MODELOS PARAMÉTRICOS

por

Gauss Moutinho Cordeiro
Departamento de Estatística e Informática,
Universidade Federal Rural de Pernambuco,
Rua Dom Manoel de Medeiros, s/n,
50.171-900 – Recife, PE, Brasil
e
Eufrásio de Andrade Lima Neto
Departamento de Estatística,
Universidade Federal da Paraíba,
Cidade Universitária, s/n,
58.051-900 – João Pessoa, PB, Brasil

Prefácio

Este texto objetiva apresentar alguns modelos de regressão para análise de dados univariados. Não se pretende abrir todos os modelos de regressão, mas sim abordar os principais modelos usados na prática de uma forma resumida e consistente.

Existe uma vasta literatura destinada a estudar – de forma isolada – os seguintes modelos: os modelos normal-linear, os modelos para a análise de dados categorizados, os modelos lineares generalizados e os modelos aditivos generalizados. O pré-requisito para a leitura deste texto é um Curso de Inferência Estatística com base em Teoria da Verossimilhança ao nível de graduação. O texto, dividido em 6(seis) capítulos, se destina prioritariamente a alunos de mestrado e doutorado embora possa também, ser utilizado por alunos dos últimos anos de graduação.

O Capítulo 1 descreve o modelo clássico de regressão e o Capítulo 2 trata dos modelos lineares generalizados. Técnicas de diagnóstico nesses modelos são descritas no Capítulo 3. Os principais modelos lineares generalizados e algumas de suas extensões são apresentados no Capítulo 4. Outros modelos de regressão importantes como o modelo normal não-linear, os modelos heterocedásticos e autocorrelacionados são tratados no Capítulo 5.

Finalmente, no Capítulo 6, apresentam-se análises de dados reais através dos sistemas S-PLUS e GLIM.

Agradecemos ao Oscar P. da Silva Neto pelo trabalho de preparação dos originais.

Recife, dezembro de 2006.

Gauss M. Cordeiro

Eufrásio de A. Lima Neto

Conteúdo

1 Modelo Clássico de Regressão **1**

1.1 Introdução 1

1.2 Estimação 2

1.3 Somas de Quadrados 6

1.4 Propriedades do EMQ e dos Resíduos 7

1.5 Modelo Normal-Linear 10

1.6 Análise de Variância 11

1.7 Seleção das Variáveis Explicativas 15

1.8 Intervalos e Regiões de Confiança 16

1.9 Técnicas de Diagnóstico 19

1.9.1 Matriz de projeção 20

1.9.2 Resíduos 21

1.9.3 Influência 23

1.9.4 Técnicas gráficas 25

1.10 Estimação de Máxima Verossimilhança 29

1.11 Exercícios 32

2 Modelos Lineares Generalizados **35**

2.1 Introdução 35
2.2 Um Esboço Sobre os MLGs 36
2.2.1 Formulação do modelo 36
2.3 As Componentes de um MLG 37
2.3.1 Componente aleatória 37
2.3.2 A componente sistemática e a função de ligação 40
2.3.3 Estatísticas suficientes e ligações canônicas 41
2.3.4 A matriz modelo 41
2.4 O Algoritmo de Estimação 43
2.5 Adequação do Modelo 47
2.6 Predição 48
2.7 Medidas de Discrepância ou Bondade de Ajuste 48
2.7.1 A função desvio 48
2.7.2 A estatística de Pearson generalizada X^2 50
2.7.3 A análise do desvio 50
2.8 Modelo Binomial 52
2.8.1 Momentos e cumulantes 53
2.8.2 Convergência para normal e Poisson 53
2.8.3 Funções de ligação apropriadas 54
2.8.4 A função de verossimilhança 60
2.8.5 Estimação dos parâmetros 61
2.8.6 A função desvio 61
2.9 Modelo de Poisson 62
2.9.1 A distribuição de Poisson 62
2.9.2 Função geratriz de momentos e cumulantes 63

2.9.3 A Função de ligação 63
2.9.4 Função desvio e principais transformações 64
2.9.5 O parâmetro de dispersão 65
2.9.6 A distribuição multinomial e a Poisson 65
2.10 Modelo Normal 66
2.10.1 Cumulantes e estimação 67
2.11 Modelo Gama 67
2.11.1 A distribuição gama 68
2.11.2 A função de variância 69
2.11.3 O desvio 69
2.11.4 A função de ligação 70
2.11.5 Estimação do parâmetro de dispersão 71
2.12 Modelo Normal Inverso 72
2.12.1 A função densidade 72
2.12.2 Principais características 72
2.13 Exercícios 73

3 Análise de Resíduos e Diagnóstico em Modelos Lineares Generalizados 77

3.1 Resíduos 77
3.1.1 Resíduo de Pearson 77
3.1.2 Resíduo de Anscombe 78
3.1.3 Desvio residual 79
3.1.4 Comparação entre os resíduos 79
3.2 Análise Residual e Medidas de Influência 82
3.2.1 O resíduo de Cox-Snell e o desvio residual 83

3.2.2 Situações assintóticas 85
3.2.3 Correção de viés para o desvio residual 85
3.3 Verificação da Distribuição dos Resíduos 87
3.3.1 Teste de normalidade 87
3.3.2 Erro de classificação na distribuição dos dados 90
3.4 Verificando a Inclusão de uma Nova Covariável 92
3.5 Verificando a Não-Linearidade em um Sub-Conjunto de Variáveis Explicativas 93
3.6 Verificando a Função de Ligação e de Variância 95
3.7 Correção de Continuidade Residual no Modelo Logístico 95
3.8 Detectando Pontos de Influência 97
3.8.1 Medidas de alavancagem 97
3.8.2 Medidas de influência 98
3.9 Exercícios 99

4 Principais Modelos Lineares Generalizados e Extensões 101
4.1 Modelos para Dados Contínuos 101
4.2 Modelo Logístico Linear 102
4.2.1 Ajuste do modelo 103
4.2.2 Bondade de ajuste 105
4.3 Modelo Log-Linear para Contagens 106
4.3.1 Modelos hierárquicos 107
4.3.2 Modelos hierárquicos para tabelas de contingência com 3 entradas 109
4.3.3 Testes de adequação 112
4.3.4 Testes de comparação entre modelos 113

4.4 Modelo para Dados Multinomiais 115
4.4.1 Momentos e cumulantes 116
4.4.2 Log verossimilhança e função desvio 116
4.5 Modelos com Parâmetros Adicionais Não-Lineares 117
4.5.1 Parâmetros na função de variância 118
4.5.2 Parâmetros na função de ligação 119
4.5.3 Parâmetros não-lineares nas covariáveis 121
4.6 Modelo de Box e Cox 122
4.7 Modelo Linear Generalizado com um Parâmetro Não-Linear Extra 126
4.8 Modelos Lineares Generalizados com Ligação Composta 127
4.9 Modelos Semi-Paramétricos 128
4.10 Modelos Aditivos Generalizados 128
4.11 Modelos de Quase-Verossimilhança 130
4.12 Modelos para Análise de Dados de Sobrevivência 136
4.12.1 Modelos de riscos proporcionais 137
4.12.2 Riscos proporcionais de Cox 139
4.13 Modelos Lineares Generalizados com Covariáveis de Dispersão . 141
4.14 Modelos Lineares Generalizados com Super-dispersão 145
4.15 Exercícios 149

5 Outros Modelos de Regressão Importantes 153
5.1 Modelos com Matriz de Covariância Não-Escalar 153
5.2 Modelo de Regressão Rígida 156
5.3 Modelo Normal Não-Linear 158
5.3.1 Estimação de máxima verossimilhança 159
5.3.2 Resultados assintóticos 161

5.3.3 Técnicas de diagnóstico 163
5.3.4 Medidas de Influência 166
5.3.5 Gráfico da Variável Adicionada 167
5.4 Modelos Heterocedásticos 167
5.5 Modelos Autocorrelacionados 172
5.6 Exercícios 174

6 Análise de Dados Reais através dos Sistemas GLIM e S-Plus177
6.1 O sistema S-Plus 177
6.2 Sistema de Avaliação - Uma Introdução 178
6.3 O Banco de Dados 179
6.4 Modelo para as Casas 180
6.5 Modelo para os Apartamentos 191
6.6 O sistema GLIM 201
6.7 Entrada dos dados 203
6.8 Uma seqüência típica de diretivas 203
6.9 Definição e Ajustamento de um MLG 205
6.10 Assinaturas de TV a Cabo 205
6.11 Demanda de Energia Elétrica 216
6.12 Importação Brasileira 222

Capítulo 1

Modelo Clássico de Regressão

1.1 Introdução

A análise de dados através da regressão linear é uma das técnicas mais usadas de estimação, existindo uma ampla literatura sobre o assunto. Os seguintes livros contém os principais tópicos relacionados com regressão linear: Scheffé (1959), Searle (1971), Rao (1973), Seber (1977), Arnold (1981), Draper e Smith (1981), Cook e Weisberg (1982), Montgomery e Peck (1982), Weisberg (1985) e Wetherill et al. (1986). O principal objetivo deste capítulo é apresentar alguns conceitos básicos de regressão linear que visam a facilitar a compreensão dos capítulos seguintes, onde serão apresentados modelos de regressão mais amplos.

O modelo clássico de regressão teve origem nos trabalhos de astronomia elaborados por Gauss no período de 1809 a 1821. É a técnica mais adequada quando se deseja estudar o comportamento de uma variável dependente y (variável resposta) em relação a outras variáveis independentes (variáveis explicativas) que são responsáveis pela variabilidade da variável resposta. O modelo clássico de regressão é definido por:

i) respostas y_i independentes (ou pelo menos não correlacionadas) para $i = 1, \ldots, n$, cada y_i tendo uma distribuição especificada de média $\mu_i =$

$E(y_i)$ e variância σ^2 constante;

ii) a média μ_i é expressa de forma linear como $\mu_i = x_i^T\beta$, onde x_i^T é um vetor 1xp com os valores de p variáveis explicativas relacionadas à i-ésima resposta y_i e β é um vetor px1 de parâmetros a serem estimados.

A estrutura i) e ii) pode também ser expressa na forma matricial $\mu = E(y) = X\beta$, onde $y = (y_1, \ldots, y_n)^T$ é um vetor nx1 cuja i-ésima componente é y_i e X é uma matriz nxp formada pelas linhas $x_1^T, \ldots, x_n^T$. Em geral, adota-se a hipótese de aditividade entre y e μ, isto é, $y = \mu + \epsilon$, onde ϵ é um vetor de erros de média zero e variância σ^2 constante. Os erros são considerados independentes ou pelos menos não-correlacionados. Os efeitos das variáveis explicativas, que formam as colunas da matriz X, sobre a variável resposta y são lineares e aditivos. Na formação da matriz modelo, considera-se geralmente a primeira coluna como um vetor de uns sendo o parâmetro correspondente denominado *intercepto.*

O objetivo inicial é estimar β a partir do vetor y de dados e da matriz modelo X conhecida, suposta de posto completo p. A estimação pelo *Método de Mínimos Quadrados* não requer qualquer hipótese sobre a distribuição das componentes do vetor y. Este método consiste em minimizar $\sum_i (y_i - \mu_i)^2$. Outras normas podem, também, ser adotadas como $min \sum_i \mid y_i - \mu_i \mid$ ou $max_i \mid y_i - \mu_i \mid$, produzindo métodos alternativos de estimação. O método de estimação M (Huber, 1973) substitui a soma de quadrados dos erros $\sum_i \epsilon_i^2$ por $\sum_i \rho(\epsilon_i)$, onde $\rho(\epsilon_i)$ é uma função simétrica. A escolha entre os métodos pode ser baseada na suposição da distribuição dos erros ϵ ou no programa computacional disponível. Entretanto, segundo as hipóteses i) e ii), o método de mínimos quadrados continua sendo o método preferido entre estes métodos de estimação.

1.2 Estimação

Adota-se a seguinte notação matricial para representar o modelo clássico de regressão

$$y = X\beta + \epsilon, \tag{1.1}$$

em que está expresso a aditividade entre os efeitos lineares sistemáticos em $\mu = X\beta$ e os efeitos aleatórios em ϵ, supondo ainda que $Cov(\epsilon) = \sigma^2 I$. A soma de quadrados dos erros SQE(β) $= \sum_i (y_i - \mu_i)^2$ correspondente ao modelo (1.1) é dada em notação matricial por

$$SQE(\beta) = (y - X\beta)^T (y - X\beta). \tag{1.2}$$

Para estimar β minimiza-se SQE(β) em relação a β, ou seja, minimiza-se o quadrado da distância entre os vetores y e $\mu = X\beta$. Esta minimização implica em resolver o sistema de p equações lineares dadas por

$$\frac{\partial SQE(\beta)}{\partial \beta_r} = 2 \sum_{i=1}^{n} x_{ir}(y_i - \mu_i) = 0, \tag{1.3}$$

para $r = 1, \ldots, p$. O sistema (1.3) em notação matricial é expresso por $X^T(y - X\beta) = 0$, ou, equivalentemente, $X^T X\beta = X^T y$. Estas p equações lineares são conhecidas como *equações normais*. Como a matriz modelo X tem posto completo, a matrix $X^T X$ é inversível e, portanto, a solução do sistema de equações normais é única. Esta solução corresponde ao *estimador de mínimos quadrados* (EMQ) de β dado por

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{1.4}$$

O EMQ $\hat{\beta}$ em (1.4), segundo o modelo (1.1), tem as seguintes propriedades: i) $\hat{\beta}$ minimiza a soma de quadrados dos erros $\sum_i \epsilon_i^2$, independentemente da distribuição proposta para os erros. Não é necessário conhecer a distribuição dos erros para estimar β mas precisa-se da normalidade para fazer inferência sobre os parâmetros em β. Esta inferência baseia-se nas distribuições t de Student e F de Snedecor; ii) as componentes do vetor $\hat{\beta}$ são funções lineares das observações e são estimadores não-viesados de menor variância dos parâmetros em β, comparando-os com quaisquer combinações lineares das observações, independentemente da distribuição considerada para os erros. O EMQ $\hat{\beta}$ em (1.4) pode ser escrito como função dos erros não observados por

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon. \tag{1.5}$$

A diferença $\hat{\beta} - \beta$ entre o EMQ e o vetor verdadeiro β de parâmetros não pode ser calculada pela equação (1.5), pois o vetor de erros ϵ não é observado. Entretanto, esta equação é importante no estudo das propriedades do EMQ $\hat{\beta}$.

No caso da matriz $A = X^T X$ ser singular, ou seja, algumas das equações normais dependem de outras equações de modo que há menos de p equações independentes para estimar os p parâmetros $\beta_1, \ldots, \beta_p$, o sistema (1.3) admitirá uma infinidade de soluções. Entretanto, se o mesmo for consistente (se existir $\hat{\beta}$), existem matrizes A^- tais que $\hat{\beta} = A^- y$ é uma solução de (1.3). As matrizes A^- dependem somente de $X^T X$ e em geral não são únicas, exceto quando $X^T X$ for não-singular. Tais matrizes são chamadas de *inversas generalizadas.*

No método de estimação de Huber (1973), citado anteriormente, a minimização de $\sum_i \rho(\epsilon_i)$ em relação a β produz o sistema de p equações não-lineares

$$\sum_{i=1}^{n} x_{ir} \rho^{(1)}(y_i - \mu_i) = 0, \tag{1.6}$$

em que $\rho^{(1)}(\epsilon) = \partial\rho(\epsilon)/\partial\mu$. Se a função $\rho(\cdot)$ é quadrática, o EMQ (1.4) segue diretamente de (1.6).

Exemplo 1.1: *Regressão Linear Simples.*

Considere uma única variável explicativa x para representar o comportamento de uma variável resposta y cuja média é dada pela equação linear $E(y) = \mu = \beta_0 + \beta_1(x - \bar{x})$. Pode-se estimar o vetor $\beta = (\beta_0, \beta_1)^T$ a partir da equação (1.4), obtendo-se o EMQ de β como

$$\hat{\beta} = \begin{pmatrix} n & \sum_i (x_i - \bar{x}) \\ \sum_i (x_i - \bar{x}) & \sum_i (x_i - \bar{x})^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_i y_i \\ \sum_i (x_i - \bar{x}) y_i \end{pmatrix}$$

que, finalmente, reduz-se à $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2} \end{pmatrix}$.

Logo, o intercepto β_0 é estimado pela média $\bar{y}$ das observações.

Exemplo 1.2: *Regressão Linear Múltipla.*

Apresentamos agora um exemplo de regressão linear mútipla na estimação do consumo de combustível nos estados americanos. Sejam as seguintes variáveis: Cons = consumo de gasolina em galões per capita-ano, Taxa = valor do imposto estadual em cents por galão de combustível, Rend = renda média em US$, Rodov = extensão em milhas da malha estadual e Licen = percentual da população habilitada a dirigir. Os dados constam de Weisberg (1985; Tabela 1.4). Assim, o interesse é estimar os cinco parâmetros do modelo de regressão linear múltipla: $E(Cons) = \beta_0 + \beta_1 Tax + \beta_2 \mathrm{Re}n + \beta_3 Rod + \beta_4 Lic$, a partir das 48 observações de cada variável.

Tabela 1.1: *Consumo de Combustível nos Estados Americanos*

Con	Tax	Ren	Rod	Lic	Con	Tax	Ren	Rod	Lic
541	9.0	3571	1976	52.5	460	8.5	4574	2619	55.1
524	9.0	4092	1250	57.2	566	9.0	3721	4746	54.4
561	9.0	3865	1586	58.0	577	8.0	3448	5399	54.8
414	7.5	4870	2351	52.9	631	7.5	3846	9061	57.9
410	8.0	4399	431	54.4	574	8.0	4188	5975	56.3
457	10.0	5342	1333	57.1	534	9.0	3601	4650	49.3
344	8.0	5319	11868	45.1	571	7.0	3640	6905	51.8
467	8.0	5126	2138	55.3	554	7.0	3333	6594	51.3
464	8.0	4447	8577	52.9	577	8.0	3063	6524	57.8
498	7.0	4512	8507	55.2	628	7.5	3357	4121	54.7
580	8.0	4391	5939	53.0	487	8.0	3528	3495	48.7
471	7.5	5126	14186	52.5	644	6.5	3802	7834	62.9
525	7.0	4817	6930	57.4	640	5.0	4045	17782	56.6
508	7.0	4207	6580	54.5	704	7.0	3897	6385	58.6
566	7.0	4332	8159	60.8	648	8.5	3635	3274	66.3
635	7.0	4318	10340	58.6	968	7.0	4345	3905	67.2
603	7.0	4206	8508	57.2	587	7.0	4449	4639	62.6
714	7.0	3718	4725	54.0	699	7.0	3656	3985	56.3
865	7.0	4716	5915	72.4	632	7.0	4300	3635	60.3
640	8.5	4341	6010	67.7	591	7.0	3745	2611	50.8
649	7.0	4593	7834	66.3	782	6.0	5215	2302	67.2
540	8.0	4983	602	60.2	510	9.0	4476	3942	57.1
464	9.0	4897	2449	51.1	610	7.0	4296	4083	62.3
547	9.0	4258	4686	51.7	524	7.0	5002	9794	59.3

1.3 Somas de Quadrados

O valor mínimo da soma de quadrados dos erros é denominado *soma de quadrados dos resíduos* (SQR), pois mede a discrepância entre o vetor de observações y e o vetor de valores ajustados (ou médias ajustadas) $\hat{\mu} = X\hat{\beta}$. Assim, SQR é expresso por

$$SQR = SQE(\hat{\beta}) = (y - X\hat{\beta})^T(y - X\hat{\beta}). \tag{1.7}$$

Pode-se verificar facilmente que $\hat{\mu} = X(X^TX)^{-1}X^Ty = Hy$, onde a matrix H é denominada *matriz de projeção.* A razão desta terminologia é que o vetor $\hat{\mu}$ dos valores ajustados é a projeção ortogonal do vetor de dados y no espaço gerado pelas colunas da matriz X.

A matriz H é simétrica ($H = H^T$), idempotente ($H^2 = H$) e tem posto p. Assim, o vetor $\hat{\beta}$ que minimiza a distância (1.2) entre y e $\mu = X\beta$ é tal que o vetor $\hat{\mu}$ dos valores ajustados é a projeção ortogonal do vetor y das observações sobre o plano gerado pelas colunas da matriz X.

O vetor de erros não-observados $\epsilon = y - X\beta$ é estimado pelo vetor de resíduos r, dado por

$$r = y - \hat{\mu} = y - X\hat{\beta}. \tag{1.8}$$

Tem-se $r = y - Hy = (I - H)y$, onde I representa a matriz identidade de ordem n. É fácil verificar que o vetor de resíduos r e o vetor $\hat{\mu}$ de valores ajustados são ortogonais. Com efeito,

$$r^T\hat{\mu} = y^T(I - H)^THy = 0,$$

pois H é simétrica e idempotente. Temos, ainda, $r^Tr = (y - \hat{\mu})^T(y - \hat{\mu}) = y^T(I - H)^T(I - H)y = y^Ty - \hat{\mu}^T\hat{\mu}$ e, portanto,

$$y^Ty = \hat{\mu}^T\hat{\mu} + r^Tr. \tag{1.9}$$

A equação (1.9) mostra que a soma de quadrados dos dados (y^Ty) iguala a soma de quadrados dos valores ajustados ($\hat{\mu}^T\hat{\mu}$) mais a soma de quadrados dos

resíduos ($r^T r$). Esta equação é uma simples aplicação do teorema de Pitágoras, onde a hipotenusa é o vetor de dados y e os catetos são os vetores das médias ajustadas $\hat{\mu}$ e dos resíduos $r = y - \hat{\mu}$. Assim, a soma de quadrados das observações $y^T y$ pode ser decomposta em duas partes: a soma de quadrados dos valores ajustados $\hat{\mu}^T \hat{\mu} = \hat{\beta}^T X^T y$ e a soma de quadrados dos resíduos $SQR = r^T r = (y - \hat{\mu})^T (y - \hat{\mu})$, que mede a variabilidade dos dados não-explicada pela regressão (vide Seção 1.6).

1.4 Propriedades do EMQ e dos Resíduos

Nesta seção apresentamos algumas propriedades de $\hat{\beta}$ que são baseadas apenas nas duas hipóteses básicas atribuídas aos dois primeiros momentos dos erros: $E(\epsilon) = 0$ e $Cov(\epsilon) = \sigma^2 I$.

a) O EMQ $\hat{\beta}$ é Não-Viesado.

A esperança do EMQ $\hat{\beta}$ é obtida de (1.5) como

$$E(\hat{\beta}) = E\{\beta + (X^T X)^{-1} X^T \epsilon\} = \beta + (X^T X)^{-1} X^T E(\epsilon) = \beta.$$

Logo, o EMQ $\hat{\beta}$ tem esperança igual ao próprio vetor β de parâmetros sendo, portanto, um estimador não-viesado.

b) Covariância do EMQ $\hat{\beta}$.

A matriz de covariância do EMQ $\hat{\beta}$ é obtida de

$$Cov(\hat{\beta}) = E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]^T\} = E\{[\hat{\beta} - \beta][\hat{\beta} - \beta]^T\}.$$

Usando (1.5) e o fato de que $E(\hat{\beta}) = \beta$, temos

$$\begin{aligned} Cov(\hat{\beta}) &= E\{(X^T X)^{-1} X^T \epsilon \epsilon^T X (X^T X)^{-1}\} \\ &= (X^T X)^{-1} X^T E(\epsilon \epsilon^T) X (X^T X)^{-1}. \end{aligned}$$

Finalmente, como $\text{Cov}(\hat{\beta}) = E(\epsilon\epsilon^T) = \sigma^2 I$, obtém-se

$$Cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}. \tag{1.10}$$

Assim, a matriz inversa $(X^T X)^{-1}$ usada para estimar β em (1.4) determina a matriz de covariância de $\hat{\beta}$ em (1.10), exceto pelo multiplicador σ^2. Os elementos da diagonal da equação (1.10) são as variâncias das estimativas de mínimos quadrados dos parâmetros em β e, portanto, representam a precisão destas estimativas.

c) Covariância do vetor $\hat{\mu}$.

A estrutura de covariância do vetor $\hat{\mu}$ das médias ajustadas segue diretamente da equação (1.10). Temos,

$$Cov(\hat{\mu}) = XCov(\hat{\beta})X^T = \sigma^2 X(X^T X)^{-1} X^T = \sigma^2 H.$$

Assim, a matriz de projeção H representa, exceto pelo escalar σ^2, a matriz de covariância de $\hat{\mu}$. Logo, $\text{Cov}(\hat{\mu}_i, \hat{\mu}_j) = \sigma^2 h_{ij}$, onde h_{ij} é o elemento (i, j) da matriz H. As propriedades desta matriz serão detalhadas na Seção 1.9.1.

d) Estimação de σ^2.

Para determinar as covariâncias de $\hat{\beta}$ e $\hat{\mu}$ torna-se necessário estimar a variância σ^2 dos erros. Para isso usamos o teorema do valor esperado de uma forma quadrática: *Se y é um vetor de média μ e matriz de covariância V, então: $E(y^T Ay) = tr(AV) + E(\mu^T A\mu)$, igualdade válida para qualquer matriz quadrada A.* Logo, de (1.7) e $r = (I - H)y$, obtém-se

$$SQR = y^T (I - H)y$$

e, portanto,

$$E(SQR) = \sigma^2 tr(I - H) + \beta^T X^T (I - H)X\beta.$$

Como $(I - H)X = 0$ e $(I - H)$ é uma matriz simétrica e idempotente, o traço de $(I - H)$ iguala ao seu posto $n - p$, implicando $E(SQR) =$

$\sigma^2(n-p)$. Assim, um estimador não-viesado de σ^2 é dado por

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{(n-p)}. \tag{1.11}$$

Estimando-se σ^2 por (1.11) pode-se calcular as covariâncias das estimativas dos parâmetros da regressão. A grande maioria dos programas computacionais de regressão apresentam as estimativas $\hat{\beta}_1, \ldots, \hat{\beta}_p$ e seus erros padrões $\mathrm{Var}(\hat{\beta}_1)^{1/2}, \ldots, \mathrm{Var}(\hat{\beta}_p)^{1/2}$, que correspondem às raízes quadradas dos elementos da diagonal da matriz (1.10).

e) Esperança e Covariância do Vetor de Resíduos r.

Determinamos agora a média e a covariância do vetor de resíduos $r = y - \hat{\mu}$. A esperança de r é nula, pois $E(r) = y - E(\hat{\mu}) = y - XE(\hat{\beta}) = 0$. O cálculo da matriz de covariância de r segue: $\mathrm{Cov}(r) = Cov(y - \hat{\mu}) = Cov((I-H)y) = (I-H)Cov(y)(I-H)^T = \sigma^2(I-H)$. Logo, a covariância entre os resíduos $r_i = y_i - \hat{\mu}_i$ e $r_j = y_j - \hat{\mu}_j$ relativos às observações de ordens i e j, é dada por

$$Cov(r_i, r_j) = \sigma^2(1 - h_{ij}).$$

Assim, embora os erros aleatórios ϵ_i tenham a mesma variância σ^2, i.e., sejam homocedásticos, o mesmo não ocorre com os resíduos, cujas variâncias dependem dos elementos da diagonal da matriz de projeção H. Tem-se, $\mathrm{Var}(r_i) = \sigma^2(1 - h_{ii})$ e, então, os resíduos definidos em (1.8) são heterocedásticos.

f) Covariância entre $\hat{\beta}$ e r.

Mostramos, agora, que os vetores $\hat{\beta}$ e r são ortogonais, ou seja, $Cov(\hat{\beta}, r) = 0$. Temos,

$$Cov(\hat{\beta}, r) = Cov((X^TX)^{-1}X^Ty, (I-H)y) = (X^TX)^{-1}X^T\sigma^2 I(I-H)^T = 0.$$

O vetor de resíduos r é, também, ortogonal ao vetor das médias ajustadas

$\hat{\mu}$. Em termos algébricos, tem-se

$$\hat{\mu}^T r = y^T H^T (I - H) y = y^T (H - H) y = 0,$$

pois a matriz de projeção H é simétrica e idempotente.

1.5 Modelo Normal-Linear

Para determinarmos a distribuição de probabilidade das estimativas de mínimos quadrados, precisamos especificar a distribuição dos erros aleatórios. A suposição de normalidade dos erros é a mais adotada e considera que os erros aleatórios $\epsilon_1, \ldots, \epsilon_n$ em (1.1) são independentes e têm distribuição normal $N(0, \sigma^2)$. O modelo (1.1) com esta suposição é denominado *modelo normal-linear*. Segundo a hipótese de normalidade dos erros, podemos deduzir as seguintes propriedades que são importantes na análise de regressão:

i) O vetor y tem distribuição normal n-variada $N_n(X\beta, \sigma^2 I)$.

ii) O EMQ $\hat{\beta}$ tem distribuição normal p-variada $N_p(\beta, \sigma^2 (X^T X)^{-1})$.

A média e a estrutura de covariância de $\hat{\beta}$ foram obtidas na Seção 1.4, itens a) e b). A normalidade de $\hat{\beta}$ decorre do fato de $\hat{\beta}$ ser uma função linear do vetor y, cuja distribuição é normal;

iii) O EMQ $\hat{\beta}$ e a soma de quadrados dos resíduos $SQR = y^T (I - H) y$ são independentes.

O vetor de resíduos $r = y - \hat{\mu} = (I - H)y$ tem distribuição normal n-variada $N_n(0, \sigma^2 (I - H))$ e é ortogonal ao EMQ $\hat{\beta}$, conforme visto na Seção 1.4, item f. Assim, como $\hat{\beta}$ e r são ortogonais e têm distribuição normal, estes vetores são independentes. Então, o EMQ $\hat{\beta}$ e a soma SQR são independentes;

iv) SQR/σ^2 tem distribuição qui-quadrado χ^2_{n-p} com $n-p$ graus de liberdade.

Para demonstrar esta propriedade usamos a seguinte decomposição da

soma de quadrados dos erros

$$\frac{\epsilon^T\epsilon}{\sigma^2} = \frac{(y - X\beta)^T(y - X\beta)}{\sigma^2} = \frac{\{r + X(\hat{\beta} - \beta)\}^T\{r + X(\hat{\beta} - \beta)\}}{\sigma^2},$$

que implica em

$$\frac{\epsilon^T\epsilon}{\sigma^2} = \frac{r^Tr}{\sigma^2} + \frac{(\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta)}{\sigma^2}, \tag{1.12}$$

pois $r^TX = 0$. O lado esquerdo de (1.12) é uma soma de quadrados de n variáveis aleatórias normais $N(0,1)$ e, portanto, tem distribuição χ^2_n com n graus de liberdade. De ii) concluímos que a forma quadrática $(\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta)/\sigma^2$ tem distribuição χ^2_p. Como $SQR = r^Tr$ e $\hat{\beta}$ são independentes, o teorema da convolução de qui-quadrados independentes implica que $SQR/\sigma^2 = r^Tr/\sigma^2$ tem distribuição qui-quadrado χ^2_{n-p} com $n - p$ graus de liberdade.

1.6 Análise de Variância

A técnica mais usada para verificar a adequação do ajuste do modelo de regressão a um conjunto de dados é a *Análise de Variância (*sigla *ANOVA)* que se baseia na seguinte identidade

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{\mu}_i - \bar{y})^2 + \sum_i (y_i - \hat{\mu}_i)^2. \tag{1.13}$$

O termo do lado esquerdo de (1.13) é a soma dos quadrados das observações em relação ao seu valor médio e representa uma medida da variabilidade total dos dados. Esta soma será denotada por $SQT = \sum_i (y_i - \bar{y})^2$. O primeiro termo do lado direito de (1.13) é a soma dos quadrados explicada pelo modelo de regressão, sendo denotada por $SQE = \sum_i (\hat{\mu}_i - \bar{y})^2$, enquanto o segundo termo é a soma de quadrados residual $SQR = \sum_i (y_i - \hat{\mu}_i)^2$, que não é explicada pelo modelo de regressão. O modelo será tanto melhor ajustado quanto maior for a variação explicada SQE em relação à variação total SQT. A dedução da equação (1.13) decorre elevando-se ao quadrado os termos da

igualdade $y_i - \bar{y} = (\hat{\mu}_i - \bar{y}) + (y_i - \hat{\mu}_i)$ e somando-se sobre as observações. Tem-se,

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{\mu}_i - \bar{y})^2 + \sum_i (y_i - \hat{\mu}_i)^2 + 2\sum_i (\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i).$$

Mostra-se agora que o último termo desta igualdade é zero. Se 1 é um vetor nx1 de uns, este termo pode ser expresso em notação matricial como

$$\begin{aligned} \sum_i (\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) &= (\hat{\mu} - \bar{y}1)^T (y - \hat{\mu}) = (y^T H - \bar{y}1^T)(I - H)y \\ &= \bar{y}1^T Hy - \bar{y}1^T y = 0, \end{aligned}$$

pois $1^T H = 1^T$ quando a matriz modelo X tem uma coluna de uns correspondente ao intercepto.

As somas de quadrados explicada $SQE = \sum_i (\hat{\mu}_i - \bar{y})^2$ e não-explicada $SQR = \sum_i (y_i - \hat{\mu}_i)^2$ pela regressão podem ser escritas em notação matricial como: $SQE = \hat{\beta}^T X^T y - n\bar{y}^2$ e $SQR = y^T(I - H)y$. Pode-se medir a adequação do ajuste do modelo comparando a soma de quadrados residual SQR (que se espera seja pequena) com a soma de quadrados devida à regressão SQE. Ou, alternativamente, comparando SQE com a soma de quadrados total $SQT = y^T y - n\bar{y}^2$. A razão desses dois termos é representada por

$$R^2 = \frac{SQE}{SQT} = \frac{\hat{\beta}^T X^T y - n\bar{y}^2}{y^T y - n\bar{y}^2}. \tag{1.14}$$

A razão (1.14) varia sempre entre 0 e 1 e R é denominado de *coeficiente de correlação múltipla de Pearson (ou coeficiente de determinação).* Este nome deve-se ao fato de R ser o coeficiente de correlação linear entre os valores observados em y e os valores ajustados em $\hat{\mu}$. Alguns pesquisadores se baseiam erroneamente apenas no valor de R^2 para escolher o melhor modelo. Entretanto, tão importante quanto termos um R^2 próximo de um, é que a estimativa de σ^2 seja também pequena, pois os intervalos de confiança para os parâmetros de interesse são proporcionais a σ.

A equação (1.13) em forma matricial é dada por

$$SQT = SQE + SQR = (\hat{\beta}^T X^T y - n\bar{y}^2) + y^T(I - H)y,$$

que é a equação básica de construção da *Tabela de Análise de Variância*. A cada soma de quadrados nesta fórmula está associado um número de graus de liberdade, que é formalmente obtido expressando a soma de quadrados correspondente em forma quadrática, cujo posto iguala o número de graus de liberdade. As somas $SQE = \hat{\beta}^T X^T y - n\bar{y}^2$ e $SQR = y^T(I - H)y$ têm distribuições $\sigma^2\chi^2_{p-1}$ e $\sigma^2\chi^2_{n-p}$, respectivamente, que são independentes.

A Tabela 1.2 apresenta a Tabela de Análise de Variância usada para testar a adequação global do Modelo de Regressão $y = X\beta + \epsilon$. Testa-se a adequação global do modelo ajustado comparando a estatística $F = \frac{MQE}{MQR}$ obtida desta tabela com o ponto crítico $F_{p-1,n-p}(\alpha)$ da distribuição $F_{p-1,n-p}$ de Snedecor com graus de liberdade $p - 1$ e $n - p$, respectivamente, supondo um nível de significância α. Se o valor da estatística F for superior ao ponto crítico, i.e., $F > F_{p-1,n-p}(\alpha)$, o efeito global de pelo menos algumas das variáveis independentes do modelo é significativo para explicar a variabilidade da variável resposta. Caso contrário, o efeito global destas variáveis para explicar o comportamento da variável dependente não é significativo.

Tabela 1.2: *Tabela de Análise de Variância*

Efeito	Soma de Quadrados	GL	Média de Quadrados	Estatística
Regressão	$SQE = \hat{\beta}^T X^T y - n\bar{y}^2$	$p-1$	$MQE = SQE/(p-1)$	$F = MQE/MQR$
Residual	$SQR = y^T(I-H)y$	$n-p$	$MQR = SQR/(n-p)$	
Total	$SQT = y^T y - n\bar{y}^2$	$n-1$		

Exemplo 1.3: *Continuação da Regressão Linear Múltipla.*

Usamos o software MINITAB para calcular as estimativas dos parâmetros da regressão

$$E(Con) = \beta_0 + \beta_1 Tax + \beta_2 \mathrm{Ren} + \beta_3 Rod + \beta_4 Lic, \tag{1.15}$$

e construir a Tabela de Análise de Variância. Os resultados do ajustamento encontram-se na Tabela 1.3, onde além da equação de regressão ajustada, aparecem em *Predictor* as variáveis explicativas, em *Coef* as estimativas ($\hat{\beta}_r$) dos parâmetros, em *StDev* seus erros padrões, ou seja, as raízes quadradas dos elementos da diagonal da matriz (1.10), ($\hat{\sigma}\sqrt{v_{rr}}$) (vide Seção 1.7) e, também, a estatística T_r.

O coeficiente de determinação de Pearson R^2 mostra que cerca de 67.8% da variabilidade do consumo de combustível nos estados americanos é explicada pelo modelo (1.15) e um menor percentual de 32.2% não é explicado por este modelo. A estatística F, obtida da tabela de análise de variância, iguala $F = 22.63$ que é muito superior ao ponto crítico $F_{4,43}(1\%) = 3.79$, ao nível de significância de 1%, da distribuição $F_{4,43}$ de Snedecor com 4 e 43 graus de liberdade. Então, concluímos que algumas das variáveis independentes em (1.15) explicam a variabilidade do consumo de combustível nos estados americanos.

Tabela 1.3: *Resultados do Ajustamento*

```
The regression equation is
Cons = 375 - 34.5 Taxa - 0.0665 Rend - 0.00240 Rodov + 13.4
Licen

Predictor        Coef       StDev          T        P
Constant        374.7       185.7       2.02    0.050
Taxa           -34.52       12.97      -2.66    0.011
Rend         -0.06653     0.01724      -3.86    0.000
Rodov       -0.002399    0.003394      -0.71    0.483
Licen          13.367       1.927       6.94    0.000

S = 66.38       R-Sq = 67.8%     R-Sq(adj) = 64.8%

Analysis of Variance

Source        DF          SS          MS         F        P
Regression     4      398906       99726     22.63    0.000
Error         43      189461        4406
Total         47      588366
```

1.7 Seleção das Variáveis Explicativas

Depois do ajustamento preliminar de um modelo de regressão, temos interesse em selecionar as variáveis explicativas que podem ser eliminadas do modelo, objetivando obter um modelo parcimonioso para explicar os dados em questão. O teste F da análise de variância permite apenas inferir que algumas das variáveis explicativas são realmente importantes para explicar a variabilidade da variável resposta. Para selecionarmos as variáveis independentes que são significativas, precisamos determinar a distribuição das estimativas dos parâmetros β e σ^2 do modelo normal-linear.

Neste modelo, a estimativa de mínimos quadrados $\hat{\beta}_r$ tem distribuição normal $N(\beta_r, \sigma^2 v_{rr})$, onde v_{rr} é o elemento (r, r) da diagonal da matriz $(X^T X)^{-1}$. Como $\hat{\beta}$ é independente de $\hat{\sigma}^2$ e a distribuição de $\hat{\sigma}^2$ é $(n-p)^{-1}\sigma^2\chi^2_{n-p}$, a estatística teste T_r definida por

$$T_r = \frac{\hat{\beta}_r - \beta_r}{\hat{\sigma}\sqrt{v_{rr}}}, \tag{1.16}$$

tem distribuição t_{n-p} de Student com $n-p$ graus de liberdade. Esta estatística permite testar se a variável explicativa x_r correspondente a β_r deve permanecer no modelo. Na prática, basta dividirmos o valor absoluto de $\hat{\beta}_r$ pelo seu erro padrão, isto é, $\hat{\sigma}\sqrt{v_{rr}}$. Se este quociente for inferior ao valor crítico $t_{n-p}(\alpha)$ da distribuição t_{n-p} de Student com $n-p$ graus de liberdade, a variável independente x_r não é significativa para explicar a variabilidade da resposta e poderá ser eliminada do modelo; caso contrário, x_r é estatisticamente significante para explicar o comportamento da variável resposta. Da Tabela 1.3, verificamos facilmente que a estatística T_r $(= Coef/StDev)$ só não é significativa para a variável independente *Rodov* $(\mid T_r \mid = 0.71 < t_{43}(5\%) = 2.02)$. Assim, podemos reajustar o modelo de regressão (1.15) à variável dependente $Cons$ excluindo a variável *Rodov*, pois a malha rodoviária estadual do estado americano não influi significativamente no consumo de combustível de seus habitantes. Reajustando o modelo de regressão (1.15) sem a variável explicativa *Rodov* obtém-se a equação da primeira regressão descrita na Tabela 1.4. Nesta equação, apenas a estimativa do intercepto (*Constant*) não é significativa, pois sua estatística T_r satisfaz $\mid T_r \mid = 1.95 < t_{44}(5\%) = 2.02$. Assim, reajustou-se um novo modelo de regressão sem o termo constante, obtendo-se a segunda

regressão descrita nesta tabela. Neste novo modelo sem intercepto, contendo apenas as variáveis explicativas $Taxa, Rend$ e $Licen$, verifica-se que a variável $Taxa$ pode ser excluída da regressão, pois $\mid T_r \mid = 1.91 < t_{45}(5\%) = 2.01$. Finalmente, a terceira regressão da Tabela 1.4, mostra que as variáveis independentes $Rend$ e $Licen$ são significativas para explicar a variabilidade do consumo de combustível per-capita por ano nos estados americanos.

A equação ajustada $E(Con) = -0.07035\mathrm{R}end + 15.344Lic$ revela que o consumo de combustível per-capita aumenta (como esperado) com o aumento do percentual da população que está habilitada a dirigir. Por exemplo, um incremento de 10% no percentual de motoristas habilitados provocaria um aumento médio de 153.44 galões no consumo per-capita anual dos habitantes de qualquer estado americano. Entretanto, nesta equação, a variável Rend aparece ajustada com sinal negativo, o que pode parecer contraditório que o consumo per-capita decresça com o aumento da renda. Uma possível explicação para este fato é que as pessoas com rendas muito altas realmente consomem menos combustível, pois procuram usar outros meios de transporte como aviões e trens para percorrer grandes distâncias. Observa-se que a última regressão contempla o maior valor da estatística F entre as regressões ajustadas, no caso $F = 1668.93$ e, então, a média de quadrados explicada pela regressão é cerca de 1669 vezes maior do que a média de quadrados residual.

1.8 Intervalos e Regiões de Confiança

Intervalos de confiança para coeficientes individuais de β ou regiões de confiança para subconjuntos e combinações lineares das componentes de β podem ser obtidos, respectivamente, utilizando os elementos da matriz $(X^TX)^{-1}$. Da estatística pivotal definida em (1.16), podemos construir um intervalo de 100(1-α)% de confiança para o verdadeiro valor β_r a partir de

$$\hat{\beta}_r \mp \hat{\sigma}\sqrt{v_{rr}}t_{n-p}(\alpha/2). \tag{1.17}$$

Os sinais menos e mais correspondem aos limites inferior e superior do intervalo, respectivamente, e as quantidades $\hat{\sigma}\sqrt{v_{rr}}$ são dadas nas Tabelas 1.3 e 1.4 na coluna $StDev$. Se o valor de σ^2 é conhecido, podemos substituir os

quantis $t_{n-p}(\alpha/2)$ da distribuição t_{n-p} de Student com $n-p$ graus de liberdade pelos correspondentes quantis da distribuição normal reduzida.

Tabela 1.4: *Três Modelos de Regressão*

Regression Analysis

```
The regression equation is
Cons = 305 - 29.3 Taxa - 0.0680 Rend + 13.7 Licen

Predictor       Coef       StDev          T        P
Constant       305.5       156.9       1.95    0.058
Taxa          -29.28       10.58      -2.77    0.008
Rend        -0.06796     0.01703      -3.99    0.000
Licen         13.747       1.839       7.47    0.000

S = 66.00       R-Sq = 67.4%     R-Sq(adj) = 65.2%
Analysis of Variance
Source       DF          SS          MS         F        P
Regression    3      396705      132235     30.36    0.000
Error        44      191662        4356
Total        47      588366
```

Regression Analysis

```
The regression equation is
Cons = - 15.2 Taxa - 0.0575 Rend + 16.4 Licen

Predictor       Coef       StDev          T        P
Noconstant
Taxa         -15.172       7.939      -1.91    0.062
Rend        -0.05751     0.01665      -3.45    0.001
Licen         16.410       1.267      12.95    0.000

S = 68.01
Analysis of Variance
Source       DF          SS          MS         F        P
Regression    3    16348097     5449366   1177.99    0.000
Error        45      208170        4626
Total        48    16556267

The regression equation is
Cons = - 0.0703 Rend + 15.3 Licen

Predictor       Coef       StDev          T        P
Noconstant
Rend        -0.07035     0.01567      -4.49    0.000
Licen         15.344       1.170      13.11    0.000

S = 69.95
Analysis of Variance
Source       DF          SS          MS         F        P
Regression    2    16331202     8165601   1668.93    0.000
Error        46      225065        4893
```

Se o objetivo é determinar uma região de 100(1-α)% de confiança para uma combinação linear $c^T\beta$ de parâmetros β, onde c é um vetor especificado de dimensão p, obtém-se de $\text{Var}(c^T\hat{\beta}) = \sigma^2 c^T(X^TX)^{-1}c$ os seguintes limites

$$c^T\hat{\beta} \mp \hat{\sigma} t_{n-p}(\alpha/2)\sqrt{c^T(X^TX)^{-1}c}, \tag{1.18}$$

onde $t_{n-p}(\alpha/2)$ é o quantil $(1-\alpha/2)$ de uma distribuição t_{n-p} de Student com $n-p$ graus de liberdade. Assim, todos os $\beta's$ que satisfizerem a equação (1.18) estarão na região de confiança desejada. Esta equação é uma generalização da equação (1.17) para os limites de confiança de um único parâmetro. Claramente, os limites de confiança dados em (1.18) corresponderão aos limites da média da variável resposta quando c corresponder aos valores das variáveis explicativas do modelo. Por outro lado, se desejarmos uma região de confiança para uma observação y_+ estimada a partir do vetor c contendo os valores das variávies explicativas, os limites dados em (1.18) serão modificados para $c^T\hat{\beta} \mp \hat{\sigma} t_{n-p}(\alpha/2)\left\{1 + c^T(X^TX)^{-1}c\right\}^{1/2}$. Estes intervalos para as observações estimadas são geralmente denominados *intervalos de tolerância.*

Finalmente, podemos obter uma região de confiança para todos os parâmetros em β a partir dos resultados descritos nos itens ii) e iv) da Seção 1.5. Com efeito, a inequação matricial

$$(\beta - \hat{\beta})X^TX(\beta - \hat{\beta}) \leq p\hat{\sigma}^2 F_{p,n-p}(\alpha), \tag{1.19}$$

onde $F_{p,n-p}(\alpha)$ é o quantil da distribuição $F_{p,n-p}$ de Snedecor com graus de liberdade p e $n-p$ cuja área à direita é α, produz uma região conjunta de confiança para todos os parâmetros em β. A inequação (1.19) representa um elipsóide de mesma dimensão p do vetor β de parâmetros. Todos os βs que satisfizerem (1.19) estarão na região de 100(1-α)% de confiança do vetor verdadeiro de parâmetros.

Exemplo 1.4: *Cálculo de intervalos de confiança.*

Inicialmente, fazemos o cálculo dos limites de confiança para os parâmetros da regressão linear simples $E(y) = \mu = \beta_0 + \beta_1(x - \bar{x})$, descrita no Exemplo

1.1. Tem-se,

$$(X^T X)^{-1} = \begin{pmatrix} 1/n & 0 \\ 0 & \frac{1}{\sum_i (x_i - \bar{x})^2} \end{pmatrix},$$

obtendo-se as variâncias das estimativas de β_0 e β_1: $\text{Var}(\hat{\beta}_0) = \sigma^2/n$ e $\text{Var}(\hat{\beta}_1) = \sigma^2/\sum_i (x_i - \bar{x})^2$. Logo, intervalos de 100(1-α)% de confiança para estes parâmetros são dados por $\hat{\beta}_0 \mp \frac{\hat{\sigma}}{\sqrt{n}} t_{n-2}(\alpha/2)$ e $\hat{\beta}_1 \mp \frac{\hat{\sigma}}{\{\sum_i (x_i - \bar{x})^2\}^{1/2}} t_{n-2}(\alpha/2)$. Se desejarmos um intervalo de tolerância para a variável resposta quando a variável explicativa é igual a x_+, obteremos

$$\hat{\beta}_0 + \hat{\beta}_1 (x_+ - \bar{x}) \mp \hat{\sigma} t_{n-p}(\alpha/2) \sqrt{1 + \frac{1}{n} + \frac{(x_+ - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}.$$

Da terceira regressão descrita na Tabela 1.4, calculamos agora os limites de confiança para os coeficientes das variáveis *Rend* e *Licen*. Da fórmula (1.17), obtemos os seguintes intervalos, ao nível de significância de 5% em que $t_{46}(0.025) = 2.01$: para a variável *Rend*, -0.07035 ∓ 0.01567 x $2.01 = (-0.102, -0.039)$ e para a variável *Licen*, 15.344 ∓ 1.170 x $2.01 = (12.922, 17.696)$. Então, podemos dizer que, com 95% de confiança, os coeficientes verdadeiros de *Rend* e *Licen* pertencem aos intervalos $(-0.102, -0.039)$ e $(12.922, 17.696)$, respectivamente.

1.9 Técnicas de Diagnóstico

As técnicas de diagnóstico são usadas para detectar problemas com o ajuste do modelo de regressão. Esses problemas são de três tipos: a) presença de observações mal ajustadas (pontos aberrantes); b) inadequação das suposições iniciais para os erros aleatórios $\epsilon_i's$ e/ou para a estrutura das médias $\mu_i s$; c) presença de observações influentes. Nesta seção desenvolvemos as principais técnicas de diagnóstico na classe dos modelos normais-lineares.

1.9.1 Matriz de projeção

A matriz de projeção H – definida na Seção 1.3 – é muito usada nas técnicas de diagnóstico em regressão. Uma característica de grande importância da matriz H é inerente aos elementos $h_{11}, \ldots, h_{nn}$ da sua diagonal. O elemento h_{ii} mede o quão distante a observação y_i está das demais $n-1$ observações no espaço definido pelas variáveis explicativas do modelo. O elemento h_{ii} só depende dos valores das variáveis explicativas, isto é, da matriz X, e não envolve as observações em y. O elemento h_{ii} representa uma *medida de alavancagem* da i-ésima observação. Se h_{ii} é grande, os valores das variáveis explicativas associados à i-ésima observação são *atípicos*, ou seja, estão distantes do vetor de valores médios das variáveis explicativas. Uma observação com h_{ii} grande poderá ter influência na determinação dos coeficientes da regressão.

Pelo fato de H ser uma matriz simétrica e idempotente, tem-se: a) $\frac{1}{n} \leq h_{ii} \leq 1$; b) $h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$; c) $tr(H) = \sum_i h_{ii} = p$. O elemento h_{ii} mede a influência da i-ésima resposta sobre o seu valor ajustado. Com efeito, se uma observação y_i tem grande alavancagem, o valor de h_{ii} é próximo de um, implicando que a variância do resíduo correspondente r_i é próxima de zero. Logo, o valor médio ajustado $\hat{\mu}_i$ é determinado praticamente pelo valor da observação y_i. Entretanto, como $\mathrm{Var}(\hat{\mu}_i) = \hat{\sigma}^2 h_{ii}$, a variabilidade da média ajustada referente à observação y_i é proporcional ao valor de h_{ii}.

Como $\sum_i h_{ii} = p$, supondo que todas as observações exerçam a mesma influência sobre os valores ajustados, espera-se que h_{ii} esteja próximo de p/n. Convém, então, examinar aquelas observações correspondentes aos maiores valores de h_{ii}. Alguns autores sugerem $h_{ii} \geq 2p/n$ como um indicador de pontos de alta alavancagem que requerem uma investigação adicional. Esta regra funciona bem na prática embora, em geral, irá detectar muitas observações de grande alavancagem. Assim, outras medidas de diagnóstico serão sempre necessárias para confirmar esse primeiro diagnóstico.

1.9.2 Resíduos

O resíduo para a i-ésima observação é definido como função $r_i = r(y_i, \hat{\mu}_i)$ que mede a discrepância entre o valor observado y_i e o valor ajustado $\hat{\mu}_i$. Observações bem (mal) ajustadas devem apresentar pequenos (grandes) resíduos. O sinal de r_i indica a direção dessa discrepância. O resíduo ordinário é definido por $r_i = y_i - \hat{\mu}_i$ mas, não é muito informativo, pois sua variância não é constante. Com efeito, r_i tem distribuição normal de média zero e variância $\text{Var}(r_i) = \sigma^2(1 - h_{ii})$ (vide Seção 1.4, item e)). Assim, observações com grande alavancagem têm resíduos de menor variabilidade do que observações de pequena alavancagem. Para comparar os resíduos devemos expressá-los em forma padronizada. Define-se, então, *resíduos padronizados* por

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}. \tag{1.20}$$

A vantagem dos resíduos padronizados é que se o modelo (1.1) está correto, todos os resíduos têm a mesma variância, mesmo não sendo independentes. As observações cujos valores absolutos dos resíduos padronizados são maiores do que 2 podem ser consideradas mal-ajustadas (*pontos aberrantes).* Estes resíduos são, também, apropriados para verificar a normalidade dos erros e a homogeneidade das variâncias. Como r_i não é independente de $\hat{\sigma}^2$, r_i^* não tem uma distribuição t de Student como deveria se esperar. Pode-se mostrar que $r_i^{*2}/(n-p)$ tem uma distribuição beta com parâmetros $1/2$ e $(n-p)/2$ e que $\text{E}(r_i^*) = 0$, $\text{Var}(r_i^*) = 1$ e $\text{Cov}(r_i^*,\ r_j^*) = -h_{ij}/\left\{(1 - h_{ii})(1 - h_{jj})\right\}^{1/2}$ para $i \neq j$.

Para contornar a dependência entre r_i e $\hat{\sigma}^2$, podemos estimar σ^2 eliminando-se a observação y_i do modelo de regressão. Assim, seja $\hat{\beta}_{(i)}$ o EMQ de β obtido quando eliminamos a observação y_i, $\hat{\mu}_{(i)} = x_i^T\hat{\beta}_{(i)}$ a média preditiva correspondente, e $\hat{\sigma}^2_{(i)}$ o estimador não-viesado da variância supondo que a observação y_i não está presente no ajustamento do modelo. Como y_i e $\hat{\mu}_{(i)}$ são independentes, a variância da diferença $y_i - \hat{\mu}_{(i)}$ é dada por

$$\text{Var}(y_i - \hat{\mu}_{(i)}) = \sigma^2\left\{1 + x_i^T(X_{(i)}^T X_{(i)})^{-1}x_i\right\},$$

onde $X_{(i)}$ representa a matriz modelo sem a linha correspondente à observação y_i. Então, define-se o *resíduo Studentizado* por

$$t_i = \frac{y_i - \hat{\mu}_{(i)}}{\hat{\sigma}_{(i)} \left\{ 1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i \right\}^{1/2}}. \tag{1.21}$$

O resíduo Studentizado tem distribuição t de Student com $n-p-1$ graus de liberdade. A desvantagem no cálculo do resíduo Studentizado pela expressão (1.21) é que teremos que ajustar n regressões adicionais (uma para cada observação retirada do modelo) para calcularmos as estimativas $\hat{\sigma}^2_{(i)}$ para $i = 1, \ldots, n$. Felizmente, podemos calcular as estimativas $\hat{\sigma}^2_{(i)}$ para $i = 1, \ldots, n$, considerando apenas a regressão original com todas as n observações, através da equação

$$\hat{\sigma}^2_{(i)} = \frac{(n-p)\hat{\sigma}^2 - r_i^2/(1-h_{ii})}{(n-p-1)}. \tag{1.22}$$

O EMQ $\hat{\beta}_{(i)}$ decorrente da eliminação da observação y_i pode ser obtido, também, da regressão com todas as observações, usando

$$\hat{\beta}_{(i)} - \hat{\beta} = -\frac{r_i}{(1-h_{ii})} (X^T X)^{-1} x_i. \tag{1.23}$$

Uma expressão bem mais simples para o resíduo Studentizado decorre da equação (1.22) e das relações $x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i = h_{ii}/(1-h_{ii})$ e $\hat{\mu}_{(i)} = x_i^T \hat{\beta}_{(i)} = \hat{\mu}_i - \frac{h_{ii} r_i}{1-h_{ii}}$. Assim, obtemos

$$t_i = \frac{y_i - \hat{\mu}_i}{\hat{\sigma}_{(i)} \sqrt{(1-h_{ii})}} = \frac{\hat{\sigma} r_i^*}{\hat{\sigma}_{(i)}}.$$

Substituindo a expressão (1.22) na equação anterior, obtém-se os resíduos Studentizados como uma função monotônica (embora não-linear) dos resíduos

padronizados, ou seja,

$$t_i = \sqrt{\frac{n-p-1}{n-p-r_i^{*2}}} r_i^*. \tag{1.24}$$

Os resíduos Studentizados definidos na equação (1.24) têm a grande vantagem de serem obtidos da regressão original com todas as observações. Estes resíduos podem ser usados para testar se há diferenças significativas entre os valores ajustados obtidos *com* e *sem* a i-ésima observação.

