MODELOS DE REGRESSÃO PARA ANÁLISE DE DADOS UNIVARIADOS

Gauss M. Cordeiro
Gilberto A. Paula

ISBN

85-244-0046-3

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO
INSTITUTO DE MATEMÁTICA PURA E APLICADA
Estrada Dona Castorina, 110
22.460 – Rio de Janeiro – RJ

PREFÁCIO

Este trabalho objetiva apresentar os modelos de regressão para análise de dados univariados. Não se pretende cobrir 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 e não-linear, os modelos para análise de dados categorizados e mesmo, mais recentemente, os modelos lineares generalizados. A idéia deste texto surgiu da inexistência, mesmo em lingua inglesa, de um único livro que abordasse todos esses modelos, mesmo que para alguns o tratamento fosse superficial. Aqui, esses modelos são resumidos em capítulos e estudos comparativos são desenvolvidos ao longo do texto.

Além disso, vários outros modelos encontrados em artigos de pesquisas recentes ao invés de livros, são aqui estudados, mesmo que em alguns casos sejam apenas referenciados os artigos mais relevantes. Entre esses modelos citam-se: os modelos aditivos generalizados, os modelos semi-paramétricos, os modelos de quase-verossimilhança e os modelos não-exponenciais não-lineares (ou modelos de dispersão).

Na grande maioria dos modelos estudados admite-se que as observações são independentes. Os modelos de regressão para análise de dados correlacionados, como os modelos de séries temporais, por requererem um estudo amplo, foram apenas citados no Capítulo 1. Também não se desenvolve a teoria dos modelos normais heterocedásticos.

O pré-requisito para a leitura é um Curso de Inferência Estatística, com base em Teoria da Verossimilhança, à nível de graduação. O texto, dividido em 8 capítulos, se destina prioritariamente a alunos de mestrado e doutorado, embora os 6 primeiros capítulos possam ser utilizados por alunos dos últimos anos de graduação.

O Capítulo 1 descreve o modelo clássico de regressão ou modelo normal-linear e o Capítulo 2 trata do modelo de Box e Cox. Os modelos log-lineares e os modelos para respostas binárias são desenvolvidos nos Capítulos 3 e 4, respectivamente. O Capítulo 5 apresenta os modelos normais não-lineares. Os modelos lineares generalizados, que engloba os modelos estudados nos Capítulos 1,2,3 e 4, são abordados no Capítulo 6. O Capítulo 7 apresenta os modelos não-exponenciais não-lineares, que são uma extensão dos modelos lineares generalizados, e o Capítulo 8 trata de modelos de regressão mais complexos, resultantes de pesquisas recentes.

O Capítulo 1 é pré-requisito de todos os capítulos e os Capítulos 7 e 8 requerem o conhecimento do assunto tratado no Capítulo 6. Não se exige outras dependências. Os Capítulos 7 e 8 são destinados a alunos de doutorado e pesquisadores. Em cada capítulo, exemplos procuram elucidar a teoria apresentada e a série de exercícios no final visa a exercitar o leitor sobre o assunto abordado.

O método de estimação de máxima verossimilhança é usado exaustivamente em todo o texto e outros métodos alternativos são apenas citados.

Muitas demonstrações foram omitidas, principalmente, quando o entendimento do assunto apresentado não requer o conhecimento das mesmas. Isto ocorre em grande parte no Capítulo 1.

Tentou-se minimizar o número de falhas e erros, os quais são de inteira responsabilidade dos autores, que ficarão agradecidos em receber quaisquer notificações sobre os mesmos.

Muitas pessoas contribuiram para este trabalho e os autores agradecem: à Comissão Organizadora do 17º Colóquio Brasileiro de Matemática pelo convite para ministrar curso baseado neste texto e à Lais Ventura Santos, Rogerio Dias trindade e Luis Alberto da Silva Santos, pelo trabalho de preparação dos originais.

Finalmente, os autores estarão recompensados, se este texto contribuir, de alguma forma, para o estudo e o desenvolvimento dos modelos de regressão no Brasil.

Rio, Maio de 1989.

Gauss M. Cordeiro
Gilberto A. Paula

CONTEÚDO

PREFÁCIO i

CONTEÚDO v

CAPÍTULO 1–MODELO NORMAL-LINEAR 1
§1.1. Introdução 1
§1.2. Estimação de Máxima Verossimilhança 2
§1.3. Análise de Variância 6
§1.4. Acréscimo na soma de Quadrados de Resíduos 10
§1.5. Estrutura Probabilística e distribuição de Formas Quadráticas 12
§1.6. Testes de Hipóteses 15
§1.7. Intervalos e Regiões de Confiança 16
§1.8. Técnicas de Diagnóstico 18
§1.8.1. Matriz de Projeção 19
§1.8.2. Resíduos 19
§1.8.3. Influência 21
§1.8.4. Técnicas Gráficas 24
§1.9. Predição da Regressão 25
§1.9.1. Intervalos de Confiança para a Média 26
§1.9.2. Intervalos de Confiança para um Conjunto de Observações 27
§1.10. Mínimos Quadrados Ponderados 28
§1.11. Dificuldades no Uso do Modelo Normal-linear 30
§1.11.1. Não-linearidade e Variância Não-constante 31
§1.11.2. Não-normalidade e Erros Correlacionados 33
§1.11.3. Erro na matriz modelo e multicolinearidade 34
§1.12. As Classificações de um e de dois Fatores 37
§1.13. Métodos de Seleção de modelos 41
§1.14. Modelos de Regressão Polinomial 44
§1.15. Modelos de Regressão "ridge" 46
§1.16. Modelos Heterocedásticos 48
§1.17. Modelos de Regressão com Estrutura de Correlação 52
§1.18. Exemplos 57
§1.19. Exercícios 64

CAPÍTULO 2–MODELO DE BOX E COX 68
§2.1. Definição 68
§2.2. Estimação da Transformação 70
§2.3. Adição de uma Covariável no Modelo Normal-linear 71
§2.4. Teste de Transformação da Variável Dependente 73
§2.5. Eliminação de Observações 76
§2.6. Teste de Transformação das Variáveis Explicativas 77

§2.7. Testes de Normalidade e Homocedasticidade 79
§2.8. Análise de Dados em Engenharia de Avaliações 82
§2.9. Exercícios 94

CAPÍTULO 3–MODELOS PARA ANÁLISE DE DADOS CATEGORIZADOS 96
§3.1. A Disgribuição de Poisson 96
§3.2. O Modelo Multinomial 99
§3.3. Inferência Sobre o Parâmetro da Distribuição de Poisson 100
§3.4. A Classificação Unidimensional 101
§3.5. A Classificação Bidimensional 104
§3.6. Modelos Log-lineares Hierárquicos 108
§3.6.1. Modelos Hierárquicos Possíveis para a Classificação Tri-dimensional . . 110
§3.6.2. Modelos Hierárquicos para a Classificação Multi-dimensional 113
§3.7. O Algoritmo de Ajustamento 115
§3.8. Testes de Adequação 118
§3.9. Dois Exemplos de Análise de Dados 121
§3.9.1. Dados de Acidentes de Trânsito 121
§3.9.2. Dados de Preferência por Partido Político 124
§3.10. Exercícios 126

CAPÍTULO 4–MODELOS PARA RESPOSTAS BINÁRIAS 131
§4.1. Introdução 131
§4.2. Distribuição Condicional para uma Única Tabela 2×2 132
§4.2.1. Exemplo 138
§4.3. Combinação de Tabelas 2×2 139
§4.3.1. Testes para os Riscos Relativos 141
§4.3.2. Exemplo 144
§4.4. Modelo Logístico Linear Simples 146
§4.4.1. Estimativas de Mínimos Quadrados para α e β 148
§4.4.2. Exemplo 150
§4.5. Modelo Logístico Linear Múltiplo 151
§4.5.1. Estimação dos Parâmetros pelo Método de Máxima Verossimilhança . 155
§4.5.2. Função Desvio e Resultados Assintóticos 156
§4.5.3. Testes de Hipóteses 157
§4.5.4. Seleção de Covariáveis 159
§4.5.5. Resíduos 161
§4.5.6. Exemplos 162
§4.6. Outros Modelos 169
§4.6.1. Modelos Probit e Complementar Log-log 169
§4.6.2. Modelo Logístico Condicional 171
§4.6.3. Modelos Logísticos Não-lineares 173
§4.7. Exercícios 176

CAPÍTULO 5–MODELO NORNAL NÃO LINEAR 178
§5.1. Introdução 178

§5.2. Estimação de Máxima Verossimilhança 183
§5.2.1. Resultados Assintóticos 187
§5.2.2. Exemplos 189
§5.3. Medidas de Não-linearidade 191
§5.3.1. Medidas de Curvatura de Bates e Watts 195
§5.3.2. Viés de Ordem n^{-1} de Box 198
§5.3.3. Aperfeiçoamento da Razão de Máxima Verossimilhança 200
§5.3.4. Exemplos 202
§5.4. Técnicas de Diagnóstico 204
§5.4.1. Matriz de Projeção 204
§5.4.2. Resíduo Projetado 204
§5.4.3. Medidas de Influência 208
§5.4.4. O Gráfico da Variável Adicionada 209
§5.4.5. Exemplos 210
§5.5. Exercícios 218

CAPÍTULO 6–MODELOS LINEARES GENERALIZADOS 222
§6.1. Definição 222
§6.2. Etapas de Trabalho com os Modelos Lineares Generalizados 227
§6.3. Estimação 229
§6.4. Medidas da Qualidade do Ajuste 232
§6.5. Análise do Desvio 236
§6.6. Distribuições Assintóticas e Regiões de Confiança 238
§6.7. Técnicas de Diagnóstico 240
§6.7.1. Análise de Resíduos 241
§6.7.2. Análise Global de Influência 243
§6.7.3. Diagnóstico Local de um único Ponto Influente 244
§6.8. Método das Covariáveis Adicionadas 246
§6.9. Análise dos Dados da Tabela 2.1 Através de um Modelo Gama 250
§6.10. Exercícios 254

CAPÍTULO 7–MODELOS NÃO-EXPONENCIAIS NÃO-LINEARES . . 258
§7.1. Uma Classificação dos Modelos de Regressão 258
§7.2. Alguns Modelos Não-exponenciais 262
§7.3. Algoritmo de Ajustamento 264
§7.4. Teste de Adequação 267
§7.5. Seleção de Covariáveis 270
§7.6. Distribuições Assintóticas 272
§7.7. Medidas de Diagnóstico 273
§7.8. Uma Análise de Dados Reais 277
§7.9. Testes de Hipóteses e Regiões de Confiança 278
§7.10. Exercícios 282

CAPÍTULO 8–MODELOS DE REGRESSÃO MAIS COMPLEXOS . . . 285
§8.1. Modelo Linear Generalizado com um Parâmetros Não-linear Extra 285
§8.2. Modelos Lineares Generalizados com Ligação Composta 287

§8.3. Modelos Aditivos Generalizados 288
§8.4. Modelos Semi-paramétricos 290
§8.5. Modelos para Análise de Dados de Sobrevivência 291
§8.5.1. Modelos de Riscos Proporcionais 292
§8.5.2. Riscos Proporcionais de Cox 295
§8.5.3. Riscos Não-proporcionais 299
§8.6. Uma Classe de Modelos Definida por duas Transformações 302
§8.7. Modelos de Quase-verossimilhança 306
§8.7.1. Modelo de Quase-verossimilhança com Função de Variância Paramétrica310
§8.7.2. Modelo de Quase-verossimilhança com Parâmetro de Dispersão Não-constante 315
§8.8. Modelos de Regressão com Estrutura de Autocorrelação Interna 310
§8.9. Outros Modelos Especiais 320
§8.10. Exercícios 322

REFERÊNCIAS 324
PALAVRAS CHAVES 342

CAPÍTULO 1

MODELO NORMAL-LINEAR

§1.1 Introdução

A análise de dados através de regressão linear é uma das técnicas mais usadas de estimação, havendo uma ampla literatura sobre o assunto. O leitor poderá encontrar nos seguintes livros os principais tópicos relacionados com regressão linear: Scheffé (1959), Rao (1965), Searle (1971), Seber (1977), Arnold (1981), Draper e Smith (1981), Cook e Weisberg (1982), Montgomery e Peck (1982) e mais recentemente 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 subseqüentes, nos quais serão apresentados modelos de regressão mais complexos.

O modelo de regressão linear clássico, também denominado *modelo normal-linear*, é definido por:

(i) Respostas y_i, $i = 1, \ldots, n$, independentes cada uma seguindo uma distribuição normal com média μ_i e variância σ^2 constante;

(ii) Cada média μ_i dada por $\eta_i = x_i^T \beta$, $i = 1, \ldots, n$, onde η_i é chamado de preditor linear, $x_i = (1, x_{i1}, \ldots, x_{i(p-1)})^T$ é um vetor $p \times 1$ com os

valores de $p-1$ variáveis explicativas e $\beta = (\beta_0, \ldots, \beta_{p-1})^T$ é o vetor de parâmetros de dimensão p a ser estimado.

A estrutura (i) e (ii) pode também ser expressa na forma matricial

$$y = X\beta + \varepsilon,$$

onde $y = (y_1, \ldots, y_n)^T$, $\varepsilon \sim N_n(0, \sigma^2 I)$, X é uma matriz $n \times p$ de linhas x_i^T, $i = 1, \ldots, n$ e I é a matriz identidade de ordem n.

Nas Seções 1.1 e 1.3 são apresentados alguns resultados básicos, assim como a notação a ser utilizada neste capítulo. Na Seção 1.4 são discutidos os modelos normais-lineares com restrições nos parâmetros na forma de igualdades lineares. A estrutura probabilística de Gauss-Markov e uma versão simplificada do teorema de Fisher-Cochran são apresentados na Seção 1.5. Testes de hipóteses e intervalos de confiança são deduzidos nas Seções 1.6 e 1.7. São enfocadas na Seção 1.8 as principais técnicas de diagnóstico para a regressão normal-linear. Previsões a partir do modelo ajustado são discutidas na Seção 1.9. O método de mínimos quadrados ponderados é apresentado na Seção 1.10. Algumas aplicações no uso do modelo normal-linear são vistas na Seção 1.11. Alguns modelos normais de uso freqüente são introduzidos nas Seções 1.12, 1.14, 1.16 e 1.17. A Seção 1.13 trata da relação de variáveis explicativas, a Seção 1.15 de regressão "ridge" e na Seção 1.18 duas análises de dados reais são apresentadas.

§1.2 Estimação de Máxima Verossimilhança

A função de probabilidade para y_i, supondo (i) e (ii) da Seção 1.1, fica

expressa na forma

$$(1.1) \qquad (2\pi)^{-1/2}\sigma^{-1}\exp\{-\frac{1}{2\sigma^2}(y_i - x_i^T\beta)^T(y_i - x_i^T\beta)\},$$

$i = 1, \ldots, n$. Seja $L(\beta)$ o logaritmo da função de verossimilhança conjunta para β. De (1.1) tem-se

$$L(\beta) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta).$$

Supõe-se σ^2 desconhecido. Derivando a expressão acima com relação a β e σ^2, respectivamente, as estimativas de máxima verossimilhança $\hat{\sigma}^2$ e $\hat{\beta}$ são obtidas de

$$(1.2) \qquad X^TX\hat{\beta} = X^Ty$$

e

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}).$$

Se X tem posto completo (X^TX) é não-singular e a solução para as equações (1.2) (equações normais) é unicamente dada por

$$(1.3) \qquad \hat{\beta} = (X^TX)^{-1}X^Ty.$$

No caso de $A = X^TX$ ser singular o sistema (1.2) admite 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 (1.2).

As matrizes A^- dependem somente de X^TX e em geral não são únicas, exceto quando X^TX for não-singular. Tais matrizes são chamadas de inversas generalizadas e há uma excelente discussão em Wetherill et al. (1986, Capítulo 5) sobre as propriedades das mesmas. Do ponto de vista prático

X^TX é sempre não-singular e ocorrem problemas sérios na obtenção de $(X^TX)^{-1}$ quando essa é mal condicionada. Alguns aspectos numéricos sobre esse problema serão discutidos na Seção 1.11.

Será mostrado na Seção 1.5 que a estimativa $\hat{\sigma}^2$ é viesada e tem valor esperado $E(\hat{\sigma}^2) = (n-p)\sigma^2/n$, o que sugere a utilização da estimativa não-viesada

$$s^2 = \frac{1}{n-p}(y - X\hat{\beta})^T(y - X\hat{\beta}).$$

A estimativa de máxima verossimilhança de β independe de σ^2. No Capítulo 7 demonstra-se que esta propriedade é também satisfeita para os modelos não-exponenciais não-lineares com $\hat{\beta}$ independente do parâmetro de dispersão, análogo à σ^2.

A solução (1.3) é também obtida aplicando o *método de mínimos quadrados*, que consiste em minimizar a soma de quadrados dos erros

$$\sum_i \varepsilon_i^2 = (y - X\beta)^T(y - X\beta).$$

Esse método não, necessariamente, produzirá as mesmas estimativas que o método de máxima verossimilhança quando ε não tiver distribuição normal. Para ilustrar, suponha que os ε_i, $i = 1, \ldots, n$, sejam independentes e distribuídos conforme uma exponencial dupla

$$f(\varepsilon_i) = (2\sigma)^{-1}e^{-|\varepsilon_i|/\sigma}, \quad i = 1, \ldots, n.$$

Então, a estimativa de máxima verossimilhança de β é obtida minimizando a soma dos erros absolutos $\sum|\varepsilon_i|$ em vez da soma de quadrados dos erros. Essa metodologia tem sido investigada qando ε_i é normal sendo denominada regressão-L_1. O método dos mínimos quadrados corresponde à regressão-L_2.

Existem vários outros métodos de estimação no modelo normal-linear que não serão estudados aqui. O método de estimação-M (Huber, 1973) substitui a soma de quadrados dos erros $\sum \varepsilon_i^2$ por $\sum \rho(\varepsilon_i)$, onde $\rho(\cdot)$ é uma função simétrica. A minimização de $\sum \rho(\varepsilon_i)$ com relação aos β's produz o sistema de p equações não-lineares

$$\sum_{i=1}^{n} \rho'(\varepsilon_i) x_i = 0,$$

cuja solução é obtida por procedimentos iterativos. Uma possível escolha para $\rho(x)$ verifica $\rho'(x) = \min\{c, \max(t, -c)\}$. O método de estimação-$M$ produz estimativas mais eficientes que a regressão-L_1 e mais robustas em relação a dados aberrantes.

Entre vários outros métodos de estimação citam-se: estimação-M generalizada; mínima mediana quadrada que corresponde a $\min_{\beta} \operatorname{med}_{i} \varepsilon_i^2$; estimação-$S$ que equivale a $\min_{\beta} S(\beta)$, onde $S(\beta)$ é um certo tipo de função definida na escala dos erros $\varepsilon_1(\beta), \ldots, \varepsilon_n(\beta)$; mínimos quadrados ponderados definido por $\min_{\beta} \sum w_i \varepsilon_i^2$ (Seção 1.10); estimação "ridge" (Seção 1.15); e estimação de James-Stein (vide Arnold, 1981).

Os valores ajustados $\hat{\mu}_i$, $1, \ldots, n$, são obtidos de $\hat{\mu} = Hy$, onde $H = X(X^T X)^{-1} X^T$ é a matriz de projeção ortogonal de vetores de $\mathbf{R}^n$ no subespaço gerado pelas colunas de X. Esta matriz é simétrica, idempotente e tem posto p.

Verifica-se que $E(\hat{\mu}^T \hat{\mu}) = E(\mu^T \mu) + p\sigma^2$ e, portanto, o comprimento de μ é superestimado pelo método de máxima verossimilhança. A estimativa $\hat{\mu}$ pode ser aperfeiçoada multiplicando por um escalar $a(\hat{\mu}, \hat{\sigma}^2)$ em $[0, 1]$. A

estimativa de μ definida por

$$a(\hat{\mu}, \hat{\sigma}^2)\hat{\mu} = \left(1 - \frac{c\hat{\sigma}^2}{\hat{\mu}^T\hat{\mu}}\right)\hat{\mu},$$

onde c é uma constante escolhida para minimizar o risco médio correspondente a uma função de perda quadrática, é denominada estimativa de James-Stein. Demonstra-se que o valor ótimo de c iguala $(p-2)(n-p)/(n-p+2)$.

Para encerrar esta seção verifica-se o efeito da escala das variáveis explicativas e da variável resposta nas estimativas dos β's em (1.3) e sobre $\hat{\sigma}^2$ ou s^2. Considere que $x_j = (x_{1j}, \ldots, x_{nj})^T$ é a j-ésima coluna de X correspondente a β_j, $j = 1, \ldots, p-1$. Se a variável explicativa x_j é substituída por $x_j' = ((x_{1j} - c_j)/d_j, \ldots, (x_{nj} - c_j)/d_j)^T$ então os coeficientes estimados $\hat{\beta}_0, \hat{\beta}_j$, $j = 1, \ldots, p-1$ tornam-se $\hat{\beta}_0' = \hat{\beta}_0 + (c_j/d_j)\hat{\beta}_j$ e $\hat{\beta}_j' = d_j\hat{\beta}_j$, $j = 1, \ldots, p-1$. As estimativas $\hat{\sigma}^2$ e s^2 e as estimativas R^2 e F's, a serem apresentadas nas Seções 1.3 e 1.6, permanecem inalteradas.

Se y é substituído por $(y-f)/g$ então $\hat{\beta}_0' = (\hat{\beta}_0 - f)/g$, $\hat{\beta}_j' = \hat{\beta}_j/g$, $j = 1, \ldots, p-1$, e todas as somas de quadrados na análise de variância (Seção 1.3) e $\hat{\sigma}^2$ ou s^2 ficam divididos por g^2. Entretanto, os testes baseados nas estatísticas F permanecem inalterados. O meio natural de transformação de escala é a padronização das variáveis, isto é, substituir x_j por $(x_j - \overline{x}_j)/DP_j$ e y por $(y - \overline{y})/DP_y$, onde DP representa desvio-padrão.

§1.3 Análise de Variância

Os afastamentos dos valores ajustados $\hat{\mu}_i$'s são avaliados examinando-se os resíduos ordinários $r_i = y_i - \hat{\mu}_i$, $i = 1, \ldots, n$ os quais serão discutidos com

detalhes na Seção 1.8. O vetor de resíduos $r = (r_1, \ldots, r_n)^T$ é também obtido de

$$r = (I - H)y, \tag{1.4}$$

ou seja, r é a projeção ortogonal do vetor y no ortocomplemento do subespaço gerado pelas colunas da matriz X.

Uma medida resumida da qualidade do ajuste é a soma de quadrados de resíduos dada por

$$SQ\,\text{Res} = \sum r_i^2 = (y - X\hat{\beta})^T(y - X\hat{\beta}).$$

Utilizando (1.4), a expressão acima fica expressa numa forma mais conveniente

$$SQ\,\text{Res} = y^T(I - H)y.$$

Mostra-se, sob a hipótese de que o modelo assumido é verdadeiro, que $y^T(I - H)y\ \sigma^2$ tem distribuição χ^2 com $(n - p)$ graus de liberdade.

A técnica mais usual para verificar a adequação do ajuste é a *Análise de Variância da Regressão*, que utiliza a seguinte identidade:

$$\sum(y_i - \overline{y})^2 = \sum(\hat{y}_i - \overline{y})^2 + \sum(y_i - \hat{\mu}_i)^2. \tag{1.5}$$

O termo do lado esquerdo da expressão acima é a soma de quadrados total (corrigida) que será denotada por SQT, enquanto o primeiro termo do lado direito é a soma de quadrados devida à regressão de y sobre as $(p - 1)$ variáveis explicativas, que será denotada por $SQReg$, e o segundo termo sendo $SQ\,\text{Res}$.

Uma maneira de se medir a adequação do ajuste é comparando a soma de quadrados residual (que se espera seja pequena) com a soma de quadrados devida à regressão. Ou, alternativamente, a soma de quadrados devida

à regressão com a soma de quadrados total. A razão desses dois termos é representada por

$$R^2 = \frac{(\hat{\beta}^T X^T y - n\overline{y}^2)}{(y^T y - n\overline{y}^2)}.$$

Essa razão varia sempre entre 0 e 1 e é também chamada de *coeficiente de correlação linear múltiplo ao quadrado.* Esse nome deve-se ao fato de R ser o coeficiente de correlação múltipla entre y e $\hat{\mu}$. Alguns pesquisadores se baseiam erroneamente apenas no valor de R^2 para escolha do melhor modelo. É tão importante, quanto ter um R^2 alto, que a estimativa de σ^2 seja também muito pequena, já que em geral os intervalos de confiança de interesse são proporcionais a σ^2.

A cada soma de quadrados de (1.5) está associado um número de graus de liberdade, que é formalmente definido expressando a soma de quadrados correspondente como uma forma quadrática. Isso será discutido na Seção 1.5. Contudo, há uma forma intuitiva de interpretar os graus de liberdade associados a cada soma de quadrados. Para ilustrar, suponha o modelo linear simples $\mu_i = \beta_0 + \beta_1 x_{i1}$, $i = 1, \ldots, n$. As estimativas de mínimos quadrados para β_0 e β_1 saem de (1.3), e são, respectivamente, dadas por

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}_1 \tag{1.6}$$

e

$$\hat{\beta}_1 = \sum_i (x_{i1} - \overline{x}_1)(y_i - \overline{y}) / \sum_i (x_{i1} - \overline{x}_1)^2. \tag{1.7}$$

O número de graus de liberdade associado à soma de quadrados total é $(n-1)$, pois essa soma pode ser obtida a partir dos $n-1$ valores $y_1 - \overline{y}, \ldots, \overline{y}_{n-1} - \overline{y}$, já que $\sum(y_i - \overline{y}) = 0$. Similarmente a soma de quadrados devida à regressão pode ser obtida, usando (1.6) e (1.7), diretamente da expressão

$\sum(\hat{y}_i - \overline{y})^2 = \hat{\beta}_1^2 \sum(x_{i1} - \overline{x}_1)^2$, que é uma função única de $y_1, y_2, \ldots, y_n$, através de $\hat{\beta}_1$. Logo, o número de graus de liberdade associado a essa soma vale 1. Por subtração o número de graus de liberdade associado à SQ Res é $(n-2)$.

No caso geral com p parâmetros, os graus de liberdade associados à $SQReg$ e SQ Res valem, respectivamente, $(p-1)$ e $(n-p)$.

Tem-se portanto a Tabela 1.1 da Análise de vavriância (ANOVA) da regressão na sua forma resumida.

Tabela 1.1: - *Tabela resumida da Análise de variância da Regressão* $y = X\beta + \varepsilon$.

Efeito	Soma de Quadrados	G.l.	Quadrado médio
Regressão			
$(\beta_1, \ldots, \beta_{p-1})/\beta_0)$	$\hat{\beta}^T X^T y - n\overline{y}^2$	$p-1$	$SQReg/(p-1)$
Residual	$y^T y - \hat{\beta}^T X^T y$	$n-p$	$SQ\,\text{Res}/(n-p)$
Total	$y^T y - n\overline{y}^2$	$n-1$	-

A estatística F definida por

$$F = \frac{SQReg/(p-1)}{SQ\,\text{Res}/(n-p)} = \frac{(n-p)R^2}{(p-1)(1-R^2)}$$

é usualmente utilizada para avaliar a significância simultânea dos coeficientes $\beta_1, \ldots, \beta_{p-1}$. A distribuição dessa estatística será discutida na Seção 1.5.

§1.4 Acréscimo na Soma de Quadrados de Resíduos

Serão abordados agora alguns aspectos da Análise de Variância no caso de se imporem restrições nos parâmetros do preditor linear η. Esses aspectos terão importância fundamental no desenvolvimento de testes de hipóteses.

Com a imposição de restrições nos parâmetros a soma de quadrados de resíduos do modelo correspondente fica inflacionada. O valor do acréscimo, em relação à soma de quadrados de resíduos sem restrições, pode ser usado para avaliar a variação na qualidade do ajuste com as restrições. Logo, é importante determinar esse acréscimo.

Suponha então o seguinte modelo:

$$y = X\beta + \varepsilon \quad \text{sujeito à} \quad C\beta = 0,$$

onde C é uma matriz $m \times p$ de constantes e $\varepsilon \sim N_n(0, \sigma^2 I)$.

Se foram observados os n valores $y_1, \ldots, y_n$, então a estimativa de mínimos quadrados de β é obtida minimizando a função Lagrangeana

$$\mathcal{L}(\beta; \Lambda) = (y - X\beta)^T(y - X\beta) + 2\Lambda^T C\beta, \tag{1.8}$$

onde $\Lambda = (\lambda_1, \ldots, \lambda_m)^T$ é o vetor de multiplicadores de Lagrange, $\lambda_j \geq 0$, $j = 1, \ldots, m$.

Após alguma álgebra chega-se à seguinte solução para (1.8):

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

onde $\hat{\Lambda} = \{C(X^T X)^{-1} C^T\}^{-1} C\hat{\beta}$.

Os valores ajustados e os resíduos ordinários serão, respectivamente, obtidos de

$$\tilde{\mu} = X\tilde{\beta} = H(y - W^T \hat{\Lambda}) \tag{1.9}$$

e

$$r = (y - X\tilde{\beta}) = (I - H)y + HW^T\hat{\Lambda},$$

onde $W = C(X^TX)^{-1}X^T$.

Assim, usando (1.9), a soma de quadrados dos resíduos fica dada por

$$SQ\,\text{Res}(C\beta = 0) = (y - X\hat{\beta})^T(y - X\hat{\beta}) + ASQ(C\beta = 0),$$

onde $ASQ(C\beta = 0) = \hat{\Lambda}^T WHW^T\hat{\Lambda}$.

Portanto, impondo as restrições $C\beta = 0$, a soma de quadrados dos resíduos do modelo irrestrito ficará acrescida da quantidade $\hat{\Lambda}^T WHW^T\hat{\Lambda}$, que também é expressa na forma

$$ASQ(C\beta = 0) = (C\hat{\beta})^T\{C(X^TX)^{-1}C^T\}^{-1}C\hat{\beta}.$$

Mostra-se que a essa soma de quadrados estão associados m graus de liberdade.

Em particular, a soma de quadrados da regressão supondo $C\beta = 0$, será dada por

$$SQReg(C\beta = 0) = \tilde{\mu}^T\tilde{\mu} = (X\tilde{\beta})^T(X\tilde{\beta}),$$

estando associados à mesma $(p-m)$ graus de liberdade. A soma de quadrados corrigida, fica nesse caso denotada por

$$SQReg(C\beta = 0/\beta_0) = (X\tilde{\beta})^T(X\tilde{\beta}) - n\overline{y}^2,$$

tendo $(p - m - 1)$ graus de liberdade.

§1.5 Estrutura Probabilística e Distribuição de Formas Quadráticas

Na Seção 1.1 foi assumida a seguinte estrutura probabilística para os erros ε_i's: (i) são mutuamente independentes e (ii) cada um tem distribuição normal com média zero e variância σ^2. Agora essa estrutura será utilizada para se determinar algumas propriedades das estatísticas definidas nas Seções 1.1 a 1.3.

É importante ressaltar que dificilmente se encontra essa estrutura na prática, havendo diversas técnicas para detectar afastamentos sérios da mesma. Algumas dessas técnicas serão discutidas nas Seções 1.8 e 1.11.

Tem-se portanto o modelo $y = X\beta + \varepsilon$ com $\varepsilon \sim N_n(0, \sigma^2 I)$. Daí segue que y tem distribuição normal multivariada com média $E(y) = X\beta$ e variância $Var(y) = \sigma^2 I$.

Mostra-se facilmente que o estimador irrestrito de mínimos quadrados $\hat{\beta}$ tem distribuição normal p-variada com média β e matriz de variância-covariância dada por

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

Apesar de os erros serem mutuamente não correlacionados, o vetor de resíduos ordinários, cujo valor esperado é dado por

$$E(r) = (I - H)X\beta = 0,$$

tem, em geral, matriz de variância-covariância não diagonal

$$Cov(r) = \sigma^2 (I - H). \tag{1.10}$$

A distribuição dos r_i's será discutida na Seção 1.8.

Como $I - H$ é idempotente, de (1.10) segue

$$E(r^T r) = \sum E(r_i^2) = \sigma^2 \sum (1 - h_{ii}) = (n-p)\sigma^2,$$

onde h_{ii} é o i-ésimo elemento da diagonal principal da matriz H. Logo, uma estimativa não-viesada para σ^2 é $s^2 = (n-p)^{-1} r^T r$.

O teorema de Fisher-Cochran (Searle, 1971) que será apresentado a seguir numa forma simplificada, é fundamental para se determinar a distribuição das somas de quadrados definidas na Tabela 1.2.

Suponha que $y \sim N_n(X\beta, \sigma^2 I)$ e sejam A_1, A_2 e A_3 matrizes $n \times n$ de postos r_1, r_2 e r_3, respectivamente, tais que $r_1 + r_2 + r_3 = n$.

Se for possível escrever

$$y^T y = y^T A_1 y + y^T A_2 y + y^T A_3 y,$$

e se essas matrizes forem tais que

$$A_i A_j = 0, \ i \neq j,$$

então

$$y^T A_i y \sim \sigma^2 \chi^2_{r_i}(\lambda_i),$$

com parâmetro de não-centralidade dado por

$$\lambda_i = (X\beta)^T A_i (X\beta), \ i = 1, 2, 3,$$

e as formas quadráticas $y^T A_i y$ são mutuamente independentes.

Utilizando os resultados das seções anteriores é possível mostrar, após alguma álgebra, que $SQ\,\mathrm{Res} = y^T A_1 y$, $ASQ(C\beta = 0) = y^T A_2 y$, $SQReg(C\beta = 0) = y^T A_3 y$, onde $A_1 = I - H$ e A_2 e A_3 são matrizes $n \times n$ de postos m e $(p - m)$, cujas expressões são omitidas aqui. Além disso

tem-se $\lambda_1 = 0$, $\lambda_2 = \lambda_3 = 0$, pois $C\beta = 0$, e $A_1A_2 = A_1A_3 = A_3A_2 = 0$. Pelo teorema de Fisher-Cochran, desde que $y^Ty = SQ\,\text{Res} + ASQ(C\beta = 0) + SQReg(C\beta = 0)$, segue o seguinte: (i) $SQ\,\text{Res} \sim \sigma^2\chi^2_{(n-p)}$; (ii) $ASQ(C\beta = 0) \sim \sigma^2\chi^2_m$; (iii) $SQReg(C\beta = 0) \sim \sigma^2\chi^2_{(p-m)}$; (iv) as formas quadráticas $SQ\,\text{Res}, ASQ(C\beta = 0)$ e $SQReg(C\beta = 0)$ são mutuamente independentes.

Mostra-se também que $SQ\,\text{Res}(C\beta = 0) \sim \sigma^2\chi^2_{(n-p+m)}$ e $SQ\,\text{Reg} \sim \sigma^2\chi^2_{(p-1)}(\lambda)$, sendo o parâmetro de não-centralidade dado por

$$\lambda = (X\beta)^T(I - J)(X\beta),$$

onde J é uma matriz $n \times n$ de 1's.

Tabela 1.2: - *Tabela usual da Análise de Variância para o modelo irrestrito* $y = X\beta + \varepsilon$ *e impondo-se as restrições* $C\beta = 0$.

Efeito	Soma de Quadrados	G.l.	Quadrado médio
Regressão			
$(\beta_1, \ldots, \beta_{p-1}/\beta_0)$	$\hat{\beta}^TX^Ty - n\overline{y}^2$	$p-1$	$SQ\,\text{Reg}/(p-1)$
Regressão			
$(C\beta = 0/\beta_0)$	$(X\tilde{\beta})^T(X\tilde{\beta}) - n\overline{y}^2$	$p-m-1$	$SQ\,\text{Reg}(C\beta = 0/\beta_0)/(p-m-1)$
Acréscimo			
Residual devido	$(C\hat{\beta})^T\{C(X^TX)^{-1}C^T\}^{-1}C\hat{\beta}$	m	$ASQ(C\beta = 0)/m$
a $C\beta = 0$			
Resíduos	$y^Ty - \hat{\beta}^TX^Ty +$		
$(C\beta = 0)$	$(C\hat{\beta})^T\{C(X^TX)^{-1}C^T\}^{-1}C\hat{\beta}$	$n-p+m$	$SQ\,\text{Res}(C\beta = 0)/(n-p+m)$
Resíduos	$y^Ty - \hat{\beta}^TX^Ty$	$n-p$	$SQ\,\text{Res}/(n-p)$
Total	$y^Ty - n\overline{y}^2$	$n-1$	-

A estatística F, definida por

$$F = \frac{ASQ(C\beta = 0)/m}{SQ\,\mathrm{Res}/(n-p)}$$

é usualmente utilizada para se avaliar a imposição das restrições $C\beta = 0$ no modelo irrestrito.

§1.6 Testes de Hipóteses

Pelos últimos resultados da seção anterior, sob a hipótese $H\colon \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, $SQReg/\sigma^2$ tem uma distribuição qui-quadrado central com $(p-1)$ graus de liberdade, já que $\lambda = 0$ sob essa hipótese. Portanto, a estatística

$$F = \frac{SQReg/(p-1)}{SQ\,\mathrm{Res}/(n-p)} = \frac{SQReg/(p-1)}{s^2}$$

tem, segundo H, uma distribuição F com $(p-1)$ e $(n-p)$ graus de liberdade. Para um nível de significância α rejeita-se H se $F \geq F_{(p-1),(n-p)}(\alpha)$, onde $F_{(p-1),(n-p)}(\alpha)$ é o quantil $(1-\alpha)$ da distribuição F com esses graus de liberdade.

Para se testar a hipótese $H\colon C\beta = 0$ utiliza-se a estatística

$$F = \frac{ASQ(C\beta = 0)/m}{s^2}, \tag{1.11}$$

que tem segundo H uma distribuição F com m e $(n-p)$ graus de liberdade.

Admite-se agora que L é uma matriz $(p-m) \times p$ que complementa C de modo que $\binom{C}{L}$ é não-singular. Verifica-se facilmente que $L\beta$ representa

$(p-m)$ funções linearmente independentes dos elementos de β. Considera-se o teste da sub-hipótese $L\beta = 0$ supondo que a hipótese $H: C\beta = 0$ é verdadeira. Demonstra-se que a seguinte estatística (vide Tabela 1.2)

$$F = \frac{SQ\,\text{Reg}(C\beta = 0/\beta_0)/(p-m-1)}{SQ\,\text{Res}(C\beta = 0)/(n-p+m)}$$

testa a sub-hipótese $L\beta = 0$ supondo que H é satisfeita. Esta estatística tem, segundo a sub-hipótese, distribuição F com $(p-m-1)$ e $(n-p+m)$ graus de liberdade.

Apenas para ilustrar a utilização das estatísticas deduzidas da Tabela 1.2, considere o modelo de regressão

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

restrito a $C\beta = 0$ com $C = (0\ 0\ 0\ 1)$. Este modelo é equivalente à hipótese $H: \beta_3 = 0$ sendo o teste realizado através de (1.11). Entretanto se o interesse é testar $\beta_1 = \beta_2 = 0$ supondo que $H: \beta_3 = 0$ é verdadeira, deve-se usar a última expressão da estatística F com $p = 4$ e $m = 1$, aqui o numerador e o denominador representando a soma de quadrados devido à regressão e dos resíduos, respectivamente, referentes ao modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. A sub- hipótese $\beta_1 = \beta_2 = 0$ será rejeitada se $F \geq F_{2,n-3}(\alpha)$.

Do que foi visto, pelo menos três hipóteses que ocorrem com freqüência na prática, poderão ser testadas utilizando as somas de quadrados da Tabela 1.2.

§1.7 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 $(X^TX)^{-1}$ ou invertendo as regiões de rejeição de alguns testes descritos na seção anterior.

Para ilustrar, considere inicialmente o modelo de regressão linear simples

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i.$$

$\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. Para se obter as variâncias de $\hat{\beta}_0$ e $\hat{\beta}_1$, é preciso inverter a matriz (X^TX), onde X é uma matriz $n \times 2$ de posto completo e de linhas dadas por $x_i = (1, x_{i1})^T$, $i = 1, \ldots, n$. Após alguma álgebra obtém-se

$$(X^TX)^{-1} = \{\sum(x_{i1} - \overline{x}_1)^2\}^{-1} \begin{pmatrix} n^{-1}\sum x_{i1}^2 & -\overline{x}_1 \\ -\overline{x}_1 & 1 \end{pmatrix}.$$

Logo, $\text{Var}(\hat{\beta}_0) = \sigma^2\{\sum(x_{i1} - \overline{x}_1)^2\}^{-1}\sum x_{i1}/n, \text{Var}(\hat{\beta}_1) = \sigma^2\{\sum(x_{i1} - \overline{x}_1)^2\}^{-1}$ e $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2\overline{x}_1\{\sum(x_{i1} - \overline{x}_1)^2\}^{-1}$, onde $\overline{x}_i = \sum x_{i1}/n$. Assim os limites de confiança de um intervalo de $100(1-\alpha)\%$ para β_1, serão dados por

$$\hat{\beta}_1 \pm t_{\alpha/2}\left\{\frac{\sum x_{i1}^2}{n\sum(x_{i1} - \overline{x}_1)^2}\right\}^{1/2} s,$$

onde $t_{\alpha/2}$ é o quantil $(1-\alpha/2)$ de uma distribuição t de student com $(n-2)$ graus de liberdade e s é a raiz quadrada da estimativa não-viesada de σ^2.

