico dos resíduos ordinários studentizados  $t_i$ 's contra os valores de  $x_2$ .*

Figura 1.6: *Gráfico dos resíduos projetados  $s_i$ 's contra os valores de  $x_2$ .*

Comparando essas figuras nota-se algumas divergências entre os diagnósticos produzidos pelos gráficos individuais, se forem adotados os mesmos critérios em cada um. Particularmente a observação #06 destaca-se como aberrante na Figura 1.6 o que parece estar em concordância com o posicionamento desse ponto na Figura 1.4.

## 2 - Técnicas de Diagnóstico para avaliar a interação entre duas drogas

(Denis)

## 3 - Modelo para explicar a fração média de cloro disponível num produto manufaturado

Um modelo utilizado para explicar a fração do cloro será um modelo normal não-linear com ligação entre  $\mu$  e  $\eta$  dada por

$$\log \mu = \eta, \quad \text{em que } \eta = \beta_1 + \beta_2 x_2.$$

Portanto tem-se

$$\tilde{X} = \left\{ \frac{d\mu}{d\eta} \frac{d\mu}{d\eta} x_2 \right\} = \{\exp(\eta) \ x_2 \exp(\eta)\}.$$

Esse modelo será utilizado para ajustar o conjunto de dados abaixo (Draper e Smith, 1981), em que  $y$  é a fração média de cloro disponível num produto manufaturado e  $x_2$  o tempo de fabricação do mesmo (em semanas):

$y$ :	0.490	0.475	0.450	0.437	0.433	0.455
	0.423	0.407	0.407	0.407	0.405	0.393
	0.405	0.400	0.395	0.400	0.390	0.390

e

$$x_2 = 2(3 + \ell), \ \ell = 1, 2, \dots, 18.$$

Utilizando o algoritmo de Newton-Raphson para o modelo normal não-linear chega-se à convergência após 7 iterações, com as estimativas (erros padrão entre parênteses)  $\hat{\beta}_1 = -0.716(0.0186)$  e  $\hat{\beta}_2 = -0.006(0.0007)$ .

A Figura 9.3 exibe o gráfico dos valores observados e ajustados versus o tempo de fabricação, mostrando que o modelo é adequado para ajustar esse conjunto de dados, com exceção apenas da observação #1 que encontra-se mais afastada da reta ajustada, em relação às outras observações.

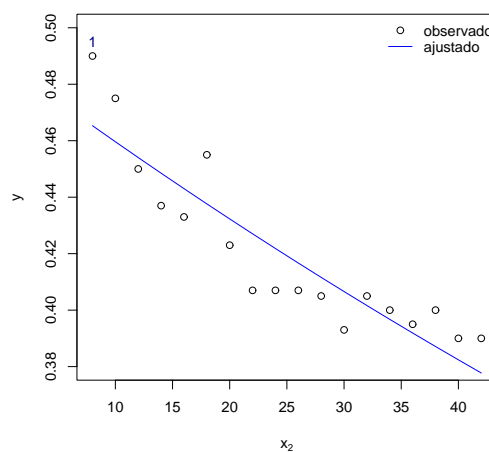


Figura 9.3: Gráfico da fração média de cloro ( $y$ ) *versus* o tempo de fabricação ( $x_2$ ).

A Figura 9.4 exibe o gráfico dos resíduos ordinários studentizados versus a co-variável tempo de fabricação do produto manufaturado. O resíduo referente a observação #1 está fora do intervalo  $(-2, 2)$  o que parece estar em concordância com o posicionamento desse ponto na Figura 9.4.

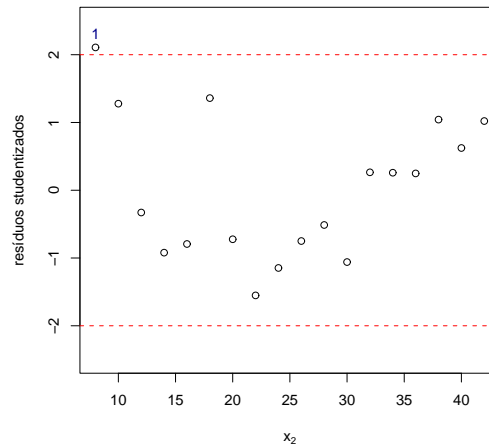


Figura 9.4: Resíduos ordinários studentizados versus os valores de  $x_2$ .

## 9.5 Exercícios

1 - Os dados abaixo referem-se à idade aproximada (em anos) e ao comprimento (em cm) de peixes de 3 espécies segundo o sexo.

Idade (em anos)	Comprimento (cm)					
	Espécie A		Espécie B		Espécie C	
	M	F	M	F	M	F
0	21.5	22.9	19.4	19.3	20.0	20.0
1	33.0	32.2	29.5	29.0	34.0	34.0
2	38.8	38.2	36.8	35.0	38.5	39.5
3	43.1	43.0	42.2	40.4	43.0	44.0
4	45.3	46.6	42.9	44.6	46.0	48.0
5	46.0	47.5	47.4	49.5		
6		50.2		51.6		
7		52.8		51.6		
8				56.9		
9				58.3		



M: macho, F: fêmea.

Resolver as seguintes questões:

- (i) Supondo o modelo de Von-Bertalanffy obter valores iniciais para os parâmetros  $\alpha$ ,  $\delta$  e  $\gamma$  para cada um dos conjuntos de dados acima.
- (ii) Ajustar o modelo de Von-Bertalanffy a cada um desses conjuntos de dados, utilizando processo iterativo dado na Seção 1.2 e os valores iniciais obtidos em (i).
- (iii) Após cada ajuste obter um intervalo de confiança de 95% para o comprimento máximo esperado e para a taxa média de crescimento.
- (iv) Verificar para cada espécie se o comprimento máximo e se a taxa média de crescimento diferem entre machos e fêmeas. Utilize nível de significância de 5%.
- (v) Explícite o cálculo do resíduo projetado para o modelo de Von-Bertalanffy.

2 - Mostre que o ponto de inflexão da curva MMF é dado por  $(x_F, \mu_F)$ , em que  $x_F = \left\{ \frac{\gamma(\delta-1)}{(\delta+1)} \right\}^{1/\delta}$  e  $\mu_F = \{\beta(\delta+1) + \alpha(\delta-1)\}/2\delta$ .

3 - Considere o seguinte conjunto de dados:

$x$	0	1	2	3	4
$y$	44.4	54.6	63.8	65.7	68.9

- (i) Obtenha os valores iniciais para o processo iterativo (1.2), supondo o modelo assintótico.
- (ii) Ajuste o modelo assintótico aos dados acima e encontre uma região de confiança de aproximadamente 95% para os parâmetros.

4 - Considere o seguinte modelo normal não-linear:

$$y = \delta\{1 - \exp(-\gamma x)\} + \varepsilon.$$

- (i) Obtenha valores iniciais para  $\delta$  e  $\gamma$ .
- (ii) Expressar a região de confiança para  $(\delta, \gamma)$  numa forma fechada e de fácil computação.
- (iii) Como fica o fator de correção de Bartlett para testar  $H : \gamma = 0$ ?
- (iv) Ajuste esse modelo ao seguinte conjunto de dados:

$x$	1	2	3	4	5	7
$y$	4.3	8.2	9.5	10.4	12.1	13.1

- (v) Teste a hipótese  $H : \gamma = 0$ . Use  $\alpha = 0.05$

5 - Ajuste o modelo logístico ao seguinte conjunto de dados:

$x$	0	1	2	3	4	5	6	8	10
$y$	1.23	1.52	2.95	4.34	5.26	5.84	6.21	6.50	6.83

Utilize a estatística  $F$  para testar a hipótese  $H : \gamma = 1$ . Use  $\alpha = 0.05$

6 - Construir uma região de confiança para o modelo

$$y = \theta_1 x^{\theta_2} + \varepsilon,$$

e ajustar esse modelo ao seguinte conjunto de dados:

$x$	4	10	17	22	25
$y$	5	20	45	66	85

Teste a hipótese  $H : \theta_2 = 1$  e adote um nível de significância de 5%.

7 - Ajuste o modelo apresentado por Ratkowsky(1983) para explicar, numa determinada reação química, a razão da reação  $y$  pelas concentrações  $x_1$  e  $x_2$  de dois reagentes (Seção 1.10 ao seguinte conjunto de dados:

$x_1$	$x_2$	$y$	$x_1$	$x_2$	$y$
1.0	1.0	0.126	3.0	0.0	0.614
2.0	1.0	0.219	0.3	0.0	0.318
1.0	2.0	0.076	3.0	0.8	0.298
2.0	2.0	0.126	3.0	0.0	0.509
1.0	0.0	0.186	0.2	0.0	0.247
3.0	0.0	0.606	3.0	0.8	0.319
0.2	0.0	0.268			

Faça uma análise completa de diagnóstico.



# Capítulo 10

## Outros Modelos de regressão

### 10.1 Introdução

### 10.2 Modelo de Box e Cox

O uso do modelo clássico de regressão é justificado admitindo-se: (i) linearidade da estrutura de  $E(Y)$ ; (ii) variância constante do erro,  $\text{Var}(Y) = \sigma^2$ ; (iii) normalidade e (iv) independência das observações. Se as suposições (i) a (iii) não são satisfeitas para os dados originais, uma transformação não-linear de  $Y$  poderá atendê-las, pelo menos aproximadamente. Em alguns problemas de regressão, deve-se transformar tanto a variável dependente quanto as variáveis explicativas para que as suposições (i) a (iv) sejam satisfeitas. Transformações das variáveis explicativas não afetam as suposições (ii), (iii) e (iv).

Se os dados  $y$  com médias  $\mu$  e variâncias  $V(\mu)$ , que dependem das médias, são transformados por  $g(y)$  para satisfazer

$$\text{Var}\{g(Y)\} = V(\mu)g'(\mu)^2 = k^2,$$

em que  $k^2$  é uma constante, a condição (ii) será satisfeita. A função estabilizadora da variância dos dados é  $g(\mu) = k \int V(\mu)^{-1/2} d\mu$ . Por exemplo, para  $V(\mu) = \mu$  e  $V(\mu) = \mu^2$ , as funções estabilizadoras são  $\sqrt{y}$  e  $\log y$ , respectivamente. Entretanto, não há garantia que  $g(y)$  escolhido desta maneira satisfaça, também, a condição (iii) de normalidade dos dados transformados. Muitas vezes os dados apresentam um ou mais pontos aberrantes

que implicam em detectar não-normalidade e heterocedasticidade. Algum cuidado deve ser tomado ainda com o mecanismo gerador dos dados e a precisão com que esses são obtidos.

Dificuldades com o modelo clássico de regressão não só ocorrem devido à violação de uma das hipóteses básicas. Muitas vezes são devidas à problemas fora do contexto da forma dos dados, como por exemplo, a multicolinearidade, quando existem relações aproximadamente lineares entre as variáveis explicativas. Esta multicolinearidade poderá causar problemas com as rotinas de inversão da matriz  $X^T X$ . Outro tipo de dificuldade ocorre quando se dispõe de um grande número de variáveis explicativas e, portanto, surge um problema de ordem combinatória para selecionar o modelo. Também é comum os dados apresentarem estruturas especiais, tais como, replicações da variável resposta em certos pontos ou mesmo ortogonalidade. Neste caso, não se deve proceder a análise usual embora, em geral, seja difícil detectar essas características em grandes massas de dados.

Nesta seção, introduz-se a classe dos modelos de Box e Cox que visa transformar a variável dependente para satisfazer as hipóteses (i) a (iv) do modelo clássico de regressão. O modelo de Box e Cox (1964) supõe que os dados  $y = (y_1, \dots, y_n)^T$  são independentes e que existe um escalar  $\lambda$  tal que os dados transformados por

$$z = z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda & \text{se } \lambda \neq 0 \\ \log y & \text{se } \lambda = 0 \end{cases} \quad (10.1)$$

satisfazem  $E(z) = \mu = X\beta$ ,  $\text{Var}(z_i) = \sigma^2$  para  $i = 1, \dots, n$  e  $z \sim N(\mu, \sigma^2 I)$ . A transformação (10.1) tem vantagem sobre a transformação potência simples  $y^\lambda$  por ser contínua em  $\lambda = 0$ . Apesar do modelo admitir a existência de um único  $\lambda$  produzindo linearidade dos efeitos sistemáticos, normalidade e constância da variância dos dados transformados, pode ser que diferentes valores de  $\lambda$  sejam necessários para alcançar tudo isso.

Um valor  $\lambda$  pode ser proposto por uma análise exaustiva ou por considerações a priori dos dados, ou ainda, por facilidade de interpretação. Alternativamente, pode-se estimar  $\lambda$  por máxima verossimilhança, embora não haja garantia de que a EMV de  $\lambda$  produza todos os efeitos desejados.

Verifica-se, facilmente, que a log-verossimilhança como função de  $\lambda, \sigma^2$  e  $\beta$  em relação às observações originais  $y$  é dada por

$$l(\lambda, \sigma^2, \beta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (z - X\beta)^T (z - X\beta) + (\lambda - 1) \sum_{i=1}^n \log y_i, \quad (10.2)$$

em que o terceiro termo é o logaritmo do Jacobiano da transformação, isto é,  $J(\lambda, y) = \prod_{i=1}^n \left| \frac{dz}{dy} \right|$ . A maximização de (10.2) em relação a  $\lambda, \sigma^2$  e  $\beta$  apresenta problemas computacionais e deve ser feita em duas etapas. Fixa-se  $\lambda$  e maximiza-se  $\ell(\lambda, \sigma^2, \beta)$  em relação aos demais parâmetros produzindo as estimativas usuais da regressão como funções de  $\lambda$ ,  $\hat{\beta}(\lambda) = (X^T X)^{-1} X^T z$  e  $\hat{\sigma}^2(\lambda) = \frac{1}{n} z^T (I - H) z$ , sendo  $H$  a matriz de projeção. O máximo da log-verossimilhança como função de  $\lambda$  vale, exceto por uma constante,

$$\hat{l}(\lambda) = -\frac{n}{2} \log \sigma^2(\lambda) + (\lambda - 1) \sum_{i=1}^n \log y_i. \quad (10.3)$$

É bastante informativo traçar o gráfico de  $\hat{l}(\lambda)$  versus  $\lambda$  para um certo conjunto de valores deste parâmetro, por exemplo, os inteiros de -3 a 3 e seus pontos médios. A estimativa de  $\lambda$  corresponderá ao ponto de maior  $\hat{l}(\lambda)$ . O único trabalho envolvido é calcular a soma dos quadrados dos resíduos na regressão de  $z$  sobre  $X$ , isto é,  $n\hat{\sigma}^2(\lambda)$ , para cada valor escolhido de  $\lambda$ . Claro está que a estimativa obtida é apenas uma aproximação da EMV de  $\lambda$ .

Objetivando a realização de inferência sobre o parâmetro  $\lambda$ , o teste da hipótese nula  $H_0 : \lambda = \lambda_0$  versus  $H_1 : \lambda \neq \lambda_0$ , em que  $\lambda_0$  é um valor especificado para  $\lambda$ , pode ser feito comparando a razão de verossimilhanças  $w = 2[\hat{l}(\lambda) - \hat{l}(\lambda_0)]$  com a distribuição assintótica  $\chi_1^2$ . Um intervalo de 100%(1 -  $\alpha$ ) de confiança para  $\lambda$  é facilmente deduzido do gráfico de  $\hat{l}(\lambda)$  versus  $\lambda$  como

$$\left\{ \lambda; l(\lambda) > \hat{l}(\hat{\lambda}) - \frac{1}{2} \chi_1^2(\alpha) \right\}. \quad (10.4)$$

Se  $\lambda = 1$  não pertencer ao intervalo (10.4) conclui-se que uma transformação dos dados será necessária e pode-se selecionar um valor conveniente neste intervalo.

No uso do modelo de Box e Cox pode-se verificar a normalidade dos dados transformados  $z_i$  a partir de um dos seguintes testes:

a) teste de Shapiro-Wilks baseado na estatística

$$W = \frac{\left\{ \sum_{i=1}^n a_i z_{(i)} \right\}^2}{\left\{ \sum_{i=1}^n (z_i - \bar{z})^2 \right\}},$$

em que  $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$  são os dados transformados ordenados e os  $a_i$ 's são constantes tabuladas juntamente com os níveis de significância para  $W$ ;

b) teste de D'Agostino

$$D = \left\{ \sum_{i=1}^n i z_{(i)} \right\} \left\{ n^{3/2} \sqrt{\sum_{i=1}^n z_i^2} \right\}.$$

c) teste de Anderson-Darling

$$A^2 = -n^{-1} \sum_{i=1}^n (2i-1) [1 + \log\{t_i(1-t_{n+1-i})\}],$$

em que  $t_i = \Phi\left(\frac{z_{(i)} - \bar{z}}{s}\right)$  e  $s^2$  é a variância amostral. Valores grandes de A são significantes.

### 10.3 Modelos lineares generalizados com função de ligação composta

Considere um modelo com distribuição (1.5), mas com componente sistemática definida por

$$\begin{aligned} E(Y) &= \mu = C\gamma, \\ f(\gamma) &= \eta = X\beta, \end{aligned} \tag{10.5}$$

em que  $\mu$  e  $y$  são vetores  $n \times 1$ ,  $C$  e  $X$  são matrizes conhecidas  $n \times m$  e  $m \times p$ , respectivamente,  $\gamma = (\gamma_1, \dots, \gamma_m)^T$ ,  $\eta = (\eta_1, \dots, \eta_m)^T$  e  $\beta = (\beta_1, \dots, \beta_p)^T$ . Uma média de  $y$  está relacionada com vários preditores lineares.

Denomina-se  $f(C^{-1}\mu) = \eta$  de *função de ligação composta*, sendo  $C^{-1}$  uma inversa generalizada de  $C$ . Quando  $C$  é a matriz identidade, obviamente a ligação composta reduz-se a uma ligação simples  $f(\mu) = \eta$ . Uma extensão de (10.5) considera uma estrutura não-linear  $\mu_i = c_i(\gamma)$  entre  $\mu$  e  $\gamma$ . O ajuste do modelo  $\mu_i = c_i(\gamma)$ ,  $f(\gamma) = \eta = X\beta$ , pode ser feito via o algoritmo descrito em (3.5) com pequenas modificações. Sem perda de generalidade trabalha-se sem o escalar  $\phi$ . Seja  $\ell(\beta)$  a log-verossimilhança para  $\beta$ . Tem-se  $\partial \ell(\beta) / \partial \beta = \tilde{X}^T V^{-1}(y - \mu)$ , em que  $V = \text{diag}\{V_1, \dots, V_n\}$ ,  $L = \{d\mu_i / d\eta_k\}$  é uma matriz  $n \times m$  e  $\tilde{X} = LX = \{\sum_{k=1}^m x_{kr} d\mu_i / d\eta_k\}$ . A informação para  $\beta$  iguala  $\tilde{X}^T V^{-1} \tilde{X}$  e o processo iterativo é expresso por

$$X^T L^{(m)T} V^{(m)-1} L^{(m)} X \beta^{(m+1)} = X^T L^{(m)T} V^{(m)-1} y^{*(m)},$$



sendo  $y^* = L\eta + y - \mu$ . A variável dependente  $y^*$ , a matriz modelo  $LX$  e os pesos  $V^{-1}$  se modificam no processo iterativo acima. Esse procedimento iterativo pode ser implementado nos *software* R, MATLAB e SAS. A inicialização pode ser feita a partir do ajuste de um modelo similar com  $C$  igual à matriz identidade. Quando  $\mu$  é linear em  $\gamma$ ,  $L = CH^{-1}$ , sendo agora  $H = \text{diag}\{d\eta_1/d\gamma_1, \dots, d\eta_m/d\gamma_m\}$  e, então,  $\tilde{X} = CH^{-1}X$  e  $y^* = CH^{-1}\eta + y - \mu$ .