1.9.3 Influência

No modelo de regressão é fundamental conhecer o grau de dependência entre o modelo ajustado e o vetor de observações y. Será preocupante se pequenas perturbações nestas observações produzirem mudanças bruscas nas estimativas dos parâmetros do modelo. Entretanto, se tais observações não alterarem os principais resultados do ajustamento, pode-se confiar mais no modelo proposto, mesmo desconhecendo o verdadeiro processo que descreve o fenômeno em estudo. As técnicas mais conhecidas para detectar esse tipo de *influência* são baseadas na exclusão de uma única observação e procuram medir o impacto dessa perturbação nas estimativas dos parâmetros. Apresentamos aqui algumas medidas de diagnóstico mais usadas na avaliação do grau de dependência entre $\hat{\beta}$ e cada uma das observações.

Inicialmente, considera-se a distância de Cook usada para detectar observações influentes. Para a i-ésima observação, a distância de Cook combina o resíduo padronizado r_i^* com a medida de alavancagem h_{ii}, sendo portanto uma medida global de quão atípica esta i-ésima observação se apresenta no ajustamento do modelo. Assim, uma medida de influência da retirada da i -ésima observação sobre as estimativas dos parâmetros do modelo é dada pela *estatística de Cook* (1977)

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X (\hat{\beta}_{(i)} - \hat{\beta})}{p\hat{\sigma}^2}. \tag{1.25}$$

A estatística D_i representa uma soma ponderada dos desvios entre as

estimativas baseadas em $\hat{\beta}$ e $\hat{\beta}_{(i)}$ em que os pesos indicam a precisão das estimativas em $\hat{\beta}$. Quanto mais precisas forem estas estimativas, maiores pesos serão alocados à diferença entre $\hat{\beta}$ e $\hat{\beta}_{(i)}$. Assim, D_i pode ser vista como uma medida da distância entre os coeficientes calculados com e sem a i-ésima observação. Esta interpretação sugere usar a distribuição F de Snedecor para decidir se a estatística de Cook é grande ou não. Valores grandes em (1.25) indicam observações que influenciam bastante as estimativas dos parâmetros do modelo. A equação (1.25) lembra a expressão (1.19), que fornece uma região de confiança simultânea para todos os parâmetros da regressão. Usando (1.23) em (1.25) pode-se obter uma expressão para D_i mais fácil de ser interpretada

$$D_i = \frac{h_{ii}}{p(1-h_{ii})} r_i^{*2}. \tag{1.26}$$

Logo, D_i será grande quando o i-ésimo resíduo padronizado for aberrante (r_i^* grande) e/ou quando a medida de alavancagem h_{ii} for próxima de um. Como visto anteriormente, r_i^{*2} mede a discrepância da i-ésima observação e h_{ii}, ou equivalentemente, o quociente $h_{ii}/(1-h_{ii})$ mede a discrepância da i-ésima linha da matriz modelo X. O efeito combinado desses indicadores de discrepância produz então a medida de influência de Cook no modelo de regressão.

A medida D_i poderá não ser adequada quando o resíduo pradronizado r_i^* for grande e h_{ii} for próximo de zero. Neste caso, $\hat{\sigma}^2$ pode estar inflacionado, e não ocorrendo nenhuma compensação por parte de h_{ii}, D_i pode ser pequeno. As observações serão consideradas influentes quando $D_i \geq F_{p,n-p}(0.50)$ e recomenda-se examinar as consequências da retirada dessas observações no ajustamento do modelo. Como para a maioria das distribuições F o quantil de 50% é próximo de um, sugere-se na prática que se o maior valor de D_i for muito inferior a um, então a eliminação de qualquer observação do modelo não irá alterar muito as estimativas dos parâmetros. Entretanto, para investigar mais detalhadamente a influência das observações com maiores valores de D_i, o analista terá que eliminar estas observações e re-computar as estimativas dos parâmetros.

Quando a i-ésima observação for detectada como um ponto aberrante

(baseando-se em r_i^*) ou como um ponto de alta alavancagem (baseando-se em h_{ii}), usa-se o valor de D_i para checar se esta observação é influente, ou seja, se quando for removida do vetor y causará mudanças apreciáveis nas estimativas de β.

Uma medida alternativa à estatística de Cook para detectar observações influentes foi proposta por Belsley et al. (1980). Esta medida, conhecida como DFFITS, é função do resíduo Studentizado t_i dado em (1.24), e da medida de alavancagem h_{ii}, sendo expressa por

$$DFFITS_i = t_i \left\{ \frac{h_{ii}}{p(1-h_{ii})} \right\}^{1/2}. \tag{1.27}$$

No caso da estatística $DFFITS_i$, os pontos influentes são aqueles em que $DFFITS_i \geq 2\left\{p/(n-p)\right\}^{1/2}$. Os comentários feitos para a estatística D_i permanecem válidos para a estatística (1.27).

Geralmente, examina-se as estatísticas D_i e $DFFITS_i$ graficamente, dando atenção àquelas observações cujas medidas têm maiores valores.

1.9.4 Técnicas gráficas

De uma forma geral, os problemas de diagnóstico a), b) e c) mencionados no início da Seção 1.9, podem ser detectados, respectivamente, através das seguintes técnicas gráficas:

a) um gráfico dos resíduos padronizados r_i^* dados em (1.20) versus a ordem das observações para detectar as observações aberrantes;

b) um gráfico dos resíduos padronizados r_i^* versus os valores ajustados $\hat{\mu}_i$ e um gráfico de probabilidade dos resíduos padronizados ordenados versus os quantis da distribuição normal reduzida. Estes quantis são definidos por $\Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$, onde $\Phi^{-1}(.)$ é a função de distribuição acumulada da normal reduzida. No primeiro gráfico dos resíduos padronizados, os pontos devem estar aleatoriamente distribuídos entre as duas retas $y = -2$ e $y = 2$ paralelas ao eixo horizontal, sem exibir uma forma definida. Se neste gráfico os pontos exibirem algum padrão, isto pode ser indicativo

de heterocedasticidade da variância dos erros ou da não-linearidade dos efeitos das variáveis explicativas nas médias das observações. No segundo gráfico, se os pontos ficarem praticamente dispostos sobre uma reta, as observações podem ser consideradas como tendo, aproximadamente, distribuição normal;

c) gráficos de h_{ii}, D_i e $DFFITS_i$ versus a ordem das observações para detectar as observações influentes.

Exemplo 1.5: *Continuação da Regressão Linear Múltipla.*

Aplicamos aqui as técnicas gráficas e de diagnóstico à terceira regressão ajustada da Tabela 1.4, ou seja, $E(Con) = -0.07035\mathrm{R}end + 15.344Lic$. Na Figura 1.1 mostramos, sucessivamente, os gráficos dos resíduos padronizados r_i^* versus a ordem das observações e versus os valores ajustados $\hat{\mu}_i$ e o gráfico de probabilidade dos resíduos padronizados ordenados versus os quantis da normal reduzida. Do primeiro destes gráficos, concluímos que duas observações (aquelas 18 e 40) têm resíduos em valor absoluto maiores do que dois, indicando que estas são observações aberrantes. O segundo gráfico dos resíduos padronizados versus os valores ajustados não apresenta nenhuma forma definida e, portanto, a variância das observações pode ser considerada constante e o modelo linear nas variáveis explicativas Rend e Lic mostra-se adequado. No terceiro gráfico da Figura 1.1, a hipótese de normalidade para o consumo de combustível é aceita pois o gráfico revela-se praticamente linear.

Na Figura 1.2 apresentamos sucessivamente gráficos das medidas de alavancagem h_{ii} e de influência D_i e $DFFITS_i$ versus a ordem das observações para o modelo de regressão em pauta. Do gráfico de h_{ii} concluímos que as

Figura 1.1: *Gráficos dos Resíduos*

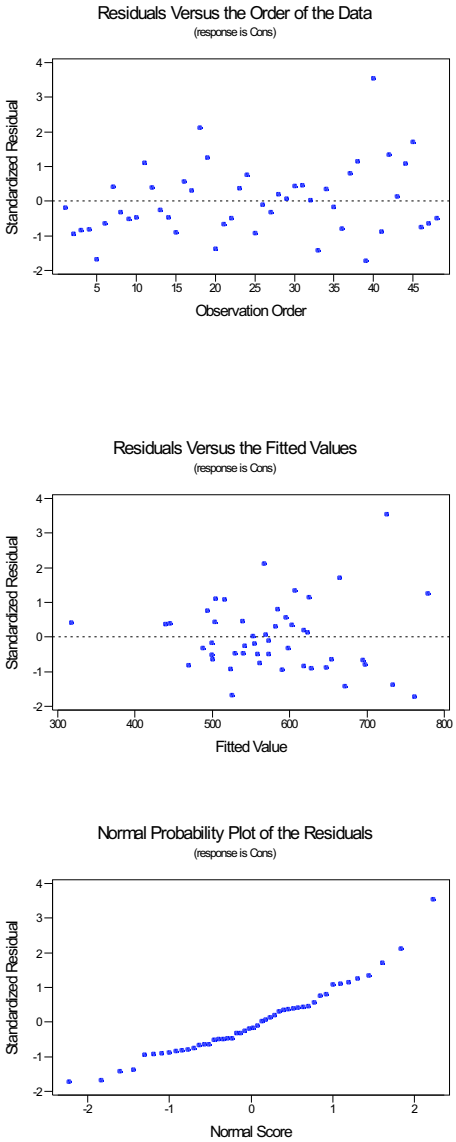
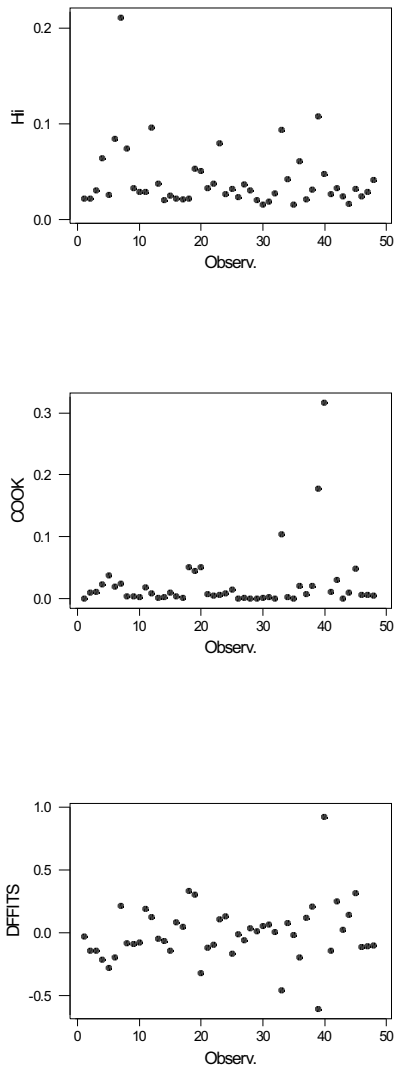


Figura 1.2: *Gráficos das Medidas de Diagnóstico*



observações $6, 7, 12, 33$ e 39 são pontos de alta alavancagem, pois seus h_{ii} são superiores ao valor crítico $2p/n = 0.083$. Pelo gráfico da estatística D_i de Cook, concluímos que as observações $33, 39$ e 40 são influentes, pois os valores de D_i são bem superiores aos demais. Note-se que a observação 40 tinha sido detectada como um ponto aberrante e as observações 33 e 39 foram detectadas como pontos de grande alavancagem. Pelo teste da estatística $DFFITS$, a conclusão é a mesma: as observações $33, 39$ e 40 são influentes, pois seus valores são superiores ao valor crítico $2\left\{p/(n-p)\right\}^{1/2} = 0.4170$.

1.10 Estimação de Máxima Verossimilhança

Apresentamos aqui o *método de estimação de máxima verossimilhança* para estimar o vetor de parâmetros β no modelo clássico de regressão (1.1). Para aplicação deste método, necessitamos supor alguma distribuição de probabilidade para o vetor y. Assim, consideramos que y tem média $\mu = X\beta$ e que suas componentes são independentes e normalmente distribuídas com mesma variância σ^2. Podemos, então, considerar que $y \sim N(X\beta, \sigma^2 I)$. A estimação de β e σ^2 por máxima verossimilhança consiste em maximizar a função de verossimilhança em relação ao vetor de parâmetros β e ao escalar σ^2. A função de verossimilhança para estes parâmetros é dada por

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right\}, \tag{1.28}$$

onde $\mu_i = x_i^T\beta$ é a média de y_i.

Maximizar a verossimilhança equivale a maximizar o logaritmo desta função $l(\beta, \sigma^2) = \log L(\beta, \sigma^2)$ que pode ser escrito na forma

$$l(\beta, \sigma^2) = -\frac{1}{2}\left\{n \log \sigma^2 + \frac{1}{\sigma^2}(y - X\beta)^T(y - X\beta)\right\}.$$

Qualquer que seja o valor de σ^2, a estimativa de máxima verossimilhança (EMV) de β minimiza a soma de quadrados acima, de modo que a EMV de β

quando os erros são normalmente distribuídos iguala à estimativa de mínimos quadrados (EMQ) $\hat{\beta} = (X^T X)^{-1} X^T y$. No modelo de regressão, a estimativa de máxima verossimilhança só coincide com a estimativa de mínimos quadrados segundo normalidade. Diferenciando a expressão acima em relação a σ^2 e igualando a zero, obtém-se a EMV de σ^2 como

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n}. \tag{1.29}$$

Note-se que a EMV de σ^2 dada em (1.29) difere da estimativa (1.11) pelo denominador. A EMV é uma estimativa viesada de σ^2, enquanto aquela proposta em (1.11) não tem viés.

A matriz de informação para β e σ^2 é calculada diferenciando a log-verossimilhança. As segundas derivadas da log-verossimilhança $l = l(\beta, \sigma^2)$ são dadas por

$$\frac{\partial^2 l}{\partial \beta_r \partial \beta_s} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ir} x_{is}, \quad \frac{\partial^2 l}{\partial \beta_r \partial \sigma^2} = \frac{1}{\sigma^4} \sum_{i=1}^{n} x_{ir} (y_i - x_i^T \beta)$$

e

$$\frac{\partial^2 l}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

Assim, os elementos da matriz de informação $I(\beta, \sigma^2)$ são calculados por

$$E\left(-\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ir} x_{is}, E\left(-\frac{\partial^2 l}{\partial \beta_r \partial \sigma^2}\right) = 0$$

e

$$E\left\{-\frac{\partial^2 l}{\partial (\sigma^2)^2}\right\} = \frac{n}{2\sigma^4}.$$

Logo, a matriz de informação para β e σ^2 pode ser escrita como

$$I(\beta, \sigma^2) = \begin{pmatrix} \sigma^{-2} X^T X & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}.$$

A inversa da matriz de informação representa a estrutura de covariância assintótica das estimativas de máxima verossimilhança. A inversa da matriz $I(\beta, \sigma^2)$ é simplesmente

$$I(\beta, \sigma^2)^{-1} = \begin{pmatrix} \sigma^2(X^T X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}.$$

No caso, o resultado assintótico é um resultado exato e a matriz $I(\beta, \sigma^2)^{-1}$ iguala à estrutura de covariância exata das estimativas de máxima verossimilhança de β e σ^2, ou seja, $Cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$, como visto em (1.10), e $\mathrm{Var}(\hat{\sigma}^2) = 2\sigma^4/n$.

Da teoria de verossimilhança, concluímos ainda que as estimativas $\hat{\beta}$ e $\hat{\sigma}^2$ têm distribuições assintóticas normais p-variada $N_p(\beta, \sigma^2(X^T X)^{-1})$ e univariada $N(\sigma^2, 2\sigma^4/n)$, respectivamente. No caso, o primeiro resultado é exato, e já tínhamos mostrado na Seção 1.5 ii) que o EMQ (idêntico ao EMV) tem distribuição normal p-variada de média β e estrutura de covariância $\sigma^2(X^T X)^{-1}$.

A estrutura bloco-diagonal da matriz $I(\beta, \sigma^2)^{-1}$ implica que as EMV $\hat{\beta}$ e $\hat{\sigma}^2$ são assintoticamente independentes. Nós tínhamos mostrado na Seção 1.5 iv) um resultado mais forte: que as estimativas $\hat{\beta}$ e $\hat{\sigma}^2$ são independentes para todo valor de n.

Mostraremos agora que as estimativas $\hat{\beta}$ e $\hat{\sigma}^2$ são estatísticas suficientes minimais para os parâmetros β e σ^2. Da equação (1.12) temos a decomposição

$$(y - X\beta)^T (y - X\beta) = SQE(\hat{\beta}) + (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta).$$

Logo, a verossimilhança (1.28) pode ser escrita como

$$L(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{ -\frac{SQE(\hat{\beta})}{2\sigma^2} - \frac{1}{2\sigma^2}(\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \right\}.$$

O critério da fatorização implica que $\hat{\beta}$ e $SQE(\hat{\beta})$ são estatísticas suficientes para os parâmetros β e σ^2, e é evidente que estas estatísticas são sufi-

cientes minimais. Embora n e X sejam necessários para calcular a verossimilhança, estas quantidades não são aleatórias e, portanto, não são partes integrantes das estatísticas suficientes.

1.11 Exercícios

1. Ajusta-se um modelo de regressão a um conjunto de dados. Mostre que:
   (i) $\sum_{i=1}^{n} Var(\hat{\mu}_i) = p\sigma^2$;
   (ii) $SQE = \hat{\mu}^T H^3 y, \quad \text{onde} \quad H = X(X^T X)^{-1} X^T$.
2. Demonstre que R^2 é igual ao quadrado da correlação entre os vetores y e $\hat{\mu}$.
3. Considere as regressões de y sobre x para os dados seguintes, especificadas por $E(y) = \beta_0 x$ e $E(y) = \beta_1 x + \beta_2 x^2$. Demonstre que $\hat{\beta}_0 = 3.077$, $\hat{\beta}_1 = 2.406$ e $\hat{\beta}_2 = 0.138$. Qual desses modelos seria o preferido?

y	5	7	7	10	16	20
x	1	2	3	4	5	6

4. Utilizando o teorema de Fisher-Cochran mostrar que as somas de quadrados $\hat{\beta}^T X^T y$ e $y^T y - \hat{\beta}^T X^T y$ são independentes e têm distribuição χ^2 com p e $(n-p)$ graus de liberdade, respectivamente.
5. O conjunto de dados abaixo corresponde à produção anual de milho (y) em *kg/ha* e a quantidade de chuva x em mm, durante 7 anos em determinado município.

Ano	1	2	3	4	5	6	7
y	1295	1304	1300	1428	1456	1603	1535
x	1094.10	1180.15	1137.30	1714.80	1289.50	1401.50	1640.40

   (i) Ajustar o modelo $y = \beta_0 + \beta_1 x + \varepsilon$ aos dados e obter $\hat{\beta}_0$, $\hat{\beta}_1$, os correspondentes desvios padrões, $\hat{\sigma}^2$ e R^2, e a tabela ANOVA;

(ii) Calcular os resíduos de Pearson $p_i = (y_i - \hat{\mu}_i)/s$ para cada observação. Verificar se há pontos aberrantes. Fazer os gráficos de p_i contra $\hat{\mu}_i$ e p_i contra i. Nota-se alguma tendência sistemática nesses gráficos?

(iii) Sugerir um novo modelo com base nos gráficos de (ii). Obter as estimativas de mínimos quadrados. Comparar $\hat{\sigma}^2$ e o R desse novo modelo com aqueles do modelo ajustado em (i);

(iv) Suponha que num determinado ano choveu 1250 mm. Calcular um intervalo de confiança de 95% para a produção de milho nesse ano, utilizando, respectivamente, os modelos ajustados em (i) e (ii). Comparar os intervalos obtidos.

6. Os dados a seguir correspondem à área de um pasto em função do tempo de crescimento. Ajustar um modelo de regressão aos mesmos.

AREA	8.93	10.80	18.59	22.33	39.35	56.11	61.72	64.62
TEMPO	9.00	14.00	21.00	28.00	42.99	57.00	63.00	70.00
AREA	67.00							
TEMPO	79.00							

7. Em 9 municípios foram observadas as seguintes variáveis: y-consumo de um determinado produto, x_1-urbanização relativa, x_2-nível educacional e x_3-percentual de jovens.
Os dados são os seguintes:

Munic.	1	2	3	4	5	6	7	8	9
x_1	41.2	48.6	42.6	39.0	34.7	44.5	39.1	40.1	45.9
x_2	41.2	10.6	10.6	10.4	9.3	10.8	10.7	10.0	12.0
x_3	31.9	13.2	28.7	26.5	8.5	24.3	18.6	20.4	15.2
y	167.1	174.4	162.0	140.8	179.8	163.7	174.5	185.7	160.6

(i) Ajustar o modelo irrestrito $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ aos dados

e esse mesmo modelo restrito à $C\beta = 0$, onde

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Formar a tabela ANOVA e testar as hipóteses $H : \beta_1 = \beta_2 = \beta_3 = 0$, $H' : C\beta = 0$ e $H'' : \beta_2 = 0$ dado $C\beta = 0$. Utilize $\alpha = 0.01$;

(ii) Para o ajuste do modelo $y = \beta_0 + \beta_2 x_2 + \varepsilon$ aos dados, calcular R^2 e $\hat{\sigma}^2$ e comparar com os valores obtidos impondo-se o modelo irrestrito corrente;

(iii) Fazer uma análise de diagnóstico completo para o ajuste de (ii).

8. Suponha um modelo de regressão $y = X\beta + \varepsilon$ contendo β_0 como intercepto e 1 o vetor $n \times 1$ de uns correspondente. Mostre que $1^T H 1 = n$, onde H é a matriz de projeção.

9. Suponha que tenhamos um modelo de regressão $y = X\beta + \varepsilon$, onde os parâmetros β estão sujeitos a restrições de igualdade do tipo $C\beta = d$. Mostre que a estimativa de mínimos quadrados (EMQ) de β é dada por

$$\tilde{\beta} = \hat{\beta} + (X^T X)^{-1} C^T (C(X^T X)^{-1} C)^{-1} (d - C\hat{\beta}),$$

onde $\hat{\beta}$ é o EMQ usual.

10. Demonstrar a desigualdade (1.19).

Capítulo 2

Modelos Lineares Generalizados

2.1 Introdução

Os Modelos Lineares Generalizados (MLGs), também denominados modelos exponenciais lineares, foram desenvolvidos por Nelder e Wedderburn (1972). Esta classe de modelos é baseada na família exponencial uniparamétrica, que possui propriedades interessantes para estimação, testes de hipóteses e outros problemas de inferência. O MLG é definido por uma distribuição de probabilidade, membro da família exponencial 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.

Várias distribuições de probabilidade importantes (discretas e contínuas) como normal, gama, Poisson, binomial, normal inversa (ou Gaussiana inversa), etc., são membros da família exponencial e os seguintes modelos são casos especiais dos MLGs:

- Modelo normal linear;
- Modelos log-lineares aplicados à análise de tabelas de contingência;

- Modelo logístico para tabelas multidimensionais de proporções;
- Modelo probit para estudo de proporções;
- Modelos estruturais com erro gama;

e outros modelos familiares. O modelo normal linear foi descrito no Capítulo 1. Os demais modelos serão descritos aqui e em capítulos posteriores.

Entretanto, os MLGs não englobam dados correlacionados e distribuições fora da família exponencial. Porém, alguns casos especiais de regressão que não são MLGs genuínos podem ser ajustados através de algoritmos iterativos, mediante pequenas alterações (Cordeiro e Paula, 1992).

2.2 Um Esboço Sobre os MLGs

2.2.1 Formulação do modelo

A formulação de um MLG compreende a escolha de uma distribuição de probabilidade para a variável resposta, das variáveis quantitativas e/ou qualitativas para representar a estrutura linear do modelo e de uma função de ligação. Para a melhor escolha da referida distribuição de probabilidade é aconselhável examinar os dados para observar algumas características, tais como: assimetria, natureza discreta ou contínua, intervalo de variação, etc. É importante salientar que os termos que compõem a estrutura linear do modelo podem ser de natureza contínua, qualitativa ou mista, e devem dar uma contribuição significativa na explicação da variável resposta.

Uma importante característica dos MLGs é a suposição de independência, ou pelo menos de não-correlação, entre as observações. Como consequência disso, dados exibindo autocorrelação no tempo, por exemplo, não devem fazer parte do contexto dos MLGs. Uma outra característica destes modelos está na distribuição da variável resposta. Considera-se uma distribuição única que deve pertencer à família exponencial. Assim, estão excluídos os modelos de análise de experimentos que têm mais de uma componente de erro explícita.

2.3 As Componentes de um MLG

De uma forma geral, a estrutura de um MLG é formada por três partes: uma *componente aleatória* composta de uma variável aleatória Y com n observações independentes, um vetor de médias μ e uma distribuição pertencente à família exponencial; uma *componente sistemática* composta por variáveis explicativas $x_1, \ldots, x_p$ tais que produzem um preditor linear η; e uma função monotônica diferenciável, conhecida como *função de ligação*, que relaciona estas duas componentes.

2.3.1 Componente aleatória

Seja um vetor de observações $y = (y_1, \ldots, y_n)^T$ referente às realizações das variáveis aleatórias $Y = (Y_1, \ldots, Y_n)^T$, independentes e identicamente distribuídas, com médias $\mu = (\mu_1, \ldots, \mu_n)^T$. A parte aleatória de um MLG supõe que cada componente de Y segue uma distribuição da família exponencial definida por

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{[y\theta - b(\theta)]}{a(\phi)} + c(y, \phi)\right\}, \tag{2.1}$$

onde $a(\cdot), b(\cdot)$ e $c(\cdot)$ são funções conhecidas; $\phi > 0$ é denominado *parâmetro de dispersão* e θ é denominado *parâmetro canônico* que caracteriza a distribuição em (2.1). Se ϕ é conhecido, a equação (2.1) representa a família exponencial uniparamétrica indexada por θ.

Assim, para a distribuição normal, temos

$$f_Y(y; \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

$$= \exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\},$$

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

Escrevendo a log-verossimilhança para uma única observação como $l =$

$l(\theta, \phi; y) = \log f_Y(y; \theta, \phi)$ temos uma função de θ e ϕ para um dado y. Assim, a média e a variância de Y podem ser calculadas facilmente por meio das seguintes relações

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \tag{2.2}$$

e

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) + E\left(\frac{\partial l}{\partial \theta}\right)^2 = 0. \tag{2.3}$$

Temos, a partir de (2.1), que

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

Logo,

$$\frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \tag{2.4}$$

e

$$\frac{\partial^2 l}{\partial \theta^2} = -\frac{b''(\theta)}{a\left(\phi\right)}. \tag{2.5}$$

Então, a partir de (2.2) e (2.4), temos $E\left(\frac{\partial l}{\partial \theta}\right) = \frac{\mu - b'(\theta)}{a(\phi)} = 0$ de modo que

$$E(Y) = \mu = b'(\theta). \tag{2.6}$$

Da equação (2.6) podemos obter, univocamente, o parâmetro canônico θ como função da média μ.

Da mesma forma, a partir de (2.3), (2.4) e (2.5), obtemos

$$-\frac{b''(\theta)}{a\left(\phi\right)} + \frac{\mathrm{Var}(Y)}{a\left(\phi\right)^2} = 0.$$

Logo,

$$\mathrm{Var}(Y) = a(\phi) b''(\theta). \tag{2.7}$$

Com isso, podemos dizer que a variância de Y é o produto de duas funções: (i) $b''(\theta)$, que depende apenas do parâmetro canônico e, por conseguinte, da média, sendo chamada de *função de variância* $V = V(\mu)$ e (ii) $a(\phi)$, que só depende de ϕ. A *função de variância* expressa como função de μ é reescrita da seguinte forma

$$V(\mu) = b''(\theta) = \frac{d\mu}{d\theta}. \tag{2.8}$$

A função $a(\phi)$ é geralmente expressa por $a(\phi) = \frac{\phi}{\lambda}$, onde ϕ (também denotado por σ^2) é um parâmetro de dispersão constante para todas as observações e λ é um peso a priori conhecido, que pode variar com as observações.

Apresentamos na Tabela 2.1 as distribuições mais importantes sob a forma (2.1) e algumas de suas principais características. Estas distribuições serão estudadas mais adiante, ou seja, normal $N(\mu, \sigma^2)$, Poisson $P(\mu)$ de média μ, binomial $B(m, \mu)$ com índice m e probabilidade de sucesso μ, gama $G(\mu, \nu)$ com média μ e parâmetro de forma ν e normal inversa $N^-(\mu, \phi)$ com média μ e parâmetro de dispersão ϕ.

Tabela 2.1: *Características de algumas distribuições da família exponencial*

Modelo	$a(\phi)$	$b(\theta)$	$c(y, \phi)$	$\mu(\theta)$	$V(\mu)$
$N(\mu, \sigma^2)$	σ^2	$\frac{\theta^2}{2}$	$-\frac{y^2}{2\phi}$ $-\{\log(2\pi\phi)\}/2$	θ	1
$P(\mu)$	1	$\exp(\theta)$	$-\log y!$	$\exp(\theta)$	μ
$\frac{B(m,\mu)}{m}$	$\frac{1}{m}$	$\log(1+e^\theta)$	$\log\binom{m}{my}$	$\frac{e^\theta}{(1+e^\theta)}$	$\mu(1-\mu)$
$G(\mu, \nu)$	ν^{-1}	$-\log(-\theta)$	$\nu\log(\nu y) - \log y$ $-\log\Gamma(\nu)$	$-\frac{1}{\theta}$	μ^2
$N^-(\mu, \phi)$	ϕ	$-(-2\theta)^{\frac{1}{2}}$	$-\frac{1}{2\phi y}$ $-\{\log(2\pi\phi y^3)\}/2$	$(-2\theta)^{-\frac{1}{2}}$	μ^3

2.3.2 A componente sistemática e a função de ligação

Inicialmente, foi dito que a função de ligação relaciona o preditor linear η à média μ do vetor de dados y. Considere, então, a estrutura linear de um modelo de regressão

$$\eta = X\beta,$$

onde $\eta = (\eta_1, \ldots, \eta_n)^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$ e X é uma matriz modelo $n{\times}p(p{<}n)$ conhecida de posto p. A função linear η dos parâmetros desconhecidos β é chamada de *preditor linear*. Além disso, outra característica da *componente sistemática* de um MLG é que a média μ do vetor y é expressa por uma função conhecida (monótona e diferenciável) de η,

$$\mu_i = g^{-1}(\eta_i), \qquad i = 1, \ldots, n$$

denominando-se $g(\cdot)$ *função de ligação*.

No modelo normal linear a média e o preditor linear são idênticos, dado que η e μ podem assumir qualquer valor na reta real $(-\infty, +\infty)$; logo, uma ligação do tipo identidade ($\eta = \mu$) é plausível para modelar dados normais. Se Y tem distribuição de Poisson, com $\mu > 0$, a função de ligação adequada é a logaritmica ($\eta = \log \mu$), pois esta tem o domínio positivo e o contradomínio na reta real. Entretanto, para modelos que assumem a distribuição binomial, onde $0 < \mu < 1$, existe a restrição de que o domínio da função de ligação esteja no intervalo (0,1), enquanto seu contradomínio é o intervalo $(-\infty, +\infty)$. As três principais funções que garantem esta restrição são:

1. logit (ou logística)

$$\eta = \log\{\mu/(1-\mu)\};$$

2. probit

$$\eta = \Phi^{-1}(\mu),$$

onde $\Phi^{-1}(\cdot)$ é a função de distribuição acumulada da normal reduzida;

3. complemento log-log

$$\eta = \log\{-\log(1-\mu)\}.$$

Finalizando, pode-se dizer que a palavra "generalizado" no MLG significa

uma distribuição mais ampla do que a normal para a variável resposta e uma função não-linear relacionando a média desta variável resposta à parte determinística do modelo.

2.3.3 Estatísticas suficientes e ligações canônicas

Cada distribuição citada na Tabela 2.1 tem uma função de ligação especial que está associada ao preditor linear $\eta = \sum_{r=1}^{p} \beta_r x_r$ e define uma estatística suficiente com a mesma dimensão de β. Estas ligações são chamadas canônicas e ocorrem quando $\theta = \eta$, onde θ é o parâmetro canônico definido em (2.1) e dado na Tabela 2.1 como argumento para a média μ. As ligações canônicas para as distribuições citadas na referida tabela são:

- normal $\eta = \mu$;
- Poisson $\eta = \log \mu$;
- binomial $\eta = \log\{\pi/(1-\pi)\}$;
- gama $\eta = \mu^{-1}$;
- normal inversa $\eta = \mu^{-2}$.

Pode-se mostrar que a estatística suficiente para o vetor de parâmetros β, supondo no modelo que a ligação é canônica, iguala $X^T y$ (em notação vetorial). Os MLGs com ligações canônicas são denominados de *modelos canônicos*.

2.3.4 A matriz modelo

A matriz modelo X é definida a partir de variáveis explicativas que podem ser contínuas, fatores qualitativos e combinações destes (McCullagh e Nelder 1989, Cap. 3).

- Variáveis Contínuas

Exemplos de variáveis contínuas são: peso, área, tempo, comprimento, etc. Cada variável contínua, ou covariável, tem uma representação algébrica

e assume uma forma no modelo. Neste caso, as respectivas representações são αX e X.

• Variáveis Qualitativas

Estas variáveis, que também são denominadas de fatores, possuem um conjunto limitado de valores conhecidos como níveis. Os níveis podem ser codificados pelos números inteiros $1, 2, \ldots, k$. O modelo $\eta = \alpha_i$ $(i = 1, \ldots, k)$ representa um fator A de k níveis. Sua forma no modelo é simplesmente A.

Para ajustar um modelo que possui fatores é necessário utilizar variáveis indicadoras. Um fator com k níveis pode ser representado por k variáveis indicadoras

$$u_i = \begin{cases} 1, & \text{se ocorre o nível i} \\ 0, & \text{caso contrário} \end{cases}$$

como

$$A = \alpha_1 u_1 + \alpha_2 u_2 + \ldots + \alpha_k u_k,$$

onde α_i= valor do i-ésimo nível.

• Termo de Interação Misto

Um termo de interação entre os fatores pode ser incluído no modelo. Em experimentos fatoriais, onde existe apenas uma observação para cada combinação dos níveis dos fatores, se são colocadas todas as interações, tem-se o modelo saturado. No caso de duas variáveis contínuas, a interação é obtida pela inclusão do termo $\beta_{12} x_1 x_2$. Se as variáveis são fatores, utiliza-se $(\alpha\beta)_{ij}$.

Além disso, pode-se ajustar uma componente que represente o efeito simultaneo de um fator e uma variável contínua. Em um modelo com o fator A e a covariável X, definidos anteriormente, ajusta-se o termo $\alpha_j X$ ao invés de αX.

• Notação Utilizada nos MLGs

Wilkinson e Rogers (1973) apresentam uma notação adequada que pode ser utilizada também em programas de computadores. Nesta notação, as primeiras letras do alfabeto $A, B, C, \ldots$ representam os fatores, enquanto que

as últimas $X, Y, Z, \ldots$ são utilizadas para as covariáveis. Esta notação é resumida na Tabela 2.2

Tabela 2.2: *Representação dos Termos nos MLGs*

Tipo do Termo	Fórmula Algébrica	Fórmula do Modelo
Covariável	λx	X
Fator	α_i	A
Misto	$\lambda_i x$	$A.X$
Composto	$(\alpha\beta)_{ij}$	$A.B$
Misto-Composto	$\lambda_{ij} x$	$A.B.X$

2.4 O Algoritmo de Estimação

Existem diversos métodos para estimar os parâmetros β, os quais podemos citar: estimação – M, Bayesiano, qui-quadrado mínimo e o método da máxima verossimilhança que será apresentado mais detalhadamente nesta seção, pelo fato de ser frequentemente utilizado nos programas computacionais.

O algoritmo de estimação dos parâmetros $\beta's$ foi desenvolvido por Nelder e Wedderburn (1972) e baseia-se em um método semelhante ao de Newton-Raphson, conhecido como *Método Escore de Fisher*. A principal diferença em relação ao modelo clássico de regressão é que as equações de máxima verossimilhança são não-lineares.

Seja $l(\beta)$ a log-verossimilhança como função de β. No método escore de Fisher utilizamos a função escore

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta},$$

e a matriz de informação de Fisher

$$K = \left\{ -E\left(\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_s} \right) \right\} = -E\left(\frac{\partial U(\beta)}{\partial \beta} \right).$$

Expandindo a função escore em série de Taylor até primeira ordem, obtém-se

$$U(\beta^{(m+1)}) = U(\beta^{(m)}) + \frac{\partial U(\beta)^{(m)}}{\partial \beta} \left[\beta^{(m+1)} - \beta^{(m)} \right] = 0$$

ou

$$\beta^{(m+1)} = \beta^{(m)} - \left[\frac{\partial U(\beta)^{(m)}}{\partial \beta} \right]^{-1} U(\beta^{(m)}),$$

onde o índice (m) significa o valor do termo na m-ésima iteração. Este é o método de Newton-Raphson para o cálculo iterativo da EMV $\hat{\beta}$ de β. Aitkin et al. (1989) apresentam um estudo completo deste algoritmo.

O método escore de Fisher (1925) é obtido pela substituição de $-\frac{\partial U(\beta)}{\partial \beta}$ pelo seu valor esperado K.

Para desenvolver o algoritmo de estimação do MLG considere a componente sistemática

$$\eta_i = g(\mu_i) = \sum_{r=1}^{p} x_{ir}\beta_r = x_i^T \beta,$$

onde x_i^T é a i-ésima linha de X.

A log-verossimilhança é dada por

$$l(\beta) = \frac{1}{a(\phi)} \sum_{i=1}^{n} \{y_i\theta_i - b(\theta_i)\} + \sum_{i=1}^{n} c(y_i, \phi).$$

Derivando $l(\beta)$ em relação ao vetor β, tem-se

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} \left\{ y_i - b'(\theta_i) \right\} \frac{\partial \theta_i}{\partial \beta}.$$

Calculando

$$\frac{\partial \theta_i}{\partial \beta} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta}$$

pela regra da cadeia e utilizando as equações (2.6), (2.7) e (2.8), obtemos

$$\mu_i = b'(\theta_i) \text{ e } V(\mu_i) = b''(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i}.$$

Como x_i^T é a i-ésima linha de X e $\eta_i = x_i^T \beta$, temos

$$\frac{\partial \eta_i}{\partial \beta} = x_i,$$

onde x_i é um vetor coluna $p \times 1$. Ainda,

$$\frac{\partial \mu_i}{\partial \eta_i} = [g'(\mu_i)]^{-1}.$$

Então, a função escore é expressa como

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} \left\{ y_i - b'(\theta_i) \right\} \frac{1}{V(\mu_i) g'(\mu_i)} x_i.$$

A matriz de informação para β é dada por

$$K = \frac{1}{a(\phi)} X^T W X, \tag{2.9}$$

onde W é uma matriz diagonal de pesos definidos por

$$w_i = V_i^{-1} g'(\mu_i)^{-2}.$$

A função escore, usando esta matriz de pesos, é expressa como

$$U(\beta) = X^T W z,$$

onde z é um vetor com dimensão n x 1 dado por

$$z_i = (y_i - \mu_i)\left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right).$$

Utilizando estes dois resultados, o algoritmo escore de Fisher para calcular a estimativa de máxima verossimilhança (EMV) de β é expresso por

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

Colocando $(X^T W^{(m)} X)^{-1}$ em evidência tem-se, finalmente,

$$\beta^{(m+1)} = (X^T W^{(m)} X)^{-1} X^T W^{(m)} y^{*(m)}, \tag{2.10}$$

onde $y^{*(m)}$ é uma variável resposta modificada denotada por

$$y^{*(m)} = X\beta^{(m)} + z^{(m)}.$$

Note que cada iteração do método escore de Fisher corresponde a uma regressão ponderada da variável dependente modificada y^* sobre a matriz modelo X, com matriz de pesos W. Com isso, quanto maior for a variância da observação, menor será seu peso no cálculo das estimativas dos parâmetros. Um resultado semelhante pode ser obtido pelo método de Newton-Raphson. A estimativa de máxima verossimilhança de β não depende do valor do parâmetro de dispersão ϕ.

Na comparação entre os dois métodos, para os modelos canônicos, tais como, modelo binomial com ligação logística, modelo de Poisson com ligação logaritmica e modelo gama com ligação inversa, eles apresentam resultados idênticos. Contudo, para os demais modelos, os erros padrão das estimativas dos parâmetros são diferentes.

Deve-se ressaltar ainda que os programas computacionais de ajustamento do MLG sempre utilizam o método escore de Fisher para calcular as estimativas dos $\beta's$. Isso deve-se ao fato de que no método de Newton-Raphson existe uma maior probabilidade do algoritmo não convergir.

2.5 Adequação do Modelo

Após formulado o modelo, torna-se necessário estimar os parâmetros e avaliar a precisão das estimativas. Nos MLGs, o processo de estimação é determinado por uma *medida (ou critério)* de bondade de ajuste entre os dados observados e os valores ajustados gerados a partir do modelo. As estimativas dos parâmetros do modelo serão aquelas que minimizam esta *medida* que equivale a maximização da log-verossimilhança descrita na Seção 2.4.

Assim, as estimativas dos parâmetros podem ser obtidas através da maximização da verossimilhança, ou log-verossimilhança, em relação aos parâmetros, supondo fixos os dados observados. Se $f_Y(y;\theta,\phi)$ é a função densidade ou função de probabilidade para a observação y dado o parâmetro θ, supondo ϕ conhecido, então a log-verossimilhança expressa como uma função do valor esperado $\mu = E(Y)$ é dada por

$$l(\mu; y) = \log f_Y(y;\theta,\phi).$$

A log-verossimilhança baseada em uma amostra de observações independentes $y_1, \ldots, y_n$ será a soma das contribuições individuais, ou seja,

$$l(\mu; y) = \sum_{i=1}^{n} \log f_{Y_i}(y_i;\theta_i,\phi),$$

onde $\mu = (\mu_1, \ldots, \mu_n)^T$ e $y = (y_1, \ldots, y_n)^T$.

Uma medida da bondade do ajuste conhecida como *desvio escalonado*, que será abordada mais adiante, é definida como

$$D^*(y;\mu) = 2l(y;y) - 2l(\mu;y).$$

Note-se que, para os modelos exponenciais, $l(y;y)$ representa a máxima verossimilhança de um ajuste exato, no qual os valores ajustados são iguais aos valores observados (modelo saturado). Assim, como $l(y;y)$ não depende dos parâmetros de interesse, maximizar a log-verossimilhança $l(\mu;y)$ é equivalente a minimizar o desvio escalonado $D^*(y;\mu)$ com relação a μ, sujeito às restrições

impostas pelo modelo. Por exemplo, para o modelo normal de regressão com variância σ^2, temos para uma única observação

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

de modo que a log-verossimilhança é dada por

$$l(\mu;y) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}.$$

Obtém-se, então, a log-verossimilhança do modelo saturado fazendo $\mu = y$. Logo,

$$l(y;y) = -\frac{n}{2}\log(2\pi\sigma^2).$$

Então, o desvio escalonado para o modelo normal iguala

$$D^*(y;\mu) = 2\left\{l(y;y) - l(\mu;y)\right\} = \frac{\sum_i (y_i-\mu_i)^2}{\sigma^2}.$$

2.6 Predição

A predição no contexto dos MLGs deve ser interpretada como uma pergunta do tipo *"o que... se... ?"*, ao contrário do contexto de séries temporais onde o valor predito está indexado pelo tempo. É importante salientar que as quantidades preditas devem estar sempre acompanhadas por medidas de precisão e que o modelo utilizado esteja correto. Para um estudo mais detalhado sobre predições, análise de variância e vários tipos de padronizações, vide Lane e Nelder (1982).

2.7 Medidas de Discrepância ou Bondade de Ajuste

2.7.1 A função desvio

Existem diversas maneiras de se construir medidas de discrepância ou bondade de ajuste. Uma destas medidas denomina-se *desvio* e equivale à diferença de

log-verossimilhanças maximizadas.

Sabemos que, dado n observações, podemos construir modelos com até n parâmetros. Porém, o modelo mais simples, chamado de *modelo nulo,* contém apenas um parâmetro que representa a média μ comum a todas as observações y's. O modelo nulo aloca toda a variação entre os y's para a componente aleatória. Por outro lado, o *modelo saturado* contém n parâmetros, um para cada observação. No modelo saturado toda a variação dos y's é alocada para a componente sistemática.

Assim, na prática, o modelo nulo é muito simples enquanto o modelo saturado é não-informativo. Porém, o modelo saturado é útil para medir a discrepância de um modelo intermediário (em investigação) com p parâmetros ($p < n$).

Seja $y = (y_1, \ldots, y_n)^T$ uma amostra aleatória com distribuição pertencente à família exponencial (2.1). Sejam $\widehat{\theta} = \theta(\widehat{\mu})$ e $\widetilde{\theta} = \theta(y)$ as estimativas dos parâmetros canônicos para o modelo em investigação e o modelo saturado, respectivamente. Seja

$$\widehat{l}_p = \sum_{i=1}^{n} l(\widehat{\theta}_i, \phi; y_i) = \sum_{i=1}^{n} \{[y_i\widehat{\theta}_i - b(\widehat{\theta}_i)]/a_i(\phi) + c(y_i, \phi)\},$$

a log-verossimilhança maximizada sobre β para ϕ fixo. Seja

$$\tilde{l}_n = \sum_{i=1}^{n} l(\widetilde{\theta}_i, \phi; y_i) = \sum_{i=1}^{n} \{[y_i\widetilde{\theta}_i - b(\widetilde{\theta}_i)]/a_i(\phi) + c(y_i, \phi)\}$$

a log-verossimilhança para o modelo saturado com n parâmetros. Assumindo ainda que $a_i(\phi) = \phi/\lambda_i$, podemos escrever

$$2(\tilde{l}_n - \widehat{l}_p) = 2\sum_{i=1}^{n} \lambda_i\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\}/\phi = D(y;\mu)/\phi = D/\phi,$$

onde

$$D = D(y;\mu) = 2\sum_{i=1}^{n} \lambda_i\{y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\}$$

é denominado *desvio* do modelo em investigação, sendo função apenas dos dados e das estimativas de máxima verossimilhança obtidas dos mesmos.

Temos a seguir as formas da *função desvio* com $\lambda_i = 1$ (caso mais comum) para as principais distribuições da família exponencial citadas na Tabela 2.1:

- normal $\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$;
- Poisson $2\sum_{i=1}^{n}\{y_i \log(y_i/\widehat{\mu}_i) - (y_i - \widehat{\mu}_i)\}$;
- binomial $2\sum_{i=1}^{n}\{y_i \log(y_i/\widehat{\mu}_i) + (m_i - y_i)\log[(m_i - y_i)/(m_i - \widehat{\mu}_i)]\}$;
- gama $2\sum_{i=1}^{n}\{\log(\widehat{\mu}_i/y_i) + (y_i - \widehat{\mu}_i)/\widehat{\mu}_i\}$;
- normal inversa $\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2/(\widehat{\mu}_i^2 y_i)$.

Maiores detalhes são dados por Nelder e Wedderburn (1972).

2.7.2 A estatística de Pearson generalizada X^2

Uma outra importante medida de discrepância do modelo ajustado em relação aos dados é a estatística de Pearson generalizada definida por

$$X^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/V(\hat{\mu}_i),$$

onde $V(\hat{\mu}_i)$ é a função de variância estimada para a distribuição proposta para os dados.

Tanto a *função desvio* quanto a *estatística de Pearson generalizada* têm, para o modelo normal linear, distribuição χ^2 exata. Resultados assintóticos são possíveis para outras distribuições. A vantagem da *função desvio* é que ela é aditiva e acrescentando-se variáveis explicativas ao modelo, o desvio deve decrescer, diferentemente de X^2. Contudo, X^2 é algumas vezes preferível pois tem uma interpretação simples.

2.7.3 A análise do desvio

A *análise do desvio* é uma generalização da análise de variância para os MLGs visando obter, a partir de uma seqüência de modelos encaixados, isto é, cada

modelo incluindo mais termos que os anteriores, os efeitos de fatores, covariáveis e suas possíveis interações.

Dois modelos M_{p_r} e M_{p_s} são encaixados ($M_{p_r} \subset M_{p_s}$) quando os termos que formam M_{p_s} incluem todos os termos que compõem M_{p_r} mais outros termos que não estão em M_{p_r}.

Considere $M_{p_1} \subset M_{p_2} \subset \ldots \subset M_{p_r}$ uma seqüência de modelos encaixados com respectivas dimensões $p_1 < p_2 < \ldots < p_r$, matrizes X_{p_1}, $X_{p_2}, \ldots, X_{p_r}$, desvios $D_{p_1} > D_{p_2} > \ldots > D_{p_r}$, todos os modelos com a mesma distribuição e função de ligação. Vale ressaltar que as desigualdades entre os desvios não são válidas para a estatística de Pearson generalizada. Logo, a comparação de modelos encaixados é feita, exclusivamente, pela função desvio.

As diferenças entre os desvios $D_{p_i} - D_{p_j}, p_i < p_j$, devem ser interpretadas como uma medida de variação dos dados, sendo explicada pelos termos que estão em M_{p_j} e não estão em M_{p_i}. Se $D_{p_i} - D_{p_j} > \chi^2_{p_j - p_i, \alpha}$ consideramos que os termos que estão em M_{p_j} e não estão em M_{p_i} são significativos.

Para entender este procedimento, tem-se um exemplo de planejamento com dois fatores A e B, com a e b níveis, respectivamente. Ajustam-se, sucessivamente, os modelos: 1 (modelo nulo), A, $A + B$, $A + B + A.B$ (modelo saturado). Na Tabela 2.3, apresenta-se a análise do desvio para esta seqüência de modelos juntamente com a interpretação dos termos.

Tabela 2.3: *Exemplo de Análise do Desvio*

Modelo	g.l.	Desvio	Diferença	g.l.	Termo
1	$ab-1$	D_1			
A	$a(b-1)$	D_A	$D_1 - D_A$	$a-1$	A ignorando B
$A+B$	$(a-1)(b-1)$	D_{A+B}	$D_A - D_{A+B}$	$b-1$	B incluído A
$A+B+A.B$	0	0	D_{A+B}	$(a-1)(b-1)$	interação $A.B$ incluídos A e B

2.8 Modelo Binomial

Esta é uma das mais antigas distribuições de probabilidade e foi desenvolvida por James Bernoulli em seu tratado Ars Conjectand, publicado em 1713. A distribuição binomial surge naturalmente em um grande número de situações, onde as observações Y são contagens não-negativas limitadas por um valor fixo. Existem duas maneiras de deduzi-la.

Supondo que Y_1 e Y_2 são variáveis aleatórias independentes de Poisson com médias μ_1 e μ_2, respectivamente, sabemos que $Y_1 + Y_2$ tem distribuição de Poisson com média $\mu_1 + \mu_2$. Assim, a distribuição condicional de Y_1 dado $Y_1 + Y_2 = m$ é expressa como

$$P(Y_1 = y \mid Y_1 + Y_2 = m) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m \tag{2.11}$$

onde $\pi = \mu_1/(\mu_1+\mu_2)$. A notação $Y \sim B(m, \pi)$ denota que Y tem distribuição binomial, expressa em (2.11), com índice m e parâmetro π.

A segunda maneira e também a mais natural, vem da distribuição de Bernoulli, expressa em (2.12), que denota um caso particular da distribuição binomial quando $m = 1$. Na distribuição de Bernoulli, Y_i assume dois valores

$$Y_i = \begin{cases} 1 \text{ se o evento de interesse ocorre na repetição } i \\ 0 \text{ caso contrário,} \end{cases}$$

tal que

$$P(Y_i = k) = \pi^k (1-\pi)^{1-k}, k = 0, 1, \tag{2.12}$$

onde π representa a probabilidade do evento de interesse ocorrer.

Assim, obtemos a distribuição binomial (2.11) para a soma $S_m = \sum_{i=1}^{m} Y_i$ de m variáveis aleatórias $Y_1, \ldots, Y_m$ de Bernoulli independentes e identicamente distribuídas conforme (2.12).

A função de probabilidade de S_m/m (proporção de sucessos) está na família exponencial (2.1) com parâmetro canônico $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, onde $\mu = E(S_m/m)$ é a probabilidade de sucesso. O parâmetro canônico representa, então, o logaritmo da razão de chances e a função de variância (2.8) iguala $V(\mu) = \frac{\mu(1-\mu)}{m}$.

2.8.1 Momentos e cumulantes

A função geratriz de cumulantes da binomial pode ser facilmente obtida a partir da soma de funções de cumulantes de variáveis aleatórias de Bernoulli independentes. A função geratriz de momentos de (2.12) é

$$M_Y(t) = E\{\exp(tY)\} = 1 - \pi + \pi\exp(t). \tag{2.13}$$

Então, temos a função geratriz de cumulantes

$$K_Y(t) = \log M_Y(t) = \log\{1 - \pi + \pi\exp(t)\}.$$

Por conseguinte, a função geratriz de momentos da soma estocástica $S_m = Y_1 + \cdots + Y_m$ é

$$M_{S_m}(t) = \{1 - \pi + \pi\exp(t)\}^m$$

e sua correspondente função geratriz de cumulantes iguala

$$\log M_{S_m}(t) = m\log\{1 - \pi + \pi\exp(t)\}. \tag{2.14}$$

Finalmente, expandindo (2.14) em série de Taylor e avaliando no ponto $t = 0$, encontramos os quatro primeiros cumulantes da distribuição binomial expressos por $\kappa_1 = m\pi$, $\kappa_2 = m\pi(1-\pi), \kappa_3 = m\pi(1-\pi)(1-2\pi)$ e $\kappa_4 = m\pi(1-\pi)\{1 - 6\pi(1-\pi)\}$.

2.8.2 Convergência para normal e Poisson

A partir da função geratriz de cumulantes (2.14) pode-se mostrar que, para m grande, todos os cumulantes de S_m são de ordem m. Logo, os cumulantes da

variável aleatória padronizada

$$Z = \frac{S_m - m\pi}{\sqrt{m\pi(1-\pi)}}$$

são: 0, para $r = 1$, e $O(m^{1-r/2})$ para $r \geq 2$. Consequentemente, quando π é fixo e $m \to \infty$, os cumulantes de Z convergem para os de uma distribuição normal padrão: $0, 1, 0, 0, \ldots$ Então, como convergência de cumulantes implica convergência em distribuição, temos que

$$P(S_m \leq y) \simeq \Phi(z^+),$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada da normal-padrão, y é um inteiro e

$$z^+ = \frac{y - m\pi + 0.5}{\sqrt{m\pi(1-\pi)}}.$$

Agora, suponha que $\pi \to 0$ e $m \to \infty$, de tal forma que $\mu = m\pi$ permanece fixo ou tende para uma constante. De (2.14), a função geratriz de cumulantes de S_m tende para

$$\frac{\mu}{\pi} \log\{1 + \pi(\exp(t) - 1)\} \to \mu\{\exp(t) - 1\}$$

que é a função geratriz de cumulantes de uma variável aleatória com distribuição de Poisson de média μ. Da mesma forma, convergência da função de cumulantes implica convergência em distribuição.