Se o interesse é encontrar um intervalo de $100(1-\alpha)\%$ de confiança para a combinação linear $C\beta = C_0\beta_0 + C_1\beta_1 + \cdots + C_{p-1}\beta_{p-1}$, tem-se

$$Var(C\hat{\beta}) = \sigma^2 C(X^TX)^{-1}C^T,$$

e os limites de confiança são

$$C\hat{\beta} \pm t_{\alpha/2}\{C(X^TX)^{-1}C^T\}^{-1/2}s.$$

No caso de C ser uma matriz $m \times p$, $m \geq 2$, em vez de se obter m intervalos simultâneos de confiança para as combinações lineares de $C\beta$, é mais usual construir regiões de confiança para $C\beta$. Essas regiões são obtidas considerando-se a região de aceitação

$$\frac{\{ASQ(C\beta = 0) - \lambda\}}{ms^2} \leq F_{m,(n-p)}(\alpha)$$

para o teste da hipótese H: $C\beta = 0$, onde $\lambda = (C\beta)^T\{C(X^TX)^{-1}C^T\}^{-1}C\beta$. Logo, uma região de $100(1-\alpha)\%$ de confiança para $C\beta$ é dada por

$$(C\hat{\beta} - C\beta)^T\{C(X^TX)^{-1}C^T\}^{-1}(C\hat{\beta} - C\beta) < s^2 m F_{m,(n-p)}(\alpha),$$

que é um elipsóide de dimensão m no subespaço das combinações lineares $C\beta$.

§1.8 Técnicas de Diagnóstico

Nesta seção serão apresentadas algumas técnicas de diagnóstico desenvolvidas para a classe dos modelos de regressão normal-linear. Essas técnicas procuram basicamente detectar problemas com o ajuste, os quais são em geral agrupados em quatro grandes classes: (i) afastamento sérios das suposições iniciais para os erros ε_i's e para as componentes η_i's; (ii) presença de observações mal ajustadas ("outliers"); (iii) presença de observações influentes; (iv) colinearidade entre as colunas da matrix X. Alguns desses problemas são tratados particularmente nas Seções 1.11 e 2.7. Há uma vasta literatura sobre o assunto, inclusive em português (vide Dachs e Carvalho, 1984). O leitor pode encontrar em Cook e Weisberg (1983) e Atkinson (1985) excelentes resenhas sobre essas técnicas.

1.8.1 - Matriz de Projeção

A matriz de projeção H definida na Seção 1.3, também conhecida como matriz "hat", pois transforma y em $\hat{\mu}$ ($\hat{\mu} = Hy$) é muito usual na detecção de pontos mais afastados dos demais (pontos com alto "leverage"). Esses pontos, além de serem potencialmente aberrantes e influentes, em geral exercem grande influência sobre a matriz $\sigma^2(X^TX)^{-1}$.

Pelo fato de H ser simétrica e idempotente é possível mostrar o seguinte: (i) $\frac{1}{n} \leq h_{ii} \leq 1$; (ii) $h_{ii} = \sum h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$; (iii) $trH = \sum h_{ii} = p$. O elemento h_{ii} de H representa a influência da i-ésima resposta sobre o i-ésimo valor ajustado. Logo, como $\sum h_{ii} = p$, supondo que todos os pontos exercem 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}. Hoaglin e Welsch (1978) sugerem $h_{ii} \geq 2p/n$ como guia para indicar pontos com um alto "leverage". Entretanto, outras medidas de diagnóstico sempre serão necessárias para confirmar esse primeiro diagnóstico.

1.8.2 - Resíduos

O resíduo para a i-ésima observação pode ser definido como uma função $r_i = r(y_i, \hat{\mu}_i)$ que procura medir a discrepância entre o valor observado e o valor ajustado. O sinal de r_i indica a direção dessa discrepância. O resíduo ordinário definido por $r_i = y_i - \hat{\mu}_i$ é um dos mais usados. De (1.4) segue que r_i tem distribuição normal com média zero e variância $\sigma^2(1 - h_{ii})$. A variância entre r_i e r_j vale $-\sigma^2 h_{ij}$.

Para comparar os resíduos r_i's deve-se expressá-los em forma padronizada. A mais usual consiste na divisão de cada r_i pela estimativa do seu desvio padrão, obtendo-se assim o resíduo ordinário studentizado

$$t_i = r_i/\{s(1 - h_{ii})^{1/2}\}, \quad i = 1, \ldots, n. \tag{1.12}$$

Como r_i não é independente de s^2, t_i não segue uma distribuição t de "Student" como era de se esperar. Demonstra-se que $t_i/(n-p)$ tem uma distribuição beta com parâmetros $1/2$ e $(n-p-1)/2$. Logo, $E(t_i) = 0$, $Var(t_i) = 1$ e $Cov(t_i, t_j) = -h_{ij}/\{(1-h_{ii})(1-h_{jj})\}^{1/2}$, $i \neq j$.

O problema mencionado acima é contornado substituindo em (1.12) s^2 por $s^2_{(i)}$, que é a estimativa de σ^2 obtida sem a i-ésima observação. Assim tem-se uma estimativa de σ^2 que é independente de r_i. Prova-se que

$$s^2_{(i)} = s^2\left(\frac{n-p-r_i^2}{n-p-1}\right), \tag{1.13}$$

sendo o novo resíduo definido por

$$t_i^* = \frac{r_i}{s_{(i)}(1-h_{ii})^{1/2}}, \tag{1.14}$$

que tem uma distribuição t de Student com $(n-p-1)$ graus de liberdade. A relação entre t_i e t_i^* é obtida substituindo (1.13) em (1.14),

$$t_i^* = t_i\left(\frac{n-p-1}{n-p-t_i^2}\right)^{1/2},$$

mostrando que t_i^{*2} é uma transformação monótona de t_i^2.

O resíduo t_i^* tem uma interpretação muito interessante quando há suspeita de falta de ajuste na i-ésima observação. Essa falta de ajuste pode ser avaliada impondo-se o seguinte modelo:

$$y = X\beta + W_i\gamma + \varepsilon,$$

$$E(\varepsilon) = 0, \quad Var(\varepsilon) = \sigma^2 I,$$

onde W_i é um vetor $n \times 1$ com o i-ésimo elemento igual a um e todos os demais iguais a zero. Prova-se (vide Cook e Weisberg, 1982, pg. 21) que a soma de quadrados de resíduos para esse modelo vale

$$(n-p)s^2 - r_i^2/(1-h_{ii}).$$

Assim, o acréscimo na soma de quadrados de resíduos do modelo devido à restrição $\gamma = 0$, é dado por $ASQ(\gamma = 0) = r_i^2/(1-h_{ii})$. A estatística F para testar $H: \gamma = 0$, fica então expressa na forma

$$F = \frac{\{r_i^2/(1-h_{ii})\}(n-p-1)}{(n-p)s^2 - r_i^2/(1-h_{ii})}, \tag{1.15}$$

que segundo H e supondo normalidade para ε, tem uma distribuição F com 1 e $(n-p-1)$ graus de liberdade. Rejeita-se H se $F \geq F_{1,(n-p-1)}(\alpha)$. Usando as expressões (1.13) e (1.14) mostra-se facilmente que $F = t_i^{*2}$.

Sob a hipótese alternativa $H: \delta \neq 0$, a estatística F dada em (1.15) segue uma distribuição F com parâmetro de não centralidade $\gamma^2(1-h_{ii})/\sigma^2$. Nota-se que para h_{ii} próximo de um esse parâmetro é próximo de zero. Logo, encontrar pontos aberrantes ("outliers") entre os pontos mais afastados será mais difícil que encontrá-los entre os pontos com h_{ii} pequeno.

1.8.3 - Influência

É fundamental num modelo de regressão conhecer o grau de dependência entre o mesmo e as observações para as quais esse é ajustado. Será preocupante se pequenas perturbações nas observações produzirem mudanças nas estimativas obtidas. No entanto, se tais perturbações não alterarem os principais resultados, 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 um único ponto e procuram medir o impacto dessa perturbação nas principais estimativas do modelo. Esse método, entretanto, pode não ser adequado se duas ou mais observações forem responsáveis conjuntamente por um termo extra no preditor linear, na variável resposta ou nas covariáveis. Em pesquisas recentes, Atkinson (1986) e Cook (1986) tratam esse tipo de problema para o caso normal-linear, com o último fazendo algumas extenções para modelos mais gerais. O método gráfico da variável adicionada, que será apresentado na Seção 2.3, é freqüentemente utilizado para avaliar a influência conjunta das observações nas estimativas individuais dos parâmetros. Esta seção, entretanto, será restrita a algumas medidas de diagnóstico usuais na avaliação do grau de dependência entre $\hat{\beta}$ e cada uma das observações.

A influência do i-ésimo ponto sobre $\hat{\beta}$ pode ser estudada através de uma avaliação do efeito causado por pequenas perturbações no termo de $L(\beta)$ correspondente a esse ponto, re-expressando $L(\beta)$ na forma

$$L(\beta) = \sum_{j=1}^{n} \delta_j L(y_j; \mu_j(\beta)),$$

onde $\mu_j(\beta) = \mu_j$, $L(\cdot;\cdot)$ representa a log-verossimilhança correspondente a cada observação, $\delta_j = \delta$ para $j = i$ e $\delta_j = 1$ para $j \neq i$. Nesse caso a estimativa de máxima verossimilhança para β fica dada por

$$\hat{\beta}_\delta = (X^T \triangle X)^{-1} X^T \triangle y,$$

onde $\triangle$ é uma matriz $n \times n$ diagonal de 1's com δ na i-ésima posição.

Desde que

$$(X^T \triangle X)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i (1-\delta) x_i^T (X^T X)^{-1}}{\{1 - (1-\delta) h_{ii}\}},$$

$\hat{\beta}_\delta$ fica expresso em função de quantidades envolvendo todas as observações e apenas o i-ésimo ponto

$$\hat{\beta}_\delta = \hat{\beta} - \frac{(X^TX)^{-1}x_i(1-\delta)}{\{1-(1-\delta)h_{ii}\}}r_i. \tag{1.16}$$

Em particular fazendo $\delta = 0$ (exclusão do i-ésimo ponto) em (1.16), obtém-se

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X^TX)^{-1}x_i}{(1-h_{ii})}r_i, \tag{1.17}$$

onde $\hat{\beta}_{(i)}$ é a estimativa de máxima verossimilhança removendo-se a i-ésima observação.

Uma medida da influência da retirada do i-ésimo ponto sobre a estimativa $\hat{\beta}_j$, que sai diretamente de (1.17), é dada por

$$\triangle_i\hat{\beta}_j = (\hat{\beta}_j - \hat{\beta}_{(i)j})/DP(\hat{\beta}_j),$$

onde $DP(\cdot)$ denota o desvio padrão e $\hat{\beta}_{(i)j}$ é a j-ésima componente do vetor $\hat{\beta}_{(i)}$. Uma outra medida de influência muito conhecida é o D de Cook (Cook, 1977) definido por

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})^T(X^TX)(\hat{\beta}-\hat{\beta}_{(i)})}{ps^2}, \tag{1.18}$$

que lembra a expressão que fornece a região de confiança simultaneamente para todos os parâmetros do modelo definida na Seção 1.1. Em (1.18) $C = I$ e $\{C(X^TX)^{-1}C^T\} = (X^TX)^{-1}$. Logo, o elipsóide de $100(1-\alpha)\%$ de confiança para todos os parâmetros fica expresso na forma simplificada.

$$(\hat{\beta}-\beta)^T(X^TX)^{-1}(\hat{\beta}-\beta) < s^2pF_{p,(n-p)}(\alpha).$$

Usando (1.17) em (1.18) tem-se uma expressão para D_i mais fácil de ser interpretada

$$D_i = \left\{ \frac{r_i}{s(1-h_{ii})^{1/2}} \right\}^2 \frac{h_{ii}}{(1-h_{ii})} \frac{1}{p}, \tag{1.19}$$

onde o termo entre chaves é o i-ésimo resíduo studentizado. Logo, D_i será grande, quando o i-ésimo ponto for aberrante, t_i grande e/ou quando h_{ii} for próximo de um.

A medida D_i poderá não ser adequada quando t_i for grande e h_{ii} for pequeno. Nesse caso s^2 pode ficar inflacionado, e não ocorrendo nenhuma compensação por parte de h_{ii}, D_i pode ficar pequeno. Uma medida mais apropriada foi proposta por Belsley et al. (1980) dada por

$$\text{DFFITS}_i = \frac{r_i}{s_{(i)}(1-h_{ii})^{1/2}} \left\{ \frac{h_{ii}}{(1-h_{ii})} \right\}^{1/2}.$$

Nota-se que essa medida é obtida diretamente de (1.19), já que $s^2_{(i)} = s^2(n-p-r_i^2)/(n-p-1)$.

Geralmente examina-se essas medidas graficamente, dando-se atenção àquelas observações com os maiores valores para D_i e DFFITS_i. Há, entretanto, sugestões de pontos críticos para essas duas medidas. Sugere-se, no caso do $DFFITS_i$, examinar as conseqüências da retirada dos pontos com $DFFITS_i \geq 2\{p/(n-p)\}^{1/2}$, e no caso do D de Cook, dos pontos com $D_i \geq F_{p,(n-p)}(\alpha)$.

1.8.4 - Técnicas Gráficas

Os problemas mencionados em (i), (ii) e (iii) no início desta seção, de uma forma geral podem ser detectados, respectivamente, através das seguintes técnicas gráficas:

(i) gráficos de probabilidades dos resíduos ordenados $t^*_{(i)}$ contra $\Phi^{-1}(\frac{i-3/8}{n+1/4})$, onde $\Phi(\cdot)$ é a função acumulativa da normal padrão. Nesses gráficos, se os pontos ficarem praticamente sobre uma reta, pode-se aceitar que os resíduos têm distribuição aproximadamente normal. Entretanto, como em geral os resíduos são correlacionados, Atkinson (1981) sugere para diminuir essa dependência que os gráficos de probabilidades sejam construídos com intervalos de confiança obtidos através de simulações com o modelo ajustado. A falta de um termo extra na componente sistemática η e/ou a necessidade de alguma transformação nas covariáveis pode ser detectada examinando os gráficos de t_i ou t^*_i contra $\hat{\mu}_i$, $i = 1, \ldots, n$, ou através do gráfico da variável adicionada (Seção 2.7), ou do gráfico de resíduos parciais (Cook e Weisberg, 1982, Capítulo 2);

(ii) gráficos de probabilidades descritos em (i) e gráficos dos resíduos t^*_i's contra a ordem das observações. Nesses últimos deve-se examinar com atenção aqueles pontos com $|t^*_i| > 2.0$;

(iii) gráficos de D_i, h_{ii} e $DFFITS_i$ contra a ordem das observações. Deve-se dar atenção às observações com $D_i \geq F_{p,(n-p)}(\alpha)$, e/ou $h_{ii} \geq 2p/n$ e/ou $DFFITS_i \geq 2\{p/(n-p)\}^{1/2}$.

Problemas decorrentes de (iv) relativos a multicolinearidade serão tratados na Seção 1.11.3.

§1.9 Predição da Regressão

Nesta seção será discutida a utilização do modelo ajustado de regressão para se fazer previsões. Embora a estimativa para o valor esperado de y, dado um conjunto de valores das variáveis explicativas, seja simplesmente obtida

substituindo esses valores na equação ajustada, a construção de intervalos de confiança não é tão simples assim.

1.9.1 - Intervalos de Confiança para a Média de y

Seja $\mu(x^{(0)})$ o valor esperado para a resposta $y(x^{(0)})$, onde $x^{(0)}$ é um vetor $p \times 1$ definido por $x^{(0)} = (1, x_1^{(0)}, \ldots, x_{p-1}^{(0)})^T$. De (1.3) a previsão para esse vetor é dada por

$$\hat{\mu}(x^{(0)}) = x^{(0)T}(X^TX)^{-1}X^Ty,$$

com variância na forma

$$Var\{\hat{\mu}(x^{(0)})\} = x^{(0)T}(X^TX)^{-1}x^{(0)}\sigma^2. \tag{1.20}$$

Logo, um intervalo de $100(1-\alpha)\%$ de confiança para $\mu(x^{(0)})$ será formado pelos limites

$$\hat{\mu}(x^{(0)}) \pm t_{\alpha/2}\{x^{(0)T}(X^TX)^{-1}x^{(0)}\}s. \tag{1.21}$$

É também usual construir regiões de confiança para a verdadeira média $\mu(x)$, onde x representa um vetor $p \times 1$ de valores arbitrários das variáveis explicativas. Nesse caso, dada a matriz X, o objetivo é construir um conjunto (de intervalos) de confiabilidade $100(1-\alpha)\%$ para $\mu(x)$, variando-se x. Esse conjunto é formado pelos limites

$$x^T(X^TX)^{-1}X^Ty \pm \{pF_{p,(n-p)}(\alpha)\}^{1/2}\{x^T(X^TX)^{-1}x\}^{1/2}s,$$

para todo x. Se x é especificado, por exemplo se $x = x^{(0)}$, então os limites acima coincidem com (1.21).

Para ilustrar considere o modelo normal linear simples

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i,$$

onde $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, n$. Substituindo a expressão para $(X^T X)^{-1}$ dada na Seção 1.7 em (1.20), obtém-se

$$Var\{\hat{\mu}(x^{(0)})\} = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_1^{(0)} - \overline{x}_i)^2}{\sum (x_{i1} - \overline{x}_1)^2} \right\},$$

onde $x^{(0)} = (1, x_1^{(0)})^T$. De (1.6) e (1.7), os valores ajustados $\hat{\mu}_i$ são obtidos de

$$\hat{\mu}_i = \overline{y} + \hat{\beta}_1 (x_{i1} - \overline{x}_1), \quad i = 1, \ldots, n.$$

Logo, um intervalo de $100(1-\alpha)\%$, $i = 1, \ldots, n$ de confiança para $\mu(x^{(0)})$ será formado pelos limites

$$\overline{y} + \hat{\beta}_1 (x_1^{(0)} - \overline{x}_1) \pm t_{\alpha/2} \left\{ \frac{1}{n} + \frac{(x_1^{(0)} - \overline{x}_1)}{\sum (x_{i1} - \overline{x}_1)^2} \right\}^{1/2} s.$$

1.9.2 - Intervalos de Confiança para um Conjunto de observações

Considere agora a situação de se fazer previsões para m observações $y_{f_1}, \ldots, y_{f_m}$, obtidas independentemente das n que estimam β, do que para as médias das mesmas dado um conjunto de valores das variáveis explicativas. Nesse caso, como cada observação y_{f_i} varia em torno de sua média com variabilidade σ^2, a variância dada em (1.20) deve ser modificada. Denotando $y_f = (y_{f_1}, \ldots, y_{f_m})^T$ e por $\hat{y}$ a previsão de y, tem-se

$$Var(\hat{y}_f) = Var(\hat{\mu}) + Var(\varepsilon_f),$$

pois o erro ε_f associado a y_f independe de y, onde Var$(\hat{\mu})$ é a variância da estimativa do valor esperado de y. Logo, se foram observados os valores $x_\ell^{(0)}$, $\ell = 1, \ldots, m$, então a variância de $\hat{y}_f$ fica expressa na forma

$$Var(\hat{y}_f) = \sigma^2 \{1 + X^{(0)T} (X^T X)^{-1} X^{(0)}\},$$

onde $X^{(0)}$ é uma matriz $m \times p$ de linhas $x_\ell^{(0)}$.

Regiões de $100(1-\alpha)\%$ de confiança para y_f são geralmente obtidas através de dois métodos, que fornecem os intervalos correspondentes para cada y_{f_i}, $i = 1, \ldots, m$. O método de Scheffé fornece os seguintes limites:

$$\hat{\mu}(x_\ell^{(0)}) \pm \{mF_{1,(n-p)}(\alpha)\}^{1/2}\{1 + x_\ell^{(0)T}(X^TX)^{-1}x_\ell^{(0)}\}^{1/2}s,$$

$\ell = 1, \ldots, m$. O outro método, que utiliza a desigualdade de Bonferroni, fornece os limites

$$\hat{\mu}(x_\ell^{(0)}) \pm t_{\alpha/2m}(1 + x_\ell^{(0)T}(X^TX)^{-1}x_\ell^{(0)})^{1/2}s,$$

$\ell = 1, \ldots, m$, onde $t_{\alpha/2m}$ é o quantil $(1-\alpha/2m)$ de uma t de Student com $(n-p)$ graus de liberdade. Ambos os métodos são conservadores e ambos os intervalos crescem em comprimento com m. O método de Bonferroni, entretanto, geralmente produz intervalos de comprimentos menores.

Em particular para uma única observação ($m = 1$), os limites para ambos os métodos são dados por

$$\hat{\mu}(x^{(0)}) \pm t_{\alpha/2}(1 + x^{(0)T}(X^TX)^{-1}x^{(0)})s. \tag{1.22}$$

Restringindo (1.22) ao caso normal linear simples, tem-se os limites

$$\overline{y} + \hat{\beta}_1(x_1^{(0)} - \overline{x}_1) \pm t_{\alpha/2}\left\{1 + \frac{1}{n} + \frac{(x_1^{(0)} - \overline{x}_1)^2}{\sum(x_{i1} - \overline{x}_1)^2}\right\}s.$$

§1.10 Mínimos Quadrados Ponderados

Em alguns problemas de regressão linear pode ocorrer que algumas das observações sejam menos precisas que as demais. Isso significa que as

variâncias das observações não serão iguais, ou seja, a matriz $Var(y)$ não será do tipo $\sigma^2 I$. Agora os elementos da diagonal principal dessa matriz devem variar.

Suponha então o seguinte modelo:

$$z = X\beta + \varepsilon', \tag{1.23}$$

onde $z = (z_1, \ldots, z_n)^T$, $\varepsilon' \sim N(0, \sigma^2 V)$ e $V = diag\{v_1, \ldots, v_n\}$. É possível mostrar que existe uma matriz não singular P tal que

$$P^T P = PP = V.$$

Logo, chamando $\varepsilon = P^{-1}\varepsilon'$, segue-se $E(\varepsilon) = 0$ e $\text{Var}(\varepsilon) = E(\varepsilon\varepsilon^T) = P^{-1}Var(\varepsilon')P^{-1} = \sigma^2 I$. Ainda ε tem distribuição normal multivariada. Então, pré-multiplicando (1.23) por P^{-1} tem-se o novo modelo

$$P^{-1}z = P^{-1}X\beta + P^{-1}\varepsilon'$$

ou

$$y = Q\beta + \varepsilon,$$

onde $y = P^{-1}z$ e $Q = P^{-1}X$.

A estimativa de mínimos quadrados ponderados para β será dada por

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} z. \tag{1.24}$$

Em particular tem-se

$$Var(\hat{\beta}) = \sigma^2 (X^T V^{-1} X)^{-1} \tag{1.25}$$

e

$$SQ\,\text{Res} = (z - X\hat{\beta})^T V^{-1} (z - X\hat{\beta}).$$

Para ilustrar considere o modelo normal linear simples

$$z_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i',$$

onde $\varepsilon_i' \sim N(0, 1/v_i)$, $i = 1, \ldots, n$. As estimativas de mínimos quadrados para $\hat{\beta}_0$ e $\hat{\beta}_1$ são, respectivamente, dadas por

$$\hat{\beta}_0 = \tilde{z} - \hat{\beta}_1 \tilde{x}_1$$

e

$$\hat{\beta}_1 = \sum v_i z_i (x_{i1} - \overline{x}_1) / \sum v_i (x_{i1} - \overline{x}_1)^2,$$

onde $\tilde{z} = \sum v_i z_i / \sum v_i$ e $\tilde{x} = \sum v_i x_{i1} / \sum v_i$.

Mostra-se também que

$$Var(\hat{\beta}_0) = \sum v_i x_{i1}^2 / \{n \sum v_i (x_{i1} - \tilde{x}_1)^2\}$$

e

$$Var(\hat{\beta}_1) = \{\sum v_i (x_{i1} - \tilde{x}_1)^2\}^{-1}.$$

Quando os v_i's forem desconhecidos deve-se estimá-los através de algum método que produza estimativas consistentes. Geralmente, utiliza-se o método de máxima verossimilhança. Na Seção 4.4.1 esse método será aplicado a um problema específico de regressão para respostas binárias.

§1.11 Dificuldades no Uso do Modelo Normal-linear

No modelo normal-linear introduzido na Seção 1.1 as seguintes suposições se verificam: (i) a estrutura para a média é linear em todas as variáveis

explicativas; (ii) a variância das observações é constante; (iii) os dados são normais; (iv) as observações são independentes. Estas hipóteses básicas serão reexaminadas no Capítulo 2, onde serão propostos testes para verificar violações nas suposições (i) (Seção 2.6), (ii) e (iii) (Seção 2.7). Além dessas hipóteses o modelo normal-linear admite que: (v) a matriz modelo X não está sujeita a erros; (iv) o conjunto de variáveis explicativas não é colinear, isto é, não existem constantes c_j's tais que $\sum_{j=1}^{p-1} c_j x_j \doteq c_0$, onde x_j é a j-ésima coluna de X.

1.11.1 - Não-linearidade e Variância Não-Constante

Com a suposição (i) sendo violada a simples transformação dos dados pode, freqüentemente, ser a ação corretiva. Os modelos normais não- lineares serão estudados em detalhes no Capítulo 5 e um teste para verificar a linearidade $x_i^T\beta$ de μ_i será apresentado na Seção 2.6. Muitos modelos aparentemente não-lineares poderão ser convertidos em lineares por transformações convenientes. O uso de logaritmos é eficaz para muitos modelos enquanto outros exigem expansões em série de Taylor até 1ª ordem para se obter a linearidade. Neste último caso a estimação dos parâmetros é realizada iterativamente resolvendo regressões lineares.

A suposição (ii) é, usualmente, violada quando a variância depende da média da resposta ou de algumas variáveis explicativas, ou possivelmente, de outras variáveis que não estão no modelo. O uso de uma regressão linear ponderada (Seção 1.10) com pesos escolhidos adequadamente como funções de variáveis explicativas poderá ser a ação corretiva. Alternativamente, pode-se usar um modelo de Box e Cox (Capítulo 2) ou aplicar alguma transformação conveniente que estabilize Var(y).

A transformação escolhida poderá ser: raiz quadrada quando os dados forem contagens; logaritmo se o intervalo de variação dos y's for positivo;

inversa para variáveis de resposta que apresentam uma taxa de variação; arco-seno-raiz quando os dados são proporções; e várias outras. O modelo de Box e Cox permite achar a transformação conveniente.

Um teste para verificar a constância da variância (homocedasticidade) será visto na Seção 2.7. Um método alternativo, bem simples, baseia-se na função de variância paramétrica para os erros

$$\text{Var}(\varepsilon_i) = \sigma^2\{\exp(z_i^T\lambda)\},$$

onde $z_i^T = (z_{i1}, \ldots, z_{iq})$ é um vetor $1 \times q$ de variáveis conhecidas, $i = 1, \ldots, n$, e λ é um vetor $q \times 1$ de parâmetros desconhecidos. Os z_i's podem estar ou não relacionados com as variáveis x_i's do modelo. Modelos heterocedásticos serão introduzidos na Seção 1.16.

Se os ε_i's são supostos independentes e normalmente distribuídos, o procedimento seguinte (Cook e Weisberg, 1983) poderá testar a hipótese de homocedasticidade $H: \lambda = 0$:

1. Obter os resíduos r_i's na regressão de y sobre X;
2. Computar os resíduos quadrados padronizados $u_i = r_i^2/\hat{\sigma}^2$, onde $\hat{\sigma}^2 = \sum r_i^2/n$;
3. Fazer a regressão de $u = (u_1, \ldots, u_n)^T$ sobre a matriz (1Z), formada pelas linhas $z_1^T, \ldots, z_n^T$ e 1 representando o intercepto, e obter $SQReg$, que terá q graus de liberdade. Se a variância tem dependência sobre a média de y, faz-se a regressão de u sobre $(1\ \hat{\mu})$ e obtém-se $SQReg$ que terá 1 grau de liberdade;
4. A hipótese de homocedasticidade será rejeitada quando $SQReg/2$ for superior a $\chi_q^2(\alpha)$, ponto da distribuição χ_q^2 correspondente ao nível α.

O procedimento anterior poderá ser substituído por um teste gráfico de r_i^2 versus $(1 - h_{ii})z^T\hat{\lambda}$ a partir do cálculo da estimativa $\hat{\lambda}$.

1.11.2 - Não-normalidade e Erros Correlacionados

A suposição de normalidade (iii) é necessária para justificar os testes t e F e os intervalos e regiões de confiança descritos nas Seções 1.6 e 1.7, respectivamente. Na Seção 2.7 desenvolvem-se testes para verificar esta suposição.

Em pequenas amostras é difícil diagnosticar a não-normalidade pelo exame dos resíduos. Verifica-se, facilmente, que

$$r = y - \hat{\mu} = (I - H)\varepsilon, \text{ pois } (I - H)X = 0,$$

isto é,

$$r_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j. \tag{1.26}$$

Pelo teorema do limite central a soma em (1.26) terá distribuição aproximadamente normal, mesmo que os ε_i's não sejam normais. Como esta soma pode ser mais importante que ε_i na determinação da distribuição de r_i, pelo menos para n pequeno, o teste de não- normalidade aplicado aos resíduos não será adequado em pequenas amostras. Entretanto, quando n cresce os h_{ii}'s tendem a zero e o termo ε_i dominará o lado direito de (1.26); neste caso, as técnicas de análise dos resíduos aplicadas aos r_i's deverão produzir os mesmos resultados como se fossem aplicadas aos próprios erros. O gráfico dos resíduos ordenados versus os quantis da $N(0,1)$ será uma linha reta se os resíduos forem normais.

A suposição (iv) de que o erro de uma observação não depende dos erros de outras observações não será satisfeita quando os dados forem ordenados no tempo ou no espaço ou quando dados adjacentes exercerem influência entre si. Testes para verificar (iv) são geralmente difíceis de serem propostos

e o diagnóstico deverá se restringir ao exame cuidadoso do processo gerador dos dados.

Para dados igualmente espaçados no tempo o teste de Durbin-Watson (1950) pode ser usado para verficiar se os erros estão correlacionados segundo um processo $AR(1)\varepsilon_t = \rho\varepsilon_{t-1} +$ ruído branco (Seção 1.17). Para isto calcula-se a estatística

$$D = \sum_{i=2}^{n}(r_i - r_{i-1})^2 / \sum_{i=1}^{n} r_i^2, \tag{1.27}$$

cuja distribuição exata segundo a hipótese de não-correlação ($H: \rho = 0$) depende da matriz modelo. Existem limites para o teste que não dependem de X mas o uso desses pode implicar no teste ser não- conclusivo.

1.11.3 - Erro na matriz modelo e multicolinearidade

Admite-se que a matriz modelo X tem elementos, possivelmente, medidos com erro. Seja X_c a matriz modelo correta que deveria ter sido observada. Tem- se $X = X_c + D$, onde D é uma matriz $n \times p$ de linhas d_i^T, $i = 1, \ldots, n$, representando os erros de cada caso. Considera-se que os d_i's são independentes e que $E(d_i) = 0$ e $S = Cov(d_i) = diag\{s_1^2, \ldots, s_p^2\}$.

A estimativa de β baseada no modelo incorreto $y = X\beta + \varepsilon = (X_c + D)\beta + \varepsilon$ é dada por $\hat{\beta} = (X^TX)^{-1}X^Ty$ e pode não ser uma estimativa razoável de β no modelo correto $y = X_c\beta + \varepsilon$. Pode-se demonstrar que o ajustamento da matriz X produz uma estimativa que é tendenciosa no modelo $y = X_c\beta + \varepsilon$, tendo a seguinte expressão para o vício:

$$E(\hat{\beta} - \beta) = -(n - p)(X^TX)^{-1}S\beta. \tag{1.28}$$

Infelizmente, os vícios das estimativas $\hat{\beta}_i$'s em (1.28) não desaparecem quando $n \to \infty$.

Pode-se ainda definir medidas de diagnóstico para decidir se os ajustamentos das matrizes X e X_c produzirão estimativas bem diferentes. Estas medidas são deduzidas, geralmente, de expansões em série de Taylor sobre $D = 0$ e os termos de ordem superior determinarão os efeitos dos erros de medida (Hodges e Moore, 1972; Beaton, Rubin e Barone, 1976).

Quando um conjunto de variáveis explicativas for colinear, isto é, pelo menos uma das variáveis deste conjunto puder ser expressa como aproximadamente uma combinação linear das demais variáveis, as variâncias das estimativas dos parâmetros β's terão valores maiores do que no caso dessas variáveis serem não-colineares. Como a matriz X é suposta de posto completo o significado de *colinear* deve ser de dependência linear aproximada.

Sejam $1, x_1, \ldots, x_{p-1}$ as p colunas de X supondo $x_i^T x_i = 1$ para $i = 1, \ldots, p-1$, isto é, que as colunas relativas aos β's estejam padronizadas. Pode-se demonstrar que

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2)} \frac{1}{\sum_i (x_{ij} - \overline{x}_j)^2}, \tag{1.29}$$

onde $\sum_i (x_{ij} - \overline{x}_j)^2$ é a soma dos quadrados em relação à média das componentes do vetor x_j e R_j^2 é o quadrado do coeficiente de correlação múltipla na regressão de x_j sobre as demais variáveis explicativas.

A quantidade $1/(1 - R_j^2)$ é denominada fator de aumento de variância de $\hat{\beta}_j$. Quando a variável x_j estiver envolvida em multicolinearidade, R_j^2 tenderá ao valor 1 e $Var(\hat{\beta}_j)$ crescerá. As variáveis x_j;s deverão ser escolhidas de maneira a se ter $R_j^2 \doteq 0$ para $j = 1, \ldots, p-1$. Isto ocorrerá, exatamente, quando os x_j's forem mutuamente ortogonais isto é, $x_i^T x_j = 0$ para $i \neq j$, o que representa a situação ideal na regressão.

Apesar dos coeficientes de correlação múltipla indicarem as variáveis envolvidas em multicolinearidade, eles não poderão indicar o número de

multicolinearidades da regressão. Este número pode ser deduzido da seguinte maneira. Sejam $\lambda_1, \ldots, \lambda_{p-1}$ os auto-valores da submatriz $\{x_i^T x_j\}$ de ordem $n-1$. Como a diagonal de $\{x_i^T x_j\}$ é formada por 1's de modo que $\sum_{i=1}^{p-1} \lambda_i = p-1$, o determinante desta matriz pode ser usado para medir o grau de mal-condicionamento dos dados.

Este determinante estará entre 0 e 1 sendo 1 para o caso das colunas de X serem ortogonais. A partir dos auto-valores definem-se para $r = 1, \ldots, p-1$

$$K_r = K_r[\{x_i^T x_j\}] = \frac{1}{\lambda_i} \max_j \{\lambda_j\}. \tag{1.30}$$

Claro que todos os K_r's são maiores ou iguais a 1.

O número de multicolinearidades em X é estimado pelo número dos maiores K_r's, podendo considerar, como o número dos K_r's maiores ou iguais a 100. A magnitude de K_1 escrita na forma 10^m indica que ocorrem m dígitos imprecisos na inversão da matriz modelo. Por exemplo, se $K_1 = 4185 \doteq 10^3$ e os dados foram medidos com 5 dígitos significativos, as estimativas dos β's não serão confiáveis a partir do 2º dígito. Se a ordem de grandeza de K_1 for 10^7 os dados deveriam ser medidos com mais de 7 dígitos significativos para que os erros provenientes da inversão de $X^T X$ não afetassem as estimativas da regressão no 1º dígito. Mesmo se a ordem de grandeza de K_1 for 10, isto poderá causar problemas nas estimativas, a menos que os dados sejam obtidos com um mínimo de duas casas decimais.

Como o cálculo de K_1 exige a obtenção dos auto-valores $\lambda_1, \ldots, \lambda_{p-1}$, este poderá ser estimado por

$$K_1 \doteq \{\max_i (1 - R_i^2)^{-1}\}^{1/2}. \tag{1.31}$$

Uma vez detectado as multicolinearidadades, elas poderão ser removidas

por transformações ou por eliminação de uma variável em cada um dos conjuntos de variáveis envolvidas em multicolinearidade.

Alternativamente, pode-se usar um modelo de regressão "ridge" (Seção 1.15).

§1.12 As Classificações de Um e de Dois Fatores

Nesta seção apresentam-se dois modelos de regressão especiais para análise de dados em classificações de um e de dois fatores. Esses modelos têm matrizes particionadas em $X = (X_1 X_2)$. Pode-se demonstrar que a inversa generalizada de $X^T X$ é dada por

$$\begin{bmatrix} I & -Z_1 X_1^T X_2 \\ 0 & I \end{bmatrix} \begin{bmatrix} Z_1 & 0 \\ 0 & G \end{bmatrix} \begin{bmatrix} I & 0 \\ -X_2^T X_1 Z_1 & I \end{bmatrix} \tag{1.32}$$

onde Z_1 e G são as inversas generalidades de $X_1^T X_1$ e $X_2^T [I - X_1 Z_1 X_1^T] X_2$, respectivamente, e os I's são matrizes identidades de ordens adequadas.

O modelo para a classificação de um fator considera p tratamentos correspondentes aos níveis do fator, cada tratamento i replicado n_i vezes $i = 1, \ldots, p$. Os dados y_{ij} para $i = 1, \ldots, p$ e $j = 1, \ldots, n_i$ são supostos normais e independentes com estrutura sistemática

$$\begin{aligned} \mu_{ij} &= E(y_{ij}) = \beta + \beta_i \\ \text{Var}(y_{ij}) &= \sigma^2. \end{aligned} \tag{1.33}$$

A matriz modelo correspondente a (1.33) pode ser escrita na forma $X = (X_1 X_2)$, onde $X_1 = (1, \ldots, 1)^T$ é um vetor de dimensão n_+ e X_2 é

uma matriz n_+ por p dada por

$$X_2 = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ & & \vdots & \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ & & \vdots & \\ 0 & 1 & \cdots & 0 \\ & & \vdots & \end{bmatrix}$$

Tem-se $\beta = (\beta, \beta_1, \cdots, \beta_p)^T$. Usando (1.32) obtém-se a inversa de X^TX, que possibilitará determinar $Cov(\hat{\beta})$,

$$(X^TX)^{-1} = \begin{bmatrix} 1/n_+ & 0 \\ 0 & A \end{bmatrix},$$

onde $A = diag\{n_1^{-1}, \cdots, n_p^{-1}\} - n_+^{-1}J$ e J é uma matriz $p \times p$ formada por 1's. As estimativas de $\beta, \beta_1, \cdots, \beta_p$ são obtidas usando (1.3) e definindo

$$\overline{y}_{i+} = \sum_j y_{ij}/n_i \text{ e } \overline{y}_{++} = \sum_i \sum_j y_{ij}/n_+$$

$$\hat{\beta} = \overline{y}_{++} \text{ e } \hat{\beta}_i = \overline{y}_{i+} - \overline{y}_{++} \cdot (1 = 1, \ldots, p). \tag{1.34}$$

De $SQ\,\text{Res} = y^T(I - H)y$ e da Tabela 2.1 obtém-se a Tabela 2.3 de análise de variância do modelo normal-linear para a classificação de um fator. Se o modelo $\mu_{ij} = \beta + \beta_i$ for verdadeiro a soma de quadrados residuais terá distribuição $\sigma^2\chi^2_{n_+-p}$. Segundo H:$\beta_1 = \cdots = \beta_p = 0$ a soma de quadrados devido aos tratamentos tem distribuição $\sigma^2\chi^2_{p-1}$.

Tabela 1.3: *Tabela de Análise de Variância para a Classificação de um Fator*

Efeito	Soma de Quadrados	G.l.
Tratamentos	$\sum_i n_i(y_{i+} - \overline{y}_{++})^2$	$p-1$
Residual	$\sum_i \sum_j (y_{ij} - \overline{y}_{i+})^2$	$n_+ - p$
Total	$\sum_i \sum_j (y_{ij} - \overline{y}_{++})^2$	$n_+ - 1$

O modelo normal-linear para a classificação de dois fatores, A com p níveis representado por linhas (blocos) e B com q níveis representado por colunas (tratamentos) com uma única observação por cela, admite dados normais satisfazendo

$$y_{ij} = \beta + \alpha_i + \beta_j + \varepsilon_{ij} \tag{1.35}$$

para $i = 1, \ldots, p, \quad j = 1, \ldots, q$, onde os erros ε_{ij}'s têm médias iguais a zero e variância σ^2. Seja $n = pq$. Pode-se provar que X^TX para este modelo é expressa na forma

$$X^TX = \begin{bmatrix} pq & q1_p^T & p1_q^T \\ q1_p & qI_p & J_{p\times q} \\ p1_q & J_{q\times p} & pI_q \end{bmatrix},$$

onde os I's são matrizes identidades e os J's são matrizes formadas por 1's com as dimensões indicadas.

A inversa generalizada de X^TX vem de (1.32)

$$(X^TX)^- = \begin{bmatrix} \frac{1}{pq} & 0 & 0 \\ 0 & \frac{1}{q}(I_p - \frac{1}{p}J_p) & 0 \\ 0 & 0 & \frac{1}{p}(I_q - \frac{1}{q}J_q) \end{bmatrix}. \tag{1.36}$$

Obtém-se $Cov(\hat{\beta}) = \sigma^2(X^TX)^-$ e de (1.3)

$$
\begin{aligned}
\hat{\beta} &= \overline{y}_{++} \\
\hat{\alpha}_i &= \overline{y}_{i+} - \overline{y}_{++} \\
\hat{\beta}_j &= \overline{y}_{+j} - \overline{y}_{++},
\end{aligned} \tag{1.37}
$$

com a notação $\overline{y}_{i+} = \sum_j y_{ij}/q$, $\overline{y}_{+j} = \sum_i y_{ij}/p$ e $\overline{y}_{++} = \sum_i \sum_j y_{ij}/pq$.