## 10.4 Modelos semi-paramétricos

Os modelos semi-paramétricos foram propostos por Green e Yandell (1985) quando definiram o preditor linear  $\eta$  como sendo a parte usual  $X\beta$  dos MLG mais uma parte  $s(t)$ , em que  $s(\cdot)$  é alguma função regular cujo argumento  $t$  pode representar uma medida de distância, tempo etc. A função  $s(t)$  é especificada por uma soma  $s(t) = \sum_{i=1}^q \gamma_i g_i(t)$  de  $q$  funções básicas  $g_1, \dots, g_q$  sendo os  $\gamma$ 's parâmetros desconhecidos. O problema de maximização consiste em definir uma log-verossimilhança penalizada como função dos parâmetros  $\beta$  e  $\gamma$  e maximizá-la

$$\max_{\beta, \gamma} [\ell\{\eta(\beta, \gamma)\} - \lambda J\{s(\gamma)\}/2],$$

em que  $J[\cdot]$  é representativo de uma penalidade sobre a não-suavidade de  $s(\cdot)$  e  $\lambda$  uma constante que indica o compromisso entre a suavidade de  $s(\cdot)$  e a maximização de  $\ell\{\eta(\beta, \gamma)\}$ . Em geral, admite-se para  $J\{\cdot\}$  a forma quadrática  $\gamma^T K \gamma$ , sendo  $K$  uma matriz de ordem  $q$  simétrica não-negativa. Se  $t$  tem dimensão um, a penalidade da não-suavidade da curva  $s(t)$  iguala  $\int \{s''(t)\}^2 dt$ , expressão comumente usada para suavizar uma curva.

Uma outra alternativa para estimar a função  $s(t)$  é usar um suavizador linear do tipo  $s(t_i) = \gamma_{0i} + \gamma_{1i}t_i$ , em que esses  $\gamma$ 's representam parâmetros ajustados por mínimos quadrados às  $n_i$  (igual ao maior inteiro  $\leq wn/2$ ) observações de cada lado de  $t_i$  e  $w$  representa a amplitude do suavizador, escolhido distante dos extremos do intervalo  $(1/n, 2)$ .

## 10.5 Modelos aditivos generalizados

Os modelos aditivos generalizados são definidos pela componente aleatória dos MLG e uma componente sistemática da forma

$$g(\mu) = \eta = \beta + \sum_{j=1}^p f_j(x_j),$$

com as restrições  $E\{f_j(x_j)\} = 0$  para  $j = 1, \dots, p$ , em que os  $f_j(x_j)$  são funções não-paramétricas a serem estimadas.

Assim, a estrutura linear  $\sum_{j=1}^p \beta_j x_j$  do MLG é substituída pela forma não-paramétrica  $\sum_{j=1}^p f_j(x_j)$ . As funções  $f_j(x_j)$  são estimadas através de um suavizador de espalhamento dos dados  $(y, x_j)$ , denotado no ponto  $x_{ij}$  por  $S(y|x_{ij})$ ,  $j = 1, \dots, p$ ,  $i = 1, \dots, n$ .

O suavizador mais usado tem a forma linear  $S(y|x_{ij}) = \hat{a}_{ij} + \hat{b}_{ij}x_{ij}$ , em que  $\hat{a}_{ij}$  e  $\hat{b}_{ij}$ , são, respectivamente, as estimativas do intercepto e da declividade na regressão linear simples ajustada somente aos pontos  $(y_e, x_{ej})$  em alguma vizinhança  $N_{ij}$  de  $x_{ij}$ . Pode-se considerar vizinhanças simétricas do tipo  $N_{ij} = \{x_{(i-r)j}, \dots, x_{ij}, \dots, x_{(i+r)j}\}$ , com o parâmetro  $r$  determinando o tamanho de  $N_{ij}$ . Tem-se

$$y\hat{b}_{ij} = \sum_{x_{ej} \in N_{ij}} (x_{ej} - \bar{x}_{ij})y_e / \sum_{x_{ej} \in N_{ij}} (x_{ej} - \bar{x}_{ij}),$$

$$\hat{a}_{ij} = \bar{y}_i - \hat{b}_{ij}\bar{x}_{ij},$$

em que  $\bar{x}_{ij}$  é a média dos valores em  $x_{ej}$  em  $N_{ij}$  e  $\bar{y}_i$  é a média dos  $y$ 's correspondentes.

Para estimar os  $f_j(x_j)$  no modelo normal linear utiliza-se o seguinte algoritmo:

1. Inicializar  $\hat{f}(x_{ij}) = 0$ ,  $\forall i, j$  e  $\hat{\beta} = \bar{y}$ ;
2. Fazer  $j = 1, \dots, p$  e  $i = 1, \dots, n$  e obter os resíduos parciais definidos por

$$r_{ij} = y_i - \hat{\beta} - \sum_{\substack{k=1 \\ k \neq j}}^p \hat{f}_k(x_{ik});$$

3. Calcular  $\hat{f}_j(x_{ij}) = S(r_j|x_{ij})$  ajustando uma regressão linear simples aos pontos  $(r_{ej}, x_{ej})$  pertencentes à uma vizinhança  $N_{ij}$  de  $x_{ij}$ ;
4. Quando  $SQR = \sum_{i=1}^n \{y_i - \hat{\beta} - \sum_{j=1}^p \hat{f}_j(x_{ij})\}^2$  convergir pára-se; caso contrário, volta-se para 2.

Observe-se que a cada etapa o algoritmo suaviza resíduos versus a covariável seguinte. Estes resíduos são obtidos removendo as funções estimadas ou efeitos de todas as outras variáveis. Propriedades interessantes deste algoritmo são discutidas por Hastie e Tibshirani (1986, 1987). A extensão do algoritmo para os MLG é baseada nas equações normais da regressão da variável dependente modificada  $y^*$  sobre  $X$ , usando pesos  $W$  (Seção 2.4). O algoritmo pode ser formulado como:

1. Inicializar  $\hat{f}_j(x_{ij}) = 0$ ,  $j = 1, \dots, p$ ,  $\hat{\beta} = g(\bar{y})$ ,  $\hat{\eta} = \hat{\beta}1$ ,  $\hat{W} = (\bar{y})$  e  $\hat{H} = H(\bar{y})$ , sendo  $W = \text{diag}\{(d\mu/\eta)^2/V\}$ ,  $H = \text{diag}\{d\eta/d\mu\}$  e  $\hat{y}^* = \hat{\beta}1 + \hat{H}(y - \hat{\beta}1)$ ;
2. Calcular os resíduos parciais  $r_j = \hat{W}\hat{y}^* - \hat{\beta}1 - \sum_{\substack{k=1 \\ k \neq j}}^p \hat{f}_k(x_k)$  para  $j = 1, \dots, p$ ;
3. Obter  $\hat{f}_j(x_{ij}) = S(r_j/x_{ij})$  através da regressão linear simples sobre os pares  $(r_{ej}, x_{ej})$  em  $N_{ij}$ ,  $i = 1, \dots, p$ ;
4. Atualizar  $\hat{\beta} = g(\frac{1^T \hat{W} \hat{y}^* 1}{n})$ ,  $\hat{\eta} = \hat{\beta} + \sum_{j=1}^p \hat{f}_j(x_j)$ ,  $\hat{u} = g^{-1}(\hat{\eta})$ ,  $\hat{H} = H(\hat{\mu})$ ,  $\hat{W} = W(\hat{\mu})$  e  $\hat{y}^* = \hat{\eta} + \hat{H}(y - \hat{\mu})$ ;
5. Calcular o desvio  $D(y; \hat{\mu})$  do modelo usando as fórmulas da Seção 2.7.1 como função de  $y$  e  $\hat{\mu}$ . Quando  $D(y; \hat{\mu})$  convergir para-se; caso contrário, volta-se para 2.

## 10.6 Modelos de quase-verossimilhança

Nos modelos de quase-verossimilhança as variáveis são consideradas independentes sem ser necessário especificar qualquer distribuição para variável resposta e o componente sistemático é dado por:

$$E(y_i) = \mu_i(\beta), \quad \text{Var}(y_i) = \phi V_i(\mu_i).$$

Aqui os  $\mu_i'$ s são funções conhecidas dos regressores, os  $V_i'$ s são funções conhecidas das médias desconhecidas (em geral  $V_i(\cdot) = V(\cdot)$  ou  $V_i(\cdot) = a_i V(\cdot)$ ) para valores conhecidos dos  $a_i'$ s e  $\phi$  é um parâmetro de dispersão, possivelmente desconhecido, podendo ainda ser uma função de regressores adicionais. Usualmente,  $\mu(\beta)$  corresponde ao componente sistemático do MLG.

Define-se a *log-quase-verossimilhança* para uma única observação apenas com a suposição de existência de sua média e de sua variância, por

$$Q = Q(y; \mu) = \frac{1}{\phi} \int (y - \mu) V(\mu)^{-1} d\mu. \quad (10.6)$$

Para  $V(\mu) = k, \mu, \mu^2, \mu(1 - \mu), \mu + \mu^2/k$  e  $\mu^3$ , com  $k$  constante, e integrando (10.6), conclui-se que, a menos de constantes, as quase-verossimilhanças são iguais aos respectivos logaritmos das distribuições normal, Poisson, gama, binomial, binomial negativa e normal inversa. Logo, os modelos de quase-verossimilhança são equivalentes aos modelos lineares generalizados para essas funções de variância. Observe-se que a função de variância

paramétrica definida por  $V_\lambda(\mu) = \mu^\lambda$ ,  $\lambda \geq 0$ , contém as variâncias das distribuições normal, Poisson, gama e normal inversa.

Wedderburn (1974) demonstrou que a log-quase-verossimilhança tem propriedades semelhantes à log-verossimilhança

$$E(\partial Q / \partial \mu) = 0, \quad E(\partial Q / \partial \mu)^2 = -E(\partial^2 Q / \partial \mu^2) = 1 / [\phi V(\mu)].$$

Uma terceira propriedade importante entre os logaritmos da verossimilhança  $\ell$  e da quase-verossimilhança  $Q$ , supondo para ambos uma mesma função de variância, é dada por

$$-E(\partial^2 Q / \partial \mu^2) \leq -E(\partial^2 \ell / \partial \mu^2). \quad (10.7)$$

Se  $Y$  seguir a família exponencial (1.5) de distribuições tem-se  $V(\mu) = d\mu/d\theta$ , e, portanto,  $Q = \frac{1}{\phi} \int (y - \mu) d\theta$ . Como  $\mu = b'(\theta)$  então  $Q$  tem expressão idêntica à log-verossimilhança da distribuição de  $y$ . A igualdade em (10.7) somente ocorre no caso de  $\ell$  ser a log-verossimilhança da família exponencial. O lado esquerdo de (10.7) é uma medida da informação quando se conhece apenas a relação entre a variância e a média dos dados enquanto o lado direito é a informação usual de Fisher obtida pelo conhecimento da distribuição dos dados. A quantidade não-negativa  $E[\partial^2(Q - \ell) / \partial \mu^2]$  é a informação que se ganha quando, ao conhecimento da relação variância-média dos dados, se acrescenta a informação da forma da distribuição dos dados. A suposição dos dados pertencer à família exponencial equivale à informação minimal obtida do simples conhecimento da relação funcional variância-média dos dados.

A log-quase-verossimilhança para  $n$  observações é igual a soma de  $n$  contribuições definidas por (10.6). As estimativas de máxima quase-verossimilhança  $\tilde{\beta}, \dots, \tilde{\beta}_p$  são obtidas maximizando esta soma. Supondo que  $\phi$  seja constante para as  $n$  observações  $y_1, \dots, y_n$ , obtém-se o sistema de equações para os  $\tilde{\beta}'s$ , que não dependem de  $\phi$

$$\sum_{i=1}^n (y_i - \mu_i)(\partial \mu_i / \partial \beta_i) / V_i(\mu_i) = 0. \quad (10.8)$$

A maximização do logaritmo da função de quase-verossimilhança generaliza o método de mínimos quadrados, que corresponde ao caso de  $V(\mu)$  constante. Pode-se demonstrar (McCullagh, 1983) que as equações de máxima quase-verossimilhança produzem as melhores estimativas lineares não-tendenciosas, o que representa uma generalização do teorema de Gauss-Markov. Os modelos de quase-verossimilhança podem ser

ajustados facilmente usando SPLUS, GENSTAT, MATLAB, BMDP ou SAS, na pior das hipóteses utilizando sub-programas especiais.

Na análise de dados na forma de contagens, trabalha-se com a distribuição de Poisson supondo que  $\text{Var}(y_i) = \phi\mu_i$ . O parâmetro  $\phi$  é estimado igualando a razão de quase-verossimilhanças  $2\{Q(y; y) - Q(y; \tilde{\mu})\}$  aos graus de liberdade  $(n - p)$  da  $\chi^2$  de referência ou então usando a expressão mais simples

$$\tilde{\phi} = (n - p)^{-1} \sum_{i=1}^n (y_i - \tilde{\mu}_i)^2 / \tilde{\mu}_i.$$

Os dados apresentarão superdispersão se  $\tilde{\phi} > 1$  e subdispersão em caso contrário. Similarmente, dados que apresentam durações de tempo com superdispersão podem ser modelados por  $\text{Var}(y_i) = \phi\mu_i^2$  supondo  $\phi > 1$  e dados na forma de contagens com superdispersão por  $V(\mu) = \mu + \lambda\mu^2$  (binomial negativa) ou por  $V(\mu) = \mu + \lambda\mu + \gamma\mu^2$ . Para proporções usa-se  $V(\mu) = \mu(1 - \mu)$  ou  $\mu^2(1 - \mu)^2$ .

A definição do logaritmo da função de quase-verossimilhança (10.6) permite fazer comparações de modelos com preditores lineares diferentes ou com funções de ligação diferentes. Entretanto, não se podem comparar, sobre os mesmos dados, funções de variância diferentes. Nelder e Pregibon (1987) propuseram uma definição de *quase-verossimilhança estendida*  $Q^+$  a partir da variância e da média dos dados, que permite fazer esta comparação, dada por

$$Q^+ = -1/2 \sum_i \log\{2\pi\phi_i V(y_i)\} - 1/2 \sum_i D(y_i; \mu_i)/\phi_i,$$

sendo o somatório sobre todas as observações e a função  $D(y; \mu)$ , denominada de *quase-desvio*, sendo uma simples extensão do desvio do MLG, definida para uma observação por

$$D(y; \mu) = -2 \int_y^\mu (y - x)V(x)^{-1} dx,$$

isto é,  $D(y; \hat{\mu}) = 2\phi\{Q(y; y) - Q(y; \hat{\mu})\}$ . A função quase-desvio para os dados iguala  $\sum_i D(y_i; \tilde{\mu}_i)$ . Para as funções de variância dos MLG, a função quase-desvio reduz-se aos desvios desses modelos.

A Tabela 10.1 apresenta logaritmo de funções de quase-verossimilhança para algumas funções de variância, com a exceção do escalar  $\phi$ , deduzidas integrando a expressão (10.6). Desta tabela os desvios são facilmente obtidos.

Tabela 10.1: Logaritmos de funções de quase-verossimilhança associados às funções de variância

Função de Variância $V(\mu)$	Log-quase-Verossimilhança $Q(y; \mu)$
$\mu^\lambda (\lambda \neq 0, 1, 2)$	$\mu^{-\lambda} \left( \frac{y\mu}{1-\lambda} - \frac{\mu^2}{2-\lambda} \right)$
$\mu(1-\mu)$	$y \log \left( \frac{\mu}{1-\mu} \right) + \log(1-\mu)$
$\mu^2(1-\mu)^2$	$(2y-1) \log \left( \frac{\mu}{1-\mu} \right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$
$\mu + \mu^2/\alpha$	$y \log \left( \frac{\mu}{\alpha+\mu} \right) + \alpha \log \left( \frac{\alpha}{\alpha+\mu} \right)$

Agora admite-se o seguinte modelo de quase-verossimilhança com função de variância paramétrica:

$$E(y_i) = \mu_i(\beta), \quad \text{Var}(y_i) = \phi V_\lambda(\mu_i),$$

em que  $\lambda$  é um parâmetro desconhecido na função de variância. Uma situação em que ocorre, naturalmente, a função de variância paramétrica, corresponde ao preditor linear  $\eta = X\beta$  tendo um componente aleatório independente extra  $\varepsilon$  de variância  $\lambda$  produzindo o preditor modificado  $\eta^* = \eta + \varepsilon$ . Até primeira ordem, obtêm-se a média e a variância modificadas  $E(y)^* = \mu + \varepsilon d\mu/d\eta$  e  $\text{Var}(y)^* = \phi V(\mu) + \lambda(d\mu/d\eta)^2$  e, portanto, a função de variância torna-se parametrizada por  $\lambda$ . Uma outra situação ocorre quando a variável resposta  $Y$  representa a soma de variáveis i.i.d. cujo número de variáveis é, também, uma variável aleatória de média  $\mu$  e variância  $V(\mu)$ . É fácil verificar que os parâmetros extras que aparecem na função de variância de  $y$  incluirão os dois primeiros momentos das variáveis i.i.d.

Para um valor fixo de  $\lambda$  pode-se ainda utilizar as equações dadas em (10.8) para obter as estimativas de máxima quase-verossimilhança dos  $\beta'$ s. A estimativa de  $\lambda$  corresponderá ao maior valor da quase-verossimilhança estendida maximizada tratada como função de  $\lambda$ , obtida de  $Q^+(\lambda)$ , ou ainda ao menor valor do *desvio estendido*  $-2Q^+(\lambda)$  dado por  $\min_\lambda -2Q^+(\lambda)$ . Seria melhor maximizar conjuntamente  $Q^+$  em relação a  $\beta$  e  $\lambda$ , embora este processo exija o cálculo da função escore em relação ao parâmetro  $\lambda$ , o que é bastante complicado.

Considera-se agora uma classe de modelos de quase-verossimilhança com

parâmetro de dispersão não-constante

$$\eta = g(\mu) = X\beta, \quad \tau = h(\phi) = Z\gamma, \quad (10.9)$$

em que  $\mu_i = E(Y_i)$ ,  $\text{Var}(Y_i) = \phi_i V(\mu_i)$ ,  $X$  e  $Z$  são matrizes  $n \times p$  e  $n \times q$  de posto completo  $p$  e  $q$ ,  $\beta$  e  $\gamma$  são vetores de parâmetros desconhecidos de dimensões  $p \times 1$  e  $q \times 1$ , respectivamente, com  $g(\cdot)$  e  $h(\cdot)$  funções de ligação conhecidas. Para  $\gamma$  fixo pode-se utilizar (10.8) para obter as estimativas de máxima quase-verossimilhança dos  $\beta$ 's e, então,  $\gamma$  será escolhido visando maximizar a quase-verossimilhança estendida maximal  $Q^+(\gamma)$  como função de  $\gamma$ . A estimativa de  $\gamma$  será o valor correspondente ao maior valor  $Q^+(\gamma)$ . A idéia básica é usar  $Q^+$  como o análogo da log-verossimilhança para se fazer inferência sobre  $\beta$  ou  $\gamma$ . As componentes quase-escore são dadas por

$$U_\beta^+ = \partial Q^+ / \partial \beta = X^T W H (y - \mu), \quad U_\gamma^+ = \partial \gamma = \frac{1}{2} Z^T L (D - \phi),$$

em que  $W = \text{diag}\{\phi^{-1} V(\mu)^{-1} g'(\mu)^{-2}\}$ ,  $H = \text{diag}\{\phi^{-2} h'(\mu)^{-1}\}$  e  $D = [D(y_1; \mu_1), \dots, D(y_n; \mu_n)]^T$ . As estimativas de quase-verossimilhança de  $\beta$  e  $\gamma$  são obtidas resolvendo o sistema não-linear resultante da igualdade de  $U_\beta^+$  e  $U_\gamma^+$  ao vetor nulo. Demonstra-se (Cordeiro e Demétrio, 1989) que as equações não-lineares para o cálculo simultâneo de  $\tilde{\beta}$  e  $\tilde{\gamma}$  podem ser dadas na forma iterativa

$$\tilde{X}^T \tilde{W}^{(m)} \tilde{X} \rho^{(m+1)} = \tilde{X}^T \tilde{W}^{(m)} \tilde{y}^{*(m)}, \quad (10.10)$$

em que

$$\tilde{X} = \begin{bmatrix} X & 0 \\ 0 & Z \end{bmatrix}, \quad \tilde{W} = \begin{bmatrix} W & 0 \\ 0 & 1/2C \end{bmatrix},$$

$$\tilde{H} = \begin{bmatrix} H & 0 \\ 0 & C^{-1}L \end{bmatrix}, \quad \tilde{y}^* = \begin{bmatrix} \eta \\ \tau \end{bmatrix} + \tilde{H} \begin{bmatrix} y - \mu \\ D - \phi \end{bmatrix},$$

$C = \text{diag}\{\phi^{-2} h'(\phi)^{-2}\}$ . A matriz  $C$  tem elementos obtidos da aproximação de primeira ordem  $E\{D(y; \mu)\} = 0$ .