2.8.3 Funções de ligação apropriadas

Para investigar a relação entre a probabilidade de sucesso π da variável resposta Y e o vetor de covariáveis $(x_1, \ldots, x_p)$ assumimos que a dependência entre π e $(x_1, \ldots, x_p)$ ocorre através da combinação linear

$$\eta = \sum_{j=1}^{p} \beta_j x_j.$$

Contudo, como $-\infty < \eta < \infty$, expressar π através de uma função linear de η seria errôneo do ponto de vista probabilístico, pois π não ficaria restrito ao intervalo (0,1). Assim, uma maneira simples e eficaz para solucionar este problema é o uso de uma transformação $g(\pi)$ que relacione o intervalo unitário à reta real, de tal forma que

$$g(\pi_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j, \qquad i = 1, \ldots, n.$$

Apresentamos abaixo algumas funções de ligação que são adequadas para dados binários, pois preservam as restrições sobre a probabilidade π:
1. Logit ou função logística

$$g_1(\pi) = \log\{\pi/(1-\pi)\};$$

2. Função probit ou inversa da distribuição acumulada da normal reduzida

$$g_2(\pi) = \Phi^{-1}(\pi);$$

3. Complemento log-log

$$g_3(\pi) = \log\{-\log(1-\pi)\}.$$

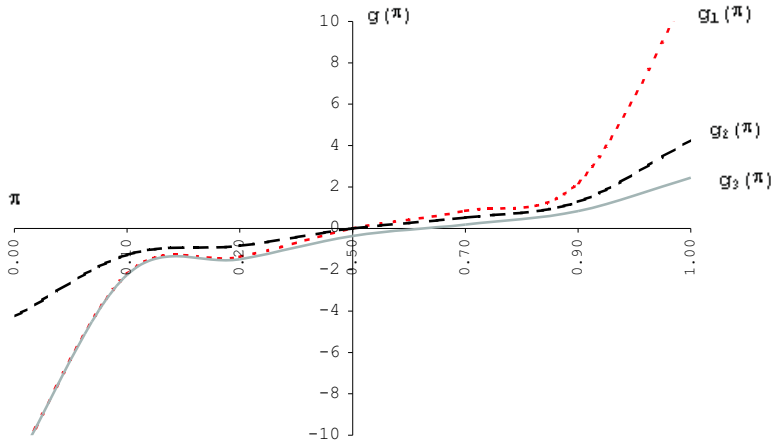
Todas as três funções possuem inversas, são contínuas e crescentes no intervalo (0,1).

Na Figura 2.1, podemos observar o comportamento das três principais ligações usualmente empregadas no modelo binomial.

As três ligações: logística, probit e complemento log-log, apresentam um comportamento praticamente linear no intervalo $0,1 \leq \pi \leq 0,9$. Para pequenos valores de π, as ligações logística e complemento log-log encontram-se bastante próximas, decaíndo mais rapidamente que a probit. Entretanto, quando π se aproxima de 1, a ligação complemento log-log cresce mais lentamente do que as ligações probit e logística. Uma característica da ligação logística é que ela decresce quando π vai para 0 e cresce quando π vai para 1 de forma bastante rápida, ou seja, quando π está próximo destes valores

limites.

Figura 2.1: *Ligações Usuais*



A função logística possui algumas características que a tornam preferida em relação às outras ligações na análise de dados binários: (i) pode ser interpretada como o logaritmo da razão de chances; (ii) apresenta propriedades teóricas mais simples; (iii) é mais conveniente para análise de dados coletados de forma retrospectiva. Entretanto, isto não quer dizer que as outras tranformações não são utilizadas na prática. Bliss (1935), utilizando um modelo binomial com ligação probit, foi quem iniciou a modelagem de proporções. A ligação logística é bastante empregada em estudos toxicológicos e epidemiológicos. A ligação complemento log-log é recomendada por Collett (1994) quando a distribuição das proporções é bastante assimétrica.

Para compreender melhor o ajuste obtido é necessário a utilização da relação entre π e o preditor linear $\eta = X\beta$. A ligação logística satisfaz

$$\log\{\pi/(1-\pi)\} = \eta = X\beta.$$

Expressando-a em termos do preditor linear, temos

$$\pi = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

Logo, se a parte sistemática do modelo para uma determinada observação tende para um valor muito negativo, sua probabilidade de sucesso tende para zero. Por outro lado, se a mesma tende para um valor muito grande, esta probabilidade tende para um.

Da mesma forma, pode-se calcular a relação entre π e η para as outras ligações:

$$\pi = g_2^{-1}(\eta) = \Phi(\eta)$$

e

$$\pi = g_3^{-1}(\eta) = 1 - \exp\{\exp(-\eta)\}.$$

Além das ligações citadas anteriormente, Aranda-Ordaz (1981) apresenta duas famílias de transformações para dados binários. A primeira é expressa por

$$T_\lambda(\pi) = \frac{2}{\lambda}\frac{\pi^\lambda - (1-\pi)^\lambda}{\pi^\lambda + (1-\pi)^\lambda}, \tag{2.15}$$

onde π denota a probabilidade de sucesso e λ representa o parâmetro da transformação.

Duas características importantes de (2.15) são $T_\lambda(\pi) = -T_\lambda(1-\pi)$ e $T_\lambda(\pi) = T_{-\lambda}(\pi)$, ou seja, T_λ trata sucesso e fracasso de forma simétrica. A família $\mathcal{F}$, como é denotada $T_\lambda(\pi)$, é chamada de *simétrica*.

A expressão (2.15) se reduz à transformação logística no limite quando $\lambda = 0$ e à transfomação linear quando $\lambda = 1$. Além disso, invertendo (2.15), obtemos

$$\pi(\eta) = \begin{cases} 0 & \left(\left|\frac{1}{2}\lambda\eta\right| \leq -1\right), \\ \dfrac{\left(1+\frac{1}{2}\lambda\eta\right)^{1/\lambda}}{\left(1+\frac{1}{2}\lambda\eta\right)^{1/\lambda} + \left(1-\frac{1}{2}\lambda\eta\right)^{1/\lambda}} & \left(\left|\frac{1}{2}\lambda\eta\right| < 1\right), \\ 1 & \left(\left|\frac{1}{2}\lambda\eta\right| \geq 1\right), \end{cases} \tag{2.16}$$

onde $\eta = X\beta$ é o preditor linear que pode assumir qualquer valor real.

Em situações onde é apropriado tratar sucesso e fracasso de forma assimétrica (Yates (1955) traz alguns exemplos), uma segunda família de transformações é proposta, sendo definida por

$$W_\lambda(\pi) = \frac{\{(1-\pi)^{-\lambda} - 1\}}{\lambda} \ . \tag{2.17}$$

Aqui, assumimos que

$$\log W_\lambda(\pi) = \eta,$$

onde η tem a mesma expressão linear citada anteriormente.

Para $\lambda = 1$, (2.17) se reduz à transformação logística, enquanto que para $\lambda = 0$ obtemos o complemento log log. Invertendo (2.17), tem-se que

$$\pi(\eta) = \begin{cases} 1 - (1 + \lambda e^\eta)^{-1/\lambda} & (\lambda e^\eta > -1), \\ \\ 1, & \text{caso contrário.} \end{cases} \tag{2.18}$$

No contexto dos MLGs, Aranda-Ordaz (1981) sugere que a função de ligação seja definida em termos das transformações inversas (2.16) ou (2.18).

A família $\mathcal{F}$ é analisada graficamente, supondo os seguintes valores arbitrários de λ: 0, 0,25 e 0,5. É importante lembrar que $T_\lambda(\pi) = T_{-\lambda}(\pi)$ e que quando λ= 0 temos a ligação logística como um caso particular. Pela Figura 2.2, podemos observar que quando $\pi < 0{,}1$ e $\pi > 0{,}8$, $T_\lambda(\pi)$ cresce ou decresce muito pouco, à medida que λ assume valores mais distantes de 0. Entretanto, para valores de $0{,}2 \leq \pi \leq 0{,}8$, praticamente não há diferença entre as ligações para os diversos λ's.

Figura 2.2: *Ligações Aranda-Ordaz (Simétricas)*

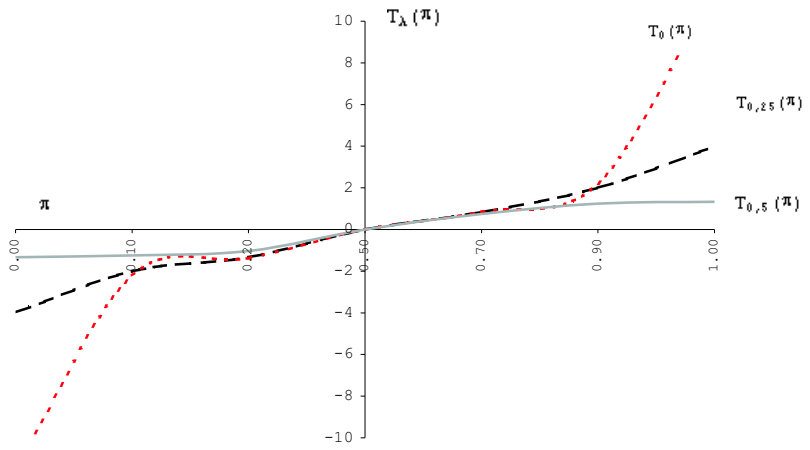
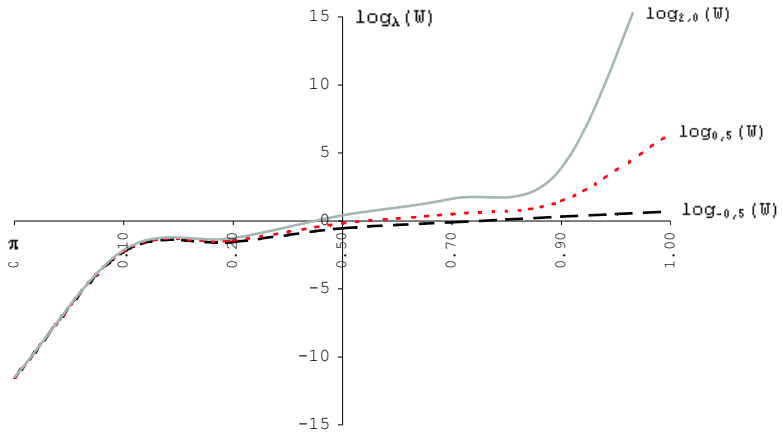


Figura 2.3: *Ligações Aranda-Ordaz (Assimétricas)*



Finalmente, na Figura 2.3, podemos visualizar algumas ligações de Aranda-Ordaz recomendadas quando tratamos sucesso e fracasso de forma

assimétrica. Os valores arbitrários de λ utilizados foram -0,5, 0,5 e 2,0. Observando a Figura 2.3 fica bastante claro que quando $\pi \leq 0{,}1$ não existe diferença entre as ligações. Porém, para valores de $\pi \geq 0{,}8$, quanto maior o valor de λ mais rapidamente $\log W_\lambda(\pi)$ cresce. Para maiores detalhes sobre estas famílias de ligação, vide Aranda-Ordaz (1981).

2.8.4 A função de verossimilhança

Considerando os dados $y_1, \ldots, y_n$ como valores observados de variáveis aleatórias independentes $Y_1, \ldots, Y_n$ com distribuição binomial de índice m_i e parâmetro π_i, respectivamente, temos, a partir de (2.11), que a log-verossimilhança de π dado y é escrita da seguinte forma

$$l(\pi; y) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right]. \tag{2.19}$$

O termo $\sum \log \binom{m_i}{y_i}$ pode ser omitido, pois não involve o parâmetro π.

A log-verossimilhança também pode ser escrita em função do preditor linear. Para isso é necessário a utilização da equação

$$g(\pi_i) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j, \qquad i = 1, \ldots, n.$$

Se a função escolhida para o modelo for a logística, obtém-se

$$g(\pi_i) = \eta_i = \log\{\pi_i/(1-\pi_i)\} = \sum_{j=1}^{p} x_{ij}\beta_j, \qquad i = 1, \ldots, n.$$

Expressando a log-verossimilhança em função dos parâmetros desconhecidos, temos

$$l(\beta; y) = \sum_{i=1}^{n}\sum_{j=1}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{n} m_i \log\left\{1 + \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)\right\}.$$

Um ponto importante que deve ser ressaltado é que a estatística $X^T y$, que aparece na log-verossimilhança, é suficiente para β, pois a ligação logística também é a ligação canônica no modelo binomial.

2.8.5 Estimação dos parâmetros

Para estimarmos os parâmetros usando o método escore de Fisher, apresentado na Seção 2.4, basta calcular a função escore e a matriz de informação de Fisher para a log-verossimilhança do modelo binomial em que $\mu = m\pi$, obtendo-se

$$U(\beta) = X^T(y - \mu)$$

e

$$K = X^T W X,$$

onde

$$W = diag\{m_i \pi_i (1 - \pi_i)\}.$$

Finalmente, o algoritmo de estimação de β é dado por

$$\beta^{(m+1)} = \beta^{(m)} + K^{(m)^{-1}} U(\beta^{(m)}).$$

É importante salientar que neste algoritmo as observações com maior variância

$$V(\pi_i) = m_i \pi_i (1 - \pi_i),$$

tem menor peso w_i para o cálculo da estimativa do vetor β.

2.8.6 A função desvio

Sabemos que a função desvio corresponde a duas vezes a diferença entre as log-verossimilhanças maximizadas, sob o modelo saturado e sob o modelo em investigação. Sob o modelo em investigação, com probabilidade estimada $\hat{\pi}$, a log-verossimilhança é dada por

$$l(\hat{\pi}; y) = \sum_i \{y_i \log \hat{\pi}_i + (m_i - y_i) \log(1 - \hat{\pi}_i)\},$$

onde $\hat{\pi}_i = \pi(\hat{\mu}_i) = \hat{\mu}_i/m_i$. No modelo saturado, a EMV de π_i é obtida por $\tilde{\pi}_i = y_i/m_i$.

Assim, a função desvio para o modelo binomial é expressa como

$$\begin{aligned} D(y;\hat{\pi}) &= 2l(\tilde{\pi};y) - 2l(\hat{\pi};y) \\ &= 2\sum_i \left\{ y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right\}, \end{aligned}$$

onde $\mu_i = m_i\pi_i$.

A variável aleatória $D(y;\hat{\pi})$ é distribuída aproximadamente como χ^2_{n-p}, onde p é o número de parâmetros ajustados segundo o modelo em investigação.

2.9 Modelo de Poisson

Ao contrário da seção anterior, em que a variável resposta assumia a forma de proporção, quando a mesma apresenta a forma de contagem, sendo as ocorrências desta variável independentes, com uma taxa que é função das variáveis que compõem X, é de se esperar que a distribuição de Poisson modele bem esses dados. O modelo de Poisson, ao contrário do modelo normal, supõe que a variância seja proporcional a média e pode ser aplicado para modelar, por exemplo, o número de acidentes diários em uma estrada, o número de pacientes infectados por uma doença específica, etc.

2.9.1 A distribuição de Poisson

Em 1837, Poisson desenvolveu esta distribuição como limite da distribuição binomial $mp = \mu$ fixo e $m \to \infty$. A distribuição de Poisson supõe que a variável de interesse assume valores inteiros não-negativos e, em particular, não existe um limite superior. A função de probabilidade de Poisson é expressa por

$$P(Y = y) = \exp(-\mu)\frac{\mu^y}{y!}, \qquad y = 0, 1, 2, \ldots$$

com $\mu > 0$.

2.9.2 Função geratriz de momentos e cumulantes

A função geratriz de momentos da distribuição de Poisson é

$$M_Y(t) = \exp\{\mu \exp(t) - 1\}.$$

Assim, a função geratriz de cumulantes é expressa por

$$K_Y(t) = \mu \exp(t) - 1,$$

cuja r-ésima derivada é igual a

$$\frac{\partial^r K_Y(t)}{\partial t^r} = \mu \exp(t), \ r \geq 1.$$

Logo, todos os cumulantes são iguais e dados por

$$\kappa_r = \mu, \qquad r \geq 1.$$

Em especial, $\mathrm{Var}(Y) = E(Y) = \mu$.

2.9.3 A Função de ligação

A ligação canônica para a distribuição de Poisson é a logaritmica

$$\eta = \log \mu.$$

É importante salientar que o modelo de Poisson com ligação logaritmica é conhecido como *Modelo Log-Linear*. Outra ligação que pode ser empregada no modelo de Poisson é a ligação potência. Cordeiro (1986; Seção 9.3.5) estuda esta opção utilizando aproximações assintóticas para o desvio.

2.9.4 Função desvio e principais transformações

Para um vetor de observações independentes com distribuição de Poisson, a log-verossimilhança é dada por

$$l(\mu; y) = \sum_{i=1}^{n} (y_i \log \mu_i - \mu_i), \tag{2.20}$$

podendo ser expressa em função dos parâmetros desconhecidos como

$$l(\beta; y) = \sum_{i=1}^{n} \sum_{j=1}^{p} \{y_i x_{ij} \beta_j - \exp(x_{ij}\beta_j)\}.$$

O valor de $\hat{\mu}_i = \exp(x_i^T \hat{\beta})$ é sempre positivo, ficando coerente com a distribuição de Poisson.

A partir da expressão (2.20) podemos obter a função desvio, expressa por

$$\begin{aligned} D(y; \hat{\mu}) &= 2l(y; y) - 2l(\hat{\mu}; y) \\ &= 2\sum_{i=1}^{n} \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}. \end{aligned}$$

Se um termo constante for incorporado ao modelo, Nelder e Wedderburn (1972) mostram que $\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0$, de tal forma que $D(y; \hat{\mu})$ reduz-se a $2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i)$, que é a estatística da razão de verossimilhanças comumente usada na análise de tabelas de contingência.

Caso haja interesse em transformar a variável resposta Y, duas sugestões para dados sob forma de contagens são $Y^{1/2}$ e $Y^{2/3}$, a segunda proposta por Anscombe (1953). A primeira transformação estabiliza a variância e possui os seguintes momentos para μ suficientemente grande: $E(Y^{1/2}) \simeq \mu^{1/2}$ e $\text{Var}(Y^{1/2}) \simeq 1/4$. A segunda transformação $Y^{2/3}$ produz uma variável aleatória mais simétrica. Um modelo alternativo é obtido a partir da suposição de normalidade para os dados transformados.

Uma terceira transformação, proposta por McCullagh e Nelder (1983, Capítulo 6), que produz simetria e estabilização da variância, é denotada a seguir:

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

Se $Y \sim P(\mu)$, então $g(Y)$ tem, aproximadamente, distribuição normal padrão.

Alternativamente, Freeman e Tukey (1950) sugerem a variável transformada

$$W = \sqrt{Y} + \sqrt{Y+1}.$$

Além disso, Anscombe (1948) propôs utilizar $2\sqrt{Y + \frac{3}{8}}$ como alternativa para melhorar a normalidade dos dados sob forma de contagens.

2.9.5 O parâmetro de dispersão

O modelo de Poisson pode ser definido com a variação para y dada por

$$\mathrm{Var}(y_i) = \phi E(y_i),$$

incluindo assim o parâmetro de dispersão ϕ que tem como objetivo explicar uma variação acima daquela estabelecida pela distribuição de Poisson. Entretanto, esta suposição não modifica a função de variância dada por

$$\mathrm{Var}(y_i) = a(\phi)V(\mu_i),$$

pois

$$V(\mu_i) = E(y_i) = \mu_i.$$

2.9.6 A distribuição multinomial e a Poisson

Ao se estudar uma variável que possui k categorias vários esquemas de amostragem são possíveis, sendo o mais simples aquele em que um número

fixado de indivíduos é escolhido aleatoriamente, implicando que as freqüências nas categorias seguem uma distribuição multinomial com probabilidades desconhecidas que devem ser estimadas.

Supondo que cada reposta segue uma distribuição de Poisson, onde $Y_1, \ldots, Y_n$ são independentes, então a distribuição conjunta de $Y_1, \ldots, Y_n$ condicionada à soma $\sum_{i=1}^{n} Y_i$ é multinomial. Portanto, escolhendo-se a função de ligação logaritmo, a verossimilhança da resposta multinomial é proporcional a verossimilhança de um modelo de Poisson supondo que as variáveis são independentes com média μ_i. Com isso, a análise de dados multinomiais pode ser feita a partir do tratamento das respostas como variáveis de Poisson independentes. Este modelo é chamado de "Poisson Trick" (Francis et al., 1993).

2.10 Modelo Normal

O modelo clássico de regressão, discutido amplamente no Capítulo 1, é o caso mais simples de MLG ocorrendo quando a distribuição dos dados é normal e a função de ligação é a identidade. A distribuição normal é utilizada em modelos para dados contínuos, embora possa ser usada como uma aproximação em modelos que tratem de quantidades discretas. Além disso, ela é frequentemente usada para modelar dados tais como: peso, altura e tempo, que são essencialmente positivos, apesar de seu domínio ser a reta real.

As hipóteses básicas do modelo normal linear são:

$$\begin{array}{lll} Y_i \sim N(\mu_i, \sigma^2) & \mu = \eta & \eta = \sum_{j=1}^{p} x_j \beta_j \\ \text{observações} & \text{função de} & \text{preditor linear baseado} \\ \text{normais independentes} & \text{ligação} & \text{nas covariáveis } x_1, \ldots, x_p \end{array} \tag{2.21}$$

onde o vetor Y, o vetor de médias μ e o preditor linear η são de dimensão n. Em (2.21), temos mais à esquerda, a componente aleatória do modelo seguida da componente sistemática que inclui a construção do preditor linear η a partir das variáveis explicativas e da função de ligação entre μ e η.

2.10.1 Cumulantes e estimação

No modelo clássico de regressão, considera-se o vetor de observações y como sendo as realizações de uma variável aleatória Y, que tem distribuição normal com $E(Y) = X\beta$ e $Cov(Y) = \sigma^2 I$. Assim, considera-se que as observações são independentes e têm igual variância.

A função geratriz de momentos da normal é dada por

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

sendo seus cumulantes $\kappa_r = 0$ para $r > 2$. Outras características desta distribuição são: média, mediana e moda iguais a μ e coeficientes de assimetria e curtose iguais a 0 e 3, respectivamente.

No modelo clássico de regressão, a EMV de β, que coincide com a de mínimos quadrados, é dada em forma fechada por

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

A função de verossimilhança depende apenas dos dados através de $\hat{\beta}$ e da soma dos quadrados dos resíduos $SQR = (y - X\hat{\beta})^T (y - X\hat{\beta})$. Sabe-se ainda que $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ e $SQR \sim \sigma^2 \chi^2_{n-p}$. Os testes estatísticos são realizados de forma exata através das estatísticas χ^2, t de Student e F como descritos no Capítulo 1.

2.11 Modelo Gama

O modelo gama é utilizado na análise de dados não-negativos de natureza contínua que apresentam uma variância crescente com a média. Além disso, assumimos que o coeficiente de variação é constante, isto é,

$$\mathrm{Var}(Y) = \sigma^2 \{E(Y)\}^2 = \sigma^2 \mu^2.$$

Note que aqui σ é o coeficiente de variação de Y e não o desvio padrão. O modelo gama também é aplicado na estimação de variâncias na análise de variância e como distribuição aproximada de medidas físicas, tempos de sobrevivência, etc.

2.11.1 A distribuição gama

O primeiro trabalho com esta distribuição foi realizado por Laplace (1836). Na família exponencial (2.1) é mais conveniente reparametrizar a sua função densidade em termos da média μ e do parâmetro de forma ν. Temos,

$$f(y;\nu,\mu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\mu}\right)^{\nu} \exp\left(-\frac{\nu y}{\mu}\right), \qquad y>0, \nu>0, \mu>0, \tag{2.22}$$

onde $\Gamma(\cdot)$ é a função gama. Assim, dizemos que $Y \sim G(\mu,\nu)$.

A partir de (2.22) pode-se encontrar a função geratriz de cumulantes como

$$K_Y(t) = -\nu \log(1-\mu t/\nu).$$

Os quatro primeiros cumulantes de Y são dados a seguir $\kappa_1 = E(Y) = \mu$, $\kappa_2 = \mathrm{Var}(Y) = \mu^2/\nu$, $\kappa_3 = E(Y-\mu)^3 = 2\mu^3/\nu^2$ e $\kappa_4 = E(Y-\mu)^4 = 6\mu^4/\nu^3$. Como $\nu = \mu^2/k_2$, ν é um parâmetro de precisão.

De forma geral, o r-ésimo cumulante pode ser obtido através de

$$\kappa_r = (r-1)!\mu^r/\nu^{r-1}.$$

A distribuição gama apresenta formas bastante diferentes sendo caracterizada pelo parâmetro de forma ν mas, aqui, estamos interessados apenas nos modelos em que este parâmetro é constante para todas as observações, de modo que as densidades de todas as observações têm a mesma forma. Por analogia aos modelos de mínimos quadrados ponderados em que as variâncias são proporcionais a constantes conhecidas, é permitido, no contexto do modelo gama, que o valor de ν varie de uma observação para outra, de modo que $\nu_i =$ constante $\times\ \lambda_i$, onde os λ_i são pesos a priori conhecidos e ν_i é o índice ou parâmetro de precisão para Y_i.

2.11.2 A função de variância

Sob a suposição da distribuição gama para a componente aleatória de um MLG, a função de variância assume forma quadrática, isto é, $V(\mu) = \mu^2$. A log-verossimilhança como função de ν e μ para uma única observação y é

$$l(\nu, \mu; y) = \nu(-y/\mu - \log \mu) + \nu \log y + \nu \log \nu - \log \Gamma(\nu),$$

onde $a(\nu) = 1/\nu$, $c(y, \nu) = \nu \log y + \nu \log \nu - \log \Gamma(\nu)$, $\theta = -1/\mu$ é o parâmetro canônico e $b(\theta) = -\log(-\theta)$ a função cumulante.

2.11.3 O desvio

Fazendo ν uma constante conhecida, a log-verossimilhança pode ser escrita como

$$l(\nu, \mu; y) = \sum_i \nu(-y_i/\mu_i - \log \mu_i)$$

para observações independentes. Se o parâmetro ν não é constante, mas pode ser escrito como $\nu_i = \nu\lambda_i$, a log-verossimilhança é expressa por

$$l(\nu, \mu; y) = \nu \sum_i \lambda_i(-y_i/\mu_i - \log \mu_i),$$

onde os λ_i são pesos a priori conhecidos.

O valor máximo da log-verossimilhança ocorre para o modelo saturado quando $\mu = y$, sendo expresso por

$$\nu \sum_i \lambda_i(1 + \log y_i),$$

que é finito para todo $y_i > 0$.

Assim, a partir da definição do desvio dada na Seção 2.5, obtemos a função desvio para o modelo gama

$$D(y; \hat{\mu}) = 2 \sum_i \lambda_i \{\log(\hat{\mu}_i/y_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\}.$$

Note que a estatística é definida apenas se todas as observações forem estritamente positivas.

De forma geral, se algumas das componentes de y assumem valor zero, podemos substituir $D(y;\mu)$ por

$$D^+(y;\hat{\mu}) = 2C(y) + 2\sum_i \lambda_i \log \hat{\mu}_i + 2\sum_i \lambda_i y_i/\hat{\mu}_i,$$

onde $C(y)$ é uma função limitada arbitrária de y.

Entretanto, note-se que a estimativa de máxima verossimilhança de ν é uma função de $D(y;\hat{\mu})$ e não de $D^+(y;\hat{\mu})$. Assim, se alguma componente de y é zero, então, $\hat{\nu} = 0$. A solução deste problema será apresentada na Seção 2.11.5, onde será mostrado um estimador alternativo para $\hat{\nu}$.

2.11.4 A função de ligação

Supondo o modelo gama, a função de ligação canônica que produz estatísticas suficientes, que são funções lineares dos dados, é expressa por

$$\eta = \mu^{-1}.$$

Contudo, para o referido modelo, a ligação canônica apresenta um grave problema: ela não garante que $\mu > 0$, implicando em restrições para as componentes do vetor de parâmetros β.

Assim, uma função de ligação comumente utilizada é

$$\eta = \log \mu,$$

que garante $\mu > 0$, pois $\mu = \exp(X\beta)$. Outra função de ligação que pode ser utilizada sob o modelo gama é a identidade $\eta = \mu$ que, também, não garante $\mu > 0$.

2.11.5 Estimação do parâmetro de dispersão

A matriz de covariância aproximada das estimativas dos parâmetros β é

$$Cov(\hat{\beta}) \simeq \sigma^2 (X^T W X)^{-1},$$

onde $W = diag\{(d\mu_i/d\eta_i)^2/V(\mu_i)\}$ é uma matriz diagonal $n \times n$ de pesos, X é a matriz modelo $n \times p$ e σ é o coeficiente de variação.

Se σ^2 é conhecido, a matriz de covariância de $\hat{\beta}$ pode ser calculada diretamente. Porém, na prática, σ^2 precisa ser estimado a partir do modelo ajustado.

Sob o modelo gama, a estimativa de máxima verossimilhança de $\nu = \sigma^{-2}$ é dada por

$$2n\{\log \hat{\nu} - \psi(\hat{\nu})\} = D(y; \hat{\mu}), \tag{2.23}$$

onde $\psi(\nu) = \Gamma'(\nu)/\Gamma(\nu)$ é a função digama.

Porém, se ν é suficientemente grande, a expressão acima pode ser expandida ignorando-se termos de ordem menor ou igual a ν^{-2}, obtendo-se, assim, uma expressão bem mais simples que pode ser usada como uma estimativa de máxima verossimilhança aproximada do parâmetro de dispersão:

$$\hat{\nu}^{-1} \simeq \frac{\bar{D}(6 + \bar{D})}{6 + 2\bar{D}}, \tag{2.24}$$

onde $\bar{D} = D(y; \hat{\mu})/n$.

Contudo, o principal problema de (2.23) e (2.24) é o fato de estarem baseadas na função desvio, pois $D(y; \hat{\mu})$ é igual a infinito quando alguma componente de y é zero. Além disso, se a suposição de distribuição gama for falsa, $\hat{\nu}^{-1/2}$ não é uma estimativa consistente para o coeficiente de variação.

Por estas razões, é aconselhável utilizar o estimador

$$\tilde{\sigma}^2 = \left\{ \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i \right\} / (n - p) = X^2/(n - p),$$

que é consistente para $\sigma^2 = \nu^{-1}$. Além disso, $\tilde{\sigma}^2$ apresenta um viés de ordem $O(n^{-1})$ se os dados são distribuídos como uma gama. O divisor $n - p$ é preferível a n, mas não é suficiente para redução do viés de $\tilde{\sigma}^2$.

2.12 Modelo Normal Inverso

2.12.1 A função densidade

A função densidade da normal inversa (ou Gaussiana inversa) $N^-(\mu, \phi)$ com média μ e parâmetro ϕ, representando o inverso de uma medida de dispersão, é dada por

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

As aplicações do modelo $N^-(\mu, \phi)$ envolvem estudo do movimento Browniano de partículas, análise de regressão com dados consideravelmente assimétricos, testes de confiabilidade, análise seqüencial e análogo da 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.

2.12.2 Principais características

As características do modelo são: função geratriz de momentos dada por

$$M(t; \mu, \phi) = \exp[\phi\mu^{-1}\{1 - (1 + 2\mu^2 t/\phi)^{1/2}\}].$$

Cumulantes para $r \geq 2$ obtidos de $\kappa_r = 1 \times 3 \times 5 \ldots (2r-1)\mu^{2r-1}\phi^{1-r}$. Coeficientes de assimetria e curtose iguais a $3\sqrt{\mu/\phi}$ e $(3+15\mu/\phi)$, respectivamente, e moda $\mu\left\{\left(\frac{1+9\mu^2}{4\phi^2}\right)^{1/2} - \left(\frac{3\mu}{2\phi}\right)\right\}$. Além disso, existe uma relação importante entre os momentos positivos e negativos dada por $E(Y^{-r}) = \frac{E(Y^{r+1})}{\mu^{2r+1}}$.

A distribuição acumulada da $N^-(\mu, \phi)$ pode ser obtida da $N(0,1)$ por

$$P(Y \leq y) = \Phi(y_1) + \exp(2\phi/\mu)\Phi(y_2),$$

onde $y_1 = (\phi/y)^{1/2}(-1 + y/\mu)$ e $y_2 = -(\phi/y)^{1/2}(1 + y/\mu)$.

A distribuição normal inversa, a gama, a log-normal e outras distribuições assimétricas, têm distribuição assintótica normal. Quando $\phi/\mu \to \infty$, $N^-(\mu, \phi)$ é assintoticamente $N(\mu, \mu^3/\phi)$.

Existem muitas analogias entre os modelos normal e normal inverso. Por exemplo, o dobro do termo do expoente com sinal negativo nas densidades normal e normal inversa, tem distribuição χ^2_1. Um estudo completo do modelo $N^-(\mu, \phi)$ é apresentado por Folks e Chhikara (1978).

2.13 Exercícios

1. Se $Y \sim P(\mu)$ demonstrar: (a) que o coeficiente de assimetria de $Y^{2/3}$ é de ordem μ^{-1} enquanto os de Y e $Y^{1/2}$ são de ordem $\mu^{-1/2}$; (b) que a log-verossimilhança para uma única observação é aproximadamente quadrática na escala $\mu^{1/3}$; (c) a fórmula do r-ésimo momento fatorial $E[Y(Y-1)\cdots(Y-r+1)] = \mu^r$; (d) que $2\sqrt{Y}$ é aproximadamente $N(0,1)$.
2. Sejam $y_i \sim B(n_i, p_i)$ e $x_i \sim B(m_i, q_i), i = 1, 2$. Mostre que a distribuição condicional de y_1 dado $y_1 + y_2 = m_1$ coincide com a distribuição condicional de x_1 dado $x_1 + x_2 = n_1$.
3. (a) Definir o algoritmo (2.10), calculando W, z e y^*, para os seguintes modelos com ligação potência $\eta = \mu^\lambda, \lambda$ conhecido: (i) normal; (ii) gama; (iii) normal inverso e (iv) Poisson;
   (b) Definir o algoritmo (2.10), calculando W, z e y^*, para o modelo binomial com ligação $\eta = \log\{[(1-\mu)^{-\lambda} - 1]\lambda^{-1}\}$, λ conhecido.
4. (a) Considere a estrutura linear $\eta_\ell = \beta x_\ell$, $\ell = 1 \ldots n$, com um único parâmetro β desconhecido e ligação $\eta = (\mu^\lambda - 1)\lambda^{-1}$, λ conhecido. Calcular a EMV de β para os modelos normal, Poisson, gama, normal inverso e binomial negativo. Fazer o mesmo para o modelo binomial com

ligação dada no exercício 3(b). Obter ainda as estimativas no caso de $x_1 = x_2 = \cdots = x_n$;
(b) Para os modelos citados acima, calcular as estimativas de MV de α e β, considerando a estrutura linear $\eta_\ell = \alpha + \beta\, x_\ell,\ \ell = 1 \ldots n$. Obter ainda a estrutura de covariância aproximada dessas estimativas.

5. Para as distribuições na família exponencial (2.1) mostre que $\kappa_3 = \kappa_2 \kappa_2'$ e $\kappa_4 = \kappa_2 \kappa_3'$ onde as derivadas são definidas em relação a μ.

6. Suponha que $Y \sim B(m, \mu)$ e que m é grande. Mostre que a variável aleatória $Z = \text{arcsen}\{(Y/m)^{1/2}\}$ tem, aproximadamente, os seguintes momentos:

$$E(Z) \doteq \text{arcsen}(\mu^{1/2}) - \frac{1 - 2\mu}{8\sqrt{m\mu(1-\mu)}}; \quad \text{Var}(Z) \doteq (4m)^{-1}.$$

7. Sejam as funções de probabilidade:

$$B(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad P(y) = \frac{e^{-\mu}\mu^y}{y!}.$$

Seja $\pi = \mu/m$. Mostre que, para μ fixo, quando $m - y \longrightarrow \infty$, temos:

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

8. Mostre que a distribuição gama tem função geratriz de cumulantes

$$K(t) = -\nu \log\left(1 - \frac{\mu t}{\nu}\right).$$

Assim, para ν grande, $\nu^{1/2}(Y - \mu)/\mu$ tem, aproximadamente, distribuição $N(0, 1)$.

9. Demonstre que a EMV do índice ν da distribuição gama é dada, aproximadamente, por

$$\nu \doteq \frac{6 + 2\overline{D}}{\overline{D}(6 + \overline{D})},$$

onde $\overline{D} = D(y; \hat{\mu})/n$ é o desvio médio.

10. Demonstrar que a ligação $\eta = \int b''(\theta)^{2/3} d\theta$ normaliza a distribuição de $\hat{\beta}$, tornando o seu coeficiente de assimetria, aproximadamente, zero.

11. Se $Y \sim B(m, \mu)$, demonstrar que a média e a variância de $\log[(Y + 1/2)/(m - Y + 1/2)]$ são $\log(\frac{\mu}{1-\mu}) + O(m^{-2})$ e $E\{(Y + 1/2)^{-1} + (m - Y + 1/2)^{-1}\} + O(m^{-3})$.

12. Caracterizar as distribuições log normal e log gama no contexto dos MLGs, definindo o algoritmo de ajustamento desses modelos com a ligação $\eta = \mu^{\lambda}$, λ conhecido.

13. 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 ligação raiz quadrada. Qual a grande vantagem deste último modelo?

14. Sejam Y_1 e Y_2 binomiais de parâmetros μ_1 e μ_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 $\mu_1, \mu_2, m_1, m_2, r$. Demonstrar que esta distribuição é um membro da família (2.1) com parâmetro $\theta = \log\{\mu_1(1 - \mu_2)/\mu_2(1 - \mu_1)\}$, $\phi = 1$ e $\mu = D_1(\theta)/D_0(\theta)$, onde $D_i(\theta) = \sum_x x^i \binom{m_1}{x} \binom{m_2}{r - x} \exp(\theta x)$ para $i = 0, 1$. Calcular a expressão do r-ésimo cumulante desta distribuição.

15. Se $Y \sim P(\mu)$ demonstrar: (a) que o coeficiente de assimetria de $Y^{2/3}$ é de ordem μ^{-1} enquanto aqueles de Y e $Y^{1/2}$ são de ordem $\mu^{-1/2}$; (b) que a log-verossimilhança para uma única observação é aproximadamente quadrática na escala $\mu^{1/3}$; (c) a fórmula do r-ésimo momento fatorial $E[Y(Y - 1) \cdots (Y - r + 1)] = \mu^r$; (d) a fórmula de recorrência entre os momentos centrais $\mu_{r+1} = r\mu\ \mu_{r-1} + \mu\ \partial\mu_r/\partial\mu$; (e) que $2\sqrt{Y}$ tem, aproximadamente, distribuição $N(0, 1)$.

16. Se $Y \sim G(\mu, \phi)$, demonstrar que: (a) quando $\phi > 1$ a densidade é zero na origem e tem uma única moda no ponto $\mu - \mu/\phi$; (b) a log-verossimilhança para uma única observação é, aproximadamente, quadrática na escala $\mu^{-1/3}$; (c) a variável transformada $3[(Y/\mu)^{1/3} - 1]$

é, aproximadamente, normal.

17. Sejam $Y_\ell \sim P(\mu_\ell)$, $\ell = 1 \ldots n$, observações 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 $\mu \to \infty$. Demonstrar que a variável aleatória $[f(Y_\ell) - f(\mu_\ell)]/\mu_\ell^{1/2} f'(\mu_\ell)$ converge em distribuição para a $N(0,1)$ quando $\mu_\ell \to \infty$. Provar ainda que a parte da log-verossimilhança que só depende dos $\mu_\ell' s$ tende assintoticamente para $-\frac{1}{2}\sum_{\ell=1}^{n}\{f(Y_\ell) - f(\mu_\ell)\}^2/Y_\ell f'(Y_\ell)^2$ quando $\mu_\ell \to \infty$, $\ell = 1 \ldots n$.
18. Se $Y \sim B(m, \mu)$, demonstrar que os momentos da estatística $Z = \pm\{2Y\ \log(Y/\mu) + 2(m-Y)\log[(m-Y)/(m-\mu)]\}^{1/2} + \{(1-2\mu)/[m\mu(1-\mu)]\}^{1/2}/6$ diferem dos correspondentes da $N(0,1)$ com erro $O(m^{-1})$.
19. A probabilidade de sucesso μ de uma distribuição binomial $B(m, \mu)$ depende de uma variável x de acordo com a relação $\mu = F(\alpha + \beta x)$, onde $F(\cdot)$ é uma função de distribuição acumulada especificada. Admite-se que para os valores $x_1 \ldots x_n$ de x, $m_1 \ldots m_n$ ensaios independentes foram realizados, sendo obtidas proporções de sucessos $y_1 \ldots y_n$, respectivamente. Comparar as estimativas $\hat{\alpha}$ e $\hat{\beta}$ para as escolhas de $F(\cdot)$: "probit", logística, arcsen $\sqrt{\ }$ e complemento $\log - \log$.
20. Sejam $y_1 \ldots y_n$ observações independentes e de mesma distribuição $G(\mu, \phi)$. Demonstrar que: (a) a estimativa de MV de ϕ satisfaz $\log \hat{\phi} - \psi(\hat{\phi}) = \log(\overline{y}/\tilde{y})$, onde $\overline{y}$ e $\tilde{y}$ são as médias aritmética e geométrica dos dados, respectivamente, e $\psi(\cdot)$ é a função digama; (b) uma solução aproximada para esta estimativa é dada por $\hat{\phi} = \overline{y}/2(\overline{y} - \tilde{y})$; (c) a variância assintótica de $\hat{\phi}$ iguala $\phi[\phi\psi'(\phi) - 1]^{-1}/n$.
21. Demonstrar que para os modelos normal e normal inverso supondo $\mu_1 = \cdots = \mu_n$, isto é, observações independentes e identicamente distribuídas, o desvio S_1 tem distribuição χ^2_{n-1}, supondo o modelo verdadeiro.
22. Demonstrar que para o modelo gama simples, em que todas as médias são iguais, o desvio reduz-se à estatística clássica $S_1 = 2n\phi\log(\overline{y}/\tilde{y})$, onde $\overline{y}$ e $\tilde{y}$ são, as médias aritmética e geométrica dos dados $y_1 \ldots y_n$, respectivamente.

Capítulo 3

Análise de Resíduos e Diagnóstico em Modelos Lineares Generalizados

3.1 Resíduos

Na modelagem estatística, a análise dos resíduos sempre se constitui numa das etapas mais importantes do processo de escolha do modelo estatístico. No contexto dos MLGs, os resíduos são usados para explorar a adequação do modelo ajustado com respeito à escolha da função de variância, da função de ligação e dos termos do preditor linear. Além disso, os resíduos são também úteis para indicar a presença de pontos aberrantes, que poderão ser influentes ou não. Os resíduos medem discrepâncias entre os valores observados $y_i's$ e os seus valores ajustados $\hat{\mu}_i's$.

3.1.1 Resíduo de Pearson

O resíduo de Pearson é definido por

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}.$$

O resíduo de Pearson recebe esse nome pois, para o modelo de Poisson, coincide com a raiz quadrada de uma componente da estatística de bondade de ajuste de Pearson $X^2 = \sum r_{P_i}^2$ (vide Seção 2.7.2).

A desvantagem deste resíduo é que sua distribuição apresenta-se, geralmente, bastante assimétrica para modelos não-normais.

3.1.2 Resíduo de Anscombe

Anscombe propôs, em 1953, uma definição para os resíduos usando uma função $A(y)$ ao invés de y, tal que $A(\cdot)$ é uma função escolhida visando tornar a distribuição de $A(Y)$ próxima à normal reduzida. Barndorff-Nielsen (1978) mostrou, em relação à família exponencial (2.1), que a função $A(\cdot)$ é dada por

$$A(\mu) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

Logo, o resíduo de Anscombe visando a normalização e estabilização da variância é expresso por

$$r_{A_i} = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}}.$$

Assim, para o modelo de Poisson, por exemplo, r_{A_i} é facilmente obtido e tem a seguinte forma

$$r_{A_i} = \frac{\frac{3}{2}(y_i^{2/3} - \hat{\mu}_i^{2/3})}{\hat{\mu}_i^{1/6}}.$$

Para o modelo gama, o resíduo de Anscombe é dado por

$$r_{A_i} = \frac{3(y_i^{1/3} - \hat{\mu}_i^{1/3})}{\hat{\mu}_i^{1/3}}.$$

3.1.3 Desvio residual

Se o desvio D é usado como uma medida de discrepância de um MLG, então, cada unidade de D contribui com uma quantidade

$$d_i = 2\nu_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\},$$

tal que $\sum_{i=1}^{n} d_i = D$. Com isso, surge uma nova definição de resíduo, a partir das componentes d_i que formam o desvio, conhecida como *desvio residual.*

Pregibon (1981) define o desvio residual como

$$r_{D_i} = sinal(y_i - \hat{\mu}_i)\sqrt{d_i},$$

pois, segundo ele, se existe uma transformação que normalize a distribuição do resíduo, então as raízes quadradas das componentes do desvio são resíduos que exibem as mesmas propriedades induzidas por esta transformação. Assim, os resíduos r_{D_i} podem ser tratados como variáveis aleatórias tendo aproximadamente distribuição normal reduzida e, conseqüentemente, $r_{D_i}^2 = d_i$ tem, aproximadamente, distribuição χ_1^2.

Assim, por exemplo, para o modelo de Poisson, temos

$$r_{D_i} = sinal(y_i - \hat{\mu}_i)\{2[y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i]\}^{1/2}.$$

Além disso, é importante enfatizar que diversas anomalias prejudiciais ao modelo são verificadas através de análises gráficas utilizando o resíduo de Anscombe e o desvio residual, dentre as quais podemos citar: falsa distribuição populacional atribuída à variável dependente Y, verificação das funções de variância e de ligação, entre outras.

3.1.4 Comparação entre os resíduos

Para o modelo normal nenhuma distinção é observada entre os três tipos de resíduos. Entretanto, o resíduo de Anscombe e o desvio residual apresentam formas funcionais muito diferentes para modelos não-normais, mas seus

valores são bastante próximos para modelos bem ajustados. O resíduo de Pearson difere em forma e valor destes dois últimos. Podemos verificar isso, considerando novamente o modelo de Poisson e fazendo $y = c\mu$ (c uma constante). Temos, a seguir, as formas funcionais para os três tipos de resíduos:

$$r_P = \hat{\mu}^{1/2}(c-1), \qquad r_A = \frac{3}{2}\hat{\mu}^{1/2}(c^{2/3}-1)$$

e

$$r_D = sign(c-1)\hat{\mu}^{1/2}[2(c\log c - c + 1)]^{1/2}.$$

Na Tabela 3.1 fazemos uma comparação entre os três resíduos citados acima para diversos valores de c.

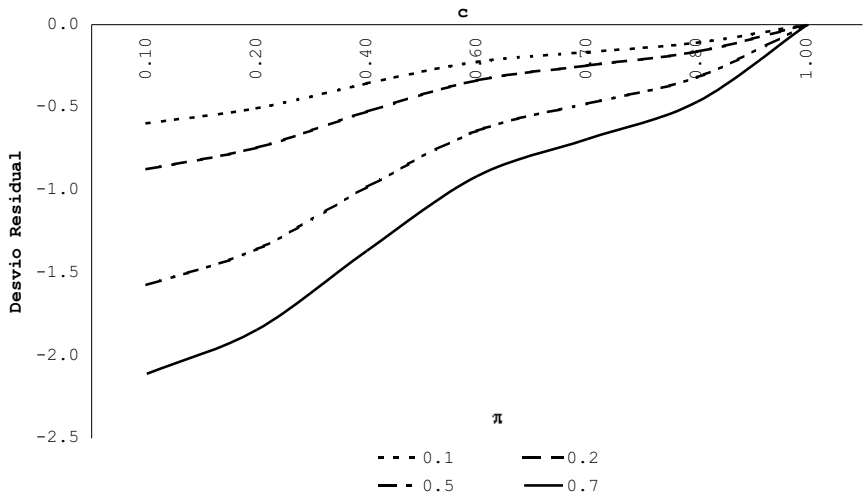
Tabela 3.1: *Comparação entre os resíduos para o modelo de Poisson*

	r_A	r_D	r_P
c	$\frac{3}{2}(c^{2/3}-1)$	$sinal(c-1)[2(c\log c - c + 1)]^{1/2}$	$(c-1)$
0.0	-1.5	-1.414	-1.0
0.2	-0.987	-0.956	-0.8
0.4	-0.686	-0.683	-0.6
0.6	-0.433	-0.432	-0.2
1.0	0.0	0.0	0.0
1.5	0.466	0.465	0.5
2.0	0.881	0.879	1.0
2.5	1.263	1.258	1.5
3.0	1.620	1.610	2.0
4.0	2.280	2.256	3.0
5.0	2.886	2.845	4.0
10.0	5.462	5.296	9.0

Nota-se que a diferença máxima entre r_A e r_D ficou em apenas 6%, registrada em $c = 0$. O resíduo de Pearson apresentou uma diferença considerável, para a grande maioria dos valores de c, em relação aos resíduos r_A e r_D. Deve-se ressaltar, porém, que para modelos mal ajustados e/ou para observações aberrantes, podem ocorrer diferenças consideráveis entre estes resíduos. Os valores de r_A e r_D são, também, bastante próximos para os modelos gama e normal inverso.

Para o modelo binomial, fazemos $y = c\ m\ \pi$, onde m representa o número de ensaios de Bernoulli, π a probabilidade de sucesso e c encontra-se no intervalo unitário devido às restrições (i) $\log c > 0$ e (ii) $\log(1 - c\pi) > 0$, provenientes da expressão para o desvio residual. Podemos observar na Figura 3.1 que o desvio residual apresenta-se menor que o resíduo de Pearson, independente do valor de π. Quando o valor de c se aproxima de 1, a diferença entre os resíduos diminui. Além disso, tanto para o desvio residual quanto para o resíduo de Pearson, à medida que π cresce o resíduo também aumenta (vide Figura 3.2).

Figura 3.1: *Desvio Residual*

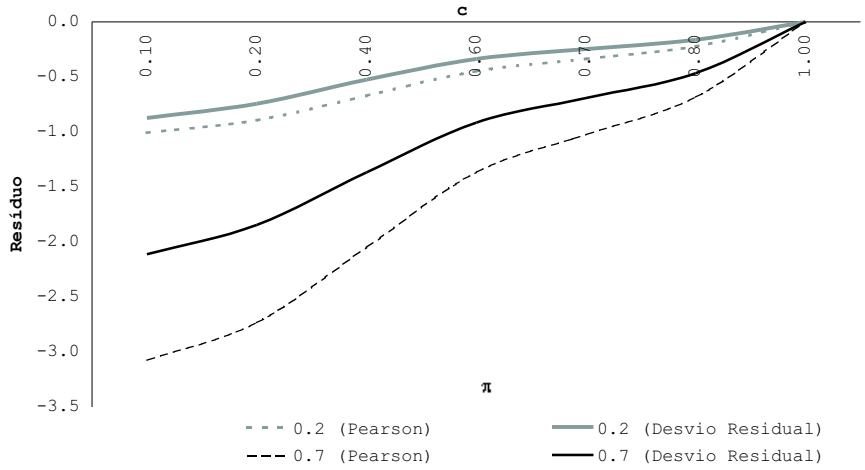


Os resultados das Figuras 3.1 e 3.2 foram obtidos considerando $m = 5$.

Entretanto, também foi analisado o comportamento quando $m = 7$ e 10. Para estes valores não houve mudanças nas conclusões e nos resultados apresentados anteriormente.

Pierce e Schafer (1986) examinam de forma mais extensiva as definições de resíduos em modelos da família exponencial.

Figura 3.2: *Comparação Entre Resíduos de Pearson e Desvio Residual*



3.2 Análise Residual e Medidas de Influência

Na escolha de um modelo estatístico a análise residual desempenha um papel muito importante. No contexto dos MLGs, os resíduos são amplamente utilizados para:

- verificar a adequação do ajustamento do modelo aos dados;
- identificar outliers e pontos influentes;
- verificar se um nova covariável pode ser introduzida no modelo;
- verificar as funções de ligação e de variância;
- avaliar a distribuição do erro aleatório.

Neste capítulo serão apresentados métodos e procedimentos relativos aos itens descritos aqui.

3.2.1 O resíduo de Cox-Snell e o desvio residual

Segundo um modelo estatístico arbitrário, Cox e Snell (1968) expressam um vetor aleatório n-dimensional Y em termos de um vetor $\beta \in \mathbf{R}^p$ de parâmetros desconhecidos e de um vetor ε de variáveis aleatórias i.i.d. não-observadas.

Supondo que cada observação y_i da componente aleatória Y_i do vetor Y depende apenas de um erro aleatório ε_i, podemos escrever de uma forma geral

$$y_i = g_i(\beta, \varepsilon_i), \qquad i = 1, \ldots, n.$$

Seja $\hat{\beta}$ a EMV de β. Suponha que a equação

$$y_i = g_i(\hat{\beta}, v_i), \qquad i = 1, \ldots, n,$$

tem como única solução

$$v_i = h_i(y_i, \hat{\beta}), \qquad i = 1, \ldots, n.$$

Então, v_i é definido como *resíduo generalizado*. No caso de variáveis aleatórias contínuas, uma definição conveniente para v_i pode ser obtida por:

$$v_i = \Phi^{-1}(F(y_i; \hat{\beta})), \qquad i = 1, \ldots, n, \tag{3.1}$$

onde $F(\cdot)$ é a função de distribuição da variável aleatória Y.

A equação (3.1) é conhecida como *resíduo de Cox-Snell* e $\Phi^{-1}(\cdot)$ é a inversa da função de distribuição acumulada da normal padrão.

A definição de desvio residual, proposta primeiramente por Pregibon (1981) no contexto dos MLGs, é desenvolvida de forma diferente do *resíduo de Cox-Snell* e pode ser aplicada a qualquer modelo estatístico.

Segundo Pregibon, seja Y o vetor aleatório n-dimensional definido anteriormente e $\theta \in \mathbf{R}^n$ um vetor de parâmetros desconhecidos. Então, podemos

expressar a observação y_i em termos dos parâmetros β_r's que pertencem a um subconjunto Θ_1 do espaço paramétrico Θ, isto é, $\theta_i = \theta_i(\beta)$, onde $\dim(\beta) = p < n$.

Para testar a hipótese $H_0 : \theta \in \Theta_1$ versus a alternativa $H_A : \theta \in \Theta$, onde $\Theta_1 \subset \Theta$, pode-se usar a razão de verossimilhanças

$$D = 2\left[\sup_{\theta\in\Theta} l(\theta; y) - \sup_{\theta\in\Theta_1} l(\theta; y)\right],$$

onde $l(\theta; y)$ é a log-verossimilhança dos parâmetros em θ supondo os dados y. Assim, temos uma medida de discrepância entre o modelo saturado (quando $\theta \in \Theta$) e o modelo restrito (quando $\theta \in \Theta_1$).

Suponha que os Y_i's são independentes, que $\tilde{\theta}$ é a EMV de θ segundo o modelo saturado, e que $\hat{\theta} = \theta(\hat{\beta})$ é a EMV de θ segundo o modelo restrito. Então, podemos escrever,

$$D = 2\sum_{i=1}^{n}[l_i(\tilde{\theta}_i; y_i) - l_i(\hat{\theta}_i; y_i)], \tag{3.2}$$

onde a quantidade (3.2) é o *desvio* do modelo. No caso do MLG, o desvio está definido na Seção 2.7.1.

Finalmente, Pregibon (1981) definiu o *desvio residual* como

$$r_D(y_i, \hat{\theta}_i) = sinal(\tilde{\theta}_i - \hat{\theta}_i)\sqrt{2[l_i(\tilde{\theta}_i; y_i) - l_i(\hat{\theta}_i; y_i)]}, \tag{3.3}$$

e demonstrou que, se existe uma transformação que normalize a distribuição dos resíduos, então as raízes quadradas das componentes do desvio são resíduos que exibem as mesmas propriedades induzidas por esta transformação.