Usando a fórmula $SQReg = \hat{\beta}^T X^T y$ deduz-se a Tabela 1.4 que fornece as somas de quadrados explicadas pelas regressões de 4 modelos.

Destas somas se obtém as somas de quadrados devidas aos efeitos dos fatores A e B. Por exemplo, a soma de quadrados devida ao fator B iguala $SQReg(4) - SQReg(2) = SQReg(3) - SQReg(1) = p\sum_j(\overline{y}_{+j} - \overline{y}_{++})^2$. Logo, os efeitos do fator B independem do fator A está incluído ou não no modelo. Neste caso, os fatores são ditos *ortogonais*, fato que ocorre nos chamados experimentos balanceados. Da Tabela 1.4 obtém-se, facilmente, a Tabela 1.5 de análise de variância do modelo (1.35).

Tabela 1.4: *SQReg para vários modelos*

	Modelo	*SQReg*
1	$\mu_{ij} = \beta$	$pq\overline{y}^2_{++}$
2	$\mu_{ij} = \beta + \beta_j$	$pq\overline{y}^2_{++} + q\sum_i(\overline{y}_{i+} - \overline{y}_{++})^2$
3	$\mu_{ij} = \beta + \alpha_i$	$pq\overline{y}^2_{++} + p\sum_j(\overline{y}_{+j} - \overline{y}_{++})^2$
4	$\mu_{ij} = \beta + \alpha_i + \beta_j$	$pqy^2_{++} + p\sum_j(\overline{y}_{+j} - \overline{y}_{++})^2 + q\sum_i(\overline{y}_{i+} - \overline{y}_{++})^2$

Tabela 1.5: *Tabela da Análise de Variância para a classficação de Dois Fatores*

Efeito	Soma de Quadrados	G.l.
Fator A	$q\sum_i(\overline{y}_{i+} - \overline{y}_{++})^2$	$p-1$
Fator B	$p\sum_j(\overline{y}_{+j} - \overline{y}_{++})^2$	$q-1$
Residual	$\sum_i\sum_j(y_{ij} - \overline{y}_{i+} - \overline{y}_{+j} + \overline{y}_{++})^2$	$(p-1)(q-1)$
Total	$\sum_i\sum_j(y_{ij} - \overline{y}_{++})^2$	$pq-1$

§1.13 Métodos de Seleção de Modelos

O leitor poderá encontrar um estudo detalhado dos métodos de seleção de modelos de regressão nos livros de Seber (1977), Weisberg (1985) e Wetherill et al. (1986). Aqui apresenta-se apenas algumas idéias sobre seleção de variáveis explicativas baseadas em procedimentos computacionais, como a regressão "stepwise", e em critérios definidos pelas estatísticas SQ Res, C_p de Mallows (1973), $PRESS$ e regra de Tukey. Existem vários outros procedimentos e critérios nos 3 livros aqui citados.

Quando p for pequeno todas as $\binom{n}{p}$ regressões poderão ser efetuadas e existem algoritmos para calculá-las eficientemente. Procedimentos de troca são usados quando o número de regressores é fixado a priori, as trocas sendo realizadas a partir de um conjunto inicial visando a minizar SQ Res. O conjunto final pode depender do inicial.

Considere um conjunto A_r de r variáveis explicativas e define-se S_r como a SQ Res na regressão de y sobre A_r. Sejam S_{r+j} e S_{r-j} as somas

de quadrados dos resíduos nas regressões de y sobre $A_r \cup A_j$ e $A_r - A_j$, respectivamente, no 1º caso A_j e A_r disjuntos e no 2º caso $A_j \subset A_r$. Nos procedimentos "stepwise" A_i será acrescido à A_r se satisfizer

$$max_j\{(S_r - S_{r+j})(n-p-j)/S_{r+j}\} > \underline{F}$$

e será retirado de A_r quando

$$min_j\{(S_{r-j} - S_r)(n-p)/S_r\} < \overline{F},$$

onde $\underline{F}$ e $\overline{F}$ são pontos críticos da distribuição F com 1, $(n-p-i)$ e 1, $(n-p)$ graus de liberdade, respectivamente.

Em geral, as variáveis são introduzidas e retiradas uma a uma satisfazendo as desigualdades anteriores.

O procedimento "stepwise" pode ser aplicado do modelo nulo (uma única média) para frente ou do modelo completo (com todas as variáveis possíveis) para trás. Diversos pacotes estatísticos, entre os quais SPSS, MINITAB, SYSTAT, BMDP e SAS, têm a regressão "stepwise".

As estatísticas R^2, C_p e $PRESS$, entre outras, são usadas como critérios de seleção de covariáveis, diferentemente dos procedimentos computacionais. O quadrado do coeficiente de correlação múltipla R^2, definido na Seção 1.3, é usado como critério de seleção de covariáveis, escolhendo estas de modo a maximizá-lo. Entretanto, como R^2 sempre cresce com a introdução de covariáveis no modelo este não penaliza a complexidade do modelo.

Uma estatística bastante simples para selecionar modelos é dada por

$$C_p = SQ\,\mathrm{Res}_p/\hat{\sigma}^2 - (n-2p), \tag{1.38}$$

sendo $SQ\,\mathrm{Res}_p$ a soma dos quadrados dos resíduos de um modelo M_p com p parâmetros e $\hat{\sigma}^2$ a estimativa não-viciada de σ^2 segundo o modelo completo. A vantagem de C_p é que depende somente dos resultados usuais da

regressão. Para o modelo completo M_m com m parâmetros $C_m = 1$. Ainda $C_p = \hat{\sigma}^{-2}(SQ\,\mathrm{Res}_p - SQ\,\mathrm{Res}_m) + p - (m-p)$ e, portanto, C_p é uma medida da distância de um modelo M_p ao modelo completo M_m.

Dois modelos podem ser comparados pelos valores de C_p; valores próximos de C_p é um indicativo da equivalência de modelos. Na Seção 7.5 apresenta-se o critério de seleção de covariáveis baseado na estatística de Akaike (1974).

Outra estatística para selecionar modelos é definida a partir dos resíduos preditivos

$$r_{(i)} = y_i - x_i^T \hat{\beta}_{(i)} = \frac{r_i}{1 - h_{ii}}, \tag{1.39}$$

onde $\hat{\beta}_{(i)}$ é a estimativa de β sem a observação y_i. Esta estatística é dada por

$$PRESS = \sum_{i=1}^{n} r_{(i)}^2, \tag{1.40}$$

e para bons modelos deverá ter valores pequenos. Como este critério é muito mais trabalhoso que o cálculo de C_p, pois envolve n regressões para cada modelo possível, pode-se inicialmente selecionar alguns modelos usando o critério C_p e depois calcular $PRESS$ apenas para estes modelos.

Encerrando esta seção apresenta se um critério muito simples de seleção das variáveis explicativas, conhecido como regra de Tukey. De $\hat{\mu} = Hy$ vem $Var(\hat{\mu}) = \sigma^2 H$ e $tr\{Var(\hat{\mu})\} = p\sigma^2$. A variância média da predição de n observações (Seção 1.9.2) é igual a $(1 + p/n)\sigma^2$. Se p for pequeno comparado com n este termo reduz-se a $n\sigma^2/(n-p)$, que pode ser consistentemente estimado por $ns^2/(n-p)$. A regra de Tukey escolhe as variáveis explicativas que minimizam esta estimativa da variância média da predição. A regra se comporta de maneira semelhante ao critério C_p.

§1.14 Modelo de Regressão Polinomial

Quando a função que relaciona a variável resposta y e o regressor x for suave mas não uma reta, o modelo de regressão linear poderá ser usado desde que transformações adequadas aplicadas a x e/ou y produzam linearidade na(s) escala(s) transformada(s). Alternativamente, pode-se adotar um modelo polinomial definido por

$$y = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \varepsilon, \tag{1.41}$$

onde os y's são independentes e normalmente distribuídos com $\mathrm{Var}(y) = \sigma^2$ constante, e p é o grau do polinômio. Se a variância não resultar constante, uma transformação que estabilize a variância deve ser necessária antes que o modelo polinomial seja ajustado.

O modelo (1.41) transforma-se no modelo normal-linear com a definição de p covariáveis $z_i = x^i$, $i = 1, \ldots, p$ e, portanto, as estimativas dos β's são em princípio calculadas via (1.3) com $X = (1\ z_1 \ldots z_p)$. Entretanto, quando p for grande, poderão existir problemas na inversão da matriz $X^T X$, pois esta conterá elementos de ordens de magnitude bem diversas.

O modelo (1.41) pode ser escrito na forma

$$y = \beta_0 + \sum_{r=1}^{p} \beta_r f_r(x) + \varepsilon, \tag{1.42}$$

onde $f_r(x) = a_{rr}x^r + \cdots + a_{r1}x + a_{r0}$ é um polinômio de grau r em x. Os f_r's são escolhidos adequadamente para satisfazer o sistema de equações

$$\sum_{i=1}^{n} f_r(x_i) = 0,\ r = 1, \ldots, p \tag{1.43}$$

e

$$\sum_{i=1}^{n} f_r(x_i) f_s(x_i) = 0 \tag{1.44}$$

para $r = 2, 3, \ldots, p$ e $s = 1, 2, \ldots, r-1$.

As equações (1.43) e (1.44) tornam as somas dos polinômios e dos produtos cruzados de polinômios, respectivamente, nulas.

Existem $(r+1)$ parâmetros em cada polinômio para satisfazer uma equação em (1.43) e $(r-1)$ equações em (1.44). O sistema terá solução única arbitrando um parâmetro em cada polinômio, geralmente, fazendo $a_{rr} = 1$, $r = 1, \ldots, p$. Os polinômios que satisfazem (1.43) e (1.44) são denominados *ortogonais* e são tabelados para n pontos espaçados em intervalos unitários e centrados em zero (vide, por exemplo, Pearson e Hartley, 1976).

A matriz $X^T X$ reduz-se a

$$X^T X = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & \sum_i f_1(x_i)^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sum_i f_p(x_i)^2 \end{bmatrix}$$

e usando (1.3) vem as estimativas de $\beta_0, \beta_1, \ldots, \beta_p$

$$\begin{aligned} \hat{\beta}_0 &= \frac{1}{n} \sum_{i-1}^{n} y_i, \\ \hat{\beta}_r &= \frac{\sum_{i=1}^{n} y_i f_r(x_i)}{\sum_{i=1} f_r^2(x_i)}, \end{aligned} \tag{1.45}$$

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

A simples inversão de X^TX produz as variâncias das estimativas

$$Var(\hat{\beta}) = \sigma^2/n$$

e

$$Var(\hat{\beta}_r) = \sigma^2 / \left\{ \sum_{i=1}^{n} f_r(x_i)^2 \right\}$$

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

A expressão de $SQReg = \hat{\beta} X^T y$ pode ser decomposta em $(p+1)$ termos estatisticamente independentes devido à ortogonalidade dos polinômios. Disto resulta a Tabela 1.6 de análise de variância para o modelo polinomial.

Tabela 1.6: *Análise de Variância para polinômios ortogonais*

Efeito	Soma de Quadrados	G.l.
Devido a β_1	$\{\sum y_i f_1(x_i)\}^2 / \{\sum f_1(x_i)^2\}$	1
$\vdots$	$\vdots$	$\vdots$
Devido a β_p	$\{\sum y_i f_p(x_i)\}^2 / \{\sum f_p(x_i)^2\}$	1
Residual	(por diferença)	$(n-p-1)$
Total	$\sum_i (y_i - \overline{y}^2)$	$(n-1)$

§1.15 Modelo de Regressão "Ridge"

O modelo de regressão "ridge" (Hoerl e Kennard, 1970) objetiva eliminar a multicolinearidade das variáveis explicativas substituindo X^TX por X^TX+

kI, onde k é uma constante positiva de valor próximo de zero. A estimativa na regressão "ridge" é obtida de

$$\beta^* = (X^TX + kI)^{-1}X^Ty \tag{1.46}$$

Sejam $\lambda_1, \ldots, \lambda_p$ os auto-valores de X^TX e $v_1, \ldots, v_p$ os auto-vetores correspondentes. Verifica-se que os auto-valores de $(X^TX+kI)^{-1}$ são $(\lambda_i+k)^{-1}$, $i = 1, \ldots, p$. Se X^TX for aproximadamente singular com mínimo auto-valor $\underline{\lambda}$, então o menor auto-valor de $(X^TX + kI)$ será $\underline{\lambda} + k$ e esta última matriz não estará tão próxima da singularidade.

Usando a decomposição de X^TX nos seus auto-vetores vem

$$Var(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \lambda_i v_i v_i^T, \tag{1.47}$$

e, portanto, os elementos em $\hat{\beta}$ que correspondem a auto-vetores não-desprezíveis deverão ter maiores variâncias e covariâncias. O erro médio quadrático (EMQ) de $\hat{\beta}$, que é não-viciado, reduz-se a

$$EMQ(\hat{\beta}) = tr\{Var(\hat{\beta})\} = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}. \tag{1.48}$$

Seja $V = (v_1, \ldots, v_p)$. De (1.46), (1.47) e (1.48) pode-se demonstrar com alguma álgebra

$$Var(\beta^*) = \sigma^2 V \Lambda_1 V^T$$

e

$$EMQ(\beta^*) = \sum_{i=1}^{p} (\lambda_i^2 + k^2\gamma_i)/(k + \lambda_i)^2,$$

onde $\Lambda_1 = diag\{\lambda_i/(k + \lambda_i)^2\}$ e $\gamma^T = (\gamma_1, \ldots, \gamma_p) = \beta^T V$.

Das expressões de $Var(\beta^*)$ e $EMQ(\beta^*)$ conclui-se que a variância de β^* é uma função decrescente de k enquanto que $E(\beta^* - \beta)$ é uma função crescente. Demonstra-se que existe um k tal que $EMQ(\beta^*) \leq EMQ(\hat{\beta})$, o que justifica o uso da regressão "ridge". Ainda $\beta^{*T}\beta^* < \hat{\beta}^T\hat{\beta}$ para todo k positivo e $\beta^{*T}\beta^*$ tende para zero quando k cresce.

O importante resultado de que β^* pode ser obtido dos auto-valores e auto-vetores de X^TX como uma função de k

$$\beta^* = \sum_{i-1}^{p} \frac{1}{\lambda_i + k} v_i^T X^T y v_i \tag{1.49}$$

permite calcular β^* para diversos valores de k. Quando $k = 0$, β^* coincide com $\hat{\beta}$. Examinando o comportamento de β^* com k crescente, pode-se escolher a estimativa de k como o valor a partir do qual as componentes do vetor β^* se estabilizam.

§1.16 Modelos Heterocedásticos

Os modelos normais heterocedásticos são aplicados em larga escala na Economia. A variância da variável resposta não é constante sendo uma função de parâmetros, provavelmente, desconhecidos.

Sejam $y_i \sim N(\mu_i, \sigma_i^2)$ $i = 1, \ldots, n$, n variáveis normais independentes cujas médias e variâncias são desconhecidas. A hipótese de independência será relaxada na Seção 1.7 e nos Capítulos 6, 7 e 8 são apresentados modelos não-normais, geralmente, heterocedásticos. Admite-se a componente linear $\eta = X\beta$ para o vetor de médias. Os modelos heterocedásticos são definidos por $(i = 1, \ldots, n)$

$$\sigma_i^2 = \sigma^2 f_i(z_i^T \delta), \tag{1.50}$$

onde os z_i^T's são vetores $1 \times q$ conhecidos e $\delta = (\delta_1, \ldots, \delta_q)^T$ é um vetor de parâmetros desconhecidos. Aqui os z_i's podem ou não estar relacionados com as colunas de X^T e os δ's e β's podem conter parâmetros em comum. A *heterocedasticidade multiplicativa* é definida por $\sigma_i^2 = \exp(z_i^T\delta)$ e a *variável-dependente* admite $\sigma_i^2 = \sigma^2(x_i^T\beta)^{2\lambda}$; em geral $\lambda = 1$. Notar que a heterocedasticidade multiplicativa envolve modelos com $\sigma_i^2 = \sigma^2 x_{ij}^{\alpha}$ bastando fazer $z_i = (1 \log x_{ij})$ e $\delta = (\log \sigma^2 \alpha)^T$.

Seja $L(\beta, \delta)$ a log-verossimilhança de um modelo heterocedástico multiplicativo supondo os dados $y = (y_1, \ldots, y_n)^T$. Pode-se escrever $L(\beta, \delta)$ em notação matricial

$$L(\beta, \delta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\Lambda^T Z\delta - \frac{1}{2}\Lambda^T\Lambda^{-1}v, \tag{1.51}$$

onde Z é uma matriz $n \times q$ formada pelas linhas z_i^T's, $\Lambda = diag\{\sigma_1^2, \ldots, \sigma_n^2\}$, $v = (v_1, \ldots, v_n)^T$ tem componentes dadas por $v_i = (y_i - \mu_i)^2$ e $1 = (1, \ldots, 1)^T$ é um vetor $n \times 1$.

Maximizando $L(\beta, \delta)$ em relação a β e δ implica na resolução do sistema de equações não-lineares

$$\begin{aligned} \hat{\beta} &= (X^T\hat{\Lambda}^{-1}X)^{-1}X^T\hat{\Lambda}^{-1}y, \\ Z\Lambda &= Z^T\hat{\Lambda}^{-1}\hat{v}. \end{aligned} \tag{1.52}$$

A solução do sistema (1.52) deve ser feita por procedimentos iterativos do tipo Newton-Raphson (escore para parâmetros, Gauss-Newton e outros). Esses procedimentos são baseados na matriz de derivadas segundas da log-verossimilhança ou em outras matrizes equivalentes à mesma. Aqui o método de escore de Fisher, que usa a matriz de informação, produz um esquema iterativo bastante simples.

A matriz de informação é bloco-diagonal sendo expressa por

$$K(\beta,\delta) = \begin{array}{c} \\ \beta \\ \delta \end{array} \begin{array}{c} \begin{array}{cc} \beta & \delta \end{array} \\ \begin{bmatrix} X^T\Lambda^{-1}X & 0 \\ 0 & \frac{1}{2}Z^TZ \end{bmatrix} \end{array}, \tag{1.53}$$

o que implica na ortogonalidade dos parâmetros β e δ. Esta ortogonalidade é a causa da simplicidade do método escore resultando no processo iterativo

$$\beta^{(m+1)} = \beta^{(m)} + (X^T\Lambda^{(m)^{-1}}X)^{-1}\Lambda^{(m)^{-1}}(y - \mu^{(m)})$$

e

$$\delta^{(m+1)} = \delta^{(m)} + (Z^TZ)^{-1}Z^T(\Lambda^{(m)^{-1}}v^{(m)} - 1),$$

onde m indica o m-ésimo passo do processo. A inicialização poderá ser feita das estimativas da regressão ordinária com $\beta^{(1)} = \hat{\beta}$, $\delta^{(1)} = (log\hat{\sigma}^2 0 \ldots 0)^T$ e $\Lambda^{(1)} = \hat{\sigma}^2 I$ se a 1ª coluna de Z for formada por 1's.

Segundo as condições usuais de regularidade, $(\hat{\beta}^T\hat{\delta}^T)^T$ têm distribuição normal assintótica com matriz de covariância $K(\beta,\delta)^{-1}$. Como $K(\beta,\delta)$ é bloco-diagonal as estimativas $\hat{\beta}$ e $\hat{\delta}$ têm distribuições independentes $N_p(\beta,(X^T\Lambda^{-1}X)^{-1})$ e $N_q(\delta,2(Z^TZ)^{-1})$, respectivamente. Testes e regiões de confiança podem ser baseados nessas distribuições assintóticas.

Estuda-se agora o modelo de heterocedasticidade variável-dependente definido por $\sigma_i^2 = \sigma^2(x_i^T\beta)^2$. Seja $L = L(\beta,\sigma^2)$ a log- verossimilhança total supondo os dados y. Pode-se verificar que a função suporte $U = U(\beta,\sigma^2) = \binom{\partial L/\partial\beta}{\partial L/\partial\sigma^2}$ tem componentes

$$\partial L/\partial\beta = \sigma^{-2}X^TV^{-1}\{z + y - (1+\sigma^2)\mu\}$$

e

$$\partial L/\partial\sigma^2 = \frac{1}{2}\sigma^{-4}1^T V^{-1}(w - \sigma^2 V1),$$

onde $z = (z_1, \ldots, z_n)^T$ com $z_i = (y_i - \mu_i)^2/\mu_i$, $w = (w_1, \ldots, w_n)^T$ com $w_i = (y_i - \mu_i)^2$, $V = diag\{\mu_1^2, \ldots, \mu_n^2\}$ e 1 é um vetor $n \times 1$ de 1's.

A matriz de informação pode ser deduzida de $E\{UU^T\}$ ou mais facilmente das segundas derivadas de L. Ela não é bloco-diagonal sendo expressa por

$$K(\beta, \sigma^2) = \begin{matrix} \\ \beta \\ \sigma^2 \end{matrix} \begin{matrix} \begin{matrix} \beta & \sigma^2 \end{matrix} \\ \begin{bmatrix} (2+\sigma^{-2})X^T V^{-1} X & \mu^T V^{-1} X \\ X^T V^1 \mu & \frac{n}{2\sigma^4} \end{bmatrix} \end{matrix}, \tag{1.54}$$

e, portanto as estimativas de β e σ^2 não são independentemente distribuídas em grandes amostras.

As estimativas $\hat{\beta}$ e $\hat{\sigma}^2$ são calculadas através do processo iterativo de escore de Fisher para parâmetros

$$\begin{bmatrix} \hat{\beta}^{(m+1)} \\ \hat{\sigma}^{(m+1)^2} \end{bmatrix} = \begin{bmatrix} \hat{\beta}^{(m)} \\ \hat{\sigma}^{(m)^2} \end{bmatrix} + K^{(m)^{-1}} U^{(m)}. \tag{1.55}$$

Testes e regiões de confiança para as componentes β e σ^2 podem ser deduzidos das estatísticas escore, Wald e razão de máxima verossimilhança, conforme será descrito nas Seções 6.6 e 6.9. Consultar também Capítulos 4, 5 e 6 de Cox e Hinkley (1979).

No lugar de usar o processo iterativo (1.55) as equações não-lineares para o cálculo de $\hat{\beta}$ e $\hat{\sigma}^2$ podem ser resolvidas por mínimos quadrados ponderados (Seção 1.10), com as seguintes etapas:

1. Inicializar com a estimativa da regressão ordinária de y sobre X: $\beta^{(1)} = (X^TX)^{-1}X^Ty$ e $m = 1$;
2. Estimar $V^{(m+1)}$ em $\hat{\beta}^{(m)}$ e usar (1.24) para obter $\hat{\beta}^{(m+1)}$ e $\hat{\sigma}^{(m+1)^2} = SQ\,\mathrm{Res}/(n-p)$, $m = m+1$;
3. Repetir 2 até as estimativas $\hat{\beta}^{(m)}$ e $\hat{\sigma}^{(m)^2}$ convergirem.

Infelizmente, as estimativas obtidas com este último processo podem diferir bastante das estimativas de máxima verossimilhança provenientes de (1.55).

§1.17 Modelos de Regressão com Estrutura de Correlação

Os modelos normais de regressão com estrutura de correlação têm a forma

$$y = X\beta + \varepsilon, \qquad E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2 V, \tag{1.56}$$

onde p é $n \times 1$, X é $n \times p$ e ε é o vetor de erros aleatórios. Esses modelos são freqüentemente usados na análise de dados temporais e em muitas aplicações econométricas. Se a matriz V for totalmente desconhecida ela envolverá muitos parâmetros, em número igual a $n(n-1)/2$, e alguns tipos de restrições serão necessárias para se ter uma solução.

Modelos do tipo (1.56) são, geralmente, necessários quando variáveis explicativas importantes estão omitidas, quer por desconhecimento quer por causa de dificuldades de medição.

Quando a matriz V for totalmente desconhecida e nenhuma estrutura especial puder ser postulada, a estimação conjunta de β e V só poderá ser

feita supondo que o modelo (1.56) evolua em vários instantes de tempo $1, 2, \ldots, T$. Em cada instante t observa-se o vetor de resposta $n \times 1$ y_t e a matriz X_t, os parâmetros permanecendo os mesmos em todos os instantes. O modelo resultante é denominado *dinâmico* e a sua log- verossimilhança pode ser expressa, supondo $\sigma^2 = 1$ sem perda de generalidade, por

$$(1.57)\quad L(\beta, V) = -\frac{Tn}{2}\log 2\pi - \frac{T}{2}\log|V| - \frac{1}{2}\sum_{t=1}^{T}(y_t - X_t\beta)^T V^{-1}(y_t - X_t\beta).$$

De (1.57) obtém-se as estimativas de máxima verossimilhança que verificam

$$(1.58)\qquad \hat{\beta} = (\sum_{t=1}^{T} X_t^T \hat{V}^{-1} X_t)^{-1}(\sum_{t=1}^{T} X_t^T \hat{V}^{-1} y_t)$$

$$(1.59)\qquad \hat{V} = \frac{1}{T}\sum_{t=1}^{T}(y_t - X_t\hat{\beta})(y_t - X_t\hat{\beta})^T.$$

Esse sistema de equações não-lineares poderá ser resolvido em dois estágios partindo das estimativas em (1.58) com $\hat{V} = I$.

Na análise de dados temporais é mais conveniente estabelecer uma estrutura para a matriz V do que supor a mesma totalmente desconhecida. Os modelos $ARMA$ (vide Box e Jenkins, 1976; Morettin e Toloi, 1981) têm uma estrutura especial para V. Um modelo $ARMA(p, q)$ é definido pela equação de regressão

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

para $t = 1, 2, \ldots, n$, onde os ε_t's são variáveis independentes normais de média zero e variância σ^2. Denomina-se a soma envolvendo os parâmetros

ϕ's de componente auto-regressiva e aquela envolvendo os parâmetros θ's de componente de médias móveis. As condições de estacionaridade, invertibilidade e demais propriedades, o cálculo das funções de autocorrelação e autocorrelação parcial, podem ser vistos nos dois livros aqui citados.

A log-verossimilhança para o modelo $ARMA(p,q)$ como função dos parâmetros $\beta = (\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)^T$ e σ^2, observada a série $y = (y_1,\ldots,y_n)^T$, é dada por

$$(1.60)\quad L = L(\beta,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|V| - \frac{1}{2\sigma^2}y^TV^{-1}y,$$

onde a estrutura de covariância da série é $\mathrm{Cov}(y) = \sigma^2V$.

A maximização de (1.60) em relação a β e σ^2 é complicada e muitos métodos têm sido propostos modificando ou aproximando a log-verossimilhança. Box e Jenkins (1976) sugerem desprezar $\log|V|$ na maximização de (1.60).

Apresenta-se aqui o método de Fisher (Equações (1.55)) para estimar β e σ^2. O método poderá ser aplicado numa vez que a matriz de informação para esses parâmetros seja determinada, pelo menos numericamente. Tem-se

$$\frac{\partial L}{\partial\beta_r} = -\frac{1}{2}tr(V^{-1}V_r) + \frac{1}{2\sigma^2}y^TV^{-1}V_rV^{-1}y$$

e

$$\frac{\partial L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}y^T\frac{1}{\sigma^2}IV^{-1}y,$$

onde β_r pode representar um parâmetro ϕ ou θ e $V_r = \partial V/\partial\beta_r$. Definem-se: $\tilde{V}_r = V^{-1}V_r$ para os parâmetros em β, isto é, $r = 1,\ldots,p+q$ e $\tilde{V}_{p+q+1} = \frac{1}{\sigma^2}I$; $\delta_r = tr(\tilde{V}_r)$ e $z_r^T = y^T\tilde{V}_r$ para $r = 1,\ldots,p+q+1$; ainda

$$Z = \begin{bmatrix} z_1^T \\ \vdots \\ z_{p+q+1}^T \end{bmatrix} \quad\text{e}\quad \delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_{p+q+1} \end{bmatrix}.$$

A função suporte $U = U(\beta, \sigma^2) = \begin{pmatrix} \partial L/\partial\beta \\ \partial L/\partial\sigma^2 \end{pmatrix}$ pode ser escrita como

$$U = -\frac{1}{2}\delta + \frac{1}{2\sigma^2} Z V^{-1} y. \tag{1.61}$$

A matriz de informação particionada em β e σ^2, deduzida das segundas derivadas da log-verossimilhança é expressa por

$$K = K(\beta, \sigma^2) = \begin{array}{c} \\ \beta \\ \sigma^2 \end{array} \begin{array}{c} \begin{array}{cc} \beta & \sigma^2 \end{array} \\ \begin{bmatrix} \{\frac{1}{2} tr(V^{-1} V_{ij})\} & \{\frac{1}{2\sigma^2} tr(\tilde{V}_i)\} \\ \{\frac{1}{2\sigma^2} tr(\tilde{V}_i)\} & \frac{n}{2\sigma^4} \end{bmatrix} \end{array}, \tag{1.62}$$

onde $V_{ij} = \partial^2 V/\partial\beta_i \partial\beta_j$. Substituindo (1.61) e (1.62) em (1.55) chega-se ao processo iterativo para o cálculo de $\phi_1, \ldots, \phi_p$, $\theta_1, \ldots, \theta_q$ e σ^2. Para qualquer modelo $ARMA$ é fácil determinar numericamente K por programas especiais e, portanto, realizar o procedimento (1.55).

Recentemente, em pesquisa ainda não publicada, Cordeiro e Klein (1989) propõem uma fórmula geral, correta até termos de ordem n^{-1}, para $E(\hat{\beta})$ e $E(\hat{\sigma}^2)$. Para modelos simples $AR(1)$, $AR(2)$, $MA(1)$, $MA(2)$ e $ARMA(1,1)$ é possível obter formas fechadas para esses valores esperados. Entretanto, isso será bastante difícil quando p e q forem grandes, face à dificuldade de achar V^{-1}. Agora, o cálculo numérico do viés de $\hat{\beta}$ e $\hat{\sigma}^2$ para qualquer modelo $ARMA$ não apresentará problemas usando o programa REDUCE.

Para ilustrar a obtenção da matriz de informação considere o modelo $AR(1)$ $y_t = \phi y_{t-1} + \varepsilon_t$. Tem-se $Var(y_t) = \sigma^2/(1-\phi^2)$, $\mathrm{Cov}\{y_{t+k}, y_t\} =$

$\sigma^2\phi^{k-1}/(1-\phi^2)$, $1, 2, \ldots, n$, e

$$V^{-1} = \begin{bmatrix} 1 & -\phi & 0 & \cdots & 0 & 0 & 0 \\ -\phi & (1+\phi^2) & -\phi & \cdots & 0 & 0 & 0 \\ & & & \vdots & & & \\ 0 & 0 & 0 & \cdots & -\phi & (1+\phi^2) & -\phi \\ 0 & 0 & 0 & \cdots & 0 & -\phi & 1 \end{bmatrix}.$$

Sendo $\tilde{V}_\phi = V^{-1}\frac{\partial V}{\partial \phi}$ é fácil verificar que

$$\tilde{V}_\phi = \frac{1}{1-\phi^2}\begin{bmatrix} \phi & 1 & \phi & \phi^2 & \cdots & \phi^{n-2} \\ (1-\phi^2) & 0 & (1-\phi^2) & \phi(1-\phi^2) & \cdots & \phi^{n-3}(1-\phi^2) \\ \phi(1-\phi^2) & 1-\phi^2 & 0 & (1-\phi^2) & \cdots & \phi^{n-4}(1-\phi^2) \\ & & & \vdots & & \\ \phi^{n-2} & \phi^{n-3} & & & \cdots & \phi \end{bmatrix}.$$

Após alguma álgebra obtém-se de (1.62) a matriz de informação para o $AR(1)$

$$K = \frac{1}{2}\begin{bmatrix} \frac{2n(1-\phi^2)+6\phi^2-2}{(1-\phi^2)^2} & \frac{2\phi}{\sigma^2(1-\phi^2)} \\ \frac{2\phi}{\sigma^2(1-\phi^2)} & \frac{n}{\sigma^4} \end{bmatrix}, \tag{1.63}$$

que poderá ser usada em (1.55) para calcular as estimativas $\hat{\beta}$ e $\hat{\sigma}^2$ uma vez avaliada a função suporte.

Pode-se usar outros métodos para estimar os parâmetros do modelo normal com estrutura de correlação dada por (1.56). Uma possibilidade é usar mínimos quadrados ponderados (Seção 1.10) iterativamente, isto é, estimando V no ponto corrente $\beta^{(m)}$ para calcular $\beta^{(m+1)}$ via (1.24). A inicialização poderá ser através da regressão ordinária.

§1.18 Exemplos

Nesta última seção analisam-se dados do consumo anual de borracha nos EUA e de vendas trimestrais de uma empresa através do modelo clássico de regressão.

1.18.1 - Consumo Anual de Borracha nos E.U.A.

Na Tabela 1.7 é apresentado um conjunto de dados do livro de Draper e Smith (1981), referente ao consumo anual de borracha (y), ao produto nacional bruto (x_1) e à renda per capita (x_2) de 1948 a 1963 nos E.U.A.. Será ilustrado, através desses dados, o cálculo das somas de quadrados descritas na Tabela 1.2 e de algumas medidas de diagnóstico apresentadas na Seção 1.8.

Inicialmente ajustou-se o modelo irrestrito $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ e $\varepsilon \sim N(0, \sigma^2)$, a esses dados, e posteriormente, ajustou-se esse modelo restrito à $C\beta = 0$, onde $C = (0\ 0\ 1)^T$ e $\beta = (\beta_0\ \beta_1\ \beta_2)^T$. As estimativas dos parâmetros com os correspondentes desvios padrões são exibidos na Tabela 1.8, enquanto a tabela ANOVA para ambos os modelos é apresentada na Tabela 1.9.

Logo, o valor da estatística F referente à hipótese $H: \beta_1 = \beta_2 = 0$ é dado por

$$F = \frac{SQReg/2}{SQ\,\mathrm{Res}/13} = \frac{0.0740}{0.0032} = 13.14$$

Tabela 1.7: *Distribuição do consumo de borracha, do produto nacional bruto e da renda per capita nos E.U.A. durante 16 anos.*

ano	y	x_1	x_2	ano	y	x_1	x_2
48	0.909	0.984	0.987	56	1.001	0.996	1.003
49	1.252	1.078	1.064	57	0.916	0.972	0.993
50	0.947	1.061	1.007	58	1.173	1.046	1.027
51	1.022	1.013	1.012	59	0.938	1.004	1.001
52	1.044	1.028	1.029	60	0.965	1.004	1.014
53	0.905	0.969	0.993	61	1.106	1.049	1.032
54	1.219	1.057	1.047	62	1.011	1.023	1.020
55	0.923	1.001	1.024	63	1.080	1.035	1.053

Fonte: Draper e Smith (1981).

que é significativo a 1% ($F_{2.13}(0.01) = 6.70$). Para a hipótese $H: C\beta = 0$ a estatística F vale

$$F = \frac{ASQ(C\beta = 0)/1}{SQ\,\text{Res}/13} = \frac{0.0230}{0.0032} = 7.19,$$

que é significativo a 5%. Finalmente, o valor de F referente à hipótese $H: \beta_1 = 0$, dado que $C\beta = 0$, é igual a

$$F = \frac{SQ\,\text{Reg}(C\beta = 0)/1}{SQ\,\text{Res}(C\beta = 0)/14} = \frac{0.1260}{0.0045} = 27.61,$$

que também é significativo a 1%.

A Figura 1.1 exibe o gráfico dos resíduos ordinários studentizados t_i's contra os valores ajustados, impondo-se o modelo irrestrito. Destaca-se como aberrante o resíduo correspondente ao ano de 50. Pelo gráfico das distâncias de Cook contra a ordem das observações, Figura 1.2, destaca-se novamente o ano de 50. O valor de h_{ii} para esse ponto, Figura 1.3, é o único acima de $(2 \times 3)/16 = 0.375$.

Tabela 1.8: *Estimativas dos parâmetros referentes ao ajuste dos modelos* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ *e restrito à* $C\beta = 0$, *aos dados da Tabela 1.7.*

	Modelo Irrestrito		Modelo Restrito	
Efeito	Parâmetros	Estimativas	Parâmetros	Estimativas
Constante	β_0	-3.253	β_0	-1.832
		(0.677)		(0.433)
Produto Nacional	β_1	2.872	β_1	2.802
Bruto		(1.090)		(0.533)
Renda Per Capita	β_2	1.227		
		(0.746)		
	s^2 =	0.0032	s^2 =	0.0045
	R^2 =	0.781	R^2 =	0.664

( ): Desvio Padrão.

Tabela 1.9: *Análise da Variância dos modelos irrestrito e restrito.*

Efeito	Soma de Quadrados	G.1.	Quadrado Médio
Regresão ($\beta_1.\beta_2\beta_0$)	0.148	2	0.0740
Regressão ($\beta_2 = 0/\beta_0$)	0.126	1	0.1250
Acréscimo devido a $\beta_2 = 0$	0.023	1	0.0230
Resíduos ($\beta_2 = 0$)	0.064	14	0.0045
Resíduos	0.041	13	0.0032
Total	0.041	15	–

Observando-se com cuidado os dados da Tabela 1.7, nota-se que o ano de 50 teve um baixo consumo de borracha e um alto produto nacional bruto, que contradiz a correlação linear positiva entre as duas variáveis. Logo, o

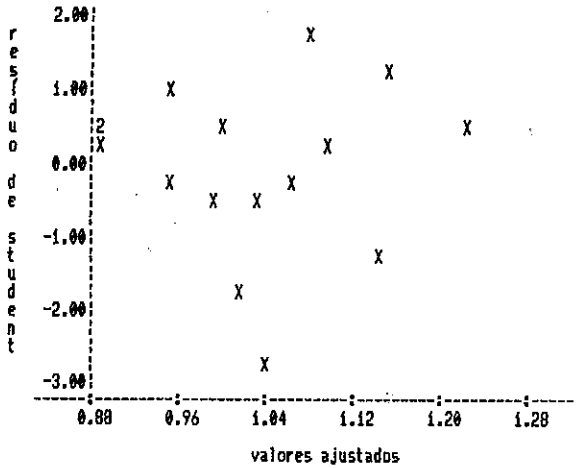

Figura 1.1: *Gráfico dos resíduos t_i's contra os $\hat{\mu}$'s impondo-se o modelo irrestrito*

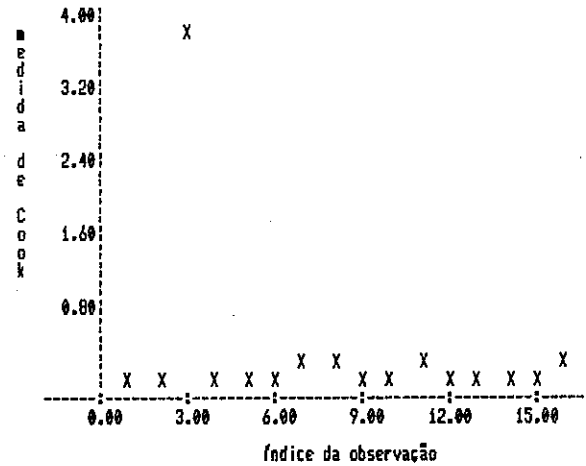

Figura 1.2:*Gráfico das distâncias de Cook contra a ordem das observações*

ponto correspondente a esse ano se situa numa região mais afastada dos demais, fazendo com que tanto h_{ii} quanto $(y_i - \hat{\mu}_i)$, para essa observação,

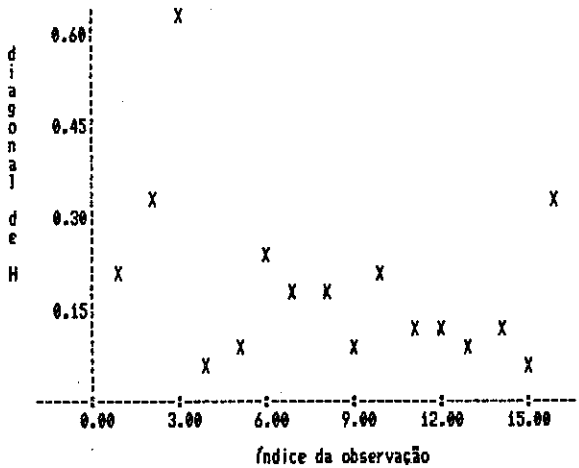

Figura 1.3: *Gráfico de h_{ii} contra a ordem das observações*

sejam valores mais extremos. Essa influência pode ser amenizada tentando-se alguma transformação na covariável x_1 e/ou na variável de resposta y.

1.18.2 - Vendas trimestrais de uma Empresa

É apresentada na Tabela 1.10 um conjunto de dados do livro de Bussab (1986), em que y representa as vendas trimestrais de uma empresa (em 1000 unidades) e x o total investido em publicidade (em mil cruzados novos).

Tabela 1.10: *Vendas e gastos trimestrais de uma empresa.*

Trimestre	x	y	Trimestre	x	y
1	11	25	5	12	25
2	5	13	6	6	12
3	3	8	7	5	10
4	9	20	8	9	15

Fonte: Bussab (1986).

O gráfico de y contra x, que é omitido aqui, sugere o modelo linear $y = \beta_0 + \beta_1 x + \varepsilon$. A suposição de normalidade para ε pode ser verificada a posteriori através do gráfico normal de probabilidades, Seção 1.8.4. O modelo ajustado, desvio padrões entre parênteses, é dado por

$$\hat{\mu} = 1.312(2.052) + 1.958(0.254)x,$$

com $s^2 = 4.646$ e $R^2 = 0.908$. A estatística F correspondente à hipótese $H: \beta_1 = 0$ vale 59.43 que é significativo a 1%, entretanto β_0 pode ser excluído do modelo. Antes, porém, é fundamental um estudo de diagnóstico.