Assim, ajustar o modelo de quase-verossimilhança (10.9) aos dados equivale a calcular repetidamente uma regressão linear ponderada de uma variável dependente modificada  $\tilde{y}^*$  sobre uma matrix  $\tilde{X}$  de dimensão  $2n \times (p + q)$  usando matrix de pesos  $\tilde{W}$  que também se modifica no processo. A implementação de (10.10) pode ser feita usando os *software* já citados nesta seção. Estas mesmas equações (10.10) continuam válidas para os *modelos lineares generalizados duplos* que são definidos pelo componente aleatório (1.5) e pelos dois componentes sistemáticos dados em (10.9).

## 10.7 Modelos para análise de dados de sobrevivência

Nesta seção, serão apresentados alguns modelos usuais para análise de dados em que a variável resposta é o tempo de sobrevivência. Por exemplo, o tempo que um certo tipo de máquina demora para quebrar ou o tempo de sobrevivência de um paciente submetido a um determinado tratamento. Geralmente esses dados apresentam uma característica específica chamada de “censura”, em virtude dos estudos terminarem quase sempre antes de se conhecer o resultado final de todas as unidades amostrais. No caso do tempo até a quebra de um certo tipo de máquina, é possível que o mesmo não seja conhecido para algumas unidades, pois as análises podem terminar antes da quebra de algumas máquinas. Os tempos dessas máquinas são tratados como censuras. Mesmo assim, esses são incorporados nos modelos de análise de sobrevivência.

O tempo de sobrevivência pode ser descrito formalmente através das seguintes funções: (i)  $f(t)$ , a densidade de probabilidade do tempo de sobrevivência; (ii)  $S(t)$ , a função de sobrevivência, sendo  $S(t) = 1 - F(t)$  e  $F(t)$  a função de distribuição acumulada; (iii)  $h(t)$ , a função de risco, que é uma medida do risco instantâneo de morte no tempo  $t$ , sendo definida por  $h(t) = F'(t)/\{1 - F(t)\}$ .

Conhecendo-se apenas uma dessas funções tem-se diretamente as outras duas. Por exemplo, para a distribuição exponencial com  $S(t) = \exp(-\lambda t)$ , fica claro que a função de risco é constante e dada por  $h(t) = \lambda$ . Para a distribuição de Weibull tem-se  $h(t) = \alpha t^{\alpha-1}$ ; logo,  $S(t) = \exp(-t^\alpha)$ . A função de risco nesse caso cresce com o tempo se  $\alpha > 1$  e decresce se  $\alpha < 1$ . O livro de Cox e Oakes (1984) apresenta um estudo completo da análise de dados de sobrevivência.

### 10.7.1 Modelos de riscos proporcionais

Em geral, a função de risco depende do tempo e de um conjunto de covariáveis, possivelmente, dependentes do tempo. O caso mais freqüente engloba uma componente que só depende do tempo, multiplicada pela componente dos efeitos das covariáveis. Esse modelo, denominado de riscos proporcionais com efeitos multiplicativos (vide Cox, 1972),



é expresso por

$$h(t; x) = \lambda(t) \exp(x^T \beta), \quad (10.11)$$

em que  $\beta = (\beta_1, \dots, \beta_p)^T$  é um vetor de parâmetros desconhecidos associados às covariáveis de  $x = (x_1, \dots, x_p)^T$ ,  $\lambda(t)$  é uma função não-negativa do tempo e  $\eta = x^T \beta$  é o preditor linear.

O modelo (10.11) implica que o quociente dos riscos para dois indivíduos num tempo qualquer, depende apenas da diferença dos preditores lineares desses indivíduos. A função de sobrevivência fica agora dada por

$$S(t; x) = \exp\{-\Lambda(t) \exp(x^T \beta)\}, \quad (10.12)$$

em que  $\Lambda(t) = \int_{-\infty}^t \lambda(u) du$ . Similarmente, a função densidade de probabilidade de  $T$  fica expressa na forma

$$f(t; x) = \Lambda'(t) \exp\{\eta - \Lambda(t) \exp(\eta)\}.$$

A distribuição do tempo de sobrevivência  $T$  do modelo acima pertence à família exponencial não-linear, mas não à família (1.5). Em particular,  $E\{\Lambda(t)\} = \exp(-\eta)$  e  $\text{Var}\{\Lambda(t)\} = \exp(-2\eta)$ .

A estimação dos  $\beta$ 's para uma função  $\lambda(t)$  especificada foi desenvolvida por Aitkin e Clayton (1980). Admite-se durante o tempo de obtenção dos dados, que foram registrados os tempos de morte de  $n - m$  indivíduos e os tempos de censura de  $m$  indivíduos. Seja uma variável dicotômica  $y_i$  que assume valor um se o indivíduo  $x_i$  morreu e valor zero se esse foi censurado no tempo  $t_i$ . Logo, um indivíduo que morreu no tempo  $t_i$  contribui com o fator  $\log f(t_i; x_i)$  para a log-verossimilhança  $\ell(\beta)$ , enquanto um indivíduo censurado em  $t_i$  contribui com  $\log S(t_i; x_i)$ . A função  $\ell(\beta)$  reduz-se à

$$\ell(\beta) = \sum_{j=1}^n \{y_i \log f(t_i; x_i) + (1 - y_i) \log S(t_i; x_i)\},$$

que pode ser expressa numa forma mais conveniente usando (10.12) como

$$\ell(\beta) = \sum_{j=1}^n (y_i \log \mu_i - \mu_i) + \sum_{j=1}^n \log\{\lambda(t_i)/\Lambda(t_i)\}, \quad (10.13)$$

em que  $\mu_i = \Lambda(t_i) \exp(\eta_i)$ . A segunda soma de (10.13) não depende dos  $\beta$ 's e, portanto, (10.13) tem a mesma forma do logaritmo da função de verossimilhança de um modelo

Tabela 10.2: Alguns modelos usuais para a análise de dados de sobrevivência

Modelo	$\lambda(t)$	densidade	“offset”
exponencial	$\lambda$	$\lambda \exp\{\eta - \lambda t \exp(\eta)\}$	$\log(\lambda t)$
Weibull	$\alpha t^{\alpha-1}$	$\alpha t^{\alpha-1} \exp\{\eta - t^\alpha \exp(\eta)\}$	$\alpha \log t$
valor-extremo	$\alpha \exp(\alpha t)$	$\alpha \exp\{\eta - t^\alpha \exp(\alpha t + \eta)\}$	$\alpha t$

de Poisson com  $n$  observações independentes  $y_1, \dots, y_n$ , médias  $\mu_1, \dots, \mu_n$ , e preditores lineares que são dados por  $\eta_i = \log \Lambda(t_i)$ ,  $i = 1, \dots, n$ .

As estimativas de máxima verossimilhança dos  $\beta'$ s podem ser obtidas pelos sistemas R, S-PLUS e MINITAB, ajustando aos dados binários  $y_i$  um modelo log-linear com “offset”  $\log \Lambda(t_i)$ . A estimação, em geral, não será um processo simples, pois o “offset” e  $\log\{\lambda(t_i)/\Lambda(t_i)\}$  podem conter os parâmetros desconhecidos definidos em  $\lambda(t)$ . Inferência sobre os  $\beta'$ s é feita da maneira usual.

A Tabela 10.2 apresenta três modelos usuais para o tempo de sobrevivência. O modelo exponencial com  $\lambda$  conhecido pode ser ajustado diretamente. Se  $\lambda$  não for conhecido, a sua estimativa de máxima verossimilhança é igual a  $(n-m)/\sum_{i=1}^n t_i \exp(\hat{\eta}_i)$ , mas os preditores estimados dependem do “offset”, que envolve  $\lambda$ . Um processo iterativo de estimação conjunta de  $\lambda$  e dos  $\beta'$ s pode ser realizado interagindo a estimativa de máxima verossimilhança de  $\lambda$  com as estimativas dos parâmetros do modelo log-linear de “offset”  $\log(\lambda t)$  especificado. Entretanto, se não há interesse em conhecer a estimativa de  $\lambda$ , o termo  $\log(\lambda)$  do “offset” pode ser incorporado à constante do preditor linear  $\eta_i$ , ficando o modelo log-linear na forma  $\log \mu_i = \log t_i + \eta_i$ , com “offset” dado por  $\log t_i$ .

Para o modelo de Weibull com  $\alpha$  desconhecido, a estimativa de máxima verossimilhança de  $\alpha$  é dada por

$$\hat{\alpha} = (n-m) / \sum_{i=1}^n (\hat{\mu}_i - y_i) \log t_i. \quad (10.14)$$

Admite-se uma estimativa inicial para  $\alpha$  e ajusta-se a  $y$ , um modelo log-linear com “offset”  $\alpha \log t$ . De (10.14) reestima-se  $\alpha$ , continuando o processo até a convergência.

O modelo de valor extremo pode ser transformado no de Weibull com a transformação  $\exp(t)$ , no lugar de  $t$ .

### 10.7.2 Riscos proporcionais de Cox

Cox (1972) iniciou uma fase importante na análise de dados de sobrevivência, definindo uma versão semi-paramétrica para o modelo de riscos proporcionais dado em (10.11). Em vez de supor que  $\lambda(t)$  é uma função regular de  $t$ , Cox definiu  $\lambda(t)$  como sendo uma função arbitrária de  $t$ , que assume valores arbitrários nos tempos em que ocorreram as falhas (mortes), porque a função de risco definida nesses intervalos não contribui para a log-verossimilhança dada em (10.14). Note que a estimativa  $\hat{\beta}$  depende somente de  $\lambda(t)$  definida nos tempos em que ocorreram as mortes.

Considere inicialmente os tempos de falhas  $t_1, t_2, \dots, t_k$  como sendo distintos, sem a ocorrência de empates. Seja  $R(t_j)$  o conjunto de risco imediatamente anterior a  $t_j$ , isto é, o conjunto de indivíduos para os quais a falha não ocorreu antes de  $t_j$ . Então, dado que ocorreu uma falha no tempo  $t_j$ , a probabilidade segundo o modelo (10.11), dessa falha ter ocorrido com o  $i$ -ésimo indivíduo, é dada por

$$P_j = \frac{\lambda(t) \exp(x_i^T \beta)}{\sum_{s \in R(t_j)} \lambda(t) \exp(x_s^T \beta)} = \frac{\exp(x_i^T \beta)}{\sum_{s \in R(t_j)} \exp(x_s^T \beta)},$$

em que o somatório é sobre o conjunto de risco  $R(t_j)$ ,

A log-verossimilhança (parcial)  $\log P_j$  pode ser expressa na forma exponencial dada em (1.5), considerando como resposta o vetor de covariáveis do indivíduo que falhou em  $t_j$ , e como fixo o conjunto de covariáveis de todos os indivíduos pertencentes à  $R(t_j)$ . Dessa forma, denotando por  $Y_i$  a resposta para esse indivíduo, tem-se

$$\log P_j = y_i^T \beta - \log \left\{ \sum_{s \in R(t_j)} \exp(x_s^T \beta) \right\},$$

que equivale à família exponencial de distribuições com parâmetro canônico  $\beta$  e  $b(\beta) = \log\{\sum_s \exp(x_s^T \beta)\}$ . A média (condicional) e a função de variância são, respectivamente, definidos por  $b'(\beta)$  e  $b''(\beta)$ . Entretanto, essa forma simplificada para  $\log P_j$  não é adequada do ponto de vista computacional, em particular no sentido de se aplicar o processo iterativo, definido na Seção 3.2 para a obtenção de  $\hat{\beta}$ . Aqui, a função de variância  $b''(\beta)$  não é uma função explícita da média, dificultando a adaptação do processo iterativo definido por (3.5).

Em McCullagh e Nelder (1989) há uma discussão sobre métodos iterativos para a estimação de  $\beta$ . Whitehead (1980) mostra que a maximização da log-verossimilhança

conjunta  $L(\beta) = \sum \log P_j$  é equivalente à maximização de uma log-verossimilhança de  $n$  variáveis de Poisson independentes. Note-se que se  $R(t_j)$  tem  $M + 1$  elementos, para todo  $j$ , então  $\ell(\beta)$  coincide com a log-verossimilhança definida em (10.13) para o modelo logístico condicional aplicado aos estudos com dados emparelhados.

O principal problema que aparece nas aplicações do modelo de Cox é a ocorrência de empates entre os tempos  $t'_j$ s. Em situações experimentais que envolvem a aplicação de drogas em animais, geralmente o tempo de sobrevivência desses animais é contado em dias, sendo inevitável a ocorrência de empates. Em outras situações práticas, esse problema também aparece com uma certa frequência.

O complicador nesses casos é que o logaritmo da função de verossimilhança  $\ell(\beta)$  pode ficar expresso numa forma bastante complexa, tornando proibitiva a aplicação de qualquer processo iterativo para estimação dos  $\beta'$ s. Para ilustrar, suponha que os indivíduos  $x_1$  e  $x_2$  falharam no mesmo tempo; logo, a probabilidade real de ocorrerem essas falhas no tempo  $t_j$  é igual à probabilidade do indivíduo  $x_i$  ter falhado antes do indivíduo  $x_2$ , mais essa mesma probabilidade no sentido inverso, isto é,

$$P_{j(\text{Real})} = \frac{\exp(x_1^T \beta)}{\sum_{s \in R(t_j)} \exp(x_s^T \beta)} \cdot \frac{\exp(x_2^T \beta)}{\left\{ \sum_{s \in R(t_j)} \exp(x_s^T \beta) - \exp(x_1^T \beta) \right\}} + \frac{\exp(x_2^T \beta)}{\sum_{s \in R(t_j)} \exp(x_s^T \beta)} \cdot \frac{\exp(x_1^T \beta)}{\left\{ \sum_{s \in R(t_j)} \exp(x_s^T \beta) - \exp(x_2^T \beta) \right\}}.$$

Cox (1975) mostra que toda a teoria usual para a estatística da razão de verossimilhanças continua valendo para os modelos de riscos proporcionais.

## 10.8 Modelos lineares generalizados com covariáveis de dispersão

Jørgensen (1987) definiu a classe dos *modelos de dispersão*, inicialmente denominada classe estendida de MLG (Jørgensen, 1983, Cordeiro, 1985), considerando um conjunto de variáveis aleatórias  $Y_1, \dots, Y_n$  com cada  $Y_\ell$  tendo função densidade (ou função de

probabilidade) na forma

$$\pi(y; \theta_i, \phi) = \exp\{\phi t(y, \theta_i) + c_1(y, \phi)\}, \quad (10.15)$$

em que  $t(\cdot, \cdot)$  e  $c_1(\cdot, \cdot)$  são funções conhecidas. Consideramos que  $\phi$  ( $\phi > 0$ ) é constante para todas as observações embora, possivelmente, desconhecido. Denominamos  $\phi^{-1}$  de parâmetro de dispersão e  $\phi$  de parâmetro de precisão. Segundo Jørgensen (1983) os modelos definidos em (10.15) incluem a possibilidade de erros correlacionados. Entretanto, se as variáveis aleatórias  $Y_1, \dots, Y_n$  forem independentes, com cada variável tendo uma distribuição da forma (10.15), a distribuição conjunta de  $Y_1, \dots, Y_n$  será também da forma (10.15).

Fazendo  $t(y, \theta) = y\theta - b(\theta)$  em (10.15), obtemos a subclasse dos *modelos exponenciais de dispersão* (Jørgensen, 1987) ou MLGs. Para  $\phi$  conhecido, os modelos exponenciais de dispersão pertencem à família exponencial de distribuições, sendo  $\theta$  o seu parâmetro canônico. Se  $\phi$  for desconhecido, estes modelos podem ou não pertencer à família exponencial de distribuições indexada por dois parâmetros.

Barndorff-Nielsen e Jørgensen (1991) definiram uma subclasse de modelos de dispersão, em que a função  $c_1(y, \phi)$  em (10.15) é aditiva, da forma  $d_1(y) + d_2(\phi)$ , os quais são denominados *modelos próprios de dispersão*. Estes modelos apresentam duas propriedades importantes. A primeira mostra que a estatística  $t(y, \theta)$  é uma estatística pivotal para  $\theta$ , isto é, a distribuição de  $t(y, \theta)$  não depende de  $\theta$  para  $\phi$  conhecido. A segunda revela que, para  $\theta$  conhecido, a função densidade (ou probabilidade) definida em (10.15) pertence à família exponencial uniparamétrica sendo  $t(y, \theta)$  uma estatística canônica.

Sejam  $Y_1, \dots, Y_n$  um conjunto de  $n$  variáveis aleatórias independentes com cada  $Y_\ell$  tendo função densidade (ou função de probabilidade) na família exponencial

$$\pi(y; \theta_i, \phi_i) = \exp[\phi_i \{y\theta_i - b(\theta_i) + c(y)\} + d_1(y) + d_2(\phi_i)], \quad (10.16)$$

em que  $b(\cdot)$ ,  $c(\cdot)$ ,  $d_1(\cdot)$  e  $d_2(\cdot)$  são funções conhecidas e  $\theta_i$  e  $\phi_i$  são, respectivamente, os  $i$ -ésimos elementos de  $\theta$  e  $\phi$ , vetores de dimensão  $n \times 1$ . A média e a variância de  $Y_i$  são  $E(Y_i) = \mu_i = db(\theta_i)/d\theta_i$  e  $\text{Var}(Y_i) = \phi_i^{-1}V_i$ , em que  $V = d\mu/d\theta$  e  $\theta = \int V^{-1}d\mu = q(\mu)$  é uma função conhecida unívoca de  $\mu$ . A componente sistemática usual para a média é  $f(\mu) = \eta = X\beta$ , em que  $f(\cdot)$  é a função de ligação,  $\eta = (\eta_1, \dots, \eta_n)^T$  é o preditor linear,  $X$  é uma matriz conhecida  $n \times p$  de posto  $p < n$  e  $\beta = (\beta_1, \dots, \beta_p)^T$  é um vetor

de parâmetros desconhecidos a ser estimado. Os parâmetros  $\theta_i$  e  $\phi_i^{-1} > 0$  são chamados de parâmetros canônico e de dispersão, respectivamente. Ambos os parâmetros variam sobre as observações através de modelos de regressão. Para as distribuições normal, gama e Gaussiana inversa, as médias e as variâncias são  $\theta_i^{-1}, -\theta_i^{-1}, (-2\theta_i)^{-1/2}$  e  $\phi_i^{-1}, \phi_i^{-1}\mu_1^2$  e  $\phi_i^{-1}\mu_1^3$ , respectivamente.

Definimos a componente sistemática do vetor de parâmetros de precisão  $\phi = (\phi_1, \dots, \phi_n)^T$  como

$$g(\phi) = \tau = S\gamma, \quad (10.17)$$

em que  $\tau$  é o preditor linear da dispersão,  $S = (s_1, \dots, s_n)^T$ , com  $s_i = (s_{i1}, \dots, s_{ip})^T$ , é uma matriz  $n \times q$  de posto  $q$  ( $q < n$ ) representando as variáveis independentes que modelam a dispersão e  $\gamma = (\gamma_1, \dots, \gamma_q)^T$  é, também, um vetor de parâmetros desconhecidos. O MLG com covariáveis de dispersão tem, portanto, dois preditores lineares:  $\eta$  – o preditor linear da média e  $\tau$  – o preditor linear da dispersão. Ambas  $f(\cdot)$  e  $g(\cdot)$  são funções um a um conhecidas e duplamente diferenciáveis. A função  $g(\cdot)$  é chamada de função de ligação da dispersão. Assume-se, também, que  $\beta$  é independente de  $\gamma$ . Temos, então,  $p+q$  parâmetros a serem estimados.