Deve-se ressaltar que o desvio residual vale em qualquer modelo estatístico e não apenas no contexto dos MLGs. A expressão (3.3) mede a discrepância entre o modelo saturado e o modelo restrito com relação à observação y_i.

3.2.2 Situações assintóticas

É importante salientar a diferença entre dois tipos de convergência assintótica: (i) quando o número de observações torna-se grande, indicada por "$n \to \infty$"; (ii) quando cada componente Y_i torna-se aproximadamente normal, esta última indicada por "$m \to \infty$", onde m pode representar, por exemplo, a média da Poisson, o parâmetro de forma da gama, os graus de liberdade da distribuição t de Student, etc. Em todos estes casos, quando $m \to \infty$, a distribuição de Y pode ser considerada aproximadamente normal.

Em nosso contexto, a principal consequência quando $n \to \infty$ é que a EMV $\hat{\theta}$ converge para θ independentemente de m. Por outro lado, quando $m \to \infty$, a distribuição da variável aleatória Y converge para a distribuição normal e, assim, $r_D(y_i, \hat{\theta}_i)$ converge para $r_D(y_i, \theta_i)$, que equivale a expressão (3.3) com θ_i no lugar de $\hat{\theta}_i$, independente do valor de n.

3.2.3 Correção de viés para o desvio residual

Quando $m \to \infty$, o desvio definido em (3.2) é assintoticamente distribuído como χ^2 com $n - p$ graus de liberdade (p corresponde a dimensão do espaço paramétrico Θ_1 sob a hipótese nula). Barndorff-Nielsen (1986) e McCullagh (1984) mostram que o desvio residual pode ser re-centrado e re-escalonado de tal forma que sua distribuição assintótica seja normal padrão até ordem $\mathrm{O}_p(m^{-3/2})$.

Quando a distribuição de Y pertence à família exponencial (2.1), temos a *função geratriz de momentos* de Y dada por

$$M_Y(t; \theta, \phi) = \exp\left\{\left[\frac{b(t\ a(\phi) + \theta) - b(\theta)}{a(\phi)}\right]\right\}. \tag{3.4}$$

Por conseguinte, a *função geratriz de cumulantes* de Y é

$$\log M_Y(t; \theta, \phi) = \frac{b(t\ a(\phi) + \theta) - b(\theta)}{a(\phi)}. \tag{3.5}$$

Logo, a fórmula geral do cumulante de ordem r de Y é

$$\kappa_r = \frac{b^{(r)}(\theta)}{a(\phi)^{1-r}}. \tag{3.6}$$

A equação (3.6) é obtida derivando-se (3.5) r vezes em relação a t e calculando a equação resultante no ponto $t = 0$.

Com isso, o termo $\rho_3(\theta)$ que representa o terceiro cumulante padronizado de Y é dado por

$$\rho_3(\theta) = E_\theta \left\{ \left[\frac{Y - \mu}{V(\mu)^{1/2}} \right]^3 \right\},$$

ou seja,

$$\rho_3(\theta) = \frac{\kappa_3}{\kappa_2^{3/2}}.$$

Em particular, McCullagh e Nelder (1983) sugerem adicionar o termo $\rho_3(\theta)/6$ na expressão do desvio residual com objetivo de remover o viés de ordem $\mathrm{O}(m^{-1/2})$ da média assintótica de r_D. Assim, o termo $\rho_3(\theta)/6$ é conhecido como *correção do viés* do desvio residual.

Finalmente, temos a expressão

$$r_{AD}(y, \theta) = r_D(y, \theta) + \rho_3(\theta)/6 \tag{3.7}$$

representando o *desvio residual ajustado*, que tem distribuição aproximadamente normal até ordem $\mathrm{O}_p(m^{-1})$.

Temos, na Tabela 3.2, os valores da correção de viés para algumas distribuições de interesse.

Tabela 3.2: *Correção de Viés*

Distribuição	$\rho_3(\theta)/6$
Gama (m, λ)	$1/(3\sqrt{m})$
t de Student (m)	0
Logística (μ, σ)	0
Laplace (μ)	0
Binomial (m, p)	$\frac{(1-2p)}{6\sqrt{mp(1-p)}}$
Poisson (m)	$1/(6\sqrt{m})$

Note que $\rho_3(\theta) = 0$ para as distribuições t de Student, logística e Laplace por causa da simetria de suas respectivas funções densidades. Para maiores detalhes sobre a normalidade assintótica do desvio residual, vide Pierce e Schafer (1986).

3.3 Verificação da Distribuição dos Resíduos

3.3.1 Teste de normalidade

Como foi previamente apresentado, no caso de variáveis aleatórias contínuas, podemos construir os resíduos de Cox e Snell a partir de uma transformação na distribuição de probabilidade $F_Y(y;\theta)$ de Y. Seja uma variável aleatória com parâmetros conhecidos, $U = F_Y(Y;\theta)$, uniformemente distribuída no intervalo unitário. Se $\Phi(\cdot)$ denota a função de distribuição de uma variável aleatória normal padrão, então

$$V = \Phi^{-1}(F_Y(Y;\theta)) = \Phi^{-1}(U)$$

tem distribuição normal padrão. Assim, assumindo θ conhecido, temos o *resíduo de Cox e Snell* dado por $v_i = \Phi^{-1}(F_Y(y_i;\theta))$ e o *desvio residual* por $r_D(y_i;\theta)$. Note-se que, na prática, o parâmetro verdadeiro não é conhecido, devendo ser substituído pela sua EMV.

No trabalho de Green (1984), Davison sugere que se $F^{-1}(\cdot;\theta)$ é conhecida, então a variável aleatória

$$G(V) = r_D(F^{-1}(\Phi(V);\theta),\theta) \tag{3.8}$$

e V podem ser comparadas.

Por exemplo, no caso em que $Y \sim N(0,1)$, temos $F_Y(y;\theta) = \Phi(y)$ e $v = y$. Conseqüentemente, $G(v) = v$, ou seja, os resíduos de Cox e Snell e o desvio residual coincidem neste caso particular. Entretanto, quando Y segue uma distribuição gama ou Weibull, por exemplo, os resíduos de Cox e Snell e o desvio residual não coincidem. Gigli (1987, Cap. 2) mostra tais resultados para outras distribuições de interesse, além destas citadas anteriormente.

Com isso, no caso onde $G(v) = v$, um gráfico de $G(v) \times v$ (conhecido por $G(v)$ *plot*) produziria uma reta de gradiente 1 passando pela origem. Isso poderia ser interpretado como um gráfico de normalidade para o desvio residual onde, no eixo das abscissas estão os quantis da normal padrão, enquanto que no eixo das ordenadas temos o desvio residual ordenado.

No caso geral, quando Y tem uma distribuição $F = F_Y(y_i;\theta)$ qualquer, ainda sabemos que V é normalmente distribuído e o gráfico de $G(v)$ versus v pode continuar sendo interpretado como um gráfico dos desvios residuais versus as estatísticas de ordem da distribuição normal. Assim, caso os pontos estejam em torno de uma reta de gradiente 1 passando pela origem, podemos considerar o desvio residual para a distribuição F como sendo aproximadamente normal.

A partir de (3.3), temos

$$G(v) = sinal(\tilde{\theta} - \hat{\theta})\sqrt{2[l(\tilde{\theta}; F^{-1}(\Phi(v);\tilde{\theta})) - l(\hat{\theta}; F^{-1}(\Phi(v);\hat{\theta}))]}. \tag{3.9}$$

Gigli (1987, Cap.2) apresenta $G(v)$ em termos da expressão (3.9) para diversas distribuições. Temos, por exemplo, quando $Y \sim Gama(m,\lambda)$

$$G(v) = sinal\left(\frac{y}{2} - m\right)\sqrt{2\left(m\log m - m\log\frac{y}{2} + \frac{y}{2} - m\right)}.$$

Se $F_Y(y;\theta)$ está bem definida e é facilmente inversível, então $G(v)$ é apenas uma função de v. De outro modo seria necessário utilizarmos uma aproximação numérica para encontrar $F^{-1}(\cdot;\theta)$ e, assim, inseri-la em $G(v)$.

Gigli (1987) também apresenta uma expressão aproximada para $G(v)$ através da expansão de $G(v)$ em série de Taylor em torno de $v = 0$. Esta aproximação deve ser utilizada quando F^{-1} não apresentar uma forma fechada.

Quando

$$v_0 = 0$$

temos

$$u_0 = \Phi(v_0) = \frac{1}{2} \text{ e } \; y_0 = F^{-1}(u_0) = y_m,$$

onde y_m é a mediana da distribuição. Se $f_Y(y;\theta)$ é a função densidade da variável aleatória Y e $r'_D(y,\theta)$, $r''_D(y,\theta)$, $r'''_D(y,\theta)$ são, respectivamente, a primeira, a segunda e a terceira derivadas de $r_D(y,\theta)$ em relação a y, temos que:

$$\begin{aligned}
G(0) &= r_D(y_m,\theta) \\
G'(0) &= \frac{1}{\sqrt{2\pi}} \frac{r'_D(y_m,\theta)}{f(y_m;\theta)} \\
G''(0) &= \frac{1}{\sqrt{2\pi}} \left\{ \frac{1}{[f(y_m;\theta)]^2} \; r''_D(y_m,\theta) - \frac{f'(y_m;\theta)}{[f(y_m;\theta)]^3} \; r'_D(y_m,\theta) \right\} \\
G'''(0) &= r'_D(y_m,\theta) \left\{ \frac{-1}{\sqrt{2\pi} f(y_m;\theta)} + \frac{3}{(\sqrt{2\pi})^3} \frac{[f'(y_m;\theta)]^2}{[f(y_m;\theta)]^5} \right. \\
&\quad \left. - \frac{1}{\left(\sqrt{2\pi}\right)^3} \frac{f''(y_m;\theta)}{[f(y_m;\theta)]^4} \right\} - \frac{3}{\left(\sqrt{2\pi}\right)^3} \frac{f'(y_m;\theta)}{[f(y_m;\theta)]^4} r''_D(y_m,\theta) \\
&\quad + \frac{1}{\left(\sqrt{2\pi}\right)^3} \frac{1}{[f(y_m;\theta)]^3} \; r'''_D(y_m,\theta)
\end{aligned}$$

e

$$G(v) = G(0) + G'(0)v + \frac{1}{2}G''(0)v^2 + \frac{1}{6}G'''(0)v^3 + O_p(v^4). \tag{3.10}$$

Assim, como foi dito anteriormente, caso o gráfico de $G(v)$ versus v seja

aproximadamente linear, temos a confirmação do quão próximo o desvio residual está do resíduo de Cox e Snell.

Gigli (1987, Cap. 2) utiliza o gráfico $G(v) \times v$ para testar a normalidade em diversas distribuições discretas e contínuas, tais como: gama, Weibull, logística, Laplace, Poisson, binomial, geométrica, etc.

3.3.2 Erro de classificação na distribuição dos dados

Nesta seção trataremos da situação em que os dados pertencem a uma certa distribuição (verdadeira), porém o investigador ajusta um modelo supondo uma distribuição falsa. Iremos nos restringir apenas ao caso em que o parâmetro de interesse é escalar, pois o caso vetorial é bastante complicado.

Suponha que Y é um vetor de variáveis aleatórias independentes pertencente a uma distribuição (verdadeira) $H(\cdot;\alpha)$. Contudo, assumimos que $Y \sim F(\cdot;\beta)$. Seja $l_F(\beta;y_i)$ a log-verossimilhança associada com a distribuição F e $l_H(\alpha;y_i)$ a log-verossimilhança associada com a distribuição H. Assim, podemos definir

$$l_F(\beta;y) = \sum_{i=1}^{n} l_F(\beta;y_i(\alpha)),$$

onde cada y_i depende de α, pois a distribuição verdadeira de Y é $H(\cdot;\alpha)$.

Note-se que a solução da equação

$$\frac{\partial l_F(\beta;y)}{\partial \beta} = 0 \tag{3.11}$$

determina $\tilde{\beta}$, a EMV irrestrita de β, que é função de α pois a distribuição verdadeira de Y é $H(\cdot;\alpha)$.

Na equação

$$E_\alpha\left[\frac{\partial l_F}{\partial \beta}|_{\beta_\alpha}\right] = 0 \tag{3.12}$$

temos β_α como função de α. A esperança em (3.12) é calculada supondo a distribuição verdadeira $H(\cdot;\alpha)$ para Y.

De acordo com a notação utilizada anteriormente, considere o seguinte exemplo: seja a variável aleatória Y com distribuição de Poisson $F(\beta) \equiv P(\beta)$ com média $E(Y) = \beta$, a falsa distribuição assumida pelo investigador. Enquanto isso, supõe-se que $H(\alpha) \equiv G(\alpha)$, distribuição geométrica com média $E(Y) = \frac{\alpha}{1-\alpha}$, a distribuição verdadeira dos dados. Temos que

$$\frac{\partial l_F(\beta; y)}{\partial \beta} = -1 + \frac{y}{\beta}.$$

A EMV irrestrita de β é $\tilde{\beta} = y$, independente de α.

Pela equação (3.12)

$$E_\alpha \left[\frac{\partial l_F}{\partial \beta}|_{\beta_\alpha} \right] = E_\alpha \left[-1 + \frac{y}{\beta} \right] \Rightarrow \hat{\beta}_\alpha = E_\alpha(y).$$

Como $E_\alpha(y) = \frac{\alpha}{1-\alpha}$, pois a esperança é calculada supondo a distribuição verdadeira $H(\cdot; \alpha)$, então

$$\hat{\beta}_\alpha = \frac{\alpha}{1-\alpha}.$$

A partir da expressão (3.8) e das soluções encontradas nas equações (3.11) e (3.12), Gigli (1987) propõe um procedimento gráfico para detectar erros de classificação na distribuição dos dados.

Seja

$$G(v) = r_D(y, \beta) = r_D(F^{-1}(\Phi(v); \beta), \beta),$$

onde F é a função de distribuição de Y, assumida pelo investigador, que depende apenas de β. Define-se uma nova função

$$G_H(v) = r_D(y; \beta_\alpha) = r_D(H^{-1}(\Phi(v); \alpha),\ \beta_\alpha). \tag{3.13}$$

Utiliza-se o seguinte procedimento para o cálculo de $G_H(v)$:
• fixa-se o valor do parâmetro α e calcula-se y, isto é, $y = H^{-1}(\Phi(v); \alpha)$;
• encontra-se $\tilde{\beta}$, a EMV irrestrita de β sob a distribuição F, resolvendo (3.11);
• encontra-se $\hat{\beta}_\alpha$ resolvendo (3.12);
• calcula-se $r_D(y, \beta_\alpha)$ a partir da definição do desvio residual da distribuição de F.

Finalmente, compara-se o gráfico $G(v)$ versus v, definido na Seção 3.3.1, com o gráfico $G_H(v)$ versus v. Caso seja visualizada alguma diferença entre os dois gráficos, podemos concluir que a distribuição F, assumida pelo investigador, tem uma maior chance de não ser a distribuição verdadeira de Y.

3.4 Verificando a Inclusão de uma Nova Covariável

Seja $y = (y_1, \ldots, y_n)^T$ um vetor $n \times 1$ de respostas com distribuição pertencente à família exponencial (2.1) e $X = (x_1, \ldots, x_p)$ a matriz modelo $n \times p$ correspondendo a p variáveis explicativas. Seja, ainda, $\eta_i = g(\mu_i) = x_i^T \beta$ o preditor linear, onde $g(\cdot)$ é a função de ligação, $\beta = (\beta_1, \ldots, \beta_p)^T$ um vetor $p \times 1$ de parâmetros desconhecidos e x_i^T a i-ésima linha de X. Temos, assim, um certo MLG de interesse.

Wang (1985) sugere um procedimento para testar se uma nova covariável $z = (z_1, \ldots, z_n)^T$ pode ser incorporada ao modelo em investigação. Para isso, basta verificar se o preditor linear

$$\eta = X\beta$$

pode assumir a seguinte forma

$$\eta = X\beta + \gamma z,$$

onde γ é um escalar.

Note que a EMV $\hat{\beta}$ de β equivale a considerar a hipótese de que $\gamma = 0$. Sejam as definições usuais

$$\begin{aligned} Var(Y_i) &= a(\phi)V_i, \\ W &= diag\left\{V_i^{-1}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\right\}, \\ H &= W^{1/2}X(X^TWX)^{-1}X^TW^{1/2} \text{ e} \end{aligned}$$

$r = r_P$ o vetor dos resíduos de Pearson generalizados, dados na Seção 3.1.1,

cujo i-ésimo elemento corresponde a $\frac{(y_i-\hat{\mu}_i)}{\hat{V}_i^{1/2}}$. Considerando todos os termos citados acima calculados em $\hat{\beta}$, Wang (1985) sugere o seguinte método para verificar se a variável z deve ser adicionada ao modelo: construir um gráfico de r_P versus $(I-\hat{H})\hat{W}^{1/2}z$, onde W e H estão avaliados em $\hat{\beta}$, e verificar se o mesmo é aproximadamente linear. Se for linear, a variável z será incorporada à matriz modelo.

De acordo com Wang (1985), este procedimento é equivalente a usar a estatística escore

$$\frac{(r_P^T z)^2}{z^T\hat{W}^{1/2}(I-\hat{H})\hat{W}^{1/2}z},$$

para testar a hipótese de que $\gamma = 0$. Esta estatística deve ser comparada ao valor crítico da distribuição χ_1^2.

3.5 Verificando a Não-Linearidade em um Sub-Conjunto de Variáveis Explicativas

Considere, sem perda de generalidade, que as últimas $p-q$ $(p > q)$ variáveis da matriz modelo X são não-lineares, de tal forma que podemos particionar X como $X = (X_1, X_2)$, onde X_2 é formada pelas referidas variáveis com suspeita de não-linearidade. Por simplicidade, considera-se as transformações possíveis à X_2 dentro da família de transformações propostas por Box e Cox (1964) e expressas por

$$X_2^{(\lambda)} = \begin{cases} (X_2^\lambda - 1)/\lambda, & \text{se} \quad \lambda \neq 0 \\ \log(X_2), & \text{se} \quad \lambda = 0. \end{cases} \tag{3.14}$$

Para verificar a não-linearidade nas variáveis contidas em X_2, segundo Wang (1987), deve-se testar a hipótese $H_0 : \lambda = 1$ no MLG com preditor linear $\eta = X_1(\beta_1, \ldots, \beta_q)^T + X_2^{(\lambda)}(\beta_{q+1}, \ldots, \beta_p)^T$.

Utilizando uma expansão linear em série de Taylor de η, podemos aprox-

imar $X_2^{(\lambda)}$ localmente por

$$X_2^{(\lambda)} + (\lambda - 1)U^{(1)},$$

onde $U^{(\lambda)} = \frac{\partial X_2^{(\lambda)}}{\partial \lambda}$.

Consequentemente, η_i pode ser aproximado por

$$x_i^T \beta + \gamma z_i, \tag{3.15}$$

onde

$$z = (z_1, \ldots, z_n)^T = U^{(1)}(\beta_{q+1}, \ldots, \beta_p)^T$$

e

$$\gamma = (\lambda - 1).$$

Note que, sob a hipótese nula H_0, a EMV $\hat{\beta}$ de β, em (3.15), é obtida pelo método de Newton-Raphson citado na Seção 2.4. Então, podemos calcular z a partir de $\hat{\beta}$. A covariável adicional z deve ser tratada como uma "*constructed variable*" (variável construída) para X_2.

Wang (1987) propõe a construção de um gráfico de r_P versus $(I - \hat{H})\hat{W}^{1/2}z$, onde r_P, $\hat{H}$ e $\hat{W}$ estão dados na Seção 3.4. Este tipo de gráfico é conhecido como "*constructed variable plot*" e $(I - \hat{H})\hat{W}^{1/2}z$ são os "*constructed residuals*" (resíduos construídos) para X_2.

A presença de uma tendência linear neste gráfico indica que $\gamma \neq 0$, ou seja, $\lambda \neq 1$. A ausência de uma tendência linear neste gráfico ($\lambda = 1$) indica que as variáveis contidas em X_2 são lineares para o MLG. Segundo Wang (1987), a estimativa $\hat{\lambda}$ de λ, dada por $1 + \hat{\gamma}$, pode ser obtida através de uma regressão linear de r sobre $(I - \hat{H})\hat{W}^{1/2}z$ e deve ser utilizada em (3.14) com o objetivo de linearizar X_2.

A estatística escore

$$\frac{(r_P^T z)^2}{z^T \hat{W}^{1/2}(I - \hat{H})\hat{W}^{1/2}z},$$

citada na Seção 3.4, também pode ser empregada para testar a hipótese

$H_0 : \gamma = 0$. Através dela podemos interpretar o grau de importância que a trasformação de Box e Cox exerce para linearizar X_2.

3.6 Verificando a Função de Ligação e de Variância

Em relação à *função de ligação*, um procedimento informal consiste na construção de um gráfico entre a variável dependente ajustada y^* e $\hat{\eta}$. Se o gráfico for aproximadamente linear, a função de ligação estará correta. Deve-se ressaltar que para dados binários este gráfico é não-informativo, sendo necessário o uso de métodos formais.

Dentre os procedimentos formais, o método proposto por Hinkley (1985) é bastante utilizado na prática. Consiste em adicionar $\hat{\eta}^2$ como uma nova covariável na matriz modelo. Se isto causar uma redução significativa no desvio, a função de ligação não é adequada. Para verificar se a redução é estatisticamente significante, pode-se utilizar o teste proposto na Seção 2.7.3.

Uma estratégia informal para verificar a adequação da *função de variância* seria construir um gráfico dos resíduos absolutos versus os valores ajustados. Caso os pontos estejam dispersos sem uma tendência (local ou global) definida, podemos considerar a função de variância adequada. Entretanto, uma tendência positiva indica que a variância está crescendo de acordo com a média. Com isso, a escolha inicial de $V(\mu) \propto \mu$ pode ser substituída por $V(\mu) \propto \mu^2$. Entretanto, uma tendência negativa indica o efeito inverso.

3.7 Correção de Continuidade Residual no Modelo Logístico

Nos últimos anos, inúmeros trabalhos têm sido publicados abordando o comportamento residual em regressão logística, dentre os quais podemos destacar: Cox e Snell (1968), Pregibon (1981), Landwehr, Pregibon e Shoemaker (1984), Jennings (1986), Copas (1988) e McCullagh e Nelder (1989).

Em particular, Pierce e Schafer (1986) sugerem uma correção de con-

tinuidade para os resíduos argumentando que $R_*(y_i \pm 1/2, p_i)$ apresenta melhor normalidade que $R_*(y_i, p_i)$, onde $* \in$ {resíduos de Pearson, Anscombe, desvio residual e desvio residual ajustado} (vide Seções 1.6 e 3.1.3). Além disso, segundo eles, quando a estimativa $\hat{p}_i$ encontra-se próxima do parâmetro p_i (desconhecido na prática), este mesmo comportamento é esperado pelos resíduos calculados a partir de $\hat{p}_i$.

Entretanto, Duffy (1990) apresenta evidência contra o uso da correção de continuidade nos resíduos em regressão logística. Através de uma análise gráfica informal, a autora conclui que a correção de continuidade age de forma a prejudicar a normalidade dos resíduos no modelo logístico. Duffy também testa a habilidade dos resíduos em detectar observações contaminadas ou outliers. Novamente, o uso da correção de continuidade prejudica a identificação de tais observações a partir dos resíduos.

Para uma análise mais detalhada sobre os problemas da correção de continuidade em modelos de regressão logística, vide Duffy (1990).

3.8 Detectando Pontos de Influência

3.8.1 Medidas de alavancagem

A idéia básica sobre os pontos de influência e de alavancagem consiste em verificar a dependência do modelo estatístico sobre as várias observações que foram coletadas e ajustadas. Tais pontos exercem um papel importante no ajuste final dos parâmetros de um modelo estatístico, ou seja, sua exclusão pode implicar mudanças substanciais dentro de uma análise estatística.

No modelo linear de regressão uma medida de alavancagem é dada pelos elementos da diagonal da matriz

$$H = X(X^T X)^{-1} X^T,$$

conhecida como matriz de projeção ou matriz *hat.*

No contexto dos MLGs, as observações conhecidas como pontos de alavancagem podem ser detectadas pelos elementos h_{ii} da matriz *hat generalizada*, definida por

$$\hat{H} = \hat{W}^{1/2} X (X^T \hat{W} X)^{-1} X^T \hat{W}^{1/2}, \tag{3.16}$$

onde $\hat{W}$ é o valor de W em $\hat{\beta}$.

Espera-se que as observações distantes do espaço formado pelas variáveis explicativas apresentem valores apreciáveis de h_{ii}. Como H é matriz de projeção, $0 \leq h_{ii} \leq 1$, vide Seção 1.9.1 para uma demostração similar. Além disso, $tr(H) = \text{posto}(H) = p$.

Hoalgin e Welsh (1978) sugerem usar $h > 2p/n$ para indicar os pontos de alavancagem. Uma ferramenta informal para visualizar tais observações consiste em usar um "*index plot*" (gráfico indexado) dos h_{ii} versus i com limite $h = 2p/n$.

3.8.2 Medidas de influência

Segundo Lee (1987), a informação de alavancagem contida em h_{ii} reflete *parcialmente* a influência de uma observação. Para verificar a completa influência da i-ésima observação, levando-se em consideração aspectos como: estimativas dos parâmetros, valores ajustados, estatísticas de bondade de ajuste, etc., torna-se necessário a comparação entre as estimativas $\hat{\beta}$ e $\hat{\beta}_{(i)}$, esta última obtida quando a referida observação é deletada. Davison e Snell (1991) propõem o uso da seguinte estatística, conhecida como *distância entre verossimilhanças*, para verificar estas observações

$$LD_i = \frac{2}{p}\{l(\hat{\beta}) - l(\hat{\beta}_{(i)})\}, \tag{3.17}$$

onde $l(\cdot)$ é a função de log-verossimilhança.

Contudo, Davison e Snell (1991) mostram que, expandindo (3.17) em série de Taylor, obtém-se

$$\hat{\beta}_{(i)} = \hat{\beta} - \hat{w}_i^{1/2}(1-h_{ii})^{1/2} r_{P_i}(X^T W X)^{-1} x_i. \tag{3.18}$$

Assim, (3.18) pode ser aproximado pela *distância generalizada de Cook* $D_i = \frac{h_{ii}}{p(1-h_{ii})} r_{P_i}^{*2}$, onde p é o posto da matriz modelo X e $r_{P_i}^* = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)(1-h_{ii})}}$ é o resíduo de Pearson padronizado.

Lee (1987) propõe julgar os pontos $D_i > \frac{\chi^2_{p,\alpha}}{p}$ como influentes. Uma ferramenta informal para visualizar tais observações é usar um "*index plot*" (gráfico indexado) dos D_i versus i com limite $\frac{\chi^2_{p,\alpha}}{p}$. Entretanto, McCullagh e Nelder (1989) propõem medir a influência de uma observação através da *estatística modificada de Cook*, sugerida por Atkinson (1981), e expressa, no contexto dos MLGs, por

$$T_i = \left\{ \frac{n-p}{p} \frac{h_i}{1-h_i} \right\}^{1/2} |r^2_{D_{(i)}}|, \tag{3.19}$$

onde $r_{D_{(i)}}$ é aproximadamente o desvio residual deletado (vide McCullagh e Nelder, 1989, Sec. 12.7.3). Aqui, $r^2_{D_{(i)}}$ é definido pela variação no desvio

residual causada pela omissão da i-ésima observação. Atkinson (1981) propõe julgar os pontos em que $T_i > 2\sqrt{\frac{p}{n}}$ como influentes.

3.9 Exercícios

1. Definir os resíduos de Pearson, Anscombe e residual para os seguintes modelos: Poisson, binomial, normal inverso, gama e binomial negativo com índice conhecido.
2. Determinar a fórmula da distância generalizada de Cook para os modelos de Poisson, gama e normal inverso com respectivas ligações canônicas.
3. Comparar os resíduos de Anscombe, Pearson e como raiz quadrada da 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.
4. Definir os resíduos de Anscombe, Pearson e como raiz quadrada da componente do desvio para o modelo binomial negativo, fazendo uma comparação entre os três resíduos.
5. Seja $Y_\ell \sim B(m_\ell, \mu_\ell)$ com a notação usual $\mu = f^{-1}(X\beta)$, $\beta = (\beta_1 \ldots \beta_p)^T$, etc. Demonstrar que os resíduos podem ser definidos por $[G(Y_\ell/m_\ell) - G(\hat{\mu}_\ell)]/G'(\hat{\mu}_\ell)[\hat{\mu}_\ell(1-\hat{\mu}_\ell)/m_\ell]^{1/2}$. Quais as vantagens das escolhas $G(\mu) = \mu$, $G(\mu) = \log[\mu/(1-\mu)]$ e $G(\mu) = \int_0^\mu x^{-1/3}(1-x)^{-1/3}dx$.
6. Justificar o uso do gráfico dos resíduos versus as seguintes escalas dependendo do tipo de erro: $\hat{\mu}(normal)$, $2\sqrt{\hat{\mu}}$ (Poisson), $2\log\hat{\mu}$ (gama) e $-2/\sqrt{\hat{\mu}}$ (normal inversa).
7. Seja $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ o análogo da matriz de projeção para um modelo linear generalizado. Demonstre que, aproximadamente,

$$V^{-1/2}(\hat{\mu} - \mu) = HV^{-1/2}(y - \mu),$$

   onde $V = \text{diag}\{V(\mu_1), \ldots, V(\mu_n)\}$ é a matriz diagonal com a função de variância.
8. Demonstre as correções de viés apresentadas na Tabela 3.2.

9. No modelo normal-linear com $\mu = E(y) = X\beta + g(z;\gamma)$, sendo $g(z;\gamma)$ aproximadamente linear, demonstrar que os resíduos parciais $\tilde{R} = Py + (I - P)z\hat{\gamma}$, onde $P = I - X(X^T X)^{-1} X^T$, podem ser expressos como combinações lineares dos resíduos $y - \hat{\mu}$ e, também, como combinações lineares dos dados y.

Capítulo 4

Principais Modelos Lineares Generalizados e Extensões

4.1 Modelos para Dados Contínuos

O modelo clássico de regressão estudado no Capítulo 1 supõe que a variância da variável resposta é constante para quaisquer valores dos parâmetros $\beta's$. Este modelo é o mais importante na análise de dados contínuos. Entretanto, é comum encontrarmos na prática dados contínuos cuja variância cresce com a média da variável resposta, ou seja:

$$Var(Y) = \sigma^2\mu^2,$$

onde σ representa o coeficiente de variação de Y. Para valores pequenos de σ, a transformação que estabiliza a variância é $log(Y)$, cujos momentos aproximados valem $E(\log Y) = \log\mu - \frac{\sigma^2}{2}$ e $Var(\log Y) = \sigma^2$. Além disso, dados contínuos positivos não podem ser modelados pelo modelo normal linear, pois não há garantia da média ser positiva.

Uma possibilidade para modelarmos dados contínuos positivos com variância constante, seria supor o modelo normal com ligação logaritmo, ou seja, $\mu = E(Y) = \exp(X\beta)$. A ligação logaritmo, então, garante a positividade

de μ. Outra alternativa seria usar a transformação logaritmo para obtermos dados modificados em $\mathbb{R}$ e, então, adotar o modelo normal para os dados transformados. Assim, os dados originais seguiriam a distribuição log normal.

Considerando-se que os dados contínuos positivos têm coeficiente de variação (e não a variância) constante para todas as observações, a melhor modelagem é geralmente obtida através da distribuição gama com uma ligação apropriada, por exemplo, logaritmo ou potência. A ligação recíproco também pode ser usada pois produz estatísticas suficientes que são funções lineares dos dados.

Em suma, dados contínuos positivos com coeficiente de variação constante podem ser modelados rotineiramente pelas distribuições gama e log normal. Se a suposição do coeficiente de variação constante for violada, os dados contínuos positivos devem ser modelados pela distribuição normal inversa ou, então, aplicando-se alguma transformação apropriada para se adotar o modelo normal aos dados modificados (vide modelo de Box e Cox, Seção 4.6).

4.2 Modelo Logístico Linear

O modelo logístico linear é um membro da classe dos MLGs servindo de alternativa para analisar respostas binárias através de um conjunto de variáveis explicativas. A relação entre a probabilidade de sucesso p e o conjunto de variáveis explicativas é dada através da função de ligação logística (vide Seção 2.8). Tal relacionamento é sigmoidal, uma vez que a relação entre o $logit(p)$ e a matriz modelo é linear. O modelo logístico linear também é conhecido na literatura como modelo de regressão logística.

Suponha que temos n observações binomiais sob a forma y_i/m_i, $i = 1, \ldots, n$, de modo que $E(y_i) = m_i p_i$, onde p_i é a probabilidade de sucesso correspondente à i-ésima observação. Assim, o *modelo logístico linear* relaciona p_i com um conjunto de p variáveis explicativas $x_{1i}, x_{2i}, \ldots, x_{pi}$, associado a i-ésima observação, sendo expresso por

$$\text{logit}(p_i) = \log\left\{\frac{p_i}{(1-p_i)}\right\} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}. \tag{4.1}$$

Podemos escrever (4.1) como

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi})}, \tag{4.2}$$

ou, denotando-se $\eta_i = \sum_j \beta_j x_{ji}$, de forma mais simples por

$$p_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

Desde que y_i seja uma observação proveniente de uma distribuição binomial com média $m_i p_i$, o valor esperado de y_i é $E(y_i) = m_i \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)$. As equações (4.1) ou (4.2) definem a componente sistemática do modelo logístico linear.

4.2.1 Ajuste do modelo

Sejam dados binomiais sob a forma de y_i sucessos em m_i ensaios de Bernoulli (vide Seção 2.8), $i = 1, \ldots, n$. A transformação logística, correspondente à probabilidade de sucesso p_i, é expressa como uma combinação linear de p variáveis explicativas $x_{1i}, x_{2i}, \ldots, x_{pi}$, sendo dada por

$$\text{logit}(p_i) = \log\left\{\frac{p_i}{(1 - p_i)}\right\} = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}.$$

A observação y_i com valor esperado $m_i p_i$ pode ser expressa como $y_i = m_i p_i + \varepsilon_i$. A componente do resíduo é dada por $\varepsilon_i = y_i - m_i p_i$ tendo valor esperado zero, contudo sua distribuição não é mais binomial. A distribuição do resíduo ε_i é conhecida como distribuição binomial modificada. Apesar de não haver relação entre a distribuição dos dados e aquela do resíduo, neste caso, é importante salientar que no ajuste do modelo é necessário apenas a distribuição de y_i.

Note que para ajustarmos o modelo logístico linear é necessário, primeiramente, estimar os $p + 1$ parâmetros $\beta_0, \beta_1, \ldots, \beta_p$. Estes parâmetros são esti-

mados através do método de máxima verossimilhança. Neste caso, a função de verossimilhança $L(\beta)$ é dada por

$$L(\beta) = \prod_{i=1}^{n} \binom{m_i}{y_i} p_i^{y_i} (1-p_i)^{m_i - y_i}.$$

A função de verossimilhança pode ser considerada função dos parâmetros $\beta's$ pois esta depende das probabilidades de sucesso desconhecidas p_i, as quais dependem dos $\beta's$ através da expressão (4.2). O problema agora é obter os valores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ que maximizam $\ell(\beta)$ ou, equivalentemente, log $L(\beta)$, expresso por

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{n} \left\{ \log \binom{m_i}{y_i} + y_i \log p_i + (m_i - y_i) \log(1-p_i) \right\} \\ &= \sum_{i=1}^{n} \left\{ \log \binom{m_i}{y_i} + y_i \eta_i - m_i \log(1+e^{\eta_i}) \right\}, \end{aligned} \tag{4.3}$$

onde $\eta_i = \sum_{j=0}^{p} \beta_j x_{ji}$ e $x_{0i} = 1$ para todo $i = 1, \ldots, n$. Para tanto, é necessário calcularmos a derivada do logaritmo da função de verossimilhança em relação aos $p+1$ parâmetros desconhecidos β, dada por

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} m_i x_{ij} e^{\eta_i} (1+e^{\eta_i})^{-1}, \; j = 0, 1, \ldots, p.$$

Assim, igualando estas derivadas a zero obtemos um conjunto de $p+1$ equações não-lineares. As estimativas $\hat{\beta}_j$ correspondem à solução deste sistema e podem ser obtidas através do algoritmo iterativo conhecido como método escore de Fisher descrito na Seção 2.4.

Uma vez calculados os $\hat{\beta}'s$, as estimativas do preditor linear do modelo são dadas por $\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$.

Conseqüentemente, as probabilidade estimadas $\hat{p}_i$ são obtidas fazendo

$$\hat{p}_i = \frac{e^{\hat{\eta}_i}}{1+e^{\hat{\eta}_i}}.$$

4.2.2 Bondade de ajuste

Existem diversas estatísticas que medem a discrepância entre as proporções observadas y_i/m_i e as proporções ajustadas $\hat{p}_i$. O *desvio* (D) é uma estatística de bondade de ajuste muito utilizada na literatura e baseia-se nas funções de log-verossimilhança maximizada sob o modelo em investigação $\widehat{l}_p$ e sob o modelo saturado $\tilde{l}_n$ (vide Seção 2.7), sendo expressa por

$$D = 2(\tilde{l}_n - \tilde{l}_p).$$

A partir desta expressão a log-verosimilhança maximizada para o modelo em investigação é dada por

$$\hat{l}_p = \sum_{i=1}^{n} \left\{ \log \binom{m_i}{y_i} + y_i \log \hat{p}_i + (m_i - y_i) \log(1 - \hat{p}_i) \right\}.$$

No modelo saturado as probabilidades ajustadas são idênticas às proporções observadas $\tilde{p}_i = y_i/m_i$. Assim, a log-verossimilhança maximizada sob o modelo saturado é dada por

$$\tilde{l}_n = \sum_{i=1}^{n} \left\{ \log \binom{m_i}{y_i} + y_i \log \tilde{p}_i + (m_i - y_i) \log(1 - \tilde{p}_i) \right\}.$$

Logo, o desvio (D) reduz-se a

$$D = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{\tilde{p}_i}{\hat{p}_i} \right) + (m_i - y_i) \log \left(\frac{1 - \tilde{p}_i}{1 - \hat{p}_i} \right) \right\}.$$

Fazendo $\hat{y}_i = m_i \hat{p}_i$, o desvio pode expresso como

$$D = 2 \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left(\frac{m_i - y_i}{m_i - \hat{y}_i} \right) \right\}.$$

É importante ressaltar, no caso onde $n_i = 1$, $i = 1, \ldots, n$, que temos

$$D = -2 \sum_{i=1}^{n} \left\{ \hat{p}_i \text{logit}\left(\hat{p}_i\right) + \log\left(1 - \hat{p}_i\right) \right\}.$$

Neste caso, o desvio torna-se uma estatística de bondade de ajuste desinformativa, pois a mesma só depende das probabilidades de sucesso ajustadas $\hat{p}_i$.

Outra estatística que pode ser empregada para verificar a adequação do modelo em investigação é a estatística X^2 de Pearson definida por

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - m_i \hat{p}_i)^2}{m_i \hat{p}_i (1 - \hat{p}_i)}.$$

Tanto o desvio (D) quanto a estatística X^2 de Pearson têm distribuição assintótica χ^2_{n-p}. Para outras informações sobre estatísticas de bondade de ajuste vide a Seção 2.7.

4.3 Modelo Log-Linear para Contagens

O modelo log-linear corresponde ao caso onde $Y \sim P(\mu)$, $\eta_i = \log \mu_i = \sum_{j=1}^{p} x_{ij}\beta_j$, $i = 1, \ldots, n$, com o parâmetro natural da distribuição de Poisson sendo igual a $\log \mu$. As quantidades x_{ij} podem ser variáveis explanatórias como no modelo logístico linear, ou binárias restritas aos valores 0 e 1 como na análise de contingência, e podem ainda ser uma mistura de variáveis explanatórias e binárias (vide Seção 1.3.4).

O algoritmo de estimação de um modelo log-linear tem a forma

$$X^T W^{(m)} X \beta^{(m+1)} = X^T W^{(m)} y^{*(m)},$$

onde $W = diag\{\mu\}$ e $y^* = \eta + W^{-1}(y - \mu)$. Estas equações podem ser escritas como $E(S_j; \mu) = s_j$, $j = 1, \ldots, p$, onde os $s_j's$ são os valores observados das estatísticas suficientes $S_j = \sum_{i=1}^{n} x_{ij} y_i$ para os parâmetros $\beta's$. Em

forma matricial $X^T\hat{\mu} = X^T y$. Quando os elementos da matriz modelo X são 0 ou 1, essas equações implicam que as estimativas das médias são obtidas igualando certas freqüências marginais totais aos seus valores esperados. De $S = (S_1, \ldots, S_p)^T = X^T y$ obtém-se $Cov(S) = X^T W X$.

Considera-se que a EMV $\hat{\beta}$ tem, aproximadamente, distribuição normal $N_p(\beta, (X^T \hat{W} X)^{-1})$ e, portanto, testes e intervalos de confiança para os parâmetros $\beta's$ podem ser obtidos com base nesta distribuição. Intervalos de confiança para os contrastes $\tau = e^T\beta$, onde $e = (e_1, \ldots, e_p)^T$ é um vetor de componentes conhecidas, podem também ser baseados na aproximação normal $\hat{\tau} = e^T\hat{\beta} \sim N_p(e^T\beta, e^T(X^T\hat{W}X)^{-1}e)$.

4.3.1 Modelos hierárquicos

Tem-se um grande interesse numa classe de modelos log-lineares, denominados *hierárquicos.* Estes modelos são baseados num método geral de parametrização, encontrado na análise de variância de experimentos fatoriais. Num modelo hierárquico, se um conjunto T é constituído por parâmetros $\beta's$ iguais a zero, então, em qualquer outro conjunto de parâmetros, gerado por termos que contenham pelo menos um termo gerador do conjunto T, todos os parâmetros deverão ser iguais a zero. Por exemplo, o modelo

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ijk}^{ABC}$$

para a classificação cruzada de três fatores A, B e C sujeitos às restrições usuais, não é hierárquico, pois a interação β_{ijk}^{ABC} está incluída sem as interações β_{ij}^{AB}, β_{ik}^{AC} e β_{jk}^{BC} estarem no modelo.

Todo modelo log-linear hierárquico corresponde a um conjunto mínimo de estatísticas suficientes representado pelos totais marginais. Existem argumentos convincentes para considerar apenas os modelos log-lineares hierárquicos na análise de dados. Em particular, existe a conveniência computacional no cálculo das estimativas de máxima verossimilhança (EMV) e, mais importantemente, uma interpretação simples. Claramente, os algoritmos de ajustamento do $GLIM$ e do $S-Plus$ não fazem qualquer distinção entre um modelo não hierárquico ou hierárquico.

Os modelos hierárquicos podem ser classificados em duas classes: a primeira, cujas estimativas $\hat{\mu}'s$ têm forma fechada, e a segunda cujas estimativas só podem ser calculadas através de técnicas iterativas. Os termos nas expressões dos $\hat{\mu}'s$ em forma fechada correspondem a certos totais marginais, que representam estatísticas suficientes para os parâmetros do modelo.

Goodman (1970, 1973) estabelece que todo modelo hierárquico, onde os $\hat{\mu}'s$ têm forma fechada, pode ser interpretado em termos de independência incondicional e/ou condicional e equiprobabilidade, mas nos modelos, onde os $\hat{\mu}'s$ não têm forma fechada, esta interpretação é, em geral, muito difícil. Algumas vezes é possível transformar o modelo não-hierárquico, associado à uma tabela de contingência, em hierárquico, através da permutação de celas.

Os modelos hierárquicos possíveis para tabelas de contingência com 3 entradas podem ser divididos em nove classes. Com exceção do modelo sem a iteração dos 3 fatores, todos os demais modelos hierárquicos têm os $\hat{\mu}'s$ em forma fechada. Em tabelas de contingência de 4 entradas, Goodman (1970) tem notado a existência de 17 modelos com os $\hat{\mu}'s$ em forma fechada, entre 27 modelos hierárquicos distintos, de um total de 170 diferentes tipos de modelos.

Goodman (1971) e Haberman (1974, Cap. 5) determinam regras para verificar se um modelo hierárquico tem $\hat{\mu}$ em forma fechada. Para modelos hierárquicos com $\hat{\mu}$ em forma fechada, o algoritmo do GLIM, em geral, não converge em uma única iteração. Haberman (1974) apresenta ainda resultados gerais para obtenção das equações de máxima verossimilhança em modelos não hierárquicos. Entretanto, essas regras não têm finalidade prática.

Para os modelos log-lineares com um número máximo de parâmetros, Bishop, Fienberg e Holland (1975) usam o método delta (Rao, 1973) para calcular a estrutura assintótica $K^{-1} = \{-k^{rs}\} = (X^T W X)^{-1}$ das estimativas dos parâmetros lineares. Lee (1977) desenvolveu regras gerais para o cálculo de expressões fechadas para as covariâncias assintóticas $-k^{rs}$, em modelos log-lineares hierárquicos, com formas fechadas para os $\hat{\mu}'s$.

4.3.2 Modelos hierárquicos para tabelas de contingência com 3 entradas

Apresentam-se, agora, todas as nove classes de modelos hierárquicos correspondentes à classificação de três fatores A, B e C. Seja $y_{ijk} \sim P(\mu_{ijk})$, o número de observações com $A = i$, $B = j$ e $C = k$, em que $1 \leq i \leq r$, $1 \leq j \leq s$ e $1 \leq k \leq t$, e utiliza-se da notação usual $y_{i++} = \sum_{j,k} y_{ijk}$, $y_{ij+} = \sum_k y_{ijk}$, etc.

O modelo saturado é definido por

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ij}^{AB} + \beta_{ik}^{AC} + \beta_{jk}^{BC} + \beta_{ijk}^{ABC}, \tag{4.4}$$

com as restrições usuais da análise de variância $\beta_+^A = \beta_+^B = \cdots = \beta_{+jk}^{ABC} = \beta_{i+k}^{ABC} = \beta_{ij+}^{ABC} = 0$. Este modelo corresponde à 1ª classe e tem-se $\hat{\mu}_{ijk} = y_{ijk}$.

A 2ª classe é definida pelo modelo (4.4) com as restrições adicionais $\beta_{ijk}^{ABC} = 0$ para todos os índices i, j, k, isto é, corresponde ao modelo sem a interação dos três fatores. A média μ_{ijk} não pode ser dada como função explícita dos totais marginais μ_{ij+}, μ_{i+k} e μ_{+jk}. Para resolver as equações de máxima verossimilhança $\hat{\mu}_{ij+} = y_{ij+}$, $\hat{\mu}_{i+k} = y_{i+k}$ e $\hat{\mu}_{+jk} = y_{+jk}$, $i = 1, \ldots, r$, $j = 1, \ldots, s$, $k = 1, \ldots, t$, onde y_{ij+}, y_{i+k} e y_{+jk} são as estatísticas suficientes minimais, necessita-se de métodos iterativos. Este modelo pode, por exemplo, ser interpretado como de interação entre A e B, dado C, independente do nível C, isto é, a razão do produto cruzado condicional $\mu_{ijk}\mu_{i'j'k}/\mu_{ij'k}\mu_{i'jk}$ independente de k.

A 3ª classe contém 3 modelos que podem ser deduzidos do modelo

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ij}^{AB} + \beta_{ik}^{AC}$$

por simples permutação. Este modelo é equivalente à hipótese que os fatores B e C são independentes, dado o fator A, isto é,

$$P(B = j, C = k \mid A = i) = P(B = j \mid A = i)P(C = k \mid A = i)$$

ou $\mu_{ijk} = \mu_{i+k}\mu_{ij+}/\mu_{i++}$. As estimativas são dadas, em forma fechada, por

$\hat{\mu}_{ijk} = y_{i+k}y_{ij+}/y_{i++}$, onde y_{i+k} e y_{ij+} são estatísticas suficientes minimais. Esta hipótese de independência condicional é análoga à correlação parcial igual a zero entre duas variáveis, dada uma terceira variável, num universo de três variáveis normais.

A 4ª classe também contém três modelos do tipo

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ij}^{AB}.$$

Este modelo equivale à hipótese que o fator C é independente do par (A, B), isto é,

$$P(A = i, B = j, C = k) = P(A = i, B = j)P(C = k)$$

ou $\mu_{ijk} = \mu_{ij+}\mu_{++k}/\mu_{+++}$. As estimativas $\hat{\mu}_{ijk} = y_{ij+}y_{++k}/y_{+++}$ são funções explícitas das estatísticas suficientes minimais y_{ij+}e y_{++k}.

A 5ª classe corresponde ao modelo

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C$$

com todas as interações nulas. Este modelo corresponde à hipótese que os três fatores são mutuamente independentes:

$$P(A = i, B = j, C = k) = P(A = i)P(B = j)P(C = k)$$

ou $\mu_{ijk} = \mu_{i++}\mu_{+j+}\mu_{++k}/\mu_{+++}^2$. As estimativas $\hat{\mu}_{ijk}$ igualam $y_{i++}y_{+j+}y_{++k}/y_{+++}^2$, onde os termos do numerador são as estatísticas suficientes minimais.

A 6ª classe tem 3 modelos obtidos de

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_k^C + \beta_{ik}^{AC}$$

por simples permutação dos fatores; este modelo equivale a cada nível de B ser igualmente equiprovável, dados A e C, isto é

$$P(B = j \mid A = i, C = k) = s^{-1}.$$

As estimativas de máxima verossimilhança são $\hat{\mu}_{ijk} = y_{i+k}/s$.

A 7ª classe também engloba 3 modelos do tipo

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_k^C.$$

Este modelo equivale às hipóteses

$$P(A = i, C = k) = P(A = i)P(C = k)$$

e

$$P(B = j \mid A = i, C = k) = s^{-1}$$

e, portanto, que os fatores A e C são independentes e, dados A e C, cada categoria de B é igualmente equiprovável. As estimativas são $\hat{\mu}_{ijk} = (y_{i++}y_{++k})/(sy_{+++})$.

A 8ª classe consiste de 3 modelos do tipo

$$\log \mu_{ijk} = \beta + \beta_i^A,$$

e este equivale à hipótese

$$P(B = j, C = k \mid A = i) = (st)^{-1},$$

que dado A, as combinações das categorias B e C são igualmente equiprováveis. Tem-se $\hat{\mu}_{ijk} = y_{i++}/st$.

A 9ª e última classe é formada pelo modelo simples

$$\log \mu_{ijk} = \beta,$$

isto é, uma única média ajustada aos dados. O modelo equivale a

$$P(A = i, B = j, C = k\ ) = (rst)^{-1},$$

isto é, todas as combinações de fatores são igualmente equiprováveis. Tem-se $\hat{\mu}_{ijk} = y_{+++}(rst)^{-1}$.

4.3.3 Testes de adequação

Para verificar a adequação do ajustamento de um modelo log-linear com p parâmetros independentes aos dados $y_1, \ldots, y_n$, utiliza-se as estatísticas

$$\begin{aligned} D(\hat{\mu}; \mathbf{y}) &= 2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i), \\ X^2 &= \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \end{aligned} \tag{4.5}$$

A primeira corresponde ao desvio que foi tratado na Seção 2.7.1 e a segunda é a estatística de Pearson generalizada apresentada na Seção 2.7.2.

As estatísticas (4.5) podem ser interpretadas como sendo a quantidade de variação dos dados não explicada pelo modelo. Supondo o modelo correto, elas têm, assintoticamente, distribuição χ^2_{n-p}.

Gart e Zweifel (1967) sugerem a adição de 0,5 às freqüências observadas em (4.5) para um aperfeiçoamento da aproximação χ^2 de referência. As distribuições de $D(\hat{\mu}; \mathbf{y})$ e X^2 se tornam mais próximas da distribuição χ^2_{n-p}, quando todas as médias $\hat{\mu}_i's$ crescem e, neste caso, a diferença $|D(\hat{\mu}; \mathbf{y}) - X^2|$ se torna cada vez menor.

As aproximações das distribuições dessas estatísticas por χ^2 são bastantes razoáveis se todos os $\hat{\mu}_i's$ forem maiores que 5. Alguns estudos de Monte Carlo (Larntz, 1978) sugerem que a estatística $D(\hat{\mu}; \mathbf{y})$ se comporta de maneira aberrante, quando a tabela tem observações muito pequenas, mas que as duas estatísticas são razoavelmente aproximadas pela distribuição χ^2, quando o menor valor dos $\hat{\mu}_i's$ for maior que 1.

Nos modelos log-lineares hierárquicos é comum usar a notação de *classe geradora*, que consiste de todos os termos de ordem mais alta que geram os parâmetros do modelo; estes termos, correspondentes a certos totais marginais, representam estatísticas suficientes de dimensão mínima. Esta notação descreve, univocamente, todos os modelos log-lineares hierárquicos.

A Tabela 4.1 apresenta os graus de liberdade $n - p$ para todas as nove

classes de modelos hierárquicos de 3 fatores, considerados anteriormente. Além disso, ainda estão especificados os termos geradores e as interpretações dos modelos.

Tabela 4.1: *Graus de liberdade das estatísticas $D(\hat{\mu};\mathbf{y})$ e X^2 para modelos log-lineares hierárquicos em tabelas de 3 entradas*

Classe Geradora	Graus de Liberdade	Descrição
1: ABC	0	modelo saturado
2: AB,AC,BC	$(r-1)(s-1)(t-1)$	associação dois a dois
3: AB,AC	$r(s-1)(t-1)$	dado A, B e C independentes
4: AB,C	$(rs-1)(t-1)$	o par (A,B) independente de C
5: A,B,C	$rst-r-s-t+2$	os três fatores independentes
6: AC	$rt(s-1)$	dados A e C, todas as categorias de B equiprováveis
7: A,C	$rst-r-t+1$	mesmo que a classe 6 com os fatores A e C independentes
8: A	$r(st-1)$	dado A, todas a combinações das categorias B e C equiprováveis
9: Nula	$rst-1$	modelo nulo

4.3.4 Testes de comparação entre modelos

A estatística $D(\hat{\mu};\mathbf{y})$ é usada para comparação de modelos log-lineares encaixados. Formula-se uma seqüência de interesse de modelos log-lineares encaixados $M_{p_1} \subset M_{p_2} \subset \cdots \subset M_{p_r}$ com parâmetros $p_1 < p_2 < \cdots < p_r$ e desvios $D_{p_1} > D_{p_2} > \ldots > D_{p_r}$. A diferença entre os desvios dos modelos encaixados

$M_{p_j} \subset M_{p_i}$ $(p_j < p_i)$ é dada por

$$D_{p_j} - D_{p_i} = 2\sum_{k=1}^{n} y_k \log(\hat{\mu}_{jk}/\hat{\mu}_{ik}), \tag{4.6}$$

onde $\hat{\mu}_{jk}$ $(\hat{\mu}_{ik})$ é a k-ésima componente estimada do vetor $\hat{\mu}_j$ $(\hat{\mu}_i)$. Esta estatística é usada para testar se a diferença entre os valores esperados ajustados, segundo os modelos M_{p_i} e M_{p_j} é, simplesmente, devido à uma variação aleatória, dado que os valores esperados verdadeiros satisfazem o modelo mais pobre M_{p_j}. Segundo $M_{p_j}, D_{p_j} - D_{p_i}$ tem distribuição assintótica $\chi^2_{p_i - p_j}$.

Se a seqüência é formada por modelos hierárquicos, Goodman (1969) demonstra, baseando-se na forma multiplicativa das estimativas das médias, que a expressão (4.6) iguala

$$D_{p_j} - D_{p_i} = 2\sum_{k=1}^{n} \hat{\mu}_{jk} \log(\hat{\mu}_{jk}/\hat{\mu}_{ik}). \tag{4.7}$$

A estatística (4.7) tem a mesma forma de um simples desvio e, mais ainda, pode ser interpretada como uma razão de verossimilhanças condicional para os parâmetros extras que estão em M_{p_i}. Portanto, o desvio em modelos log-lineares hierárquicos tem a propriedade de *aditividade*, que geralmente não é verificada para a estatística X^2. Por esta razão, o desvio é a estatística preferida.