Examinando-se os gráficos usuais, nenhum ponto se destaca como aberrante e/ou influente; entretanto, o gráfico dos resíduos t_i's contra t (trimestre), Figura 1.4, indica para a falta de um termo linear em t na componente sistemática.

Ajusta-se, então, o modelo $y = \beta_0 + \beta_1 x + \beta_2 t + \varepsilon$, obtendo-se

$$\hat{\mu} = 4.826(0.983) + 1.948(0.097)x - 0.763(0.128)t, \tag{1.64}$$

com $s^2 = 0.684$ e $R^2 = 0.989$.

Nota-se uma redução substancial em s^2 e um aumento em R^2. Além disso, agora todos os parâmetros são significativamente (a 1%) diferentes de zero. Os gráficos dos resíduos t_i's contra t e contra os valores ajustados não apresentam nenhum ponto aberrante ou qualquer tendência sistemática. Analogamente, os gráficos de D_i e h_i contra a ordem das observações não indicam pontos influentes.

Apesar de não serem recomendadas extrapolações (previsões com valores das covariáveis fora dos limites amostrais) em regressão linear, suponha como ilustração que a empresa deseja obter um intervalo de 95% para as vendas no nono trimestre, se forem investidos 10 mil cruzados novos. A

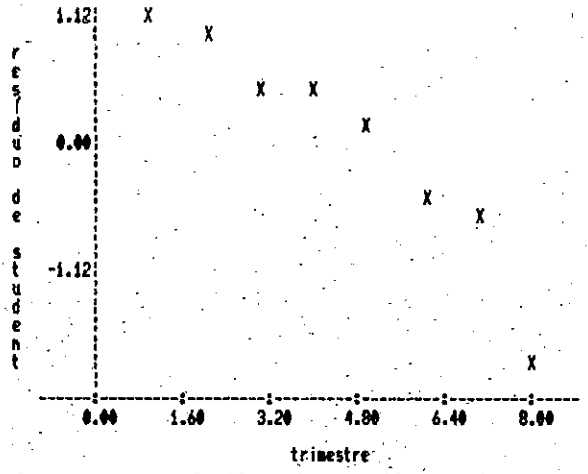

Figura 1.4 - *Gráfico dos resíduos t_i's contra t (trimestre), impondo-se o modelo* $y = \beta_0 + \beta_1 x + \beta_2 t + \varepsilon$

estimativa do valor esperado desse trimestre, usando (1.64), vale

$$\hat{\mu}(x^{(0)}) = 4.826 + 1.948 \times 10 - 0.763 \times 9 = 17.439,$$

onde $x^{(0)} = (1\ 10\ 9)^T$. A variância de $\hat{\mu}(x^{(0)})$ é obtida de (1.20), onde X é a matriz do novo modelo, $Var\{\hat{\mu}(x^{(0)})\} = 0.684 \times 0.696 = 0.476$ e, portanto, a variância da previsão $\hat{y}(x^{(0)})$ iguala $Var\{\hat{y}(x^{(0)})\} = 0.684 + 0.476 = 1.160$. Um intervalo de 95% de confiança para $y(x^{(0)})$ será formado pelos limites (vide Seção 1.9.2) $\hat{\mu}(x^{(0)}) \pm t_{0.025}\{1 + x^{(0)^T}(X^TX)^{-1}x^{(0)}\}s$, que como $t_{0.025=2.571}$, correspondem à 14.67 e 20.21.

§1.19 Exercícios

1. Ajustar aos dados do Exercício 2 da Seção 2.9 o modelo $V = \beta_0 + \beta_1 A + \beta_2 D + \varepsilon$ calculando as estimativas de β_0, β_1 e β_2. Testar as hipóteses $H: \beta_1 = 0$ versus $A: \beta_1 \neq 0$ e $H: \beta_2 = 0$ versus $A: \beta_2 \neq 0$.

2. Ajustar um modelo de regressão polinomial aos dados do Exercício 3 da Seção 2.9.

3. Aos dados da Tabela 2.1 ajustar os seguintes modelos normais- lineares:

(a) $V = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 N + \varepsilon$;

(b) $V = \beta_0 + \beta_1 T + \beta_2 P + \beta_4 C + \beta_5 U + \varepsilon$;

(b) $V = \beta_0 + \beta_1 T + \beta_2 P + \beta_4 C + \beta_5 U + \varepsilon$;

(c) $V = \beta_0 + \beta_1 logT + \beta_2 logP + \beta_3 N + \varepsilon$;

(d) $V = \beta_0 + \beta_1 T + \beta_2 P + \beta_3 N + \beta_4 C + \beta_U + \beta_6 R_1 + \beta_7 R_2 + \varepsilon$;

(e) Analisar os resíduos nos modelos ajustados (a) a (d).

4. Considere as regressões de y sobre x para os dados abaixo, 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

5. Os dados abaixo são os valores da OTN (em cruzados novos) nos anos de 1987 e 1988.

(a) Estimar o valor da OTN em Dezembro de 1989 e obter um intervalo de 95% de confiança para este valor;

	Janeiro	Fevereiro	Março	Abril	Maio	Junho
1989	0.106	0.106	0.182	0.210	0.252	0.310
1988	0.597	0.696	0.820	0.952	1.135	1.337

	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro
1987	0.367	0.378	0.402	0.402	0.463	0.523
1988	1.598	1.982	2.392	2.966	3.774	4.791

(b) Incluir nesses dados o valor 6.170 correspondente às OTNs de Janeiro a Março de 1989 e estimar novamente a OTN em Dezembro de 89, calculando o intervalo de 95% de confiança; (c) O que aconteceu com as estimativas dos parâmetros ao se incluir esses 3 dados?; (d) Comparar os ajustamentos (globalmente e localmente) nos dois casos.

6. Demonstrar: (a) a expressão (1.13); (b) a fórmula (1.16); (c) o vício (1.28); (d) a fórmula da inversa generalizada dada em (1.32); (e) a inversa (1.36); (f) a fórmula (1.49); (g) as expressões das matrizes de informação dadas em (1.53), (1.54), (1.62) e o caso particular (1.63).

7. 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.

8. 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 um 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 $\hat{\beta}, \hat{\beta}_1$, os correspondentes desvios padrões, s^2 e R^2, e a tabela ANOVA.

(ii) Calcular os resíduos de Pearson $p_i = (y_i - \hat{\mu}_i)/r$ 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. Compare o r^2 e o R^2 desse novo modelo com aqueles do modelo ajustado em (i).

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

9. Obter o algoritmo de ajustamento baseado em (1.55) para os modelos MA(1), MA(2), AR(2) e ARMA(1,1) calculando expressões fechadas para as matrizes de informação desses modelos.

10. Em 9 municípios foram observadas as seguintes variáveis: y- consumo de um determinado produto, x_1-urbanização relativa, x_2-nível educacional.

Os dados são os seguintes:

Municípios	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 ANOVA da Tabela 1.2 e teste 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 s^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).

CAPÍTULO 2

MODELO DE BOX E COX

§2.1 Definição

O uso do modelo clássico de regressão introduzido na Seção 1.1 é justificado admitindo-se (Seção 1.11): (i) linearidade da estrutura de $E(y)$; (ii) constância da variância do erro, $\text{Var}(y) = \sigma^2$; (iii) normalidade e (iv) independência das observações. Se as suposições (i) a (iii) não são satisfeitas para os dados originais, uma transformação não linear de y poderá 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

$$\text{Var}\{g(y)\} = V(\mu)g'(u)^2 = k^2,$$

onde k^2 é uma constante, a condição (ii) será satisfeita. A função estabilizadora da variância dos dados segue de $g(\mu) = k\int V(\mu)^{-1/2}d\mu$. 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 à 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 dever ser tomado ainda com o mecanismo gerador dos dados e a precisão com que estes são obtidos.

Dificuladades 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 existirem relações aproximadamente lineares entre as variáveis explicativas. Esta multicolinearidade causará problemas com as rotinas de inversão da matriz X^TX (Seções 1.11.3 e 1.15). 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 a transformar a variável dependente para satisfazer todas as hipóteses (i) a (iv) do modelo clássico de regressão. Na Seção 2.2 apresenta-se a estimação da transformação, em 2.3 acrescenta-se uma covariável ao modelo normal-linear e em 2.4 testa-se a transformação da variável dependente. A eliminação de observações e o teste de transformação das variáveis explicativas são, discutidos nas Seções 2.5 e 2.6, respectivamente. Na Seção 2.7 apresentam-se os testes de normalidade e homocedasticidade para completar o estudo feito nas Seções 1.11.1 e 1.11.2. Finalmente, uma análise de dados em Engenharia de Avaliações é feita na Seção 2.8 ilustrando as

potencialidades do modelo de Box e Cox.

O modelo de Box e Cox (1964) admite que os dados $y = (y_1, \ldots, y_n)^T$ são independentes e que existe um escalar λ tal que os dados transformados por

$$z = z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \text{ se } \lambda \neq 0 \\ \log y \text{ se } \lambda = 0 \end{cases} \tag{2.1}$$

satisfazem $E(z) = \mu = X\beta$, $\text{Var}(z_i) = \sigma^2$ para $i = 1, \ldots, n$ e $z \sim N(\mu, \sigma^2 I)$. A transformação (2.1) 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 constância da variância dos dados transformados, pode ser que diferentes valores de λ sejam necessários para alcançar tudo isso.

§2.2 Estimação da Transformação

Um valor para 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

$$\begin{aligned} L(\lambda, \sigma^2, \beta) = &-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(z - X\beta)^T(z - X\beta) \\ &+ (\lambda - 1)\sum_{i=1}^{n} log y_i, \end{aligned} \tag{2.2}$$

onde o terceiro termo é o logaritmo do Jacobiano da transformação, isto é, $J(\lambda; y) = \Pi_{i=1}^{n} \left|\frac{dz}{dy}\right|$. A maximização de (2.2) em relação a λ, σ^2 e β apresenta problemas computacionais e deve ser feita em duas etapas. Fixa-se λ e maximiza-se $L(\lambda, \sigma^2, \beta)$ em relação aos demais parâmetros produzindo as estimativas usuais da regressão como funções de λ, $\hat{\beta}(\lambda) = (X^T X)^{-1} X^T z$ e $\hat{\sigma}^2(\lambda) = \frac{1}{n} z^T (I - H) z$, sendo H a matriz de projeção. O máximo da log-verossimilhança como função de λ vale, exceto por uma constante,

$$\hat{L}(\lambda) = -\frac{n}{2} \log \sigma^2(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log y_i. \tag{2.3}$$

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 os 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.

Em termos computacionais é mais econômico calcular $\hat{\sigma}^2(\lambda)$ como $\frac{1}{n}[z^T z - w^T (X^T X)^{-1} w]$ com $w = X^T z$ do que como $\frac{1}{n} z^T (I - H) z$.

Este método será facilmente estendido se o parâmetro da transformação for um vetor embora apresentará inconvenientes na representação gráfica.

§2.3 Adição de uma Covariável no Modelo Normal Linear

Considera-se o modelo normal-linear definido na Seção 1.1 com a adição de uma variável explicativa extra w correspondente ao parâmetro adicional γ,

produzindo a componente sistemática particionada

$$E(y) = \mu = X\beta + w\gamma. \tag{2.4}$$

O interesse é deduzir um teste de significância para o termo extra, ou seja, para $H: \gamma = 0$ versus $A: \gamma \neq 0$.

Demonstra-se que a inversa da matriz $(X\ \omega)^T\ (X\ \omega)$ é dada por

$$\begin{pmatrix} (X^TX)^{-1} + mdd^T & -md \\ -md & m \end{pmatrix}, \tag{2.5}$$

onde $d = (X^TX)^{-1}X^T\omega$ é a estimativa do parâmetro na regressão de ω sobre X e o escalar m iguala $1/(\omega^T\omega + \omega^THw)$, sendo H a matriz de projeção do modelo $E(y) = X\beta$. A matriz (2.5), exceto pelo multiplicador σ^2, é a estrutura de covariância de $\begin{pmatrix} \hat{\beta} \\ \hat{\gamma} \end{pmatrix}$. Os valores ajustados segundo o modelo (2.4) têm estrutura de covariância igual a

$$\sigma^2[H + m(I - H)\omega\omega^T(I - H)].$$

Demonstra-se que $\hat{\gamma}$ pode ser dado em termos dos resíduos das regressões de w e y sobre X, definidos por $R_y = (I - H)y$ e $R_w = (I - H)w$,

$$\hat{\gamma} = R_w^TR_y/R_w^TR_w = \frac{w^T(I - H)y}{w^T(I - H)w}. \tag{2.6}$$

Esta expressão mostra que $\hat{\gamma}$ é a estimativa da declividade da regressão passando pela origem da variável dependente y sobre R_w. Um gráfico de y versus R_w, freqüentemente denominado *gráfico da variável adicionada* pode ser usado para verificar a necessidade de termos adicionais na componente sistemática, bem como, indicar observações aberrantes ou influentes.

Define-se o produto escalar $S(a,b) = a^T(I-H)b$ como a soma residual de produtos dos vetores a e b. A variância de $\hat{\gamma}$ iguala $\sigma^2/w^T(I-H)w$ e, portanto, uma estatística para o teste de $H\colon \gamma = 0$ é expressa dividindo $\hat{\gamma}$ pelo seu erro padrão

$$T_w(0) = S(w,y)/\hat{\sigma}S(w,w)^{1/2}. \tag{2.7}$$

Em (2.7) usa-se como estimativa da variância do erro $\hat{\sigma}^2$ a expressão

$$\hat{\sigma}^2 = \frac{S(w,w)S(y,y) - S(w,y)^2}{(n-p-1)S(w,w)}. \tag{2.8}$$

Observa-se então que para testar se o efeito da adição de uma variável extra é significativo, necessita-se somente das somas de quadrados residuais dos y's e de produtos cruzados dos y's e w's. O teste é realizado comparando $T_w(0)$ com a distribuição t de Student com $n-p-1$ graus.

§2.4 Teste de transformação da variável dependente

Retorna-se ao modelo de Box e Cox discutido nas Seções 2.1 e 2.2, onde o objetivo aqui é realizar inferência sobre λ. O teste da hipótese nula composta $H\colon \lambda = \lambda_0$ versus $A\colon \lambda \neq \lambda_0$, onde λ_0 é um valor especificado para λ, pode ser feito comparando a razão de verossimilhança $RV = 2[\hat{L}(\hat{\lambda}) - \hat{L}(\lambda_0)]$ com a distribuição assintótica χ^2_1. A vantagem deste é de ser independente da parametrização adotada mas requer o cálculo de $\hat{\gamma}$. Pode-se ainda trabalhar com a raíz quadrada de RV que, assintoticamente segundo H, tem distribuição $N(0,1)$.

Um intervalo de $100(1-\alpha)\%$ de confiança para λ é facilmente deduzido do gráfico de $\hat{L}(\lambda)$ versus λ como

$$\{\lambda; L(\lambda) > \hat{L}(\hat{\lambda}) - \frac{1}{2}\chi_1^2(\alpha)\}. \tag{2.9}$$

Na Seção 2.8 ilustra-se o cálculo deste intervalo numa análise de dados reais. Se $\lambda = 1$ não pertencer ao intervalo (2.9) conclui-se que uma transformação dos dados será necessária e pode-se selecionar um valor conveniente (inteiro ou metade de inteiro) neste intervalo.

Alternativamente, testes podem ser desenvolvidos a partir dos resíduos nas regressões de z e $w = \partial z/\partial\lambda$ sobre X, no lugar de serem baseados na verossimilhança, seguindo a teoria da covariável adicionada desenvolvida na seção anterior. Para isto é mais fácil trabalhar com a transformação padronizada definida por $z = z(\lambda) = (y^\lambda - 1)/\lambda J(\lambda; y)^{1/n}$, que reduz-se a

$$z = z(\lambda) = \begin{cases} \frac{(y^\lambda - 1)}{\lambda \tilde{y}^{\lambda-1}} & se\ \lambda \neq 0 \\ \tilde{y}\log y & se\ \lambda = 0, \end{cases} \tag{2.10}$$

onde $\tilde{y}$ é a média geométrica dos $y_i's$. Para esta transformação padronizada o Jacobiano é um e, portanto, $\hat{L}(\lambda) = -\frac{n}{2}\log\hat{\sigma}^2(\lambda)^2$, isto é, a log-verossimilhança maximizada sobre os parâmetros β's como função de λ depende dos dados somente através da soma de quadrados dos resíduos dos $z_i's$.

Expandindo (2.10) em série de Taylor ao redor de um valor λ_0 obtém-se a aproximação

$$z(\lambda) = z(\lambda_0) + w(\lambda_0)(\lambda - \lambda_0), \tag{2.11}$$

onde a covariável adicionada vale

$$w(\lambda) = \frac{\partial z(\lambda)}{\partial\lambda} = \frac{y^\lambda \log y - (y^\lambda - 1)(1/\lambda + \log\tilde{y})}{\lambda\tilde{y}^{\lambda-1}}. \tag{2.12}$$

Calculando o valor esperado de (2.11) e desde que $E[z(\lambda)] = X\beta$ para algum λ vem

$$E\{z(\lambda_0)\} = X\beta - w(\lambda_0)(\lambda - \lambda_0)$$

e, portanto, usando os resultados de (2.4) e (2.6), a estimativa aproximada de λ é expressa por

$$\tilde{\lambda} = \lambda_0 - \frac{S(w, z; \lambda_0)}{S(w, w; \lambda_0)}, \tag{2.13}$$

onde $S(w, z; \lambda_0) = w(\lambda_0)^T(I-H)z(\lambda_0)$ e $S(w, w; \lambda_0) = w(\lambda_0)^T(I-H)w(\lambda_0)$ são produtos escalares de $w(\lambda_0)$ pelos resíduos das regressões de $z(\lambda_0)$ e $w(\lambda_0)$ sobre X, respectivamente.

A estimativa $\tilde{\lambda}$ não é a $EMV\hat{\lambda}$ de λ e pode diferir bastante desta. Entretanto, se a expressão (2.13) for usada iterativamente as estimativas sucessivas de λ convergirão para $\hat{\lambda}$. A precisão de $\hat{\lambda}$, isto é, $\mathrm{Var}(\hat{\lambda})$, ainda não é conhecida devido à dificuldade de inversão da matriz de informação particionada em λ, β e σ^2. Draper e Cox (1969) propuseram uma fórmula para $\mathrm{Var}(\hat{\lambda})$ que infelizmente está incorreta. Agora, a partir da aproximação (2.13) deduz-se

$$\mathrm{Var}(\tilde{\lambda}) = \hat{\sigma}_z^2 S(w, w; \lambda_0), \tag{2.14}$$

onde $\hat{\sigma}_z^2$ é obtido por analogia a (2.8)

$$\hat{\sigma}_z^2 = \frac{1}{n-p-1} S(z, z; \lambda_0) - \frac{S^2(w, z; \lambda_0)}{S(w, w; \lambda_0)}. \tag{2.15}$$

Ressalta-se que o cálculo de expressões do tipo $S(a, b; \lambda_0)$ é mais vantajoso a partir de $a^T b - a^T\hat{b}$, onde $\hat{b}$ representa os valores ajustados na regressão de b sobre X.

Um teste alternativo aproximado para a transformação $H: \lambda = \lambda_0$ versus $A: \lambda \neq \lambda_0$ resulta de (2.7)

$$T_w(\lambda_0) = \frac{-S(w, z; \lambda_0)}{\{\hat{\sigma}_z^2 S(w, w; \lambda_0)\}^{1/2}}. \tag{2.16}$$

Um valor significante para $T_w(\lambda_0)$ indica que a transformação requerida para os dados deve ser diferente do valor especificado λ_0. A estatística (2.16) tem aproximadamente distribuição $t_{(n-p-1)}$.

§2.5 Eliminação de observações

Para detectar observações influentes no modelo transformado via λ_0 usa-se o gráfico de $z(\lambda_0)$ versus $w(\lambda_0)$ este definido em (2.12); espera-se que uma observação situada fora da tendência geral deste gráfico, ao ser eliminada, tenha um efeito apreciável sobre a transformação adotada. Determina-se agora o efeito da eliminação de um conjunto de m observações indexada pelo vetor L sobre $\tilde{\lambda}$ e $T_w(\lambda_0)$.

Eliminando-se m observações no modelo $E(y) = X\beta$, a soma dos quadrados dos resíduos original $(n-p)s^2 = y^T(I-H)y$ reduz-se a

$$(n-p-m)s_L^2 = (n-p)s^2 - R_L^T(I-H)^{-1}R_L,$$

onde R_L representa os m resíduos indexados por L e H_L é a submatriz $m \times m$ correspondente de H. Define-se

$$S_L(a, b; \lambda_0) = S(a, b; \lambda_0) - a_L^{*T}(I - H_L)^{-1}b_L^*,$$

onde a_L^* e b_L^* são as m componentes relevantes dos resíduos das regressões dos vetores a e b sobre X. No modelo aumentado (2.11), a estimativa

do coeficiente da regressão de $w(\lambda_0)$ após a eliminação das m observações, segue de (2.6)

$$\hat{\gamma}_L = S_L(w, y; \lambda_0)/S_L(w, w; \lambda_0).$$

Usando esta expressão obtém-se como estimativa da transformação após essa eliminação

$$\tilde{\lambda}_L = \tilde{\lambda}_L(\lambda_0) = \lambda_0 - \frac{S_L(w, z; \lambda_0)}{S_L(w, w; \lambda_0)}. \tag{2.17}$$

A dedução de (2.17) inclui uma aproximação adicional, pois as variáveis $z(\lambda)$ são funções da média geométrica dos dados e esta não permanece constante com a eliminação das m observações. Entretanto, quando m for pequeno comparado com n a mudança em $\tilde{y}$ poderá ser negligenciada. O efeito da eliminação na estatística (2.16) é dado por

$$T_{wL}(\lambda_0) = \frac{-S_L(w, z; \lambda_0)}{\{\hat{\sigma}^2_{zL} S_L(w, w; \lambda_0)\}^{1/2}}, \tag{2.18}$$

onde aqui $\hat{\sigma}^2_{zL}$ é obtido por analogia a (2.15) com a eliminação dos dados indexados por L.

§2.6 Teste de transformação das variáveis explicativas

A verificação da aditividade das variáveis explicativas do modelo pode ser feita através do teste de Tukey (1949). Para isto calcula-se a covariável adicionada $w = \hat{\mu} \otimes \hat{\mu}$, onde $\hat{\mu} = X\hat{\beta}$ e $\otimes$ representa o produto direto e forma-se o modelo estendido $\mu = X\beta + w\gamma$. A não-aditividade das covariáveis em X é medida pela significância do coeficiente γ no modelo estendido, sendo o

teste realizado com base na estatística $t = \hat{\gamma}/\operatorname{Var}(\hat{\gamma})^{1/2}$ que tem distribuição t de Student com $n-p-1$ graus.

No caso do coeficiente γ ser significativo, uma transformação da variável resposta como no modelo de Box e Cox poderá apresentar aditividade na escala transformada. O teste de Tukey estima a transformação por $\tilde{\lambda} = 1 - 2t\overline{y}$ (vide exemplo na Seção 2.8).

Box e Tidwell (1962) apresentaram um procedimento iterativo para obter transformações adequadas das variáveis explicativas $x_1, \ldots, x_p$ através da relação funcional

$$\mu = g(z_1, \ldots, z_p), \tag{2.19}$$

onde $z_i = x_i^{\alpha_i}$ se $\alpha_i \neq 0$ e $z_i = \log x_i$ se $\alpha_i = 0$ e $g(\cdot)$ é uma função linear ou quadrática nos z_i's.

Os parâmetros a serem estimados são os α_i's e os coeficientes dos z_i's. Arbitrando valores iniciais iguais a 1 para os α_i's tem-se a expansão

$$\mu = g(x_1, \ldots, x_p) + \sum_{i=1}^{p} (\alpha_i - 1) \frac{dg}{d\alpha_i}\Big|_{\alpha_i = 1}, \tag{2.20}$$

onde $\frac{dg}{d\alpha_i} = \frac{dg}{dz_i} x_i \log x_i$ se simplifica para $\beta_i x_i \log x_i$ por ser $g(\cdot)$ linear ou quadrática nos z_i's. Os ajustamentos dos dois modelos $\mu = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$ e $\mu = \beta_0' + \sum_{i=1}^{p} \beta_i' x_i + \sum_{i=1}^{p} \gamma_i x_i \log x_i$ produzem as estimativas $\hat{\beta}_i, \hat{\beta}_i'$, $i = 0, \ldots, p$ e $\gamma_i, i = 1, \ldots, p$ e, portanto, estima-se os α_i's por $\hat{\alpha}_i = \hat{\gamma}_i/\hat{\beta}_i + 1$. Este procedimento deverá ser repetido para as novas variáveis explicativas $z_i = x_i^{\alpha_i}$, $i = 1, \ldots, p$, até a convergência.

Ramsey (1969) desenvolveu ainda vários testes de adequação da componente sistemática do modelo normal-linear. A inadequação desta componente pode ser devida à omissão de uma variável explicativa importante ou à não-inclusão de funções das variáveis no modelo.

Um teste aproximado de H: a componente $E(y) = 1\alpha + X\beta$ é adequada, onde $(1\ X)$ tem posto completo p e admite-se que $1^T X = 0$, isto é, as variáveis explicativas estão centradas, segue da regressão dos resíduos $r = y - \hat{\mu}$ obtidos com o ajustamento de $(1\ X)$ a y, sobre a matriz formada por potências dos valores ajustados $Q = (\hat{\mu}^{(2)}\hat{\mu}^{(3)}\hat{\mu}^{(4)})$, onde $\hat{\mu}^{(\ell)}$ é o vetor da ℓ-ésima potência dos valores ajustados $\hat{\mu}_i$'s.

Cada elemento de Q é subtraído da média da coluna correspondente para formar $\tilde{Q}$ e calcular C^TC e C^Tr em $C^TC = \tilde{Q}^T\tilde{Q} - (\tilde{Q}^TX)(X^TX)^{-1}(\tilde{Q}^TX)^T$ e $C^Tr = \tilde{Q}^Tr$. Estas duas expressões permitem obter $S_2 = (C^Tr)^T(C^TC)^{-1}(C^Tr)$, que representa a soma dos quadrados explicada pela regressão de r sobre $C = (I - H)Q$, sendo aqui $H = n^{-1}11^T + X(X^TX)^{-1}X^T$, e daí a estatística

$$F = (n - p - 3)S_2/3(S_1 - S_2), \tag{2.21}$$

onde $S_1 = r^Tr$. O teste de H compara (2.21) com a distribuição $F_{3,(n-p-3)}$. Se F é significativo a componente sistemática em H mostra-se inadequda.

§2.7 Testes de normalidade e homocedasticidade

Nesta seção alguns testes são apresentados para verificar normalidade e homocedasticidade dos dados ao se adotar o modelo clássico de regressão. Um estudo detalhado desses testes pode ser encontrado nos Capítulos 8 e 9 de Wetherill et al. (1986) e nos artigos aqui citados. Todos esses testes só são válidos supondo observações independentes e identicamente distribuídas, mas podem ser aplicados, como meras aproximações, aos resíduos da regressão, pois estes não são independentes.

Sejam $t_i = (y_i - \hat{\mu}_i)/s(1-h_{ii})^{1/2}$, como definidos em (1.12), os resíduos studentizados da regressão e $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$ os seus valores ordenados crescentemente. Seja m_i o valor esperado da i- ésima estatística de ordem em n normais reduzidas. Aproximadamente

$$m_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right). \tag{2.22}$$

Portanto, um gráfico de $t_{(i)}$ versus m_i próximo da 1ª bissetriz é um indicativo de aceitação da hipótese de normalidade H.

Um teste analítico é baseado na estatística (D'Agostino, 1971)

$$D^2 = \frac{[\sum_{i=1}^{n} b_i t_{(i)}]^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{2.23}$$

onde $b_i = [i - (n+1)/2]/n^{3/2}$, que equivale à razão de duas estimativas de σ^2. Segundo H, D^2 tem distribuição que independe de σ^2. Valores de D próximos de zero sugerem rejeição de H.

Demonstra-se que os dois primeiros momentos (μ_1' e μ_2) e os coeficientes de assimetria (γ_1) e curtose (γ_2) de D têm as séries assintóticas

$$\mu_1' = 0.2821 - 0.0705/n + 0.0088/n^2 + 0.0110/n^3 - 0.0029/n^4,$$
$$\mu_2' = 0.0009/n - 0.0005/n^2 - 0.0050/n^3 + 0.0031/n^4,$$
$$\gamma_1 = -8.5836/\sqrt{n} + 33.8083/n\sqrt{n} - 63.0418/n^2\sqrt{n},$$
$$\gamma_2 = 114.7320/n - 961.4588/n^2.$$

Estas expansões permitem determinar o ponto crítico d_α de D em relação ao ponto crítico z_α da $N(0,1)$ na fórmula de Cornish-Fisher

$$d_\alpha = \mu_1' + \sqrt{\mu_2}\left\{z_2 + \frac{\gamma_1}{6}(z_\alpha^2 - 1)\right.$$
$$\left. + \frac{\gamma_2}{24}(z_\alpha^3 - 3z_\alpha) - \frac{\gamma_1^2}{36}(2z_\alpha^3 - 5z_\alpha)\right\}. \tag{2.24}$$

O teste bilateral da hipótese de normalidade H ao nível α é realizado calculando $d_{\alpha/2}$ e $d_{1-\alpha/2}$ sendo a região de rejeição de H dada pelo intervalo $(d_{\alpha/2},\ d_{1-\alpha/2})$.

Um teste de normalidade mais simples, alternativo ao anterior, é baseado na assimetria e curtose dos resíduos padronizados. Calculam-se os momentos centrais dos resíduos $m_r = n^{-1}\sum_{i=1}^{n}(t_i - \bar{t})^r$, $\bar{t}$ é o resíduo médio amostral, e os seus coeficientes amostrais de assimetria $\sqrt{b_1} = m_3/m_2^{3/2}$ e curtose $b_2 = m_4/m_2^2 - 3$. Espera-se que b_1 e b_2 sejam próximos de zero se os dados forem normais. A estatística $S = n(b_1/6 + b_2^2/24)$ (Bowman e Shenton, 1975) segundo H tem distribuição assintótica χ_2^2. Então, se $S > \chi_2^2(\alpha)$ H é rejeitada.

O objetivo agora é testar a hipótese de homocedasticidade dos dados. Violações desta hipótese deve-se a dados representativos de médias ou com variâncias que dependem da média ou de variáveis explicativas, ou é devido ao mecanismo gerador dos dados ou ainda a presença de observações aberrantes.

Gráficos dos resíduos versus covariáveis (possivelmente associadas com heterocedasticidade) são usualmente eficazes para rejeitar H quando se detecta que resíduos ou as suas variâncias estão variando com a escala adotada no eixo horizontal. Um teste analítico de H é baseado na estatística

$$\phi = \frac{[\sum_{i=1}^{n}(\hat{\mu}_i - \overline{y})r_i^2]^2}{\sum_{i=1}^{n}(\hat{\mu}_i - \overline{y})^2 \sum_{i=1}^{n}(r_i^2 - \overline{r^2})}, \tag{2.25}$$

onde $r_i = y_i - \hat{\mu}_i$ e $\overline{r^2}$ é a média aritmética dos r_i^2.

A hipótese de homocedasticidade é rejeitada ao nível de significância α se $\phi > \chi_1^2(\alpha)$.

Caso haja interesse em testar heterocedasticidade específica a uma

combinação de variáveis $w_1, \ldots, w_q$ utiliza-se a estatística (Godfrey, 1978)

$$G = \frac{1}{2}\tilde{r}^T W(W^T W)^{-1} W^T \tilde{r}, \tag{2.26}$$

onde $W = (1, w_1, \ldots, w_q)$, $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_n)^T$ com $\tilde{r}_i = n(y_i - \hat{\mu}_i)/\sum_{i=1}^{n}(y_1 - \hat{\mu}_i)^2 - 1$, que tem distribuição assintótica χ^2_q. Nota-se que $2G$ é a soma dos quadrados devida à regressão de $\tilde{r}$ sobre W. A matriz W pode ser um subconjunto de variáveis explicativas de X.

§2.8 Análise de Dados em Engenharia de Avaliações

A Engenharia de Avaliações utiliza essencialmente o método comparativo, pelo qual o valor de um bem é obtido por comparação com outros de características similares. Ocorre que, após a coleta dos elementos de referência, o avaliador está geralmente de posse de uma amostra aleatória formada por dados de características heterogêneas, sendo imprescindível a sua homogeneização para aplicação adequada do método comparativo. Observa-se no Brasil, que a grande maioria dos profissionais que atuam neste ramo, têm aplicado esta prática sem os cuidados necessários, utilizando-se de fatores empíricos e ponderações de ordem subjetiva, o que implica numa sensível perda do nível de precisão dos seus trabalhos.

Um caminho para resolver a grande maioria dos problemas em Engenharia de Avaliações é a utilização de modelos de regressão. A Tabela 2.1 apresenta os dados de 50 lotes urbanos, situados nos bairros de Casa Forte, Torre e Iputinga da cidade de Recife, utilizados pela Caixa Econômica Federal para avaliação de diversas glebas, a serem adquiridas para

Tabela 2.1: *Dados de terrenos em três bairros da cidade de Recife com as características*: T = testada, P = profundidade, N = natureza do evento, C = contemporaneidade, U = nível de urbanização, R_1, R_2, R_3 = localização e V = valor por m^2 do terreno.

T	P	N	C	U	R_1	R_2	R_3	V
19	30	0	31	6	1	0	0	2.982
17	29	1	22	7	1	0	0	2.695
36	29	0	32	6	1	0	0	3.831
10	20	0	26	4	1	0	0	2.250
12	36	1	25	6	0	1	0	3.588
11	31	1	20	7	0	1	0	1.261
10	30	1	14	8	0	1	0	1.587
15	20	0	31	4	0	1	0	2.667
10	30	0	31	4	0	1	0	3.333
13	36	0	31	6	0	1	0	4.273
16	34	0	52	7	0	1	0	29.994
9	52	1	27	8	0	0	1	4.288
24	39	0	31	6	0	0	1	7.478
15	82	0	31	8	0	0	1	10.339
36	39	1	3	6	0	0	1	2.422
16	135	1	32	8	0	0	1	5.734
22	22	1	39	7	0	0	1	4.019
40	233	1	50	8	0	0	1	16.094
14	84	1	42	8	0	0	1	9.267
16	33	0	54	8	0	0	1	32.567
17	27	1	45	6	0	0	1	9.918
17	30	1	29	8	0	0	1	7.843
12	247	1	26	8	0	0	1	7.093

17	21	0	55	5	0	0	1	29.790
25	37	0	55	5	0	0	1	32.258
24	30	0	44	5	1	0	0	4.047
14	26	1	29	5	1	0	0	3.710
12	40	0	44	5	1	0	0	6.250
13	30	0	44	8	1	0	0	10.970
14	28	0	43	4	1	0	0	5.639
13	41	0	43	7	1	0	0	9.259
10	25	0	44	7	1	0	0	10.737
9	22	0	43	4	1	0	0	4.067
15	32	0	43	5	1	0	0	5.208
15	30	0	32	5	1	0	0	3.333
15	64	0	32	5	1	0	0	2.083
66	29	0	8	5	1	0	0	2.712
15	25	0	31	5	1	0	0	5.333
14	24	1	23	7	1	0	0	3.535
25	80	1	20	7	1	0	0	2.306
18	30	1	44	7	1	0	0	13.789
10	28	1	43	6	1	0	0	10.714
9	30	0	57	8	0	1	0	30.476
14	30	0	57	3	0	1	0	28.571
14	71	0	57	7	0	1	0	35.211
9	25	0	57	4	0	1	0	14.846
15	66	0	57	6	0	1	0	15.344
16	25	1	49	7	0	1	0	8.040
18	23	0	54	5	0	0	1	24.154
14	30	0	30	5	1	0	0	3.333

implantação do projeto Recife-Programa de Revitalização do Rio Capiba-

ribe. Esses dados correspondem aos valores unitários dos terrenos coletados em $NCz\$/m^2(V)$, as suas respectivas características físicas de testada efetiva (T), profundidade equivalente (P) e nível de Urbanização (V), a época da ocorrência da avaliação (C), a natureza do evento que deu origem à avaliação (N) e a localização: Iputinga (R_1), Torre (R_2) e Casa Forte (R_3) (Dantas e Cordeiro, 1989).

As variáveis T, P, C, U e V são quantitativas e N, R_1, R_2 e R_3 variáveis do tipo $0-1$. O nível de urbanização (U) é uma covariável variando de 1 a 8 que cresce com o nível de beneficiamento existente na vizinhança do terreno; a natureza do evento (N) é 0 para ofertas e 1 para negociações; a contemporaneidade (C) representa o número de meses ocorridos entre uma data base e a efetivação da avaliação; e a localização do terreno é dada pelos valores de R_1, R_2 e R_3.

Considera-se aqui que a variável resposta V tem distribuição normal. Uma outra distribuição para V será proposta na Seção 6.9. Fazendo a regressão de V sobre as demais variáveis independentes, obtém-se a média ajustada (erros padrões das estimativas dos parâmetros estão entre parênteses)

$$\begin{aligned}\hat{\mu}_i = &-15.760(6.130) + 0.179(0.091)T_i - 0.020(0.020)P_i \\ &+ 0.517(0.071)C_i - 3.394(2.140)N_i \\ &+ 1.381(0.705)U_i - 5.999(2.168)R_{1i} - 0.536(2.350)R_{2i}.\end{aligned} \tag{2.27}$$

Observar que a coluna de R_3 é uma combinação linear de R_1 e R_2 e, portanto, foi suprimida. Neste ajustamento $SQRes = 1289.6$, $\hat{\sigma}^2 = 31.45$ e $R^2 = 0.72$. Assim, 72% da variação dos valores dos lotes é explicada pelo modelo normal-linear. A Tabela 2.2 de análise de variância mostra que pelo menos algumas das variáveis independentes do modelo realmente explicam

a variação dos valores unitários dos terrenos, pois $F = 15.76$ excede por larga margem o ponto crítico $F(7, 42) = 3.10$ ao nível de 1%

Tabela 2.2: *Análise de Variância do Modelo Normal-linear para V*

Fonte	G.l.	SQ	MQ	F
Regressão	7	3386.4	483.77	15.76
Resíduo	42	1289.6	30.70	
Total	49	4676.0		

O gráfico dos valores dos terrenos (V) versus as médias ajustadas ($\hat{\mu}$) da Figura 2.1, revela a inadequação do modelo normal-linear para V, pois a grande maioria dos pontos estão mal ajustados, havendo 40 observações com resíduos em valor absoluto superiores a 2. Ainda, várias médias ajustadas (aquelas correspondentes às observações 2,4,7,15,27,39 e 40) são negativas revelando uma situação impossível de acontecer.

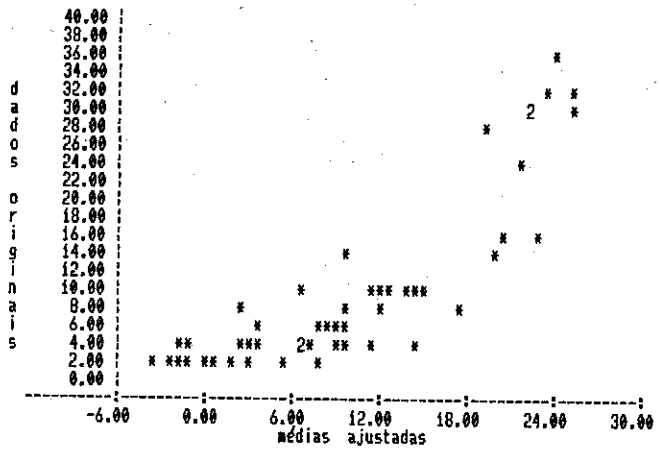

Figura 2.1: *Gráfico dos valores dos terrenos versus as médias ajustadas.*

Propõe-se então um modelo de Box e Cox para esses dados. A transformação pode ser obtida pelo método de Tukey descrito na Seção 2.6. Acrescenta-se ao modelo original a covariável $w = \hat{\mu} \otimes \hat{\mu}$ e obtém-se como estimativa do parâmetro γ correspondente, $\hat{\gamma} = 0.0572$, com erro padrão 0.0086, implicando no valor $t = 6.18$, resultado superior ao ponto crítico da distribuição t com 40 graus ao nível de 5% (2.23). Assim, alguma transformação da variável resposta é necessária sendo estimada por $\tilde{\lambda} = 1 - 2t\overline{y} = -0.084$. Como este valor é próximo de zero, pode-se trabalhar com a transformação logarítmica.

A transformação a ser aplicada aos valores unitários dos terrenos pode ser também determinada pelo método gráfico de Box e Cox (Seções 2.2 e 2.4).

Considerando a transformação definida em (2.10), a Figura 2.2 mostra o gráfico da log-verossimilhança maximizada $\hat{L}(\lambda) = -\frac{n}{2}\log\hat{\sigma}^2(\lambda)$ versus λ. Deste gráfico se conclui que o máximo de $\hat{L}(\lambda)$ ocorre próximo de $\tilde{\lambda} = -0.1$, correspondente a $\hat{L}(-0.1) = -43.21$, e praticamente coincide com o valor encontrado pelo método de Tukey. Usando a expressão (2.9) obtém-se um intervalo de 95% de confiança para λ como $\{\lambda; L(\lambda) > -45.13\}$, que equivale a $-0.35 < \lambda < 0.07$, e inclui a transformação logarítmica e exclui a identidade.