Considere o logaritmo da função de verossimilhança total como função de  $\beta$  e  $\gamma$

$$\ell(\beta, \gamma) = \sum_{i=1}^n \{\phi_i[y_i\theta_i - b(\theta_i) + c(y_i)] + d_1(y_i) + d_2(\phi_i)\},$$

sendo o vetor de dados  $y = (y_1, \dots, y_n)^T$  fixado, em que  $y_i$  denota o valor observado da variável aleatória  $Y_i$ . Na expressão acima,  $\theta$  está associado a  $\beta$  através da função de ligação  $f(\cdot)$  ( $\theta$  é uma função de  $\mu$ ) e  $\phi$  está relacionado com  $\gamma$  através de  $g(\cdot)$ .

Denotamos a função escore total por

$$\mathbf{U} = \mathbf{U}(\beta, \gamma) = \begin{pmatrix} \partial\ell(\beta, \gamma)/\partial\beta \\ \partial\ell(\beta, \gamma)/\partial\gamma \end{pmatrix}$$

cujos componentes são

$$\partial\ell(\beta, \gamma)/\partial\beta = \mathbf{X}^T \Phi \mathbf{W}^{1/2} \mathbf{V}^{-1/2} (\mathbf{y} - \boldsymbol{\mu}) \quad \text{e} \quad \partial\ell(\beta, \gamma)/\partial\gamma = \mathbf{S}^T \Phi_1 \mathbf{v},$$

em que  $\Phi = \text{diag}\{\phi_1, \dots, \phi_n\}$ ,  $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$  com  $w_i = V_i^{-1}(d\mu_i/d\eta_i)^2$ ,  $\mathbf{V} = \text{diag}\{V_1, \dots, V_n\}$ ,  $\Phi_1 = \text{diag}\{\phi_{1i}, \dots, \phi_{1n}\}$  com  $\phi_{1i} = \partial\phi_i/\partial\eta_i$  e  $\mathbf{v} = (v_1, \dots, v_n)^T$  com  $v_i = y_i\theta_i - b(\theta_i) + c(y_i) + \partial d_2(\phi_i)/\partial\phi_i$ .

A partição  $(\beta^T, \gamma^T)$  induz uma correspondente matriz de informação particionada para estes parâmetros. A matriz de informação total de Fisher  $\mathbf{K} = \mathbf{K}(\beta, \gamma)$  pode ser deduzida de  $E\{\mathbf{U}(\beta, \gamma)\mathbf{U}^T(\beta, \gamma)\}$ . Esta matriz é bloco-diagonal dada por

$$\mathbf{K}(\beta, \gamma) = \begin{bmatrix} \mathbf{K}_{\beta, \beta} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{\gamma, \gamma} \end{bmatrix}$$

em que  $\mathbf{K}_{\beta, \beta} = \mathbf{X}^T \mathbf{W} \Phi \mathbf{X}$  e  $\mathbf{K}_{\gamma, \gamma} = -\mathbf{S}^T \mathbf{D}_2 \Phi_1^2 \mathbf{S}$ , sendo  $\mathbf{D}_2 = \text{diag}\{d_{21}, \dots, d_{2n}\}$ ,  $d_{2i} = \partial^2 d_2(\phi_i) / \partial \phi_i^2$  e  $\Phi_1^2 = \text{diag}\{\phi_{11}^2, \dots, \phi_{1n}^2\}$ , são as matrizes de informação para  $\beta$  e  $\gamma$ , respectivamente. Os parâmetros  $\beta$  e  $\gamma$  são globalmente ortogonais e suas estimativas de máxima verossimilhança são assintoticamente independentes (Cox e Reid, 1987).

Os estimadores de máxima verossimilhança  $\hat{\beta}$  e  $\hat{\gamma}$  podem ser calculados através do processo iterativo score de Fisher, resolvendo as seguintes equações

$$\begin{bmatrix} \hat{\beta}^{(m+1)} \\ \hat{\gamma}^{(m+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}^{(m)} \\ \hat{\gamma}^{(m)} \end{bmatrix} + \mathbf{K}^{(m)-1} \mathbf{U}^{(m)}. \quad (10.18)$$

As equações (10.18) implicam na solução iterativa do sistema de equações

$$\tilde{\mathbf{X}}^T \tilde{\mathbf{W}}^{(m)} \tilde{\mathbf{X}} \rho^{(m+1)} = \tilde{\mathbf{X}}^T \tilde{\mathbf{W}}^{(m)} \tilde{\mathbf{y}}^{*(m)}, \quad (10.19)$$

em que

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & -\mathbf{S} \end{bmatrix}, \quad \tilde{\mathbf{W}} = \begin{bmatrix} \Phi \mathbf{W} & \mathbf{0} \\ \mathbf{0} & D_2 \Phi_1^2 \end{bmatrix},$$

$$\tilde{\Phi} = \begin{bmatrix} \mathbf{W}^{-1/2} \mathbf{V}^{-1/2} & \mathbf{0} \\ \mathbf{0} & -D_2^{-1} \Phi_1^{-1} \end{bmatrix}, \quad \rho = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$$

e

$$\tilde{\mathbf{y}}^* = \begin{bmatrix} \eta \\ \tau \end{bmatrix} + \tilde{\Phi} \begin{bmatrix} \mathbf{y} - \mu \\ \mathbf{v} \end{bmatrix}.$$

Em geral, na resolução iterativa de 10.19 tem-se que fazer a regressão da variável dependente modificada  $\tilde{\mathbf{y}}^*$  sobre a matriz modelo  $\tilde{\mathbf{X}}$  usando os pesos modificados definidos por  $\tilde{\mathbf{W}}$ . A variável dependente modificada  $\tilde{\mathbf{y}}^*$  também varia durante o procedimento iterativo e deve ser recalculada em toda repetição do processo iterativo. O procedimento inicial é feito pela escolha de valores arbitrários para  $\beta$  e  $\gamma$ .

## 10.9 Modelos lineares generalizados com superdispersão

Na prática o fenômeno de super-dispersão não é incomum, e foi considerado amplamente na literatura, particularmente em relação às distribuições binomial e Poisson. Pelo termo de super-dispersão queremos dizer que a variância da variável resposta excede a variância da variável nominal (McCullagh e Nelder, 1989). A incidência e o grau de super-dispersão encontrados dependem do campo de aplicação. Há diferentes causas de super-dispersão. Em algumas circunstâncias a causa pode ser do processo de coleta de dados, correlação entre respostas individuais e variáveis omitidas. Uma consequência da super-dispersão é que os erros-padrão das estimativas do modelo estarão incorretos e, também, que os desvios serão muito grandes conduzindo à seleção de modelos complexos. O problema da super-dispersão é fácil de reconhecer mas difícil de estudar em generalidade. Aplicando os MLG com uma relação variância-média especificada e com um parâmetro de dispersão multiplicativo, muitas vezes obtém-se um ajustamento do modelo em que a variância é maior do que o preditor da média.

Dey et al. (1997) definiram uma classe de *MLG com super-dispersão* em que as variáveis aleatórias  $Y_1, \dots, Y_n$  são independentes e cada  $Y_i$  tem densidade (ou função de probabilidade) com dois parâmetros pertencente à família exponencial

$$\pi(y; \mu, \phi) = A(y) \exp\{(y - \mu)\psi^{(1,0)}(\mu, \phi) + \phi T(y) + \psi(\mu, \phi)\}, \quad (10.20)$$

em que  $A(\cdot)$ ,  $T(\cdot)$  e  $\psi(\cdot, \cdot)$  são funções conhecidas e  $\psi^{(r,s)} = \partial^r \psi^{(1,0)}(\mu, \phi) / \partial \mu^r \partial \phi^s$ . A média e a variância de  $Y$  são  $E(Y) = \mu$  e  $\text{Var}(Y) = \psi^{(2,0)^{-1}}$ , e a média e a variância de  $T(Y)$  são  $E\{T(Y)\} = -\psi^{(0,1)}$  e  $\text{Var}\{T(Y)\} = -\psi^{(0,2)}$ . Além disso,  $\text{Cov}(Y, T(Y)) = 0$ .

Gelfand e Dalal (1990) mostraram que se (10.20) é integrável em relação a  $y$  e se a função  $T(y)$  é convexa, tendo a média  $\mu$  fixa, então  $\text{Var}(Y)$  aumenta com  $\phi$  e, portanto, esse parâmetro representa uma dispersão.

A família exponencial uniparamétrica é obtida de (10.20) com  $\phi = 0$ , conduzindo a forma

$$\pi(y; \phi, 0) = A(y) \exp\{y\theta - b(\theta)\},$$

em que  $\theta = \psi^{(1,0)}(\mu, 0)$  e  $b(\theta) = -\psi(\mu, 0) + \mu\psi^{(1,0)}(\mu, 0)$ .

Considera-se MLG com superdispersão que têm duas componentes sistemáticas que são parametrizadas como  $f(\mu) = \eta = X\beta$  e  $g(\phi) = \tau = S\gamma$ , sendo  $X$  e  $S$  matrizes  $n \times p$  e  $n \times q$ , de postos  $p$  e  $q$ , respectivamente e  $\beta = (\beta_1, \dots, \beta_p)^T$  e  $\gamma = (\gamma_1, \dots, \gamma_q)^T$



vetores de parâmetros desconhecidos a serem estimados. Considera-se que  $f(\cdot)$  e  $g(\cdot)$  são funções monótonas conhecidas e diferenciáveis e que  $\beta$  é independente de  $\gamma$ . A função  $g(\cdot)$  é uma função de ligação adicional chamada de função de ligação de dispersão. O MLG é baseado na família exponencial (1.5) de um parâmetro assumindo  $\phi$  fixo sendo  $\theta = q(\mu)$  o parâmetro natural,  $\mu = db\theta)/d\theta$  a média e  $\phi$  o parâmetro de dispersão comum para todas as observações, embora possivelmente desconhecido. As únicas distribuições contínuas da forma (1.5) são baseadas nas distribuições normal, gama e Gaussiana inversa.

Note-se que a família de distribuições em (1.5) é uma sub-família simples de (10.20) e difere desta no sentido de que tem uma forma geral de dois parâmetros para modelos exponenciais, enquanto (1.5) é apenas um modelo exponencial de um parâmetro  $\theta$  quando  $\phi$  é mantido fixo. Entretanto, como um modelo de dois parâmetros  $(\theta, \phi)$ , (10.20) não tem a forma do modelo exponencial. Deste modo, o MLG com super-dispersão, como definido acima, é uma extensão dos MLGs.

Para um determinado MLG com super-dispersão, o objetivo é calcular as estimativas dos parâmetros  $\beta$  e  $\gamma$  simultaneamente, desde que eles representam os efeitos das variáveis explicativas da média e do parâmetro de dispersão, respectivamente. Denotamos a amostra aleatória por  $y_1, \dots, y_n$  e o logaritmo da função de verossimilhança total por

$$\ell(\beta, \gamma) = \sum_{i=1}^n \{(y_i - \mu_i)\psi^{(1,0)}(\mu, \phi_i) + \phi_i T(y_i) + \psi(\mu_i, \phi_i)\} + \sum_{i=1}^n \log A(y_i). \quad (10.21)$$

Esta função é suposta regular (Cox e Hinkley, 1974; Capítulo 9) com relação às derivadas em  $\beta$  e  $\gamma$  até terceira ordem. A inferência sobre  $\beta$  e  $\gamma$  pode ser feita através do método de verossimilhança, análogos aos dos MLG com covariáveis de dispersão (Cordeiro e Botter, 2000). O vetor escore é dado na forma

$$\mathbf{U} = \mathbf{U}(\beta, \gamma) = \begin{bmatrix} \frac{\partial \ell(\beta, \gamma)}{\partial \beta} \\ \frac{\partial \ell(\beta, \gamma)}{\partial \gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \boldsymbol{\psi}^{(2,0)} \mathbf{M}_1 (\mathbf{y} - \boldsymbol{\mu}) \\ \mathbf{S}^T \boldsymbol{\Phi}_1 \boldsymbol{\nu} \end{bmatrix}, \quad (10.22)$$

em que  $\mathbf{y} - \boldsymbol{\mu} = (y_1 - \mu_1, \dots, y_n - \mu_n)^T$  e  $\mathbf{v} = (v_1, \dots, v_n)^T$  com  $v_i = \psi_i^{(1,1)}(y_i - \mu_i) + T(y_i) + \psi_i^{(0,1)}$ . E mais,  $m_{ri} = \frac{d^r \mu_i}{d\eta_i^r}$  e  $\phi_{ri} = \frac{d^r \phi_i}{d\tau_i^r}$  são, respectivamente, as derivadas das funções de ligação inversas  $\mu = f^{-1}(\eta)$  e  $\phi = g^{-1}(\tau)$ ,  $r = 1, 2$  e  $i = 1, \dots, n$ . Definem-se, também, as seguintes matrizes diagonais  $n \times n$ :  $\mathbf{M}_r = \text{diag}\{m_{r1}, \dots, m_{rn}\}$  e  $\boldsymbol{\Phi}_r = \text{diag}\{\phi_{r1}, \dots, \phi_{rn}\}$  para  $r = 1, 2$  e  $\boldsymbol{\psi}^{(2,0)} = \text{diag}\{\psi_1^{(2,0)}, \dots, \psi_n^{(2,0)}\}$  e  $\boldsymbol{\psi}^{(0,2)} = \text{diag}\{\psi_1^{(0,2)}, \dots, \psi_n^{(0,2)}\}$ .

A partição  $(\beta^T, \gamma^T)^T$  induz uma matriz de informação total para estes parâmetros que são de interesse para a inferência de verossimilhança. A matriz de informação bloco-diagonal é dada por

$$\mathbf{K}(\beta, \gamma) = \begin{bmatrix} \mathbf{K}_{\beta, \beta} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{\gamma, \gamma} \end{bmatrix}, \quad (10.23)$$

em que  $\mathbf{K}_{\beta, \beta} = \mathbf{X}^T \psi^{(2,0)} \mathbf{M}_1^2 \mathbf{X}$  e  $\mathbf{K}_{\gamma, \gamma} = \mathbf{S}^T \psi^{(0,2)} \Phi_1^2 \mathbf{S}$  são as matrizes de informação de  $\beta$  e  $\gamma$ , respectivamente. Deste modo, os parâmetros  $\beta$  e  $\gamma$  são ortogonais e suas estimativas de máxima verossimilhança  $\hat{\beta}$  e  $\hat{\gamma}$  são assintoticamente independentes.

As EMV  $\hat{\beta}$  e  $\hat{\gamma}$  satisfazem equações não-lineares  $\mathbf{U}(\hat{\beta}, \hat{\gamma}) = 0$  que derivam de (10.22) e (10.23) e que podem ser resolvidos pelo método score de Fisher. Com isso, Cordeiro e Botter (2000) obtiveram as seguintes equações para estimar iterativamente  $\beta$  e  $\gamma$

$$\begin{aligned} \mathbf{X}^T \psi^{(2,0)(m)} \mathbf{M}_1^{(m)^2} \mathbf{X} \beta^{(m+1)} &= \mathbf{X}^T \psi^{(2,0)(m)} \mathbf{M}_1^{(m)^2} \varepsilon_1^{(m)}, \\ \mathbf{S}^T \psi^{(0,2)(m)} \Phi_1^{(m)^2} \mathbf{S} \gamma^{(m+1)} &= \mathbf{S}^T \psi^{(0,2)(m)} \Phi_1^{(m)^2} \varepsilon_2^{(m)}, \end{aligned} \quad (10.24)$$

em que  $\varepsilon_1 = \boldsymbol{\eta} + \mathbf{M}_1^{-1}(\mathbf{y} - \boldsymbol{\mu})$  e  $\varepsilon_2 = \boldsymbol{\tau} + \psi^{(0,2)^{-1}} \Phi_1^{-1}$  são vetores  $n \times 1$ .

As equações (10.24) mostram que qualquer *software* que permitir ajustar uma regressão linear ponderada pode ser usado para calcular as estimativas  $\hat{\beta}$  e  $\hat{\gamma}$ . Em termos gerais, temos que fazer a regressão iterativa da variável dependente modificada  $\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$  sobre a matriz modelo  $(\mathbf{X} \ \mathbf{S})$  com os pesos modificados definidos por

$$\begin{bmatrix} \psi^{(2,0)} \mathbf{M}_1^2 & \mathbf{0} \\ \mathbf{0} & \psi^{(0,2)} \Phi_1^2 \end{bmatrix}.$$

Esse ciclo deve ser repetido até se obter convergência. O procedimento de iteração em (10.23) é fácil de ser executado usando um algoritmo iterativo em algum *software* tipo R, MATLAB, SAS e Ox, seguindo as mesmas linhas descritas em Cordeiro e Paula (1989) e Cordeiro e Demétrio (1989). O início do procedimento é executado escolhendo valores arbitrários para  $\beta$  e  $\gamma$ .

## 10.10 Modelos com matriz de covariância não-escalar

Considera-se o modelo de regressão

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Psi}, \quad (10.25)$$

em que ambos  $\sigma^2$  e  $\boldsymbol{\Psi}$  são desconhecidos. No caso mais geral,  $\boldsymbol{\Psi}$  conterá  $n(n+1)/2 - 1$  parâmetros distintos, igual ao número de elementos da diagonal mais metade daqueles fora da diagonal menos um, um sendo subtraído pois está fatorado em  $\sigma^2 \boldsymbol{\Psi}$ . Dois casos especiais importantes de (10.25) são os modelos heterocedásticos e os modelos de autocorrelação descritos nas Seções 10.12 e 10.13, respectivamente. Se  $\boldsymbol{\Psi}$  for conhecido, o *estimador de mínimos quadrados generalizado* (EMQG) será  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Psi}^{-1} \mathbf{y}$  que é o estimador de mínima variância na classe dos estimadores lineares não-viesados de  $\boldsymbol{\beta}$ . Se  $\boldsymbol{\varepsilon}$  tem, também, distribuição normal, então  $\hat{\boldsymbol{\beta}}$  é o EMV sendo de mínima variância na classe dos estimadores não-viesados. Adicionalmente,  $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) / n$  é o estimador viesado de  $\sigma^2$ . Se o interesse é testar a hipótese nula de restrições lineares  $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{0}$ , em que  $\mathbf{R}$  é uma matriz  $r \times p$  de coeficientes conhecidos, a estatística

$$F = \hat{\boldsymbol{\beta}}^T \mathbf{R}^T [\mathbf{R}(\mathbf{X}^T \boldsymbol{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{R}^T]^{-1} \mathbf{R}\hat{\boldsymbol{\beta}} / r \hat{\sigma}^2$$

tem distribuição nula  $F_{r, n-p}$ , que pode ser usada tanto para testar  $H_0$  quanto na estimação restrita de intervalos para  $\boldsymbol{\beta}$ .

Quando  $\boldsymbol{\Psi}$  é desconhecido, situação mais comum na prática, o EMQG dado anteriormente é inviável. Neste caso, pode-se formar o estimador

$$\hat{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T \hat{\boldsymbol{\Psi}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\Psi}}^{-1} \mathbf{y}, \quad (10.26)$$

em que a matriz de covariância desconhecida  $\boldsymbol{\Psi}$  é substituída em (10.26) por um estimador consistente  $\hat{\boldsymbol{\Psi}}$ . Como o número de parâmetros desconhecidos em  $\boldsymbol{\Psi}$  é de ordem  $O(n)$ , em geral restringe-se o número desses parâmetros supondo que  $\boldsymbol{\Psi}$  é função de um vetor  $\boldsymbol{\gamma}$  de  $q+1$  parâmetros desconhecidos.