A propriedade de aditividade é a base para testar a significância de adicionar termos a um modelo. Tem-se: $D_{p_j} = (D_{p_j} - D_{p_i}) + D_{p_i}$, onde $D_{p_i}(D_{p_j})$ é a quantidade de variação dos dados, não explicada pelo modelo $M_{p_i}(M_{p_j})$, e $D_{p_j} - D_{p_i}$ é a variação explicada pelos termos extras no modelo M_{p_i}. O método de partição da estatística D_p para os modelos log-lineares hierárquicos, foi desenvolvido por Ku e Kullback (1968). Esta partição possibilita apresentar os resultados na forma de tabelas de análise de variância.

Pode-se definir uma medida de comparação entre modelos encaixados, análoga ao coeficiente de correlação múltipla dos modelos de regressão. Na comparação dos modelos encaixados $M_{p_j} \subset M_{p_i}(p_j < p_i)$, esta medida é $(D_{p_j} -$

$D_{p_i})/D_{p_j}$ e representa um índice de qualidade relativa dos ajustamentos dos modelos aos dados. Esta estatística é limitada entre 0 e 1; um valor próximo de um sugere que M_{p_j} é muito melhor que M_{p_i}, e um valor próximo a zero é indicativo que os dois modelos proporcionam, aproximadamente, ajustamentos equivalentes.

Rao (1973) propõe a estatística

$$R = \sum_{k=1}^{n} \frac{(\hat{\mu}_{jk} - \hat{\mu}_{ik})^2}{\hat{\mu}_{ik}} \tag{4.8}$$

que é análoga a (4.7) e tem a mesma forma da estatística X^2. Entretanto, o seu uso, na prática, não é difundido.

4.4 Modelo para Dados Multinomiais

Se a resposta de um indivíduo ou item está restrita a um conjunto de possíveis opções ou categorias pré-estabelecidas, dizemos que a variável de interesse é politômica, sendo a distribuição multinomial comumente usada para representar tal variável.

Suponha que indíviduos numa população de interesse possuam uma, e apenas uma, de p características $A_1, \ldots, A_p$. Tais características podem ser, por exemplo, cor do cabelo, posição sócio-econômica, causa da morte, etc. Se a população é suficientemente grande e se uma amostra aleatória de tamanho n é sorteada, quantos indivíduos poderemos esperar que apresentem a característica A_j? A resposta pode ser dada através da distribuição multinomial, expressa por

$$P(Y_1 = y_1, \ldots, Y_p = y_p; n, \pi) = \binom{n}{\mathbf{y}} \pi_1^{y_1} \ldots \pi_p^{y_p}, \tag{4.9}$$

onde $\pi_1, \ldots, \pi_p$ são as proporções populacionais de cada característica e

$$\binom{n}{\mathbf{y}} = \frac{n!}{y_1! \ldots y_p!}.$$

Outra derivação da distribuição multinomial é a seguinte. Suponha que $Y_1, \ldots, Y_p$ são variáveis aleatórias de Poisson independentes com médias $\mu_1, \ldots, \mu_p$. Então, a distribuição condicional conjunta de $Y_1, \ldots, Y_p$, supondo que $Y_+ = n$, é dada por (4.9) com $\pi_j = \mu_j/\mu_+$.

A distribuição multinomial onde $\pi_j = 1/p$ é conhecida como distribuição multinomial uniforme.

4.4.1 Momentos e cumulantes

A função geratriz de momentos da distribuição multinomial $M(n, \pi)$ é expressa por

$$M_Y(t) = E\ \exp\left(\sum t_j Y_j\right) = \left\{\sum \pi_j \exp(t_j)\right\}^n.$$

Em seguida, apresentamos a função geratriz de cumulantes

$$K_Y(t) = n \log\left\{\sum \pi_j \exp(t_j)\right\}.$$

Os três primeiros cumulantes de uma distribuição multinomial são:

$$E(Y_r) = n\pi_r$$

$$cov(Y_r, Y_s) = \begin{cases} n\pi_r(1-\pi_r) & \text{se} \quad r = s \\ -n\pi_r\pi_s & \text{se} \quad r \neq s. \end{cases}$$

$$\kappa_3(Y_r, Y_s, Y_t) = \begin{cases} n\pi_r(1-\pi_r)(1-2\pi_r) & \text{se} \quad r = s = t \\ -n\pi_r\pi_t(1-2\pi_r) & \text{se} \quad r = s \neq t \\ 2n\pi_r\pi_s\pi_t & \text{se} \quad r \neq s \neq t \end{cases}$$

4.4.2 Log verossimilhança e função desvio

Suponha n vetores independentes, cada um com p categorias, denotados por $y_1, \ldots, y_n$, onde $y_i = (y_{i1}, \ldots, y_{ip})$ e $\sum_j y_{ij} = m_i$, i.e. $Y_i \sim M(m_i, \pi_i)$ com $\pi_i =$

$(\pi_{i1}, \ldots, \pi_{ip})$. Podemos denotar, para a i-ésima observação y_i, a contribuição da log-verossimilhança como

$$l(\pi_i; \mathbf{y}_i) = \sum_{j=1}^{p} y_{ij} \log \pi_{ij}.$$

Vale ressaltar que as observações e probabilidades estão sujeitas às seguintes restrições $\sum_j y_{ij} = m_i$ e $\sum_j \pi_{ij} = 1$.

A log-verossimilhança total é obtida através da soma das contribuições individuais, em virtude da suposição de independência das n observações. A log-verossimilhança total pode ser expressa por

$$l(\pi; \mathbf{y}) = \sum_{i,j} y_{ij} \log \pi_{ij}.$$

O desvio residual é obtido pela diferença entre a log-verossimilhança do modelo saturado e a log-verossimilhança do modelo em investigação. No modelo multinomial, obtemos a log-verossimilhança do modelo saturado quando $\tilde{\pi}_{ij} = y_{ij}/m_i$. Dessa forma,

$$\begin{aligned} D(\mathbf{y}; \pi) &= 2\left\{l(\tilde{\pi}; \mathbf{y}) - l(\hat{\pi}; \mathbf{y})\right\} \\ &= 2\sum y_{ij} \log \tilde{\pi}_{ij} - 2\sum y_{ij} \log \hat{\pi}_{ij} \\ &= 2\sum y_{ij} \log(y_{ij}/\hat{\mu}_{ij}), \end{aligned}$$

onde $\hat{\pi}_{ij} = \hat{\mu}_{ij}/m_i$.

4.5 Modelos com Parâmetros Adicionais Não-Lineares

Neste capítulo serão abordados modelos caracterizados pela inclusão de parâmetros desconhecidos em sua função de variância, em sua função de ligação ou em ambas. Adicionalmente, também será abordada a inclusão de covariáveis com uma estrutura não-linear no modelo.

4.5.1 Parâmetros na função de variância

Nos MLGs apresentados na Seção 2.3.1 foram abordadas cinco distribuições para a variável resposta. Dentre elas, a normal, a normal inversa e a gama, contém parâmetro de dispersão explícito. Por outro lado, as distribuições discretas em suas formas padrões não contém tal parâmetro. Além disso, supondo que o parâmetro de dispersão é constante, o mesmo não é utilizado na solução das equações de máxima verossimilhança de $\hat{\beta}$.

A distribuição binomial negativa é um exemplo de distribuição que apresenta um parâmetro desconhecido na função de variância. Esta distribuição discreta pode ser expressa da seguinte forma:

$$P(Y = y; \alpha, k) = \frac{(y+k-1)!}{y!(k-1)!} \frac{\alpha^y}{(1+\alpha)^{y+k}}; \qquad y = 0, 1, 2, \ldots$$

A média e a variância são dadas, respectivamente, por

$$E(Y) \quad = \quad \mu = k\alpha, \quad var(Y) = k\alpha + k\alpha^2 = \mu + \mu/k^2.$$

A log-verossimilhança pode ser expressa da seguinte forma

$$l = y \log\{\alpha/(1+\alpha)\} - k \log(1+\alpha) + (\text{função de } y \text{ e } k),$$

a qual, para k fixo, tem a forma de um MLG com ligação canônica

$$\eta = \log\left(\frac{\alpha}{1+\alpha}\right) = \log\left(\frac{\mu}{\mu+k}\right),$$

e função de variância

$$V = \mu + \mu^2/k.$$

O termo μ pode ser interpretado como a função de variância de uma Poisson e o termo μ^2/k como uma componente extra resultante da combinação de uma distribuição de Poisson com uma distribuição gama, no processo de obtenção da binomial negativa. A princípio k é desconhecido e, claramente, não se trata de um parâmetro de dispersão. Estimativas de k para amostras univariadas e multivariadas foram discutidas por Anscombe (1949). A sua esti-

mativa de máxima verossimilança requer a solução de uma equação não-linear envolvendo a função digama. Além disso, a utilização da ligação canônica é problemática, pois torna o preditor linear função do parâmetro da função de variância. Assim, o uso da binomial negativa em aplicações é bastante raro. Para maiores informações vide McCullagh e Nelder (1989).

Um outro exemplo de parâmetros adicionais na função de variância ocorre quando modelamos um conjunto de observações com erro gama e supomos que estes dados são coletados sob uma medida absoluta de erro. McCullagh e Nelder (1989) apresentam, neste caso, a seguinte função de variância:

$$V = \tau + \sigma^2 \mu^2.$$

O primeiro termo da expressão refere-se a medida absoluta de erro enquanto que o segundo termo corresponde a suposição da distribuição gama.

4.5.2 Parâmetros na função de ligação

Normalmente, no contexto dos MLGs, a função de ligação do modelo é suposta como conhecida. Entretanto, em algumas situações, pode ser útil assumir que a ligação provém de uma classe de funções indexadas por um ou mais parâmetros desconhecidos. Um teste de bondade de ajuste, em função deste(s) parâmetro(s), pode ser utilizado para detectar qual o intervalo de valores viáveis destes parâmetros é mais adequado para os dados. Além disso, se um particular valor é de interesse, pode-se, através de um teste de bondade de ligação (Pregibon, 1980), comparar seu desvio com o desvio do melhor ajuste. Outro teste que pode ser utilizado neste caso é o teste escore.

Uma classe de funções de ligação bastante conhecida é a função potência, expressa por

$$\eta = \begin{cases} \mu^\lambda, & \text{para} \quad \lambda \neq 0 \\ \log \mu, & \text{para} \quad \lambda = 0, \end{cases}$$

ou, supondo continuidade em $\lambda = 0$,

$$\eta = \frac{\mu^\lambda - 1}{\lambda}.$$

Esta classe de funções, utilizada para transformar os dados ao invés dos valores ajustados, foi definida por Box e Cox (1964) (vide Seção 4.6). Para um dado valor de λ, o modelo pode ser ajustado utilizando a ligação potência e, em seguida, seu desvio respectivo é calculado normalmente. Repetindo este procedimento para diferentes valores de λ, pode-se construir um gráfico dos respectivos desvios versus λ e visualizar qual o intervalo de valores de λ é mais adequado para os dados observados.

Pode-se otimizar η em relação a λ através do processo de linearização proposto por Pregibon (1980), pelo qual a função de ligação é expandida em série de Taylor sobre um valor fixo λ_0. Assim, para a classe de funções potência temos

$$\begin{aligned} g(\mu;\lambda) &= \mu^\lambda \simeq g(\mu;\lambda_0) + (\lambda - \lambda_0) g'_\lambda(\mu;\lambda) \\ &= \mu^{\lambda_0} + (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu, \end{aligned} \tag{4.10}$$

tal que podemos aproximar a função de ligação $\eta = \mu^\lambda$ por

$$\eta_0 = \mu^{\lambda_0} = \mu^\lambda - (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu = \sum \beta_j x_j - (\lambda - \lambda_0)\mu^{\lambda_0} \log \mu.$$

Dessa forma, dado um valor inicial λ_0 de λ, com os respectivos valores ajustados $\hat{\mu}_0$, é possível incluir no modelo a nova covariável $-\hat{\mu}_0^{\lambda_0} \log \hat{\mu}_0$. Ao ajustarmos este novo modelo, a estimativa do parâmetro pode ser interpretada como uma correção de primeira ordem para o valor inicial de λ_0. A redução significativa do desvio, em função da inclusão da nova covariável, pode ser utilizada como teste para verificar se λ_0 é um valor adequado para λ. Para obter a EMV de λ deve-se repetir o processo acima. A convergência não é garantida, contudo, sendo necessário que o valor de λ_0 seja próximo do valor de $\hat{\lambda}$ para que a expansão linear (4.10) seja adequada.

O método abordado anteriormente pode ser estendido no caso de mais de um parâmetro na função de ligação. Para cada parâmetro λ, adicionamos

uma covariável extra

$$-\left(\frac{\partial g}{\partial \lambda}\right)_{\lambda=\lambda_0}$$

na matriz modelo, sendo a estimativa do parâmetro da covariável uma correção de primeira ordem para o valor incial de λ_0. Pregibon (1980) discute dois exemplos com dois parâmetros. O primeiro é dado por

$$g(\mu;\alpha,\lambda) = \frac{(\mu+\alpha)^{\lambda}-1}{\lambda},$$

isto é, a família potência indexada por λ, mas, adicionando um parâmetro α de locação. Note que $g(\mu;1,1) = \mu$, de forma que a ligação identidade é um membro desta família.

O segundo exemplo é útil em modelos baseados em distribuições de tolerância. A função de ligação generalizada é dada por

$$g(\mu;\lambda,\delta) = \frac{\pi^{\lambda-\delta}-1}{\lambda-\delta} - \frac{(1-\pi)^{\lambda+\delta}-1}{\lambda+\delta},$$

onde π é a proporção de sucessos, ou seja, μ/m. Esta família contém a ligação logística quando $\lim\limits_{\lambda,\delta\to 0} g(\mu;\lambda,\delta)$.

A família de ligação uniparamétrica utilizada para dados binomiais

$$g(\mu;\lambda) = \log\left[\left\{\frac{\left(\frac{1}{(1-\pi)}\right)^{\lambda}-1}{\lambda}\right\}\right]$$

contém as ligações logística ($\lambda = 1$) e complemento log-log ($\lambda \to 0$) como casos especiais.

4.5.3 Parâmetros não-lineares nas covariáveis

Uma função de x como, por exemplo, e^{kx} pode ser incluída na matriz modelo substituindo-se, simplesmente, x por e^{kx} (desde que k seja conhecido). Entretanto, se k precisa ser estimado, então temos um problema de não-linearidade.

Neste caso, Box e Tidwell (1962) apresentam a seguinte técnica de linearização: seja $g(x;\theta)$ uma covariável não-linear, onde θ é desconhecido. Através de sua expansão em torno de um valor inicial θ_0, obtemos a seguinte aproximação linear

$$g(x;\theta) \simeq g(x;\theta_0) + (\theta - \theta_0)\left[\frac{\partial g}{\partial \theta}\right]_{\theta=\theta_0}.$$

Assim, se a covariável não-linear, pertencente ao preditor linear, é dada por

$$\beta g(x;\theta),$$

é possível reescrevê-la em função de

$$\beta u + \gamma v,$$

onde $u = g(x;\theta_0)$, $v = \left[\frac{\partial g}{\partial \theta}\right]_{\theta=\theta_0}$ e $\gamma = \beta(\theta - \theta_0)$.

Após o ajuste do modelo, contendo u e v como covariáveis adicionais, temos

$$\theta_1 = \theta_0 + \hat{\beta}/\hat{\gamma}$$

como um estimador iterativo. A convergência não é garantida para valores iniciais arbitrários muito distantes da solução. Maiores detalhes, vide McCullagh e Nelder (1989).

4.6 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, $Var(y) = \sigma^2$; (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 verificá-las, pelo menos aproximadamente. Em alguns problemas de regressão deve-se transformar tanto a variável dependente quanto as variáveis explicativas para que as suposições acima sejam satisfeitas. Transformações das variáveis explicativas não afetam as suposições (ii), (iii) e (iv).

Se os dados y com médias μ e variâncias $V(\mu)$, que dependem das médias, são transformados por $g(y)$ para satisfazer

$$Var\{g(y)\} = V(\mu)g'(u)^2 = k^2,$$

onde 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 de dados e a precisão com que estes 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 de 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 $y = (y_1, \ldots, y_n)^T$ são independentes e que existe um escalar λ tal que os dados transformados por

$$z = z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda & \text{se} \quad \lambda \neq 0 \\ \log y & \text{se} \quad \lambda = 0 \end{cases} \tag{4.11}$$

satisfazem $E(z) = \mu = X\beta$, $\mathrm{Var}(z_i) = \sigma^2$ para $i = 1, \ldots, n$ e $z \sim N(\mu, \sigma^2 I)$. A transformação (4.11) tem vantagem sobre a transformação potência simples y^λ por ser contínua em $\lambda = 0$. Apesar do modelo admitir a existência de um único λ produzindo linearidade dos efeitos sistemáticos, normalidade e variância constante dos dados transformados, pode ser que diferentes valores de λ sejam necessários para alcançar tudo isso.

Um valor λ pode ser proposto por uma análise exaustiva ou por considerações a priori dos dados, ou ainda, por facilidade de interpretação. Alternativamente, pode-se estimar λ por máxima verossimilhança, embora não haja garantia de que a EMV de λ produza todos os efeitos desejados.

Verifica-se, facilmente, que a log-verossimilhança como função de λ, σ^2 e β em relação às observações originais y é dada por

$$l(\lambda, \sigma^2, \beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z - X\beta)^T(z - X\beta) + (\lambda - 1)\sum_{i=1}^{n}\log y_i, \quad (4.12)$$

onde 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 (4.12) em relação a λ, σ^2 e β apresenta problemas computacionais e deve ser feita em duas etapas. Fixa-se λ e maximiza-se $\ell(\lambda, \sigma^2, \beta)$ em relação aos demais parâmetros produzindo as estimativas usuais da regressão como funções de λ, $\hat{\beta}(\lambda) = (X^TX)^{-1}X^Tz$ 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 λ vale, exceto por uma constante,

$$\hat{l}(\lambda) = -\frac{n}{2}\log\sigma^2(\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log y_i. \quad (4.13)$$

É bastante informativo traçar o gráfico de $\hat{l}(\lambda)$ versus λ para um certo conjunto de valores deste parâmetro, por exemplo, os inteiros de -3 a 3 e seus pontos médios. A estimativa de λ 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 λ. Claro está que a estimativa obtida é apenas uma aproximação da EMV de λ.

Objetivando a realização de inferência sobre o parâmetro λ, o teste da hipótese nula $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$, onde λ_0 é um valor especificado para λ, pode ser feito comparando a razão de verossimilhanças $w = 2[\hat{l}(\lambda) - \hat{l}(\lambda_0)]$ com a distribuição assintótica χ^2_1. Um intervalo de $100\%(1-\alpha)$ de confiança para λ é facilmente deduzido do gráfico de $\hat{l}(\lambda)$ versus λ como

$$\left\{\lambda; l(\lambda) > \hat{l}(\hat{\lambda}) - \frac{1}{2}\chi^2_1(\alpha)\right\}. \tag{4.14}$$

Se $\lambda = 1$ não pertencer ao intervalo (4.14) 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

$$W = \frac{\left\{\sum_{i=1}^{n} a_i z_{(i)}\right\}^2}{\left\{\sum_{i=1}^{n} (z_i - \bar{z})^2\right\}},$$

onde $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$ 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

$$A^2 = -n^{-1}\sum_{i=1}^{n}(2i-1)\left[1 + \log\{t_i(1 - t_{n+1-i})\}\right],$$

onde $t_i = \Phi\left(\frac{z_{(i)}-\bar{z}}{s}\right)$ e s^2 é a variância amostral. Valores grandes de A são significantes.

4.7 Modelo Linear Generalizado com um Parâmetro Não-Linear Extra

Este modelo é um caso especial da forma mais geral apresentada na Seção 4.5. Um parâmetro não-linear extra α aparece nos modelos lineares generalizados, mais freqüentemente, nas seguintes situações:

a) na função de ligação visando definir uma família paramétrica de ligações;

b) como parâmetro de transformação da variável resposta ou de variáveis explicativas;

c) na função de variância dos modelos de quase-verossimilhança (Seção 4.11) ou em certas distribuições como a binomial negativa, onde $V = \mu + \mu^2/\alpha$ depende de um parâmetro α que não é de escala e, em geral, é desconhecido;

d) no modelo logístico com probabilidade de sucesso da forma

$$\mu = \alpha + (1-\alpha)\exp(\eta)/[1+\exp(\eta)];$$

e) em distribuições especiais como o parâmetro de forma da Weibull.

A estimação conjunta de α e dos $\beta's$ geralmente é bastante complicada e só deverá ser feita quando for necessário conhecer a covariância conjunta entre as estimativas $\hat{\beta}$ e $\hat{\alpha}$. Se este não for o caso, deve-se estimar os $\beta's$ condicionalmente ao parâmetro α, isto é, calculando o desvio fixando $\alpha(D_p(\alpha))$. Um gráfico de $D_p(\alpha)$ versus α possibilitará escolher a estimativa $\tilde{\alpha}$ como o valor de α correspondente ao menor $D_p(\alpha)$. Deve-se esperar que $\tilde{\alpha}$ esteja próximo de $\hat{\alpha}$.

4.8 Modelos Lineares Generalizados com Ligação Composta

Considere um modelo com distribuição (2.1), mas com componente sistemática definida por

$$\begin{aligned} E(y) &= \mu = C\gamma, \\ f(Y) &= \eta = X\beta, \end{aligned} \tag{4.15}$$

onde μ e y são vetores $n \times 1$, C e X são matrizes conhecidas $n \times m$ e $m \times p$, respectivamente, $\gamma = (\gamma_1, \ldots, \gamma_m)^T$, $\eta = (\eta_1, \ldots, \eta_m)^T$ e $\beta = (\beta_1, \ldots, \beta_p)^T$. Uma média de y está relacionada com vários preditores lineares.

Denomina-se $f(C^-\mu) = \eta$, onde C^- é uma inversa generalizada de C, de *função de ligação composta*. Quando C é a matriz identidade, obviamente a ligação composta reduz-se a uma ligação simples $f(\mu) = \eta$. Uma extensão de (4.15) considera uma estrutura não-linear $\mu_i = c_i(\gamma)$ entre μ e γ. O ajustamento do modelo $\mu_i = c_i(\gamma)$, $f(\gamma) = \eta = X\beta$, pode ser feito via o algoritmo descrito em (2.4) com pequenas modificações. Sem perda de generalidade trabalha-se sem o escalar ϕ. Seja $\ell(\beta)$ a log-verossimilhança para β. Tem-se $\partial\ell(\beta)/\partial\beta = \tilde{X}^T V^{-1}(y-\mu)$, onde $V = \text{diag}\{V_i, \ldots, V_n\}$, $L = \{d\mu_i/d\eta_k\}$ 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 β iguala $\tilde{X}^T V^{-1}\tilde{X}$ e o processo iterativo é expresso por

$$X^T L^{(m)T} V^{(m)^{-1}} L^{(m)} X\beta^{(m+1)} = X^T L^{(m)T} V^{(m)^{-1}} y^{*(m)},$$

onde $y^* = L\eta + y - \mu$. A variável dependente y^*, a matriz modelo LX e os pesos V^{-1} se modificam no processo iterativo acima. O sistema GLIM não pode ser usado diretamente e o usuário deve trabalhar com programas especiais. A inicialização pode ser feita a partir do ajustamento de um modelo similar com C igual à matriz identidade. Quando μ é linear em γ, $L = CH^{-1}$, sendo agora $H = \text{diag}\{d\eta_1/d\gamma_1, \ldots, d\eta_m/d\gamma_m\}$ e, então, $\tilde{X} = CH^{-1}X$ e $y^* = CH^{-1}\eta + y - \mu$.

4.9 Modelos Semi-Paramétricos

Os modelos semi-paramétricos foram propostos por Green e Yandell (1985) quando definiram o preditor linear η como sendo a parte usual $X\beta$ dos MLGs mais uma parte $s(t)$, onde $s(\cdot)$ 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 $g_1, \ldots, g_q$ sendo os $\gamma's$ parâmetros desconhecidos. O problema de maximização consiste em definir uma log-verossimilhança penalizada como função dos parâmetros β e γ e maximizá-la

$$\max_{\beta,\gamma}[\ell\{\eta(\beta,\gamma)\} - \lambda J\{s(\gamma)\}/2],$$

onde $J[\cdot]$ é representativo de uma penalidade sobre a não-suavidade de $s(\cdot)$ e λ 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 $\gamma^T K\gamma$, com 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 $s(t_i) = \gamma_{0i} + \gamma_{1i} t_i$, onde 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 $(1/n, 2)$.

4.10 Modelos Aditivos Generalizados

Os modelos aditivos generalizados são definidos pela componente aleatória dos MLGs 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 $j = 1, \ldots, p$, onde 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 $S(y|x_{ij}),\ j = 1, \ldots, p,\ i = 1, \ldots, n$.

O suavizador mais usado tem a forma linear $S(y|x_{ij}) = \hat{a}_{ij} + \hat{b}_{ij} x_{ij}$, onde $\hat{a}_{ij}$ 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 $N_{ij} = \{x_{(i-r)j}, \ldots, x_{ij}, \ldots, x_{(i+r)j}\}$, onde o parâmetro r determina o tamanho de N_{ij}. Tem-se

$$\hat{b}_{ij} = \sum_{x_{ej} \in N_{ij}} (x_{ej} - \overline{x}_{ij}) y_e / \sum_{x_{ej} \in N_{ij}} (x_{ej} - \overline{x}_{ij}),$$

$$\hat{a}_{ij} = \overline{y}_i - \hat{b}_{ij} \overline{x}_{ij},$$

onde $\overline{x}_{ij}$ é a média dos valores em x_{ej} em N_{ij} e $\overline{y}_i$ 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 $\hat{f}(x_{ij}) = 0, \quad \forall i, j$ e $\hat{\beta} = \overline{y}$;
2. Fazer $j = 1, \ldots, p$ e $i = 1, \ldots, n$ 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 MLGs é baseada nas equações normais da regressão da variável dependente modificada y^* sobre X usando pesos W (Seção 2.4). O algoritmo fica sendo:

1. Inicializar $\hat{f}_j(x_{ij}) = 0,\ j = 1, \ldots, p,\ \hat{\beta} = g(\overline{y}),\ \hat{\eta} = \hat{\beta}1,\ \hat{W} = (\overline{y})$ e $\hat{H} = H(\overline{y})$, 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 $j = 1, \ldots, p$;
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 $N_{ij},\ i = 1, \ldots, p$;
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 pára-se; caso contrário, volta-se para 2.

4.11 Modelos de Quase-Verossimilhança

Nos modelos de quase-verossimilhança as variáveis são consideradas independentes sem ser necessário especificar qualquer distribuição para o erro e a componente sistemática é dada por:

$$E(y_i) = \mu_i(\beta), \qquad \text{Var}(y_i) = \phi V_i(\mu_i).$$

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 ϕ é um parâmetro de dispersão, possivel-

mente desconhecido, podendo ainda ser uma função de regressores adicionais. Usualmente $\mu(\beta)$ equivale à componente sistemática do MLG.

Define-se a *log-quase-verossimilhança* para uma única observação apenas com a suposição de existencia de sua média e de sua variância, por

$$Q = Q(y;\mu) = \frac{1}{\phi}\int (y-\mu)V(\mu)^{-1}d\mu. \tag{4.16}$$

Para $V(\mu) = k, \mu, \mu^2, \mu(1-\mu),\ \mu+\mu^2/k$ e μ^3, com k constante, e integrando (4.16), 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

$$E\{\partial Q/\partial\mu\} = 0,\ E\{\partial Q/\partial\mu\}^2 = -E\{\partial^2 Q/\partial\mu^2\} = 1/[\phi V(\mu)].$$

Uma terceira propriedade importante entre os logaritmos da verossimilhança ℓ e da quase-verossimilhança Q, supondo para ambos uma mesma função de variância, é dada por

$$-E\{\partial^2 Q/\partial\mu^2\} \leq -E\{\partial^2\ell/\partial\mu^2\}. \tag{4.17}$$

Se y seguir a família exponencial (2.1) 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 (4.17) somente ocorre no caso de ℓ ser a log-verossimilhança da família exponencial. O lado esquerdo de (4.17) 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 (4.16). As estimativas de máxima quase-verossimilhança $\tilde{\beta}, \ldots, \tilde{\beta}_p$ são obtidas maximizando esta soma. Supondo que ϕ seja constante para as n observações $y_1, \ldots, y_n$, obtém-se o sistema de equações para os $\tilde{\beta}'s$, que não dependem de ϕ

$$\sum_{i=1}^{n}(y_i-\mu_i)(\partial\mu_i/\partial\beta_i)/V_i(\mu_i)=0. \tag{4.18}$$

A maximização da log-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 o SPLUS, GENSTAT, GLIM, BMDP ou SAS, na pior das hipóteses utilizando sub-programas especiais.

Na análise de dados na forma de contagens trabalha-se com o erro de Poisson supondo que $\mathrm{Var}(y_i)=\phi\mu_i$. O parâmetro ϕ é estimado igualando a razão de quase-verossimilhanças $2\{Q(y;y)-Q(y;\tilde{\mu})\}$ aos graus de liberdade $(n-p)$ da χ^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 super-dispersão se $\tilde{\phi}>1$ e sub-dispersão em caso contrário. Similarmente, dados que apresentam durações de tempo com super-dispersão podem ser modelados por $\mathrm{Var}(y_i)=\phi\mu_i^2$ supondo $\phi>1$ e dados na

forma de contagens com sub-dispersão 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 da log-quase-verossimilhança (4.16) permite fazer comparações de modelos com preditores lineares diferentes ou com funções de ligação diferentes. Entretanto, não se pode 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 $\mathrm{D}(y;\mu)$, 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 MLGs, a função quase-desvio reduz-se aos desvios desses modelos.

A Tabela 4.2 apresenta log-quase-verossimilhanças para algumas funções de variância, com a exceção do escalar ϕ, deduzidas integrando (4.16). Desta tabela os desvios são facilmente obtidos.

Agora admite-se o seguinte modelo de quase-verossimilhança com função de variância paramétrica:

$$E(y_i) = \mu_i(\beta), \quad \mathrm{Var}(y_i) = \phi V_\lambda(\mu_i),$$

onde λ 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 uma componente aleatória independente extra ε de variância λ produzindo o preditor modificado $\eta^* = \eta + \varepsilon$. Até primeira

Tabela 4.2: *Log-quase-verossimilhanças associadas às funções de variância*

Função de Variância $V(\mu)$	Log-quase-Verossimilhança $Q(y;\mu)$
$\mu^{\lambda}(\lambda \neq 0,1,2)$	$\mu^{-\lambda}\left(\frac{y\mu}{1-\lambda} - \frac{\mu^2}{2-\lambda}\right)$
$\mu(1-\mu)$	$y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$
$\mu^2(1-\mu)^2$	$(2y-1)\log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$
$\mu + \mu^2/\alpha$	$y\log\left(\frac{\mu}{\alpha+\mu}\right) + \alpha\log\left(\frac{\alpha}{\alpha+\mu}\right)$

ordem, obtém-se a média e a variância modificadas $E(y)^* = \mu + \varepsilon d\mu/d\eta$ e $\mathrm{Var}(y)^* = \phi V(\mu) + \lambda(d\mu/d\eta)^2$ e, portanto, a função de variância torna-se parametrizada por λ. 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 μ 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 λ pode-se ainda utilizar as equações dadas em (4.18) para obter as estimativas de máxima quase-verossimilhança dos $\beta's$. A estimativa de λ corresponderá ao maior valor da quase-verossimilhança estendida maximizada tratada como função de λ, 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 β e λ, embora este processo exija o cálculo da função escore em relação ao parâmetro λ, o que é bastante complicado.

Considera-se agora uma classe de modelos de quase-verossimilhança com

parâmetro de dispersão não-constante

$$\eta = g(\mu) = X\beta, \quad \tau = h(\phi) = Z\gamma, \tag{4.19}$$

onde $\mu_i = E(y_i)$, $\mathrm{Var}(y_i) = \phi_i V(\mu_i)$, X e Z são matrizes $n \times p$ e $n \times q$ de posto completo p e q, β e γ 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 γ fixo pode-se utilizar (4.18) para obter as estimativas de máxima quase-verossimilhança dos $\beta's$ e, então, γ será escolhido visando maximizar a quase-verossimilhança estendida maximal $Q^+(\gamma)$ como função de γ. A estimativa de γ 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 β ou γ. As componentes quase-escore são dadas por

$$U_\beta^+ = \partial Q^+/\partial\beta = X^T W H(y-\mu), \quad U_\gamma^+ = \partial\gamma = \frac{1}{2} Z^T L(D-\phi),$$

onde $W = \mathrm{diag}\{\phi^{-1}V(\mu)^{-1}g'(\mu)^{-2}\}$, $H = \mathrm{diag}\{\phi^{-2}h'(\mu)^{-1}\}$ e $D = (D(y_1;\mu_1),\ldots,D(y_n;\mu_n))^T$. As estimativas de quase-verossimilhança de β e γ são obtidas resolvendo o sistema não-linear resultante da igualdade de U_β^+ e U_γ^+ 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

$$\tilde{X}^T \tilde{W}^{(m)} \tilde{X} \rho^{(m+1)} = \tilde{X}^T \tilde{W}^{(m)} \tilde{y}^{*(m)}, \tag{4.20}$$

onde

$$\tilde{X} = \begin{pmatrix} X & 0 \\ 0 & Z \end{pmatrix}, \ \tilde{W} = \begin{pmatrix} W & 0 \\ 0 & 1/2C \end{pmatrix},$$

$$\tilde{H} = \begin{pmatrix} H & 0 \\ 0 & C^{-1}L \end{pmatrix}, \quad \tilde{y}^* = \begin{pmatrix} \eta \\ \tau \end{pmatrix} + \tilde{H} \begin{pmatrix} y-\mu \\ D-\phi \end{pmatrix},$$

$C = \mathrm{diag}\{\phi^{-2}h'(\phi)^{-2}\}$. A matriz C tem elementos obtidos da aproximação de primeira ordem $E\{D(y;\mu)\} = 0$.

Assim, ajustar o modelo de quase-verossimilhança (4.19) 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

matriz de pesos $\tilde{W}$ que também se modifica no processo. A implementação de (4.20) pode ser feita usando os softwares já citados nesta seção. Estas mesmas equações (4.20) continuam válidas para os *modelos lineares generalizados duplos* que são definidos pela componente aleatória (2.1) e pelas duas componentes sistemáticas dadas em (4.19).

4.12 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, onde $S(t) = 1 - F(t)$, sendo $F(t)$ a função de distribuição acumulada de t; (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 descresce se $\alpha < 1$. O livro de Cox e Oakes (1984) apresenta um estudo completo da análise de dados de sobrevivência.

4.12.1 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

$$h(t;x) = \lambda(t)\exp(x^T\beta), \tag{4.21}$$

onde $\beta = (\beta, \ldots, \beta_p)^T$ é um vetor de parâmetros desconhecidos associados às covariáveis de $x = (x_1, \ldots, x_p)^T$, $\lambda(t)$ é uma função não-negativa do tempo e $\eta = x^T\beta$ é o preditor linear.

O modelo (4.21) 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

$$S(t;x) = \exp\{-\Lambda(t)\exp(x^T\beta)\}, \tag{4.22}$$

onde $\Lambda(t) = \int_{-\infty}^{t} \lambda(u)du$. Similarmente, a 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 (2.1). Em particular, $E\{\Lambda(t)\} = \exp(-\eta)$ e $\mathrm{Var}\{\Lambda(t)\} = \exp(-2\eta)$.

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 (4.22) como

$$\ell(\beta) = \sum_{j=1}^{n} (y_i \log\ \mu_i - \mu_i) + \sum_{j=1}^{n} \log\{\lambda(t_i)/\Lambda(t_i)\}, \tag{4.23}$$

onde $\mu_i = \Lambda(t_i) \exp(\eta_i)$. A segunda soma de (4.23) não depende dos $\beta's$ e, portanto, (4.23) tem a mesma forma da log-verossimilhança de um modelo de Poisson com n observações independentes $y_1, \ldots, y_n$, médias $\mu_1, \ldots, \mu_n$, e preditores lineares que são dados por $\eta_i = \log\ \Lambda(t_i),\ i = 1, \ldots, n$.

As estimativas de máxima verossimilhança para os $\beta's$ podem ser obtidas pelos sistemas GLIM e S-PLUS, 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 4.3 apresenta três modelos usuais para o tempo de sobrevivência. O modelo exponencial com λ conhecido pode ser ajustado diretamente. Se λ 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 λ. Um processo iterativo de estimação conjunta de λ e dos $\beta's$ pode ser realizado interagindo a estimativa de máxima verossimilhança de λ 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 λ, o termo $\log(\lambda)$ do "offset" pode ser incorporado à constante do preditor linear η_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 α desconhecido, a estimativa de máxima

verossimilhança de α é dada por

$$\hat{\alpha} = (n-m)/\sum_{i=1}^{n}(\hat{\mu}_i - y_i)\log t_i. \tag{4.24}$$

Admite-se uma estimativa inicial para α e ajusta-se a y, um modelo log-linear com "offset" $\alpha \log t$. De (4.24) reestima-se α, continuando o processo até a convergência.

Tabela 4.3: *Alguns modelos usuais para a análise de dados de sobrevivência*

Modelo	$\lambda(t)$	densidade	"offset"
exponencial	λ	$\lambda \exp\{\eta - \lambda t \exp(\eta)\}$	$log(\lambda t)$
Weibull	$\alpha t^{\alpha-1}$	$\alpha t^{\alpha-1} \exp\{\eta - t^{\alpha} \exp(\eta)\}$	$\alpha \log t$
valor-extremo	$\alpha \exp(\alpha t)$	$\alpha \exp\{\eta - t^{\alpha} \exp(\alpha t + \eta)\}$	αt

O modelo de valor extremo pode ser transformado no de Weibull com a transformação $\exp(t)$, no lugar de t.

4.12.2 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 (4.21). 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 (4.24). 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 $t_1, t_2, \ldots, t_k$ 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 (4.21), dessa falha ter ocorrido com o i-ésimo indivíduo, é dada por

$$P_j = \frac{\lambda(t)\exp(x_i^T\beta)}{\sum\limits_{s\in R(t_j)} \lambda(t)\exp(x_s^T\beta)} = \frac{\exp(x_i^T\beta)}{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta)},$$

onde 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 (2.1), 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 β 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 2.4 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 (2.11).

Em McCullagh e Nelder (1989) há uma discussão sobre métodos iterativos para a estimaçao de β. 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 (4.23) 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 a log-verossimilhança $\ell(\beta)$ pode ficar expressa 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,

$$\begin{aligned} P_{j(\text{Real})} &= \frac{\exp(x_1^T\beta)}{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta)} \cdot \frac{\exp(x_2^T\beta)}{\left\{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta) - \exp(x_1^T\beta)\right\}} \\ &+ \frac{\exp(x_2^T\beta)}{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta)} \cdot \frac{\exp(x_1^T\beta)}{\left\{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta) - \exp(x_2^T\beta)\right\}}. \end{aligned}$$

Cox (1975) mostra que toda a teoria usual para a estatística da razão de verossimilhanças continua valendo para os modelos de riscos proporcionais.

4.13 Modelos Lineares Generalizados com Covariáveis de Dispersão

Jørgensen (1987) definiu a classe dos *modelos de dispersão*, inicialmente denominada classe estendida de MLGs (Jørgensen, 1983), considerando um conjunto de variáveis aleatórias $Y_1, \ldots, Y_n$ com cada Y_ℓ tendo função densidade

(ou função de probabilidade) na forma

$$\pi(y;\theta_l,\phi) = \exp\{\phi t(y,\theta_l) + c_1(y,\phi)\}, \tag{4.25}$$

onde $t(\cdot\ ,\ \cdot)$ e $c_1(\cdot\ ,\ \cdot)$ são funções conhecidas. Consideramos que $\phi\,(\phi > 0)$ é constante para todas as observações embora, possivelmente, desconhecido. Denominamos ϕ^{-1} de parâmetro de dispersão e ϕ de parâmetro de precisão. Segundo Jørgensen (1983) os modelos definidos em (4.25) incluem a possibilidade de erros correlacionados. Entretanto, se as variáveis aleatórias $Y_1,\ldots,Y_n$ forem independentes, com cada variável tendo uma distribuição da forma (4.25), a distribuição conjunta de $Y_1,\ldots,Y_n$ será também da forma (4.25).

Fazendo $t(y,\theta) = y\theta - b(\theta)$ em (4.25), obtemos a subclasse dos *modelos exponenciais de dispersão* (Jørgensen, 1987) ou MLGs. Para ϕ conhecido, os modelos exponenciais de dispersão pertencem à família exponencial de distribuições, sendo θ o seu parâmetro canônico. Se ϕ 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, onde a função $c_1(y,\phi)$ em (4.25) 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 θ, isto é, a distribuição de $t(y,\theta)$ não depende de θ para ϕ conhecido. A segunda revela que, para θ conhecido, a função densidade (ou probabilidade) definida em (4.25) pertence à família exponencial uniparamétrica sendo $t(y,\theta)$ uma estatística canônica.

Sejam $Y_1,\ldots,Y_n$ um conjunto de n variáveis aleatórias independentes com cada Y_ℓ tendo função densidade (ou função de probabilidade) na família exponencial

$$\pi(y;\theta_l,\phi_l) = \exp[\phi_l\{y\theta_l - b(\theta_l) + c(y)\} + d_1(y) + d_2(\phi_l)], \tag{4.26}$$

onde $b(\cdot)$, $c(\cdot)$, $d_1(\cdot)$ e $d_2(\cdot)$ são funções conhecidas e θ_l e ϕ_l são, respectivamente, os l-ésimos elementos de θ e ϕ, vetores de dimensão $n\times 1$. A média e a variância de Y_l são $E(Y_l) = \mu_l = db(\theta_l)/d\theta_l$ e $\mathrm{Var}(Y_l) = \phi_l^{-1}V_l$, onde

$V = d\mu/d\theta$ e $\theta = \int V^{-1} d\mu = q(\mu)$ é uma função conhecida unívoca de μ. A componente sistemática usual para a média é $f(\mu) = \eta = X\beta$, onde $f(\cdot)$ é a função de ligação, $\eta = (\eta_1, \ldots, \eta_n)^T$ é o preditor linear, X é uma matriz conhecida $n \times p$ de posto $p < n$ e $\beta = (\beta_1, \ldots, \beta_p)^T$ é um vetor de parâmetros desconhecidos a ser estimado. Os parâmetros θ_l e $\phi_l^{-1} > 0$ 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 $\theta_l^{-1}, -\theta_l^{-1}, (-2\theta_l)^{-1/2}$ e $\phi_l^{-1}, \phi_l^{-1}\mu_1^2$ e $\phi_l^{-1}\mu_1^3$, respectivamente.

Definimos a componente sistemática do vetor de parâmetros de precisão $\phi = (\phi_1, \ldots, \phi_n)^T$ como

$$g(\phi) = \tau = S\gamma, \tag{4.27}$$

onde τ é o preditor linear da dispersão, $S = (s_1, \ldots, s_n)^T$, com $s_l = (s_{l1}, \ldots, s_{lp})^T$, é uma matriz $n \times q$ de posto q $(q < n)$ representando as variáveis independentes que modelam a dispersão e $\gamma = (\gamma_1, \ldots, \gamma_q)^T$ é, também, um vetor de parâmetros desconhecidos. O MLG com covariáveis de dispersão tem, portanto, dois preditores lineares: η – o preditor linear da média e τ – 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 β é independente de γ. Temos, então, $p + q$ parâmetros a serem estimados.

Considere a log-verossimilhança total como função de β e γ

$$\ell(\beta, \gamma) = \sum_{l=1}^{n} [\phi_l \{y_l \theta_l - b(\theta_l) + c(y_l)\} + d_1(y_l) + d_2(\phi_l)],$$

sendo o vetor de dados $y = (y_1, \ldots, y_n)^T$ fixado, onde y_l denota o valor observado da variável aleatória Y_l. Na expressão acima, θ está associado a β através da função de ligação $f(\cdot)$ (θ é uma função de μ) e ϕ está relacionado com γ através de $g(\cdot)$.

Denotamos a função escore total por

$$U = U(\beta, \gamma) = \begin{pmatrix} \partial\ell(\beta,\gamma)/\partial\beta \\ \partial\ell(\beta,\gamma)/\partial\gamma, \end{pmatrix}$$

cujas componentes são

$$\partial\ell(\beta,\gamma)/\partial\beta = X^T\Phi W^{1/2}V^{-1/2}(y-\mu) \quad \text{e} \quad \partial\ell(\beta,\gamma)/\partial\gamma = S^T\Phi_1 v,$$

onde $\Phi = \text{diag}\{\phi_1, \ldots, \phi_n\}$, $W = \text{diag}\{w_1, \ldots, w_n\}$ com $w_l = V_l^{-1}(d\mu_l/d\eta_l)^2$, $V = \text{diag}\{V_1, \ldots, V_n\}$, $\Phi_1 = \text{diag}\{\phi_{1l}, \ldots, \phi_{1n}\}$ com $\phi_{1l} = \partial\phi_l/\partial\eta_l$ e $v = (v_1, \ldots, v_n)^T$ com $v_l = y_l\theta_l - b(\theta_l) + c(y_l) + \partial d_2(\phi_l)/\partial\phi_l$.

A partição (β^T, γ^T) induz uma correspondente matriz de informação particionada para estes parâmetros. A matriz de informação total de Fisher $K = K(\beta,\gamma)$ pode ser deduzida de $E\{U(\beta,\gamma)U^T(\beta,\gamma)\}$. Esta matriz é bloco-diagonal dada por

$$K(\beta,\gamma) = \begin{pmatrix} K_{\beta,\beta} & 0 \\ 0 & K_{\gamma,\gamma}, \end{pmatrix}$$

onde $K_{\beta,\beta} = X^TW\Phi X$ e $K_{\gamma,\gamma} = -S^TD_2\Phi_1^2S$, sendo $D_2 = \text{diag}\{d_{21}, \ldots, d_{2n}\}$, $d_{2l} = \partial^2 d_2(\phi_l)/\partial\phi_l^2$ e $\Phi_1^2 = \text{diag}\{\phi_{1l}^2, \ldots, \phi_{1n}^2\}$, são as matrizes de informação para β e γ, respectivamente. Os parâmetros β e γ são globalmente ortogonais e suas estimativas de máxima verossimilhança são assintoticamentes independentes (Cox e Reid, 1987).

Os estimadores de máxima verossimilhança $\hat{\beta}$ e $\hat{\gamma}$ podem ser calculados através do processo iterativo escore de Fisher, resolvendo as seguintes equações

$$\begin{bmatrix} \hat{\beta}^{(m+1)} \\ \hat{\gamma}^{(m+1)} \end{bmatrix} = \begin{bmatrix} \hat{\beta}^{(m)} \\ \hat{\gamma}^{(m)} \end{bmatrix} + K^{(m)-1}U^{(m)}. \tag{4.28}$$

As equações (4.28) implicam na solução iterativa do sistema de equações

$$\tilde{X}^T\tilde{W}^{(m)}\tilde{X}\rho^{(m+1)} = \tilde{X}^T\tilde{W}^{(m)}\tilde{y}^{*m},$$

onde

$$\tilde{X} = \begin{bmatrix} X & 0 \\ 0 & -S \end{bmatrix}, \quad \tilde{W} = \begin{bmatrix} \Phi W & 0 \\ 0 & D_2\Phi_1^2 \end{bmatrix},$$

$$\tilde{\Phi} = \begin{bmatrix} W^{-1/2}V^{-1/2} & 0 \\ 0 & -D_2^{-1}\Phi_1^{-1} \end{bmatrix}, \quad \rho = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$$

e

$$\tilde{y}^* = \begin{bmatrix} \eta \\ \tau \end{bmatrix} + \tilde{\Phi} \begin{bmatrix} y - \mu \\ v \end{bmatrix}. \tag{4.29}$$

Em geral, temos que fazer a regressão da variável dependente modificada dada por (4.29) na matriz modelo $\tilde{X}$ usando os pesos modificados definidos por $\tilde{W}$. A variável dependente modificada $\tilde{y}^*$ também varia durante o procedimento iterativo e deve ser recalculada em toda repetição. O ajuste do modelo com covariáreis de dispersão no GLIM é feito usando quatro macros, definindo o modelo pelo usuário. O procedimento inicial é feito pela escolha de valores arbitrários para β e γ.

4.14 Modelos Lineares Generalizados com Super-dispersão

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 conseqüência 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 MLGs com uma relação variância-média especificada e com um parâmetro de dispersão multiplicativo, muitas vezes obtém-se um ajustamento do modelo onde a variância é maior do que o preditor da média.

Dey et al. (1997) definiram uma classe de *MLGs com super-dispersão* onde as variáveis aleatórias $Y_1, \ldots, Y_n$ são independentes e cada Y_i tem densidade (ou função de probabilidade) com dois parâmetros pertencente à família exponencial

$$\pi(y; \mu, \phi) = A(y) \exp\{(y - \mu)\psi^{(1,0)}(\mu, \phi) + \phi T(y) + \psi(\mu, \phi)\}, \tag{4.30}$$

onde $A(\cdot)$, $T(\cdot)$ e $\psi(\cdot, \cdot)$ são funções conhecidas e $\psi^{(r,s)} = \partial\psi^{r+s}(\mu, \phi)/\partial\mu^r\partial\phi^s$. A média e a variância de Y são $E(Y) = \mu$ e $\mathrm{Var}(Y) = \psi^{(2,0)^{-1}}$, e a média e a variância de $T(Y)$ são $E\{T(Y)\} = -\psi^{(0,1)}$ e $\mathrm{Var}\{T(Y)\} = -\psi^{(0,2)}$. Além disso, $\mathrm{Cov}(Y,\ T(Y)) = 0$.

Gelfand e Dalal (1990) mostraram que se (4.30) é integrável em relação a y e se a função $T(y)$ é convexa, tendo a média μ fixa, então a $\mathrm{Var}(Y)$ aumenta com ϕ.

A família exponencial uniparamética é obtida de (4.30) com $\phi = 0$, conduzindo a forma

$$\pi(y; \phi, 0) = A(y) \exp\{y\theta - b(\theta)\},$$

onde $\theta = \psi^{(1,0)}(\mu, 0)$ e $b(\theta) = -\psi(\mu, 0) + \mu\psi^{(1,0)}(\mu, 0)$.

Considera-se MLGs com super-dispersão que têm duas componentes sistemáticas que são parametrizadas como $f(\mu) = \eta = X\beta$ e $g(\phi) = \tau = S\gamma$, onde X e S são matrizes $n \times p$ e $n \times q$, de postos p e q, respectivamente, $\beta = (\beta_1, \ldots, \beta_p)^T$ e $\gamma = (\gamma_1, \ldots, \gamma_q)^T$ são vetores de parâmetros desconhecidos a serem estimados. Considera-se que $f(\cdot)$ e $g(\cdot)$ são funções monótonas conhecidas e diferenciáveis e que β é independente de γ. A função $g(\cdot)$ é uma função de ligação adicional chamada de função de ligação de dispersão. O MLG é baseado na família exponencial (2.1) de um parâmetro assumindo ϕ fixo onde $\theta = q(\mu)$ é o parâmetro natural, $\mu = \frac{db(\theta)}{d\theta}$ é a média e ϕ é o parâmetro de precisão comum para todas as observações, embora possivelmente descon-

hecido. As únicas distribuições contínuas da forma (2.1) são baseadas nas distribuições normal, gama e Gaussiana inversa.

Note-se que a família de distribuições em (2.1) é uma sub-família simples de (4.30) e difere desta no sentido de que tem uma forma geral de dois parâmetros para modelos exponenciais, enquanto (2.1) é apenas um modelo exponencial de um parâmetro θ quando ϕ é mantido fixo. Entretanto, como um modelo de dois parâmetros (θ, ϕ), (4.30) 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 β e γ simultaneamente, desde que eles representam os efeitos das variáveis explicativas da média e do parâmetro de dispersão, respectivamente. Denotamos a amostra aleatória por $y_1, \ldots, y_n$ e a função de log-verossimilhança total por

$$\ell(\beta,\gamma) = \sum_{l=1}^{n}\{(y_l-\mu_l)\psi^{(1,0)}(\mu,\phi_l)+\phi_l T(y_l)+\psi(\mu_l,\phi_l)\}+\sum_{l=1}^{n}\log\, A(y_l). \quad (4.31)$$

Esta função é suposta regular (Cox e Hinkley, 1974; Capítulo 9) com relação às derivadas em β e γ até terceira ordem. A inferência sobre β e γ pode ser feita através do método de verossimilhança, análogos aos dos MLGs com covariáveis de dispersão (Cordeiro e Botter, 2000). O vetor escore é dado na forma

$$U = U(\beta,\gamma) = \begin{pmatrix} \dfrac{\partial \ell(\beta,\gamma)}{\partial \beta} \\ \\ \dfrac{\partial \ell(\beta,\gamma)}{\partial \gamma} \end{pmatrix} = \begin{pmatrix} X^T \psi^{(2,0)} M_1 (y-\mu) \\ \\ S^T \Phi_1 \nu \end{pmatrix}, \quad (4.32)$$

onde $y-\mu = (y_1-\mu_1, \ldots, y_n-\mu_n)^T$ e $v = (v_1, \ldots, v_n)^T$ com $v_\ell = \psi_\ell^{(1,1)}(y_\ell - \mu_\ell) + T(y_\ell) + \psi_\ell^{(0,1)}$. E mais, $m_{il} = \dfrac{d^i \mu_l}{d\eta_l^i}$ e $\phi_{il} = \dfrac{d^i \phi_l}{d\tau_l^i}$ são, respectivamente, as derivadas das funções de ligação inversas $\mu = f^{-1}(\eta)$ e $\phi = g^{-1}(\tau)$, $\quad i = 1, 2$

e $l = 1, \ldots, n$. Definimos, também, as seguintes matrizes diagonais $n \times n$: $M_i = \text{diag}\{m_{i1}, \ldots, m_{in}\}$ e $\Phi_i = \text{diag}\{\phi_{i1}, \ldots, \phi_{in}\}$ para $i = 1, 2$ e $\psi^{(2,0)} = \text{diag}\{\psi_1^{(2,0)}, \ldots, \psi_n^{(2,0)}\}$ e $\psi^{(0,2)} = \text{diag}\{\psi_1^{(0,2)}, \ldots, \psi_n^{(0,2)}\}$.

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

$$K(\beta, \gamma) = \begin{pmatrix} K_{\beta,\beta} & 0 \\ 0 & K_{\gamma,\gamma} \end{pmatrix}, \tag{4.33}$$

onde $K_{\beta,\beta} = X^T \psi^{(2,0)} M_1^2 X$ e $K_{\gamma,\gamma} = S^T \psi^{(0,2)} \Phi_1^2 S$ são as matrizes de informação de β e γ, respectivamente. Deste modo, os parâmetros β e γ são ortogonais e suas estimativas de máxima verossimilhança $\hat{\beta}$ e $\hat{\gamma}$ são assintoticamente independentes.

As EMVs $\hat{\beta}$ e $\hat{\gamma}$ satisfazem equações não-lineares $U(\hat{\beta}, \hat{\gamma}) = 0$ que derivam de (4.32) e que podem ser resolvidos pelo método escore de Fisher. Com isso, Cordeiro e Botter (2000) obtiveram as seguintes equações para estimar iterativamente β e γ

$$\begin{aligned} X^T \psi^{(2,0)(m)} M_1^{(m)^2} X \beta^{(m+1)} &= X^T \psi^{(2,0)(m)} M_1^{(m)^2} \varepsilon_1^{(m)}, \\ S^T \psi^{(0,2)(m)} \Phi_1^{(m)^2} S \gamma^{(m+1)} &= S^T \psi^{(0,2)(m)} \Phi_1^{(m)^2} \varepsilon_2^{(m)}, \end{aligned} \tag{4.34}$$

onde $\varepsilon_1 = \eta + M_1^{-1}(y - \mu)$ e $\varepsilon_2 = \tau + \psi^{(0,2)^{-1}} \Phi_1^{-1}$ são vetores $n \times 1$.