A regressão da variável transformada $\log V$ sobre as mesmas variáveis independentes do modelo anterior, implica na seguinte expressão para as médias ajustadas de $\log V$ (erros padrões entre parênteses).

$$\begin{aligned}\hat{E}(\log V_i) = &- 0.780(0.433) + 0.013(0.006)T_i - 0.991(0.001)P_i \\ &+ 0.056(0.005)C_i - 0.193(0.151)N_i + 0.127(0.050)U_i \\ &- 0.479(0.153)R_{1i} - 0.291(0.166)R_{2i}.\end{aligned} \tag{2.28}$$

A Tabela 2.3 mostra a análise de variância relativa ao ajustamento do

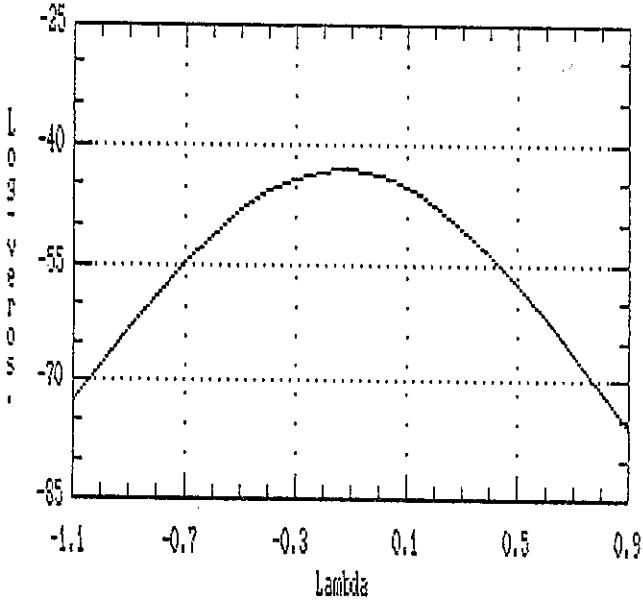

Figura 2.2: *Gráfico da log-verossimilhança maximizada (vertical) versus o parâmetro de transformação (horizontal).*

modelo normal-linear para $\log V$. A estimativa da variância é $\hat{\sigma}^2 = 0.153$ e o modelo explica 84% ($R^2 = 0.84$) da variação total dos dados, ocorrendo um incremento de 12% com a transformação da variável resposta em relação ao modelo anterior.

As Figuras 2.3, 2.4 e 2.5 apresentam os gráficos de $\log V_i$ versus $\hat{E}(\log V_i)$, dos resíduos $\log V_i - \hat{E}(\log V_i)$ versus $\hat{E}(\log V_i)$ e dos resíduos ordenados versus os quantis da $N(0,1)$, respectivamente. Das Figuras 2.3 e 2.4 se constata que todos os resíduos estão no intervalo $(-1.0, 0.75)$ e que 84% deles estão entre -0.5 e 0.5 indicando condições favoráveis ao ajustamento. Os pontos na Figura 2.4 apresentam-se distribuídos aleatoriamente e então aceita-se a hipótese de independência e variância constante dos resíduos. Os pontos da Figura 2.5 estão bastante próximos da bissetriz do

1º quadrante e indica que a distribuição normal para $\log V$ parece razoável.

As Figuras 2.6 e 2.7 correspondem aos gráficos dos elementos h_{ii} da diagonal da matriz de projeção H versus as médias ajustadas e da estatística de Cook $C_i = \{\log V_i - \hat{E}(\log V_i)\}^2 h_{ii}/(1-h_{ii})^2$ versus o índice i das observações, respectivamente. Estes gráficos servem para detectar pontos influentes no modelo adotado. Da Figura 2.6 nota-se que nenhum elemento da diagonal da matriz "hat" é superior a p/n, isto é 0.16, e portanto, não deve haver pontos influentes.

Tabela 2.3: *Análise de variância do modelo normal-linear para* $\log V$.

Fonte	G.l.	SQ	MQ	F
Regressão	7	32.71	4.67	31.13
Resíduo	42	6.43	0.15	
Total	49	39.14		

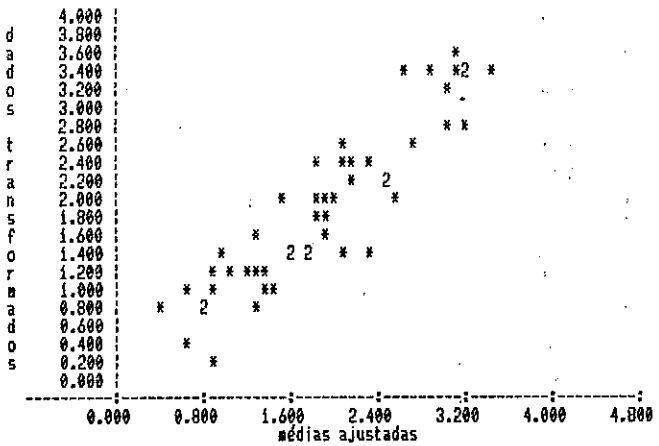

Figura 2.3: *Logarítmos dos dados versus médias ajustadas supondo que os primeiros são normais.*

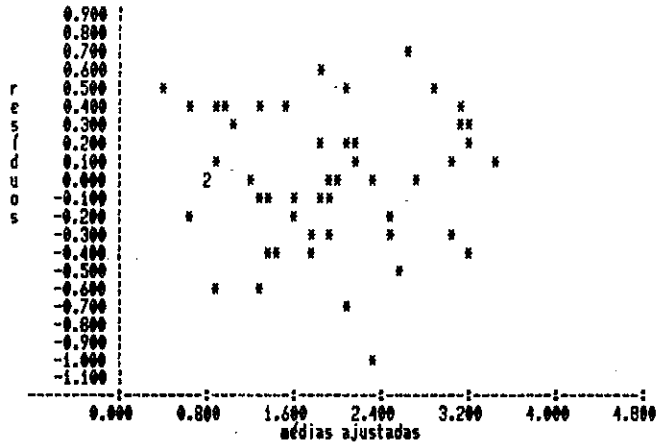

Figura 2.4: *Resíduos versus médias ajustadas supondo* $\log V$ *normal.*

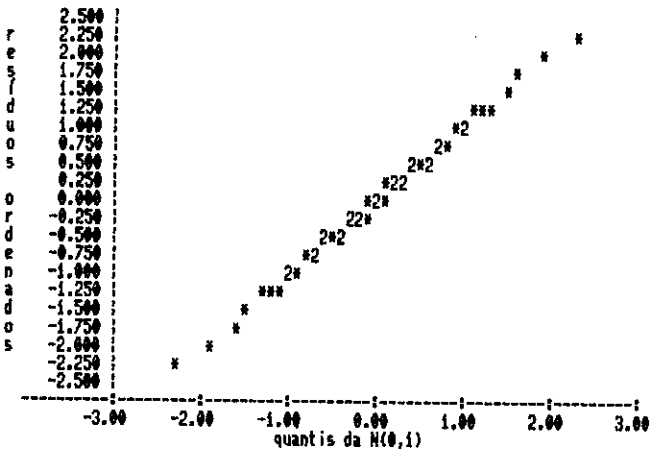

Figura 2.5: *Resíduos ordenados versus quantis da* $N(0,1)$ *supondo* $\log V$ *normal.*

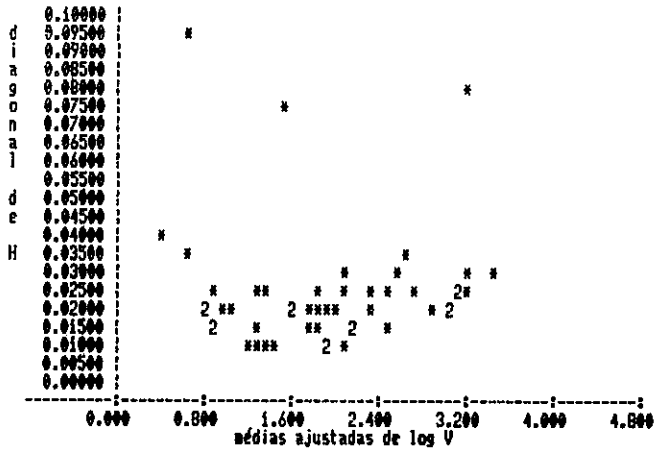

Figura 2.6: *Gráfico de h_{ii} versus médias ajustadas.*

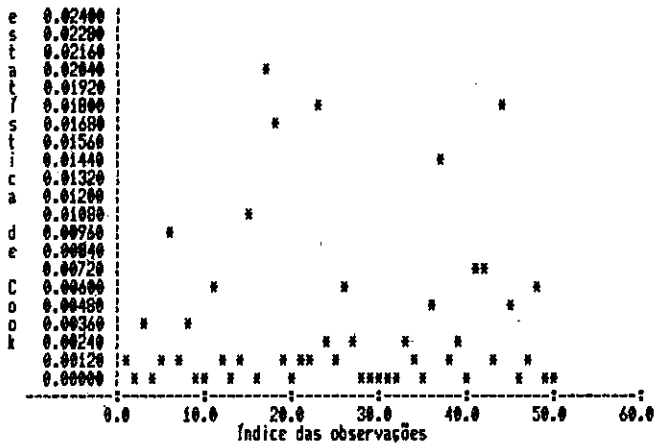

Figura 2.7: *Gráfico de C_i versus i.*

Como a hipótese de $\log V$ normal foi aceita, V tem distribuição log-normal de média $\exp\{\hat{E}(\log V_i) + \frac{1}{2}.0.15\}$, que com o uso de (2.28) reduz-se à expressão

(2.29)

$$\hat{\mu}_i = 0.494\ 1.013^{T_i}\ 0.999^{P_i}\ 1.057^{C_i}\ 0.824^{N_i} \times 1.135^{U_i}\ 0.619^{R_{1i}}\ 0.748^{2i}.$$

A fórmula (2.29) poderá ser usada para estimar os valores por m^2 de novos terrenos na área abrangida pela pesquisa. De (2.29) observa-se que os valores dos lotes crescem de forma diretamente proporcional à testada, ao nível de urbanização e ao tempo; decrescem com o aumento da profundidade equivalente, e são mais valorizados no sentido Iputinga, Torre e Casa Forte. Verifica-se ainda, que, em média, os terrenos são negociados por 82.4% do valor de oferta inicial, a taxa de valorização territorial no período pode ser estimada em 5.7% a.m., os valores unitários crescem de 13.5% a cada nível de urbanização e são em Iputinga e Torre, 61.9% e 74.8% dos valores de terrenos similares em Casa Forte, respectivamente. Todos estes resultados são bastante coerentes com a crença a priori que se tinha desses efeitos.

A variância de V_i pode ser encontrada por $\mathrm{Var}(V_i) = 0.188\exp(2\mu_i)$ com μ_i estimado de (2.29).

A Figura 2.8 mostra o gráfico dos dados originais (V) versus as médias ajustadas obtidas da expressão (2.29). Os resíduos padronizados $(V_i - \hat{\mu}_i)/\hat{V}ar(V_i)$ versus médias ajustadas são apresentados na Figura 2.9. Estas duas figuras indicam a adequação do modelo log-normal para V. Com a exceção da observação 44 todos os resíduos estão no intervalo $(-1.65, 1.65)$, com 38 pontos em $(-1, 1)$.

Este exemplo ilustra as potencialidades do modelo de Box e Cox em situações onde as hipóteses do modelo normal-linear não são satisfeitas.

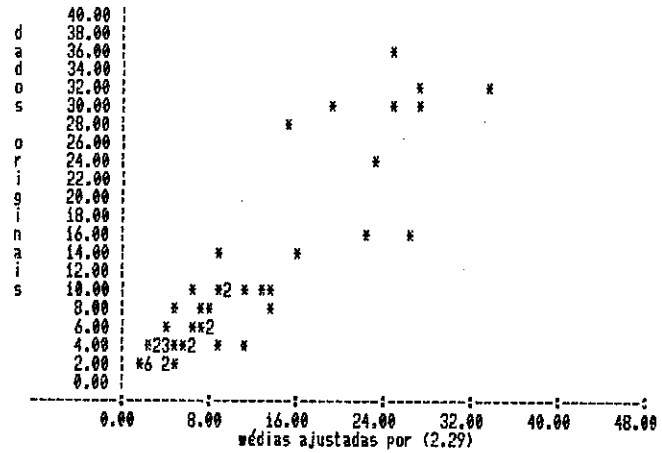

Figura 2.8: *Dados originais versus médias ajustadas no modelo log-normal para V.*

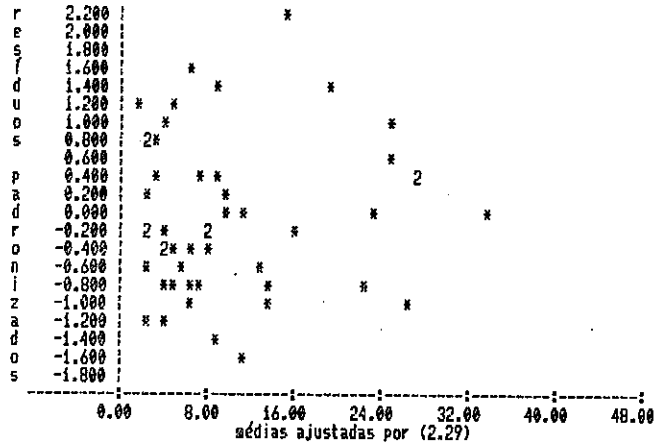

Figura 2.9: *Resíduos padronizados versus médias ajustadas no modelo log-normal para V.*

§2.9 Exercícios

1. Aplicar o processo de estimação de Box e Cox (Seção 2.2) para estimar os parâmetros da transformação

$$z(\lambda) = \begin{cases} \{(y+\lambda_2)^{\lambda_1} - 1\}/\lambda_1 \; se \; \lambda_1 \neq 0 \\ \log(y+\lambda_2) \; se \; \lambda_1 = 0. \end{cases}$$

2. Ajustar um modelo de Box e Cox 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				

3. Os dados a seguir correspondem à área de um pasto em função do tempo de crescimento. Ajustar um modelo de Box e Cox 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

4. Obter uma fórmula aproximada para a variância de $\hat{\lambda}$ (Seção 2.4) invertendo a matriz de informação particionada em β, σ^2 e λ.
5. Aplicar o teste de transformação das variáveis explicativas desenvolvido na Seção 2.6 para estimar α_1 e α_2 considerando a média do volume V do Exercício 2 expressa por $\mu = \beta_0 + \beta_1 A^{\alpha_1} + \beta_2 D^{\alpha_2}$. Estimar também os β's.
6. Aplicar o teste de Ramsey (Seção 2.6) para verificar a adequação do modelo $E(V) = \beta_0 + \beta_1 A + \beta_2 D$ para os dados do Exercício 2.
7. Aplicar os testes de normalidade e homocedasticidade aos valores unitários dos terrenos da Tabela 2.1
8. Para os dados analisados na Seção 2.8 propor um modelo Box e Cox supondo que a testada e a profundidade dos terrenos são covariáveis medidas na escala logarítmica, como sugerido pelas normas brasileiras 5676. Verificar a adequação do modelo.
9. Verificar a existência de observações influentes no conjunto de dados do Exercício 2 usando o método descrito na Seção 2.5.
10. Analisar os dados do Exercício 5 da Seção 6.10 através de um modelo de Box e Cox.

CAPÍTULO 3

MODELOS PARA ANÁLISE DE DADOS CATEGORIZADOS

§3.1 A Distribuição de Poisson

Os modelos para análise de *dados categorizados* representando freqüências são baseados na distribuição de Poisson. Esta distribuição pode ser deduzida teoricamente por princípios elementares com um mínimo de suposições. Além de descrever dados experimentais representando contagens, ela pode modelar o número de eventos em qualquer intervalo de tempo fixado, desde que estes ocorram aleatoriamente e independentemente no tempo com taxa de ocorrência constante.

Aqui são citadas algumas propriedades da distribuição de Poisson. Em 3.2 o modelo multinomial é introduzido. A inferência sobre o parâmetro da distribuição de Poisson é considerada na Seção 3.2. As classificações unidimensional e bidimensional são tratadas nas Seções 3.4 e 3.5, respectivamente. Os modelos log-lineares hierárquicos são introduzidos em 3.6. O algoritmo de ajustamento dos modelos log-lineares e testes de adequação são desenvolvidos em 3.7 e 3.8, respectivamente. Finalmente, na Seção 3.9, dois exemplos de análise de dados reais são apresentados.

Uma variável Y tem distribuição de Poisson com parâmetro $\mu (Y \sim P(\mu))$ quando

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \; y = 0, 1, \ldots . \tag{3.1}$$

A função geratriz de momentos da distribuição de Poisson iguala

$$M(t; \mu) = \exp[\mu\{\exp(t) - 1\}] \tag{3.2}$$

e os seus momentos centrais seguem a lei de recorrência com $\mu = E(Y)$ e $\mu_0 = 1$.

$$\mu_{r+1} = r\mu\mu_{r-1} + \mu\frac{\partial \mu_r}{\partial \mu}, r \geq 1.$$

Em particular, $\text{Var}(Y) = \mu_2 = \mu$, $\mu_3 = \mu$ e $\mu_4 = \mu + 3\mu^2$. Quando $\mu \to \infty$ a variável padronizada $(Y - \mu)/\mu^{1/2}$ converge em distribuição para $N(0, 1)$. A função de distribuição de Y pode ser aproximada por

$$P(Y \leq y) = \Phi\{(y + 0.5 - \mu)/\mu^{1/2}\},$$

onde $\Phi(\cdot)$ é a distribuição acumulada da $N(0, 1)$. O fator 0.5 é conhecido como correção de continuidade. Uma aproximação mais razoável é dada por $P(Y \leq y) = 1 - \Phi(z)$, onde

$$z = 3(y + 1)^{-1/2}\left\{\left(\frac{\mu}{y+1}\right)^{3/2} - 1 + \frac{1}{9(y+1)}\right\}.$$

A distribuição acumulada da Poisson pode também ser expressa em termos da acumulativa da qui-quadrado

$$P(Y \leq y) = P(\chi^2_{2(y+1)} > 2\mu). \tag{3.3}$$

De $P(Y = y+1)/P(Y = y) = \mu/(y+1)$, nota-se que $P(Y = y)$ cresce com y até o máximo correspondente a $y^* = [\mu]$, maior inteiro menor ou igual a μ, ou a dois máximos correspondentes a $y^* = \mu - 1$ e $y^* = \mu$ se μ é um inteiro, e depois decresce com y crescente.

A expansão de Cornish-Fisher, com $Z = (Y - \mu)\mu^{-1/2}$, é dada por

$$Z - \frac{1}{3}(Z^2 - 1)\mu^{-1/2} + \frac{1}{36}(7Z^3 - Z)\mu^{-1}, \tag{3.4}$$

sendo aproximadamente $N(0,1)$. Uma aproximação mais simples é considerar $2\sqrt{Y}$ ou $2\sqrt{Y + 3/8}$ como $N(0,1)$.

Probabilidades individuais de Poisson podem ser calculadas, aproximadamente, por $P(Y = y) = \Phi(y_+) - \Phi(y_-)$, onde $y_+ = (y - \mu + 0.5)\mu^{-1/2}$ e $y_- = (y - \mu - 0.5)\mu^{-1/2}$. A soma dessas probabilidades individuais pode ser expressa por

$$P(y_1 \leq Y \leq y_2) = \Phi(z_2) - \Phi(z_1) + \frac{1}{6}(2\pi\mu)^{-1/2}\{H(z_2) - H(z_1)\}, \tag{3.5}$$

onde $H(z) = (1 - z^2)\exp(-z^2/2)$, $z_1 = (y_1 - \mu - 0.5)\mu^{-1/2}$ e $z_2 = (y_2 - \mu + 0.5)\mu^{-1/2}$.

O modelo de Poisson desempenha na análise de dados categorizados o mesmo papel do normal na análise de dados contínuos. A diferença fundamental é que a estrutura multiplicativa para as médias do modelo de Poisson é mais apropriada do que a estrutura aditiva das médias do modelo normal. Tem-se constatado, na análise de dados categorizados, que a média μ pode ser expressa como um produto de outras médias marginais, que se tornam os parâmetros lineares do modelo quando se adota a escala logarítmica para μ. Por exemplo, independência de dois fatores numa tabela de contingência $r \times s$ equivale ao modelo $\mu_{ij} = \mu_{i+}\mu_{+j}/\mu_{++}$, com a

notação usual para a soma de índices, e isto implica que o logaritmo de μ_{ij} é igual à uma estrutura linear formada pelos efeitos principais dos fatores sem a interação.

§3.2 O Modelo Multinomial

Sejam $Y_1, \ldots, Y_r$ variáveis independentes com cada Y_i tendo distribuição $P(\mu_i)$, $i = 1, \ldots, r$. Usando a função geratriz de probabilidades demonstra-se, facilmente, que $\sum_{i=1}^{r} Y_i \sim P(\sum_{i=1}^{r} \mu_i)$. Considera-se a distribuição conjunta de $Y_1, \ldots, Y_r$ condicionada ao total de observações $\sum_{i=1}^{r} Y_i = y_+$. Tem-se

$$P(Y_1 = y_1, \ldots, Y_r = y_r \mid \sum Y_i = y_+)$$
$$= \Pi(\mu_i^{y_i} e^{-\mu_i}/y_i!) / \left\{ \left(\sum \mu_i \right)^{y_+} e^{-\sum \mu_i} / y_+! \right\},$$

com os somatórios e o produtório variando de 1 a r. Definindo $p_i = \mu_i / \sum \mu_j$ vem

$$P(Y_1 = y_1, \ldots, Y_r = y_r) = \frac{y_+!}{\Pi y_i!} \Pi p_i^{y_i}, \tag{3.6}$$

o que resulta na distribuição multinomial de índice y_+ e probabilidades $p_1, \ldots, p_r$.

Quando $r = 2$, a distribuição de Y_1 condicionada a $Y_1 + Y_2 = y_+$ é binomial com índice y_+ e probabilidade $\mu_1/(\mu_1 + \mu_2)$.

Este resultado de que a distribuição multinomial pode ser deduzida da distribuição conjunta de variáveis de Poisson independentes condicionada ao total observado, mostra a equivalência do modelo de Poisson para as freqüências com o modelo de resposta multinomial para o estudo de proporções.

§3.3 Inferência Sobre o Parâmetro da Distribuição de Poisson

Admite-se que $Y_1, \ldots, Y_r$ são independentes com cada variável tendo distribuição $P(\mu)$. Seja $\overline{Y} = \sum_{i=1}^{r} Y_i/r$ a média e $\overline{y}$ o seu valor amostral. A normalidade assintótica de $(Y - \mu)/\mu^{1/2}$ implica que $(\overline{Y} - \mu)(r/\mu)^{1/2}$ converge em distribuição para $N(0,1)$ quando r é fixo e $\mu \to \infty$. Ainda, de acordo com o teorema do limite central, $(\overline{Y} - \mu)(r/\mu)^{1/2}$ converge para $N(0,1)$ quando μ é fixo e $r \to \infty$. Assim, para r grande ou μ grande

$$P\{(\overline{Y} - \mu)^2 r/\mu \leq z_{\alpha/2}^2\} = (1 - \alpha), \tag{3.7}$$

onde $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. Os valores de μ que satifazem a expressão (3.7) constituem um intervalo de $100(1-\alpha)\%$ de confiança para μ, cujos limites são

$$\overline{y} + z_{\alpha/2}^2/2r \pm z_{\alpha}/2 \left(\frac{\overline{y}}{r} + \frac{z_{\alpha/2}^2}{4r^2} \right)^{1/2}.$$

Para r grande esses limites reduzem-se a $\overline{y} \pm z_{\alpha/2}(\overline{y}/r)^{1/2}$.

A log-verossimilhança como função de μ supondo os dados $y_1, \ldots, y_r$ iguala

$$L(\mu) = r\overline{y} log\mu - r/\mu - \sum \log y_i!,$$

implicando que $\overline{Y}$ é suficiente para μ. Ainda $\partial L(\mu)/\partial\mu = r(\overline{y} - \mu)/\mu$ e, então, $\overline{Y}$ é o único estimador de μ não-tendencioso e de variância mínima μ/r.

Uma maneira alternativa de encontrar um intervalo de confiança para μ baseia-se no fato de $\sum_{i=1}^{r} Y_i \sim P(r\mu)$. Os limites inferior $\underline{\mu}$ e superior $\overline{\mu}$ para μ seguem de

$$\exp(-r\underline{\mu}) \sum_{y=n\overline{y}}^{\infty} (r\underline{\mu})^y/y! = \alpha/2$$

e

$$\exp(-r\overline{\mu}) \sum_{y=0}^{n\overline{y}} (r\overline{\mu})^y/y! = \alpha/2.$$

Usando (3.3) vem

$$\underline{\mu} = \frac{1}{2r}\chi^2_{2r\overline{y},\alpha/2}, \quad \overline{\mu} = \frac{1}{2r}\chi^2_{2(r\overline{y}+1),1-\alpha/2}, \tag{3.8}$$

que são calculados através da interpolação na tabela χ^2 central.

§3.4 A Classificação Unidimensional

Considere que n unidades de uma população são classificadas em r categorias que podem representar diferenças qualitativas ou quantitativas ou uma ordem natural de classificação. Uma unidade selecionada ao acaso na população tem probabilidade p_i de estar na categoria i, $i = 1, \ldots, r$. Seja Y_i a freqüência de unidades classificadas na categoria i. Supondo que, $Y_1, \ldots, Y_r$ são variáveis de Poisson independentes associadas às categorias com parâmetros $\mu_1, \ldots, \mu_r$, respectivamente, a distribuição de $Y_1, \ldots, Y_r$ condicionada a $\sum_{i=1}^{r} Y_i = n$, como visto na Seção 3.2, é multinomial de índice n e probabilidades $p_1 = \mu_1/n, \ldots, p_r = \mu_r/n$.

Seja a hipótese de homogeneidade de médias H: $\mu_1 = \cdots = \mu_r = \mu$ a ser testada contra A: μ_i não é constante. Supondo H verdadeira, a distribuição conjunta de $Y_1, \ldots, Y_r$ condicionada a $\sum_{i=1}^{r} Y_i = n$ é uma multinomial simétrica com $p_i = 1/r$, $i = 1, \ldots, r$, e a freqüência média $\overline{Y} = \sum_{i=1}^{r} Y_i/r$ é uma estatística suficiente para o valor comum μ das médias. Ainda, neste caso, a distribuição da estatística

$$X^2 = \sum_{i=1}^{r} (Y_i - \overline{Y})^2/\overline{Y} \tag{3.9}$$

condicionada a $\overline{Y}$, independe de μ, dependendo somente de n e r.

Salienta-se aqui que no estudo de dados categorizados normalmente ocorrem dois tipos de situações assintóticas: a 1ª quando μ é fixo e $n \to \infty$ e a 2ª quando n permanece fixo e $\mu \to \infty$.

A distribuição de X^2 condicionada a $\overline{Y}$ converge para χ^2_{r-1} quando $n \to \infty$ com os dois primeiros momentos exatos expressos por $E(X^2 \mid \overline{Y}) = r - 1$ e $\mathrm{Var}(X^2 \mid \overline{Y}) = 2(r-1)(1 - n^{-1})$.

Esta estatística X^2 é um caso especial da estatística de Pearson $X^2 = \sum (o-e)^2/e$, onde $o(e)$ representa a freqüência observada (esperada), comumente usada para testar a adequação entre os valores esperados segundo uma hipótese e os valores observados. Como $X^2/(r-1)$ representa o quociente da variância amostral pela média amostral, o valor de $X^2/(r-1)$ será aproximadamente um se a amostra for proveniente de uma única distribuição de Poisson mas irá exceder 1 se os Y_i's tiverem distribuições de Poisson diferentes ou mesmo uma única distribuição de Poisson mista.

O teste de H versus A pode também ser baseado na razão de verossimilhança

$$Y^2 = 2 \sum_{i=1}^{r} Y_i \log(Y_i/\overline{Y}), \tag{3.10}$$

assintoticamente equivalente a X^2. Esta estatística é um caso bem simples da forma (3.28) $2\sum o\log(o/e)$ a ser tratada na Seção 3.8. Por sua vez esta última expressão é um caso particular do desvio do modelo de Poisson obtido da fórmula geral (6.6).

Considera-se agora uma reparametrização das médias dos $Y_i's$: $\beta = \log\mu_r$ e $\beta_i = \log(\mu_i/\mu_r)$ para $i = 1,\ldots,r$. Assim, a média de Y_i fica expressa por

$$\log\mu_i = \beta + \beta_i. \tag{3.11}$$

Observar que a hipótese de homogeneidade pode ser colocada na forma H: $\beta_i = 0$ $i = 1,\ldots,r-1$ já que $\beta_r = 0$.

Nesta parametrização a distribuição conjunta dos Y_i's é dada por

$$\exp\{e^{-\beta}\sum e^{\beta_i} + y_{++}\beta + \sum y_i\beta_i\}/\Pi y_i!$$

e, portanto, a soma $\sum_{i=1}^{r} Y_i$ e as freqüências individuais $Y_1,\ldots,Y_r$ são estatísticas suficientes para β e $\beta_1,\ldots,\beta_r$, respectivamente. As estimativas de máxima verossimilhança das médias são iguais às freqüências correspondentes, pois o modelo é saturado. Então, maximizando a exponencial anterior vem as estimativas $\hat{\beta} = \log y_r$ e $\hat{\beta}_i = \log(y_i/y_r)$. O modelo (3.11) é saturado (número de parâmetros igual ao número de observações) e, portanto, $\hat{\mu}_i = y_i$ de onde se obtém diretamente as expressões de $\hat{\beta}$ e dos $\hat{\beta}_i$'s.

A distribuição conjunta de $Y_1,\ldots,Y_r$ condicionada a

$$\sum_{i=1}^{r} Y_i = n$$

claramente é multinomial e não envolve o parâmetro β

$$P(Y_1 = y_1,\ldots,Y_r = y_r \mid \sum_{i=1}^{r} Y_i = n) = \frac{n!\exp(\sum y_i\beta_i)}{\Pi y_i!(\sum e^{\beta_i})^n}.$$

§3.5 A Classificação bidimensional

Uma amostra de tamanho n é extraída de uma população e classificada por cada um de dois métodos A e B, contendo r e s categorias, respectivamente. Seja Y_{ij} o número de unidades classificadas na categoria i de A e j de B. Supondo que os Y_{ij}'s são variáveis independentes de Poisson de médias μ_{ij}'s, a distribuição conjunta dos Y_{ij}'s condicionada a $\sum_{i,j} Y_{ij} = n$ é multinomial de índice n e probabilidades dadas por $p_{ij} = \mu_{ij}/n$. Adotando-se a seguinte parametrização $\log \mu_{rs} = \beta$, $\beta_i^A = \log(\mu_{is}/\mu_{rs})$, $\beta_j^B = \log(\mu_{rj}/\mu_{rs})$ e $\beta_{ij}^{AB} = \log(\mu_{ij}/\mu_{is}\mu_{rj})$, obtém-se o modelo

$$\log \mu_{ij} = \beta + \beta_i^A + \beta_j^B + \beta_{ij}^{AB}, \tag{3.12}$$

para $i = 1, \ldots, r$ e $j = 1, \ldots, s$. Notar que $\beta_r^A = \beta_s^B = 0$ e $\beta_{is}^{AB} = 0$ para $i = 1, \ldots, r$ e $\beta_{rj}^{AB} = 0$ para $j = 1, \ldots, s$.

Os modelos (3.11) e (3.12) expressam uma estrutura linear para os logaritmos das médias de variáveis de Poisson, denominados *modelos log-lineares*, os quais são representados de uma forma geral por

$$\log \mu_i = \beta + \sum_{j=1}^{p} x_{ij}\beta_j \ , \tag{3.13}$$

onde os x_{ij}'s são 0's ou 1's.

Estes modelos usados na análise de dados categorizados, que representam contagens em classificações denominadas de *tabelas de contingência*, tiveram grande desenvolvimento na década dos 70, principalmente, na Universidade de Chicago. Os seguintes livros: Bishop, Fienberg e Holland

(1975), Haberman (1974, 1978a,b), Fienberg (1980) e Plackett (1981) apresentam a teoria dos modelos log-lineares. Presentemente, eles são estudados dentro do contexto dos modelos lineares generalizados (Capítulo 6), pois são um caso especial destes modelos.

O modelo (3.12) tem um parâmetro $\beta, r-1$ parâmetros relativos ao método A (β_i^A's), $s-1$ parâmetros relativos ao método B (β_j^B's) e $(r-1)(s-1)$ parâmetros associados aos dois métodos (β_{ij}^{AB}'s), totalizando rs parâmetros independentes, mesmo número de médias desconhecidas, sendo portanto saturado. As estimativas de máxima verossimilhança desses parâmetros são $\hat{\beta} = \log y_{rs}, \hat{\beta}_i^A = \log(y_{is}/y_{rs})$, $\hat{\beta}_j^B = \log(y_{rj}/y_{rs})$ e $\hat{\beta}_{ij}^{AB} = \log(y_{ij}y_{rs}/y_{is}y_{rj})$ para $i = 1, \ldots, r$ e $j = 1, \ldots, s$. Quando o modelo log-linear é saturado as estimativas dos parâmetros em (3.13) são obtidas diretamente das expressões que formam a parametrização arbitrária das médias, pois as freqüências esperadas são iguais às observadas.

De maneira análoga à Seção 3.4 verifica-se, facilmente, da distribuição conjunta dos Y_{ij}'s reparametrizada nesses β's, que a freqüência total Y_{++}, as freqüências marginais totais Y_{i+}'s e Y_{+j}'s e as freqüências individuais Y_{ij}'s são estatísticas suficientes para os parâmetros $\beta, \beta_i^{A\prime}s, \beta^{B\prime}s$ e $\beta_{ij}^{AB\prime}s$, respectivamente.

O objetivo usual no modelo (3.12) é testar a hipótese de que os métodos de classificação A e B são independentes, isto é, testar $H: p_{ij} = p_{i+}p_{+j}$, ou equivalentemente, $H: \mu_{ij} = \mu_{i+}\mu_{+j}/\mu_{++}$ para todo par (i,j). Na reparametrização esta hipótese fica expressa por $H: \beta_{ij}^{AB} = 0$ para todo par (i,j). Para se fazer inferência sobre os parâmetros β_{ij}^{AB}'s que medem a associação entre os métodos A e B, trabalha-se com a distribuição conjunta dos Y_{ij}'s condicionada a $Y_{i+} = y_{i+}$ $(i = 1, \ldots, r)$ e $Y_{+j} = y_{+j}$ $(j = 1, \ldots, s)$.

Demonstra-se facilmente que

$$f(y_{ij}'s \mid y_{i+}'s, y_{+j}'s; \beta_{ij}^{AB}'s)$$

$$= \frac{\exp(\sum_{i,j} y_{ij}\beta_{ij}^{AB})}{\Pi_{i,j} y_{ij}!} \left\{ \sum_{y_{ij}'s} \frac{\exp(\sum_{i,j} y_{ij}\beta_{ij}^{AB})}{\Pi_{i,j} y_{ij}!} \right\}^{-1}, \tag{3.14}$$

onde $\sum_{y_{ij}'s}$ representa um somatório para todos os valores possíveis dos $y_{ij}'s$ satisfazendo $\sum_j y_{ij} = y_{i+}$ e $\sum_i y_{ij} = y_{+j}$. A densidade (3.14) é bastante trabalhosa de ser calculada requerendo programas de computador. Quando H é verdadeira, isto é, todos os β_{ij}^{AB}'s nulos, (3.14) reduz-se a

$$f(y_{ij}'s \mid y_{i+}'s, y_{+j}'s; 0) = \frac{1}{n!}\Pi_{i,j}\frac{y_{i+}!y_{+j}!}{y_{ij}!} \tag{3.15}$$

implicando ainda $E(Y_{ij}) = y_{i+}y_{+j}/n$ e $Cov(Y_{ij}, Y_{k\ell}) = \frac{y_{i+}y_{+j}}{n^2(n-1)}(\delta_{ik}n - y_{k+})(\delta_{j\ell}n - y_{+\ell})$, sendo δ_{ij} o delta de Kronecker. Quando $r = s = 2$ a densidade (3.14) é denominada hipergeométrica generalizada (vide Seção 4.2), pois segundo H se reduz à distribuição hipergeométrica.

Seja $\beta^{AB} = (\beta_{ij}^{AB})$ uma matriz $(r-1) \times (s-1)$ cujos elementos são os parâmetros desconhecidos da distribuição condicionada (3.14). Define-se por $\tilde{\beta}^{AB}$ a estimativa de máxima verossimilhança condicional de β^{AB} baseada em (3.14). Pode-se demonstrar que as equações para determinar $\tilde{\beta}^{AB}$ são expressas na forma

$$y_{ij} = E(Y_{ij}; \tilde{\beta}^{AB}) = m(\tilde{\beta}^{AB}, y_{i+}'s, y_{+j}'s)$$

para $i = 1, \ldots, r$ e $j = 1, \ldots, s$. As funções $m(\beta^{AB}, y_{i+}'s, y_{+j}'s)$ dos parâmetros desconhecidos e das freqüências totais marginais são bastante complicadas. A solução $\tilde{\beta}^{AB}$, quando existir, é única e pode ser obtida por

otimização numérica usando o procedimento do simplex (Nelder e Mead, 1965).

A estimação incondicional de β^{AB} é bem simples sendo expressa por

$$\hat{\beta}_{ij}^{AB} = \log(\frac{y_{ij}y_{rs}}{y_{is}y_{rj}}), \tag{3.16}$$

para $i = 1, \ldots, r$ e $j = 1, \ldots, s$.

Uma vez estimado β^{AB} as médias são univocamente determinadas resolvendo o sistema de rs equações

$$\sum_j \mu_{ij}(\beta^{AB}) = y_{i+}, \; i = 1, \ldots, r, \tag{3.17}$$

$$\sum_i \mu_{ij}(\beta^{AB}) = y_{+j}, \; j = 1, \ldots, s, \tag{3.18}$$

$$\log\left\{\frac{\mu_{ij}(\beta^{AB})\mu_{rs}(\beta^{AB})}{\mu_{is}(\beta^{AB})\mu_{rj}(\beta^{AB})}\right\} = \beta_{ij}^{AB} \tag{3.19}$$

para $i = 1, \ldots, r-1$ e $j = 1, \ldots, s-1$.

Além da estimação de β^{AB} é de interesse prático o teste de $H: \beta^{AB} = \overline{\beta}^{AB}$ versus $A: \beta^{AB} \neq \overline{\beta}^{AB}$, onde $\overline{\beta}^{AB}$ é uma matriz especificada de associação entre os métodos A e B.

Inicialmente deve-se calcular de (3.17), (3.18) e (3.19) as estimativas das médias para $\beta^{AB} = \overline{\beta}^{AB}$, isto é, $\mu_{ij}(\overline{\beta}^{AB})$, $i = 1, \ldots, r$ e $j = 1, \ldots, s$. O teste de H versus A pode ser baseado em qualquer uma das estatísticas seguintes, que têm distribuição assintótica $\chi^2_{(r-1)(s-1)}$,

$$X^2 = \sum_{i,j}\{y_{ij} - \mu_{ij}(\overline{\beta}^{AB})\}/\mu_{ij}(\overline{\beta}^{AB})$$

$$Y^2 = 2\sum_{i,j} y_{ij}\log\{y_{ij}/\mu_{ij}(\overline{\beta}^{AB})\}.$$

A hipótese nula mais freqüente é a de independência H: $\beta^{AB} = 0$, existindo uma forma explícita para as estimativas das médias segundo H

$$\mu_{ij}(0) = \frac{y_{i+}y_{+j}}{n}. \tag{3.20}$$

Regiões de confiança para β^{AB} podem ser construídas da aproximação $\chi^2_{(r-1)(s-1)}$ para X^2 ou Y^2.

Toda a teoria das Seções 3.4 e 3.5 pode ser generalizada para classificações multidimensionais, embora com maior dificuldade na álgebra e na parte computacional (Plackett,1981, Capítulo 7). Na seção seguinte apresenta-se uma parametrização dos modelos log-lineares, encontrada na análise de variância de experimentos fatoriais, que permite com maior facilidade, estudar a teoria para classificações multidimensionais.

§3.6 Modelos Log-lineares Hierárquicos

Considere o modelo log-linear (3.12), discutido na Seção 3.5, para a classificação de dois fatores A e B com as seguintes restrições sobre os β's, usadas comumente na análise de variância: $\beta^A_+ = 0, \beta^B_+ = 0, \beta^{AB}_{i+} = 0,\ i = 1, \ldots, r$ e $\beta^{AB}_{+j} = 0,\ j = 1, \ldots, s$. Este modelo é saturado, pois inclui todos os termos possíveis e, então, as estimativas dos β's são diretamente obtidas de (3.12) por somas simples em i e j, $\hat{\beta} = \frac{1}{rs}\sum_{i,j} \log y_{ij}, \hat{\beta}^A_i = \frac{1}{s}\sum_j \log y_{ij} - \hat{\beta}, \hat{\beta}^B_j = \frac{1}{s}\sum_i \log y_{ij} - \hat{\beta}$ e $\hat{\beta}_{ij} = \log y_{ij} - \hat{\beta} - \hat{\beta}^A_i - \hat{\beta}^B_j$.

A Tabela 3.1 apresenta uma lista de alguns modelos log-lineares relativos à classificação bidimensional. O modelo 1 significa que uma observação tem a mesma chance de cair em qualquer cela, isto é, $\mu_{ij} = \frac{n}{rs}$.