Vamos considerar a estimação de máxima verossimilhança (MV) de  $\boldsymbol{\beta}$ ,  $\sigma^2$  e  $\boldsymbol{\gamma}$  no modelo

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{\Psi}(\boldsymbol{\gamma})), \quad (10.27)$$

em que enfatizamos em (10.27) que a matriz  $\boldsymbol{\Psi}$  depende de um vetor  $q \times 1$  de parâmetros extras desconhecidos. A estimação de MV de  $\boldsymbol{\beta}$  e  $\sigma^2$  condicional a  $\boldsymbol{\gamma}$  produz os estimadores

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = (\mathbf{X}^T \boldsymbol{\Psi}(\boldsymbol{\gamma})^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Psi}(\boldsymbol{\gamma})^{-1} \mathbf{y} \quad (10.28)$$

e

$$\tilde{\sigma}(\gamma)^2 = (\mathbf{y} - \mathbf{X}\tilde{\beta}(\gamma))^T \Psi(\gamma)^{-1} (\mathbf{y} - \mathbf{X}\tilde{\beta}(\gamma)) / n. \quad (10.29)$$

Usa-se a notação  $\tilde{\beta}(\gamma)$ ,  $\tilde{\sigma}^2(\gamma)$  e  $\Psi(\gamma)$  acima para enfatizar a dependência dessas quantidades em  $\gamma$ . O logaritmo da função de verossimilhança perfilada para  $\gamma$  é

$$\ell_p(\gamma) = -n \log\{\tilde{\sigma}(\gamma)^2\} - \log\{\Psi(\gamma)\}. \quad (10.30)$$

A maximização de (10.30), em geral, não produz forma fechada para  $\tilde{\gamma}$  e procedimentos iterativos devem ser usados para obter o EMV  $\tilde{\gamma}$ , e, então,  $\tilde{\Psi} = \Psi(\tilde{\gamma})$ . Os estimadores incondicionais de  $\beta$  e  $\sigma^2$  são facilmente deduzidos de (10.28) – (10.29) como  $\tilde{\beta} = \tilde{\beta}(\tilde{\gamma})$  e  $\tilde{\sigma}^2 = \tilde{\sigma}(\tilde{\gamma})^2$ .

Pode-se demonstrar que a matriz de informação conjunta para  $\theta = (\beta^T, \sigma^2, \gamma^T)^T$  é dada por

$$I(\theta) = \begin{bmatrix} \sigma^{-2} \mathbf{X}^T \Psi^{-1} \mathbf{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{n}{2\sigma^4} & \frac{1}{2} \sigma^{-2} \text{vec}(\Psi^{-1})^T \mathbf{A} \\ \mathbf{0} & \frac{1}{2} \sigma^{-2} \mathbf{A}^T \text{vec}(\Psi^{-1}) & \frac{1}{2} \mathbf{A}^T (\Psi^{-1} \otimes \Psi^{-1}) \mathbf{A} \end{bmatrix},$$

em que  $\mathbf{A} = \mathbf{A}(\gamma) = \text{vec}(\partial \Psi(\gamma) / \partial \gamma^T)$ ,  $\otimes$  representa o produto de Kronecker e o operador  $\text{vec}(\cdot)$  transforma as colunas de uma matriz em vetor.

No modelo (10.25), deseja-se agora testar a hipótese geral

$$H_0 : g(\theta) = 0 \quad \text{versus} \quad H_1 : g(\theta) \neq 0,$$

em que  $g$  é um vetor  $r \times 1$ . Seja  $\mathbf{F}$  a matriz  $(p + q + 1) \times r$  dada por  $\mathbf{F} = \partial g(\theta)^T / \partial \theta$ . A estatística de Wald é definida por

$$W = g(\hat{\theta})^T (\hat{\mathbf{F}}^T \mathbf{I}(\hat{\theta})^{-1} \hat{\mathbf{F}})^{-1} g(\hat{\theta}),$$

em que  $\hat{\theta}$  é o EMV irrestrito de  $\theta$ ,  $\hat{\mathbf{F}}$  é a matriz  $\mathbf{F}$  avaliada em  $\theta = \hat{\theta}$  e  $\mathbf{I}(\hat{\theta})$  é a informação em  $\hat{\theta}$ . A distribuição nula assintótica de  $W$  é  $\chi_r^2$ .

Uma estatística alternativa a de Wald é a estatística escore de Rao que envolve o EMV restrito  $\tilde{\theta}$ . Seja  $\mathbf{U}(\theta)$  a função escore para  $\theta$ , i.e.,  $\mathbf{U}(\theta) = \partial \ell(\theta) / \partial \theta$ . A estatística escore para testar  $H_0$  é dada por

$$S_R = \mathbf{U}(\tilde{\theta})^T \mathbf{I}(\tilde{\theta})^{-1} \mathbf{U}(\tilde{\theta}),$$

que, também, tem distribuição nula assintótica igual a  $\chi_r^2$ .

O teste da razão de verossimilhanças equivale ao uso da estatística

$$w = 2\{\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\}.$$

As três estatísticas  $W$ ,  $S_R$  e  $w$  têm propriedades assintóticas, em geral, equivalentes. Em vários modelos de regressão do tipo (10.25), os EMV restritos são mais fáceis de serem computados, o que representa uma vantagem de  $S_R$  em relação a  $w$  e  $W$ .

Suponha agora que as restrições são lineares apenas em  $\boldsymbol{\beta}$ , ou seja,  $H_0 : \mathbf{R}\boldsymbol{\beta} = 0$  e que  $\sigma^2$  e  $\boldsymbol{\psi}$  são conhecidos. Neste caso, as três estatísticas de teste,  $W$ ,  $S_R$  e  $w$  são idênticas e reduzem-se a

$$W = S_R = w = \tilde{\boldsymbol{\beta}}^T \mathbf{R}^T [\mathbf{R}(\mathbf{X}^T \boldsymbol{\psi}^{-1} \mathbf{X})^{-1} \mathbf{R}^T]^{-1} \mathbf{R} \tilde{\boldsymbol{\beta}} / \sigma^2,$$

em que  $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\psi}^{-1} \mathbf{y}$  é o EMV de  $\boldsymbol{\beta}$  quando  $\boldsymbol{\psi}$  é conhecido.

## 10.11 Modelo de regressão rígida

O modelo de regressão rígida objetiva superar os problemas de multicolinearidade das variáveis explicativas adicionando-se uma pequena constante positiva  $k$  aos termos da matriz  $\mathbf{X}^T \mathbf{X}$ . Outra alternativa para superar a multicolinearidade é aplicar transformações do tipo Box e Cox às variáveis explicativas. O estimador de regressão rígida é obtido resolvendo-se  $(\mathbf{X}^T \mathbf{X} + k\mathbf{I})\hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{y}$ , que produz  $\boldsymbol{\beta}^* = (\mathbf{X}^T \mathbf{X} + k\mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$ . Sejam  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$  os autovalores ordenados de  $\mathbf{X}^T \mathbf{X}$  e  $v_1, \dots, v_p$  seus autovetores correspondentes. Pode-se demonstrar que

$$(\mathbf{X}^T \mathbf{X} + k\mathbf{I})^{-1} = \sum_{i=1}^p (\lambda_i + k)^{-1} v_i v_i^T,$$

revelando que se  $\mathbf{X}^T \mathbf{X}$  é quase singular com  $\lambda_p$  pequeno, então, o menor autovalor de  $\mathbf{X}^T \mathbf{X} + k\mathbf{I}$  será  $\lambda_p + k$  e esta última matriz não será tão próxima da singularidade.

Sejam  $\mathbf{V}$  e  $\boldsymbol{\Lambda}$  as matrizes dos autovetores e autovalores de  $\mathbf{X}^T \mathbf{X}$ , ou seja,  $\mathbf{V} = (v_1, \dots, v_p)$  e  $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_p\}$ . O erro médio quadrático (EMQ) de  $\boldsymbol{\beta}^*$  é dado por

$$\text{EMQ}(\boldsymbol{\beta}^*) = \text{tr}(\mathbf{V}(\boldsymbol{\beta}^*)) + \{\mathbf{E}(\boldsymbol{\beta}^*) - \boldsymbol{\beta}\}^T \{\mathbf{E}(\boldsymbol{\beta}^*) - \boldsymbol{\beta}\},$$

em que  $\mathbf{V}(\boldsymbol{\beta}^*) = \sigma^2 \mathbf{W} \mathbf{X}^T \mathbf{X} \mathbf{W}$  e  $\mathbf{W} = \mathbf{V}(\boldsymbol{\Lambda} + k\mathbf{I})^{-1} \mathbf{V}^T$ . Tem-se, ainda,  $\mathbf{V}(\boldsymbol{\beta}^*) = \sigma^2 \mathbf{V} \boldsymbol{\Lambda}_* \mathbf{V}^T$ , em que  $\boldsymbol{\Lambda}_* = \text{diag}\{\lambda_i(\lambda_i + k)^{-2}\}$  e, então,  $\text{tr}(\mathbf{V}(\boldsymbol{\beta}^*)) = \sum \lambda_i(\lambda_i + k)^{-2}$ . Mas  $\boldsymbol{\beta}^* = \mathbf{W} \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}$ , em que  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  é o estimador de MQ de  $\boldsymbol{\beta}$ .

Assim,

$$E[\{E(\beta^*) - \beta\}^T \{E(\beta^*) - \beta\}] = \beta^T \mathbf{V} \wedge_+ \mathbf{V}^T \beta,$$

em que  $\wedge_+ = \text{diag}\{k^2(\lambda_i + k)^{-2}\}$ . Finalmente,

$$EMQ(\beta^*) = \sum (\lambda_i^2 + \gamma_i k^2)(\lambda_i + k)^{-2},$$

em que  $\gamma = (\gamma_1, \dots, \gamma_p)^T = \beta^T V$ .

Temos que a variância de  $\beta^*$  é uma função decrescente de  $k$  enquanto o seu viés é uma função crescente de  $k$ . Pode-se demonstrar que existe um  $k$  tal que  $EMQ(\beta^*) \leq EMQ(\hat{\beta})$ . Essa é a principal justificativa do uso da regressão rígida. Pode-se mostrar, ainda, que  $\beta^{*T} \beta^* < \hat{\beta}^T \hat{\beta}$ ,  $\forall k > 0$  e que  $\beta^{*T} \beta^* \rightarrow 0$  quando  $k$  cresce. Assim, o estimador de regressão rígida tende a origem quando  $k$  cresce. Temos ainda que

$$\beta^* = \sum_{i=1}^p \frac{1}{\lambda_i + k} d_i \mathbf{v}_i,$$

em que  $d_i = \mathbf{v}_i^T \mathbf{X}^T \mathbf{y}$ . Assim, determinando-se os autovalores e autovetores de  $\mathbf{X}^T \mathbf{X}$ , os estimadores de regressão rígida serão obtidos para qualquer valor de  $k$ . Define-se o traço rígido como um gráfico de  $\beta^*$  versus  $k$  para valores crescentes de  $k$ . Quando  $k = 0$ , tem-se o estimador de MQ de  $\beta$ . Com base no traço rígido pode-se escolher como valor de  $k$  o ponto em que as estimativas em  $\beta^*$  estão estabilizadas.

## 10.12 Modelos heterocedásticos

A heterocedasticidade é muito importante na modelagem de dados reais, pois a constância de variância (homocedasticidade) pode ser uma suposição forte em determinadas situações. Para o modelo de regressão geral (10.25), a heterocedasticidade estará presente se os elementos da diagonal de  $\Psi$  não são todos idênticos. Se, adicionalmente,  $\varepsilon$  está livre da autocorrelação,  $\Psi$  pode ser escrito como uma matriz diagonal cujo  $i$ -ésimo elemento é  $\sigma_i^2$ . A heterocedasticidade pode surgir das seguintes formas: (i) uso de dados sobre médias; (ii) variâncias que dependem das médias; (iii) variâncias que dependem de variáveis explicativas; (iv) diferentes observadores, locais de obtenção dos dados, etc; (v) pontos aberrantes. Se a heterocedasticidade está presente, precisamos investigar a sua forma e sua modelagem. Outra alternativa é tentar uma transformação do tipo Box-Cox com o objetivo de obter uma resposta modificada que se ajuste ao modelo clássico de regressão.

Um teste bastante usado para detectar heterocedasticidade é baseado na estatística de Anscombe

$$A = \frac{\sum_i r_i^2 (\hat{\mu}_i - \bar{y})}{s^2 \sum_{i,j} (\delta_{ij} - h_{ij})^2 (y_i - \bar{y})(y_j - \bar{y})}, \quad (10.31)$$

em que  $\delta_{ij} = 1$  se  $i = j$  e  $\delta_{ij} = 0$  se  $i \neq j$ ,  $h_{ij}$  são os elementos da matriz de projeção  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ ,  $\hat{\mu} = \mathbf{H}\mathbf{y}$ ,  $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ ,  $\bar{y} = (n-p)^{-1} \sum_i (1 - h_{ii})\hat{\mu}_i$  e  $s^2 = (n-p)^{-1} \sum_i r_i^2$ . Se (10.31) diferir significativamente de zero, pode-se supor a heterocedasticidade dos  $y'_i$ s.

Antes de considerar formas específicas de heterocedasticidade suponha que  $\Psi = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ . O estimador de mínimos quadrados generalizado (EMQG)  $\hat{\beta}$  é obtido de  $\hat{\beta} = (\mathbf{X}^T \Psi^{-1} \mathbf{X})^{-1} \mathbf{X}^T \Psi^{-1} \mathbf{y}$ . Quando  $\sigma^2$  depende de parâmetros desconhecidos, o EMQG de  $\beta$  pode ser obtido da equação acima substituindo-se  $\sigma_i^2$  por uma estimativa consistente  $\hat{\sigma}_i^2$  produzindo  $\hat{\beta} = (\mathbf{X}^T \hat{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Psi}^{-1} \mathbf{y}$ .

De agora em diante, denota-se por  $\dot{\mathbf{A}}$  a matriz contendo os quadrados dos elementos da matriz  $\mathbf{A}$ . Uma forma simples de estimar o vetor  $\dot{\sigma} = (\sigma_1^2, \dots, \sigma_n^2)^T$  contendo as variâncias desconhecidas é

$$\hat{\sigma} = \dot{\mathbf{M}}^{-1} \dot{\mathbf{r}}, \quad (10.32)$$

em que  $\dot{\mathbf{r}}$  é o vetor dos quadrados dos resíduos  $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$  e  $\mathbf{M} = \mathbf{I} - \mathbf{H}$  é uma matriz idempotente de posto  $n-p$ . Assim, (10.32) revela que  $\hat{\sigma}$  é obtido como uma transformação linear de  $\dot{\mathbf{r}}$ .

É fácil verificar que o EMQ  $\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  satisfaz  $E(\hat{\beta}) = \beta$  e  $\text{Cov}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \Psi \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}$ .

As principais formas de modelar a heterocedasticidade são:

- (i)  $\sigma_i^2 = (\mathbf{z}_i^T \boldsymbol{\gamma})^2$ , ou seja, o desvio padrão de  $y_i$  é uma função linear de variáveis exógenas;
- (ii)  $\sigma_i^2 = \sigma^2 (\mathbf{x}_i^T \mathbf{b})^{2\delta}$ , ou seja, a variância é proporcional a uma potência (em geral, par) do valor esperado;
- (iii)  $\sigma_i^2 = \exp(\mathbf{z}_i^T \boldsymbol{\gamma})$ , ou seja, o logaritmo da variância é uma função linear de variáveis exógenas. Esta última suposição define o *modelo heterocedástico multiplicativo*.

Apresenta-se agora o processo de estimação dos  $\beta$ 's e dos parâmetros das funções de variância acima, supondo que os dados são não-correlacionados.

$$(i) \quad y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma_i^2 = (\mathbf{z}_i^T \boldsymbol{\gamma})^2.$$

Neste caso, o EMQG de  $\boldsymbol{\beta}$  é

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^n (\mathbf{z}_i^T \boldsymbol{\gamma})^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i=1}^n (\mathbf{z}_i^T \boldsymbol{\gamma})^{-2} \mathbf{x}_i y_i. \quad (10.33)$$

Existem três estimadores possíveis para  $\boldsymbol{\gamma}$ : o estimador de MQ  $\hat{\boldsymbol{\gamma}}$ , o EMQG  $\hat{\hat{\boldsymbol{\gamma}}}$  e o EMV  $\tilde{\boldsymbol{\gamma}}$ , e, então, correspondente a cada um desses estimadores, teremos o EMQG  $\hat{\hat{\boldsymbol{\beta}}}$  obtido de (10.33) substituindo-se  $\boldsymbol{\gamma}$  por  $\hat{\boldsymbol{\gamma}}$ ,  $\hat{\hat{\boldsymbol{\gamma}}}$  e  $\tilde{\boldsymbol{\gamma}}$ . As variáveis padronizadas  $\sigma_1^{-1}\varepsilon_1, \dots, \sigma_n^{-1}\varepsilon_n$  são iid com média zero e variância um. Tem-se  $E(\sigma_i^{-1}|\varepsilon_i|) = c$ , em que  $c$  independe de  $i$  e depende somente da distribuição de  $\varepsilon_i$ . Assim,  $E(|\varepsilon_i|) = c\sigma_i$  e, portanto,

$$|r_i| = c\mathbf{z}_i^T \boldsymbol{\gamma} + v_i,$$

em que  $r_i = y_i - \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  e  $v_i = |r_i| - E(|\varepsilon_i|)$  é o novo erro do modelo correspondente ao parâmetro  $\boldsymbol{\gamma}$ . Logo,

$$c\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T |\mathbf{r}|$$

com  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$  e  $|\mathbf{r}| = (|r_1|, \dots, |r_n|)^T$ . O inconveniente do estimador  $\hat{\boldsymbol{\gamma}}$  é que este não tem as “propriedades do EMQ” pois, em geral, os  $v_i$ 's são heterocedásticos e autocorrelacionados e não têm média zero. Note-se que  $\hat{\hat{\boldsymbol{\beta}}}$  independe de  $c$ . O EMQG  $\hat{\hat{\boldsymbol{\gamma}}}$  é obtido do EMQ  $\hat{\boldsymbol{\gamma}}$  a partir da equação

$$c\hat{\boldsymbol{\gamma}} = \left( \sum_{i=1}^n (\mathbf{z}_i^T \hat{\boldsymbol{\gamma}})^{-1} \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \sum_{i=1}^n (\mathbf{z}_i^T \hat{\boldsymbol{\gamma}})^{-2} \mathbf{z}_i |r_i|.$$

O método de MV fornece a 3ª alternativa para estimar  $\boldsymbol{\gamma}$ . Se os  $\varepsilon_i$ 's são normais, o logaritmo da função de verossimilhança para  $\boldsymbol{\beta}$  e  $\boldsymbol{\gamma}$  é

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = - \sum_i \log \mathbf{z}_i^T \boldsymbol{\gamma} - \frac{1}{2} \sum_{i=1}^n \left( \frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\mathbf{z}_i^T \boldsymbol{\gamma}} \right)^2.$$

Obtendo-se a função escore para  $\boldsymbol{\beta}$  e  $\boldsymbol{\gamma}$  e igualando-a a zero, tem-se um sistema não-linear para calcular  $\tilde{\boldsymbol{\beta}}$  e  $\tilde{\boldsymbol{\gamma}}$  iterativamente. Suponha agora que  $\boldsymbol{\gamma} = (\gamma_1, \boldsymbol{\gamma}^{*T})^T$ , sendo  $\boldsymbol{\gamma}^* = (\gamma_2, \dots, \gamma_q)^T$ . Os  $\varepsilon_i$ 's são homocedásticos quando  $\boldsymbol{\gamma}^* = 0$  e um teste de homocedasticidade pode ser deduzido da razão de verossimilhanças  $w = 2\{\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) - \ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}_1)\}$ , em que os dois tils representam estimativas de MV restritas a  $\boldsymbol{\gamma}^* = 0$ , ou seja,  $\tilde{\boldsymbol{\gamma}}_1 = n^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$  e  $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ . Sob a hipótese  $\boldsymbol{\gamma}^* = 0$ ,  $w$  tem distribuição assintótica igual a  $\chi_{q-1}^2$ . Testes baseados nas estatísticas de Wald e escore podem, também, ser construídos.



(ii)  $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ ,  $E(\varepsilon_i) = 0$ ,  $\text{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2(\mathbf{x}_i^T \boldsymbol{\beta})^2$  (considerando o caso  $\delta = 1$ ).