As equações (4.34) mostram que qualquer software contendo uma regressão linear ponderada pode ser usado para calcular as estimativas $\hat{\beta}$ e $\hat{\gamma}$. Em termos gerais, temos que fazer a regressão da variável dependente modificada $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$ sobre a matriz modelo $(X \;\; S)$ com os pesos modificados

definidos por

$$\begin{pmatrix} \psi^{(2,0)} M_1^2 & 0 \\ 0 & \psi^{(0,2)} \Phi_1^2 \end{pmatrix}.$$

Este ciclo será repetido até convergência. O procedimento de iteração em (4.33) é mais fácil de ser executado usando o algoritmo em linguagem GLIM seguindo as mesmas linhas descritas em Cordeiro e Paula (1989) e Cordeiro e Demétrio (1989). Para definir o MLG com super-dispersão no GLIM usa-se a diretiva que declara o próprio modelo do usuário por quatro macros. O início do procedimento é executado escolhendo valores arbitrários para β e γ.

4.15 Exercícios

1. Ajustar um modelo de regressão aos dados do volume V de árvores de cereja preta em termos da altura A e do diâmetro D (Ryan et al., 1985) apresentados abaixo:

V	8.300	11.200	13.700	17.900	8.600	11.300	13.800
A	70.00	75.00	71.00	80.00	65.00	79.00	64.00
D	10.30	19.90	25.70	58.30	10.30	24.20	24.90
V	18.00	8.800	11.400	14.000	18.000	10.500	11.400
A	80.00	63.00	76.00	78.00	80.00	72.00	76.00
D	51.50	10.20	21.00	34.50	51.00	16.40	21.40
V	14.20	20.600	10.700	11.700	14.500	10.800	12.000
A	80.00	87.00	81.00	69.00	74.00	83.00	75.00
D	31.70	77.00	18.80	21.30	36.30	19.70	19.10
V	16.000	11.000	12.900	16.300	11.000	12.900	17.300
A	72.00	66.00	74.00	77.00	75.00	85.00	81.00
D	38.30	15.60	22.20	42.60	18.20	33.80	55.40
V	11.100	13.300	17.500				
A	80.00	86.00	82.00				
D	22.60	27.40	55.70				

   Fazer uma análise desses dados via o modelo de Box e Cox (1964).

2. Analisar os dados seguintes (Freedman, Pisani e Purves, 1978) sobre a admissão de estudantes em 6 cursos de graduação da Universidade da

Califórnia.

	Homens		Mulheres	
Curso	Inscritos	Admitidos	Inscritos	Admitidos
A	825	512	108	89
B	560	353	25	17
C	325	121	393	134
D	417	138	375	131
E	191	53	393	94
F	373	22	341	24

3. Ajuste o modelo logístico linear simples ao seguinte conjunto de dados:

x_i	0	20	25	30	35	40
y_i	0	2	5	6	6	7
n_i	7	8	8	8	8	8

4. (a) Mostrar que os 9 modelos hierárquicos abaixo, correspondentes à classificação dos 4 fatores A,B,C e D não têm forma fechada; (b) Verificar ainda as expressões dos graus de liberdade do desvio;

Classe geradora	Graus de liberdade
AB, AC, BC, D	IJKL-IJ-JK-IK-L+I+J+K
AB, AC, BC, CD	IJKL-IJ-JK-IK-KL+I+J+2K-1
AB, AC, BC, BD, CD	IJKL-IJ-JK-IK-JL-KL+I+2J+2K+L-2
AB, AC, AD, BC, BD, CD	IJKL-IJ-IK-IL-JK-JL-KL+2(I+J+K+L)-3
ABC, BD, CD	IJKL-IJK-JL-KL+J+K+L-1
ABC, AD, BD, CD	IJKL-IJK-IL-JL-KL+I+J+K+2L-2
ABC, ABD, CD	(IJ-1)(K-1)(L-1)
ABC, ABD, BCD	(IJ-J+1)(K-1)(L-1)
ABC, ABD, ACD, BCD	(I-1)(J-1)(K-1)(L-1)

(c) Interpretar os modelos acima.

5. Demonstrar que para o modelo logístico-linear o desvio reduz-se à expressão $S_p = -2\sum_{\ell=1}^{n}[\hat{\mu}_\ell \log \hat{\mu}_\ell + (1-\hat{\mu}_\ell)\log(1-\hat{\mu}_\ell)]$.

6. Demonstrar que o desvio do modelo correspondente à hipótese de interação zero entre os três fatores de uma classificação de três entradas numa tabela $I \times J \times K$, é dado por:

$$\begin{aligned} S_p \;=\; 2\Bigg(& \sum_{i,j,k} y_{ijk} \log y_{ijk} - \sum_{j,k} y_{+jk} \log y_{+jk} - \sum_{i,k} y_{i+k} \log y_{i+k} - \\ & \sum_{i,j} y_{ij+} \log y_{ij+} + \sum_{i} y_{i++} \log y_{i++} + \sum_{j} y_{+j+} \log y_{+j+} + \\ & \sum_{k} y_{++k} \log y_{++k} - y_{+++} \log y_{+++} \Bigg), \end{aligned}$$

onde $p = IJK - (I-1)(J-1)(K-1)$. Demonstrar que S_p converge em distribuição para a variável $\chi^2_{(I-1)(J-1)(K-1)}$ quando y_{+++} tende para ∞, se e somente se, a tabela é perfeita, no sentido de que $\mu_{ijk} = \mu_{+jk}\mu_{i+k}\mu_{ij+}/\mu_{i++}\mu_{+j+}\mu_{++k}$.

7. Analisar os dados abaixo referentes a quatorze estudos retrospectivos sobre a associação entre o fumo e o câncer no pulmão.

	Estudo	1	2	3	4	5	6	7
Pacientes	total	86	93	136	82	444	605	93
com câncer	não-fumantes	3	3	7	12	32	8	5
	total	86	270	100	522	430	780	186
controle	não-fumantes	14	43	19	125	131	114	12
	Estudo	8	9	10	11	12	13	14
Pacientes	total	1357	63	477	728	518	490	265
com câncer	não-fumantes	7	3	18	4	19	39	5
	total	1357	133	615	300	518	2365	287
controle	não-fumantes	61	27	81	54	56	636	28

8. Analisar os dados abaixo referentes as freqüências observadas de moças da Nova Zelândia por faixa etária e pelo estágio de desenvolvimento do busto (1 = imaturo, 5 = completamente desenvolvido).

		Idade				
		10-10.99	11-11.99	12-12.99	13-13.99	14-14.99
	1	621	292	132	50	27
Desenvolvimento	2	251	353	273	182	69
do busto	3	50	214	337	397	273
	4	7	72	160	333	501
	5	0	5	39	132	289

9. Analisar os dados abaixo referentes aos números de acidentes com motoristas, sem acompanhantes, classificados por tipo e severidade do acidente, peso do carro e estado do motorista após o acidente.

		classificação do acidente			
	motorista jogado	colisão		capotagem	
peso do carro	para fora	grave	não-grave	grave	não-grave
	sim	23	26	80	19
pequeno	não	150	350	112	60
	sim	161	111	265	22
padrão	não	1022	1878	404	148

10. Analisar os dados seguintes relativos aos números de crianças do 1º grau da cidade do Recife, classificadas por escola e pela renda familiar mensal dos pais. As escolas A e B são particulares e C, D e E são públicas. Os dados foram coletados em junho/1985 (Cordeiro, 1986, Capítulo 6)

	Renda familiar mensal em salário mínimos				
Escola	1 – 4	5 – 8	9 – 12	13 – 16	17 ou mais
A	3	74	108	124	56
B	0	47	95	171	112
C	108	147	121	19	5
D	189	127	8	2	0
E	37	98	137	34	7

Capítulo 5

Outros Modelos de Regressão Importantes

Neste capítulo descrevemos cinco tipos de modelos de regressão bastante usados na análise de dados. Os modelos são: modelos com matriz de covariância não-escalar (Seção 5.1), modelo de regressão rígida (Seção 5.2), modelo normal não-linear (Seção 5.3), modelos heterocedásticos (Seção 5.4) e modelos autocorrelacionados (Seção 5.5).

5.1 Modelos com Matriz de Covariância Não-Escalar

Considera-se o modelo de regressão

$$y = X\beta + \varepsilon, \quad E(\varepsilon) = 0, \quad Cov(\varepsilon) = \Psi = \sigma^2\psi, \tag{5.1}$$

onde ambos σ^2 e ψ são desconhecidos. No caso mais geral, ψ 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 $\Psi = \sigma^2\psi$. Dois casos especiais importantes de (5.1) são os modelos

heterocedásticos e os modelos de autocorrelação descritos nas Seções 5.4 e 5.5, respectivamente. Se ψ for conhecido, o *estimador de mínimos quadrados generalizado* (EMQG) será $\hat{\beta} = (X^T\psi^{-1}X)^{-1}X^T\psi^{-1}y$ que é o estimador de mínima variância na classe dos estimadores lineares não-viesados de β. Se ε tem, também, distribuição normal, então $\hat{\beta}$ é o EMV sendo de mínima variância na classe dos estimadores não-viesados. Adicionalmente, $\hat{\sigma}^2 = (y - X\hat{\beta})^T\psi^{-1}(y - X\hat{\beta})/n$ é o estimador viesado de σ^2. Se o interesse é testar a hipótese nula de restrições lineares $H_0 : R\beta = 0$, onde R é uma matriz $r \times p$ de coeficientes conhecidos, a estatística

$$F = \hat{\beta}^T R^T [R(X^T\psi^{-1}X)^{-1}R^T]^{-1}R\beta/r\hat{\sigma}^2$$

tem distribuição nula $F_{r,\,n-p}$, que pode ser usada tanto para testar H_0 quanto na estimação restrita de intervalos para β.

Quando ψ é desconhecido, situação mais comum na prática, o EMQG dado anteriormente é inviável. Neste caso, pode-se formar o estimador

$$\hat{\hat{\beta}} = (X^T\hat{\psi}^{-1}X)^{-1}X^T\hat{\psi}^{-1}y, \tag{5.2}$$

onde a matriz de covariância desconhecida ψ é substituída em (5.2) por um estimador consistente $\hat{\psi}$. Como o número de parâmetros desconhecidos em ψ é de ordem $O(n)$, em geral restringe-se o número desses parâmetros supondo que ψ é função de um vetor γ de $q+1$ parâmetros desconhecidos.

Vamos considerar a estimação de máxima verossimilhança (MV) de β, σ^2 e γ no modelo

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2\psi(\gamma)), \tag{5.3}$$

onde enfatizamos em (5.3) que a matriz ψ depende de um vetor $q \times 1$ de parâmetros extras desconhecidos. A estimação de MV de β e σ^2 condicional a γ produz os estimadores

$$\tilde{\beta}(\gamma) = (X^T\psi(\gamma)^{-1}X)^{-1}X^T\psi(\gamma)^{-1}y \tag{5.4}$$

e

$$\tilde{\sigma}(\gamma)^2 = (y - X\tilde{\beta}(\gamma))^T\psi(\gamma)^{-1}(y - X\tilde{\beta}(\gamma))/n. \tag{5.5}$$

Usamos a notação $\tilde{\beta}(\gamma), \tilde{\sigma}^2(\gamma)$ e $\psi(\gamma)$ acima para enfatizar a dependência destas quantidades em γ. A log-verossimilhança perfilada para γ é

$$\ell_p(\gamma) = -n \ \log\{\tilde{\sigma}(\gamma)^2\} - \log\{\psi(\gamma)\}. \tag{5.6}$$

A maximização de (5.6), 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 β e σ^2 são facilmente deduzidos de (5.4) – (5.5) 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

$$I(\theta) = \begin{pmatrix} \sigma^{-2} X^T \psi^{-1} X & 0 & 0 \\ 0 & \frac{n}{2\sigma^4} & \frac{1}{2}\sigma^{-2}\mathrm{vec}(\psi^{-1})^T A \\ 0 & \frac{1}{2}\sigma^{-2} A^T \mathrm{vec}(\psi^{-1}) & \frac{1}{2} A^T (\psi^{-1} \otimes \psi^{-1}) A \end{pmatrix},$$

onde $A = A(\gamma) = \mathrm{vec}\left(\frac{\partial \psi(\gamma)}{\partial \gamma^T}\right)$, $\otimes$ representa o produto de Kronecker e o operador vec $(\cdot)$ transforma as colunas de uma matriz em vetor.

No modelo (5.1), deseja-se agora testar a hipótese geral

$$H_0 : g(\theta) = 0 \quad versus \quad H_1 : g(\theta) \neq 0,$$

onde g é um vetor $r \times 1$. Seja F a matriz $(p+q+1) \times r$ dada por $F = \frac{\partial g(\theta)^T}{\partial \theta}$. A estatística de Wald é definida por

$$W = g(\hat{\theta})^T (\hat{F}^T I(\hat{\theta})^{-1} \hat{F})^{-1} g(\hat{\theta}),$$

onde $\hat{\theta}$ é o EMV irrestrito de θ, $\hat{F}$ é a matriz F avaliada em $\theta = \hat{\theta}$ e $I(\hat{\theta})$ é a informação em $\hat{\theta}$. A distribuição nula assintótica de W é χ^2_r.

Uma estatística alternativa a de Wald é a estatística escore de Rao que envolve o EMV restrito $\tilde{\theta}$. Seja $U(\theta)$ a função escore para θ, i.e., $U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$. A estatística escore para testar H_0 é dada por

$$S_R = U(\tilde{\theta})^T I(\tilde{\theta})^{-1} U(\tilde{\theta}),$$

que, também, tem distribuição nula assintótica igual a χ^2_r.

O teste da razão de verossimilhanças equivale ao uso da estatística

$$w = 2\{\ell(\hat{\theta}) - \ell(\tilde{\theta})\}.$$

As três estatísticas W, S_R e w têm propriedades assintóticas, em geral, equivalentes. Em vários modelos de regressão do tipo (5.1), 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 β, ou seja, $H_0 : R\beta = 0$ e que σ^2 e ψ são conhecidos. Neste caso, as três estatísticas de teste, W, S_R e w são idênticas e reduzem-se a

$$W = S_R = w = \tilde{\beta}^T R^T [R(X^T \psi^{-1} X)^{-1} R^T]^{-1} R\tilde{\beta}/\sigma^2,$$

onde $\tilde{\beta} = (X^T \psi^{-1} X)^{-1} X^T \psi^{-1} y$ é o EMV de β quando ψ é conhecido.

5.2 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 $X^T X$. 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 $(X^T X + kI)\hat{\beta} = X^T y$, que produz $\beta^* = (X^T X + kI)^{-1} X^T y$. Sejam $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ os autovalores ordenados de $X^T X$ e $v_1, \ldots, v_p$ seus autovetores correspondentes. Pode-se demonstrar que

$$(X^T X + kI)^{-1} = \sum_{i=1}^{p} (\lambda_i + k)^{-1} v_i v_i^T,$$

revelando que se $X^T X$ é quase singular com λ_p pequeno, então, o menor autovalor de $X^T X + kI$ será $\lambda_p + k$ e esta última matriz não será tão próxima da singularidade.

Sejam V e $\wedge$ as matrizes dos autovetores e autovalores de X^TX, ou seja, $V = (v_1, \ldots, v_p)$ e $\wedge = diag\{\lambda_1, \ldots, \lambda_p\}$. O erro médio quadrático (EMQ) de β^* é dado por

$$\text{EMQ}(\beta^*) = tr(V(\beta^*)) + \{E(\beta^*) - \beta\}^T\{E(\beta^*) - \beta\},$$

onde $V(\beta^*) = \sigma^2 W X^T X W$ e $W = V(\wedge + kI)^{-1}V^T$. Tem-se, ainda, $V(\beta^*) = \sigma^2 V \wedge_* V^T$, onde $\wedge_* = diag\{\lambda_i(\lambda_i+k)^{-2}\}$ e, então, $tr(V(\beta^*)) = \sum \lambda_i(\lambda_i+k)^{-2}$. Mas $\beta^* = WX^TX\hat{\beta}$, onde $\hat{\beta} = (X^TX)^{-1}X^Ty$ é o estimador de MQ de β.

Assim,

$$E[\{E(\beta^*) - \beta\}^T\{E(\beta^*) - \beta\}] = \beta^T V \wedge_+ V^T\beta,$$

onde $\wedge_+ = diag\{k^2(\lambda_i + k)^{-2}\}$. Finalmente,

$$EMQ(\beta^*) = \sum(\lambda_i^2 + \gamma_i k^2)(\lambda_i + k)^{-2},$$

onde $\gamma = (\gamma_1, \ldots, \gamma_p)^T = \beta^T V$.

Temos que a variância de β^* 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 $EMQ(\beta^*) \leq EMQ(\hat{\beta})$. Esta 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 ${\beta^*}^T\beta^* \longrightarrow 0$ quando k cresce. Assim, o estimador de regressão rígida tende a origem quando k cresce. Temos ainda que

$$\beta^* = \sum_{i=1}^{p} \frac{1}{\lambda_i + k} d_i v_i,$$

onde $d_i = v_i^T X^T y$. Assim, determinando-se os autovalores e autovetores de X^TX, os estimadores de regressão rígida serão obtidos para qualquer valor de k. Define-se o traço rígido como um gráfico de β^* versus k para valores crescentes de k. Quando $k = 0$, tem-se o estimador de MQ de β. Com base no traço rígido pode-se escolher como valor de k o ponto onde as estimativas em β^* estão estabilizadas.

5.3 Modelo Normal Não-Linear

Até o início da década de 70 as principais técnicas desenvolvidas para os modelos de regressão não-lineares se restringiam à suposição de normalidade para a variável resposta. Em 1972, Nelder e Wedderburn ampliaram a distribuição da variável resposta para a família exponencial de distribuições, definindo os Modelos Lineares Generalizados. Mesmo assim, os modelos normais não-lineares continuaram recebendo um tratamento especial, surgindo diversos trabalhos nas décadas de 70 e 80, destacando-se o livro de Ratkowsky (1983).

A principal característica dos modelos não-lineares é que eles são deduzidos a partir de suposições teóricas (quase sempre equações diferenciais) e os parâmetros resultantes são interpretáveis. Assim, aproximá-los pelos 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 aproximadamente normal. Isso levou ao desenvolvimento de técnicas específicas de diagnóstico para os modelos normais não-lineares (Cook e Tsai, 1985). Similarmente, as propriedades das somas de quadrados contidas nas tabelas clássicas de análise de variância 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

$$y_i = f_i(\beta; x) + \varepsilon_i, i = 1, \ldots, n, \tag{5.7}$$

onde os $\varepsilon_i's$ são distribuídos normalmente com média zero e variância constante σ^2, as $f_i's$ são funções diferenciáveis, $\beta = (\beta_1, \ldots, \beta_p)^T$ contém os parâmetros desconhecidos a serem estimados e $x = (x_1, \ldots, x_q)^T$ representa os valores de q variáveis explicativas.

Esses modelos são aplicáveis nas mais diversas áreas, tais como Ecologia, Agricultura, Farmacologia, Biologia, etc. A seguir, serão citados dois modelos não-lineares com suas respectivas áreas de maior aplicação:

(i) Modelo para avaliar a mistura de duas drogas

Esse modelo é geralmente aplicado na área de Farmacologia e é dado por

$$y = \alpha + \delta \log\{x_1 + \rho x_2 + k(\rho x_1 x_2)^{1/2}\} + \varepsilon,$$

onde x_1 e x_2 representam, respectivamente, as log-doses de duas drogas A e B, δ é a inclinação comum da relação log-dose-resposta, ρ é a potência da droga B em relação a 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.

(ii) Modelo de Von-Bertalanffy

Freqüentemente aplicado na área Ecológica para explicar o comprimento de um peixe pela sua idade. A forma mais conhecida desse modelo é dada por

$$y = \alpha[1 - \exp\{-\delta(x - \gamma)\}] + \varepsilon,$$

onde x representa a idade do peixe, α é o comprimento máximo esperado para a espécie, δ é a taxa média de crescimento e γ é um valor nominal em que o comprimento do peixe é zero.

5.3.1 Estimação de máxima verossimilhança

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes com a estrutura dada em (5.7). Será apresentado a seguir o algoritmo de Newton-Raphson para a obtenção da estimativa de mínimos quadrados de β, que coincide com a estimativa de máxima verossimilhança. Essa estimativa é obtida minimizando a

função quadrática

$$S(\beta) = \sum_{i=1}^{n} \{y_i - \eta_i(\beta)\}^2,$$

onde $\eta_i(\beta) = f_i(\beta; x)$. Expandindo $S(\beta)$ em série de Taylor em torno de um valor β^0 até a segunda ordem, chega-se ao seguinte processo iterativo para obter $\hat{\beta}$:

$$\beta^{(m+1)} = \beta^{(m)} + \{\tilde{X}^{(m)T}\tilde{X}^{(m)}\}^{-1}\tilde{X}^{(m)T}\{y - \eta(\beta^{(m)})\}, \tag{5.8}$$

$m = 0, 1, \ldots,$ onde $\tilde{X}$ é a matriz Jacobiana da transformação de $\eta(\beta)$ em β. Esse processo iterativo, também conhecido como algoritmo de Newton-Raphson para o modelo normal não-linear, deve continuar até que uma certa norma $\| \beta^{(m+1)} - \beta^{(m)} \| < \epsilon$, onde ϵ é um valor arbitrário suficientemente pequeno.

A convergência de (5.8) em geral depende dos valores iniciais para os parâmetros do vetor β. Isso pode evitar que problemas relacionados com a estrutura paramétrica do modelo, tais como a não-linearidade acentuada e/ou mal condicionamento da matriz $\tilde{X}$, prejudiquem a convergência do processo iterativo. Em Souza (1998) 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 β, as quais serão aplicadas a seguir para os modelos descritos na seção anterior.

(i) Modelo para avaliar a mistura de duas drogas

Como α e δ 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 α_0 e δ_0, os valores iniciais para os demais parâmetros podem ser obtidos através das estimativas de mínimos quadrados do modelo linear simples

$$z_0 = \rho x_2 + \theta t + \varepsilon,$$

onde $z_0 = \exp\{(y - \alpha_0)/\delta_0\} - x_1, \theta = k\rho^{1/2}$ e $t = (x_1 x_2)^{1/2}$.

Uma maneira alternativa, quando não for possível conhecer α_0 e δ_0 pela forma acima, é através da fixação de estimativas para ρ e k, com os demais valores iniciais sendo dados pelas estimativas de mínimos quadrados do modelo

$$y = \alpha + \delta t + \varepsilon,$$

onde $t = \log\{x_1 + \rho_0 x_2 + k_0(\rho_0 x_1 x_2)^{1/2}\}$. Se os valores obtidos não levarem (5.8) à convergência deve-se tentar novas estimativas para ρ e k e repetir o procedimento.

(ii) Modelo de Von-Bertalanffy

O primeiro passo nesse caso é obter um valor inicial para α. Como este parâmetro representa a assíntota, ou o tamanho máximo esperado para a espécie, um valor inicial razoável para α pode ser $\alpha_0 = y_{\max}$. Conhecendo α_0 e substituindo o mesmo na parte sistemática do modelo, obtém-se a seguinte relação: $z_0 = \theta - \delta x$, onde $\theta = \gamma\delta$ e $z_0 = \log\{1 - (\mu/\alpha_0)\}$. Logo, valores iniciais para γ e δ podem ser obtidos da regressão linear simples de $\log\{1 - (y/\alpha_0)\}$ sobre x. Se as estimativas de α_0, γ_0 e δ_0 não levarem (5.8) à convergência, deve-se tentar uma nova estimativa para α e repetir o procedimento.

5.3.2 Resultados assintóticos

Nesta seção serão apresentados os resultados assintóticos mais relevantes relacionados com a estimação e testes de hipóteses para o parâmetro $\beta = (\beta_1, \ldots, \beta_p)^T$ do modelo normal não-linear.

A verossimilhança do modelo (5.7), como função de β, é expressa na forma

$$L(\beta) = (2\pi\sigma^2)^{-n/2} \exp\{-S(\beta)/2\pi\sigma^2\}.$$

A EMV $\hat{\beta}$ é obtida pelo processo iterativo dado em (5.8). Esta estimativa é consistente e tem assintoticamente distribuição normal p variada de média β 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 σ^2 é dada por

$s^2 = S(\hat{\beta})/(n-p)$, onde $S(\hat{\beta})$ é a soma de quadrados dos resíduos do modelo ajustado. Logo, um intervalo de 100(1-α)% para β_j, será formado pelos limites

$$\hat{\beta}_j \pm t_{\alpha/2}(-\hat{k}^{jj})^{1/2},$$

onde $t_{\alpha/2}$ é o quantil (1-$\alpha/2$) de uma distribuição 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-α)% de confiança para β foi proposta por Beale (1960), e é formada pelos contornos de $S(\beta)$ tais que

$$S(\beta) = S(\hat{\beta})\{1 + \frac{p}{n-p}F_{p,\,n-p}(\alpha)\}.$$

Em particular, se $L(\beta)$ for aproximadamente quadrática, a região de confiança acima é bem aproximada por

$$(\hat{\beta} - \beta)^T(\tilde{X}^T\tilde{X})(\hat{\beta} - \beta) \leq s^2 p F_{p,\,n-p}(\alpha),$$

onde $F_{p,\,n-p}(\alpha)$ é o quantil (1-α) 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 $H : \beta \in B$, onde B é um subconjunto do espaço paramétrico, utiliza-se usualmente a estatística da razão de verossimilhanças, dada por

$$-2\log\lambda = n\log\{S(\tilde{\beta}) - S(\hat{\beta})\},$$

onde $S(\tilde{\beta})$ é a soma dos quadrados dos resíduos para o modelo ajustado em H. Sob essa hipótese, a estatística acima tem assintoticamente distribuição χ^2 com $(p-m)$ graus de liberdade, onde $m = \dim(B)$.

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.

5.3.3 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. No caso normal não-linear utiliza-se na detecção de pontos mais afastados dos demais, possivelmente pontos influentes, a matriz de projeção local dada por

$$\hat{H} = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T,$$

onde $\tilde{X}$ 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 $h_{ii} \geq 2p/n$ continua sendo adotado como guia para detectar pontos suspeitos de serem influentes.

Os resíduos ordinários no caso normal não-linear são definidos por $r_i = y_i - \eta_i(\hat{\beta})$, $i = 1, \ldots, n$. A distribuição desses resíduos agora é intratável, 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, nestes casos, os critérios de diagnóstico da regressão normal não-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 daquelas do resíduo ordinário da regressão normal-linear.

Ao expandir $\eta'(\hat{\beta})$ e $\eta(\hat{\beta})$ por série de Taylor em torno de β até a primeira e segunda ordem, respectivamente, Cook e Tsai (1985) encontraram a seguinte aproximação para r:

$$r \cong (I - H)r - \tilde{X}\sum_{i=1}^{n} r_i W_i \Delta - \frac{1}{2}(I - H)\Delta^T W \Delta, \tag{5.9}$$

onde H é o projetor ortogonal em $C(\tilde{X})$ (subespaço gerado pelas colunas de $\tilde{X}$), $\Delta = \hat{\beta} - \beta$ e W é uma matriz $p \times p$ com i-ésima face dada por $W_i = \left(\frac{\partial^2 \eta_i}{\partial \beta_r \partial \beta_s}\right)$, $r, s = 1, \ldots, p$.

Uma aproximação quadrática para r é obtida substituindo a primeira aproximação linear para r e Δ, respectivamente, em (5.9), mostrando que

$$E(r) \cong (I - H)f$$

e

$$Cov(r, \eta(\hat{\beta})) \cong NN^T\sigma^2 - Var(r),$$

onde f é um vetor $n \times 1$ de elementos $f_i = -\frac{1}{2}\sigma^2 tr(W_i)$ $i = 1, \ldots, n$, 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 $Var(r) = NN^T\sigma^2 +$ parte positiva. Logo, a covariância entre r e $\eta(\hat{\beta})$ tende a ser negativa, o que pode dificultar a interpretação dos gráficos padrões baseados em r.

Mostra-se que o segundo termo em (5.9) está em $C(\tilde{X})$, enquanto o terceiro termo está em $C(W^*)$, onde W^* é um "vetor" $n \times p \times p$ cuja (k,j)-ésima coluna é a projeção de $\tilde{X}_{kj} = (\partial^2\eta_1/\partial\beta_k\partial\beta_j, \ldots, \partial^2\eta_n/\partial\beta_k\partial\beta_j)^T$ 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 (5.9), Cook e Tsai (1985) definiram o resíduo projetado

$$(I - H_2)r = (I - H)\varepsilon - (I - H_1)\varepsilon. \tag{5.10}$$

O primeiro termo em (5.10) é a aproximação linear para o resíduo ordinário r, enquanto o segundo termo reflete a perda de informação necessária para se remover as componentes não-lineares de (5.7). Se $q = posto(H_1)$ for pequeno em relação a $(n - p)$, então essa perda também será pequena.

De (5.10) vem $E\{(I-H_2)r\} = 0$, $Var\{(I-H_2)r\} = \sigma^2(I-H_2)$ e $E\{r^T(I-H_2)r\} = \sigma^2 tr(I-H_2)$. Logo, uma estimativa alternativa para σ^2 é dada por

$$\tilde{\sigma}^2 = \frac{r^T(I-\hat{H}_2)r}{tr(\hat{H}_2)}.$$

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 na componente sistemática.

É importante lembrar que os operadores utilizados acima dependem de β, 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 sobre $(I-\hat{H})\tilde{X}_{kj}$, $k, j = 1, \ldots, p$.

Na regressão linear, mesmo para erros não-correlacionados e de variância constante, os resíduos são correlacionados e com variâncias 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 Studentizados, cuja i-ésima componente será dada por

$$s_i = \frac{\{(I-\hat{H}_1)r\}_i}{\tilde{\sigma}\{(I-\hat{H}_2)r\}_{ii}^{1/2}}, \quad i = 1, \ldots, n. \tag{5.11}$$

Para avaliar se os erros ε_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, onde $\Phi(\cdot)$ é a função acumulativa da normal padrão. A análise dos resíduos em (5.11) procede-se similarmente ao modelo normal linear.

5.3.4 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 $\hat{\beta}^1_{(i)}$, que é obtida inicializando o processo iterativo (5.8) 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 $\hat{\beta}^1_{(i)}$ 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

$$\hat{\beta}^1_{(i)} = \hat{\beta} - \frac{(\tilde{X}^T\tilde{X})^{-1}}{(1-\hat{h}_{ii})}\tilde{x}_i r_i, \tag{5.12}$$

onde $\tilde{X}$ e $\tilde{x}_i$ são avaliados em $\hat{\beta}$ e $\tilde{x}_i$ é a i-ésima coluna de $\tilde{X}$. Logo, $\hat{\beta}^1_{(i)}$ depende de quantidades correspondentes ao i-ésimo ponto e de quantidades conhecidas que envolvem todas as observações.

A distância de Cook é expressa por

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})^T(\tilde{X}^T\tilde{X})(\hat{\beta}_{(i)} - \hat{\beta})/ps^2,$$

onde s^2 foi definido anteriormente. Usando (5.12) na expressão acima, obtém-se a forma aproximada

$$D^1_i = \frac{\hat{t}_i^2}{p}\frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})},$$

onde $\hat{t}_i^2 = r_i/\{s(1-\hat{h}_{ii})^{1/2}\}$ é o i-ésimo resíduo ordinário Studentizado, $i = 1, \ldots, n$. Os critérios de calibração para a regressão normal linear podem ser estendidos para o caso não-linear desde que os contornos de $S(\beta) = \sum\{y_i - \eta_i(\beta)\}^2$ sejam aproximadamente elípticos. Isso porque em muitos problemas de regressão normal não-linear as regiões de confiança usuais para β 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 permite detectar àqueles pontos com os valores de D_i^1 correspondentes mais afastados dos demais. Se o interesse é detectar pontos influentes nas estimativas individuais $\hat{\beta}_j$, $j = 1, \ldots, p$, 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.

5.3.5 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 de modelos normais não-lineares, entretanto, de uma forma um pouco diferente. Num modelo normal não-linear faz sentido incluir um novo parâmetro na parte sistemática, que em muitos casos pode significar uma interação, do que uma nova variável.

Suponha então o preditor não-linear $\eta(\beta)$ para o modelo reduzido e o preditor não-linear $\eta(\beta, \gamma)$ com um parâmetro γ a ser incluído no modelo. Seja $\tilde{X}_\gamma$ um vetor $n \times 1$ com as derivadas parciais de $\eta(\beta, \gamma)$ em relação a γ. Giltinan et al. (1988) sugerem o gráfico de $r = y - \eta(\hat{\beta})$ contra $(I - \hat{H})\tilde{X}_{\hat{\gamma}}$, onde $\hat{H}$ é a matriz de projeção correspondente ao modelo reduzido e $\tilde{X}_{\hat{\gamma}}$ é o vetor $\tilde{X}_\gamma$ computado sob a hipótese nula $H : \gamma = 0$. A estimativa $\hat{\gamma}$ corresponde à estimativa do parâmetro da regressão linear simples, passando pela origem, de $y - \eta(\hat{\beta})$ sobre $(I - \hat{H})\tilde{X}_{\hat{\gamma}}$. Logo, o gráfico proposto pode revelar como as observações estão contribuindo nessa relação e como estão se afastando dela.

5.4 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 (5.1), a hetero-

cedasticidade estará presente se os elementos da diagonal de Ψ não são todos idênticos. Se, adicionalmente, ε está livre da autocorrelação, Ψ pode ser escrito como uma matriz diagonal cujo i-ésimo elemento é σ_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 como modelá-la. 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

$$A = \frac{\sum_i r_i^2(\hat{\mu}_i - \tilde{y})}{s^2 \sum_{i,j} (\delta_{ij} - h_{ij})^2 (y_i - \overline{y})(y_j - \overline{y})}, \tag{5.13}$$

onde $\delta_{ij} = 1$ se $i = j$ e $\delta_{ij} = 0$ se $i \neq j$, h_{ij} são os elementos da matriz de projeção $H = X(X^TX)^{-1}X^T$, $\hat{\mu} = Hy$, $r = (I - H)y$, $\tilde{y} = (n-p)^{-1}\sum_i(1 - h_{ii})\hat{\mu}_i$ e $s^2 = (n-p)^{-1}\sum_i r_i^2$. Se (5.13) diferir significativamente de zero, pode-se supor a heterocedasticidade dos $y_i's$.

Antes de considerar formas específicas de heterocedasticidade suponha que $\Psi = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. O estimador de mínimos quadrados generalizado (EMQG) $\hat{\beta}$ é obtido de $\hat{\beta} = (X^T\Psi^{-1}X)^{-1}X^T\Psi^{-1}y$. Quando σ^2 depende de parâmetros desconhecidos, o EMQG de β pode ser obtido da equação acima substituindo-se σ_i^2 por uma estimativa consistente $\hat{\sigma}_i^2$ produzindo $\hat{\hat{\beta}} = (X^T\hat{\Psi}^{-1}X)^{-1}X^T\hat{\Psi}^{-1}y$.

De agora em diante, denota-se por $\dot{A}$ a matriz contendo os quadrados dos elementos da matriz A. Uma forma simples de estimar o vetor $\dot{\sigma} = (\sigma_1^2, \ldots, \sigma_n^2)^T$ contendo as variâncias desconhecidas é

$$\hat{\dot{\sigma}} = \dot{M}^{-1}\dot{r}, \tag{5.14}$$

onde $\dot{r}$ é o vetor dos quadrados dos resíduos $r = (I - H)y$ e $M = I - H$ é uma matriz idempotente de posto $n - p$. Assim, (5.14) revela que $\hat{\dot{\sigma}}$ é obtido como uma transformação linear de $\dot{r}$.

É fácil verificar que o EMQ $\hat{\beta} = (X^T X)^{-1} X^T y$ satisfaz $E(\hat{\beta}) = \beta$ e $\mathrm{Cov}(\hat{\beta}) = (X^T X)^{-1} X^T \Psi X (X^T X)^{-1}$.

As principais formas de modelar a heterocedasticidade são:

(i) $\sigma_i^2 = (z_i^T \gamma)^2$, ou seja, o desvio padrão de y_i é uma função linear de variáveis exógenas;

(ii) $\sigma_i^2 = \sigma^2 (x_i^T \beta)^{2\delta}$, ou seja, a variância é proporcional a uma potência (em geral par) do valor esperado;

(iii) $\sigma_i^2 = \exp(z_i^T \gamma)$, 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) $\quad y_i = x_i^T \beta + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma_i^2 = (z_i^T \gamma)^2.$

Neste caso, o EMQG de β é

$$\hat{\beta} = \left(\sum_{i=1}^{n} (z_i^T \gamma)^{-2} x_i x_i^T \right)^{-1} \sum_{i=1}^{n} (z_i^T \gamma)^{-2} x_i y_i. \tag{5.15}$$

Existem três estimadores possíveis para γ: o estimador de MQ $\hat{\gamma}$, o EMQG $\hat{\hat{\gamma}}$ e o EMV $\tilde{\gamma}$, e, então, correspondente a cada um desses estimadores, teremos o EMQG $\hat{\hat{\beta}}$ obtido de (5.15) substituindo-se γ por $\hat{\gamma}, \hat{\hat{\gamma}}$ e $\tilde{\gamma}$. As variáveis padronizadas $\sigma_1^{-1}\varepsilon_1, \ldots, \sigma_n^{-1}\varepsilon_n$ são iid com média zero e variância um. Tem-se $E(\sigma_i^{-1}|\varepsilon_i|) = c$, onde c independe de i e depende somente da distribuição de ε_i. Assim, $E(|\varepsilon_i|) = c\sigma_i$ e, portanto,

$$|r_i| = c z_i^T \gamma + v_i,$$

onde $r_i = y_i - x_i^T(X^TX)^{-1}X^Ty$ e $v_i = |r_i| - E(|\varepsilon_i|)$ é o novo erro do modelo correspondente ao parâmetro γ. Logo,

$$c\hat{\gamma} = (Z^TZ)^{-1}Z^T|r|$$

com $Z = (z_1, \ldots, z_n)$ e $|r| = (|r_1|, \ldots, |r_n|)^T$. O inconveniente do estimador $\hat{\gamma}$ 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 $\hat{\hat{\beta}}$ independe de c. O EMQG $\hat{\hat{\gamma}}$ é obtido do EMQ $\hat{\gamma}$ a partir da equação

$$c\hat{\hat{\gamma}} = \left(\sum_{i=1}^{n}(z_i^T\hat{\gamma})^{-1}z_iz_i^T\right)^{-1}\sum_{i=1}^{n}(z_i^T\hat{\gamma})^{-2}z_i|r_i|.$$

O método de MV fornece a 3ª alternativa para estimar γ. Se os $\varepsilon_i's$ são normais, a log-verossimilhança para β e γ é

$$\ell(\beta,\gamma) = -\sum_i \log\ z_i^T\gamma - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\beta}{z_i^T\gamma}\right)^2.$$

Obtendo-se a função escore para β e γ e igualando-a a zero, tem-se um sistema não-linear para calcular $\tilde{\beta}$ e $\tilde{\gamma}$ iterativamente. Suponha agora que $\gamma = (\gamma_1, \gamma^{*T})^T$, onde $\gamma^* = (\gamma_2, \ldots, \gamma_q)^T$. Os $\varepsilon_i's$ são homocedásticos quando $\gamma^* = 0$ e um teste de homocedasticidade pode ser deduzido da razão de verossimilhanças $w = 2\{\ell(\tilde{\beta},\tilde{\gamma}) - \ell(\tilde{\tilde{\beta}},\tilde{\tilde{\gamma}}_1)\}$, onde os dois tils representam estimativas de MV restritas a $\gamma^* = 0$, ou seja, $\tilde{\tilde{\gamma}}_1 = n^{-1}(y - X\tilde{\tilde{\beta}})^T(y - X\tilde{\tilde{\beta}})$ e $\tilde{\tilde{\beta}} = (X^TX)^{-1}X^Ty$. Sob a hipótese $\gamma^* = 0$, w tem distribuição assintótica igual a χ^2_{q-1}. Testes baseados nas estatísticas de Wald e escore podem, também, ser construídos conforme apresentado na Seção 5.1.

(ii) $y_i = x_i^T\beta + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2(x_i^T\beta)^2$ (considerando o caso $\delta = 1$).

A matriz de covariância de ε é, simplesmente, $\mathrm{Cov}(\varepsilon) = \Psi = \sigma^2\mathrm{diag}\{(x_i^T\beta)^2\}$. O EMQG $\hat{\beta} = (X^T\Psi^{-1}X)^{-1}X^T\Psi^{-1}y$ é inviável, pois Ψ depende de β. Entretanto, pode-se usar o EMQ de β para obter o estimador

$\hat{\Psi}$ de Ψ e, então, definir $\hat{\hat{\beta}}$. Um estimador conveniente para a matriz de covariância assintótica de $\hat{\hat{\beta}}$ é $\hat{\Sigma}_{\hat{\hat{\beta}}} = \hat{\sigma}^2(X^T\hat{\Psi}^{-1}X)^{-1}$, onde

$$\hat{\sigma}^2 = (n-p)^{-1}(y - X\hat{\hat{\beta}})^T\hat{\Psi}^{-1}(y - X\hat{\hat{\beta}}).$$

Se y tem distribuição normal multivariada, pode-se usar o método de MV para estimar conjuntamente β e Ψ. A dependência de Ψ sobre β implica que tanto a função $(y - X\beta)^T\Psi^{-1}(y - X\beta)$ quanto a log-verossimilhança não são agora funções quadráticas de β. Métodos iterativos são necessários para obter os EMV neste caso.

(iii) $\quad y_i = x_i^T\beta + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma_i^2 = \exp(z_i^T\gamma),$
onde z_i^T é um vetor $1\times q$ contendo variáveis explicativas adicionais para estimar $\gamma \in \mathbb{R}^q$. O primeiro elemento de z_i é comumente 1. O EMQG de β é

$$\hat{\beta} = \left\{\sum_{i=1}^n \exp(-z_i^T\gamma)x_ix_i^T\right\}^{-1}\sum_{i=1}^n \exp(-z_i^T\gamma)x_iy_i. \tag{5.16}$$

A partir dos resíduos $r = (I - H)y$ de mínimos quadrados pode-se definir o modelo

$$\log r_i^2 = z_i^T\gamma + v_i,$$

onde $v_i = \log(\varepsilon_i^2/\sigma_i^2)$, e obter o EMQ de γ como

$$\hat{\gamma} = \left(\sum_{i=1}^n z_iz_i^T\right)^{-1}\sum_{i=1}^n z_i\log r_i^2. \tag{5.17}$$

O problema com o estimador (5.17) é que os v_i não têm média zero e são heterocedásticos e autocorrelacionados. Com o estimador (5.17) inserido em (5.16), obter-se-á o estimador $\hat{\hat{\beta}}$ de β.

Pode-se demonstrar que a covariância assintótica de $\hat{\gamma}$ é, simplesmente, $\Sigma_{\hat{\gamma}} = 4.9348(Z^TZ)^{-1}$. Se $\gamma^T = (\gamma_1, \gamma^{*T})$, um teste de homocedasticidade

$(H_0 : \gamma^* = 0)$ pode ser realizado através da estatística

$$g = 0.2026\hat{\gamma}^{*T}(Z^T Z)^{-1}\gamma^*$$

que tem, aproximadamente, distribuição nula igual a χ^2_{q-1}.

O método de MV pode, também, ser usado para estimar conjuntamente β e γ a partir da maximização de

$$\ell(\beta, \gamma) = -\frac{1}{2}\sum_{i=1}^{n} z_i^T\gamma - \frac{1}{2}\sum_{i=1}^{n}\exp(-z_i^T\gamma)(y_i - x_i^T\beta)^2.$$

O método escore de Fisher é baseado na informação conjunta dada por

$$K = \begin{pmatrix} X^T\Psi^{-1}X & 0 \\ 0 & \frac{1}{2}Z^T Z \end{pmatrix}.$$

A ortogonalidade entre β e γ facilita o cálculo da estrutura de covariância assintótica dos EMV de β e γ bastando inverter K.

5.5 Modelos Autocorrelacionados

Considere o modelo $y = X\beta + \varepsilon$ em que $E(\varepsilon) = 0$ e $\mathrm{Cov}(\varepsilon) = \Psi = \sigma^2\psi$ com ψ 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

$$y_i = x_i^T\beta + \varepsilon_i, \quad \varepsilon_i = \rho\varepsilon_{i-1} + v_i, \tag{5.18}$$

onde $E(v_i) = 0$, $\mathrm{Var}(v_i) = \sigma_v^2$ e $E(v_i v_j) = 0$ para $i \neq j$ e $|\rho| < 1$. A matriz de covariância de ε é $\mathrm{Cov}(\varepsilon) = \sigma_v^2\psi$ dada por

$$\Psi = \sigma_v^2\psi = \frac{\sigma_v^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ & & \vdots & & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}. \tag{5.19}$$

A inversa de Ψ é

$$\Psi^{-1} = \sigma_v^{-2}\psi^{-1} = \sigma_v^{-2}\begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}.$$

Se ρ é conhecido, o EMQG $\hat{\beta} = (X^T\psi^{-1}X)^{-1}X^T\psi^{-1}y$ é facilmente obtido usando $\hat{\beta} = (X^{*T}X^*)^{-1}X^{*T}y^*$, que é o EMQ aplicado ao modelo transformado $y^* = X^*\beta + \varepsilon^*$, onde $y^* = Py,\ X^* = PX,\ \varepsilon^* = P\varepsilon$ e

$$P = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ & & \vdots & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}$$

é definida de $P^TP = \psi^{-1}$.

Quando ρ é desconhecido, deve-se estimá-lo por $\hat{\rho}$ para obter o estimador $\hat{\hat{\beta}} = (X^T\hat{\psi}^{-1}X)^{-1}X^T\hat{\psi}^{-1}y$, onde $\hat{\psi}$ é a matriz (5.19) avaliada em $\hat{\rho}$. Algumas formas para estimar ρ estão dadas a seguir:

(a) coeficiente de correlação amostral

$$\hat{\rho}_1 = \sum_{i=2}^{n} r_i r_{i-1} \Big/ \sum_{i=1}^{n} r_i^2,$$

onde $r = (I - H)y$ são os resíduos de mínimos quadrados;

(b) estatística de Durbin-Watson

$$\hat{\rho}_2 = 1 - 0.5\sum_{i=2}^{n}(r_i - r_{i-1})^2 \Big/ \sum_{i=1}^{n} r_i^2;$$

(c) estatística de Theil-Nagar

$$\hat{\rho}_3 = \frac{n^2\hat{\rho}_2 + p^2}{n^2 - p^2}.$$

5.6 Exercícios

1. Considere o modelo heterocedástico $y_i = x_i^T\beta + \varepsilon_i,\ E(\varepsilon_i) = 0,\ \mathrm{Var}(\varepsilon_i) = \sigma_i^2 = \sigma^2(x_i^T\beta)^2$. Calcular a matriz de informação conjunta de β e σ^2 supondo que ε_i tem distribuição normal, lognormal e gama.

2. Considere o modelo heterocedástico multiplicativo $y_i = x_i^T\beta + \varepsilon_i,\ E(\varepsilon_i) = 0$ e $\mathrm{Var}(\varepsilon_i) = \exp(z_i^T\gamma)$. Deduzir a matriz de informação conjunta para β e γ supondo que ε_i tem distribuição gama. Quais as formas das estatísticas de Wald e escore para testar hipóteses relativas a: (a) um subconjunto de parâmetros em β; (b) um subconjunto de parâmetros em γ.

3. Seja o modelo de regressão (5.3) supondo $\sigma^2 = 1$. Calcular as formas das estatísticas escore, Wald e razão de verossimilhanças para testar hipóteses relativas: (a) a um subconjunto de parâmetros em β; (b) a um subconjunto de parâmetros em γ.

4. Considere o modelo de Gompertz $\mu = \exp\{\alpha - \exp(\delta - \gamma x)\}$ para explicar o comprimento médio de um certo tipo de feijoeiro em função da quantidade de água x na raiz do mesmo. A partir do conjunto de dados abaixo:

$$\begin{aligned} y_i \;=\; & 1.3,\ 1.3,\ 1.9,\ 3.4,\ 5.3,\ 7.1,\ 10.6,\ 16.0, \\ & 16.4,\ 18.3,\ 20.9,\ 20.5,\ 21.3,\ 21.2,\ 20.9 \end{aligned}$$

   e $x_i = 0.5 + i,\ i = 0, \ldots, 14$, mostre que iniciando o processo iterativo (5.8) com os valores iniciais $\alpha_0 = 3.0,\ \delta_0 = 2.1$ e $\gamma_0 = 0.4$ chega-se à convergência após 7 iterações com as estimativas $\hat{\alpha} = 3.114(0.037),\ \hat{\delta} = 2.106(0.235)$ e $\hat{\gamma} = 0.388(0.046)$, erros padrão entre parênteses, indicando que os parâmetros estão bem determinados.

5. Considere o modelo de autocorrelação com erros AR(2) especificado por

$$y_i = x_i^T \beta + \varepsilon_i, \quad \varepsilon_i = \theta_1 \theta_{i-1} + \theta_2 \varepsilon_{i-2} + v_i,$$

onde $E(v_i) = 0$, $\text{Var}(v_i) = \sigma_v^2$ e $E(v_i v_j) = 0$ para $i \neq j$. O processo é estacionário quando $\theta_1 + \theta_2 < 1$, $\theta_2 - \theta_1 < 1$ e $-1 < \theta_2 < 1$. Se $Cov(\varepsilon) = \sigma_v^2 \psi$ demonstre que

$$\psi^{-1} = \begin{pmatrix} 1 & -\theta_1 & -\theta_2 & \cdots & 0 \\ -\theta_1 & 1+\theta_1^2 & -\theta_1 + \theta_1\theta_2 - \theta_2 & \cdots & 0 \\ -\theta_2 & -\theta_1 + \theta_1\theta_2 - \theta_2 & 1+\theta_1^2+\theta_2^2 & \cdots & 0 \\ & & \vdots & & \\ 0 & 0 & 0 & \cdots & -\theta_1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Sendo $P^T P = \psi^{-1}$ mostre que

$$P = \begin{pmatrix} \sigma_v/\sigma_e & 0 & 0 & 0 & \cdots & 0 & 0 \\ -\rho_1\sqrt{1-\theta_2^2} & \sqrt{1-\theta_2^2} & 0 & 0 & \cdots & 0 & 0 \\ -\theta_2 & -\theta_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\theta_2 & -\theta_1 & 1 & \cdots & 0 & 0 \\ & & & \vdots & & & \\ 0 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & -\theta_1 & 1 \end{pmatrix},$$

onde $\dfrac{\sigma_v}{\sigma_e} = \left\{ \dfrac{(1+\theta_1)}{(1-\theta_2)} [(1-\theta_2)^2 - \theta_1^2] \right\}^{1/2}$ e $\rho_1 = \theta_1/(1-\theta_2)$.

6. Para o conjunto de dados a seguir, estime os parâmetros do modelo $y = \theta_1 x^{\theta_2} + \varepsilon$ e construa três estatísticas para testar a hipótese de linearidade

$H_0 : \theta_2 = 1$.

x	4	10	17	22	25
y	5	20	45	66	85

7. Considere o modelo parcialmente não-linear $\mu = E(y) = -\alpha + \frac{\delta}{\gamma + x}$ para explicar a resistência y de um termostato pela temperatura x. Utilize o conjunto de dados:

y_i:	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 $x_i = 50 + 5i,\ i = 0, 1, \ldots, 15$.

Mostre utilizando o algoritmo iterativo (5.8) que as estimativas dos parâmetros são $\hat{\alpha} = 5.145,\ \hat{\delta} = 6.14 \times 10^5$ e $\hat{\gamma} = 3.44 \times 10^4$.

10. Considere o modelo normal não-linear

$$y = \delta\{1 - \exp(-\gamma x)\} + \varepsilon$$

ajustado 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

(a) Obter as estimativas de MV de δ e γ;

(b) Testar a hipótese $H_0 : \gamma = 0$.

Capítulo 6

Análise de Dados Reais através dos Sistemas GLIM e S-Plus

6.1 O sistema S-Plus

O S-plus consiste em um ambiente de trabalho para realização de análises estatísticas. Dentre as diversas técnicas estatísticas disponíveis no software podemos citar: análise exploratória de dados, modelagem estatística (modelo normal linear, regressão robusta, MLGs, entre outros), análise de cluster, análise de sobrevivência, controle de qualidade, análise de séries temporais, visualização de dados, etc.

O S-Plus corresponde a uma versão ampliada e aprimorada da linguagem S, orientada para objetos e ambiente de análise de dados. A linguagem S começou como um projeto de computação estatística nos laboratórios da AT&T Bell (atualmente Lucent Technologies) no final da década de 70, com o objetivo de desenvolver um ambiente interativo para análise de dados. Na década de 80, o pesquisador R. Douglas Martin da *University of Washinghton* iniciou a Statistical Science, Inc. (StatSci) para ampliar e aprimorar a lin-

guagem S, criando assim, a primeira versão do S-Plus.

Como foi dito, o S-Plus é uma versão expandida e aprimorada da linguagem S, com as seguintes características: (i) é uma linguagem interpretativa que permite a análise interativa de dados; (ii) pode ser ampliado por funções construídas pelo usuário; (iii) é orientada para objetos e vetorizado, fazendo com que seja fácil implementar algoritmos; (iv) suporta funções escritas nas linguagens C e FORTRAN. Maiores detalhes sobre os recursos do software podem ser encontrados no manual do usuário, no help ou nos manuais on-line presentes no programa.

O ajuste de um MLG através do software S-Plus ocorre de forma rápida e simples. O primeiro passo consiste em selecionar, através do menu principal as seguintes opções: Statistics ▶ Regression ▶ Generalized Linear. Em seguida será possível definirmos: a variável dependente e variáveis independentes do modelo, a distribuição do erro, a função de ligação, tabela ANOVA, valores ajustados, devio residual e resíduo de Pearson. O usuário também poderá escolher alguns gráficos para diagnóstico, tais como: resíduos versus valores ajustados, valores observados versus valores ajustados e QQ-Plot.

Nas seções 6.4 e 6.5 apresentaremos, detalhadamente, uma análise de dados reais utlizando o software S-Plus. Posteriormente, também serão abordadas análises realizadas através de uma outra ferramenta, adequada para ajustar MLGs, conhecida como GLIM ("Generalized Linear Interactive Modelling").

6.2 Sistema de Avaliação - Uma Introdução

Um Sistema de Avaliação reúne um conjunto amplo de conhecimentos na área de engenharia e arquitetura, bem como em outras áreas de ciências sociais, exatas e da natureza, com o objetivo de determinar tecnicamente o valor de um bem, de seus direitos, frutos e custos de reprodução, etc. Os Sistemas de Avaliação são empregados para subsidiar tomadas de decisão com respeito aos valores, custos e alternativas de investimento, envolvendo bens de qualquer natureza, tais como: imóveis, máquinas e equipamentos, automóveis, móveis e

utensílios, obras de arte, empreendimentos de base imobiliária como shopping centers, hotéis, parques temáticos, cinemas, etc., além de seus frutos e direitos.

Os Sistemas de Avaliação são de grande interesse para diversos agentes do mercado imobiliário, tais como: imobiliárias, bancos de crédito imobiliário, compradores ou vendedores de imóveis. Ainda para empresas seguradoras, o poder judiciário, os fundos de pensão, os incorporadores, os construtores, prefeituras, investidores, etc.

O principal objetivo de um Sistema de Avaliação é a determinação técnica do valor de um bem, dos seus custos, frutos ou direitos sobre ele. Dessa forma, a metodologia de Modelos Lineares Generalizados será aplicada para avaliar imóveis (apartamentos e casas) situados em uma área pré-determinada da Região Metropolitana de Recife (RMR), a partir de um conjunto de variáveis explicativas. Através do modelo será estimado o valor do imóvel com o objetivo de calcular o Imposto Predial e Territorial Urbano (IPTU).