O modelo 2(3) corresponde à hipótese de que todos os níveis do fator $B(A)$ são equiprováveis dentro dos níveis do fator $A(B)$, isto é, $\mu_{ij} = \mu_{i+}/s(\mu_{ij} = \mu_{+j}/r)$. O modelo 4 significa que A e B são independentes e é equivalente à hipótese $\mu_{ij} = \mu_{i+}\mu_{+j}/\mu_{++}$.

Nesta tabela não foram incluídos modelos log-lineares do tipo $\log \mu_{ij} = \beta + \beta_i^A + \beta_{ij}^{AB}$ ou $\log \mu_{ij} = \beta + \beta_j^B + \beta_{ij}^{AB}$. Estes modelos não têm importância prática. Os modelos log-lineares de interesse são denominados *hierárquicos* e têm a seguinte propriedade: se um conjunto T for constituído por parâmetros β*ts* iguais a zero, então, em qualquer outro conjunto de parâmetros gerado por termos que contenham pelo menos

Tabela 3.1: *Modelos log-lineares para a classificação bidimensional. Os números de parâmetros independentes estão entre parênteses.*

	Modelo	Descrição
1	$\beta \quad (1)$	A e B nulos
2	$\beta + \beta_i^A \quad (r)$	B nulo
3	$\beta + \beta_j^B \quad (s)$	A nulo
4	$\beta + \beta_i^A + \beta_j^B \quad (r+s-1)$	A e B independentes
5	$\beta + \beta_i^A + \beta_j^B + \beta_{ij}^{AB} \quad (rs)$	A e B dependentes

um termo gerador do conjunto T, todos os parâmetros deverão ser iguais a zero. Assim, o modelo $\log \mu_{ij} = \beta + \beta_i^A + \beta_{ij}^{AB}$ não é hierárquico, pois a interação β_{ij}^{AB} está incluída sem o termo β_j^B estar presente.

Um outro exemplo de modelo não-hierárquico é o modelo $\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ijk}^{ABC}$ para a classificação tridimensional em três fatores A, B e C sujeito às restrições usuais, pois a interação β_{ijk}^{ABC} está incluída sem as 3 interações de 2ª ordem desses fatores 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 das estimativas de MV e, mais importantemente, uma interpretação simples.

O algoritmo de ajustamento dos modelos log-lineares a ser apresentado na Seção 3.7 não faz qualquer distinção entre um modelo não-hierárquico ou hierárquico.

Os modelos hierárquicos podem ser classificados em duas classes: a 1ª cujas estimativas $\hat{\mu}$*s* estão em forma fechada, e a 2ª 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 $\hat{\beta}$*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 de permutação das celas. Regras gerais que permitem verificar se um modelo hierárquico tem $\hat{\mu}$ em forma fechada são dadas por Goodman (1971a,b) e Haberman (1974, Capítulo 5).

3.6.1 - Modelos hierárquicos possíveis para a classificação tri-dimensional

Os modelos hierárquicos possíveis para a classificação tri-dimensional

de 3 fatores A, B e C podem ser divididos em 9 classes descritas a seguir. Com a exceção do modelo sem a interação dos três fatores, todos os demais modelos hierárquicos têm os $\hat{\mu}$'s em forma fechada.

Seja $Y_{ijk} \sim P(\mu_{ijk})$, o número de observações com $A = i$, $B = j$ e $C = k$, onde $1 \leq i \leq r$, $1 \leq j \leq s$ 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

$$\log \mu_{ijk} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{ij}^{AB} + \beta_{ik}^{AC} + \beta_{jk}^{BC} + \beta_{ijk}^{ABC}, \tag{3.21}$$

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 $\hat{\mu}_{ijk} = y_{ijk}$. As estimativas dos β's são facilmente deduzidas do (3.21) aplicando simples somas em i, j e k. Este modelo corresponde à hipótese de que todo par de fatores varia com o nível do terceiro fator.

A 2ª classe é definida pelo modelo (3.21) com as restrições adicionais $\beta_{ijk}^{ABC} = 0$ para todos os i, j, k, isto é, representa o modelo sem a interação dos três fatores. A média μ_{ijk} não pode ser dada como uma função explícita dos totais marginais μ_{ij+}, μ_{i+k} e μ_{+jk}. Para resolver as equações de MV $\hat{\mu}_{ij+} = y_{ij}, \hat{\mu}_{i+k} = y_{i+k}$ 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 de C, isto é, a razão do produto cruzado condicional $\mu_{ijk}\mu_{i'j'k}/\mu_{ij'k}\mu_{i'jk}$ independe 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 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 equivale à hipótese que os três fatores são independentes: $P(A = i, B = j, C = k) = P(A = i)P(B = j)(C = k)$ 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 MV 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 $\log \mu_{ijk} = \beta + \beta_i^A$; este equivale à hipótese $P(B = j, C = k \mid A = i) = (st)^{-1}$, que dado A, as

combinações de categorias de 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 $\hat{\mu}_{ijk} = y_{+++}/rst$.

3.6.2 - Modelos Hierárquicos para a Classificação Multi-dimensional

Os modelos hierárquicos para a análise de dados categorizados em uma tabela de contingência multi-dimensional $r \times s \times t \times u \ldots$ correspondente à classificação dos fatores $A, B, C, D \ldots$ podem ser deduzidos do modelo saturado, cujos termos são os seguintes: β (média global), $\beta_i^A, \beta_j^B, \ldots$ (efeitos de um fator), $\beta_{ij}^{AB}, \beta_{ik}^{AC}, \ldots$ (efeitos de dois fatores), $\beta_{ijk}^{ABC}, \beta_{ij\ell}^{ABD}, \ldots$ (efeitos de três fatores) $, \ldots, \beta_{ijk\ell\ldots}^{ABCD\cdots}$ (efeitos de todos os fatores). Os modelos hierárquicos para esta classificação são deduzidos do modelo saturado igualando a zero alguns dos parâmetros que medem efeitos dos fatores principais ou de interações dos fatores.

Se o modelo hierárquico tiver os $\hat{\mu}$*ts* em forma fechada as estimativas dos β*ts* serão diretamente obtidas da fórmula do modelo pelas simples somas nos níveis dos fatores.

Os termos de 1ª ordem ou efeitos principais $\beta_i^A, \beta_j^B, \ldots$ refletem às diferenças nas freqüências marginais dos fatores. Os termos $\beta_{ij}^{AB}, \beta_{ik}^{AC}, \ldots$ refletem diferenças nas freqüências das celas devidas às associações entre pares de fatores. Os termos $\beta_{ijk}^{ABC}, \beta_{ij\ell}^{ABD}, \ldots$ refletem associações entre pares de fatores cujas intensidades diferem de acordo com o nível do terceiro fator especificado. Os termos de ordem superior podem ser descritos de forma análoga apesar de apresentarem, muitas vezes, dificuldades de

interpretação, e serem improváveis na prática.

O número de modelos hierárquicos associados à uma tabela multidimensional cresce geometricamente com o número de dimensões de tabela. Apesar da existência de procedimentos do tipo "stepwise" (Seção 1.13) para seleção de modelos, informações a priori sobre a estrutura dos dados, são guias úteis na formulação dos modelos.

Para tabelas de 4 dimensões existem 170 diferentes tipos de modelos. Goodman (1970) determina que 113 modelos são hierárquicos e podem ser classificados em 27 hipóteses diferentes, entre as quais 17 hipóteses têm estimativas $\hat{\mu}$'s em forma fechada.

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 3.2 apresenta alguns modelos hierárquicos com estimativas em forma fechada com seus números de parâmetros independentes associados às classificações $r \times s \times t \times u$ de quatro fatores $A, B, C,$ e D, onde a observação da cela (j, j, k, ℓ) é $y_{ijk\ell}$. Para exemplificar, o modelo correspondente à classe geradora AB, BC, D é $\log \mu_{ijk\ell} = \beta + \beta_i^A + \beta_j^B + \beta_k^C + \beta_\ell^D + \beta_{ij}^{AB} + \beta_{jk}^{BC}$. Os números de parâmetros são facilmente obtidos da forma fechada de $\hat{\mu}_{ijk\ell}$.

Tabela 3.2: *Alguns modelos herárquicos com forma fechada para classificação de 4 fatores* A, B, C e D.

classe geradora	estimativas	número de parâmetros
A, B, C, D	$y_{i+++}y_{+j++}y_{++k+}y_{+++\ell}/y^3_{++++}$	$r+s+t+u-3$
AB, C, D	$y_{ij++}y_{++k+}y_{+++\ell}/y^2_{++++}$	$rs+t+u-2$
AB, CD	$y_{ij++}y_{++k\ell}/y_{++++}$	$rs+tu-1$
AB, BC, D	$y_{ij++}y_{+jk+}y_{+++\ell}/y_{+j++}y_{++++}$	$rs+ts+u-s-1$
AB, BC, CD	$y_{ij++}y_{+jk+}y_{++k\ell}/y_{+j++}y_{++k+}$	$rs+st+tu-s-t$
AB, AC, AD	$y_{ij++}y_{i+k+}y_{i++\ell}/y^2_{i+++}$	$rs+rt+ru-2r$
ABC, D	$y_{ijk+}y_{+++\ell}/y_{++++}$	$rst+u-1$
ABC,AD	$y_{ijk+}y_{i++\ell}/y_{i+++}$	$rst+ru-r$
ABC, ABD	$y_{ijk+}y_{ij+\ell}/y_{ij++}$	$rst+rsu-rs$

§3.7 O Algoritmo de Ajustamento

O modelo log-linear foi definido na Seção 3.5 e corresponde a

$$Y \sim P(\mu), \eta_i = \log \mu_i = \sum_{r=1}^{p} x_{ir}\beta_r, \tag{3.22}$$

onde as quantidades x_{ir} podem ser variáveis explanatórias como na regressão logística (Seções 4.4 e 4.5), ou binárias restritas aos valores 0 e 1 como na análise de tabelas de contingência, e podem ainda ser uma mistura de variáveis explanatórias e binárias.

Sejam $y_1, \ldots, y_n$ as freqüências observadas em uma tabela de contingência com p parâmetros associados. A log-verossimilhança para os parâmetros é dada por

$$L(\beta) = \sum_{i=1}^{n} [-\mu_i + y_i \log \mu_i - \log y_i!], \tag{3.23}$$

cuja derivada reduz-se a

$$\partial L(\beta)/\partial \beta_r = \sum_{i=1}^{n} (y_i - \mu_i) x_{ir}.$$

A função escore $\partial L(\beta)/\partial \beta$, cujas componentes são dadas pela expressão anterior, é escrita como

$$\partial L(\beta)/\partial \beta = X^T (y - \mu) \tag{3.24}$$

e, portanto, as equações de máxima verossimilhança para calcular os $\hat{\beta}$*s* são

$$X^T \hat{\mu} = Xy. \tag{3.25}$$

Inserindo $\log \mu_i$ dado em (3.22) na expressão (3.23) verifica-se que as estatísticas $S_r = \sum_{i=1}^{n} x_{ir} Y_i$, $r = 1, \ldots, p$ são suficientes de dimensão mínima p para os parâmetros β*s*. Sejam $s_r, r = 1, \ldots, p$, os valores observados destas estatísticas. As equações (3.25) podem ser escritas $E(S_r; \hat{\mu}) = s_r$, $r = 1, \ldots, p$, implicando que as estimativas são obtidas igualando os valores observados das estatísticas suficientes aos seus valores esperados. Analogamente, quando os elementos de X são 0 ou 1, as equações (3.25) implicam que as estimativas das médias são obtidas igualando certas freqüências marginais totais aos seus valores esperados.

As equações (3.25) para determinar os β's poderão ser colocadas na forma iterativa

$$X^T W^{(m)} X \beta^{(m+1)} = X^T W^{(m)} y^{*(m)}, \tag{3.26}$$

onde $W = diag\{\mu\}$ e $y^* = \eta + W^{-1}(y - \mu)$, que equivale a fazer repetidamente uma regressão linear ponderada de uma variável dependente modificada y^* sobre X. A demonstração deste processo iterativo num contexto mais amplo é dada na Seção 6.3.

Diferenciando $\partial L(\beta)/\partial\beta_r$ em relação a β_s, multiplicando por -1, achando o valor esperado e escrevendo em notação matricial, obtém-se a matriz de informação de Fisher

$$K = X^T W X. \tag{3.27}$$

A matriz K^{-1} representa a estrutura de covariância assintótica das estatísticas dos parâmetros β's. Para os modelos log-lineares saturados, Bishop, Fienberg e Holland (1975) usam o método delta (Rao, 1973) para calcular K^{-1}. Lee (1977) desenvolve regras gerais para o cálculo de expressões fechadas para os elementos de K^{-1} em modelos log-lineares hierárquicos com forma fechada para os $\hat{\mu}$'s.

A estimativa $\hat{\beta}$ dos parâmetros em (3.22) tem distribuição assintótica $N_p(\hat{\beta}, \hat{K}^{-1})$. sendo $\hat{K}$ a matriz (3.27) avaliada no ponto $\hat{\beta}$. Testes e intervalos de confiança para os parâmetros β's podem ser deduzidos desta 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 baseadas na aproximação normal $\hat{\tau} \sim N(\hat{\tau}, e^T\hat{K}^{-1}e)$. Todos os resultados de distribuições assintóticas e regiões de confiança (Seção 6.6), das técnicas de diagnóstico (Seção 6.7) e do método das covariáveis adicionadas (Seção 6.8) podem ser particularizados aqui para os modelos log-lineares.

§3.8 Testes de Adequação

Para verificar a adequação do ajustamento de um modelo log-linear com p parâmetros independentes aos dados $y_1, \ldots, y_n$, utilizam-se as estatíticas

$$D_p = 2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i), \tag{3.28}$$

$$X_p^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/\hat{\mu}_i. \tag{3.29}$$

A estatística D_p é a razão de verossimilhança (desvio) e X_p^2 é uma forma generalizada da estatística de Pearson. As expressões (3.28) e (3.29) são casos especiais das fórmulas (6.6) e (6.13), respectivamente.

Estas estatísticas são interpretadas como a quantidade de variação nos dados não-explicada pelo modelo. Supondo o modelo verdadeiro, elas convergem em distribuição para a variável χ^2_{n-p}.

A variação total dos dados $\sum_{i=1}^{n} y_i \log(y_i/\overline{y})$ pode ser decomposta em duas parcelas: $E_p = \sum_{i=1}^{n} y_i \log(\hat{\mu}_i/\overline{y})$ correspondente à variação explicada pelo modelo e $D_p = \sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i)$ relativa à variação não-explicada. O quociente $R^2 = E_p/(E_p + D_p)$ é um análogo do coeficiente de determinação do modelo normal-linear (Seção 1.3) e poderá ser utilizado para verificar a qualidade do modelo ajustado.

Gart e Zweifel (1967) sugerem a adição de 0.5 às freqüências observadas em (3.28) e (3.29) para um aperfeiçoamento da aproximação χ^2

de referência. As distribuições de D_p e X^2 se tornam mais próximas da variável χ^2_{n-p}, quando todas as médias $\mu_i's$ crescem e, neste caso, a diferença $\left|X^2_p - D_p\right|$ se torna cada vez menor.

As aproximações das distribuições dessas estatísticas por χ^2 serão bastante razoáveis, quando todos os $\mu_i's$ forem maiores que 5. Estudos de Monte Carlo (Larntz, 1978) sugerem que a estatística (3.28) se comporta de maneira aberrante, quando a tabela tem observações muito pequenas, mas que as duas estatísticas são razoávelmente aproximadas pela variável χ^2, quando o menor valor dos $\mu_i's$ for maior que 1.

Associado às estatísticas D_p e X^2_p define-se os graus de liberdade do modelo log-linear por $\vartheta = n - p$, isto é, como a diferença entre o número de dados e o número de termos $\beta's$ no modelo. Se D_p ou X^2_p for superior a $\chi^2_{n-p}(\alpha)$ o modelo em investigação deverá ser rejeitado. A Tabela 3.3 apresenta os graus de liberdade para os modelos discutidos na Seção 3.1.

A estatística (3.28) é usada para comparação de modelos log-lineares encaixados. Formula-se uma seqüência de interesse de modelos log-lineares encaixados, $M_{p_1}, M_{p_2} \ldots M_{p_r}$, onde cada modelo contém os anteriores como casos especiais, com parâmetros $p_1 < p_2 < \cdots < p_r$ e freqüências estimadas $\hat{\mu}_1, \hat{\mu}_2 \ldots \hat{\mu}_r$, respectivamente.

Os desvios desses modelos satisfazem as desigualdades $D_{p_1} > D_{p_2} > \cdots > D_{p_r}$. Este resultado, em geral, não é verdadeiro para a estatística (3.29).

A diferença entre os desvios dos modelos M_{p_ℓ} e $M_{p_m}(p_m > p_\ell)$ é

$$D_{p_\ell} - D_{p_m} = 2\sum_{i=1}^{n} y_i \log\left(\frac{\hat{\mu}_i^{(m)}}{\hat{\mu}_i^{(\ell)}}\right), \tag{3.30}$$

onde $\hat{\mu}_i^{(\ell)}(\hat{\mu}_i^{(m)})$ é a i-ésima média estimada segundo $M_{p_\ell}(M_{p_m})$. A estatística (3.30) é usada para testar se a diferença entre os valores esperados

ajustados segundo os modelos M_{p_ℓ} e M_{p_m} é, simplesmente, devido a uma variação aleatória, dado que os valores esperados verdadeiros satifazem o modelo M_{p_ℓ}. Se $D_{p_\ell} - D_{p_m} > \chi^2_{p_m - p_\ell}(\alpha)$ os termos extras em M_{p_m} são significativos para explicar os dados.

Toda a discussão de análise do desvio na Seção 6.5 se aplica diretamente aqui para o estudo dos modelos log-lineares encaixados.

Retornando aos modelos M_{p_ℓ} e M_{p_m} encaixados com $p_m > p_\ell$ tem-se:

$$D_{p_\ell} = (D_{p_\ell} - D_{p_m}) + D_{p_m}, \tag{3.31}$$

onde $D_{p_\ell}(D_{p_m})$ é a quantidade de variação dos dados não-explicada por

Tabela 3.3: *Graus de liberdade das estatísticas D_p e X^2_p para os 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 3 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 as combinações de categorias de B e C equiprováveis
9: Nula	$rst-1$	modelo nulo

$M_{p_\ell}(M_{p_m})$ e $D_{p_\ell} - D_{p_m}$ é a variação explicada pelos termos em M_{p_m} que não estão em M_{p_ℓ}. Pode-se demonstrar que $D_{p_\ell} - D_{p_m}$ tem uma forma equivalente a (3.28) sendo expressa por

$$D_{p_\ell} - D_{p_m} = 2\sum_{i=1}^{n} \hat{\mu}_i^{(m)} \log\left(\frac{\hat{\mu}_i^{(m)}}{\hat{\mu}_i^{(\ell)}}\right), \tag{3.32}$$

e, portanto, pode ser interpretada como uma razão de MV condicional para os parâmetros extras que estão em M_{p_m}.

A propriedade de aditividade que é satisfeita por (3.28), sendo decorrente das expressões (3.31) e (3.32), é a base para testar a significância de adicionar termos a um modelo log-linear.

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 $M_{p_e}, M_{p_m}(p_m > p_e)$, esta medida é $(D_{p_e} - D_{p_m})/D_{p_e}$ e representa um índice de qualidade relativa dos ajustamentos dos modelos aos dados. Esta estatística é limitada por 0 e 1; um valor próximo de um sugere que M_{p_m} é muito melhor que M_{p_e}, e um valor próximo de zero é indicativo que os dois modelos proporcionam, aproximadamente, ajustamentos equivalentes.

§3.9 Dois Exemplos de Análise de Dados

3.9.1 - Dados de acidentes de trânsito

A Tabela 3.4 apresenta as freqüências (Y) de acidentes com motoristas, sem acompanhantes, classificadas segundo os fatores: PC = peso do

carro (1− leve, 1− padrão), MF = motorista jogado para fora (1− sim, 2− não), CO = conseqüência do acidente (1− grave, 2− não grave) e AC = tipo de acidente (1− colisão, 2-capotagem). Os ajustamentos de modelos log-lineares hierárquicos a estes dados são apresentados a seguir. Para isto utilizou-se o programa GLIM (Seção 6.3) de ajustamento dos modelos lineares generalizados, embora isso possa ser feito por outros programas: SAS, GENSTAT, BMDP, entre outros.

Tabela 3.4: *Dados de acidentes de trânsito quanto a vários fatores envolvidos descritos em Kihlberg, Narragon e Campbell (1964).*

	CO	1		2	
	AC	1	2	1	2
PC	MF				
1	1	23.00	80.00	26.00	19.00
	2	150.00	112.00	350.00	60.00
2	1	161.00	265.00	111.00	22.00
	2	1022.00	404.00	1878.00	148.00

O ajustamento do modelo com apenas os fatores principais ($PC+MF+AC+CO$) produz $D_5 = 1193.1$. O algoritmo descrito em (3.6) convergiu em 4 iterações. Este modelo é fortemente rejeitado. Adiciona-se ao mesmo todas as interações entre dois fatores . O modelo resultante ($PC+MF+AC+CO+PC.MF+PC.AC+PC.CO+MF.AC+MF.CO+AC.CO$) apresenta razão de verossimilhança $D_{11} = 7.33$ sendo portanto aceito ao nível de 5% ($7.33 < 11.07$). Ajustam-se agora 6 modelos obtidos do modelo anterior pela eliminação de uma única interação entre dois fatores. Os aumentos na razão de verossimilhança $D_{10} - D_{11}$ decorrentes da eliminação de cada interação foram os seguintes: $PM \cdot MF - 1.69, PC \cdot AC - 57.49, PC \cdot CO - 15.58, MF \cdot AC - 220.24, MF \cdot CO) - 114.48, AC \cdot CO - 441.89$. Assim a

única interação não-significativa é $PC \cdot MF$ e, portanto, pode ser eliminada do modelo.

O ajustamento do modelo $PC+MF+AC+CO+PC \cdot AC+PC \cdot CO+MF \cdot AC + MF \cdot CO + AC \cdot CO$ com 10 parâmetros produz $D_{10} = 9.02$ que é inferior ao ponto crítico $\chi^2_6(0.05) = 12.59$. As médias estimadas correspondentes $\hat{\mu}$ estão apresentadas na Tabela 3.5 juntamente com as freqüências observadas. Esta tabela revela uma boa concordância entre $\hat{\mu}$ e y.

Tabela 3.5: *Médias ajustadas segundo o modelo com fatores principais e todas as interações de 2ª ordem exceto $PC \cdot MF$, e valores observados.*

		CO	1		2	
		AC	1	2	1	2
PC	MF					
1	1	y	23.00	80.00	26.00	19.00
		$\hat{\mu}$	23.36	78.07	21.16	13.74
	2	y	150.00	112.00	350.00	60.00
		$\hat{\mu}$	143.29	120.28	361.19	58.91
2	1	y	161.00	265.00	111.00	22.00
		$\hat{\mu}$	166.73	260.83	109.74	33.36
	2	y	1022.00	404.00	1878.00	148.00
		$\hat{\mu}$	1022.61	401.82	1872.91	142.99

3.9.2 - Dados de Preferência por Partido Político

A Tabela 3.6 apresenta os 1257 dados de Butter e Stokes (1957) representando as freqüências de votos classificados pelos seguintes fatores: *P*-partido político (1-conservador, 2-trabalhista), C-classe social (1-média alta, 2-média baixa, 3- operários) e S-sexo (1-homens, 2-mulheres).

Tabela 3.6: *Freqüências observadas de votos por partido político da Grã-Bretanha, por sexo e por classe social.*

	S	1		2	
C	P	1	2	1	2
1		82	30	96	30
2		79	53	101	34
3		118	252	155	227

Um programa GLIM para ajustamento dos seguintes modelos pode ser visto em Cordeiro (1986, Seção 7.1.5):

Modelo	Termos no modelo	Estatística (3.28)	ν
1	$P+S+C$	165.0	7
2	$P+S+C+P\cdot S$	154.2	6
3	$P+S+C+P\cdot C$	13.65	5
4	$P+S+C+S\cdot C$	164.7	5
5	$P+S+C+P\cdot C+S\cdot C$	13.27	3
6	$P+S+C+P\cdot S+P\cdot C$	2.76	4
7	$P+S+C+P\cdot S+S\cdot C$	153.8	4
8	$P+S+C+P\cdot S+P\cdot C+S\cdot C$	1.87	2

Desta relação se conclui que apenas os modelos $P+S+C+P\cdot S+P\cdot C$ e $P+S+C+P\cdot S+P\cdot C+S\cdot C$ são aceitos comparando-os com os qui-quadrados correspondentes. O interesse agora é saber se o termo adicional no 2º modelo (interação $S\cdot C$) implica numa melhoria significativa no ajustamento obtido com o 1ºmodelo. A diferença 0.89 entre as razões de verossimilhança desses modelos é inferior ao valor crítico $\chi^2_2(0.05)=5.99$. Logo, a inclusão da interação $S\cdot C$ no modelo $P+S+C+P\cdot S+P\cdot C$ não produz um aperfeiçoamento substancial no ajustamento.

Para testar a significância de um termo, pode-se considerar vários pares de modelos hierárquicos. Cada par, em geral, produzirá um valor diferente para a estatística-teste do efeito do termo de interesse. O par geralmente é escolhido incluindo um modelo que tenha um bom ajustamento. Para testar o efeito da interação partido×classe ($P\cdot C$), pode-se trabalhar com os seguintes pares de modelos: 1 e 3, 4 e 5, 7 e 8. Neste caso, para qualquer par, o efeito da interação $P\cdot C$ é altamente significativo. Para verificar o efeito no ajustamento ao adicionar a interação partido×sexo ($P\cdot S$), pode-se tomar os pares 1 e 2, 3 e 6, 5 e 8 e, em qualquer caso, o efeito de $P\cdot S$ é significativo. Constata-se, ainda, que o efeito de $S\cdot C$ continuará insignificante se forem comparados os modelos 1 e 4 ou 3 e 5.

Seja a seqüência de modelos encaixados 1,3,6 e 8 com a partição da estatística (3.28): $165.0=(165.0-13.65)+(13.65-2.76)+(2.76-1.87)+1.87=151.35+10.89+0.89+1.87$. Os números $165.0, 151.35, 10.89, 0.89$ e 1.87 representam as quantidades de variação nos dados com os significados respectivos: não-explicada pelo modelo 1, explicada pelo modelo 3, explicada pelo modelo 6, explicada e não-explicada pelo modelo 8. A percentagem da variação nos dados explicada pelo modelo 6 comparando com o modelo 1 é $100(165.0-2.76)/165.0=98\%$. O quociente $100(2.76-1.87)/2.76=32\%$ representa o decréscimo na variação não-explicada pelo modelo 6 quando a

interação $S \cdot C$ é adicionada ao mesmo.

Segundo o modelo $P + S + C + P \cdot S + P \cdot C$ as médias ajustadas (as freqüências observadas correspondentes estão entre parênteses) são: 78.70(82), 32.11(30), 99.30(96), 27.89(30), 79.59(79), 46.56(53), 100.4(101), 40.44(34), 120.7(118), 256.3(252), 152.3(155) e 222.7(227), revelando uma razoável concordância com os valores observados. Destas médias pode-se calcular a razão entre as preferências pelos partidos conservador e trabalhista nas três classes sociais e por sexo dada abaixo:

classe social	homens	mulheres
média alta	2.45	3.56
média baixa	1.71	2.48
operários	0.47	0.68

Assim, a classe operária prefere o partido trabalhista e as classes média alta e média baixa o partido conservador.

§3.10 Exercícios

1. Demonstrar as expressões dos graus de liberdade para os modelos hierárquicos citados na Tabela 3.1. Calcular a estrutura de covariância assintótica das estimativas dos parâmetros β*s* para cada um desses modelos.
2. Demonstrar que (3.28) para o 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: $D_p = 2(\sum_{i,j,k} y_{ijk} \log y_{ijk} -$

$\sum_{j,k} y_{+jk} \log y_{+jk} - \sum_{i,k} y_{i+k} \log y_{i+k} - \sum_{i,j} y_{ij+} \log y_{ij+} + \sum_i y_{i++} \log y_{i++} + \sum_j y_{+j+} \log y_{+j+} + \sum_k y_{++k} \log y_{++k} - y_{+++} \log y_{+++})$, onde $p = IJK - (I-1)(J-1)(K-1)$. Demonstrar que D_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 $\mu_{ijk} = \mu_{+jk}\mu_{i+k}\mu_{ij+}/\mu_{i++}\mu_{+j+}\mu_{++k}$.

3. Demonstrar que os 9 modelos hierárquicos seguintes descritos pelas suas classes geradores e associados a uma tabela de contingência $I \times J \times K \times L$ relativa à classificação de 4 fatores A, B, C e D, não têm forma fechada. Verificar ainda as expressões dos graus de liberdade correspondentes e apresentar interpretações para os modelos

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-2
AB, AC, BC, BD, CD	IJKL-IJ-JK-IK-JL-KL+I+2J+2K+L-1
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).

4. Analisar os dados seguintes relativos aos números de casais dos EUA classificados em 1974 por sexo do cônjuge entrevistado e graus educacionais do homem e da mulher (Haberman, 1978a e Cordeiro, 1986, Seção 7.1.5).

sexo do entrevistado	Grau do Homem	Grau da Mulher		
		primário	2º grau	Graduação ou pós-graduação
homem	primário	135	60	1
	2º grau	43	151	19
	graduação	4	35	12
	pós-graduação	2	24	23
Mulher	primário	124	63	1
	2º grau	39	219	18
	graduação	1	24	26
	pós-graduação	0	17	14

5. 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

6. Seja uma distribuição multinomial com probabilidades $\pi_1 \ldots, \pi_m$ dependendo de um parâmetro θ desconhecido. Considere uma amostra de tamanho n. Calcular a forma das estatísticas de máxima verossimilhança,

escore e de Wald para o teste de $H_0: \theta = \theta^{(0)}$ versus $H: \theta \neq \theta^{(0)}$, onde $\theta^{(0)}$ é um valor especificado.

7. Considere uma tabela de contingência $r \times s$, onde $y_{ij} \sim P(\mu_{ij})$, $i = 1, \ldots, r$, $j = 1, \ldots, s$. Para o teste da hipótese de independência linha-coluna versus uma alternativa geral, calcular a forma das estatísticas escore, de Wald e da razão de MV.

8. 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

9. 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 $E[Y(Y-1)\ldots(Y-r+1)] = \mu^r$; (d) que $2\sqrt{Y}$ é aproximadamente $N(0,1)$.

10. Calcular a forma da matriz de informação para o modelo log-linear associado a uma tabela de contingência com dois fatores sem interação e uma observação por cela.

CAPÍTULO 4

MODELO PARA RESPOSTAS BINÁRIAS

§4.1 Introdução

Em muitas situações práticas, o fenômeno estudado admite apenas dois resultados possíveis, os quais são classificados como "sucesso" ou "fracasso". Geralmente é chamado de "sucesso" o resultado mais importante ou aquele que se pretende relacionar com outras variáveis de interesse. Tais fenômenos, por apresentarem essa característica dicotômica são chamados de binários. Seguem abaixo alguns exemplos.

(i) O resultado de um exame médico, positivo ou negativo;

(ii) A cura ou não, após um período fixo, de um paciente submetido a um certo tratamento;

(iii) Se um componente eletrônico é classificado ou não como defeituoso, após uma inspeção;

(iv) A sobrevivência ou não, após um período fixo, de um animal que recebeu uma dose de uma certa droga.

A resposta de cada fenômeno binário será representada pela variável aleatória y. Comumente atribui-se valor um para y se ocorrer "sucesso" e valor zero se ocorrer "fracasso", sendo a mesma observada em n indivíduos, usualmente supostos independentes.

O principal objetivo deste capítulo é apresentar diversos métodos para avaliar a relação entre a resposta y e diversas variáveis explicativas, as quais podem representar fatores de interesse, grupos de indivíduos ou mesmo variáveis quantitativas. Inicialmente apresenta-se nas Seções 4.2 e 4.3 os principais métodos clássicos para tabelas 2×2, com um forte enfoque na área biomédica. Na Seção 4.4 é introduzido o modelo logístico linear simples e o método de mínimos quadrados para estimação dos parâmetros. O modelo logístico linear múltiplo é discutido de uma forma bastante abrangente na Seção 4.5. Finalmente, outros modelos usuais com respostas binárias são apresentados na Seção 4.6.

§4.2 Distribuição Condicional para uma Única Tabela 2×2

Suponha uma amostra de n indivíduos, dos quais n_1 receberam um novo tratamento e $n_0 = n - n_1$ um tratamento padrão para a cura de uma certa doença. Após um determinado período observou-se o número de indivíduos curados em cada categoria. Cada indivíduo será classificado segundo duas variáveis binárias: (i) x, tipo de tratamento recebido ($x = 0$ tratamento padrão, $x = 1$ novo tratamento) e (ii) y, diagnóstico da doença após o período ($y = 1$ curado, $y = 0$ não curado).

Sejam $p_1 = P(y = 1|x = 1)$ e $p_0 = P(y = 1|x = 0)$; logo, $1 - p_1 = P(y = 0|x = 1)$ e $1 - p_0 = P(y = 0|x = 0)$. Se as observações nos indivíduos são feitas independentemente e as probabilidades p_1 e p_0 não variam de um indivíduo para o outro é razoável supor $y_1 \sim B(n_1, p_1)$ e $y_0 \sim B(n_0, p_0)$, onde y_1 e y_0 representam o total de indivíduos curados

com o novo tratamento e com o tratamento padrão, respectivamente. Os n indivíduos são resumidos segundo a Tabela 4.1.

Tabela 4.1: *Distribuição dos n indivíduos em um estudo de seguimento com um fator de exposição binário.*

Diagnóstico	Tratamento		Total
da doença	Novo	Padrão	
Curado	a	$m-a$	m
Não curado	n_1-a	n_0-m+a	$n-m$
Total	n_1	n_0	n

Nesse tipo de estudo, também conhecido como estudo de seguimento, muito comum na área biomédica, o principal interesse é saber quanto o novo tratamento é melhor que o padrão, ou em outras palavras, qual a chance de um indivíduo que recebeu o novo tratamento ser curado, em relação a um indivíduo que recebeu o tratamento padrão. Essa razão de chances é comumente aproximada pela razão de produtos cruzados, (Cornfield, 1951, 1956) também chamada de risco relativo ("Odds Ratio")

$$\psi = p_1(1-p_0)/\{p_0(1-p_1)\},$$

cuja estimativa não condicional de máxima verossimilhança é dada por

$$\tilde{\psi} = a(n_0-m+a)/\{(n_1-a)(m-a)\}.$$

A medida de associação ψ pode ser estendida para qualquer tabela 2×2 do tipo acima, assim como as interpretações apresentadas. Inferências sobre

ψ são geralmente feitas utilizando-se a distribuição condicional de y_1 dado $y_1 + y_0 = m$, resultando na hipergeométrica não-central (ou generalizada – Seção 3.5) que depende somente de ψ (Cox, 1970)

$$(4.1) \qquad P\{a|n_1, n_0, m; \psi\} = \binom{n_1}{a}\binom{n_0}{m-a}\psi^a \Big/ \sum_s \binom{n_1}{s}\binom{n_0}{m-s}\psi^s,$$

onde o somatório no denominador varia de $s = \max(0, m - n_0)$ até $r = \min(n_1, m)$.

O logaritmo da verossimilhança condicional para ψ fica então dado por

$$(4.2) \qquad L(\psi; a) = a \log \psi - \log P_0(\psi),$$

onde $P_0(\psi) = \sum_s \binom{n_1}{s}\binom{n_0}{m-s}\psi^s$. Portanto, outra estimativa de máxima verossimilhança para ψ pode ser obtida maximizando (4.2), que é equivalente a encontrar a solução positiva da equação

$$(4.3) \qquad a = E\{y_1|\hat{\psi}\} = P_1(\hat{\psi}),$$

onde $E\{y_1|\psi\}$ é o valor esperado da distribuição (4.1) e $P_r(\psi) = \sum_s s^r \binom{n_1}{s}\binom{n_0}{m-s}\psi^s = E\{y_1^r|\psi\}P_0(\psi)$, $r = 1, 2 \ldots$.

Quando os totais marginais $n_1, n_0, n-m$ e m crescem $P_r(\psi)$ define um polinômio de grau alto, cuja solução é obtida somente através de métodos numéricos. Será apresentado na Seção 4.3, um método numérico aproximado para resolver a equação (4.3) sem precisar extrair as raízes do polinômio $P_1(\psi)$. Um algoritmo para calcular a solução exata de (4.3), que envolve poucas operações aritméticas, foi proposto por Gail et al. (1981).

Testes exatos e intervalos de confiança para ψ são geralmente obtidos diretamente de (4.1) para pequenas amostras. Breslow e Day (1980, Cap.

4) apresentam alguns exemplos. Em particular quando $\psi = 1$, que equivale a $p_1 = p_0$, (4.1) reduz-se à distribuição hipergeométrica central (Seção 3.5)

$$P\{a|n_1, n_0, m; \psi = 1\} = \binom{n_1}{a}\binom{n_0}{m-a}\Big/\binom{n_1+n_0}{m}, \tag{4.4}$$

que pode ser utilizada para testar $H{:}\psi = 1$. Nesse caso, se foi observado $y_1 = a$, o nível descritivo do teste para H, também conhecido como teste exato de Fisher (Cox, 1970), é dado por $P = \min\{P_I, P_S\}$, onde

$$P_I = \sum_{u \leq a} P\{u|n_1, n_0, m; \psi = 1\}$$

e

$$P_S = \sum_{u \geq a} P\{u|n_1, n_0, m; \psi = 1\}.$$

Quando os totais marginais da Tabela 4.1 têm valores altos, a distribuição condicional (4.1) pode ser aproximada por uma distribuição normal com média $\mu = \mu(\psi)$ e variância $v = v(\psi)$ (Hannan e Harknen, 1963), dadas por

$$\psi = \mu(n_0 - m + \mu)/\{(n_1 - \mu)(m - \mu)\} \tag{4.5}$$

e

$$v = \{\frac{1}{\mu} + \frac{1}{n_1 - \mu} + \frac{1}{m - \mu} + \frac{1}{n_0 - m + \mu}\}^{-1}. \tag{4.6}$$

Quando $\psi \neq 1$, a equação (4.5) é quadrática em μ, entretanto, somente uma raiz satisfaz a condição $\max(0, m - n_0) \leq \mu \leq \min(n_1, m)$. Essa raiz (Paula, 1982) é sempre expressa na forma

$$\mu = | \, |r| - s \, |,$$

onde $r = [\frac{1}{2}\{(n_1 + n_0)/(\psi - 1) + m + n_1\}]$ e $s = \{r^2 - mn_1\psi/(\psi - 1)\}^{1/2}$.

Para $\psi = 1$ a solução de (4.5) é simplesmente $\mu(1) = mn_1/n$, que coincide com a média exata da distribuição hipergeométrica central (4.4). A variância reduz-se a $v(1) = n_1 n_0 m(n-m)/n^3$, que difere muito pouco da variância exata de (4.4), dada por $v_*(1) = n_1 n_0 m(n-m)/\{n^2(n-1)\}$. Portanto, para valores altos de n_1, n_0, m e $n - m$, hipóteses do tipo $H{:}\,\psi = \psi_0$ podem ser testadas utilizando-se os valores críticos da distribuição normal padrão. Em particular, para $H{:}\,\psi = 1$, as probabilidades P_I e P_S, admitindo-se correção para continuidade (Breslow e Day 1980, pg. 131), serão aproximadas, respectivamente, por

$$P_I \cong 1 - \Phi\left(\frac{a - \mu(1) - 1/2}{v(1)}\right)$$

$$P_S \cong \Phi\left(\frac{a - \mu(1) + 1/2}{v(1)}\right),$$

onde $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão. Analogamente, o teste exato de Fisher será aproximado pela estatística

$$\begin{aligned} X^2 &= (|a - \mu(1)| - 1/2)^2/v(1) \\ &= (|ad - bc| - n/2)^2(n-1)/\{n_1 n_0 m(n-m)\}, \end{aligned}$$

onde $b = m - a$, $c = n_1 - a$ e $d = n_0 - m + a$, que segundo $H{:}\,\psi = 1$ tem aproximadamente distribuição χ^2 com 1 grau de liberdade.

É possível obter também intervalos de $100(1-\alpha)\%$ de confiança, $0 < \alpha < 1$, para ψ. Um intervalo aproximado é formado pelos limites

$$\hat{\psi}_I, \hat{\psi}_S = \tilde{\psi}^{(1 \pm z_{\alpha/2}/x)},$$

onde $z_{\alpha/2}$ é o quantil $(1 - \alpha/2)$ da normal padrão e x é o valor observado de $\sqrt{X^2}$.

Nos estudos de caso e controle, também muito freqüentes na área biomédica, apesar dos indivíduos serem amostrados de forma retrospectiva, todos os resultados descritos anteriormente são também aplicáveis. Em tais estudos, amostra-se n_1 indivíduos, os casos, dentre aqueles portadores de uma certa doença e $n_0 = n - n_1$ indivíduos, os controles, dentre aqueles livres da doença. Os casos são codificados com $y = 1$ e os controles com $y = 0$.

Para cada indivíduo amostrado observa-se a presença ou a ausência de algum fator de interesse, que será representado pela variável x, $x = 0$ ausência e $x = 1$ presença do fator. Define-se então $q_1 = P(x = 1|y = 1)$ e $q_0 = P(x = 1|y = 0)$; logo $1 - q_1 = P(x = 0|y = 1)$ e $1 - q_0 = P(x = 0|y = 0)$. Os n indivíduos amostrados podem ser sumarizados conforme a Tabela 4.2.