A matriz de covariância de  $\boldsymbol{\varepsilon}$  é, simplesmente,  $\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} = \sigma^2 \text{diag}\{(\mathbf{x}_i^T \boldsymbol{\beta})^2\}$ . O EMQG  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Psi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Psi}^{-1} \mathbf{y}$  é inviável, pois  $\boldsymbol{\Psi}$  depende de  $\boldsymbol{\beta}$ . Entretanto, pode-se usar o EMQ de  $\boldsymbol{\beta}$  para obter o estimador  $\hat{\boldsymbol{\Psi}}$  de  $\boldsymbol{\Psi}$  e, então, definir  $\hat{\hat{\boldsymbol{\beta}}}$ . Um estimador conveniente para a matriz de covariância assintótica de  $\hat{\hat{\boldsymbol{\beta}}}$  é  $\hat{\Sigma}_{\hat{\hat{\boldsymbol{\beta}}}} = \hat{\sigma}^2 (\mathbf{X}^T \hat{\boldsymbol{\Psi}}^{-1} \mathbf{X})^{-1}$ , sendo

$$\hat{\sigma}^2 = (n - p)^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Psi}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}).$$

Se  $\mathbf{Y}$  tem distribuição normal multivariada, pode-se usar o método de MV para estimar conjuntamente  $\boldsymbol{\beta}$  e  $\boldsymbol{\Psi}$ . A dependência de  $\boldsymbol{\Psi}$  sobre  $\boldsymbol{\beta}$  implica que tanto a função  $(\mathbf{y} - \mathbf{X} \boldsymbol{\beta})^T \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})$  quanto ao logaritmo de verossimilhança não são agora funções quadráticas de  $\boldsymbol{\beta}$ . Métodos iterativos são necessários para obter os EMV nesse caso.

(iii)  $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$ ,  $E(\varepsilon_i) = 0$ ,  $\text{Var}(\varepsilon_i) = \sigma_i^2 = \exp(\mathbf{z}_i^T \boldsymbol{\gamma})$ ,

sendo  $\mathbf{z}_i^T$  um vetor  $1 \times q$  contendo variáveis explicativas adicionais para estimar  $\boldsymbol{\gamma} \in \mathbb{R}^q$ .

O primeiro elemento de  $\mathbf{z}_i$  é comumente 1. O EMQG de  $\boldsymbol{\beta}$  é

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^n \exp(-\mathbf{z}_i^T \boldsymbol{\gamma}) \mathbf{x}_i \mathbf{x}_i^T \right\}^{-1} \sum_{i=1}^n \exp(-\mathbf{z}_i^T \boldsymbol{\gamma}) \mathbf{x}_i y_i. \quad (10.34)$$

A partir dos resíduos  $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$  de mínimos quadrados pode-se definir o modelo

$$\log r_i^2 = \mathbf{z}_i^T \boldsymbol{\gamma} + v_i,$$

sendo  $v_i = \log(\varepsilon_i^2 / \sigma_i^2)$ , e obter o EMQ de  $\boldsymbol{\gamma}$  como

$$\hat{\boldsymbol{\gamma}} = \left( \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^T \right)^{-1} \sum_{i=1}^n \mathbf{z}_i \log r_i^2. \quad (10.35)$$

O problema com o estimador (10.35) é que os  $v_i$  não têm média zero e são heterocedásticos e autocorrelacionados. Com o estimador (10.35) inserido em (10.34), obter-se-á o estimador  $\hat{\hat{\boldsymbol{\beta}}}$  de  $\boldsymbol{\beta}$ .

Pode-se demonstrar que a covariância assintótica de  $\hat{\boldsymbol{\gamma}}$  é, simplesmente,  $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}} = 4.9348(\mathbf{Z}^T \mathbf{Z})^{-1}$ . Se  $\boldsymbol{\gamma}^T = (\gamma_1, \gamma^*)^T$ , um teste de homocedasticidade ( $H_0 : \gamma^* = 0$ ) pode ser realizado através da estatística

$$\mathbf{g} = 0.2026 \hat{\boldsymbol{\gamma}}^{*T} (\mathbf{Z}^T \mathbf{Z})^{-1} \boldsymbol{\gamma}^*$$

que tem, aproximadamente, distribuição nula igual a  $\chi_{q-1}^2$ .

O método de MV pode, também, ser usado para estimar conjuntamente  $\beta$  e  $\gamma$  a partir da maximização de

$$\ell(\beta, \gamma) = -\frac{1}{2} \sum_{i=1}^n \mathbf{z}_i^T \gamma - \frac{1}{2} \sum_{i=1}^n \exp(-\mathbf{z}_i^T \gamma) (y_i - \mathbf{x}_i^T \beta)^2.$$

O método escore de Fisher é baseado na informação conjunta dada por

$$\mathbf{K} = \begin{bmatrix} \mathbf{X}^T \Psi^{-1} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} \mathbf{Z}^T \mathbf{Z} \end{bmatrix}.$$

A ortogonalidade entre  $\beta$  e  $\gamma$  facilita o cálculo da estrutura de covariância assintótica dos EMV de  $\beta$  e  $\gamma$  bastando inverter  $\mathbf{K}$ .

### 10.13 Modelos autocorrelacionados

Considere o modelo  $\mathbf{y} = \mathbf{X}\beta + \varepsilon$  em que  $E(\varepsilon) = \mathbf{0}$  e  $\text{Cov}(\varepsilon) = \Psi = \sigma^2 \psi$  com  $\psi$  não-diagonal, isto é, as observações são correlacionadas. Várias estruturas de correlação para os  $\varepsilon$ 's são possíveis como os processos AR( $p$ ), MA( $q$ ) e ARMA( $p, q$ ). Nesta seção abordaremos apenas o caso mais simples, ou seja, o processo AR(1). O modelo de regressão com erros AR(1) pode ser escrito como

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad \varepsilon_i = \rho \varepsilon_{i-1} + v_i, \quad (10.36)$$

em que  $E(v_i) = 0$ ,  $\text{Var}(v_i) = \sigma_v^2$  e  $E(v_i v_j) = 0$  para  $i \neq j$  e  $|\rho| < 1$ . A matriz de covariância de  $\varepsilon$  é  $\text{Cov}(\varepsilon) = \sigma_v^2 \psi$  dada por

$$\Psi = \sigma_v^2 \psi = \frac{\sigma_v^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}. \quad (10.37)$$

A inversa de  $\Psi$  é

$$\Psi^{-1} = \sigma_v^{-2} \psi^{-1} = \sigma_v^{-2} \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

Se  $\rho$  é conhecido, o EMQG  $\hat{\beta} = (\mathbf{X}^T \psi^{-1} \mathbf{X})^{-1} \mathbf{X}^T \psi^{-1} \mathbf{y}$  é facilmente obtido usando  $\hat{\beta} = (\mathbf{X}^{*T} \mathbf{X}^*)^{-1} \mathbf{X}^{*T} \mathbf{y}^*$ , que é o EMQ aplicado ao modelo transformado  $\mathbf{y}^* = \mathbf{X}^* \beta + \varepsilon^*$ ,

em que  $\mathbf{y}^* = \mathbf{P}\mathbf{y}$ ,  $\mathbf{X}^* = \mathbf{P}\mathbf{X}$ ,  $\boldsymbol{\varepsilon}^* = \mathbf{P}\boldsymbol{\varepsilon}$  e

$$\mathbf{P} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}$$

é definida de  $\mathbf{P}^T\mathbf{P} = \boldsymbol{\psi}^{-1}$ .

Quando  $\rho$  é desconhecido, deve-se estimá-lo por  $\hat{\rho}$  para obter o estimador  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\boldsymbol{\psi}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\boldsymbol{\psi}}^{-1}\mathbf{y}$ , em que  $\hat{\boldsymbol{\psi}}$  é a matriz (10.37) avaliada em  $\hat{\rho}$ . Algumas formas para estimar  $\rho$  estão dadas a seguir:

- (a) coeficiente de correlação amostral

$$\hat{\rho}_1 = \sum_{i=2}^n r_i r_{i-1} / \sum_{i=1}^n r_i^2,$$

em que  $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$  são os resíduos de mínimos quadrados;

- (b) estatística de Durbin-Watson

$$\hat{\rho}_2 = 1 - 0.5 \sum_{i=2}^n (r_i - r_{i-1})^2 / \sum_{i=1}^n r_i^2;$$

- (c) estatística de Theil-Nagar

$$\hat{\rho}_3 = \frac{n^2 \hat{\rho}_2 + p^2}{n^2 - p^2}.$$



# Capítulo 11

## Modelos não-lineares da família exponencial

### 11.1 Introdução



# Capítulo 12

## Modelos simétricos não-lineares

### 12.1 Introdução





## APÊNDICE

## A.1 Programa R para os dados do Exemplo 6.1 - Volume de árvore

```
# Libraries needed #
library(MASS)
library(car)

# Minitab Cherry Tree Data
# =====

# Volume of usable wood in 31 black cherry trees from
# Minitab Student Handbook (1985), Ryan, Joiner and Ryan.
#   D = diameter at 4.5 ft from ground (inches)
#   H = height (feet)
#   V = volume (cubic feet)

##trees<-read.table("Tree.dat", header=TRUE)
##require(trees)
data(trees, package='datasets')
data() # lists all datasets
attach(trees)

D<-trees[,1]
H<-trees[,2]
V<-trees[,3]

# first examine the data
par(mfrow=c(2,3))
hist(D, main="Girth")
hist(H, main="Height")
hist(V, main="Volume")
boxplot(D, main="Girth")
boxplot(H, main="Height")
boxplot(V, main="Volume")
```

```

#Scatterplot
pairs(trees)
plot(trees)
scatterplot.matrix(trees) # uses library(car)

## Fitted models ##
mod1<-lm(V~1)
mod2<-lm(V~D)
summary(mod2)
mod3<-lm(V~H)
summary(mod3)
mod4<-lm(V~D+H)
summary(mod4)
anova(mod1, mod4)
anova(mod1, mod2, mod4)
anova(mod1, mod3, mod4)

#####
# A set of four plots for diagnostics #
#####
n<-dim(trees)[1] # number of data
par(mfrow=c(2,2))
# Observed against fitted values #
plot(fitted(mod4),V)
identify(fitted(mod4),V)    #click on the point, use ESC  in R to esc

# abs(DFFitS) vs index #
plot(abs(dffits(mod4)))
abline(3*sqrt(mod4$rank/mod4$df.residual),0,lty=2)
identify(1:n,abs(dffits(mod4)))    #click on the point, use ESC  in R to esc

```

```

# QQplot with a simulated envelope #
qq.plot(mod4,simulate=TRUE, reps=19)

# Likelihood profile plot for Box-Cox f#
boxcox(mod4)      # needs library(MASS)
#####

## Log transformed data ##
#Scatterplots
logD<-log(D)
logH<-log(H)
logV<-log(V)
ltrees<-cbind(logD,logH,logV)
pairs(ltrees)

## Fitted models ##
mod1<-lm(logV~1)
mod2<-lm(logV~logD)
summary(mod2)
mod3<-lm(logV~logH)
summary(mod3)
mod4<-lm(logV~logD+logH)
summary(mod4)
anova(mod1, mod4)
anova(mod1, mod2, mod4)
anova(mod1, mod3, mod4)

#####
# A set of four plots for diagnostics #
#####
n<-dim(trees)[1] # number of data

```

```
par(mfrow=c(2,2))
# Observed against fitted values #
plot(fitted(mod4),V)
identify(fitted(mod4),V)

# abs(DFFitS) vs index #
plot(abs(dffits(mod4)))
abline(3*sqrt(mod4$rank/mod4$df.residual),0,lty=2)
identify(1:n,abs(dffits(mod4)))

# QQplot with simulated envelope #
qq.plot(mod4,simulate=TRUE, reps=19)

# Likelihood profile plot for Box-Cox #
boxcox(mod4)
#####
```

## A.2 Programa R para os dados do Exemplo 6.2 - Gordura no leite

```

# Average daily fat yields (kg/day) from milk
# from a single cow for each of 35 weeks
# McCulloch (2001)
# Ruppert, Cressie, Carroll (1989) - Biometrics, 45: 637-656

fatyield.dat<-scan(what=list(yield=0))
0.31 0.39 0.50 0.58 0.59 0.64
0.68 0.66 0.67 0.70 0.72 0.68
0.65 0.64 0.57 0.48 0.46 0.45
0.31 0.33 0.36 0.30 0.26 0.34
0.29 0.31 0.29 0.20 0.15 0.18
0.11 0.07 0.06 0.01 0.01

fatyield.dat$week=1:35
attach(fatyield.dat)
lweek<-log(week)
plot(weeks,yield, pch="*", xlab="Weeks", ylab="Fat yield (kg/day)",
main="Figura 1. Observed fat yield (kg/day) for each week")

## Normal model for log(fat)
lyield<-log(yield)
mod1<-lm(lyield~week+lweek)
summary(mod1)
fit1<-fitted(mod1)

## Normal model with log link
mod2<-glm(yield~week+lweek, (family=gaussian(link="log")))
fit2<-fitted(mod2)

# Plotting

```

```
plot(c(0,35), c(0,0.9), type="n", xlab="Weeks", ylab="Fat yield (kg/day)")
points(week,yield, pch="*")
w<-seq(1,35,0.1)
lines(w, predict(mod2,data.frame(week=w, lweek=log(w)),type="response"),
col="green", lty=1)
lines(w, exp(predict(mod1,data.frame(week=w, lweek=log(w)),type="response")),
col="red", lty=2)
legend(20,0.9,c("Observed","Log transformation", "Log link"), lty=c(-1,2,1),
pch=c("*"," "," "), col=c("black","red","green"),cex=.6)
title(sub="Figura 1. Fat yield (kg/day) for each week")
```

## A.3 Programa R para os dados do Exemplo 6.3 - Importação.

```
importa.dat <- scan()
5482      1.629      82.17
4046      1.423     109.40
5749      1.517      88.80
5495      1.356     111.36
6043      1.331      87.94
5173      1.244     105.50
5679      1.181      85.28
4576      1.046      97.60
5605      1.315      82.06
4265      1.091      96.39
5565      1.217      86.49
5474      1.091     106.01
5610      1.177      82.62
6345      1.300     100.01
5309      1.135      78.30
4330      1.380      91.70
4804      1.434      78.34
5034      1.354     104.02
4872      1.306      87.11
5614      1.314     108.26
5071      1.209      85.77
6015      1.452     101.05
4646      1.156      80.91
4630      1.499      97.02
3824      1.740      75.88
4725      1.626     101.71
3651      2.004      83.65
5221      1.467     103.80
3907      1.957      82.80
```

5976	1.441	101.30
4044	1.959	80.10
5230	1.421	99.90
3155	1.971	79.10
6007	1.388	106.90
3406	2.015	87.59
7328	1.340	108.92
3730	2.024	87.19
6914	1.305	106.01
3623	2.027	85.94
6049	1.283	104.01
3094	2.036	84.55
7087	1.279	109.66
3016	2.219	92.47
8023	1.075	115.30
3132	2.201	95.23
11814	0.957	116.45
3925	2.131	94.44
12065	0.942	113.92
3352	2.013	90.69
13651	0.955	116.09
2760	2.023	99.48
11917	0.951	115.67
3661	1.991	102.87
12030	0.970	114.93
4270	1.924	101.15
10738	0.980	111.63
3565	1.832	97.65
12478	0.995	118.06
3610	1.792	106.21
14235	1.012	122.90



3987	1.914	103.45
15837	1.030	120.69
3888	1.789	101.10
13150	1.049	116.90
3516	1.692	97.72
15405	1.067	123.85
3349	1.657	105.78
16930	1.086	126.37
3776	1.643	105.84
15873	1.106	122.55
3963	1.607	98.87
13415	1.126	118.11
3548	1.557	95.01
14591	1.147	125.74

```
importa.dat <-data.frame(matrix(importa.dat,74,3, byrow=T))
```

```
IM <- importa.dat[,1]
```

```
TCI <- importa.dat[,2]
```

```
RN <- importa.dat[,3]
```

```
invIM <- 1/IM
```

```
logIM <- log(IM)
```

```
importa.dat <-data.frame(IM,logIM,invIM,TCI,RN)
```

```
pairs(importa.dat)
```

```
# IM Normal como função de ligação identidade
```

```
importa.Nident<-glm(formula = IM ~ TCI + RN, family = gaussian(identity),data = importa.dat)
anova(importa.Nident, test="F")
```

```
summary(importa.Nident)
```

```
plot(fitted(importa.Nident),IM, xlab="valor ajustado", ylab="valor observado",
main="Normal(identidade)")
```

```
plot(fitted(importa.Nident), rstudent(importa.Nident), xlab="valor ajustado", ylab="residuals")
```

```

main="Normal(identidade)")
plot(TCI, rstudent(importa.Nident), xlab="TCI", ylab="resíduos",
main="Normal(identidade)")
plot(RN, rstudent(importa.Nident), xlab="RN", ylab="resíduos",
main="Normal(identidade)")
qqnorm(resid(importa.Nident),ylab="Residuo",main="Normal(identidade)")
qqline(resid(importa.Nident))

# log(IM) Normal como função de ligação identidade
importa.logNident<-glm(formula = logIM ~ TCI + RN, family = gaussian(identity),
data = importa.dat)
anova(importa.logNident, test="F")
summary(importa.logNident)
plot(fitted(importa.logNident),logIM, xlab="valor ajustado", ylab="valor observado",
main="logNormal(identidade)")
plot(fitted(importa.logNident), rstudent(importa.logNident), xlab="valor ajustado",
main="logNormal(identidade)")
plot(TCI, rstudent(importa.logNident), xlab="TCI", ylab="resíduos",
main="logNormal(identidade)")
plot(RN, rstudent(importa.logNident), xlab="RN", ylab="resíduos",
main="logNormal(identidade)")
qqnorm(resid(importa.logNident),ylab="Residuo",main="logNormal(identidade)")
qqline(resid(importa.logNident))

# IM Normal como função de ligação logarítmica
importa.Nlog<-glm(formula = IM ~ TCI + RN, family = gaussian(log),data = importa.dat)
anova(importa.Nlog, test="F")
summary(importa.Nlog)
plot(fitted(importa.Nlog),IM, xlab="valor ajustado", ylab="valor observado",
main="Normal(log)")
plot(fitted(importa.Nlog), rstudent(importa.Nident), xlab="valor ajustado", ylab="re

```

```
main="Normal(log)")
plot(TCI, rstudent(importa.Nlog), xlab="TCI", ylab="resíduos",
main="Normal(log)")
plot(RN, rstudent(importa.Nlog), xlab="RN", ylab="resíduos",
main="Normal(log)")
qqnorm(resid(importa.Nlog),ylab="Residuo",main="Normal(log)")
qqline(resid(importa.Nlog))

# IM Gama como função de ligação inversa
importa.Ginv<-glm(formula = IM ~ TCI + RN, family = Gamma(inverse),data = importa.da
anova(importa.Ginv, test="F")
summary(importa.Ginv)
plot(fitted(importa.Ginv),IM, xlab="valor ajustado", ylab="valor observado",
main="Gamma(inversa)")
plot(fitted(importa.Ginv), importa.Ginv$deviance.resid, xlab="valor ajustado",
ylab="resíduos", main="Gamma(inversa)")
plot(TCI, importa.Ginv$deviance.resid, xlab="TCI", ylab="resíduos",
main="Gamma(inversa)")
plot(RN, importa.Ginv$deviance.resid, xlab="RN", ylab="resíduos",
main="Gamma(inversa)")
qqnorm(resid(importa.Ginv),ylab="Residuo",main="Gamma(inversa)")
qqline(resid(importa.Ginv))

# IM Gama como função de ligação identidade
importa.Gident<-glm(formula = IM ~ TCI + RN, family = Gamma(identity),data = importa
anova(importa.Gident, test="F")
summary(importa.Gident)
plot(fitted(importa.Gident),IM, xlab="valor ajustado", ylab="valor observado",
main="Gamma(identity)")
plot(fitted(importa.Gident), importa.Ginv$deviance.resid, xlab="valor ajustado", yla
main="Gamma(identity)")
```

```
plot(TCI, importa.Gident$deviance.resid, xlab="TCI", ylab="resíduos",
main="Gamma(identity)")
plot(RN, importa.Gident$deviance.resid, xlab="RN", ylab="resíduos",
main="Gamma(identity)")
qqnorm(resid(importa.Gident),ylab="Residuo",main="Gamma(identity)")
qqline(resid(importa.Gident))

# IM Gama como função de ligação logarítmica
importa.Glog<-glm(formula = IM ~ TCI + RN, family = Gamma(log),data = importa.dat)
anova(importa.Glog, test="F")
summary(importa.Glog)
plot(fitted(importa.Glog),IM, xlab="valor ajustado", ylab="valor observado",
main="Gamma(log)")
plot(fitted(importa.Glog), importa.Ginv$deviance.resid, xlab="valor ajustado", ylab=
main="Gamma(log)")
plot(TCI, importa.Glog$deviance.resid, xlab="TCI", ylab="resíduos",
main="Gamma(log)")
plot(RN, importa.Glog$deviance.resid, xlab="RN", ylab="resíduos",
main="Gamma(log)")
qqnorm(resid(importa.Glog),ylab="Residuo",main="Gamma(log)")
qqline(resid(importa.Glog))
```