6.3 O Banco de Dados

Foram analisados dois bancos de dados que podem ser solicitados aos autores. O primeiro, chamado de ND1CA, corresponde a 376 casas de uma área pré-derterminada da Região Metropolitana do Recife (RMR). O segundo, chamado de ND1AP, corresponde a 847 apartamentos de uma área pré-derterminada da RMR. Em ambos, a variável dependende corresponde ao *Valor do Imóvel em Reais*, sendo expressa por *val.* Inicialmente, um total de 17 variáveis explicativas de natureza qualitativa - dicotômica (0: ausência; 1: presença) ou categórica - e quantitativa foram utilizadas, sendo expressas por:

Variáveis dicotômicas

- *pri* - o imóvel encontra-se situado em uma via primária de tráfego;
- *sec* - o imóvel encontra-se situado em uma via secundária de tráfego;
- *col* - o imóvel encontra-se situado em uma via coletora;
- *loc* - o imóvel encontra-se situado em uma via de tráfego local;
- *cor* - o imóvel encontra-se situado em um corredor;

- *res* - o imóvel localiza-se em uma área residencial;
- *pre* - o imóvel localiza-se em uma área de preservação;
- *z4* - presença de similaridade com um local do bairro de Boa Viagem;
- *z6* - presença de similaridade com um local do bairro de Boa Viagem;
- *z7* - presença de similaridade com um local do bairro de Boa Viagem;
- *z8* - presença de similaridade com um local do bairro de Boa Viagem;
- *ord* - o imóvel encontra-se situado em uma área de ocupação ordenada;
- *des* - o imóvel encontra-se situado em uma área de ocupação desordenada;

Variáveis quantitativas
- *are* - área construída;
- *ida* - idade do imóvel;

Variáveis categóricas
- *pad* - padrão do imóvel (E=1, D=2, C=3, B=4, A=5);
- *con* - estado de conservação do imóvel (1=péssimo, 2=ruim, 3=bom, 4=muito bom, 5=excelente);

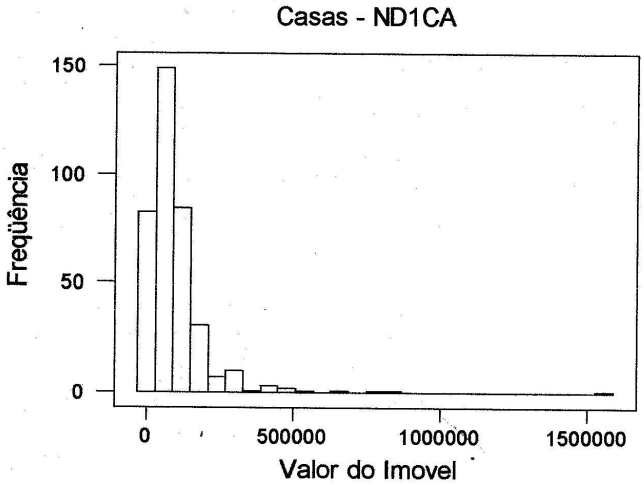
As variáveis *z4*, *z6*, *z7* e *z8* indicam setores do bairro de Boa Viagem.

A seguir, serão apresentadas todas as etapas que levaram ao ajuste final dos modelos nos bancos de dados ND1CA e ND1AP, respectivamente, incluindo a seleção de variáveis, escolha da componente aleatória, verificação da parte sistemática, análise residual, medidas de diagnóstico, etc.

6.4 Modelo para as Casas

Inicialmente, sabemos que a variável dependente é de natureza contínua. Além disso, note pela Figura 6.1 a existência de uma grande concentração de pontos à esquerda da distribuição. A partir disso, sugerimos um modelo gama para explicar o comportamento do valor do imóvel em função das variáveis explicativas.

Figura 6.1:



A respeito da função de ligação, utilizamos a ligação logarítmica devido aos problemas que podem ocorrer com a ligação canônica no modelo gama.

Da análise de uma seqüência de modelos encaixados, podemos medir a importância de cada variável no modelo.

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

Call: glm(formula = val ~ pri + sec + col + loc + cor + res + pre + z4 + z6 +
      z7 + z8 + ord + des + are + ida + pad + con, family = Gamma(
      link = log), data = ND1CA, na.action = na.exclude, control =
      list(epsilon = 0.0001, maxit = SO, trace = F))
Deviance Residuals:
      Min         1Q     Median        3Q      Max
-3.033592 -0.4221872 -0.1599562 0.2275395 2.437416

Coefficients: (3 not defined because of singularities)
                   Value   Std. Error      t value
(Intercept)  9.938317887 0.8057553370  12.33416328
        pri  1.036251839 0.4844657258   2.13895800
        sec  1.085290107 0.5088970966   2.13263175
        col  0.904922666 0.5413592708   1.67157508
        loc  0.854571040 0.5405260795   1.58099872
        cor -0.428454651 0.2045992322  -2.09411661
```

```
        res -0.356266106 0.1279168269  -2.78513871
        pre           NA           NA           NA
         z4  0.317888997 0.4644359886   0.68446246
         z6 -0.030744086 0.4707628986  -0.06530694
         z7 -0.186307728 0.4646634260  -0.40095200
         z8           NA           NA           NA
        ord -0.207511917 0.2298971411  -0.90262939
        des           NA           NA           NA
        are  0.002500768 0.0002238083  11.17370479
        ida -0.002069625 0.0019946574  -1.03758398
        pad  0.122875582 0.0394806771   3.11229673
        con 0.062027793  0.0440757947   1.40729835

Dispersion Parameter for Gamma family taken to be 0.4110574
    Null Deviance: 348.8528 on 375 degrees of freedom
Residual Deviance: 170.0874 on 361 degrees of freedom
Number of Fisher Scoring Iterations: 4

Analysis of Deviance Table
Gamma model
Response: val

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   375   348.8528
 pri  1  3.77921       374   345.0736
 sec  1  5.53890       373   339.5347
 col  1  6.21285       372   333.3218
 loc  1  0.00741       371   333.3144
 cor  1  0.87731       370   332.4371
 res  1  0.19446       369   332.2426
 pre  0  0.00000       369   332.2426
  z4  1 49.42537       368   282.8173
  z6  1 10.82964       367   271.9876
  z7  1  0.46197       366   271.5256
  Z8  0  0.00000       366   271.5256
 ord  1  0.77265       365   270.7530
 des  0  0.00000       365   270.7530
 are  1 95.15600       364   175.5970
 ida  1  0.27850       363   175.3185
 pad  1  4.44874       362   170.8698
 con  1  0.78239       361   170.0874
```

Inicialmente, devemos salientar que as variáveis *pre*, *z8* e *des* foram retiradas pois estão correlacionadas linearmente com variáveis que já estão incluídas no modelo. Além disso, note-se que as variáveis *pri*, *loc*, *cor*, *res*, *z7*,

ord, *ida* e *con* apresentam desvio residual inferior a $\chi^2_{1,0.05} = 3,841$, sendo excluídas do modelo.

Após os ajustes citados anteriormente, obtemos o seguinte modelo:

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

Call: glm(formula = val ~ sec + col + z4 + z6 + are + pad, family = Gamma(
      link = log), data = ND1CA, na.action = na.exclude, control =
      list(epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
       Min         1Q     Median        3Q      Max
 -3.032018 -0.4416364 -0.1736085 0.2251939 3.089711

Coefficients:

                   Value   Std. Error     t value
(Intercept) 10.298084396 0.0723736857 142.2904512
        sec  0.320533577 0.1474347065   2.1740714
        col  0.051003938 0.0912540983   0.5589222
         z4  0.473329928 0.0868856134   5.4477365
         z6  0.149000473 0.1064030890   1.4003397
        are  0.002530693 0.0002362635  10.7113171
        pad  0.114787876 0.0413550475   2.7756678

Dispersion Parameter for Gamma family taken to be 0.4784381
    Null Deviance: 348.8528 on 375 degrees of freedom
Residual Deviance: 178.379 on 369 degrees of freedom
Number of Fisher Scoring Iterations: 4

Analysis of Deviance Table
Gamma model
Response: val

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   375   348.8528
 sec  1  6.63122       374   342.2216
 col  1  5.28585       373   336.9357
 z4   1 46.24801       372   290.6877
 z6   1 11.07410       371   279.6136
 are  1 97.77260       370   181.8410
 pad  1  3.46200       369   178.3790
```

Porém, pelos resultados acima, a variável *pad* que antes apresentava um desvio residual satisfatório, deve ser retirada do modelo face a redução no seu

desvio residual ficando, o mesmo, inferior a $3,841$. Assim, finalmente, obtemos o seguinte modelo:

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

Call: glm(formula = val ~ sec + col + z4 + z6 + are, family = Gamma(link =
       log), data = ND1CA, na.action = na.exclude, control = list(
       epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
      Min         1Q    Median        3Q      Max
-3.005159 -0.4661969 -0.173288 0.2051069 2.900241

Coefficients:
                   Value   Std. Error     t value
(Intercept) 10.376359954 0.0616451911 168.3239159
        sec  0.345406454 0.1453474799   2.3764186
        col  0.041265966 0.0899448826   0.4587917
         z4  0.515412831 0.0840398079   6.1329606
         z6  0.181897100 0.1047599108   1.7363236
        are  0.002957677 0.0002059731  14.3595312

(Dispersion Parameter for Gamma family taken to be 0.4653562)

    Null Deviance: 348.8528 on 375 degrees of freedom

Residual Deviance: 181.841 on 370 degrees of freedom

Number of Fisher Scoring Iterations: 4

Analysis of Deviance Table

Gamma model

Response: val

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   375   348.8528
 sec 1   6.63122       374   342.2216
 col 1   5.28585       373   336.9357
  z4 1  46.24801       372   290.6877
  z6 1  11.07410       371   279.6136
 are 1  97.77260       370   181.8410
```

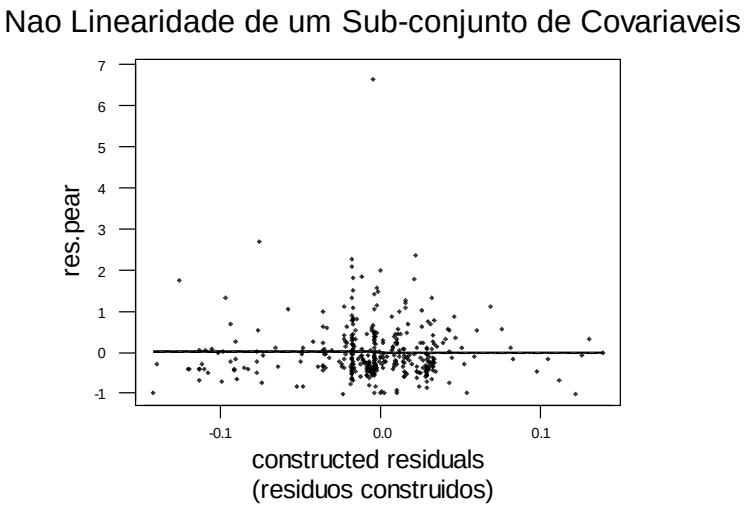
Note-se que o desvio residual do modelo (181,841) é inferior ao valor crítico $\chi^2_{370,0.05} = 415,85$, o que nos leva a aceitá-lo em princípio. Além disso, para to-

das as variáveis explicativas, seus respectivos desvios residuais apresentam-se superiores a $\chi^2_{1,0.05} = 3,841$, sinalizando que as mesmas são importantes para o modelo. O número reduzido de iterações pelo Método Escore de Fisher, necessárias para a convergência das estimativas dos parâmetros, é outro indicador positivo.

Em seguida, ilustramos o método proposto por Wang (1985) para inclusão de uma nova covariável ao modelo, apresentado na Seção 3.4. Suponha que desejamos incluir a variável idade (*ida*) ao modelo. A partir da Figura 6.2, temos que a mesma não deve ser adicionada devido a ausência de uma tendência (não necessariamente linear) nesta Figura.

Figura 6.2:

Como vimos anteriormente, as covariáveis *ida*, *pad* e *con* foram eliminadas do modelo. Um dos motivos da eliminação pode ser a presença de não-linearidade. Wang (1987) propõe um método para verificar a presença e corrigir a não-linearidade das variáveis, apresentado na Seção 3.5. Entretanto, a ausência de uma relação linear na Figura 6.3 e a análise dos resultados apresentados a seguir, indicam que a exclusão de tais covariáveis não ocorreu devido a presença de não-linearidade.



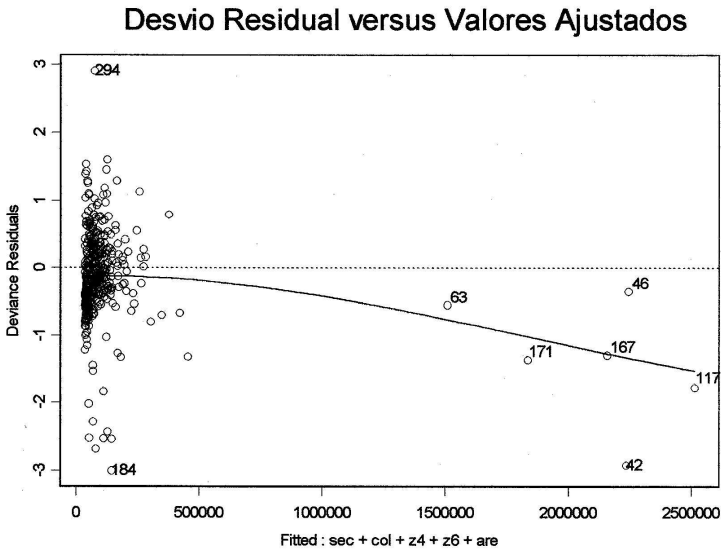
Regression Analysis

```
The regression equation is
res.pearson = -0.135 constr

Predictor       Coef       StDev         T        p
Noconstant
constr       -0.1350      0.8477     -0.16    0.874

S = 0.6822
Analysis of Variance

Source       DF           ss          ms         F        p
Regression    1       0.0118      0.0118      0.03    0.874
Error       375     174.5466      0.4655
Total       376     174.5584
```

Através da Figura 6.4, podemos observar que as observações 184 e 294 apresentam um elevado desvio residual, próximo a ± 3. Além disso, fica visível a presença de um conjunto de pontos distante da massa de dados, localizados à direita da figura. Para todas estas observações será medido o grau de influência e de alavancagem sobre o modelo proposto utilizando as medidas de Cook

modificada (T_i) e de alavanca (h_{ii}). Caso a observação não seja influente nem de alavancagem esta deverá ser retirada do modelo, configurando-se num outlier.

Figura 6.4:



Entretanto, através das Figuras 6.5 e 6.6, verifica-se que as observações 42, 63, 117, 167 e 171 configuram-se como pontos de influência e de alavancagem no modelo. A observação 46 configura-se apenas como um ponto de alavanca. Por fim, as observações 184 e 294, que apresentam um desvio residual elevado, devem ser consideradas apenas influentes. As estatísticas de corte para a verificação dos pontos de influência e de alavanca são as seguintes:

$$T = 0,2527 \quad \text{e} \quad h = 0,0319,$$

onde $p = 6$ e $n = 376$. No total foram registrados 29 pontos de alavancagem e 219 pontos de influência.

Figura 6.5:

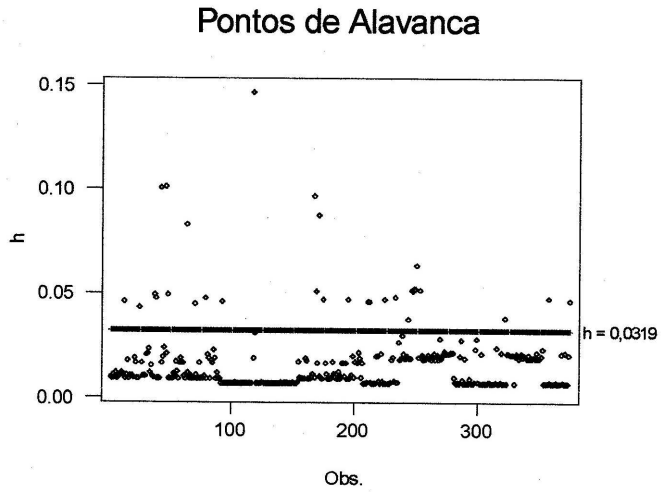
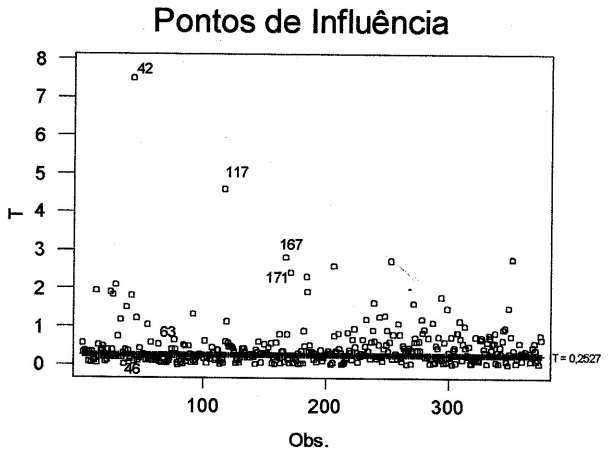


Figura 6.6:

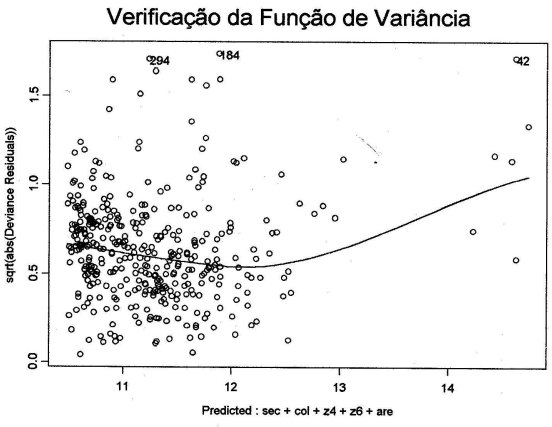


Baseando-se no método da variável adicionada proposto por Hinkley (Seção 3.6), testou-se a adequação da função de ligação logarítmica utilizada neste modelo. Fica evidente, observando os resultados a seguir, que a inclusão de $\hat{\eta}^2$ (neta.2) como uma nova covariável ao modelo proporciona uma redução

significativa no desvio. Este resultado pode implicar que algumas das variáveis explicativas apareçam sob forma não-linear. Entretanto, deve-se salientar que para as demais ligações o método iterativo de Fisher não obteve convergência.

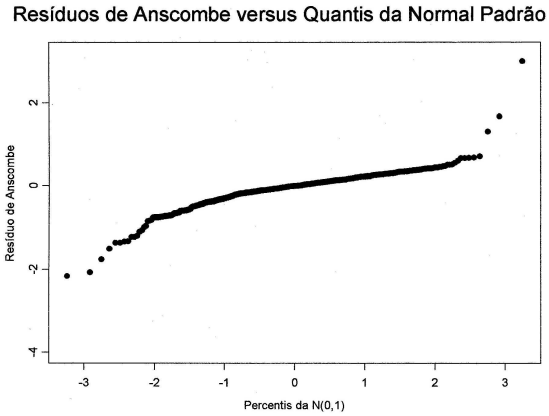
```
        *** Generalized Linear Model ***
Coefficients:
                   Value  Std. Error   t value
(Intercept)  40.29830814 4.300141234  9.371392
        sec   2.59200294 0.352392682  7.355439
        col   0.26403989 0.090640058  2.913060
         z4   3.68476056 0.456969306  8.063475
         z6   1.17327321 0.172771254  6.790905
        are   0.02318127 0.002955007  7.844742
     neta.2  -0.28126320 0.040448882 -6.953547

Residual Deviance: 165.0336 on 369 degrees of freedom
Number of Fisher Scoring Iterations: 4

Analysis of Deviance Table
Terms added sequentially (first to last)
       Df Deviance Resid.   Df Resid. Dev
  NULL                     375   348.8528
    V2  1  6.63122         374   342.2216
    V3  1  5.28585         373   336.9357
    V8  1 46.24801         372   290.6877
    V9  1 11.07410         371   279.6136
   V15  1 97.77260         370   181.8410
neta.2  1 16.80741         369   165.0336
```

Figura 6.7:



Através da Figura 6.7 conclui-se que a função de variância é adequada em virtude dos pontos estarem dispersos de forma aleatória. Deve-se ressaltar que as observações que estão à direita da massa de dados são os mesmos pontos de influência e de alavanca ao qual nos referimos anteriormente. Pela Figura 6.8, a distribuição proposta inicialmente para os dados é aceita de forma razoável. Entretanto, nota-se que os pontos situados à direita da figura ficam mais afastados da primeira bissetriz, sinalizando alguma fragilidade na função de variância que pode ser causada pelos pontos de influência e de alavanca que apresentavam desvio residual elevado.

Figura 6.8:



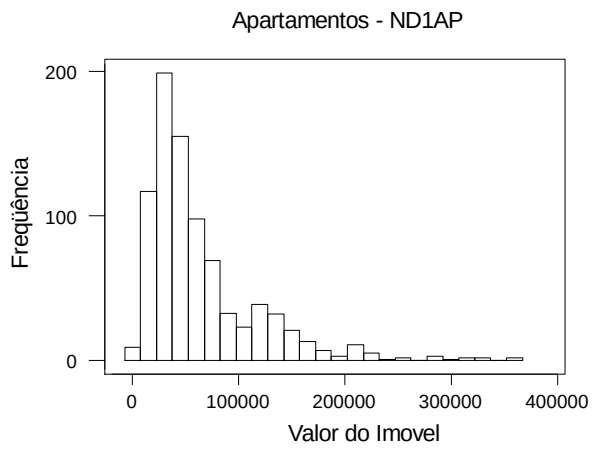
Adicionalmente, verificamos que as observações 42 e 117, que apresentam os maiores valores para a estatística T_i, realmente alteram as estimativas dos parâmetros do modelo. Ajustando o modelo final sem estas observações, verificamos uma queda de 0,46% na estimativa do *intercepto*, um aumento de 9,87% na estimativa do parâmetro da variável *sec*, reduções de 3,82%, 6,84% e 43,92% nas estimativas dos parâmetros das variáveis *z4*, *z6*, e *col*, respectivamente, e um aumento de 10,35% na estimativa do parâmetro das variável *are*. As estimativas dos parâmetros do modelo final, sem as observações 42 e 117, encontra-se a seguir:

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

Call: glm(formula = val ~ sec + col + z4 + z6 + are, family = Gamma(link =
      log), data = ND1CA, na.action = na.exclude, control = list(
        epsilon = 0.0001, maxit = 50, trace = F))
Deviance Residuals:
          Min         1Q     Median        3Q      Max
    -3.016678 -0.4536586 -0.1628268 0.2171578 2.813623

Coefficients:
                   Value   Std. Error     t value
(Intercept) 10.328597186 0.0618087635 167.1057080
        sec  0.379493473 0.1445126334   2.6260228
        col  0.023141292 0.0886555875   0.2610246
         z4  0.495733437 0.0827690689   5.9893562
         z6  0.169455490 0.1028688034   1.6472972
        are  0.003263856 0.0002241293  14.5623815
```

6.5 Modelo para os Apartamentos

Novamente, devido a natureza contínua da variável dependente e da grande concentração de pontos à esquerda da distribuição (vide Figura 6.9) foi sugerido um modelo gama aos dados.

Figura 6.9:



A ligação logarítmica foi utilizada devido aos problemas que podem ocorrer com a ligação canônica no modelo gama. Em relação à importância de cada variável, sabemos que pode ser medida através de uma análise de desvio para uma seqüência de modelos encaixados. Estes resultados são apresentados a seguir.

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

Call: glm(formula = val ~ pri + sec + col + loc + cor + res + pre + z4 + z6 +
      z7 + z8 + ord + des + are + ida + pad + con, family = Gamma(
      link = log), data = ND1AP, na.action = na.exclude, control =
      list(epsilon = 0.0001, maxit = 50, trace = F))

Deviance Residuals:

       Min         1Q       median        3Q      Max
 -2.386343 -0.1522035 -0.003152855 0.1456799 3.027486

Coefficients: (3 not defined because of singularities)

                   Value   Std. Error     t value
(Intercept)  9.782878704 0.6326473324  15.4633999
        pri -0.523239183 0.3882009237  -1.3478566
        sec -0.339563690 0.3955977960  -0.8583559
        col -0.425028464 0.3981416096  -1.0675309
        loc -0.446994646 0.3977108488  -1.1239187
        cor -0.267121977 0.0623660334  -4.2831324
        res -0.111789602 0.0543454901  -2.0570171
        pre           NA           NA          NA
         z4  0.311544824 0.1074950889   2.8982238
         z6 -0.021372095 0.1065759066  -0.2005340
         z7 -0.037993131 0.1115294913  -0.3406555
         Z8           NA           NA          NA
        ord  0.103537981 0.3783041202   0.2736898
        des           NA           NA          NA
        are  0.005259481 0.0002037686  25.8110547
        ida  0.010170396 0.0014754314   6.8931679
        pad  0.057711376 0.0201323875   2.8665938
        con -0.053680467 0.0430135225  -1.2479905

Dispersion Parameter for Gamma family taken to be 0.1415032
    Null Deviance: 504.3072 on 846 degrees of freedom
Residual Deviance: 105.1213 on 832 degrees of freedom
Number of Fisher Scoring Iterations: 5

Analysis of Deviance Table
```

```
Gamma model
Response: val

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   846 504.3072
 pri 1    8.3309       845 495.9763
 sec 1   17.0703       844 478.9060
 col 1    3.7657       843 475.1403
 loc 1    0.5409       842 474.5994
 cor 1    4.2090       841 470.3904
 res 1   19.3292       840 451.0612
 pre 0    0.0000       840 451.0612
  z4 1  175.6190       839 275.4423
  z6 1    0.1511       838 275.2912
  z7 1    0.6240       837 274.6672
  Z8 0    0.0000       837 274.6672
 ord 1    0.8410       836 273.8262
 des 0    0.0000       836 273.8262
 are 1  154.2938       835 119.5325
 ida 1   13.0076       834 106.5249
 pad 1    1.1794       833 105.3454
 con 1    0.2241       832 105.1213
```

Novamente, as variáveis *pre*, *z8* e *des* foram retiradas pois encontram-se correlacionadas linearmente com variáveis que já estão incluídas no modelo. Além disso, concluímos que as variáveis *col*, *loc*, *z6* , *z7*, *ord* e *con* devem ser excluídas do modelo pois apresentam seus respectivos desvios residuais inferiores ao valor crítico $\chi^2_{1,0.05} = 3,841$. A variável *pad* não será excluída, inicialmente, pois apresenta um valor significativo em sua estatística t.

Após as alterações sugeridas anteriormente, obtemos o modelo abaixo, onde a variável *pad* apresenta desvio residual inferior ao valor crítico $\chi^2_{1,0.05} = 3,841$, devendo ser excluída do modelo.

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

Call: glm(formula = val ~ pri + sec + cor + res + z4 + are + ida + pad,
      family = Gamma(link = log), data = ND1AP, na.action =
      na.exclude, control = list(epsilon = 0.0001, maxit = 50, trace
       = F))
```

```
Deviance Residuals:

       Min         1Q       Median        3Q      Max
 -2.379498 -0.1561799 -0.002628095 0.1450961 3.034516

Coefficients:
                   Value   Std. Error    t value
(Intercept)  9.167502083 0.1053944762  86.982757
        pri -0.099153328 0.0874509834  -1.133816
        sec  0.093994106 0.0374541477   2.509578
        cor -0.258538798 0.0613611537  -4.213395
        res -0.103498193 0.0539980078  -1.916704
         z4  0.331896269 0.0342282873   9.696549
        are  0.005265093 0.0002036334  25.855742
        ida  0.010045124 0.0014284290   7.032288
        pad  0.054710266 0.0194984129   2.805883

Dispersion Parameter for Gamma family taken to be 0.1435626
    Null Deviance: 504.3072 on 846 degrees of freedom
Residual Deviance: 105.5722 on 838 degrees of‘freedom
Number of Fisher Scoring Iterations: 4

Analysis of Deviance Table
Gamma model
Response: val

Terms added sequentially (first to last)
    Df Deviance Resid. Df Resid. Dev
NULL                  846   504.3072
 pri 1   8.3309       845   495.9763
 sec 1  17.0703       844   478.9060
 cor 1   5.2173       843   473.6887
 res 1  19.8910       842   453.7977
  z4 1 177.6851       841   276.1126
 are 1 155.9254       840   120.1872
 ida 1  13.5113       839   106.6759
 pad 1   1.1037       838   105.5722
```

Finalmente, após as últimas alterações, obtemos o modelo abaixo:

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

Call: glm(formula = val ~ pri + sec + cor + res + z4 + are + ida, family =
      Gamma(link = log), data = ND1AP, na.action = na.exclude,
      control = list(epsilon = 0.0001, maxit = 50, trace = F))
```

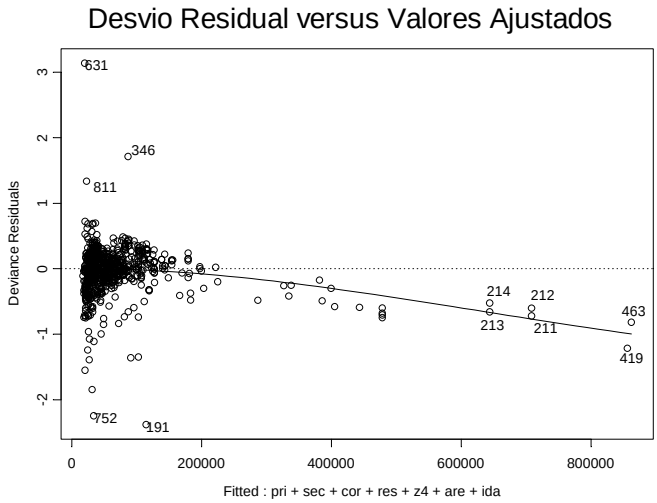
```
Deviance Residuals:
       Min         1Q       Median        3Q     Max
 -2.381841 -0.1651449 -0.002783275 0.1451711 3.13695

Coefficients:
                   Value   Std. Error    t value
(Intercept)  9.095695865 0.1029743775  88.329700
        pri -0.113232118 0.0893524666  -1.267252
        sec  0.095537348 0.0382937153   2.494857
        cor -0.264712457 0.0627466859  -4.218748
        res -0.105657393 0.0552163316  -1.913517
         z4  0.353537504 0.0341939889  10.339171
        are  0.005547092 0.0001869459  29.672174
        ida  0.011735283 0.0012856030   9.128233

Dispersion Parameter for Gamma family taken to be 0.1501208
    Null Deviance: 504.3072 on 846 degrees of freedom
Residual Deviance: 106.6759 on 839 degrees of freedom
Number of Fisher Scoring Iterations: 4

Analysis of Deviance Table
Gamma model
Response: val
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   846   504.3072
 pri  1   8.3309       845   495.9763
 sec  1  17.0703       844   478.9060
 cor  1   5.2173       843   473.6887
 res  1  19.8910       842   453.7977
  z4  1 177.6851       841   276.1126
 are  1 155.9254       840   120.1872
 ida  1  13.5113       839   106.6759
```

Da mesma forma que no caso das casas, o desvio residual do modelo para os apartamentos (106,676) está abaixo do valor crítico $\chi^2_{839,0.05} = 907,50$, levando-nos a aceitar o modelo proposto. Tem-se, ainda, que todas as variáveis explicativas incluídas são significantes devido aos seus respectivos desvios estarem acima do valor crítico $\chi^2_{1,0.05} = 3,841$. Além disso, o número reduzido de iterações até a convergência das estimativas dos parâmetros colabora com o modelo ajustado.

Pela Figura 6.10 podemos observar que as observações 191, 346, 631, 752 e 811 encontram-se afastadas da massa de dados por apresentarem desvios

residuais, em valor absoluto, elevados. Além disso, como no modelo para as casas, temos a presença de um conjunto de pontos situados à direita da massa de dados. Para todas essas observações será medido o grau de influência e de alavancagem através das medidas de Cook modificada (T_i) e de alavanca (h_{ii}). Caso a observação não seja influente nem de alavanca esta deverá ser retirada do modelo, configurando-se num outlier.

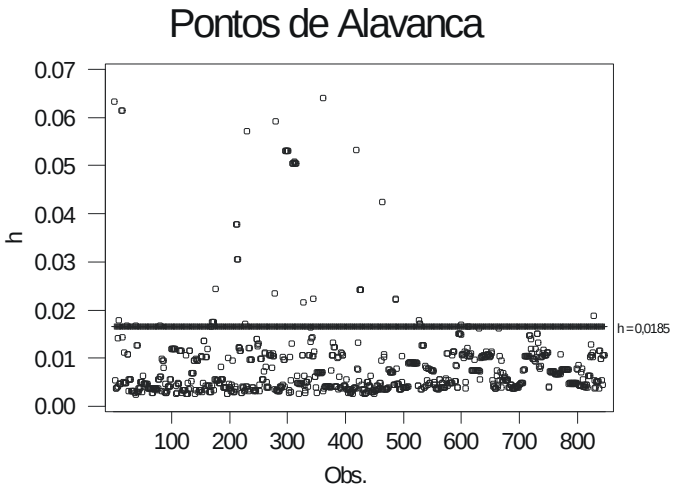
Figura 6.10:

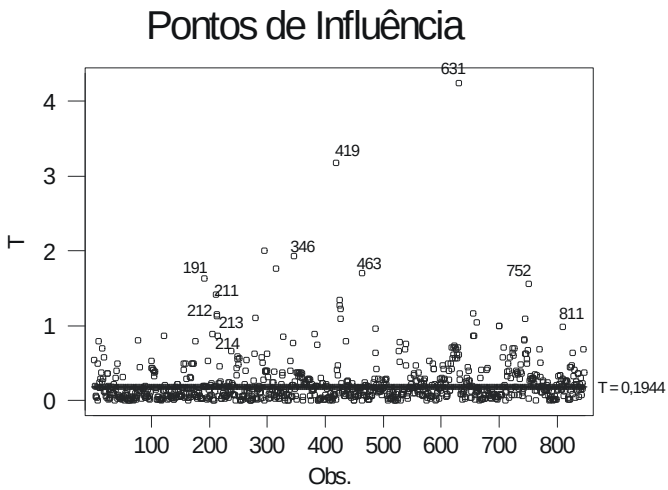


Através das Figuras 6.11 e 6.12 verificamos que as observações 191, 346, 631, 752 e 811 se caracterizam como influentes. As observações 211, 212, 213, 214, 419 e 463, além de influentes, representam pontos de alavancagem no modelo. Neste caso, as estatísticas de corte para verificar a influência e o poder de alavanca das observações são

$$T = 0,1944 \quad \text{e} \quad h = 0,0189,$$

onde $p = 8$ e $n = 847$. No total foram registrados 50 pontos de alavancagem e 333 pontos de influência.



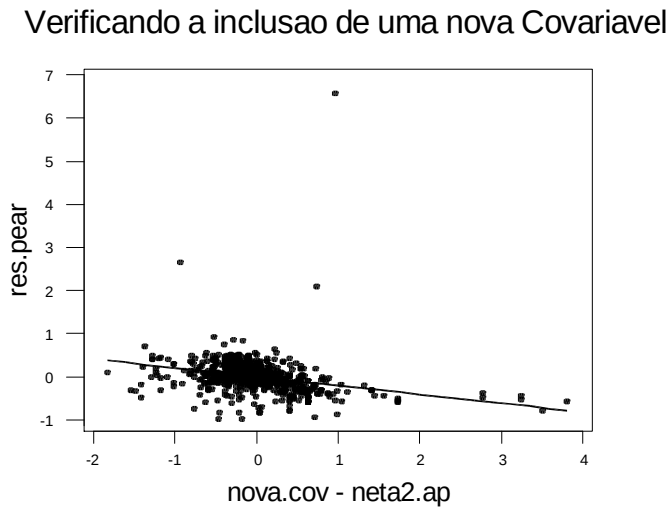


Em seguida, testamos a adequação da função de ligação através do método da variável adicionada. Fica evidente, através dos resultados a seguir, que a

inclusão de $\hat{\eta}^2$ (neta2.ap) no modelo proporciona uma redução significativa no desvio.

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

Call: glm(formula = val ~ pri + sec + cor + res + z4 + are + ida + neta2.ap,
      family = Gamma(link = log), data = ND1AP, na.action =
      na.exclude, control = list(epsilon = 0.0001, maxit = 50, trace
       = F))
Deviance Residuals:
       Min         1Q       Median        3Q      Max
 -2.392411 -0.1562116 -0.008551645 0.1220453 3.676017

Coefficients:
                   Value  Std. Error    t value
(Intercept)  27.93498860 2.068248783  13.506590
        pri  -0.79471882 0.122676308  -6.478177
        sec   0.47215794 0.057370926   8.229917
        cor  -0.26973313 0.066893504  -4.017929
        res   0.15092937 0.065877635   2.291056
         z4   1.93465476 0.175324754  11.034693
        are   0.03726713 0.003511148  10.613943
        ida   0.06997701 0.006547747  10.687188
   neta2.ap  -0.24354429 0.026740149  -9.107813

Dispersion Parameter for Gamma family taken to be 0.1702137
    Null Deviance: 504.3072 on 846 degrees of freedom
Residual Deviance: 94.03792 on. 838 degrees of freedom
Number of Fisher Scoring Iterations: 5

Analysis of Deviance Table
Gamma model
Response: val

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
    NULL                   846   504.3072
     pri  1   8.3309       845   495.9763
     sec  1  17.0703       844   478.9060
     cor  1   5.2173       843   473.6887
     res  1  19.8910       842   453.7977
      z4  1 177.6851       841   276.1126
     are  1 155.9254       640   120.1872
     ida  1  13.5113       839   106.6759
neta2.ap  1  12.6380       838    94.0379
```

Figura 6.13:



Uma outra maneira de verificar a adequação da função de ligação seria através do método proposto por Wang (1985), para a inclusão de uma nova covariável ao modelo. Considerando $\hat{\eta}^2$ (neta2.ap) como esta nova covariável, nota-se, pela Figura 6.13, a presença de uma tendência linear nos dados. Sendo assim, a nova covariável deverá ser incluída no modelo provocando, conseqüentemente, uma redução significativa no desvio. Este resultado pode implicar que algumas das variáveis explicativas apareçam sob forma não linear. Entretanto, ressalte-se que para as demais ligações o método iterativo de Fisher não obteve convergência ou o modelo apresentou desvio superior ao modelo com ligação logarítmica.

Finalmente, pela Figura 6.14, conclui-se que a função de variância é adequada devido a aleatoriedade dos pontos e a ausência de uma tendência predominante. Pela Figura 6.15, verificamos que os pontos estão bem ajustados e a distribuição proposta inicialmente aos dados é aceita de forma satisfatória. Já os pontos situados nas extremidades encontram-se mais afastados da primeira bissetriz. Entretanto, tais pontos correspondem as observações influentes e de alavanca detectados nas Figuras 6.11 e 6.12.

Figura 6.14:

Verificação da Função de Variância

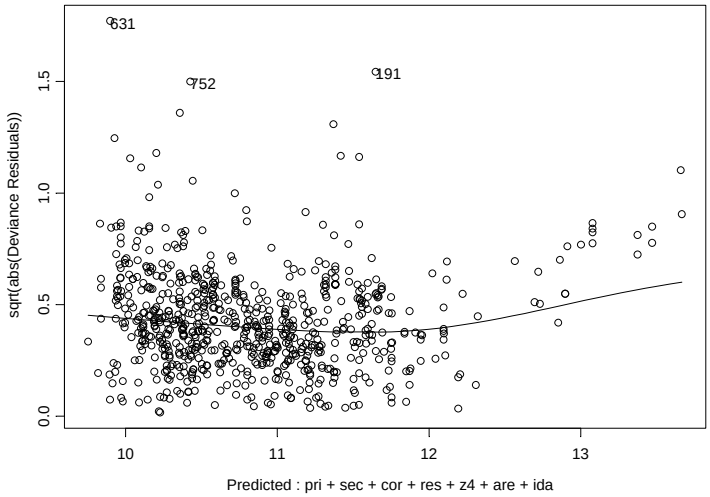
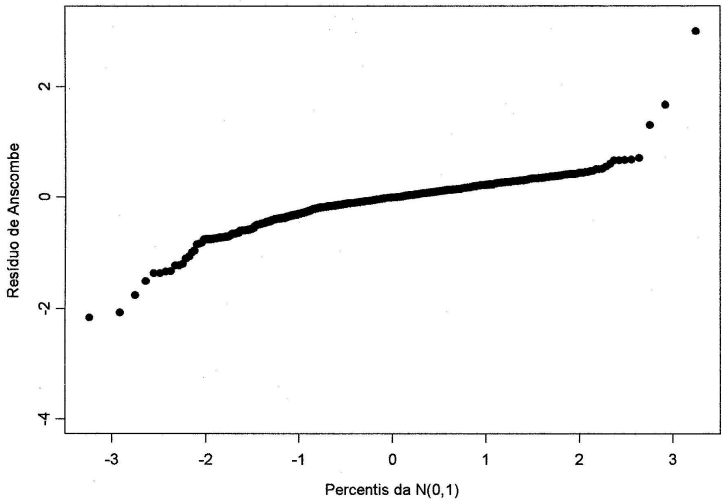


Figura 6.15:

Resíduos de Anscombe versus Quantis da Normal Padrão



Novamente, verificamos o peso que as observações influentes e/ou de alavanca exercem sobre as estimativas dos parâmetros. As observações 419 e 631, que apresentam os maiores valores para a estatística T_i, alteram de forma substancial as estimativas dos parâmetros do modelo. Ajustando o modelo final sem estas observações verificamos uma queda de 2,67% na estimativa do intercepto, um aumento de 10,88% na estimativa do parâmetro da variável *pri*, reduções de 9,06%, 31,40%, 49,45% e 95,90% nas estimativas dos parâmetros das variáveis *z4, sec, cor* e *res*, respectivamente, e um aumento de 2,39% e 15,55% nas estimativas dos parâmetros das variáveis *are* e *ida*, respectivamente. As estimativas dos parâmetros do modelo final sem as observações 419 e 631 encontram-se a seguir:

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

Call: glm(formula = val ~ pri + sec + cor + res + z4 + are + ida, family =
      Gamma(link = log), data = ND1AP, na.action = na.exclude,
      control = list(epsilon = 0.0001, maxit = 50, trace = F))

Deviance Residuals:

       Min         1Q       Median        3Q      Max
 -2.394092 -0.1526442 -0.002598195 0.1474848 1.764249

Coefficients:
                   Value   Std. Error     t value
(Intercept)  8.852951230 0.0806847864  109.722683
        pri -0.125556510 0.0693735106   -1.809862
        sec  0.065535991 0.0297427843    2.203425
        cor -0.133806162 0.0489597815   -2.732981
        res -0.004335227 0.0430210308   -0.100770
         z4  0.321495927 0.0265842665   12.093466
        are  0.005679812 0.0001488017   38.170352
        ida  0.013560026 0.0010057020   13.483144
```

6.6 O sistema GLIM

O sistema GLIM foi desenvolvido pelo grupo de computação da Royal Statistical Society. O GLIM possui um bom manual de utilização que contém um resumo da teoria dos modelos lineares generalizados, um guia completo das diretivas com exemplos de utilização, aplicações a dados reais e bibliografia.

O sistema é constituído de uma seqüência de definições, declarações e comandos, também chamados de diretivas, iniciados e terminados pelo símbolo $. Nenhum espaço deve existir entre este símbolo e a palavra que o acompanha. O símbolo $ pode indicar, simultaneamente, o fim de uma diretiva anterior e o início de uma outra. As diretivas do GLIM são formadas por letras latinas maiúsculas, dígitos de 0 a 9, espaço em branco, parênteses, operadores (· , + - * / =) e os símbolos especiais: $ (símbolo da diretiva), % (símbolo de fun ções, escalares e vetores, definidos pelo sistema), : (símbolo de repetição), # (símbolo de substituição), ! (final do registro), e outros caracteres menos importantes. Em geral, uma diretiva é predefinida pelo sistema e constituída de um nome (iniciado pelo símbolo $), com somente os três primeiros caracteres armazenados.

Um identificador pode representar um dos cinco objetos seguintes: escalar, vetor, função, macro e sub-arquivo. Os identificadores podem ser de dois tipos: definidos pelo usuário ou pelo sistema. Aqueles definidos pelo sistema consistem do símbolo de função %, seguido por uma ou duas letras, e os do usuário são formados por uma letra seguida de letras e/ou dígitos, onde somente os 4 primeiros caracteres são significantes.

Os escalares são simples números destinados a armazenar características do modelo e do ajustamento como, por exemplo, os graus de liberdade do modelo, a estatística de Pearson generalizada, o desvio após cada ajustamento, entre outras. Um vetor no GLIM pode representar uma covariável com valores arbitrários ou um fator com valores restritos aos inteiros 1, 2, ..., n, onde n é o número de níveis do fator. Alguns vetores já são predefinidos pelo sistema como os valores ajustados, as componentes do desvio, os preditores lineares estimados, entre outros. As funções são definidas pelo sistema e usadas em cálculos com vetores e escalares, enquanto que as macros constituem em subrotinas do programa, que podem conter um conjunto de instruções do GLIM ou um texto a ser impresso. Todas as macros são definidas pelo usuário. Por último, os sub-arquivos permitem ao usuário guardar conjuntos distintos de dados, conjuntos de instruções de um programa, etc., que fazem parte de um arquivo e referenciar, a qualquer tempo, somente as seções do arquivo desejadas. Para mais detalhes sobre os identificadores, vide Cordeiro (1986).

6.7 Entrada dos dados

Admitindo-se que o sistema GLIM está pronto para ser usado, o primeiro passo será a entrada dos dados. A maneira mais simples de entrada de dados no GLIM ocorre quando o número de observações é pequeno e sua entrada é realizada via teclado. Muitas variáveis nos MLGs representam vetores de um mesmo comprimento, usualmente, o número observado de casos. Para especificar que o número de dados é INT usa-se a diretiva $UNITS INT $.

Com a definição do comprimento padrão dos vetores, deve-se citar aqueles que correspondem aos dados que serão lidos e, depois, inserir esses dados. Isto é feito através das diretivas $DATA [INT] LISTA DE VETORES $ READ ... DADOS ...$. O comando READ implica numa leitura cíclica dos dados na ordem mencionada pela declaração DATA.

Entretanto, normalmente estamos interessados em analisar uma grande quantidade de dados armazenados em arquivo. Neste caso, a leitura das observações será realizada através do comando $DINPUT INT1 [INT2] $, onde INT2 é a largura declarada, opcionalmente, do arquivo INT1. Para checar os valores lidos o comando $LOOK [INT1 [INT2]] LISTA DE VETORES $ imprime, em paralelo, as componentes sucessivas, entre as posições INT1 e INT2, dos vetores lidos.

6.8 Uma seqüência típica de diretivas

Na Tabela 6.1 apresentamos uma seqüência típica de diretivas do GLIM. Os exemplos mais simples de análise de dados, via GLIM, têm uma forma similar a esta seqüência.

Tabela 6.-1: *Seqüência Típica de Diretivas do GLIM*

$UNITS	definir o número de dados
$FACTOR	identificar as variáveis independentes qualitativas e definir as suas quantidades de níveis
$DATA	rotular as variáveis cujos valores serão lidos
$READ	introduzir estes valores
$CALCULATE	calcular os níveis dos fatores
$PRINT	checar os dados de entrada ou que já foram calculados
$PLOT	observar a relação funcional entre as variáveis
$CALCULATE	transformar algumas variáveis
$PLOT	observar novamente a relação funcional entre variáveis
$YVARIATE	definir a variável dependente
$ERROR	definir a distribuição da variável resposta
$LINK	definir a ligação
$FIT	realizar um ajustamento
$FIT	introduzir mais variáveis independentes na estrutura linear e determinar seus efeitos
$DISPLAY	obter as estimativas dos parâmetros, valores ajustados, resíduos, etc
$PLOT	examinar mais cuidadosamente os resíduos
$END	terminar o programa corrente
$STOP	sair do GLIM

6.9 Definição e Ajustamento de um MLG

A definição de um MLG no GLIM requer as seguintes diretivas: YVARIATE (especifica a variável resposta), ERROR (define a distribuição do erro), LINK (define a ligação), WEIGHT (especifica pesos a priori para os dados), SCALE (especifica o parâmetro de entrada ϕ) e OFFSET (fixa valores para uma parte linear conhecida do modelo). O ajustamento de um modelo, previamente definido, é realizado pelo comando $FIT [ESTRUTURA LINEAR DO MODELO] $, onde a estrutura linear do modelo é uma fórmula que pode envolver o escalar do sistema %GM, variáveis independentes qualitativas (fatores), quantitativas (covariáveis) e mistas.

O comando FIT produz os seguintes resultados imediatos: número de iterações do algoritmo até a convergência, valor do desvio e seus graus de liberdade. Para realizar cálculos com os resultados do ajustamento pode-se usar, diretamente, os escalares do sistema: %DF (graus de liberdade do modelo), %DV (desvio após cada ajustamento), %PL (número de parâmetros linearmente independentes do modelo), %X2 (estatística de Pearson generalizada), %ML (número de elementos da matriz de covariância dos estimadores dos parâmetros linearmente independentes do modelo), %SC (parâmetro de escala dado ou estimado) e os vetores do sistema: %FV (valores ajustados), %LP (preditores lineares), %WT (pesos do processo iterativo estimados), %WV (variável dependente modificada estimada), %DR (estimativa da derivada do preditor linear em relação a média), %VA (função de variância estimada), %DI (componentes do desvio), %GM (média geral usada nos ajustamentos dos modelos) e %RE (pesos para gráficos ou para obtenção de características estimadas do modelo).

Nas próximas seções apresentaremos alguns exemplos de ajustes de MLGs a dados reais utilizando o pacote GLIM.

6.10 Assinaturas de TV a Cabo

Esta parte do livro tem como objetivo desenvolver modelos lineares generalizados para analisar dados de assinaturas de TV a cabo, demanda de energia elétrica e importação brasileira.

O primeiro modelo estima uma equação para o número de assinantes (em milhares) de TV a Cabo (ASSIN) em 40 áreas metropolitanas (Ramanathan, 1993), tendo como variáveis explicativas o número de domicílios (em milhares) na área (DOMIC), a renda per capita (em US$) por domicílio com TV a cabo (RENDA), a taxa de instalação (TAXA), o custo médio mensal de manutenção (CUSTO), o número de canais a cabo disponíveis na área (CADI) e o número de canais não pagos com sinal de boa qualidade disponíveis na área (CANAIS). Apresentam-se a seguir as observações de todas as variáveis do modelo.

```
$DATA 40 OBSER ASSIN DOMIC RENDA TAXA CUSTO CADI CANAIS $READ

  1  105.000   350.000   9839   14.95   10.00   16   13
  2   90.000   255.631  10606   15.00    7.50   15   11
  3   14.000    31.000  10455   15.00    7.00   11    9
  4   11.700    34.840   8958   10.00    7.00   22   10
  5   46.000   153.434  11741   25.00   10.00   20   12
  6   11.217    26.621   9378   15.00    7.66   18    8
  7   12.000    18.000  10433   15.00    7.50   12    8
  8    6.428     9.324  10167   15.00    7.00   17    7
  9   20.100    32.000   9218   10.00    5.60   10    8
  10   8.500    28.000  10519   15.00    6.50    6    6
  11   1.600     8.000  10025   17.50    7.50    8    6
  12   1.100     5.000   9714   15.00    8.95    9    9
  13   4.355    15.204   9294   10.00    7.00    7    7
  14  78.910    97.889   9784   24.95    9.49   12    7
  15  19.600    93.000   8173   20.00    7.50    9    7
  16   1.000     3.000   8967    9.95   10.00   13    6
  17   1.650     2.600  10133   25.00    7.55    6    5
  18  13.400    18.284   9361   15.50    6.30   11    5
  19  18.708    55.000   9085   15.00    7.00   16    6
  20   1.352     1.700  10067   20.00    5.60    6    6
  21 170.000   270.000   8908   15.00    8.75   15    5
  22  15.388    46.540   9632   15.00    8.73    9    6
  23   6.555    20.417   8995    5.95    5.95   10    6
  24  40.000   120.000   7787   25.00    6.50   10    5
  25  19.900    46.390   8890   15.00    7.50    9    7
  26   2.450    14.500   8041    9.95    6.25    6    4
  27   3.762     9.500   8605   20.00    6.50    6    5
  28  24.882    81.980   8639   18.00    7.50    8    4
  29  21.187    39.700   8781   20.00    6.00    9    4
  30   3.487     4.113   8551   10.00    6.85   11    4
  31   3.000     8.000   9306   10.00    7.95    9    6
  32  42.100    99.750   8346    9.95    5.73    8    5
  33  20.350    33.379   8803   15.00    7.50    8    4
  34  23.150    35.500   8942   17.50    6.50    8    5
```

```
$DATA 40 OBSER ASSIN DOMIC RENDA TAXA CUSTO CADI CANAIS $READ
    35   9.866    34.775   8591   15.00    8.25   11    4
    36  42.608    64.840   9163   10.00    6.00   11    6
    37  10.371    30.556   7683   20.00    7.50    8    6
    38   5.164    16.500   7924   14.95    6.95    8    5
    39  31.150    70.515   8454    9.95    7.00   10    4
    40  18.350    42.040   8429   20.00    7.00    6    4
```

Iniciaremos com o modelo supondo erro normal e as ligações identidade e logarítmica, respectivamente. O comando FIT ajusta o modelo com todas as variáveis.

```
$UNITS 40 $
$YVAR ASSIN $
$FIT DOMIC+RENDA+TAXA+CUSTO+CADI+CANAIS $
deviance = 5791.4
    d.f. =   33
$YVAR ASSIN $ERR N $LIN L $
model changed
$FIT DOMIC+RENDA+TAXA+CUSTO+CADI+CANAIS $
 deviance =  4632. at cycle  5
     d.f. =    33
```

Os modelos não são aceitos pelo valor tabelado da distribuição qui-quadrado com 33 graus de liberdade ao nível de 5%. Com isso, iremos usar um modelo com erro gama e ligação identidade para tentar obter um melhor ajuste. O comando DIS apresenta as características do modelo ajustado.

```
$YVAR ASSIN $ERR G $LIN I $
model changed
$FIT DOMIC+RENDA+TAXA+CUSTO+CADI+CANAIS $
deviance =  4.3142 at cycle  4
    d.f. = 33
$DIS MEC $
Current model:
  number of units is  40
  y-variate  ASSI
  weight     *
  offset     *
  probability distribution is GAMMA
             link function is IDENTITY
           scale parameter is to be estimated by the mean deviance
```

```
 terms =  1 + DOMI + REND + TAXA + CUST + CADI + CANA
        estimate        s.e.     parameter
   1      -5.512       5.723     1
   2      0.4092     0.03281     DOMI
   3   0.0005349   0.0007075     REND
   4      0.1165     0.09404     TAXA
   5     -0.5457      0.2513     CUST
   6      0.4692      0.1739     CADI
   7     -0.2028      0.1861     CANA
   scale parameter taken as  0.1307
Correlations of parameter estimates
 1   1.0000
 2  -0.3953   1.0000
 3  -0.9146   0.3332   1.0000
 4   0.3750  -0.1360  -0.6858   1.0000
 5  -0.3081   0.2810   0.2872  -0.3151   1.0000
 6  -0.0304  -0.1103  -0.2091   0.6441  -0.6990   1.0000
 7   0.5148  -0.2857  -0.6558   0.5410  -0.5684   0.4165   1.0000
        1        2        3        4        5        6        7
```

Com o desvio de 4.3142 o modelo gama com ligação identidade é aceito, pois esta estatística é muito inferior ao ponto crítico da distribuição qui-quadrado com 33 graus de liberdade. A Figura 6.16, mostra que os dados foram bem ajustados pelo modelo gama com ligação identidade.

Figura 6.16: *Valores ajustados versus valores observados.*

```
$PLOT FV YV '*'$
   160.00 |
   152.00 |
   144.00 |                         *
   136.00 |
   128.00 |
   120.00 |
   112.00 |                                         *
   104.00 |                     *
    96.00 |
    88.00 |
    80.00 |
    72.00 |
    64.00 |           *
    56.00 |
    48.00 |         *
    40.00 |    *     *        *
    32.00 |     * *
    24.00 |    *     *
    16.00 | *225*
     8.00 |333
     0.00 53
----------:---------:---------:---------:---------:---------:---------:
         0.0      40.0      80.0     120.0     160.0     200.0     240.0
```

Para verificar se a função de ligação é adequada, usamos uma covariável adicional Z

```
$CAL Z=LP*LP $
$YVAR ASSIN $ERR G $LIN I $
model changed
$FIT DOMIC+RENDA+TAXA+CUSTO+CADI+CANAIS+Z $
 deviance =  4.3120 at cycle  4
    d.f. = 32
```

A redução no desvio (acima), provocada pela inclusão da variável Z, não é significativa, indicando que a ligação identidade está correta, sendo isso confirmado pela Figura 6.17.

Figura 6.17: *Variável dependente modificada versus preditor linear.*

```
$PLOT WV LP '*' $
    200.0 |
    190.0 |
    180.0 |
    170.0 |                                  *
    160.0 |
    150.0 |
    140.0 |
    130.0 |
    120.0 |
    110.0 |                                              *
    100.0 |
     90.0 |                                *
     80.0 |             *
     70.0 |
     60.0 |
     50.0 |                      *
     40.0 |        *   *   *
     30.0 |         *
     20.0 |    34  * * *
     10.0 | *423*
      0.0 36*
----------:---------:---------:---------:---------:---------:---------:
         0.0      30.0      60.0      90.0     120.0     150.0     180.0
```

Na Figura 6.18 observamos um comportamento próximo à reta $Y = X$ (1ª bissetriz), mostrando que a distribuição gama para o erro está adequada.