Tabela 4.2: *Distribuição dos n indivíduos em um estudo de caso e controle com fator de exposição binário.*

	Fator		
Doença	Presença	Ausência	Total
Presença	a	$n_1 - a$	n_1
Ausência	$m - a$	$n_0 - m + a$	n_0
Total	m	$n - m$	

O principal interesse nesses estudos é estimar a chance de um indivíduo com presença do fator ter a doença em relação a um indivíduo livre do fator. Analogamente aos estudos de seguimento, essa razão de chances é aproximada por

$$\psi = q_1(1 - q_0)/\{q_0(1 - q_1)\},$$

cuja estimativa não condicional de máxima verossimilhança coincide com a estimativa $\tilde{\psi}$.

Aqui também a inferência sobre ψ é feita através de uma distribuição condicional, de x_1 dado $x_1 + x_0 = n_1$, onde x_1 e x_0 são, respectivamente, os números de casos e de controles com presença do fator. Supondo que $x_1 \sim B(m_1, q_1)$ e $x_0 \sim B(m_0, q_0)$ mostra-se que essa distribuição condicional coincide com (4.1). Logo, todos os resultados deduzidos de (4.1) são estendidos para os estudos de caso e controle.

4.2.1 - Exemplo

Para ilustrar alguns dos métodos descritos nesta seção, será utilizado um subconjunto dos dados de Rothman et al. (1979) referentes a um estudo de caso e controle, cujo objetivo principal foi avaliar a associação entre a exposição a certas drogas por parte da mãe, antes e durante a gravidez, e a ocorrência de doenças congênitas do coração no recém-nascido.

Os casos consistiram de 390 mães que tiveram crianças com doenças congênitas do coração durante o período de 1973 a 1975 no estado de Massachusetts (E.U.A.). Os controles consistiram de 1254 mães selecionadas aleatoriamente dentre aquelas que tiveram crianças nesse período naquele estado, livres desse tipo de doença. A Tabela 4.3 apresenta a distribuição dos casos e controles segundo a exposição ou não por parte da mãe à droga Diazepan.

Tabela 4.3: *Distribuição dos 390 casos e dos 1254 controles segundo a exposição à droga* Diazepan.

	Expostos	Não Expostos	Total
Casos	15	375	390
Controles	22	1232	1254
Total	37	1607	1644

Fonte: Rothman et al. (1979).

Tem-se $a = 15, m_1 = 390, m_0 = 1254$ e $n_1 = 37$. A estimativa não condicional de máxima verossimilhança vale $\tilde{\psi} = (15 \times 1232)/(22 \times 375) = 2.24$. A obtenção da estimativa condicional envolve a extração das raízes de um polinômio de grau $r = 37$, que é proibitivo computacionalmente.

A estatística X^2, adaptada ao estudo de caso e controle, vale $X^2 = (|15 \times 1232 - 22 \times 375| - 1644/2)^2(1644-1)/(37 \times 1607 \times 390 \times 1254) = 5.00$, cujo nível descritivo é $P = 0.013$.

Um intervalo de 95% de confiança para ψ será formado pelos seguintes limites:

$$\hat{\psi}_I = \hat{\psi}^{(1-1.96/x)} = 2.24^{(1-0.876)} = 1.105$$
$$\hat{\psi}_S = \hat{\psi}^{(1+1.96/x)} = 2.24^{(1+0.876)} = 4.540,$$

onde x é o valor observado de $\sqrt{X^2}$.

§4.3 Combinação de Tabelas 2×2

Suponha agora que as variáveis x e y definidas na seção anterior sejam observadas em k estratos, que podem representar, por exemplo, k faixas

etárias. Na i-ésima faixa etária aplica-se o tratamento novo a um número n_{1i} de indivíduos e o tratamento padrão em $n_{0i} = n_i - n_{1i}$ indivíduos, observando-se após o período os curados y_{1i} e y_{0i} com o novo tratamento e com o tratamento padrão, respectivamente. Na área biomédica, a inclusão de estratos visa principalmente evitar que a associação entre a resposta y e o fator x seja distorcida, em virtude da não inclusão de algum outro fator que esteja associado com x e y. Esses fatores são comumente chamados de fatores de confundimento e a forma mais eficaz de controle dos mesmos é através da utilização de algum modelo para prever $P\{y = 1\}$. Nesses modelos, que serão discutidos nas próximas seções, todos os fatores suspeitos de confundimento são controlados simultaneamente.

Serão denotadas por p_{1i} e p_{0i} as probabilidades de cura com os tratamentos novo e padrão, respectivamente, no i-ésimo estrato. Supondo que as observações são feitas de forma independente e que as probabilidades p_{1i} e p_{0i} não variam de um indivíduo para o outro, é muito razoável supor $y_{1i} \sim B(n_{1i}, p_{1i})$ e $y_{0i} \sim B(n_{0i}, p_{0i})$. O risco relativo no i-ésimo estrato é aproximado por

$$\psi_i = p_{1i}(1 - p_{0i})/\{p_{0i}(1 - p_{1i})\},$$

sendo $\tilde{\psi}_i = a_i(n_{0i} - m_i + a)/\{(n_{1i} - a_i)(m_i - a_i)\}$ a correspondente estimativa de máxima verossimilhança não-condicional.

Aqui também será utilizada a distribuição condicional de y_{1i} dado $y_{1i} + y_{0i} = m_i$ denotada por $P\{a_i | n_{1i}, n_{0i}, m_i; \psi_i\}$, assim como o logaritmo da verossimilhança condicional para ψ_i

$$L(\psi_i; a_i) = a_i \log \psi_i - \log P_0^{(i)}(\psi_i),$$

onde $P_0^{(i)}(\psi_i)$ é o polinômio $P_0(\psi)$ definido na seção anterior, restrito ao i-ésimo estrato.

4.3.1 - Testes para os Riscos Relativos

O primeiro passo no caso estratificado é verificar se os riscos relativos $\psi_i ts$ são homogêneos nos estratos, que equivale a testar $H: \psi_1 = \psi_2 = \cdots = \psi_k$. Sob essa hipótese, o logaritmo da função de verossimilhança conjunta para o risco relativo comum ψ, fica expresso na forma

$$\log \psi \sum_i a_i - \sum_i \log\{P_0^i(\psi)\}.$$

Diferenciando a expressão acima em ψ, a estimativa de máxima verossimilhança $\hat{\psi}$ é calculada da equação

$$\sum_i a_i = \sum_i E(y_{1i}/\hat{\psi}). \tag{4.7}$$

Infelizmente $E(y_{i1}|\psi)$, $i = 1, \ldots, k$, é não-linear em ψ, o que exige a utilização de algum método iterativo para se obter $\hat{\psi}$. Aplicando o método de Newton-Raphson chega-se ao seguinte processo iterativo para resolver (4.7):

$$\log \psi^{(m+1)} = \log \psi^{(m)} - V^{-1}(\psi^{(m)})\{a - E(\psi^{(m)})\}, \tag{4.8}$$

$m = 0, 1, 2, \ldots$, onde $E(\psi) = \sum_i \mu_i(\psi)$, $V(\psi) = \sum_i v_i(\psi)$, $a = \sum_i a_i$, $\mu_i(\psi) = E\{y_{1i}|\psi\}$ e $v_i(\psi) = \mathrm{Var}\{y_{1i}|\psi\}$.

É importante lembrar que μ_i e v_i dependem dos polinômios $P_0^{(i)}(\psi)$, $P_1^{(i)}(\psi)$ e $P_2^{(i)}(\psi)$, cujos graus dependem dos totais marginais da i-ésima tabela. Logo, tem-se $3k$ polinômios para serem resolvidos em cada passo do processo iterativo (4.8), o que aumenta a complexidade computacional do mesmo. A estimativa de Mantel-Haenszel

$$\hat{\psi}_{MH} = \sum_i \{a_i(n_{0i} - m_i + a_i)/n_i\} / \sum_i \{n_{1i} - a_i)(m_i - a_i)/n_i\},$$

é geralmente utilizada como valor inicial $\psi^{(0)}$.

McDullagh e Nelder (1983, Cap. 4) sugerem um método numérico para obter $V(\psi)$ e $E(\psi)$ em cada passo de (4.8), o qual dispensa a extração das raízes dos polinômios $P_0^{(i)}(\psi)$, $P_1^{(i)}(\psi)$ e $P_2^{(i)}(\psi)$. Esse método baseia-se na relação abaixo entre a média $\mu = E\{y_1|\psi\}$ e a variância $v = \mathrm{Var}\{y_1|\psi\}$ de (4.1) (vide Mantel e Hankey, 1975).

$$\{\mu(n_0 - m + \mu) + v\}/\{(n_1 - m)(m - \mu) + v\} = \psi. \tag{4.9}$$

Para v e ψ fixos tem-se portanto uma equação quadrática em μ. Logo, conhecendo $v^{(0)}$ e $\psi^{(0)}$, obtém-se $\mu^{(0)}$ facilmente de (4.9). Um valor inicial para v pode ser a variância condicional de y_1 quando $\psi = 1$, dada por

$$v^{(0)} = n_1 n_0 m(n - m)/\{n^2(n - 1)\}. \tag{4.10}$$

Assim, no passo inicial de (4.8) tem-se $E(\psi^{(0)}) = \sum_i \mu_i^{(0)}$ e $V(\psi^{(0)}) = \sum_i v_i^{(0)}$, onde $\mu_i^{(0)}$ e $v_i^{(0)}$, $i = 1, \ldots, k$, são calculados de (4.9) e (4.10), respectivamente. Substituindo esses valores no processo iterativo (4.8) obtém-se $\psi^{(1)}$.

A estimativa de v_i nessa iteração e nas demais, pode ser obtida da aproximação assintótica para a variância condicional de y_1, dada em (4.6). Segundo McDullagh e Nelder (1983, Cap. 5) a aproximação numérica descrita acima produz resultados satisfatórios quando os totais marginais são pelo menos iguais a 5 em cada estrato. Para pequenas amostras, o algoritmo proposto por Gail et al. (1981), que calcula a estimativa de máxima verossimilhança exata para ψ, pode ser utilizado a um custo computacional relativamente baixo.

Como foi visto na seção anterior, a estimativa condicional de máxima verossimilhança para ψ_i sai da equação não-linear

$$a_i = E\{y_{1i}|\hat{\psi}_i\}, \quad i = 1, \ldots, k,$$

que depende dos polinômios $P_0^{(i)}(\psi)$ e $P_1^{(i)}(\psi)$.

Analogamente a $\hat{\psi}$, a estimativa $\hat{\psi}_i$ pode também ser obtida através de um processo iterativo similar a (4.8), dado por

$$(4.11) \quad \log \psi_i^{(m+1)} = \log \psi_i^{(m)} - \{v_i^{(m)}\}^{-1}\{a_i - \mu_i^{(m)}\}, \quad m = 0, 1, 2, \ldots.$$

O método numérico proposto para o cálculo de $\mu_i^{(m)}$ e $v_i^{(m)}$ em cada passo de (4.8) é também aplicável em (4.11), evitando assim a extração das raízes dos polinômios $P_0^{(i)}(\psi)$ e $P_1^{(i)}(\psi)$. A estimativa não condicional $\tilde{\psi}_i$ é geralmente utilizada para iniciar (4.11). Logo, a estatística da razão de máxima verossimilhança para testar H, fica expressa na forma

$$-2\left[\sum_i a_i \log(\hat{\psi}/\hat{\psi}_i) + \sum_i \log\{P_0^{(i)}(\hat{\psi}_i)/P_0^{(i)}(\hat{\psi})\}\right].$$

Sob H e para $n \to \infty$, onde $n = \sum_i n_i$, essa estatística tem aproximadamente distribuição χ^2 com $(k-1)$ graus de liberdade.

Uma estatística alternativa (vide Breslow e Day, 1980, pg. 142) para testar a homogeneidade dos riscos relativos nos estratos é dada por

$$(4.12) \qquad X^2 = \sum_i (a_i - \mu_i(\hat{\psi}))^2 / v_i(\hat{\psi}),$$

onde $\hat{\psi}$ é a estimativa de máxima verossimilhança sob a hipótese de homogeneidade. Se n é muito maior que o número de estratos, (4.12) pode ser aproximada por uma χ^2 com $(k-1)$ graus de liberdade. Muitos pesquisadores têm utilizado $\hat{\psi}_{MH}$ no lugar de $\hat{\psi}$ em (4.12), com resultados bastante satisfatórios. As estimativas $\mu_i(\hat{\psi})$ e $v_i(\hat{\psi})$ são obtidas de (4.5) e (4.6), respectivamente.

Se a hipótese H não é rejeitada tem-se um mesmo risco relativo ψ em todos os estratos. Testes de hipóteses e intervalos de confiança para ψ podem ser desenvolvidos utilizando-se a distribuição condicional

$$\sum_i P\{a_i|n_{1i}, n_{0i}, m_i; \psi\}.$$

Para o caso particular de $\psi = 1$, tem-se em cada estrato uma distribuição hipergeométrica central. Um teste aproximado para $H: \psi = 1$, que utiliza a aproximação normal para a distribuição condicional de y_{1i}, $i = 1 \ldots, k$, foi desenvolvido por Mantel e Haenszel (1959), e é dado por

$$X^2 = \left(\left| \sum_i a_i - \sum_i \mu_i(1) \right| - 1/2 \right)^2 \Big/ \sum_i v_i(1). \tag{4.13}$$

Sob H, a estatística acima pode ser aproximada por uma χ^2 com 1 grau de liberdade. Um intervalo de confiança usual para ψ de $100(1-\alpha)\%$ de confiança que utiliza a estatística de Mantel-Haenszel $\hat{\psi}_{MH}$, é formado pelos limites

$$\hat{\psi}_I, \hat{\psi}_S = \hat{\psi}_{MH}^{(1 \pm z_{\alpha/2}/x)},$$

onde x é o valor observado de $\sqrt{X^2}$.

4.3.2 - Exemplo

Na Tabela 4.4 são apresentados os dados de um estudo de seguimento para avaliar a associação entre drogas orais anti-diabéticas e morte cardiovascular. Dois grupos de indivíduos receberam tratamentos diferentes durante um tempo fixo e nesse período verificou-se em cada grupo o número de mortes. Um grupo recebeu o tratamento com Tolbutamide e o outro com Placebo, e os indivíduos foram subdivididos em duas faixas etárias, até 55 anos e mais de 55 anos.

Tabela 4.4: *Distribuição dos indivíduos do experimento com Tolbutamide e Placebo.*

Tratamento	Até 55 anos		Mais de 55 anos	
	Mortos	Sobreviventes	Mortos	Sobreviventes
Tolbutamide	8	98	22	76
Placebo	5	115	16	69

Fonte: Rothman (1977).

A estimativa de Mantel-Haenszel é dada por

$$\hat{\psi}_{MH} = (8 \times 115/226 + 22 \times 69/183)/(98 \times 5/226 + 76 \times 16/183) = 1.403.$$

De (4.5) e (4.6) tem-se $\mu_1(\hat{\psi}_{MH}) = 7.132$, $\mu_2(\hat{\psi}_{MH}) = 22.856$, $v_1(\hat{\psi}_{MH}) = 3.035$ e $v_2(\hat{\psi}_{MH}) = 7.278$.

Logo, a estatística X^2 para testar a homogeneidade dos riscos relativos nas duas faixas-etárias vale

$$X^2 = (8 - 7.132)^2/3.035 + (22 - 22.856)^2/7.278 = 0.349,$$

que é não-significativa ao nível de 10%. O valor da estatística (4.13) para testar a homogeneidade dos dois tratamentos, $H: \psi = 1$, onde ψ é o risco relativo comum nas duas faixas-etárias, é dado por

$$X^2 = \{8 + 22 - (6.097 + 20.350)\}/(3.051 + 7.489) = 0.337,$$

que também é não-signiticativo ao nível de 10% . Portanto, os dois tratamentos são homogêneos na causa de mortes cardiovasculares nas duas faixas-etárias.

§4.4 Modelo Logístico Linear Simples

O modelo logístico linear é derivado da função matemática

$$f(x) = (1 + e^{-x})^{-1}, -\infty < x < \infty,$$

que varia monotonicamente de 0 a 1 à medida que x cresce, sendo simétrica em torno de $x = f^{-1}(1/2)$. O termo linear refere-se à propriedade da transformação logit em $f(x)$

$$\text{logit } f(x) = \log[f(x)/\{1 - f(x)\}] = x,$$

que é linear em x. Além disso, a forma sigmoidal de $f(x)$, muito parecida com a forma de diversas curvas de dose e resposta, nas quais para cada dose x tem-se uma resposta, geralmente uma proporção no intervalo $[0, 1]$, sugere a utilização de $f(x)$ para ajustar tais curvas.

Suponha inicialmente uma situação bastante simples que já foi descrita na Seção 4.2, onde se tentou estudar a associação entre uma resposta binária representada por y (por exemplo $y = 1$ curado e $y = 0$ não-curado) e um fator dicotômico, representado por x (por exemplo, $x = 0$ tratamento padrão e $x = 1$ novo tratamento). Naquela situação foram apresentados alguns métodos clássicos para avaliar essa associação. Aqui será utillizada uma estrutura mais complexa que consiste na imposição de um modelo de regressão (logístico linear) para prever a probabilidade $p(x) = P\{y = 1|x\}$. Esse modelo é dado por

$$\text{logit}p(x) = \eta$$

ou

$$\begin{aligned} p(x) &= \exp(\eta)/\{1 + \exp(\eta)\} \\ &= \{1 + \exp(-\eta)\}^{-1}, \end{aligned} \tag{4.14}$$

onde $\eta = \alpha + \beta x$ é chamado de preditor linear e α e β são parâmetros desconhecidos a serem estimados.

Desse modelo segue que $p(0) = P\{y = 1|x = 0\} = \exp(\alpha)/\{1 + \exp(\alpha)\}$, $p(1) = P\{y = 1|x = 1\} = \exp(\alpha+\beta)/\{1+\exp(\alpha+\beta)\}$, $1-p(0) = P\{y = 0|x = 0\} = \{1 + \exp(\alpha)\}^{-1}$ e $1 - p(1) = p\{y = 0/x = 1\} = \{1+\exp(\alpha+\beta)\}^{-1}$. Portanto, a razão de produtos cruzados (risco relativo) é agora expressa em função do parâmetro β

$$\psi = p(1)\{1 - p(0)\}/[p(0)\{1 - p(1)\}] = \exp(\beta)$$

ou

$$\log \psi = \beta.$$

A variável explicativa x pode também representar a doses de uma determinada droga que é aplicados em algum tipo de animal e verificado após um período fixo se esse animal sobreviveu ($y = 0$) ou não ($y = 1$). Nesse caso, faz sentido calcular o risco de um animal que a recebeu a dose x^*, não sobreviver, em relação à um animal que recebeu a dose x. Esse risco é aproximado por

$$\psi = p(x^*)\{1 - p(x)\}/[p(x)\{1 - p(x^*)\}] = \exp\{\beta(x^* - x)\}$$

ou

$$\log \psi = \beta(x^* - x).$$

Portanto, o logaritmo do risco relativo ψ é proporcional à diferença entre as doses. Se $\beta > 0$ o risco de não sobrevivência cresce com o aumento da dose e se $\beta < 0$ ocorre o contrário. A seguir é apresentada uma forma bastante simples de estimar os parâmetros α e β.

4.4.1 - Estimativas de Mínimos Quadrados para α e β.

Sem perda de generalidade, suponha que k doses $x_1, x_2, \ldots, x_k$ são aplicadas, respectivamente, em $n_1, n_2, \ldots, n_k$ animais, observando-se após um período fixo o número de animais y_i que não sobreviveram com a i-ésima dose x_i, $i = 1, \ldots, k$.

Seja a transformação logit

$$\text{logit } \tilde{p}_i = \log\{\tilde{p}_i/(1-\tilde{p}_i)\},$$

onde $\tilde{p}_i = \tilde{p}(x_i) = y_i/n_i$ é a proporção de animais que morreram com a dose x_i. Supondo que $y_i \sim B(n_i, p_i)$, onde $p_i = p(x_i)$, para n_i grande mostra-se que (Bickel e Doksum, 1977, Cap. 8)

$$\sqrt{n_i}[\log\{\tilde{p}_i/(1-\tilde{p}_i)\} - (\alpha + \beta x_i)]$$

segue uma distribuição aproximadamente normal com média 0 e variância $\sigma_i^2 = \sigma^2(p_i)$ dada por

$$\sigma_i^2 = \{p_i(1-p_i)\}^{-1}.$$

Chamando $z_i =$ logit $\tilde{p}_i$, então para n_i grande

$$z_i = \alpha + \beta x_i + \varepsilon_i', \tag{4.15}$$

onde $\varepsilon_i' \sim N(0, \sigma_i^2/n_i)$. Logo, se as drogas são aplicadas independentemente tem-se um problema de regressão linear com variâncias desiguais e dependentes de parâmetros desconhecidos. Para σ_i^2 conhecido o modelo de regressão linear (4.15) pode ser expresso numa forma mais conveniente

$$\frac{z_i}{w_i} = \alpha\frac{1}{w_i} + \beta\frac{x_i}{w_i} + \varepsilon_i,$$

onde $w_i = \sigma_i/\sqrt{n_i}$ e $\varepsilon_i \sim N(0,1)$, que permite a aplicação direta do método de mínimos quadrados descrito no Capítulo 1, para estimar α e β.

Entretanto, como em geral σ_i^2 é desconhecido, é razoável substituir w_i pela correspondente estimativa de máxima verossimilhança

$$\tilde{w}_i = \{n_i\tilde{p}_i(1-\tilde{p}_i)\}^{-1/2}.$$

Dessa forma, as estimativas de mínimos quadrados para α e β ficam dadas por

$$\tilde{\beta} = \sum s_i z_i(x_i - \tilde{x})/\sum s_i(x_i - \tilde{x})^2,\ \tilde{\alpha} = \tilde{z} - \tilde{\beta}\tilde{x},$$

onde $\tilde{z} = \sum z_i s_i/\sum s_i$, $\tilde{x} = \sum x_i s_i/\sum s_i$ e

$$s_i = \tilde{w}_i^{-2} = n_i\tilde{p}_i(1-\tilde{p}_i).$$

Bickel e Doksum (1977, Cap. 8) sugerem a utilização da transformação $\log\{(y_i+0.5)/(n_i-y_i+0.5)\}$ no lugar de logit $\tilde{p}_i$ para evitar dificuldades nos casos em que todos os animais submetidos à dose x_i sobreviverem ($y_i = 0$) ou morrerem ($y_i = n_i$). Tratando os $\tilde{w}_i$'s como constantes é possível aplicar a teoria clássica de regressão linear para o caso binomial acima. Por exemplo, se as suposições feitas são corretas, a variância de $\tilde{\beta}$ pode ser estimada por

$$\mathrm{Var}(\tilde{\beta}) = \{\sum s_i(x_i - \tilde{x})^2\}^{-1}.$$

Logo, um intervalo de $100(1-\alpha)\%$ de confiança para a taxa β fica dado por $\tilde{\beta} \pm z_{\alpha/2}\{\mathrm{Var}(\tilde{\beta})\}^{1/2}$, onde $z_{\alpha/2}$ representa o quantil $(1-\alpha/2)$ de uma normal padrão. Esses intervalos podem ser utilizados para testar $H\colon \beta = 0$. Nesse caso, rejeita-se a hipótese H se e somente se

$$|\tilde{\beta}|/\{\mathrm{Var}(\tilde{\beta})\}^{1/2} > z_{\alpha/2}.$$

4.4.2 - Exemplo

Serão utilizados agora parte dos dados de um estudo (Paula et al., 1988) desenvolvido para avaliar o efeito de diversos extratos vegetais na mortalidade de embriões de Biomphalaria Glabrata (hospedeiro da equistossomose). Em particular para o extrato vegetal aquoso quente de folhas de L. Nobilis, foram utilizados $k = 6$ grupos de embriões. Cada um dos n_i embriões do i-ésimo grupo recebeu uma dose x_i (em partes por milhão) do extrato vegetal, observando-se após o 20º dia o número y_i de embriões mortos. Esses dados, bem como as estimativas s_i, $\tilde{p}_i$ e z_i, são resumidos na Tabela 4.5.

Tabela 4.5: *Resultado do experimento com o extrato vegetal de folhas de L. Nobilis na morte do hospedeiro da equistossomose.*

x_i	n_i	y_i	$\tilde{p}_i$	z_i	s_i
0	295	14	0.0474	-3.00	13.32
50	273	30	0.1099	-2.09	26.70
100	301	75	0.2491	-1.10	56.30
150	365	263	0.7205	0.95	73.50
200	272	244	0.8970	2.16	26.13
300	379	378	0.9974	5.95	0.98

Fonte: Paula et al. (1988).

Utilizando os resultados da Tabela 4.5, obtém-se as estimativas $\tilde{\beta} = 0.029$ e $\tilde{\alpha} = -3.568$. Logo, o modelo logístico linear ajustado fica dado por

$$\text{logit}\, \hat{p}(x) = -3.568 + 0.029x$$

ou

$$\hat{p}(x) = \exp(-3.568 + 0.029x)/\{\exp(-3.568 + 0.029x)\},$$

onde $p(x)$ é a probabilidade de um embrião submetido à dose x não sobreviver após o 20º dia.

A Figura 4.1 exibe o gráfico dos valores observados $\tilde{p}(x_i)$ e ajustados $\hat{p}(x_i)$ contra as doses x_i, $i = 1, \ldots, 6$. Dos dados da Tabela 4.5 obtém-se também $\tilde{\beta}/\{\text{Var}(\tilde{\beta})\}^{1/2} = 22.32$. Logo, para um nível de significância de 1% rejeita-se a hipótese $H: \beta = 0$. Nesses estudos, um dos objetivos é estimar a dose letal que mata 50% dos embriões, denotada por $LD(50)$. Do modelo logístico linear (4.14) mostra-se facilmente que essa dose é estimada por $LD(50) = -\tilde{\alpha}/\tilde{\beta} = 123.03$.

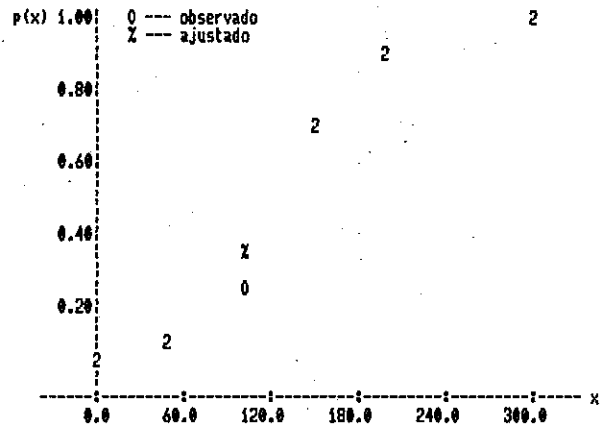

Figura 4.1: *Gráfico das freqüências relativas y_i/n_i e das probabilidades ajustadas $\hat{p}(x_i)$ contra as doses x_i do extrato vegetal L. Nobilis*

§4.5 Modelo Logístico Linear Múltiplo

Em geral nos problemas com respostas binárias o principal interesse é es-

tudar a relação entre a resposta y e diversas variáveis explicativas, as quais podem representar fatores de interesse, fatores suspeitos de confundimento, variáveis quantitativas, ou mesmo estratos. Suponha inicialmente um modelo logístico linear com duas variáveis explicativas dicotômicas, denotadas por x_2 e x_3. Por exemplo com x_2 representando os pacientes submetidos a um tratamento padrão ($x_2 = 0$) e novo ($x_2 = 1$), respectivamente, e x_3 um fator de interesse ($x_3 = 0$ ausente e $x_3 = 1$ presente).

O modelo logístico linear, sem interação, fica expresso na forma

$$\text{logit}p(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3, \tag{4.16}$$

onde $p(x) = P\{y = 1|x\}$ e $x = (x_2, x_3)^T$. Para facilitar a interpretação dos parâmetros de (4.16) as probabilidades $p(x)$ serão denotadas por $P_{ij} = P\{y = 1|x_2 = i, x_3 = j\}$, $i, j = 0, 1$. Tem-se de (4.16),

$$\begin{aligned} &P_{00} = \exp(\beta_1)/\{1 + \exp(\beta_1)\}, P_{10} = \exp(\beta_2)/\{1 + \exp(\beta_2)\}, \\ &P_{01} = \exp(\beta_3)/\{1 + \exp(\beta_3)\} \quad \text{e} \\ &P_{11} = \exp(\beta_2 + \beta_3)/\{1 + \exp(\beta_2 + \beta_3)\}. \end{aligned} \tag{4.17}$$

A chance de um paciente na condição $(i, j)^T$, ser curado, em relação a um paciente na condição $(0, 0)^T$, será aproximada por

$$\psi_{ij} = P_{ij}(1 - P_{00})/\{P_{00}(1 - P_{ij})\}, \quad i, j = 0, 1.$$

Substituindo as probabilidades descritas em (4.17) na expressão acima, obtém-se

$$\psi_{10} = \exp(\beta_2), \psi_{01} = \exp(\beta_3) \text{ e } \psi_{11} = \exp(\beta_2 + \beta_3). \tag{4.18}$$

Portanto, a hipótese de não-interação entre x_2 e x_3 assumida no modelo (4.16), leva à relação multiplicativa entre os riscos relativos.

$$\psi_{11} = \psi_{10}\psi_{01}.$$

Considere agora o modelo logístico linear saturado

$$\text{logit } p(x) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2 x_3. \tag{4.19}$$

Os riscos relativos ψ_{10} e ψ_{01} continuam expressos como em (4.18), entretanto ψ_{11} dependerá também do parâmetro β_4

$$\psi_{11} = \exp(\beta_2 + \beta_3 + \beta_4).$$

Logo,

$$\log\{\psi_{10} \cdot \psi_{01} | \psi_{11}\} = \beta_4.$$

Assim, testar a hipótese $H: \beta_4 = 0$ (interação nula) em (4.19) equivale a testar a hipótese multiplicativa $H: \psi_{11} = \psi_{10} \cdot \psi_{01}$.

Se a variável x_3 representa um determinado estrato, o risco relativo (chance relativa) de cura, entre o tratamento novo e o padrão no primeiro estrato ($x_3 = 0$) será dado por $\psi_{10} = \exp(\beta_2)$. Esse risco relativo no segundo estrato ($x_3 = 1$) vale $\psi_{11}/\psi_{01} = \exp(\beta_2 + \beta_4)$. Logo, $H: \beta_4 = 0$ equivale também à hipótese de homogeneidade dos riscos relativos nos estratos.

Como foi mostrado na seção anterior, os experimentos de dose e resposta, quando a resposta é uma proporção, podem ser analisados através de um modelo logístico linear simples. Entretanto, em tais experimentos, se a resposta refere-se à cura ($y = 1$) ou não ($y = 0$) de uma certa doença, geralmente a eficiência da droga decresce após uma certa dosagem, em virtude da toxidade da mesma, como é mostrado pela Figura 4.2.

Nesse caso a relação entre a probabilidade de cura $p(x)$ e a dose x pode ser aproximada pelo modelo

$$p(x) = [1 + \exp\{-(\alpha + \beta x + \gamma x^2)\}]^{-1},$$

onde $p(x) = P\{y = 1|x\}$, e α, β e γ são parâmetros interpretáveis. O parâmetro α está relacionado com a probabilidade de cura quando $x = 0$, β

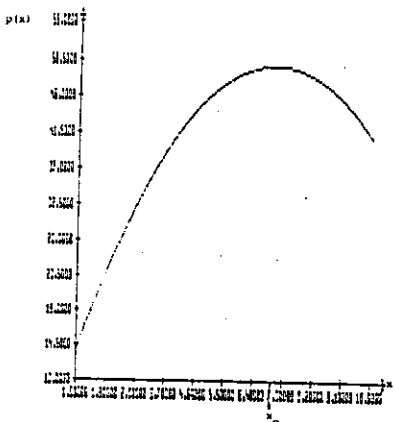

Figura 4.2: *Descrição de uma curva de dose e resposta, com toxidade da droga após uma dosagem* x_0.

está relacionado com a taxa de cura da droga, enquanto γ está relacionado com a toxidade da mesma.

Finalmente, considere uma situação em que a resposta binária anterior seja observada segundo um fator dicotômico $x_1(x_1 = 1$ presença, $x_1 = 0$ ausência) em k estratos. Os k estratos serão representado por $k-1$ variáveis dicotômicas, $x_2, \ldots, x_k$, tais que $x_j = 1$ se a resposta foi observada no j-ésimo estrato e $x_j = 0$ em caso contrário, $j = 2, \ldots, k$.

O modelo logístico linear, nesse caso, fica dado por

$$p(x) = \{1 + \exp(-\eta)\}^{-1}, \tag{4.20}$$

onde $\eta = \alpha + \beta_1 x_1 + \sum_{j=2}^{k} \beta_j x_j + \sum_{j=2}^{k} \gamma_j x_1 x_j, P(x) = P\{y = 1|x\}$ e $x = (x_1, \ldots, x_k)^T$. Chamando de ψ_j o risco relativo de cura no j-ésimo estrato, entre um indivíduo com presença do fator e um indivíduo sem o fator, obtém-se

$$\psi_1 = \exp(\beta_1) \text{ e } \psi_j = \exp(\beta_1 + \gamma_j),\ j = 2, \ldots, k.$$

Note que $\log\{\psi_j/\psi_1\} = \gamma_j$, logo γ_j pode ser interpretado como um incremento no logaritmo do risco relativo do j-ésimo estrato em relação ao primeiro estrato. Portanto, testar a homogeneidade dos riscos relativos nos estratos equivale a testar $H: \gamma_2 = \cdots = \gamma_k = 0$. Se essa hipótese não é rejeitada o modelo (4.20) fica expresso numa forma mais simples

$$p(x) = [1 + \exp\{-(\alpha_j + \beta_1 x_1)\}]^{-1}, \tag{4.21}$$

onde $\alpha_1 = \alpha$ e $\alpha_j = \alpha + \beta_j$, $j = 2, \ldots, k$.

4.5.1 - Estimação dos Parâmetros pelo Método de Máxima Verossimilhança

O modelo logístico linear múltiplo pode sempre ser expresso na forma geral

$$p(x) = \{1 + \exp(-\eta)\}^{-1},$$

onde $\eta = x^T\beta$ é um proditor linear, $x = (x_1, \ldots, x_p)^T$ representa p variáveis explicativas, incluindo interações, e $\beta = (\beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros a ser estimado.

Suponha que a resposta y seja observada em n indivíduos subdivididos em k grupos. No i-ésimo grupo, representado por $x_i = (x_{i1}, \ldots, x_{ip})^T$, são observados n_i indivíduos, sendo denotada por y_{ij} a resposta do j-ésimo indivíduo, $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$. Se no i-ésimo grupo as observações são independentes e se a probabilidade $p(x_i)$ não varia de um indivíduo para outro, então é razoável supor $y_i = \sum_j y_{ij} \sim B(n_i, p_i)$, onde $p_i = p(x_i)$, $i = 1, \ldots, k$. Para $n_1, \ldots, n_k$ grandes o método de mínimos quadrados descrito na Seção 4.4, pode ser estendido para o caso múltiplo, assim como os resultados assintóticos apresentados no Capítulo 1.

Contudo, será discutido nesta seção o método usual de máxima verossimilhança para estimação dos parâmetros de β, que consiste na maximização

do logaritmo da função de verossimilhança de β, dado por

$$L(\beta) = \sum_i \sum_j y_i x_{ij} \beta_j - \{\sum_i n_i \log(1 + \exp \sum_j x_{ij} \beta_j)\}.$$

Seja $U(\beta) = \partial L(\beta)/\partial \beta$ a função escore para β. A estimativa de máxima verossimilhança $\hat{\beta}$ é obtida de $U(\hat{\beta}) = 0$, que equivale a resolver o seguinte sistema de equações não-lineares:

$$X^T(y - \mu(\hat{\beta})) = 0,$$

onde X é uma matriz $n \times p$ de linhas x_i^T, $y = (y_1, \ldots, y_k)^T$, $\mu(\beta) = (\mu_1, \ldots, \mu_k)^T$, $\mu_i = n_i p_i = n_i \exp(\eta_i)/\{1 + \exp(\eta_i)\}$, $\eta_i = x_i^T \beta$, $i = 1, \ldots, k$.

O sistema acima para obter a estimativa de máxima verossimilhança $\hat{\beta}$, exige a utilização de um processo iterativo. Aplicando o método de Newton-Raphson chega-se ao seguinte algoritmo:

$$(4.22) \quad \beta^{(m+1)} = \beta^{(m)} + (X^T W^{(m)} X)^{-1} X^T W^{(m)} y^{*(m)}, \; m = 0, 1, 2, \ldots,$$

onde $W = \text{diag}\{w_1, \ldots, w_k\}$, $y^* = (y_1^*, \ldots, y_k^*)^T$, $w_i = n_i p_i (1 - p_i)$ e $y_i^* = \eta_i + (y_i - n_i p_i)/\{n_i p_i (1 - p_i)\}$, $i = 1, \ldots, k$.

O processo iterativo (4.22) geralmente é iniciado em $\beta^{(0)} = 0$ que equivale a $p_i^{(0)} = 1/2$, $i = 1, \ldots, k$. Na Seção 6.3, esse processo é estendido para o cálculo das estimativas nos modelos lineares generalizados.

4.5.2 - Função Desvio e Resultados Assintóticos

Para avaliar a qualidade do ajuste de um modelo logístico linear, utiliza-se frequentemente a função desvio ("deviance") introduzida por Nelder e Wedderburn (1972), que será discutida na Seção 6.4. Essa função para

o caso binomial é expressa por

$$D(y;\hat{\mu}) = 2\sum_i [y_i \log(y_i/\hat{\mu}_i) + (n_i - y_i)\log\{(n_i - y_i)/(n_i - \hat{\mu}_i)\}], \tag{4.23}$$

onde $\hat{\mu} = \mu(\hat{\beta})$.

Finalmente, sob a hipótese de que o modelo assumido é verdadeiro, (4.23) tem aproximadamente distribuição χ^2 com $(k-p)$ graus de liberdade em duas situações muito frequentes na prática: (i) quando $N \to \infty$, onde $N = \min\{n_1, \ldots, n_k\}$, k mantendo-se fixo; (ii) quando $k \to \infty$ e os n_i's se mantém fixos. Similarmente, nessas duas situações, $(\hat{\beta} - \beta)$ tem aproximadamente distribuição normal p- variada com média zero e estrutura de covariância dada por $K^{-1} = (X^T W X)^{-1}$. Logo, denotando-se $K^{-1} = \{-k^{ij}\}$, intervalos de confiança para β_j podem ser formados da maneira usual. Por exemplo, um intervalo de $100(1-\alpha)\%$ de confiança para β_j será dado por

$$\hat{\beta}_j \pm t_{\alpha/2}(-k^{jj})^{1/2} ,$$

onde $t_{\alpha/2}$ é o quantil $(1-\alpha/2)$ de uma t de Student com $(k-p)$ graus de liberdade.

Dos resultados acima, um intervalo assintótico de $100(1-\alpha)\%$ de confiança para o risco relativo $\psi(x)$, será formado pelos limites

$$\exp[\log\hat{\psi}(x) \pm z_{\alpha/2} \times \{\mathrm{Var}(\log\hat{\psi}(x))\}^{1/2}],$$

onde $\mathrm{Var}(\log\hat{\psi}(x)) = x^T \hat{K}^{-1} x$.

4.5.3 - Testes de Hipóteses

Serão apresentados nesta seção as estatísticas da razão de máxima verossimilhança, de Wald e escore para testar a significância dos parâmetros

de um modelo logístico linear. Essas três estatísticas, apesar de serem expressas em formas diferentes, têm assintoticamente a mesma distribuição segundo a hipótese nula.

Suponha então que o vetor de parâmetros β seja particionado na forma $\beta = (\beta_q^T \beta_{p-q}^T)^T$, $q < p$, e considere a hipótese H: $\beta_q = 0$. As três estatísticas acima, que serão descritas a seguir, têm assintoticamente, sob H, distribuição χ^2 com q graus de liberdade. Maiores detalhes dessas estatísticas podem ser vistos em Cox e Hinkley (1979).

A. Estatística da razão de máxima verossimilhança

Definida por

$$-2\{L(0, \hat{\beta}_{p-q}^{(0)}) - L(\hat{\beta}_q, \hat{\beta}_{p-q})\},$$

onde $\hat{\beta}_{p-q}^{(0)}$ é a estimativa de máxima verossimilhança, sob H, de β_{p-q}. A estatística acima pode ser expressa como a diferença entre as funções desvio do modelo sob H e do modelo com p parâmetros

$$D_{p-q}(y; \hat{\mu}^{(0)}) - D_p(y; \hat{\mu}),$$

onde $\hat{\mu}^{(0)}$ é a estimativa de máxima verossimilhança, sob H, de $\mu = \mu(\beta)$. Essas funções serão denotadas por D_{p-q} e D_p.

B. Estatística de Wald

Definida por

$$W = \hat{\beta}_q^T \hat{K}_{qq}^{-1} \hat{\beta}_q,$$

onde $\hat{K}_{pq}^{-1} = K_{qq}^{-1}(\hat{\beta})$ é uma submatriz $q \times q$ de $\hat{K}^{-1}$ com as estimativas das variâncias e covariâncias dos elementos de $\hat{\beta}_q$.