A.?? Programa R para os dados de malária.

```
## Leitura dos dados
```

```
Modelo Nacional
```

```
## Lendo os dados
```

```
dados= read.table(file="C:/Documents and Settings/Administrador/Meus  
documentos/Artigo Malária/dados.txt",header=T)  
attach(dados)
```

```
## Transformação Box e Cox
```

```
library(MASS)
```

```
mal=malaria+1e-20
```

```
#Gráfico da log-verossimilhança maximizada
```

```
boxcox(mal ~ -1 + medicos + gini + trans_fed + tx_analf, lambda = seq(-1, 1, 1/100),  
plotit = TRUE, ylab = "log-verossimilhança")
```

```
# Valor de lambda que maximiza a log-verossimilhança
```

```
c <- boxcox(mal ~ -1 + medicos + gini + trans_fed + tx_analf, lambda = seq(-1, 1,  
1/100), plotit = FALSE, ylab = "log-verossimilhança")
```

```
# Maior valor encontrado na log-verossimilhança maximizada
```

```
M <- cbind(c$x,c$y)
```

```
M <- as.matrix(M)
```

```
b <- M[which.max(M[,2]),]
```

```
b
```

```
# Intervalo de confiança
```

```
L <- b[2]
a <- L - (0.5*qchisq(0.95,1))
a
M[M[,2] > a]

## Poisson

poisson<-glm(malaria ~ renda + pop_total + saude_san + gini + idh_ed + tx_analf +
medicos + hab_urb
+ trans_fed + trans_est, family=poisson(link = "log
summary(poisson)

# Resíduos - Componentes do desvio
rd<-residuals(poisson, type="deviance")
plot(rd, xlab="Índice", ylab="Componentes do desvio")
abline(-2,0,lty=2, col="red")
abline(2,0,lty=2, col="red")

## Binomial Negativa
library(MASS)

# Modelo - Nacional
malaria.NB<-glm.nb(malaria ~ gini + tx_analf + trans_fed, link = log)
summary(malaria.NB)

qchisq(0.95,malaria.NB$df.residual)

# Valor ajustado
mi <- fitted(malaria.NB)
```

```
# Resíduos - Componentes do desvio
rd<-residuals(malaria.NB,type="deviance")

# Gráfico do resíduo componente do desvio versus índice das observações
plot(rd, xlab="Índice", ylab="Componentes do desvio", ylim=c(-2.5,2.5))
abline(-2,0,lty=2, col="red")
abline(2,0,lty=2, col="red")
identify(rd, col="blue", label=municipio, cex=0.7)

# Pontos Influentes - Distância de Cook
di <- cooks.distance(malaria.NB)
plot(di, xlab="Índice", ylab="Distâncias de Cook", main="")
identify(di, col="blue", label=municipio, cex=0.7)

# Pontos de Alavanca
p=length(malaria.NB$coef)
n=length(malaria)

h <- lm.influence(malaria.NB)$hat
plot(h, xlab="Índice", ylab="Diagonal de H", main="")
abline((2*p)/n,0,lty=2, col="red")
identify(h, col="blue", label=municipio, cex=0.7)

# Gráfico do componente do desvio padronizado versus o preditor linear
eta <- predict(malaria.NB)
rd<-residuals(malaria.NB,type="deviance")
h <- lm.influence(malaria.NB)$hat
cdp = rd/(1-h)^0.5
plot(eta, cdp, xlab=expression(paste(hat(eta))), ylab="Componentes do desvio
padronizado")
abline(-2,0,lty=2, col="red")
```

```
abline(2,0,lty=2, col="red")
identify(eta, cdp, col="blue", label=municipio, cex=0.7)

### Modelo - Norte
## Lendo os dados
dados= read.table(file="C:/Documents and Settings/Administrador/Meus
documentos/Artigo Malária/norte.txt",header=T)
attach(dados)

malaria.NB<-glm.nb(malaria ~ -1 + renda + pop_total + saude_san + gini + trans_est,
link = log)
summary(malaria.NB)

### Modelo - Sudeste
## Lendo os dados
dados= read.table(file="C:/Documents and Settings/Administrador/Meus
documentos/Artigo Malária/sudeste.txt",header=T)
attach(dados)

malaria.NB<-glm.nb(malaria ~ medicos + hab_urb, link = log)
summary(malaria.NB)

### Modelo - Sul
## Lendo os dados
dados= read.table(file="C:/Documents and Settings/Administrador/Meus
documentos/Artigo Malária/sul.txt",header=T)
attach(dados)
```



```
malaria.NB<-glm.nb(malaria ~ -1 + gini + hab_urb + trans_fed, link = log)
summary(malaria.NB)
```

A.?? Programa R para os dados do Exemplo 4.5 - Rotenone.

```
## Leitura dos dados
dose <- c(0,2.6,3.8,5.1,7.7,10.2)
y <- c(0,6,16,24,42,44)
m <-c(49,50,48,46,49,50)
Rotenone.dat <- data.frame(dose, y, m)
attach(Rotenone.dat)

## Gráfico de dispersão
plot(dose,y/m, xlab="Dose", ylab="Proporções observadas", pch="*")

## Análise do desvio
resp<-cbind(y,m-y)
Rotenon1<-glm(resp~1, family=binomial)
Rotenon2<-glm(resp~dose, family=binomial)
summary(Rotenon2)
anova(Rotenon1, Rotenon2, test="Chisq")

## Gráfico
plot(c(0,10.2), c(0,1), type="n", xlab="Dose", ylab="Proporção")
points(dose,y/m,pch="*")
x<-seq(0,10.2,0.2)
lp<-predict(Rotenon2,data.frame(dose=x))
pe<-exp(lp)/(1+exp(lp))
lines(x,pe,lty=1)
```

## A.4 Programa R para os dados do Exemplo 7.2 - Cypermethrin.

```

y <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex<-factor(rep(c("M","F"), c(6,6))) ldose<-rep(0:5,2)
dose<-2**ldose dose<-factor(dose) Cyper.dat <- data.frame(sex,
dose, ldose, y) attach(Cyper.dat)
plot(ldose,y/20, pch=c(rep("*",6),rep("+",6)),col=c(rep("green",6),
rep("red",6)), xlab="log(dose)", ylab="Proportion killed")
resp<-cbind(y,20-y)

mod1<-glm(resp~1, family=binomial) mod2<-glm(resp~dose,
family=binomial) mod3<-glm(resp~sex, family=binomial)
mod4<-glm(resp~dose+sex, family=binomial) anova(mod1, mod2, mod4,
test="Chisq") anova(mod1, mod3, mod4, test="Chisq")

mod5<-glm(resp~ldose, family=binomial) mod6<-glm(resp~sex+ldose-1,
family=binomial) mod7<-glm(resp~ldose/sex, family=binomial)
mod8<-glm(resp~ldose*sex, family=binomial) anova(mod1, mod5, mod6,
mod8, test="Chisq") anova(mod1, mod5, mod7, mod8, test="Chisq")
summary(mod6)

plot(c(1,32), c(0,1), type="n", xlab="log(dose)",
ylab="Proportions", log="x") points(2**ldose,y/20,
pch=c(rep("*",6),rep("+",6)), col=c(rep("green",6),rep("red",6)))
ld<-seq(0,5,0.1) lines(2**ld, predict(mod6,data.frame(ldose=ld,
sex=factor(rep("M",length(ld)),levels=levels(sex))), type="response"),
col="green")
lines(2**ld, predict(mod6,data.frame(ldose=ld,
sex=factor(rep("F",length(ld)),levels=levels(sex))), type="response"),
col="red")

```

## A.5 Programa R para os dados do Exemplo 7.3 - Tribolium.

```

# Collett(2002)- pag. 103
dose <- c(2.00,2.64,3.48,4.59,6.06,8.00)
dose <- c(rep(dose,3))
d <- as.factor(dose)
ldose<-log(dose)
y <- c(3,5,19,19,24,35,2,14,20,27,41,40,28,37,46,48,48,50)
m <- c(50,49,47,50,49,50,50,49,50,50,50,50,50,50,50,50,50)
insecticid<-as.factor(c(rep("DDT",6),rep("BHC",6),rep("DDT+BHC",6)))
Tribolium.dat <- data.frame(dose, y, m, insecticid)
attach(Tribolium.dat)
resp<-cbind(y,m-y)
pe=y/m
plot(dose,pe, pch=c(rep("*",6),rep("+",6), rep("-",6)),
col=c(rep("green",6), rep("red",6),rep("blue",6)),
xlab="Dose", ylab="Proporções de insetos mortos")
plot(ldose,pe, pch=c(rep("*",6),rep("+",6), rep("-",6)),
col=c(rep("green",6), rep("red",6),rep("blue",6)),
xlab="Log(Dose)", ylab="Proporções de insetos mortos")

mod1<-glm(resp~1, family=binomial)
1-pchisq(deviance(mod1), df.residual(mod1))
print(X2<-sum(residuals(mod1, 'pearson')^2))
1-pchisq(X2, df.residual(mod1))
mod2<-glm(resp~d, family=binomial)
1-pchisq(deviance(mod2), df.residual(mod2))
print(X2<-sum(residuals(mod2, 'pearson')^2))
1-pchisq(X2, df.residual(mod2))
mod3<-glm(resp~insecticid, family=binomial)
1-pchisq(deviance(mod3), df.residual(mod3))
print(X2<-sum(residuals(mod3, 'pearson')^2))

```

```
1-pchisq(X2, df.residual(mod3))
mod4<-glm(resp~insecticid+d, family=binomial)
1-pchisq(deviance(mod4), df.residual(mod4))
summary(mod4)
print(X2<-sum(residuals(mod4, 'pearson')^2))
1-pchisq(X2, df.residual(mod4))
anova(mod4, test="Chisq")
mod5<-glm(resp~insecticid+ldose-1, family=binomial)
1-pchisq(deviance(mod5), df.residual(mod5))
summary(mod5, test="Chisq")
print(X2<-sum(residuals(mod5, 'pearson')^2))
1-pchisq(X2, df.residual(mod5))
anova(mod1, mod2, mod4, test="Chisq")
anova(mod1, mod3, mod4, test="Chisq")
mod6<-glm(resp~ldose/insecticid, family=binomial)
1-pchisq(deviance(mod6), df.residual(mod6))
summary(mod6)
mod7<-glm(resp~ldose, family=binomial)
1-pchisq(deviance(mod7), df.residual(mod7))
print(X2<-sum(residuals(mod7, 'pearson')^2))
1-pchisq(X2, df.residual(mod7))
mod8<-glm(resp~insecticid*ldose-insecticid, family=binomial)
1-pchisq(deviance(mod8), df.residual(mod8))
print(X2<-sum(residuals(mod8, 'pearson')^2))
1-pchisq(X2, df.residual(mod8))
mod9<-glm(resp~insecticid+ldose, family=binomial)
1-pchisq(deviance(mod9), df.residual(mod9))
print(X2<-sum(residuals(mod9, 'pearson')^2))
1-pchisq(X2, df.residual(mod9))
summary(mod9)
anova(mod9, test="Chisq")
```

```

mod10<-glm(resp~insecticid*ldose, family=binomial)
1-pchisq(deviance(mod10), df.residual(mod10))
summary(mod10)
print(X2<-sum(residuals(mod10, 'pearson')^2))
1-pchisq(X2, df.residual(mod10))

plot(c(1.8,8), c(-3,3.5), type="n", xlab="dose", ylab="Logit(proporções)", log="x")
points(dose,log(pe/(1-pe)), pch=c(rep("*",6),rep("+",6), rep("-",6)),
col=c(rep("green",6), rep("red",6),rep("blue",6)))
ld<-seq(log(2),log(8),0.005)
lines(exp(ld), predict(mod5,data.frame(ldose=ld,
insecticid=factor(rep("DDT",length(ld)),levels=levels(insecticid))),
type="link"), col="green")
lines(exp(ld), predict(mod5,data.frame(ldose=ld,
insecticid=factor(rep("BHC",length(ld)),levels=levels(insecticid))),
type="link"), col="red")
lines(exp(ld), predict(mod5,data.frame(ldose=ld,
insecticid=factor(rep("DDT+BHC",length(ld)),levels=levels(insecticid))),
type="link"), col="blue")

plot(c(1.8,8), c(0,1), type="n", xlab="dose", ylab="proporções", log="x")
points(dose,pe, pch=c(rep("*",6),rep("+",6), rep("-",6)),
col=c(rep("green",6), rep("red",6),rep("blue",6)))
ld<-seq(log(2),log(8),0.005)
lines(exp(ld), predict(mod5,data.frame(ldose=ld,
insecticid=factor(rep("DDT",length(ld)),levels=levels(insecticid))),
type="response"), col="green")
lines(exp(ld), predict(mod5,data.frame(ldose=ld,
insecticid=factor(rep("BHC",length(ld)),levels=levels(insecticid))),
type="response"), col="red")
lines(exp(ld), predict(mod5,data.frame(ldose=ld,

```

```
insecticid=factor(rep("DDT+BHC",length(ld)),levels=levels(insecticid)),  
type="response", col="blue")
```

## A.6 Programa R para os dados do Exemplo 7.4 - Contagem de bactérias

```
# *** Bacterial concentrations of stored micro-organisms
# Glim4 Manual p531
ltime <- log((tim <- c(0, 1, 2, 6, 12))+ 0.1)
lcount <- log(count <- c(31, 26, 19, 15, 20))

bacteria.dat <- data.frame(tim, count, ltime, lcount)
attach(bacteria.dat)

par(mfrow=c(1,2))
plot(tim, count, xlab="Time in months", ylab="Counts")
plot(ltime,lcount, xlab="Log(time in months)", ylab="Log(counts)")
par(mfrow=c(1,1))

mod1<-glm(count ~ tim, family=poisson)
anova(mod1, test="Chisq")

mod2<-glm(count ~ ltime, family=poisson)
anova(mod2, test="Chisq")

plot(c(0,12), c(15,31), type="n", xlab="Time in months", ylab="Counts")
points(tim,count,pch="*")
x<-seq(0,12,0.1)
lp<-predict(mod2,data.frame(ltime=log(x+0.1)), type="response")
lines(x,lp,lty=1)
```



## A.7 Programa R para os dados do Exemplo 7.5 - Armadilhas adesivas

```
# 2 by 2 table - Traps
y <- c(246, 17, 458, 32)
armcor <- factor(c(1, 1, 2, 2))
sexo <- factor(c(1, 2, 1, 2))

count.dat <- data.frame(armcor, sexo, y)
attach(count.dat)

# calculate observed odds ratio
246*32/(17*458)

# now set up log linear model
mod1<-glm(y ~ armcor*sexo, family=poisson)
print(sum(residuals(mod1, 'pearson')^2))
anova(mod1, test="Chisq")
summary(mod1)

# note that this model reproduces the data
# also the fitted log-odds ratio is 0.01098 giving a
# fitted odds-ratio of
exp(mod1$coef[4])

# the interaction term is not significant, so we
# cannot reject the hypothesis that the odds-ratio is 1,
# i.e. trap's colour and sex are independent.

# refit simplest adequate model
# the main effects, or independence, model
mod2<-glm(y ~ armcor+sexo, family=poisson)
print(sum(residuals(mod1, 'pearson')^2))
```

```
anova(mod2, test="Chisq")  
1-pchisq(deviance(mod2), df.residual(mod2))  
summary(mod2)
```

A.?? Programa R para os dados do Exemplo ?? -

```
#Leitura dos dados
Drogas<-read.table("F://Backup/UFRPE/MLG/InteracaoDrogas.txt", header=T)
attach(Drogas)
Drogas
#Estimacao dos valores iniciais
rho0<-0.05
kappa0<-0.5
t<-log(x1+rho0*x2+kappa0*(rho0*x1*x2)^0.5)
mod0<-lm(y~t)
alpha0<-mod0$coef[[1]]
delta0<-mod0$coef[[2]]
#Modelo Proposto
Model.Drogas<-nls(y~alpha+delta*log(x1+rho*x2+kappa*(rho*x1*x2)^0.5),
data=Drogas, start=list(alpha=alpha0,delta=delta0,rho=rho0,kappa=kappa0) )
summary(Model.Drogas)
#Modelo Nulo (assumindo k=0)
Model.Drogas.Null<-nls(y~alpha+delta*log(x1+rho*x2), data=Drogas,
start=list(alpha=alpha0,delta=delta0,rho=rho0) )
summary(Model.Drogas.Null)
anova(Model.Drogas.Null,Model.Drogas)
#Analises graficas
alpha<-coef(Model.Drogas)[[1]]
delta<-coef(Model.Drogas)[[2]]
rho<-coef(Model.Drogas)[[3]]
kappa<-coef(Model.Drogas)[[4]]
y<-Drogas$y
x1<-Drogas$x1
x2<-Drogas$x2
d1<-1
d2<-log(x1+rho*x2+kappa(rho*x1*x2)^0.5)
```

```

d3<-delta*( x2+0.5*kappa*((x1*x2)^0.5)*(rho^(-0.5)) ) / (x1+rho*x2+kappa*(rho*x1*x2)
d4<- delta*( (rho*x1*x2)^0.5)/(x1+rho*x2+kappa*(rho*x1*x2)^0.5)
Xtilde<-cbind(d1,d2,d3,d4)
s<-summary(Model.Drogas)$sigma
H<-solve(t(Xtilde)%*%Xtilde)
H<-Xtilde%*%H%*%t(Xtilde)
r<-as.vector(resid(Model.Drogas))
t<-r /(s*sqrt(1-diag(H)) )
#Resíduos Studentizados
plot(t,xlab="Índice",ylab="Resíduos Studentizados")
identify(t)
D<-as.vector((t^2/(length(coef(Model.Drogas))))*(diag(H)/(1-diag(H))))
#Distancia de Cook
plot(D,xlab="Índice",ylab="Distância de Cook")
identify(D)
beta<-as.vector(coef(Model.Drogas))
beta0<-beta
for (j in 1:(n-1)) beta0<-rbind(beta0,beta)
beta.1step<-solve((t(Xtilde)%*%Xtilde))
beta.1step<- beta0-((Xtilde*r)%*%beta.1step)
beta.1step
DP_kappa<-sqrt(s^2*solve((t(Xtilde)%*%Xtilde))[4,4])
k.1step<-as.vector(kappa-beta.1step[,4])/DP_kappa
#Variação em Kappa
plot(k.1step,xlab="Índice",ylab="Kappa 1a ordem")
identify(k.1step)
#Gráfico da variável adicionada
alpha.null<-coef(Model.Drogas.Null)[[1]]
delta.null<-coef(Model.Drogas.Null)[[2]]
rho.null<-coef(Model.Drogas.Null)[[3]]
kappa.null<-0

```

```
d1.null<-1
d2.null<-log(x1+rho.null*x2)
d3.null<-delta.null*( x2 ) / (x1+rho.null*x2)
d4.null<- delta.null*((rho.null*x1*x2)^0.5)/(x1+rho.null*x2+kappa.null*(rho.null*x1*
Xtilde.null<-cbind(d1.null,d2.null,d3.null)
s.null<-summary(Model.Drogas.Null)$sigma
H.null<-solve(t(Xtilde.null)%*%Xtilde.null)
H.null<-Xtilde.null%*%H.null%*%t(Xtilde.null)
vec<-H.null%*%d4.null
plot(vec,res,xlab="",ylab="Resíduo")
identify(vec,res)
```