```
$CAL NN=ND((GL(40,1)-0.5)/40) $
$CAL A=3*(YV**(1/3)-FV**(1/3))/FV**(1/3) $
$SORT ORD A $
```

Figura 6.18: *Resíduos ordenados de Anscombe versus quantis da normal* $N(0,1)$.

```
$PLOT ORD NN '*' $
  0.8000 |
  0.7200 |                                                  *
  0.6400 |
  0.5600 |
  0.4800 |                                            **  *
  0.4000 |                                        ***
  0.3200 |                                      **
  0.2400 |
  0.1600 |                                    **
  0.0800 |                                **2**
  0.0000 |                               **
 -0.0800 |                             *2
 -0.1600 |                          **2
 -0.2400 |                       *2**
 -0.3200 |                   ****
 -0.4000 |                  *
 -0.4800 |                *
 -0.5600 |            *  *
 -0.6400 |
 -0.7200 |        *
 -0.8000 |
----------:---------:---------:---------:---------:---------:---------:
       -3.00     -2.00     -1.00      0.00      1.00      2.00      3.00
```

As covariáveis RENDA, TAXA e CANAIS não são significativas, com isso iremos ajustar um novo modelo retirando as covariáveis RENDA e CANAIS, mas supondo o mesmo erro e a mesma ligação.

Considera-se agora um novo modelo, retirando as covariáveis RENDA e CANAIS, que não são significativas.

```
$YVAR ASSIN $ERR G $LIN I $
model changed
$FIT DOMIC+TAXA+CUSTO+CADI $
deviance =  4.4586 at cycle  4
     d.f. = 35


$DIS ME $
Current model:

  number of units is  40

  y-variate  ASSI
  weight     *
  offset     *
```

```
probability distribution is GAMMA
          link function is IDENTITY
         scale parameter is to be estimated by the mean deviance

terms =  1 + DOMI + TAXA + CUST + CADI


       estimate        s.e.     parameter
  1      -2.190       2.117     1
  2      0.4006     0.03043     DOMI
  3      0.1786     0.06360     TAXA
  4     -0.6937      0.2153     CUST
  5      0.5508      0.1602     CADI
  scale parameter taken as  0.1274
```

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 iremos também retirar 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.

```
$YVAR ASSIN $ERR G $LIN I $
model changed
$FIT DOMIC+CUSTO+CADI $
deviance =  5.2985 at cycle  8
    d.f. = 36


$DIS ME $
Current model:

  number of units is  40

  y-variate  ASSI
  weight     *
  offset     *

  probability distribution is GAMMA
```

```
            link function is IDENTITY
          scale parameter is to be estimated by the mean deviance
  terms =  1 + DOMI + CUST + CADI
        estimate        s.e.     parameter
    1      3.131        1.365        1
    2      0.3979       0.03300     DOMI
    3     -0.5235       0.2345      CUST
    4      0.1458       0.1085      CADI
    scale parameter taken as  0.1472
```

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, ou seja, se tivermos um aumento de 10% no número de domicílios (DOMI), o número de assinantes crescerá em cerca de 9,44%. Já um aumento de 10% no custo de manutenção (CUSTO), implica num decréscimo de 1,567% no número de assinantes de TV a cabo. Mostramos na Figura 6.19 os valores ajustados versus valores observados, revelando uma boa adequação do modelo.

Figura 6.19: *Valores ajustados versus valores observados.*

```
$PLOT FV YV '*' $
   160.00 |
   152.00 |
   144.00 |
   136.00 |                        *
   128.00 |
   120.00 |
   112.00 |                                          *
   104.00 |                     *
    96.00 |
    88.00 |
    80.00 |
    72.00 |
    64.00 |           *
    56.00 |
    48.00 |         *
    40.00 |    *     *        *
    32.00 |     * *
    24.00 |    *     *
    16.00 | *325*
     8.00 |432
     0.00 52
----------:---------:---------:---------:---------:---------:---------:
         0.0      40.0      80.0     120.0     160.0     200.0     240.0

$CAL R=(ASSIN - FV)/ FV $
```

Figura 6.20: *Resíduos de Pearson versus valores ajustados.*

```
$PLOT R FV '*' $
    1.200 |
    1.100 |
    1.000 |            *
    0.900 |
    0.800 |
    0.700 |
    0.600 |        *                        *
    0.500 *  2 2
    0.400 |    *
    0.300 |*
    0.200 |     *
    0.100 | *   *   *
    0.000 *   * *       *
   -0.100 |2  *                           *
   -0.200 |*  * * *  *    *                           *
   -0.300 | 2****             *
   -0.400 |*
   -0.500 |*           *
   -0.600 | *
   -0.700 |
   -0.800 |
----------:---------:---------:---------:---------:---------:---------:
         0.0      30.0      60.0      90.0     120.0     150.0     180.0
```

Os resíduos acima 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, iremos definir uma nova covariável, com o objetivo de obter o sinal desejado para a mesma.

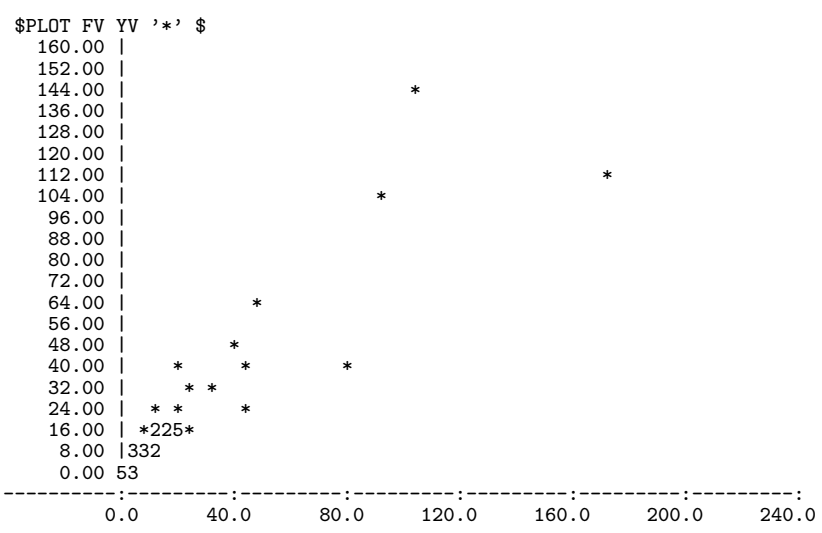
$C Definindo nova variável.

```
$CAL TX2 = TAXA**2 $

$YVAR ASSIN $ERR G $LIN I $
model changed
$FIT DOMIC+CUSTO+CADI+TAXA+TX2 $
deviance =  4.3325 at cycle  4
     d.f. = 34


 $DIS ME $
 Current model:

   number of units is  40
```

```
y-variate  ASSI
weight     *
offset     *

probability distribution is GAMMA
           link function is IDENTITY
         scale parameter is to be estimated by the mean deviance

terms =  1 + DOMI + CUST + CADI + TAXA + TX2

       estimate        s.e.     parameter
  1      0.5643        3.372        1
  2      0.4037        0.03030     DOMI
  3     -0.6899        0.2015      CUST
  4      0.5050        0.1608      CADI
  5     -0.1212        0.2954      TAXA
  6    0.008338        0.008228     TX2
  scale parameter taken as  0.1274
```

O modelo é aceito pelo teste qui-quadrado ao nível de 5%. Temos que as covariáveis TAXA e TX2 não são significativas mas o sinal da covariável TAXA agora apresenta-se correto às custas da não-linearidade do modelo.

Figura 6.21: *Valores ajustados versus valores observados.*



Na Figura 6.21 os pontos apresentam-se de forma linear, indicando que os dados foram bem ajustados.

```
$CAL R=(ASSIN - FV)/ FV $
```

Figura 6.22: *Resíduos de Pearson versus valores ajustados.*

```
$PLOT R FV '*' $
   1.0400 |
   0.9600 |
   0.8800 |             *
   0.8000 |
   0.7200 |
   0.6400 |
   0.5600 |   **   *                          *
   0.4800 |  *
   0.4000 |  * *
   0.3200 |*
   0.2400 |*
   0.1600 |     *
   0.0800 *     2   *
   0.0000 *2   *        *
  -0.0800 *
  -0.1600 | *** *                           *
  -0.2400 | ** *    *     *                            *
  -0.3200 |  * *   *            *
  -0.4000 |      *
  -0.4800 |**          *
  -0.5600 |
----------:---------:---------:---------:---------:---------:---------:
         0.0      30.0      60.0      90.0     120.0     150.0     180.0
```

Os pontos da Figura 6.22 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. Pode-se efetuar também uma análise de sensibilidade com o objetivo de medir os impactos de cada variável no número de assinaturas de TV a cabo nas 40 regiões metropolitanas. Assim, o melhor modelo para explicar os dados acima é dado por:

$$\text{ASSIN} = 3.131 + 0.3979\text{DOMIC} - 0.5235\text{CUSTO} + 0.1458\text{CADI}.$$

Com este modelo pode-se concluir que: com um aumento de 10% no número de domicílios obtém-se um aumento de 9.83% no número de assinantes. Entretanto, um aumento de 10% no custo de manutenção provoca uma redução de 1.56% no número de assinantes.

6.11 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.

```
$DATA 92 ANO T ELAR PER  PGR  RECA D1 D2 D3 D4 $READ
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
```

```
$DATA 92 ANO T ELAR PER  PGR  RECA D1 D2 D3 D4 $READ
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
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
```

```
$DATA 92 ANO T ELAR PER  PGR  RECA D1 D2 D3 D4 $READ
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
```

$C Definição dos fatores

```
$UNITS 92 $FACT 92 D 4 $
$CAL D = GL(4,1) $
```

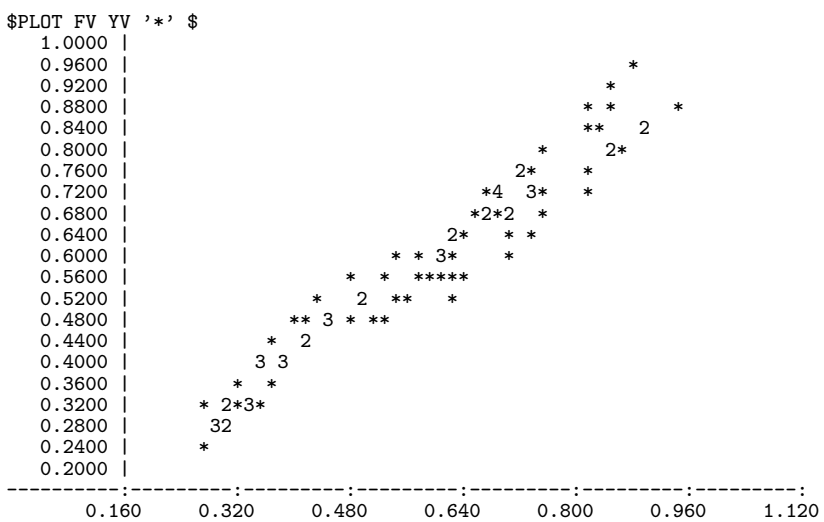
O ajuste do modelo será iniciado usando erro normal e as ligações identidade e logarítmica, respectivamente.

```
$YVAR ELAR $
$FIT PER+PGR+RECA+D1+D2+D3 $

deviance =  0.21417
    d.f. = 85

$YVAR ELAR $ERR N $LIN L $
model changed
$FIT PER+PGR+RECA+D1+D2+D3 $

deviance =  0.17169 at cycle  3
    d.f. = 85

$DIS ME $

Current model:
 number of units is  92

  y-variate  ELAR
  weight     *
  offset     *
```

```
probability distribution is NORMAL
           link function is LOGARITHM
           scale parameter is to be estimated by the mean deviance

terms =  1 + PER + PGR + RECA + D1 + D2 + D3

       estimate        s.e.     parameter
  1      -2.228      0.2395     1
  2     -0.1125     0.02396     PER
  3     0.07300     0.02012     PGR
  4       163.0       14.16     RECA
  5      0.1262     0.02217     D1
  6    -0.04949     0.02409     D2
  7      0.1102     0.02369     D3
  scale parameter taken as  0.002020
```

Os dois modelos são aceitos pelo valor tabelado da distribuição qui-quadrado com 85 graus de liberdade ao nível de 5%, sendo o melhor ajuste aquele de menor desvio. Todas as covariáveis são significativas. Observa-se que a diferença entre os valores observados e os valores ajustados é muito pequena, indicando que os dados estão bem ajustados, conforme melhor observado na Figura 6.23.

Figura 6.23: *Valores ajustados versus valores observados.*



```
$CAL Z=LP*LP $
$Yvar ELAR $ERR N $LIN L $
model changed
$FIT PER+PGR+RECA+D1+D2+D3+Z $
deviance =  0.16957 at cycle  3
    d.f. = 84
```

A redução no desvio (acima), provocada pela inclusão da variável Z, não é significativa, indicando que a ligação identidade está correta, sendo confirmada pela Figura 6.24, pois esta se apresenta de forma linear.

Figura 6.24: *Variável dependente modificada versus preditor linear.*

```
$PLOT WV LP '*' $
   0.0000 |
  -0.0800 |                                          2 *
  -0.1600 |                                       * 3 ** 2 *
  -0.2400 |                                     *  *    *
  -0.3200 |                                 * *2*2**2**
  -0.4000 |                                    **24*
  -0.4800 |                            * **2322
  -0.5600 |                            **  *2
  -0.6400 |                          2   *
  -0.7200 |                        * **  *
  -0.8000 |                        *2*
  -0.8800 |                      2* *
  -0.9600 |                *  ****
  -1.0400 |             *    3
  -1.1200 |           2 ***
  -1.2000 |       *2 * *
  -1.2800 |  *  2   *
  -1.3600 |
  -1.4400 |
  -1.5200 |
  -1.6000 |
----------:---------:---------:---------:---------:---------:---------:
       -1.500    -1.200    -0.900    -0.600    -0.300    0.000     0.300
```

```
$CAL NN= ND((GL(92,1)-0.5)/92) $
$CAL A=YV - FV $
$SORT ORD A $
```

Figura 6.25: *Resíduos ordenados de Anscombe versus quantis da normal* $N(0,1)$.

```
$PLOT ORD NN '*' $
   0.1000 |                                                  *
   0.0900 |                                              * *
   0.0800 |                                             *
   0.0700 |                                            **
   0.0600 |                                         22
   0.0500 |                                       22
   0.0400 |                                      3
   0.0300 |                                  4433
   0.0200 |                               *4
   0.0100 |                              53
   0.0000 |                            *4
  -0.0100 |                            *4
  -0.0200 |                          243
  -0.0300 |                        *42
  -0.0400 |                      332
  -0.0500 |                   *22
  -0.0600 |                2
  -0.0700 |              **
  -0.0800 |
  -0.0900 |        *  * *
  -0.1000 |
----------:---------:---------:---------:---------:---------:---------:
       -3.60     -2.40     -1.20      0.00      1.20      2.40      3.60
```

Os pontos na Figura acima apresentam o comportamento de uma reta, indicando que a distribuição normal para o erro é adequada para representar os dados.

```
$CAL R=(ELAR - FV) $
```

A Figura 6.26 apresenta pontos dispersos de forma aleatória indicando que pode ser aceita a hipótese de independência e variância constante para os resíduos.

Com base nos dados e resultados acima pode-se concluir que uma equação para explicar a demanda de energia elétrica é dada por:

$$\log(\text{ELAR}) = -\ 2.228 - 0.1125\text{PER} + 0.073\text{PGR} + 163\text{REC} + 0.1262\text{D1} + 0.04949\text{D2} + 0.1102\text{D3},$$

o que é razoável, pois espera-se um aumento na demanda de eletricidade (ELAR) quando seu preço (PER) diminuir, quando o preço do gás natural

(PGR) aumentar e quando a renda per capita (REC) aumentar. Isto pode ser analisado pela sensibilidade marginal, isto é, para cada 1% de aumento do preço da tarifa implicará uma redução de cerca de 10% da demanda de eletricidade; entretanto, um aumento de 1% no preço do gás natural acarretaria um aumento de 7,57% na demanda de eletricidade.

Figura 6.26: *Resíduos de Pearson versus valores ajustados.*

```
$PLOT R FV '*' $
   0.1600 |
   0.1440 |                    *
   0.1280 |
   0.1120 |                         *  *
   0.0960 |   *                    *         *
   0.0800 |                        *           *
   0.0640 |     22              *       * **    **  * *
   0.0480 |      *  *          *    *                *
   0.0320 |        2           * * *     3  *   *
   0.0160 |        **         *      * 2     *
   0.0000 |          * *            * 22      *     *
  -0.0160 |                    *     *    *    *
  -0.0320 |        **    2                22*       *  *
  -0.0480 |             2 * 2**             2
  -0.0640 |           *        * *  * *         2 *    ** *
  -0.0800 |              *   *
  -0.0960 |
  -0.1120 |                *       *
  -0.1280 |                   * *
  -0.1440 |
  -0.1600 |
----------:---------:---------:---------:---------:---------:---------:
       0.160     0.320     0.480     0.640     0.800     0.960     1.120
```

6.12 Importação Brasileira

O impacto das variáveis que influenciam a balança comercial tem sido amplamente discutido após a abertura econômica diante do processo de inserção da economia brasileira na globalização dos anos 90. Do ponto de vista da política econômica é importante identificar estes impactos, bem como, o efeito dinâmico de políticas monetárias e cambiais frente aos setores que se relacionam com o comércio internacional.

Dentro deste contexto, há um particular interesse em examinar detalhadamente a dinâmica da desvalorização e/ou valorização cambial sobre as importações, dado a evidência empírica no sentido de que esse efeito possa

ser negativo (Braga e Rossi, 1987). Para isso, utiliza-se os instrumentais estatísticos tradicionais de regressão comparativamente ao método que trata os erros de estimação de forma aleatória.

A violação de pressupostos sobre o erro, muitas vezes é inevitável pelo critério tradicional e, por isso, utiliza-se neste trabalho a metodologia dos modelos lineares generalizados com a expectativa de melhorar as estimativas das relações de importações no Brasil. O objetivo é encontrar uma equação para a importação brasileira (IM), tendo como variáveis explicativas a taxa de câmbio (TCI) e o Produto Interno Bruto representando a renda nacional (RN). O modelo é calculado com dados trimestrais das contas externas do Brasil no período de 1980 à 1998 (Banco Central). As importações estão especificadas em milhões de dólares, 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). Segue-se todas as observações das variáveis do modelo.

```
$DATA 74 IM TCI RN $READ
     5482  1.629    82.17
     5749  1.517    88.80
     6043  1.331    87.94
     5679  1.181    85.28
     5605  1.315    82.06
     5565  1.217    86.49
     5610  1.177    82.62
     5309  1.135    78.30
     4804  1.434    78.34
     4872  1.306    87.11
     5071  1.209    85.77
     4646  1.156    80.91
     3824  1.740    75.88
     3651  2.004    83.65
     3907  1.957    82.80
     4044  1.959    80.10
     3155  1.971    79.10
     3406  2.015    87.59
     3730  2.024    87.19
     3623  2.027    85.94
     3094  2.036    84.55
     3016  2.219    92.47
     3132  2.201    95.23
```

```
     3925  2.131    94.44
     3352  2.013    90.69
$DATA 74 IM TCI RN $READ
     2760  2.023    99.48
     3661  1.991   102.87
     4270  1.924   101.15
     3565  1.832    97.65
     3610  1.792   106.21
     3987  1.914   103.45
     3888  1.789   101.10
     3516  1.692    97.72
     3349  1.657   105.78
     3776  1.643   105.84
     3963  1.607    98.87
     3548  1.557    95.01
     4046  1.423   109.40
     5495  1.356   111.36
     5173  1.244   105.50
     4576  1.046    97.60
     4265  1.091    96.39
     5474  1.091   106.01
     6345  1.300   100.01
     4330  1.380    91.70
     5034  1.354   104.02
     5614  1.314   108.26
     6015  1.452   101.05
     4630  1.499    97.02
     4725  1.626   101.71
     5221  1.467   103.80
     5976  1.441   101.30
     5230  1.421    99.90
     6007  1.388   106.90
     7328  1.340   108.92
     6914  1.305   106.01
     6049  1.283   104.01
     7087  1.279   109.66
     8023  1.075   115.30
    11814  0.957   116.45
    12065  0.942   113.92
    13651  0.955   116.09
    11917  0.951   115.67
    12030  0.970   114.93
    10738  0.980   111.63
    12478  0.995   118.06
    14235  1.012   122.90
    15837  1.030   120.69
```

```
    13150  1.049   116.90
    15405  1.067   123.85
$DATA 74 IM TCI RN $READ
    16930  1.086   126.37
    15873  1.106   122.55
    13415  1.126   118.11
    14591  1.147   125.74
```

Primeiramente, a análise do modelo será feita nos moldes tradicionais que especifica o modelo levando em consideração os erros distribuídos normalmente. A função, em termos da notação original, é a seguinte:

$$\hat{E}(\mathrm{IM}) = -3203.3 - 4210.7\mathrm{TCI} + 158.92\mathrm{RN},$$

tendo como desvio $D = 0.31177E + 09$, (no caso soma dos quadrados dos resíduos), indicando que a variância dos dados é muito grande. O coeficiente de determinação $R^2 = 0.7106$ indica que as duas variáveis explicativas (TCI e RN) são responsáveis por 71.06% da variação total da importação (IM). A estatística de Durbin-Watson $d = 0.2715$ detectou a presença de autocorrelação positiva.

Numa análise gráfica verifica-se que a variância não é constante ao longo do tempo, indicando a presença de heterocedasticidade. E foi feita uma transformação logarítmica nos dados com o objetivo de corrigir a heterocedasticidade, mas não corrigiu a autocorrelação. Para eliminar os efeitos da autocorrelação foi feito uma transformação nas variáveis, com isso obtemos uma estimativa corrigida da equação original, implicando na seguinte equação corrigida:

$$\hat{E}(\mathrm{LIM}) = 0.044203 - 0.26109\mathrm{LTCI} + 1.9123\mathrm{LRN},$$

com desvio $D = 1.2203$. O coeficiente de determinação $R^2 = 0.9321$ indica que 93.21% da variação total da importação é explicada pelas covariáveis LTCI e LRN. A estatística de Durbin-Watson $d = 2.2317$ indica que não há autocorrelação dos erros.

Usando o GLIM também fizemos a análise do modelo com erro normal e ligações identidade e logarítmica, respectivamente. O comando FIT ajusta o modelo com todas as variáveis explicativas.

```
$units 74 $

$YVAR IM
$FIT TCI+RN $
deviance = 315765888.
    d.f. =        71

$DIS MEC $
Current model:

  number of units is  74

  y-variate  IM
  weight     *
  offset     *

  probability distribution is NORMAL
             link function is IDENTITY
           scale parameter is to be estimated by the mean deviance

  terms =  1 + TCI + RN
         estimate        s.e.     parameter
    1      -2284.       2941.     1
    2      -4441.       777.1     TCI
    3       152.5       21.70     RN
    scale parameter taken as  4447407.

Correlations of parameter estimates
  1   1.0000
  2  -0.7728   1.0000
  3  -0.9410   0.5245   1.0000
         1        2        3

$YVAR IM $ERR N $LIN L $
$FIT TCI+RN $
deviance = 146543440. at cycle  4
     d.f. =        71

 $DIS MERC $
 Current model:
   number of units is  74

   y-variate  IM
   weight     *
   offset     *

   probability distribution is NORMAL
              link function is LOGARITHM
```

```
            scale parameter is to be estimated by the mean deviance

   terms =  1 + TCI + RN
         estimate        s.e.     parameter
     1       7.037      0.3855     1
     2     -0.8180      0.1161     TCI
     3     0.02744    0.002559     RN
     scale parameter taken as  2063992.
```

Os dois modelos usando erro normal não são aceitos pelo valor tabelado da qui-quadrado com 71 graus de liberdade ao nível de 5%. Iremos ajustar um novo modelo usando erro gama, ligações identidade e logarítmica.

```
$YVAR IM $ERR G $LIN I $
$FIT TCI+RN $
 deviance =  6.1914 at cycle  7
     d.f. = 71

 $DIS MEC $
 Current model:

   number of units is  74
  y-variate  IM
  weight     *
  offset     *

   probability distribution is GAMMA
              link function is IDENTITY
            scale parameter is to be estimated by the mean deviance

   terms =  1 + TCI + RN
         estimate        s.e.     parameter
    1       3424.       2143.     1
    2      -3706.       527.6     TCI
    3       83.00       17.09     RN
    scale parameter taken as  0.08720

 Correlations of parameter estimates
   1   1.0000
   2  -0.7411   1.0000
   3  -0.9192   0.4272   1.0000
          1        2        3
```

O modelo com desvio de 6.1914 é aceito pelo teste qui-quadrado ao nível de 5%. As estimativas dos parâmetros são significativas, o que pode ser obser-

vado pela estatística T, calculada pela macro TVAL, disponível na biblioteca do GLIM. Os sinais dos parâmetros estão corretos, isto é, a medida que aumentamos a renda nacional (RN), a importação brasileira aumentará, diferentemente da taxa de câmbio (TCI), pois a importação sofrerá uma diminuição com o aumento da TCI.

Pela Figura 6.27 observamos que os dados não foram bem ajustados.

```
$USE TVAL $

T values
     +----------+
     |    TV_   |
 +---+----------+
 | 1 |   1.598  |
 | 2 |  -7.023  |
 | 3 |   4.857  |
 +---+----------+
```

Figura 6.27: *Valores observados versus valores ajustados.*

```
$PLOT YV FV $
   17600. |
   16800. |                                                  I
   16000. |                                                II
   15200. |                                                  I
   14400. |                                                 I I
   13600. |                                              I  I
   12800. |                                               I I
   12000. |                                                3I
   11200. |
   10400. |                                              I
    9600. |
    8800. |
    8000. |                                             I
    7200. |                                   II I
    6400. |                          I   I  I II
    5600. |                I     I I II2 I2     II  I
    4800. |                   I      I3I     I  II
    4000. |       222I       2 II  2 II      II
    3200. |      I2 2 I       I  I I  I
    2400. |                I
    1600. |
----------:---------:---------:---------:---------:---------:---------:
       1600.     3200.     4800.     6400.     8000.     9600.     11200.
```

O diagrama de dispersão dos valores ajustados e o preditor linear (Figura 6.28) indicam que a função de ligação não está correta, o que é comprovado,

formalmente, pela redução significativa no desvio que ocorre com a inclusão da variável explicativa Z no modelo.

Figura 6.28: *Valores ajustados versus preditor linear.*

```
$PLOT WV LP '*' $
    9.800 |
    9.700 |                                          *  *  *
    9.600 |                                           *  * *
    9.500 |                                     *  *    *
    9.400 |                                          2*2
    9.300 |                                       *
    9.200 |
    9.100 |
    9.000 |                                     *
    8.900 |                     *  *
    8.800 |                  *  *
    8.700 |     * *    2*   *  *        *
    8.600 |           * 23*     2*
    8.500 |        2   * *  *     *  *
    8.400 | *        **  *  *
    8.300 |* **2 **
    8.200 |* 22**  2
    8.100 |**** *  *
    8.000 |*
    7.900 |*
    7.800 |
----------:---------:---------:---------:---------:---------:---------:
       8.100     8.400     8.700     9.000     9.300     9.600     9.900

$CAL Z=LP*LP $
$YVAR IM $ERR G $LIN I $
$FIT TCI+RN+Z $
 deviance =  2.5015 at cycle  4
     d.f. = 70
```

Como a função de ligação não é adequada para o modelo iremos trabalhar com a ligação logarítmica mantendo o mesmo erro, pois pela Figura 6.29, os pontos apresentam um comportamento próximo à reta $y = x$, mostrando que a distribuição gama para o erro está adequada.

```
$CAL NN=ND((GL(74,1)-0.5)/74) $
$CAL A=3*(YV**(1/3)-FV**(1/3))/FV**(1/3) $
$SORT ORD A $
```

Figura 6.29: *Resíduos ordenados de Anscombe versus quantis da* $N(0,1)$.

```
$PLOT ORD NN '*' $
   0.4500 |
   0.4000 |
   0.3500 |                                                  *    *
   0.3000 |                                                *
   0.2500 |                                            *2* *
   0.2000 |                                        22*
   0.1500 |                                      22
   0.1000 |                                    22
   0.0500 |                                 *32*
   0.0000 |                            *32332
  -0.0500 |                        22332
  -0.1000 |                      22*
  -0.1500 |                    *
  -0.2000 |                  22*
  -0.2500 |               2**
  -0.3000 |              *
  -0.3500 |
  -0.4000 |           **
  -0.4500 |
  -0.5000 |    *    *
  -0.5500 |
----------:---------:---------:---------:---------:---------:---------:
       -3.00     -2.00     -1.00      0.00      1.00      2.00      3.00
```

Modelo com erro gama e ligação logarítmica.

```
$YVAR IM $ERR G $LIN L $
$FIT TCI+RN $
deviance =  3.9075 at cycle  3
    d.f. = 71
$DIS MEC $
Current model:
  number of units is  74
  y-variate  IM
  weight     *
  offset     *

  probability distribution is GAMMA
            link function is LOGARITHM
          scale parameter is to be estimated by the mean deviance

  terms =  1 + TCI + RN
        estimate        s.e.     parameter
    1       8.132      0.3272     1
    2     -0.7633     0.08645     TCI
    3     0.01650    0.002414     RN
    scale parameter taken as  0.05504
```

```
Correlations of parameter estimates
  1   1.0000
  2  -0.7728   1.0000
  3  -0.9410   0.5245   1.0000
         1        2        3

$USE TVAL $
T values
     +-----------+
     |    TV_    |
 +---+-----------+
 | 1 |   24.857  |
 | 2 |   -8.829  |
 | 3 |    6.835  |
 +---+-----------+
```

Com um desvio de 3.9075 o modelo tem um bom ajuste pois esta estatística é muito inferior ao ponto crítico da qui-quadrado com 71 graus de liberdade. Pela estatística T observa-se que todas as estimativas dos parâmetros são significativas. A Figura abaixo indica que não houve um bom ajuste dos dados, sendo necessário ajustar um novo modelo.

Figura 6.30: *Valores observados versus valores ajustados.*

```
$PLOT YV FV '*' $
   17600. |
   16800. |                                                  *
   16000. |                                               **
   15200. |                                                 *
   14400. |                                                * *
   13600. |                                            *   *
   12800. |                                              * *
   12000. |                                              22
   11200. |
   10400. |                                             *
    9600. |
    8800. |
    8000. |                                            *
    7200. |                                   ** *
    6400. |                          *   *  * **
    5600. |                *     * * * *2 2*     **   *
    4800. |                   *      4*      *  **
    4000. |    **2*2       *** *  ****       **
    3200. |   ***** *       *   * *  *
    2400. |             *
    1600. |
----------:---------:---------:---------:---------:---------:---------:
       7.800     8.100     8.400     8.700     9.000     9.300     9.600
```

Faz-se um novo ajuste com erro gama, ligações identidade e logarítmica, usando transformação logarítmica nos dados.

```
$CAL LIM=LOG(IM) $
$CAL LTCI=LOG(TCI) $
$CAL LRN=LOG(RN) $

$YVAR LIM $ERR G $LIN I $
$FIT LTCI+LRN $
deviance =  0.051764 at cycle  3
    d.f. = 71

$DIS MEC $
Current model:
  number of units is  74

  y-variate  LIM
  weight     *
  offset     *

  probability distribution is GAMMA
             link function is IDENTITY
           scale parameter is to be estimated by the mean deviance

  terms =  1 + LTCI + LRN
         estimate        s.e.     parameter
    1       3.348       1.112     1
    2      -1.236      0.1249     LTCI
    3       1.245      0.2371     LRN
    scale parameter taken as  0.0007291

Correlations of parameter estimates
  1   1.0000
  2  -0.5411   1.0000
  3  -0.9991   0.5110   1.0000
         1        2        3

$USE TVAL $

T values
     +----------+
     |    TV_   |
 +---+----------+
 | 1 |   3.011  |
 | 2 |  -9.894  |
 | 3 |   5.251  |
 +---+----------+
```

Figura 6.31: *Valores observados versus valores ajustados.*

```
$PLOT YV FV '*' $
    9.800 |
    9.700 |                                          * 2
    9.600 |                                          * * *
    9.500 |                                        * *   *
    9.400 |                                            23
    9.300 |                                          *
    9.200 |
    9.100 |
    9.000 |                                        *
    8.900 |                                  * *
    8.800 |                                * *
    8.700 |                    *    * 2   * *
    8.600 |             *        *   *32     2*     *
    8.500 |                *    *  *  *  *
    8.400 |               *        ** *         * *
    8.300 |     ** *       * *   *        *
    8.200 |      22       **  ***  *
    8.100 |    *  * *             *
    8.000 |     ***
    7.900 |            *
    7.800 |
----------:---------:---------:---------:---------:---------:---------:
       7.800     8.100     8.400     8.700     9.000     9.300     9.600

$YVAR LIM $ERR G $LIN L $
$FIT LTCI+LRN $
deviance =  0.049192 at cycle  3
    d.f. = 71

$DIS MERC $
Current model:

   number of units is  74

   y-variate  LIM
   weight     *
   offset     *

   probability distribution is GAMMA
              link function is LOGARITHM
            scale parameter is to be estimated by the mean deviance

   terms =  1 + LTCI + LRN

          estimate        s.e.     parameter
     1       1.525      0.1262     1
     2     -0.1441     0.01430     LTCI
     3      0.1479     0.02687     LRN
     scale parameter taken as  0.0006928
```

Os dois modelos são aceitos pelo valor tabelado da qui-quadrado com 71 graus de liberdade ao nível de 5%. O segundo modelo, com ligação logarítmica, apresenta-se melhor ajustado, o que pode ser observado pela pequena diferença entre os valores observados e os valores ajustados ao comparar a Figura 6.32 em relação a Figura 6.31.

Figura 6.32: *Valores observados versus valores ajustados.*

```
$PLOT YV FV '*' $
    9.800 |
    9.700 |                                          ***
    9.600 |                                         * * *
    9.500 |                                        * *  *
    9.400 |                                            23
    9.300 |                                          *
    9.200 |
    9.100 |
    9.000 |                                        *
    8.900 |                                 * *
    8.800 |                                * *
    8.700 |                    *    * 2  * *
    8.600 |                *      *   *4*    2*    *
    8.500 |                   *    * * *  *
    8.400 |                  *      ** *       * *
    8.300 |        ** *       **  *       *
    8.200 |         *3      **  *** *
    8.100 |        *  **            *
    8.000 |         2*
    7.900 |               *
    7.800 |
 ----------:---------:---------:---------:---------:---------:---------:
        2.0400    2.0800    2.1200    2.1600    2.2000    2.2400    2.2800
```

A redução no desvio resultante da inclusão da variável explicativa Z não é significativa, comprovando formalmente a adequação da função de ligação que também pode ser verificado pela Figura 6.33, que se apresenta de forma linear.

```
$CAL Z=LP*LP $
$YVAR LIM $ERR G $LIN L $
$FIT LTCI+LRN+Z $

deviance =  0.033916 at cycle  3
    d.f. = 70
```

Figura 6.33: *Variável dependente modificada versus preditor linear.*

```
$PLOT WV LP '*' $
   2.2920 |
   2.2800 |                                        *
   2.2680 |                                    *   2
   2.2560 |                                *  *      2
   2.2440 |                                    *   **3
   2.2320 |                                      *
   2.2200 |
   2.2080 |
   2.1960 |                                  *
   2.1840 |                    2 *
   2.1720 |                  *
   2.1600 |          *  **2 3*  2
   2.1480 |        *      2***  * *      *
   2.1360 |          *** *  *          *
   2.1240 |       *      *           *
   2.1120 |    222*  **     *
   2.1000 |    32* 2 2
   2.0880 |    2*
   2.0760 |    **
   2.0640 |
   2.0520 |
----------:---------:---------:---------:---------:---------:---------:
       2.0800    2.1200    2.1600    2.2000    2.2400    2.2800    2.3200
```

Figura 6.34: *Resíduos de Pearson versus valores ajustados.*

```
$CAL R=(LIM - FV)/FV $
$PLOT R FV '*'$
  0.05400 |
  0.04800 |
  0.04200 |    *                                  *
  0.03600 |       *                            *  *    *
  0.03000 |           *
  0.02400 |          * 2      *                      **
  0.01800 |   *  *            *                   *
  0.01200 | *2     *     * **    *
  0.00600 |  ***    *  2 2*
  0.00000 |  2* *      * *    *                          *
 -0.00600 |  2 *   *                                   *  *
 -0.01200 | *    *    *  * *  **                     *  *
 -0.01800 | *2   *  *                                     *2
 -0.02400 |        *            *              *
 -0.03000 |  *     *
 -0.03600 |               *
 -0.04200 |
 -0.04800 |                              *
 -0.05400 |
 -0.06000 |                         *  *
 -0.06600 |
----------:---------:---------:---------:---------:---------:---------:
       8.100     8.400     8.700     9.000     9.300     9.600     9.900
```

A Figura 6.34 dos resíduos versus os valores ajustados, apresenta pontos de forma aleatória em torno da reta horizontal que passa pela origem, indicando que pode ser aceita a hipótese de variância constante para os resíduos.

Figura 6.35: *Resíduos de Anscombe versus ordem das observações.*

```
$CAL I=GL(74,1) $
$PLOT A I '&'$
   0.4500 |
   0.4000 |
   0.3500 |2
   0.3000 |                                                &
   0.2500 | &                &                           && &
   0.2000 |  &            &   &                           & &
   0.1500 |     &           &             &               &
   0.1000 |                               & &     &  &
   0.0500 |  &2               & &        &     &
   0.0000 |    &&  && &   &               & & &  &&&2
  -0.0500 |      &&& & &&    &   &         & & & &
  -0.1000 |        &        &    &       &        &
  -0.1500 |              &
  -0.2000 |             &         &&     2
  -0.2500 |           &          &  &             &
  -0.3000 |                         &
  -0.3500 |
  -0.4000 |                        &  &
  -0.4500 |
  -0.5000 |                          2
  -0.5500 |
----------:---------:---------:---------:---------:---------:---------:
         0.0      16.0      32.0      48.0      64.0      80.0      96.0
```

Os pontos da Figura 6.35 apresentam-se de forma aleatória indicando que os resíduos são independentes.

Através das análises feitas anteriormente a estimação da equação da importação brasileira é mostrada a seguir: $\hat{E}(\text{LIM}) = 1.525 - 0.1441\text{LTCI} + 0.1479\text{LRN}$, sendo os resultados obtidos satisfatórios. A variável explicativa taxa de câmbio (TCI) apresenta coeficiente estimado com o sinal teoricamente correto e estatisticamente significativo ao nível de 5% de significância.

Com isso, temos que, para cada aumento (ou redução) de uma unidade no logaritmo da taxa de câmbio, corresponderá um decréscimo (ou elevação) de 0.1441 unidades no logaritmo das importações brasileiras, mantidos constantes os demais fatores. Para cada aumento (ou redução) de uma unidade no logaritmo da renda nacional, corresponderá um aumento (ou decréscimo) de 0.1479 unidades no logaritmo das importações brasileiras.

Em termos de sensibilidade percentual, temos que 1% de aumento na taxa de câmbio implicará, praticamente, em 1% (0.998%) de aumento nas importações brasileiras. O mesmo ocorre com a renda nacional, um aumento de 1% na renda nacional, corresponderá um aumento de 1% nas importações brasileiras.

Os modelos finais mais adequados são:

Modelo 1: $\hat{E}(\text{LIM}) = 0.044203 - 0.26109\text{LTCI} + 1.9123\text{LRN}$,
com erro normal;

Modelo 2: $\hat{E}(\text{IM}) = -2284 - 4441\text{TCI} + 152.5\text{RN}$, via GLIM,
com erro normal;

Modelo 3: $\hat{E}(\text{LIM}) = 1.525 - 0.1441\text{LTCI} + 0.1479\text{LRN}$, via GLIM,
com erro gama.

A literatura econômica sugere modelos com erros com distribuição normal. Considerando a estimação no GLIM para testar os erros, observou-se que os erros não têm distribuição normal. Assim, testou-se vários procedimentos obtendo-se como melhor especificação aquela com distribuição gama.

Observando os parâmetros estimados verifica-se diferenças significativas entre os modelos, isto é, com um aumento de uma unidade no logaritmo da taxa de câmbio do modelo 1, temos um decréscimo de 0.2610 unidades no logaritmo das importações, enquanto um mesmo aumento na taxa de câmbio do modelo 3, teremos uma redução menor de 0.1441 unidades no logaritmo das importações brasileiras. Como o modelo 3 apresenta uma menor redução nas importações, podemos considerá-lo o melhor modelo dentre os três modelos apresentados.

Bibliografia

[1] Aitkin, M., Anderson, D., Francis, B. e Hinde, J. (1989). *Statistical modelling in GLIM.* Clarendon Press, Oxford, UK.

[2] Aitkin, M. e Clayton, D. (1980). The fitting of exponencial, Weibull and extreme value distributions to complex censored survival data using GLIM. *Appl. Statist.*, **29**, 156-163.

[3] Anscombe, F.J. (1948). The transformation of Poisson, binomial and negative binomial data. *Biometrika*, **37**, 358-383.

[4] Anscombe, F.J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics*, **15**, 229-230.

[5] Anscombe, F.J. (1953). Contribution to the discussion of H. Hotelling's paper. *J. R. Statist. Soc. B*, **15**, 229-230.

[6] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. e Tukey, J.W. (1972). *Robust estimates of location.* Princeton University Press, Princeton, N.J.

[7] Andrews, D.F. e Pregibon, D. (1978). Finding the outliers that matter. *J. R. Statist. Soc. B*, **40**, 87-93.

[8] Aranda-Ordaz, F. (1981). On the families of transformations to additivity for binary response data. *Biometrika*, **68**, 357-363.

[9] Arnold, S.F. (1981). T*he theory of linear models and multivariate analysis.* John Wiley, New York.

[10] Atkinson, A.C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

[11] Barndorff-Nielsen, O.E. (1978). *Information and exponencial families in statistical theory.* Wiley, Chichester.

[12] Barndorff-Nielsen, O.E. e Jϕrgensen, B. (1991). Proper dispersion models. *Aarhus, Department of Statistics - Aarhus University.* (Research Report, 200).

[13] Bates, D.M. e Watts, D.G. (1980). Relative curvature measures of nonlinearity. *J. R. Statist. Soc. B*, **42**, 1-25.

[14] Beale, E.M.L. (1960). Confidence region in nonlinear estimation. *J. R. Statist. Soc. B*, **22**, 41-76.

[15] Bernoulli, J. (1713). *Ars conjectandi.* Thurnisius, Basilea.

[16] Belsley, D.A. , Kuh, E. e Welsch, R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity.* John Wiley, New York.

[17] Bishop, Y.M.M., Fienberg, S.E. e Holland, P.W. (1975). *Discrete multivariate analysis: theory and pratice.* MIT Press, Cambridge, MA.

[18] Bliss, C.I. (1935). The calculator of the dosage-mortality curve. *Ann. Appl. Biol.*, **22**, 134-167.

[19] Box, G.E.P. e Cox, D.R. (1964). An analysis of transformation. *J. R. Statist. Soc. B*, **26**, 211-252.

[20] Box, G.E.P. e Tidwell, P.W. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531-550.

[21] Braga, N.C. e Rossi, J.W. (1987). A dinâmica da balança comercial do Brasil, 1970-84. *Revista Brasileira de Economia*, **41**, 237-248.

[22] Collet, D. (1994). *Modelling binary data.* Chapman and Hall, London.

[23] Cook, R.D. (1977). *Detection of influential observations in linear regression.* Technometrics, **19**, 15-18.

[24] Cook, R.D. e Tsai, C.L. (1985). Residual in nonlinear regression. *Biometrika*, **72**, 23-29.

[25] Cook, R.D. e Weisberg, S. (1982). *Residuals and influence in regression.* Chapman and Hall, London.

[26] Copas, J.B. (1988). Binary regression models for contaminated data (with discussion). *J. R. Statist. Soc. B*, **50**, 225-265.

[27] Cordeiro, G.M. (1986). *Modelos lineares generalizados.* VII SINAPE, UNICAMP.

[28] Cordeiro, G.M. e Demétrio, C.G.B. (1989). An algorithm for fitting a quasi-likelihood model with a non-constant dispersion parameter. *Lecture Notes in Statistics*, Proceedings of the GLIM'89 International Conference. Springer-Verlag, Berlin.

[29] Cordeiro, G.M e Paula, G.A. (1989). Fitting non-exponencial family nonlinear models in GLIM by using the offset facilities. *Lecture Notes in Statistics*, **57**, 105-144.

[30] Cordeiro, G.M e Botter, D. (1998). Improved Estimators for Generalized Linear Models with Dispersion Covariates. *Journal Statistical Computation and Simulation*, **62**, 91-104.

[31] Cordeiro, G.M e Paula, G.A. (1992). Estimation, large-samples parametric tests and diagnostics for non-exponencial family nonlinear models. *Communications in Statistics, Simulation and Computation*, **21**, 149-172.

[32] Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B*, **74**, 187-220.

[33] Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

[34] Cox, D.R. e Hinkley, D.V. (1974). *Theoretical Statistics.* Chapmand and Hall, London.

[35] Cox, D.R. e Oakes, D. (1984). *Analysis of survival data.* Chapman and Hall, London.

[36] Cox, D.R. e Snell, E.J. (1968). A general definition of residual (with discussion). *J. R. Statist. Soc. B*, **30**, 248-275.

[37] Dey, D.K., Gelfand, A.E. e Peng, F. (1997). Overdispersion generalized linear models. *Journal of Statistical Planning and Inference*, **68**, 93-107.

[38] Draper, N.R. e Smith, H. (1981). *Applied regression analysis.* John Wiley, New York.

[39] Duffy, D.E. (1990). On continuity-corrected residuals in logistic regression. *Biometrika*, **77**, 2, 287-293.

[40] Fisher, R.A. (1925). *Statistical methods for research workres.* Oliver and Boyd, Edinburgh.

[41] Folks, J.L. e Chhikara, R.S. (1978). The inverse Gaussian distribution and its statistical application, a review. *J. R. Statist. Soc. B*, **40**, 263-289.

[42] Francis, B., Green, M. e Payne, C. (1993). *The GLIM system generalized linear iteractive modelling.* New York.

[43] Gart, J.J. e Zweifel, J.R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassy. *Biometrika*, **54**, 181-187.

[44] Gelfand, A.E. e Dalal, S.R. (1990). A note on overdispersed exponencial families. *Biometrika*, **77**, 55-64.

[45] Gigli, A. (1987). *A comparasion between Cox & Snell residuals and deviance residuals.* MSc thesis, Imperial College, London.

[46] Giltnan, D.M., Capizzi, T.P. e Malani, H. (1988). Diagnostic tests for similar action of two compunds. *Appl. Statist.*, **37**, 39-50.

[47] Goodman, L.A. (1969). On partitioning χ^2 and detecting partial association in three-way contingency tables. *J. R. Statist. Soc. B*, **31**, 486-498.

[48] Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classification. *Journal of American Statistical Association*, **65**, 226-256.

[49] Goodman, L.A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classification. *Technometrics*, **13**, 33-61.

[50] Goodman, L.A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, **60**, 179-192.

[51] Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives (with discussion). *J. R. Statist. Soc. B*, **46**, 149-192.

[52] Green, P.J. e Yandell, B.S. (1985). Semi-parametric generalized linear models. *Lecture Notes in Statistics*, **32**, 44-55, Springer-Verlag, Berlin.

[53] Haberman, S.J. (1974). *The analysis of frequence data*. Univ. of Chicago Press, Chicago, Illinois.

[54] Hastie, T. e Tibshirani, R. (1986). Generalized aditive models. *Statistical Science*, **1**, 297-318.

[55] Hastie, T. e Tibshirani, R. (1987). Generalized aditive models. Some applications. *Journal of the American Statistical Association*, **82**, 371-386.

[56] Hinkley, D.V. (1985). Transformation diagnostic for linear models. *Biometrika*, **72**, 487-496.

[57] Hoaglin, D.C. e Welsch, R. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, **32**, 17-22.

[58] Huber, P. (1973). Robust regression: asymptotics, conjectures and monte carlo. *Ann. Statist.,* **1**, 799-821.

[59] Jennrich, R.I. (1969). Asymptotic properties of nonlinear least-squares estimation. *Annals Math. Statist.*, **20**, 633-643.

[60] Jϕrgensen, B. (1983). Maximum likelihood estimates and large samples inference for generalized linear and nonlinear regression models. *Biometrika*, **70**, 19-28.

[61] Jϕrgensen, B. (1987). Exponencial disperson models (with discussion). *J. R. Statist. Soc. B*, **49**, 127-162.

[62] Ku, H.H. e Kulback, S. (1968). Interaction in multidimensional contingency tables: an information theoretic approach. *J. Res. Nat. Bur. Standards*, **78B**, 159-199.

[63] Landwehr, J.M., Pregibon, D. e Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *Journal of American Statistical Association*, **79**, 61-83.

[64] Lane, P.W. e Nelder, J.A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, **73**, 13-22.

[65] Laplace, P.S. (1836). *Théore analytique des probabilités*. Supplement to Third Edition, Couvier, Paris.

[66] Larntz, K. (1978). Small samples comparisons of exact levels for chi-square goodness of fit statistics. *Journal of the American Statistical Association*, **73**, 362, 253-263.

[67] Lee, A.H. (1987). Diagnostic displays for assessing leverage and influence in generalized linear models. *Austral. J. Statist.*, **29**, 233-243.

[68] Lee, K. (1977). On the asymptotic variances of $\hat{\mu}$ terms in log-linear models of multidimensional contingency tables. *Journal of the American Statistical Association*, **72**, 358, 412-419.

[69] McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, **11**, 59-67.

[70] McCullagh, P. (1984). *On the conditional distribution of goodness-of-fit statistics for discrete data*. Unpublished Manuscript.

[71] McCullagh, P. e Nelder, J.A. (1983, 1989). *Generalized linear models.* Chapman and Hall, London.

[72] Montgomery, D.C. e Peck, E. A. (1982). *Introduction to linear regression analysis.* John Wiley, New York.

[73] Nelder, J.A. e Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221-232.

[74] Nelder, J.A. e Wedderburn, R.W.M (1972). Generalized linear models. *J. R. Statist. Soc. A*, **135**, 370-384.

[75] Pierce, D.A. e Schafer, D.W. (1986). Residual in generalized linear models. *Journal of the American Statistical Association*, **81**, 977-986.

[76] Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Appl. Statist.*, **29**, 15-24.

[77] Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, **9**, 705-724.

[78] Ramanthan, R. (1993). *Statistical methods in econometrics.* Academic Press, New York.

[79] Rao, C.R. (1973). *Linear statistical inference and its applications.* John Wiley, New York.

[80] Ratkowsky, D.A. (1983). *Nonlinear regression modelling.* Marcel Dekker, New York.

[81] Scheffé, H. (1959). *The analysis of variance.* John Wiley, New York.

[82] Searle, S.R. (1971). *Linear models.* John Wiley, New York.

[83] Seber, G.A.F. (1977). *Linear regression analysis.* John Wiley, New York.

[84] Sousa, D.G. (1986). *Algumas considerações sobre regressão não-linear.* Dissertação de Mestrado, IME-USP, São Paulo.

[85] Wang, P.C. (1985). Adding a variable in generalized linear models. *Technometrics*, **27**, 273-276.

[86] Wang, P.C. (1987). Residual plots for detecting nonlinearity in generalized linear models. *Technometrics*, **29**, 435-438.

[87] Wedderburn, R.W.M. (1974). Quasi-likelihood function, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-477.

[88] Weisberg, S. (1985). *Applied linear regression.* John Wiley, New York.

[89] Wetherill, G.B. , Duncombe, P., Kenward, M., Kollerstrom, J., Paul, S. R.e Vowden, B.J. (1986). *Regression analysis with applications.* Chapman and Hall.

[90] Wilkinson, G.N. e Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Appl. Statist.*, **22**, 392-399.

Índice

adequação do modelo, 47, 77, 104, 210
análise de variância, 11, 13–15, 48, 50, 68, 72, 105, 107, 112, 156
análise do desvio, 50, 51

componente aleatória, 37, 49, 66, 69, 83, 126, 131, 134, 178
componente sistemática, 37, 40, 44, 49, 66, 101, 125, 126, 128, 129, 141, 163

desvio residual, 79, 81–89, 91, 95, 97, 115, 181, 182, 184, 185, 188, 191, 193
distribuição de Poisson, 62
distribuição binomial, 40, 52, 53, 60, 62, 76, 101, 116
distribuição de Poisson, 40, 52, 54, 62–66, 90, 104, 116
distribuição gama, 68, 69, 71, 74, 87, 100, 116, 117, 172, 207, 227, 235

equações normais, 3, 4, 128
estatística de Cook, 23–25
estatística modificada de Cook, 97
estatísticas suficientes, 31, 32, 41, 70, 100, 104–108, 110
estimação de máxima verossimilhança, 29, 152

função de ligação, 35–37, 40, 41, 51, 55, 58, 66, 70, 77, 91, 94, 100, 115, 117–119, 124, 141, 144, 176, 179, 186, 195, 197, 207, 226, 227, 232
função de variância, 39, 50, 53, 65, 69, 77, 95, 98, 115–117, 124, 129, 131, 132, 138, 188, 197, 203
função desvio, 48, 50, 51, 61, 62, 64, 69, 71, 114

ligações canônicas, 41, 98

método de mínimos quadrados, 2, 130
método escore de Fisher, 43, 44, 46, 61, 102, 146, 170, 183
medida de alavancagem, 20, 23–25, 96
medidas de influência, 82, 96, 164
modelo de Box e Cox, 100, 120, 121, 123, 147
modelo de regressão rígida, 154
modelo gama, 46, 67–71, 76, 78, 178, 179, 189, 190, 206
modelo log-linear, 63, 75, 104, 105, 110, 136, 137

modelo logístico linear, 100, 101, 104, 148
modelo normal, 66, 79, 99, 100
modelo normal inverso, 72
modelo normal não-linear, 156, 159, 164, 165
modelo normal-linear, 10, 98, 127
modelos aditivos generalizados, 126
modelos autocorrelacionados, 151, 170
modelos de quase-verossimilhança, 128–130, 132
modelos de riscos proporcionais, 135, 137, 139
modelos heterocedásticos, 151, 152, 165
modelos hierárquicos, 105–107, 111, 112, 148
modelos lineares generalizados, 35, 77, 99, 125, 139, 143, 156, 177
modelos semi-paramétricos, 126

quase-verossimilhança estendida, 131–133

regressão linear, 1, 94, 133, 146, 156, 159, 161, 163, 164
regressão linear múltipla, 5, 13, 26
regressão linear simples, 4, 18, 127, 128, 159, 165
resíduo de Anscombe, 78, 79
resíduo de Cox-Snell, 83
resíduo de Pearson, 77, 78, 80, 81, 97, 176
resíduo Studentizado, 22, 25
resíduos padronizados, 21, 23, 25, 26

soma de quadrados dos resíduos, 6, 7, 10, 160

técnicas de diagnóstico, 19, 20, 156, 161, 163
teste de normalidade, 87