A.?? Programa R para os dados do Exemplo ?? -

A.?? Programa R para os dados do Exemplo 9.1 - Modelo de Gompertz para explicar o comprimento de um certo tipo de feijoeiro

```
## Lendo os dados Gompertz <-
read.table(file="C:/UFRPE/MLG/Análises/dados_gompertz.txt",header=T)
attach(Gompertz)
```

```
## Ajustando um modelo de regressão normal não-linear # Valor
inicial a=3.0 mod1 <- nls(y ~ a * exp(- exp(b - g * x)),
start=list(a=3.0, b=2.1, g=0.4)) summary(mod1)
```

Formula:  $y \sim a * \exp(-\exp(b - g * x))$

Parameters:

	Estimate	Std. Error	t value	Pr(> t )
a	22.5066	0.8373	26.882	4.32e-12 ***
b	2.1063			0.2351
g	8.959	1.16e-06 ***	0.3881	0.0459
	8.455	2.12e-06 ***		

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.024 on 12 degrees of freedom

Number of iterations to convergence: 10 Achieved convergence  
tolerance: 5.4e-06

```
# Valor inicial a=21.3 mod2 <- nls(y ~ a * exp(- exp(b - g * x)),
start=list(a=21.3, b=2.1, g=0.4)) summary(mod2)
```

Formula:  $y \sim a * \exp(-\exp(b - g * x))$

Parameters:

```

      Estimate Std. Error t value Pr(>|t|)
a  22.5066      0.8373  26.882 4.32e-12 *** b   2.1063      0.2351
8.959 1.16e-06 *** g   0.3881      0.0459   8.455 2.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1.024 on 12 degrees of freedom

Number of iterations to convergence: 7 Achieved convergence  
tolerance: 8.316e-06

```

## Gráfico valores observados e ajustados versus a covariável x
plot(x, y, ylim=c(0,22), ylab="") lines(x, fitted(mod2), lty=1,
col="blue") legend("bottomright", c("observado","ajustado"),
text.col="black", lty=c(-1,1),
                pch=c(1, -1), col=c("black","blue"), bty = "n")

```



A.?? Programa R para os dados do Exemplo ?? - Modelo para explicar a resistência de um termostato

```
## Lendo os dados termostato <-
read.table(file="C:/UFRPE/MLG/Análises/dados_termostato.txt",
header=T) attach(termostato)

## Transformando y ly <- log(y)

## Ajustando um modelo de regressão normal não-linear mod1 <-
nls(ly ~ -a + d/(g + x), start=list(a=1.0, d=3.0, g=2.0))
summary(mod1)

Formula: ly ~ -a + d/(g + x)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
a 1.205e+01  1.693e-02   712.0    <2e-16 ***
d 6.149e+03  1.455e+01
422.6    <2e-16 ***
g 3.441e+02  5.073e-01   678.3    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0001724 on 13 degrees of freedom

Number of iterations to convergence: 10 Achieved convergence
tolerance: 5.573e-06

## Gráfico valores observados e ajustados versus a covariável x
plot(x, y, ylab="") lines(x, exp(fitted(mod1)), lty=1, col="blue")
```

```
legend("topright", c("observado","ajustado"), text.col="black",  
lty=c(-1,1),  
pch=c(1, -1), col=c("black","blue"), bty="n")
```

A.?? Programa R para os dados do Exemplo ?? - Modelo para explicar a fração média de cloro disponível num produto manufaturado

```
## Lendo os dados cloro=
read.table(file="C:/UFRPE/MLG/Análises/dados_cloro.txt",header=T)
attach(cloro)
```

```
## Ajustando um modelo de regressão não linear mod1<- nls(y ~
exp(b1 + b2 * x2), start=list(b1=0.6, b2=0.01)) summary(mod1)
```

Formula:  $y \sim \exp(b1 + b2 * x2)$

Parameters:

	Estimate	Std. Error	t value	Pr(> t )
b1	-0.7159925	0.0186375	-38.417	< 2e-16 ***
b2	-0.0061323	0.0007214	-8.501	2.5e-07 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0133 on 16 degrees of freedom

Number of iterations to convergence: 7 Achieved convergence  
tolerance: 4.224e-07

```
## Gráfico valores observados e ajustados versus a covariável x2
plot(x2, y, ylim=c(0.38,0.50), xlab=expression(paste(x[2])),
ylab="y") lines(x2, fitted(mod1), lty=1, col="blue")
legend("topright", c("observado","ajustado"), text.col="black",
lty=c(-1,1),
      pch=c(1, -1), col=c("black","blue"), bty = "n")
```

```

identify(x2, y, col="darkblue")

X <-
matrix(c(exp(coef(summary(mod1))[1,1]+coef(summary(mod1))[2,1]*x2),
          exp(coef(summary(mod1))[1,1]+coef(summary(mod1))[2,1]*x2)*x2),
        length(x2))
Xt <- t(X)
XtX <- Xt*%*%X
H <- X*%*%solve(XtX)*%*%Xt
h <- diag(H) r <- summary(mod1)$resid s <- summary(mod1)$sigma

# Resíduo Ordinário Studentizado ti <- r/(s*sqrt(1-h))

## Gráfico dos resíduos studentizados versus valores da covariável
x2 plot(x2, ti, ylim=c(-2.5,2.5), xlab=expression(paste(x[2])),
        ylab="resíduos studentizados")
abline(-2,0,lty=2, col="red") abline(2,0,lty=2, col="red")
identify(x2, ti, col="darkblue")

```

A.?? Programa R para os dados do Exemplo ?? -



# Referências Bibliográficas

- Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, New York, second edition.
- Aitkin, M.; Francis, B.; Hinde, J. (2005). *Statistical modelling in GLIM 4*. Oxford University Press, Oxford.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Auto Cntl AC-19*, **6**, 716–723.
- Andrews, D. F.; Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society B*, **40**, 87–93.
- Anscombe, F. J. (1953). Contribution to the discussion of h. hotelling’s paper. *J. R. Statist. Soc. B*, **15**, 229–230.
- Anscombe, F. J. (1964). Normal likelihood functions. *Ann. Inst. Statist. Math.*, **16**, 1–19.
- Aranda-Ordaz, F. (1981). On the families of transformations to additivity for binary response data. *Biometrika*, **68**, 357–363.
- Armstrong, R. D.; Frome, E. L. (1976). A branch and bound solution of a restricted least squares problem. *Technometrics*, **18**, 447–450.
- Ashton, W. D. (1972). *The Logit Transformation with Special Reference to its Uses in Bioassay*. Griffin, London.
- Atkinson, A. C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13–20.

- Atkinson, A. C. (1985). *Transformations and Regression*. Oxford University Press, Oxford.
- Atkinson, A. C.; Davison, A. C.; Nelder, J. A.; O'Brien, C. M. (1989). *Model Checking*. Imperial College, London.
- Barndorff-Nielsen, O. E. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons, New York.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. A*, **160**, 268–282.
- Bartlett, M. S. (1947). The use of transformations. *Biometrics*, **3**, 39–52.
- Bates, D. M.; Watts, D. G. (1980). Relative curvature measures of nonlinearity. *J. R. Statist. Soc. B*, **42**, 1–25.
- Beale, E. M. L. (1960). Confidence region in nonlinear estimation. *J. R. Statist. Soc. B*, **22**, 41–76.
- Belsley, D. A.; Kuh, E.; Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. John Wiley, New York.
- Berkson, J. (1944). Application of the logistic function to bioassay. *J. R. Statist. Soc. B*, **39**, 357–365.
- Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. R. Statist. Soc. B*, **25**, 220–233.
- Bliss, C. I. (1935). The calculator of the dosage-mortality curve. *Ann. Appl. Biol.*, **22**, 134–167.
- Box, G. E. P.; Cox, D. R. (1964). An analysis of transformation. *J. R. Statist. Soc. B*, **26**, 211–252.
- Box, G. E. P.; Tiao, P. W. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, London.
- Box, G. E. P.; Tidwell, P. W. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531–550.



- Box, M. J. (1971). Bias in non-linear estimation (with discussion). *J. R. Statist. Soc. B*, **33**, 171–201.
- Buse, A. (1982). The likelihood ratio, wald and lagrange multiplier tests: An expository note. *The American Statistician*, **36**, 153–157.
- Collet, D. (2002). *Modelling binary data*. Chapman and Hall, London, second edition.
- Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.
- Cook, R. D.; Tsai, C. L. (1985). Residual in nonlinear regression. *Biometrika*, **72**, 23–29.
- Cook, R. D.; Tsai, C. L.; Wei, B. C. (1986). Bias in non-linear regression. *Biometrika*, **73**, 615–623.
- Cook, R. D.; Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall, London.
- Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *J. Roy. Statist. Soc. B*, **45**, 401–413.
- Cordeiro, G. M. (1986). *Modelos lineares generalizados*. VII SINAPE, UNICAMP.
- Cordeiro, G. M. (1987). On the corrections to the likelihood ratio statistics. *Biometrika*, **74**, 265–274.
- Cordeiro, G. M. (1993). Bartlett corrections and bias correction for two heteroscedastic regression models. *Communications in Statistics, Theory and Methods*, **22**, 169–188.
- Cordeiro, G. M. (1995). Performance of a bartlett-type modification for the deviance. *Journal of Statistical Computation and Simulation*, **51**, 385–403.
- Cordeiro, G. M. (1999). *Introdução à teoria assintótica*. 22º Colóquio Brasileiro de Matemática, IMPA.
- Cordeiro, G. M. (2004a). Corrected likelihood ratio tests in symmetric nonlinear regression models. *Journal of Statistical Computation and Simulation*, **74**, 609–620.

- Cordeiro, G. M. (2004b). On pearson's residuals in generalized linear models. *Statistics and Probability Letters*, **66**, 213–219.
- Cordeiro, G. M. (2004c). Second-order covariance matrix of maximum likelihood estimates in generalized linear models. *Statistics and Probability Letters*, **66**, 153–160.
- Cordeiro, G. M.; Barroso, L. P. (2007). A third-order bias corrected estimate in generalized linear models. *Test*, **16**, 76–89.
- Cordeiro, G. M.; Cribari-Neto, F.; Aubin, E. Q.; Ferrari, S. L. P. (1995). Bartlett corrections for one-parameter exponential family models. *Journal of Statistical Computation and Simulation*, **53**, 211–231.
- Cordeiro, G. M.; McCullagh, P. (1991). Bias correction in generalized linear models. *J. Roy. Statist. Soc. B*, **53**, 629–643.
- Cordeiro, G. M.; Paula, G. A. (1989a). Fitting non-exponential family nonlinear models in glim by using the offset facilities. *Lecture Notes in Statistics*, **57**, 105–144.
- Cordeiro, G. M.; Paula, G. A. (1989b). Improved likelihood ratio statistics for exponential family nonlinear models. *Biometrika*, **76**, 93–100.
- Cox, D. R. (1970). *Analysis of binary data*. Chapman and Hall, London.
- Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B*, **74**, 187–220.
- Cox, D. R.; Hinkley, D. V. (1986). *Theoretical Statistics*. University Press, Cambridge.
- Cox, D. R.; Snell, E. J. (1968). A general definition of residual (with discussion). *J. R. Statist. Soc. B*, **30**, 248–275.
- Demétrio, C. G. B.; Hinde, J. (1997). Half normal plots and overdispersion. *GLIM Newsletter*, **27**, 19–26.
- Dey, D. K.; Gelfand, A. E.; Peng, F. (1997). Overdispersion generalized linear models. *Journal of Statistical Planning and Inference*, **68**, 93–107.

- Dobson, A. J. (2001). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, London, second edition.
- Dyke, G.; Patterson, H. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics*, **8**, 1–12.
- Fahrmeir, L.; Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, **13**, 342–368.
- Fahrmeir, L.; Tutz, G. (1994). *Multivariate Statistical Modelling based on Generalized Linear Models*. Springer-Verlag, New York.
- Feigl, P.; Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, **21**, 826–838.
- Finney, D. (1952). *Probit Analysis*. Cambridge University Press, London, second edition.
- Firth, D. (1991). Generalized linear models. In Hinkley, D.; Reid, N.; Snell, E., editors, *Statistical Theory and Modelling*, pages 55–82. Chapman & Hall.
- Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, **222**, 309–368.
- Fisher, R. (1925). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.
- Fisher, R. (1935). The case of zero survivors (appendix to bliss, c.i. (1935)). *Ann. Appl. Biol.*, **22**, 164–165.
- Fisher, R.; Yates, F. (1970). *Statistical Tables for Biological, Agricultural and Medical Research*. Oliver and Boyd, Edinburgh.
- Folks, J.; Chhikara, R. (1978). The inverse gaussian distribution and its statistical application, a review. *J. R. Statist. Soc. B*, **40**, 263–289.
- Francis, B.; Green, M.; Payne, C. (1993). *The GLIM system generalized linear interactive modelling*. New York.

- Gasser, M. (1967). Exponential survival with covariance. *Journal of the American Statistical Association*, **62**, 561–568.
- Gelfand, A.; Dalal, S. (1990). A note on overdispersed exponential families. *Biometrika*, **77**, 55–64.
- Gentle, J.; Kennedy, W. (1979). Algorithms for linear regression with linear restrictions. In Gentlemen, J., editor, *Proceedings of the Computer Science and Statistics: 12th Annual Symposium of the Interface*. University of Waterloo, Waterloo.
- Giltinan, D.; Capizzi, T.; Malani, H. (1988). Diagnostic tests for similar action of two compounds. *Appl. Statist.*, **37**, 39–50.
- Guttman, I.; Meeter, D. A. (1965). On beale's measures of nonlinearity. *Technometrics*, **7**, 623–637.
- Haberman, S. (1970). *The general log-linear model. PhD dissertation*. Univ. of Chicago Press, Chicago, Illinois.
- Haberman, S. (1974). *The analysis of frequency data*. Univ. of Chicago Press, Chicago, Illinois.
- Haberman, S. (1978). *Analysis of quantitative data*, volume 1. Academic Press, New York.
- Hauck, W.; Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, **72**, 851–853.
- Hinde, J.; Demétrio, C. G. B. (1998a). *Overdispersion: Models and Estimation*. XIII SINAPE, São Paulo.
- Hinde, J.; Demétrio, C. G. B. (1998b). Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.
- Johnson, N. L.; Kotz, S.; Balakrishnan, N. (2004). *Continuous univariate distributions*. John Wiley & Sons, New York, second edition.
- Jørgensen, B. (1987). Maximum likelihood estimates and large samples inference for generalized linear and nonlinear regression models. *Biometrika*, **70**, 19–28.

- Judge, G. G.; Griffiths, W. E.; Hill, R. C.; Lütkepohl, H.; Lee, T.-C. (1985). *The theory and practice of Econometrics*. John Wiley & Sons, New York.
- Larsen, W. A.; McCleary, S. J. (1972). The use of partial residual plots in regression analysis. *Technometrics*, **14**, 781–790.
- Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, **43**, 295–303.
- Lee, Y.; Nelder, J. A.; Pawitan, Y. (2006). *Generalized Linear Models with Random Effects. Unified Analysis via H-likelihood*. Chapman & Hall/CRC, London.
- Mallows, C. L. (1966). *Choosing a subset regression*. Presented at Annual A.S.A. Meetings, Los Angeles.
- Mantel, N.; Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.*, **22**, 719–748.
- Martin, J. T. (1942). The problem of the evaluation of rotenone-containing plants. vi: The toxicity of 1-elliptone and of poisons applied jointly, with further observations on the rotenone equivalent method of assessing the toxicity of derris root. *Annals of Applied Biology*, **29**, 69–81.
- McCullagh, P.; Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, second edition.
- McCulloch, C. E.; Searle, S. R. (2000). *Generalized, Linear, and Mixed Models*. John Wiley & Sons, New York.
- McDonald, J. W.; Diamond, . I. D. (1982). On the fitting of generalized linear models with non-negativity parameter constraints using the kuhn tucker conditions of nonlinear programming. *Unpublished*.
- McDonald, J. W.; Diamond, . I. D. (1983). Fitting generalized linear models with linear inequality parameter constraints. *GLIM Newsletter*, **8**, 29–36.
- Mendenhall, P.; Scheaffer, R. L.; Wackerly, D. D. (1981). *Mathematical Statistics with Applications*. Duxbury, Boston.

- Molenberghs, G.; Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer-Verlag, New York.
- Morgan, C. N. (1992). *Analysis of Quantal Response Data*. Chapman and Hall, London.
- Morris, C. N. (1982). Natural exponential families with quadratic variance functions: statistical theory. *Annals of Statistics*, **11**, 515–529.
- Nelder, J. A. (1966). Inverse polynomials, a useful group of multifactor response functions. *Biometrics*, **22**, 128–141.
- Nelder, J. A.; Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, A*, **135**, 370–384.
- Paula, G. A. (1987). Projected residuals in glim. *Biometrika*, **15**, 28–31.
- Paula, G. A. (2004). *Modelos de Regressão com Apoio Computacional*. IME/USP, São Paulo.
- Paulino, C. D.; Singer, J. M. (2006). *Análise de dados categorizados*. Editora Edgard Blücher, São Paulo.
- Phelps, K. (1982). Use of the complementary log-log function to describe dose response relationship in insecticide evaluation field trials. In *GLIM 82: Proceedings of the International Conference on Generalized Linear Models. Lecture notes in Statistics*, volume 14, pages 155–163. Springer-Verlag, New York.
- Pregibon, D. (1979). *Data analytic methods for generalized linear models. PhD Thesis*. University of Toronto, Toronto.
- Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.*, **29**, 15–24.
- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rao, C. R. (1973). *Linear statistical inference and its applications*. John Wiley, New York.

- Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut, Copenhagen.
- Ratkowsky, D. A. (1983). *Nonlinear regression modelling*. Marcel Dekker, New York.
- Ridout, M. S. (1990). *Using Generalized Linear Models to Analyze Data from Agricultural, and Horticultural Experiments*. Departamento de Matemática e Estatística da ESALQ/USP, Piracicaba (não publicado).
- Ridout, M. S.; Demétrio, C. G. B. (1992). Generalized linear models for positive count data. *Revista de Matemática e Estatística*, **10**, 139–148.
- Ridout, M. S.; Demétrio, C. G. B.; Hinde, J. (1998). Models for count data with many zeros. *Proceedings of XIXth International Biometrics Conference, Cape Town, Invited Papers*, pages . 179–192.
- Ridout, M. S.; Fenlon, J. (1998). *Statistics in Microbiology*. Horticultural Station, East Malling (Notes for workshop).
- Ridout, M. S.; Hinde, J.; Demétrio, C. G. B. (2001). A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57**, 219–223.
- Ryan, B. F.; Joiner, B. L.; Ryan Jr., T. A. (1976). *Minitab Student Handbook*. Duxbury Press, New York.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
- Searle, S. (1982). *Linear models*. John Wiley, New York.
- Silveira Neto, S.; Nakano, O.; Barbin, D.; Villa Nova, N. (1976). *Manual de Ecologia dos Insetos*. Ed. Agronômica 'Ceres', São Paulo.
- Silvey, S. (1975). *Statistical Inference*. Chapman and Hall', London, second edition.
- Smyth, G. (1989). Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society B*, **51**, 47–60.

- Theil, H. (1965). The analysis of disturbances in regression analysis. *Journal of the American Statistical Association*, **60**, 1067–1079.
- Theil, H.; Van de Panne, C. (1960). Quadratic programming as an extension of classical quadratic maximization. *Management Science*, **7**, 1–20.
- Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics*, **5**, 232–242.
- Vaeth, M. (1985). On the use of wald's test in exponential families. *Int. Statistical Review*, **53**, 199–214.
- Vieira, A.; Hinde, J.; Demétrio, C. (2000). Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, **27**, 373–389.
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, **54**, 426–482.
- Wang, P. (1985). Adding a variable in generalized linear models. *Technometrics*, **27**, 273–276.
- Wang, P. (1987). Residual plots for detecting nonlinearity in generalized linear models. *Technometrics*, **29**, 435–438.
- Waterman, M. (1977). Least squares with non negative regression coefficients. *Journal of Statistical Computation and Simulation*, **6**, 67–70.
- Weisberg, S. (2005). *Applied linear regression*. John Wiley, New York, third edition.
- Wilks, S. (1937). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.*, **9**, 60–62.
- Zippin, C.; Armitage, P. (1966). Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, **22**, 665–672.