C. Estatística escore

Definida por

$$E = U^T(\hat{\beta}^{(0)}) K^{-1}(\hat{\beta}^{(0)}) U(\hat{\beta}^{(0)}),$$

onde $U(\hat{\beta}^{(0)})$ e $K^{-1}(\hat{\beta}^{(0)})$ são, respectivamente, $U(\beta)$ e $K^{-1}(\beta)$ avaliados em $\hat{\beta}^{(0)} = (O^T, \hat{\beta}_{p-q}^{(0)T})^T$.

Alguns casos especiais são de interesse. Em particular para testar $H: \beta_j = 0$, a estatística de Wald fica simplificada na forma

$$W = \hat{\beta}_j^2/(-k^{jj}),$$

que assintoticamente pode ser aproximada por uma χ^2 com 1 grau de liberdade.

Day e Byar (1979) mostraram que a estatística escore para testar $H: \beta_1 = 0$ no modelo (4.21) coincide com a estatística de Mantel-Haenszel dada em (4.13).

4.5.4 - Seleção de Covariáveis

Os métodos de seleção de covariáveis da regressão linear não são diretamente aplicáveis ao modelo logístico-linear. Agora o principal objetivo não é mais selecionar um conjunto de covariáveis que explique bem a resposta e também produza boa precisão. No modelo logístico linear os parâmetros são interpretáveis e muitas vezes para que um efeito seja bem entendido é necessário manter no modelo interações não significativas envolvendo o mesmo. É o caso do princípio hierárquico, discutido na Seção 3.6, em que a manutenção de uma certa interação implica na manutenção de todas as subinterações, assim como dos efeitos principais correspondentes. Como as interações envolvendo quatro ou mais fatores quase sempre são de difícil interpretação e podem causar singularidade na matriz do modelo, é comum restringir-se à seleção de covariáveis que envolvam no máximo interações com três fatores.

Nem sempre nas aplicações do modelo logístico linear todas as interações possíveis de uma mesma ordem são consideradas. Isso ocorre em

particular na área Epidemiológica (vide Kleinbaum et al., 1982), em que se tem fatores de exposição e fatores de confundimento, sendo consideradas apenas aquelas interações entre esses fatores que são facilmente interpretáveis e que tenham algum significado no problema.

Analogamente à regressão linear, não há um critério ótimo de seleção de covariáveis para o modelo logístico linear. Entretanto, será apresentado a seguir um procedimento simples, que pode ser aplicado em muitas situações práticas. Esse procedimento tem os seguintes passos:

1ºPasso

Considere como modelo inicial aquele incluindo todos os fatores principais e todas as interações de interesse envolvendo dois fatores. Por exemplo, suponha a existência de apenas um fator de exposição, denotado por E, e cinco fatores $C_1, C_2, \ldots, C_5$ suspeitos de confundimento. Alguns desses cinco fatores podem representar interações. Por exemplo pode-se ter $C_5 = C_1C_2$. Então esse modelo deve incluir os efeitos principais $E, C_1, \ldots, C_5$ e as interações $EC_1, EC_2, \ldots, EC_5$. Se esses fatores são dicotômicos o modelo fica expresso na forma

$$\begin{aligned} \text{logit } p = {} & \alpha + \beta E + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4 + \beta_5 C_5 \\ & + \gamma_1 EC_1 + \gamma_2 EC_2 + \gamma_3 EC_3 + \gamma_4 EC_4 + \gamma_5 EC_5. \end{aligned} \tag{4.24}$$

2º Passo

Utilizando alguma das estatísticas descritas na seção anterior, testar a inclusão individual de cada uma das interações $EC_1C_2, EC_1C_3, \ldots, EC_4C_5$ no modelo (4.24). Incluir no modelo apenas aquelas interações que forem significativas e que não causarem singularidade na matriz do modelo.

3º Passo

Testar agora a exclusão de cada interação envolvendo dois fatores que não seja subinteração das interações selecionadas no passo anterior. Por

exemplo, se no 2º passo foram selecionadas as interações EC_1C_2 e EC_2C_3, então pelo princípio hierárquico devem ser mantidas no modelo as interações EC_1, EC_2 e EC_3 e os efeitos principais E, C_1, C_2 e C_3. Logo, sobram para serem testadas as interações EC_4 e EC_5.

4º Passo

Suponha que a interação EC_4 foi considerada não-significativa, então o modelo adotado no 3º passo é

(4.25)

$$\text{logit}p = \alpha + \beta E + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4 + \beta_5 C_5 + \gamma_1 EC_1 + \gamma_2 EC_2 + \gamma_3 EC_3 + \gamma_5 EC_5 + \delta_1 EC_1C_2 + \delta_2 EC_2C_3.$$

Respeitando o princípio hierárquico, testar a exlcusão dos efeitos principais que sobrarem (se existirem). No caso acima deve-se testar a exclusão de C_4 no modelo (4.25).

4.5.5 - Resíduos

A função desvio definida na Seção 4.5.2, que é utilizada para avaliar a qualidade do ajuste, não fornece informações sobre os afastamentos do modelo ajustado. Esses afastamentos são usualmente investigados através de resíduos, utilizando-se gráficos.

Dentre os resíduos mais conhecidos, destaca-se o resíduo de Pearson, definido por

$$r_i = (y_i - n_i\hat{p}_i)/\{n_i\hat{p}_i(1-\hat{p}_i)\}^{1/2},$$

$i = 1, \ldots, k$. Esse resíduo é muito usual para n_i grande, entretanto se n_i é pequeno ou se p_i é próximo de 0 ou 1, a i-ésima componente da função desvio, definida por $\pm d_i^{1/2}$ sendo

$$d_i = [2y_i \log(y_i/n\hat{p}_i) + 2(n_i - y_i)\log\{(n_i - y_i)/(y_i - n_i\hat{p}_i)\}]$$

é mais recomendada. O sinal de $d_i^{1/2}$ é o mesmo de $y_i - n_i\hat{p}_i$.

As componentes $\pm d_i^{1/2}$'s são geralmente melhores aproximadas pela distribuição normal que os resíduos r_i's. Entretanto na situação extrema em que $n_i = 1$, é mais conveniente utilizar esses resíduos agrupando-se os dados. Landwehr et al. (1984) apresentam um método gráfico, que envolve a simulação da distribuição dos resíduos, para essa situação.

Utilizando Cox e Snell (1968) mostra-se que uma padronização mais adequada para r_i, que leva em conta a variabilidade de $\hat{p}_i$, é dada por $r_i/\{1-h_{ii}\}^{1/2}$, onde h_{ii} é o i-ésimo elemento da diagonal principal da matriz $H = \hat{W}^{1/2}X(X^T\hat{W}X)^{-1}X^T\hat{W}^{1/2}$. Analogamente $\pm d_i^{1/2}/\{1-h_{ii}\}^{1/2}$ se aproxima melhor da normal padrão que $\pm d_i^{1/2}$.

Williams (1987) mostra que a variação na função desvio após a exclusão da i-ésima observação é aproximada por

$$r_{Gi}^2 = (1-h_{ii})\{d_i/(1-h_{ii})\} + h_{ii}\{r_i^2/(1-h_{ii})\},$$

e define o resíduo $\pm r_{Gi}$, onde o sinal de r_{Gi} é o mesmo de $y_i - n_i\hat{p}_i$. Williams (1987) sugere a utilização de r_{Gi} no gráfico normal de probabilidades, que é usual para detectar possíveis afastamentos das suposições feitas para as variáveis y_i's.

Outros métodos de diagnóstico serão descritos no Capítulo 6, os quais podem ser particularizados para os modelos discutidos neste capítulo.

4.5.6 - Exemplo

Para ilustrar os métodos descritos nesta seção, serão utilizados os dados apresentados em Paula et al. (1984) referentes a um estudo de caso e controle realizado no setor de Anatomia e Patologia do hospital Heliópolis em São Paulo, entre 1970 e 1982. A finalidade desse estudo foi avaliar a as-

sociação entre algumas variáveis histológicas e o tipo de processo infeccioso pulmonar, maligno ou benigno.

Foram considerados como casos todos os pacientes diagnosticados entre 1970 e 1982 nesse hospital, que tiveram o diagnóstico confirmado de processo maligno (71 pacientes). Os controles consistiram de uma amostra de 104 pacientes de um total de 270, que também foram diagnosticados no mesmo período no referido hospital e tiveram confirmado o diganóstico infeccioso pulmonar benigno.

Diversas variáveis histológicas foram observadas em fragmentos de tecidos retirados da região pulmonar de cada um dos 175 pacientes. Essas análises, entretanto, serão restritas a duas variáveis histológicas (dois tipos de células), histiócitos-linfócitos e fibrose frouxa e a duas variáveis potenciais de confundimento, sexo e idade. Portanto, tem-se as seguintes variáveis: (i) tipo do processo ($y = 1$ maligno , $y = 0$ benigno), (ii) HL, histiócitos-linfócitos ($HL = 1$ presença , $HL = 0$ ausência), (iii) FF, fibrose frouxa ($FF = 1$ presença , $FF = 0$ ausência), (iv) SEXO (SEXO $= 1$ masculino , SEXO $= 0$ feminino) e (v) IDADE (em anos).

Como nos estudos de caso e controle os indivíduos são amostrados retrospectivamente, ou seja, dado y, o modelo logístico para explicar $p(x) = P\{y = 1|x\}$ é expresso numa forma um pouco diferente dos estudos de seguimento (vide McDullagh e Nelder, 1983, pg. 78), dada por

$$\text{logit } p(x) = \alpha^* + \eta,$$

onde $\eta = x^T\beta$, $\alpha^* = \log(\pi_0/\pi_1)$, $\pi_0 = P\{Z = 1|y = 0\}$, $\pi_1 = P\{Z = 1|y = 1\}$ e Z é uma variável dicotômica representando a presença ($Z = 1$) ou ausência ($Z = 0$) de um indivíduo da população na amostra de casos e controles. Os parâmetros de β são estimados diretamente de (4.22), enquanto α^*, se π_1 e π_0 forem conhecidos, pode ser introduzido posteriormente. Os

riscos relativos ficam invariantes com a introdução de α^*.

Será aplicado a seguir o método de seleção de covariáveis descrito na Seção 4.5.4. Como o interesse aqui não é prever $p(x)$ e sim estudar a associação entre o tipo de processo e as variáveis histológicas, serão ajustados modelos logísticos sem a constante α^*. O modelo inicial, modelo I, é dado por

$$\begin{aligned}\text{logit } p = \alpha &+ \beta_1 * \text{ SEXO } + \beta_2 * \text{ IDADE } + \beta_3 * HL + \beta_4 * FF \\ &+ \gamma_1 * HL * \text{ SEXO } + \gamma_2 * HL * \text{ IDADE } + \gamma_3 * FF * \text{ SEXO} \\ &+ \gamma_4 * FF * \text{ IDADE } .\end{aligned}$$

Serão adicionadas individualmente as interações $HL*$ SEXO $*$ IDADE e $FF*$ SEXO $*$ IDADE no modelo I. Como as interações envolvendo três fatores são em geral de difícil interpretação, será adotado um nível de significância de 5% nos testes de inclusão das mesmas. Nos demais testes será adotado um nível de 10%. Em Kupper et al. (1976) há uma discussão sobre o controle do nível de significância na seleção de covariáveis em modelos de regressão. Pelos resultados da Tabela 4.6 nenhuma das interações acima é significativa ao nível de 5%; logo, não serão incluídas no modelo I.

Tabela 4.6: *Valores da função desvio com a inclusão individual das interações $HL*SEXO*IDADE$ e $FF*SEXO*IDADE$ no modelo I.*

Modelo	Graus de liberdade	Desvio	Diferença
Modelo I	166	151.28	–
Modelo I+ HL*IDADE*SEXO	165	147.55	3.73
Modelo I+ $HL*$IDADE$*$SEXO	165	148.50	2.78

O passo seguinte consiste em testar a exclusão de cada interação envolvendo dois fatores do modelo I. Os resultados são descritos na Tabela 4.7 e pode-se observar que somente a interação $HL *$ IDADE é significativa a 10%. Dentre as restantes, $HL *$ SEXO tem o maior nível descritivo, logo deve ser excluída do modelo 1.

Tabela 4.7: *Valores da função desvio com a exclusão individual do modelo I das interações envolvendo dois fatores.*

Modelo	Graus de liberdade	Desvio	Diferença
Modelo I	166	151.28	-
Modelo I $-HL *$ IDADE	167	155.23	3.95*
Modelo I-$HL *$ SEXO	167	152.23	0.95
Modelo I-$FF *$ IDADE	167	152.90	1.62
Modelo I-$FF *$ SEXO	167	153.61	2.33

*: significativo a 10%.

O modelo resultante, modelo II, é dado por

$$\text{logit}\, p = \alpha + \beta_1 * \text{SEXO} + \beta_2 * \text{IDADE} + \beta_3 * HL + \beta_4 * FF + \gamma_2 * HL * \text{IDADE} + \gamma_3 * FF * \text{SEXO} + \gamma_4 * FF * \text{IDADE}.$$

Testa-se agora a exclusão de cada interação envolvendo dois fatores do modelo II.

Pelos resultados, Tabela 4.8, ao nível de 10% a interação $FF *$ IDADE deve ser excluída do modelo II.

Chega-se portanto ao modelo III dado por

$$\text{logit}\, p = \alpha + \beta_1 * \text{SEXO} + \beta_2 * \text{IDADE} + \beta_3 * HL + \beta_4 * FF + \gamma_2 * HL * \text{IDADE} + \gamma_3 * FF * \text{SEXO}.$$

Tabela 4.8: *Valores da função desvio com a exclusão individual do modelo II das interações $HL * IDADE, FF * IDADE$ e $FF * SEXO$.*

Modelo	Graus de liberdade	Desvio	Diferença
Modelo II	167	152.23	–
Modelo II-$HL *$ IDADE	168	156.57	4.34*
Modelo II-$FF *$ IDADE	168	153.96	1.73
Modelo II-$FF *$ SEXO	168	156.06	3.63*

*: significativo a 10%.

O passo seguinte consiste em testar a exlcusão das interações $HL *$ IDADE e $FF *$ SEXO do modelo III. Pelos resultados descritos na Tabela 4.9, nenhuma desses interações deve ser excluída.

Tabela 4.9: *Resultados referentes à exclusão individual das interações $HL * IDADE$ e $FF * SEXO$ do modelo III.*

Modelo	Graus de liberdade	Desvio	Diferença
Modelo III	168	153.96	–
Modelo III-$HL *$ IDADE	169	160.16	6.20*
Modelo III-$HL *$ SEXO	169	157.49	3.53

*: significativo a 10%.

Logo, como não sobra nenhum efeito principal (pelo princípio hierárquico), o modelo III será adotado como modelo final. As estimativas dos parâmetros e dos desvios padrões assintóticos são apresentadas na Tabela 4.10.

Tabela 4.10: *Estimativas dos parâmetros do modelo III.*

Efeito	Parâmetro	Estimativas
Constante	α	-10.204
		(0.967)
Sexo	β_1	-1.239
		(0.544)
Idade	β_2	0.036
		(0.016)
HL	β_3	-5.163
		(1.631)
FF	β_4	-2.797
		(1.378)
HL*Idade	γ_2	0.068
		(0.029)
FF*Sexo	γ_3	2.487
		(1.439)

( ): desvio padrão assintótico

Seja ψ_{HL} o risco de um paciente com presença de HL, estar com processo maligno, em relação a um paciente sem a presença da célula. Supondo que esses pacientes tem o mesmo sexo, a mesma idade e o mesmo nível de FF, então ψ_{HL} é estimado por

$$\hat{\psi}_{HL} = \exp\{-5.163 + 0.068 * \text{ IDADE }\}. \tag{4.26}$$

De (2.26) conclui-se que o risco de processo maligno é em geral menor para os pacientes com presença da célula que para os pacientes com ausência

da mesma. Esse risco relativo cresce, entretanto, com o aumento da idade dos pacientes, ficando próximo de 1 quando os pacientes têm aproximadamente 75 anos.

Para ilustrar, suponha que dois pacientes de 40 anos e mesmo sexo tenham sido submetidos a exames no hospital Heliópolis para diagnosticar o tipo de processo infeccioso pulmonar. Após os exames, constatou-se o mesmo nível de FF para ambos os pacientes, entretanto apenas um apresentou presença da célula HL. Logo, o risco do paciente, cujo exame não detectou a célula HL, estar com processo maligno, em relação ao outro paciente, é aproximado por

$$\hat{\psi}_{HL}^{-1} = \exp(5.163 - 0.068 \times 40) = 11.51.$$

Um intervalo do $100(1-\alpha)\%$ de confiança para ψ_{HL} é formado pelos limites

$$\hat{\psi}_{HL}^{I}, \hat{\psi}_{HL}^{S} = \exp\{\log \hat{\psi}_{HL} \pm 1.96 \times \mathrm{Var}(\log \hat{\psi}_{HL})\},$$

onde $\mathrm{Var}(\log \hat{\psi}_{HL}) = \mathrm{Var}(\hat{\alpha}) + (IDADE)^2 \mathrm{Var}(\hat{\gamma}_2) + 2(IDADE)Cov(\hat{\alpha}, \hat{\gamma}_2)$.

Analogamente seja ψ_{FF} o risco de um paciente com presença da célula FF, estar com processo maligno, em relação a um paciente com ausência da mesma. Supondo que esses pacientes são semelhantes nas demais variáveis, ψ_{FF} pode ser estimado por

$$\hat{\psi}_{FF} = \exp(-2.797 + 2.487 * SEXO).$$

Logo, se são comparados pacientes do sexo feminino esse risco vale aproximadamente $\exp(-2.797) = 0.06$. Para pacientes do sexo masculino, o risco é estimado por $\exp(-2.797+2.487) = 0.73$. Isso mostra que a presença da célula FF é mais importante entre os pacientes do sexo feminino do que entre os pacientes do sexo masculino. Diversos outros riscos relativos podem

ser estimados da Tabela 4.10, e conseqüentemente outras interpretações serão obtidas.

§4.6 Outros Modelos

A seguir são apresentados outros modelos para análise de dados binários, como os modelos probit e complementar log-log, modelo logístico condicional e modelos logísticos não-lineares.

4.6.1 - Modelos Probit e Complementar Log-Log

Suponha novamente a variável explicativa x representando a dose de uma certa droga e uma resposta y, denotando a sobrevivência ($y = 0$) ou não ($y = 1$) de um animal após um determinado período em que o mesmo recebeu uma dosagem da droga. Como foi visto na Seção 4.4.1, o modelo logístico linear pode ser aplicado para ajustar curvas resultantes desses experimentos. Entretanto, há situações em que o mesmo não é o mais adequado, havendo dois outros modelos que podem ajustar melhor os dados. Esses modelos são geralmente menos usuais que o logístico, em virtude de não possibilitarem a interpretação direta dos parâmetros.

O primeiro desses modelos, o probit, é definido por

$$\Phi^{-1}(p(x)) = \eta, \tag{4.27}$$

onde $\eta = \alpha + \beta x$ e $\Phi^{-1}(p)$ é a inversa da função acumulativa da distribuição normal-padrão. O outro modelo, o complementar log-log, é expresso na forma

$$\log\{1 - \log(1 - p(x))\} = \eta, \tag{4.28}$$

onde $\log\{1-\log(1-p)\}$ é a função acumulativa da distribuição do valor extremo.

Os modelos logístico e probit são simétricos em torno de $p = 0.5$ e muito semelhantes no intervalo $0.1 \leq p \leq 0.9$. O complementar $\log - \log$ apesar de não ser simétrico, é muito parecido com o logístico para valores pequenos de p, e à medida que p tende para 1 o mesmo tende mais lentamente para infinito que os outros dois.

Os parâmetros α e β dos modelos probit e complementar $\log - \log$ podem ser estimados analogamente ao logístico, pelo método de mínimos quadrados descrito na Seção 4.4.1. Para k doses diferentes da droga, aplicadas respectivamente a $n_1, \ldots, n_k$ animais, tem-se as transformações empíricas $\Phi^{-1}(\tilde{p}_i)$ e $\log\{1-\log(1-\tilde{p}_i)\}$, onde $\tilde{p}_i = y_i/n_i,\ i = 1, \ldots, k$.

Exceto as expressões das variâncias $\sigma^2(p_i)$ que devem mudar (vide Bickel e Doksum,1977, Cap. 7) todas as demais expressões da Seção 4.4.1 são válidas para os modelos probit e complementar $\log - \log$. Com o preditor linear na forma geral $\eta_i = x_i^T\beta,\ i = 1, \ldots, k$, a estimativa de máxima verossimilhança $\hat{\beta}$ impondo-se (4.27) ou (4.28) é também obtida de um processo iterativo na forma (4.22) com pesos w's e variáveis dependentes modificadas y^*'s, respectivamente, dados por

$$w = n\Phi'^2/\{p(1-p)\},\ y^* = \eta + (y-np)/(n\Phi') \text{ com } \phi' = d\phi(\eta)/d\eta$$

e

$$w = n(1-p)\log^2\{(1-p)/p\},\ y^* = \eta + (y-np)/[n(1-p)\{1-\log(1-p)\}].$$

Em Paula et al. (1988) são comparados os ajustes dos modelos logístico, probit e complementar $\log - \log$ em 10 experimentos do tipo dose e resposta.

4.6.2 - Modelo Logístico Condicional

Em alguns estudos de caso e controle ou de seguimento o número de estratos formados pode ser relativamente grande. Isso ocorre em particular nos estudos emparelhados de caso e controle, em que a influência de fatores suspeitos de confundimento é controlada através de emparelhamentos de casos com controles, segundo alguns níveis desses fatores. Para cada emparelhamento tem-se um estrato. Nessas situações, se é adotado um modelo logístico linear, além dos parâmetros correspondentes aos efeitos incluídos no modelo, tem-se uma constante α para cada estrato formado. Nos casos de estratos com poucas observações, o número de parâmetros a serem estimados podem ser da ordem do número total de observações, o que em geral leva a estimativas sèriamente viesadas (Cox e Hinkley, 1979, pg. 292).

Para ilustrar, suponha um estudo de caso e controle com k emparelhamento $1-1$ (1 caso por 1 controle) segundo k níveis de um fator suspeito de confundimento. Seja a resposta denotada por $y(y = 1$ caso, $y = 0$ controle) e suponha ainda uma variável dicotômica x representando um certo fator ($x = 1$ presença, $x = 0$ ausência). Se é adotado um modelo logístico linear com riscos relativos constantes nos estratos formados, tem-se

$$\text{logit } p_i(x) = \alpha_i + \beta x, \tag{4.29}$$

onde $p_i(x) = P\{y_i = 1|x\}$, $i = 1, \ldots, k$.

A forma mais usual de evitar o problema mencionado acima, consiste na eliminação dos parâmetros α_i's através de uma distribuição condicional apropriada, similar àquela dada em (4.1) (vide Prentice e Breslow, 1978).

Suponha que no i-ésimo estrato formado, que tem $n_i = 2$ elementos, observou-se para a variável explicativa x os valores x_{i1} e x_{i2}, sem ser especificado qual desses valores corresponde ao caso e qual corresponde ao

controle. Portanto, tem-se uma tabela 2×2 com as margens fixas. A distribuição condicional de que x_{i1} de fato corresponde ao caso e x_{i2} corresponde ao controle é dada por

$$\frac{P(x_{i1}|y_i = 1)P(x_{i2}|y_i = 0)}{P(x_{i1}|y_i = 1)P(x_{i2}|y_i = 0) + P(x_{i2}|y_i = 1)P((x_{i1}|y_i = 0)} \tag{4.30}$$

Utilizando a relação $P(x|y) = P(y|x)P(x)/P(y)$ e (4.29), mostra-se facilmente que (4.30) fica simplificado na forma

$$\exp(\beta x_{i1})/\{\exp(\beta x_{i1}) + \exp(\beta x_{i2})\}.$$

Portanto, para k estratos a função de verossimilhança conjunta para β é dada por

$$\prod_{i=1}^{k}[\exp(\beta x_{i1})/\{\exp(\beta x_{i1}) + \exp(\beta x_{i2})\}], \tag{4.31}$$

onde x_{i1} é o valor de x observado para o i-ésimo caso.

Generalizando, suponha um total de k emparelhamentos do tipo $1 - M$ (1 caso com M controles) e p variáveis explicativas, denotadas por $x_1, \ldots, x_p$, que podem representar interações. Seja o modelo logístico

$$\text{logit}p_i(x) = \alpha_i + x^T\beta, \tag{4.32}$$

onde $x = (x_1, \ldots, x_p)^T$ e $\beta = (\beta_1, \ldots, \beta_p)^T$, $i = 1, \ldots, k$. Considere ainda que no i-ésimo estrato formado foram observados os conjuntos de valores $x_{i0}^T, x_{i1}^T, \ldots, x_{iM}^T$, onde $x_{i\ell}^T = (x_{i\ell 1}, \ldots, x_{i\ell p})$, $\ell = 0, \ldots, M$, para os $n_i = 1 + M$ casos e controles, sem ser especificado qual desses conjuntos corresponde ao caso e quais aos controles. Logo, a distribuição condicional de que o

primeiro conjunto x_{i0}^T de fato pertence ao caso e os demais aos controles, fica agora dada por

$$\frac{P(x_{i0}|y_i=1)\prod_{\ell=1}^{M}P(x_{i\ell}|y_i=0)}{\sum_{j=0}^{M}P(x_{ij}|y_i=1)\prod_{\substack{\ell=0\\ \ell\neq j}}P(x_{i\ell}|y_i=0)}. \tag{4.33}$$

Utilizando o teorema de Bayes e o modelo (4.32) em (4.33), a distribuição condicional conjunta, em função de β, fica expressa na forma

$$\begin{aligned}&\prod_{i=1}^{k}\left\{\exp(x_{i0}^T\beta)/\sum_{\ell=0}^{M}\exp(x_{i\ell}^T\beta)\right\}\\ &=\prod_{i=1}^{k}\left[1+\sum_{\ell=1}^{M}\exp\{(x_{i\ell}-x_{i0})^T\beta\}\right]^{-1}.\end{aligned} \tag{4.34}$$

Aplicando o logaritmo em (4.34), obtém-se

$$L(\beta)=\sum_{i=1}^{k}x_{i0}^T\beta-\log\{\sum_{\ell}\exp(x_{i\ell}^T\beta)\}. \tag{4.35}$$

A estimativa de máxima verossimilhança $\hat{\beta}$ é obtida maximizando $L(\beta)$. Pregibon (1984) apresenta um processo iterativo para se obter $\hat{\beta}$, assim como discute a definição de um resíduo e diversas técnicas de diagnóstico para o modelo condicional (4.34). A função desvio é agora obtida de (4.35), entretanto pouco se sabe sobre a distribuição nula da mesma. A função $L(\beta)$ acima será novamente discutida na Seção 8.5, onde serão tratados os modelos para análise de dados de sobrevivência.

4.6.3 - Modelos Logísticos Não-Lineares

Nos experimentos de tipo dose e resposta pode ocorrer que a transformação logit ou qualquer outra usual, não seja suficiente para produzir a

linearização desejada, sendo necessário impor alguma estrutura não- linear na parte sistemática. Por outro lado, pode ocorrer de uma estrutura não-linear e interpretável em experimentos com respostas contínuas, apresentar resultados satisfatórios para o caso binomial com transformação logit.

Serão apresentados a seguir três modelos logísticos não-lineares que se enquadram em alguma das situações acima.

O primeiro desses modelos (vide Giltinan et al. 1988) é frequentemente utilizado para avaliar a atividade conjunta de duas drogas A e B, separadamente ativas, na mortalidade de insetos. Sejam y a resposta binária de interesse ($y = 1$ o inseto morreu, $y = 0$ o inseto sobreviveu) e x e z as doses das drogas A e B, respectivamente. O modelo é dado por

$$logitp = \alpha + \beta \log\{x + \rho z + k(\rho x z)^{1/2}\},$$

onde $p = P\{y = 1|x, z\}$, β é a inclinação da relação dose resposta, ρ é a potência da droga B em relação à droga A e k é a interação entre as duas drogas, que tem a seguinte interpretação: $k = 0$ aditividade, $k < 0$ antagonismo e $k > 0$ sinergismo.

O segundo modelo (vide McCullagh e Nelder, 1983, pg. 205), tem sido aplicado para avaliar o efeito da mistura de um inseticida com um reagente químico na mortalidade de insetos. Aqui o reagente é inativo na ausência do inseticida e é utilizado para aumentar a toxidade do mesmo. O modelo é definido por

$$\text{logit}p = \alpha + \beta_1 \log(z - \theta) + \beta_2 x/(\phi + x),$$

onde x e z são respectivamente as doses do reagente e do inseticida, e os parâmetros β_1, β_2 e ϕ são interpretáveis. O parâmetro θ é incluído em alguns casos e representa um limite inferior para as doses do inseticida. Os

dois modelos acima podem ser ajustados pelo processo iterativo que será descrito na Seção 7.3.

Finalmente, será apresentado o modelo logístico generalizado proposto por Stukel (1988), que generaliza a maioria dos modelos usuais para respostas binárias. Esse modelo é definido por

$$logitp = h_\alpha(\eta),$$

onde $h_\alpha(\eta)$ é uma função estritamente crescente em η indexada por dois parâmetros α_1 e α_2, os quais definem a forma da curva. O preditor $h_\alpha(\eta)$ é definido da seguinte maneira:

Para $\eta \geq 0 (p \geq 1/2)$

$$\begin{aligned} h_\alpha &= \alpha_1^{-1}\{\exp(\alpha_1 |\eta|) - 1\}, & \alpha_1 > 0 \\ &= \eta, & \alpha_1 = 0 \\ &= -\alpha_1^{-1} \log(1 - \alpha_1 |\eta|), & \alpha_1 < 0 \end{aligned}$$

e para $\eta \leq 0 (p \leq 1/2)$

$$\begin{aligned} h_\alpha &= -\alpha_2^{-1}\{\exp(\alpha_2 |\eta|) - 1\}, & \alpha_2 > 0 \\ &= \eta, & \alpha_2 = 0 \\ &= \alpha_2^{-1} \log(1 - \alpha_2 |\eta|), & \alpha_2 < 0. \end{aligned}$$

Em particular quando $\alpha_1 = \alpha_2 = 0$ tem-se o modelo logístico linear múltiplo. Para $\alpha_1, \alpha_2 > 0$, a função h tem uma forma exponencial, e para $\alpha_1, \alpha_2 < 0$ uma forma logarítmica. Ocorre simetria somente quando $\alpha_1 = \alpha_2$. Se α_1 e α_2 são conhecidos tem-se um Modelo Linear Generalizado (Nelder e Wedderburn, 1972) com ligação $h^{-1}[\log\{p/(1-p)\}] = \eta$. A função h e as derivadas correspondentes da mesma são contínuas em α e

η. Os modelos probit e complementar log-log são obtidos aproximadamente quando $\alpha_1 = \alpha_2 = 0.165$, e $\alpha_1 = 0.62$ e $\alpha_2 = -0.037$, respectivamente.

Stukel (1985) apresenta um processo iterativo para obtenção das estimativas de máxima verossimilhança de β e $\alpha = (\alpha_1, \alpha_2)^T$ pelos sistemas GLIM ou GENSTAT.

§4.7 Exercícios

1 - Considere uma Tabela 2×2 com os seguintes valores: $a = 2$, $n_1 = 5$, $n_2 = 2$ e $m = 3$. Obtenha a estimativa de máxima verossimilhança condicional para ψ.

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 - Os dados abaixo são de um estudo de seguimento para avaliar a associação entre duas técnicas cirúrgicas A e B para cura de uma certa doença, e a ocorrência de problemas graves pós-operatórios.

Problemas graves pós-operatórios	Até 50 anos		Acima de 50 anos	
	Técnica A	Técnica B	Técnica A	Técnica B
Sim	6	7	7	4
Não	14	23	9	12

Verifique se o risco relativo (técnica A em relação à técnica B) é homogêneo nos estratos, ao nível de 10%. Utilize a estatística (4.12) com $\hat{\psi}_{MH}$ no lugar de $\hat{\psi}$.

4 - 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

5 - Mostre que o parâmetro β do modelo logístico condicional dado em (4.31) pode ser estimado pelo processo itetativo (4.22).

6 - Mostrar que a densidade (4.1) pertence à família exponencial de distribuições (6.1) e que seus comulantes $K_i(\psi)$, $i = 1, 2, \ldots$ seguem a relação

$$K_i(\psi) = \frac{1}{\psi}\frac{d}{d\psi}K_{i-1}(\psi),\ i = 2, 3, \ldots$$

7 - Considere a regressão linear $\alpha + \beta x_i$ para representar a probabilidade de sucesso μ_i da i-ésima observação binária y_i, $i = 1, \ldots, n$. Calcular as estimativas de máxima verossimilhança e de mínimos quadrados dos parâmetros α e β nos dois casos: (a) sem restrições; (b) com a restrição dos valores ajustados pertencerem ao intervalo (0,1). Obter as estatísticas escore, de Wald e da razão de máxima verossimilhança nos seguintes testes: $H_1: \alpha = 0$ versus $A_1: \alpha \neq 0$ e $H_2: \beta = 0$ versus $A_2: \beta \neq 0$.

8 - Sejam dois grupos de indivíduos que são comparados via o modelo logístico linear com uma covariável adicional x, isto é, supondo as estruturas $\alpha+\beta x$ e $\alpha+\beta x+\triangle$. Obter as formas das três estatísticas discutidas na Seção 4.5.3 para os testes: $H: \triangle = 0$ versus $A: \triangle \neq 0$, $H_1: \beta = 0$ versus $A_1: \beta \neq 0$ e $H_2: \alpha = 0$ versus $A_2: \alpha \neq 0$.

CAPÍTULO 5

MODELO NORMAL NÃO-LINEAR

§5.1 Introdução

Até o início da década de 70 as principais técnicas desenvolvidas para os modelos de regressão não-linear se restringiam à suposição de normalidade para a variável de resposta. Em 1972, Nelder e Wedderburn ampliaram a distribuição da variável de resposta para a família exponencial de distribuições, definindo os Modelos Lineares Generalizados (vide Capítulo 6). Mesmo assim, os modelos normais não-lineares continuaram recebendo um tratamento especial, surgindo diversos trabalhos na década de 70 e alguns nesta década. Particularmente, destaca-se o livro de Ratkowsky (1983), onde vários modelos normais não-lineares são discutidos segundo diversos aspectos.

A principal característica desses modelos é que os mesmos em geral são deduzidos a partir de suposições teóricas (quase sempre equações diferenciais) e os parâmetros resultantes são interpretáveis. Assim, aproximá-los para os modelos normais lineares, mesmo que sejam alcançados ajustes satisfatórios, prejudicaria bastante a obtenção de estimativas mais realistas dos parâmetros de interesse.

Nem sempre os modelos normais não-lineares são expressos numa forma paramétrica adequada, que facilite a convergência rápida dos processos iterativos utilizados na estimação dos parâmetros, sendo necessário procurar, em muitos casos, uma parametrização mais apropriada.

Embora as técnicas de diagnóstico da regressão normal não-linear sejam simples extensões das técnicas da regressão linear, as interpretações não são diretamente aplicadas, particularmente em virtude dos resíduos ordinários não terem mais distribuição aproximadamente normal. Isso levou ao desenvolvimento de técnicas específicas de diagnóstico para os modelos normais não-lineares (vide Cook e Tsai, 1985). Similarmente, as propriedades das somas de quadrados contidos nas tabelas clássicas de análise da variância (ANOVA), apresentadas no Capítulo 1, 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

$$y_i = f_i(\beta; x) + \varepsilon_i = \eta_i(\beta) + \varepsilon_i, \ i = 1, \ldots, n, \tag{5.1}$$

onde os ε_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 covariáveis.

Esses modelos são aplicados nas mais diversas áreas, tais como Ecologia, Agricultura, Farmacologia, Biologia, etc. A seguir são apresentados alguns modelos e a(s) respectiva(s) área(s) em que cada um é mais utilizado.

1 - Modelo para avaliar a mistura de duas drogas

Esse modelo é geralmente aplicado na área Farmacológica 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 à droga A e k representa a interação entre as drogas, sendo interpretado da seguinte maneira: $k = 0$ significa que há ação similar entre as duas drogas, $k > 0$ representa sinergismo e $k < 0$ significa antagonismo.

2 - Modelo de Von-Bertalanffy

Frequentemente aplicado na área Ecológica para explicar o comprimento de um peixe pela sua idade. A forma mais conhecida desse modelo é dada por

$$y = \alpha[1 - \exp\{-\delta(x - \gamma)\}] + \varepsilon,$$

onde x representa a idade do peixe, α é 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.

3 - Modelos sigmoidais

Fenômenos produzindo curvas sigmoidais na forma de S são freqüentemente encontrados na Agricultura, em Biologia, Ecologia, Engenharia e Economia. Essas curvas começam em algum ponto fixo e crescem monotonicamente até um ponto de inflexão, a partir daí a taxa de crescimento começa a diminuir até a curva se aproximar de um valor final, chamado de assíntota. Na Tabela 5.1 são relacionados alguns modelos usuais com essa forma.

Tabela 5.1: *Alguns modelos do tipo sigmoidal.*

Modelo	Componente sistemático
Gompertz	$\alpha \exp\{-\exp(\beta - \gamma x)\}$
logístico	$\alpha/\{1 + \exp(\beta - \gamma x)\}$
Richards	$\alpha/[\{1 + \exp(\beta - \gamma x)\}^{1/\delta}]$
Morgan-Mercer-Flodin (MMF)	$(\beta\gamma + \alpha x^{\delta})/(\gamma + x^{\delta})$
Weibull	$\alpha - \beta \exp(-\gamma x^{\delta})$

Fonte: Ratkowsky (1983).

Nesses modelos o parâmetro α é o valor máximo esperado para a resposta, ou assíntota. O parâmetro β está relacionado com o intercepto, isto é, com o valor de $\mu = E(y)$ correspondente a $x = 0$. Para todos os modelos da Tabela 5.1 esse parâmetro pelo menos determina o intercepto. O parâmetro γ está relacionado com a taxa média de crescimento da curva, e finalmente o parâmetro δ, que aparece em alguns modelos, é utilizado para aumentar a flexibilidade dos mesmos no ajuste dos dados.

4 - Modelos parcialmente não-lineares

Os modelos normais parcialmente não-lineares aparecem muito freqüentemente na prática e por terem uma estrutura simples diversas técnicas usuais são simplificadas. Esses modelos são expressos na forma

$$y = x^T\alpha + \delta f(\gamma) + \varepsilon,$$

onde $\alpha = (\alpha_1, \ldots, \alpha_{p-2})^T$, x é um vetor que contém os valores de $p-2$ covariáveis, δ e γ são escalares e as f_i's são funções diferenciáveis. Alguns exemplos são apresentados a seguir.

Gallant (1975) aplica o modelo $\mu = \alpha_1 x_1 + \alpha_2 x_2 + \delta \exp(\gamma z)$ em um delineamento com um fator, onde x_1 e x_2 representam dois tratamentos e z o tempo, que afeta exponencialmente a resposta. Darby e Ellis (1976) utilizam o modelo $\mu = \alpha + \delta \log(z_1 + \gamma z_2)$ para avaliar a atividade conjunta de duas drogas. Stone (1980) sugere um modelo parecido dado por $\mu = \alpha + \delta \log\{z_1/(\gamma + z_2)\}$, onde z_1 e z_2 representam, respectivamente, as concentrações de uma droga ativa e de um reagente. Outro exemplo é o modelo assintótico de regressão $\mu = \alpha - \delta\gamma^z$ (Ratkowsky, 1983) que tem sido intensivamente aplicado na Agricultura, assim como, na Biologia, Engenharia e particularmente na Ecologia para explicar o comprimento y de um peixe pela idade z do mesmo. Um modelo parcialmente não-linear utilizado para explicar a resistência y de um termostato pela temperatura z é dado por $\mu = -\alpha + \delta/(\gamma + z)$.

Os seguintes modelos com estrutura não-linear mais geral são encontrados em Ratkowsky: (i) $\mu = \alpha \exp\{-\beta/(\gamma + x)\}$, para explicar a idade de um certo tipo de coelho selvagem pelo peso x dos olhos e (ii) $\mu = \theta_1\theta_3 x_1/(1 + \theta_1 x_1 + \theta_2 x_2)$, para explicar, numa determinada reação química, a razão da reação denotada por y pelas concentrações x_1 e x_2 de dois reagentes. Ratkowsky também apresenta diversas formas alternativas de reparametrização para a maioria dos modelos mencionados acima, com o intuito de diminuir o viés das estimativas e facilitar a convergência do processo iterativo utilizado.

Na Seção 5.2 é apresentado o processo iterativo de Newton-Raphson para a obtenção da estimativa de mínimos quadrados de β assim como alguns resultados assintóticos. Medidas de não-linearidade e algumas técnicas relacionadas com esse assunto são discutidas na Seção 5.3. As principais técnicas de diganóstico utilizadas na regressão normal não- linear são apresentadas na Seção 5.4, seguidas de algumas ilustrações.

§5.2 Estimação de Máxima Verossimilhança

Sejam $y_1, \ldots, y_n$ variáveis aleatórias independentes com a estrutura dada em (5.1). 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

$$S(\beta) = \sum_{i=1}^{n} \{y_i - \eta_i(\beta)\}^2.$$

Expandindo $\eta(\beta)$ por série de Taylor em torno de um valor $\beta^{(0)}$ até segunda ordem, chega-se ao seguinte processo iterativo para obter $\hat{\beta}$:

$$\beta^{(m+1)} = \beta^{(m)} + \{\tilde{X}^{(m)T}\tilde{X}^{(m)}\}^{-1}\tilde{X}^{(m)^T}\{y - \eta(\beta^{(m)}\}, \tag{5.2}$$

$m = 0, 1, \ldots$, 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 $\left|(\beta^{(m+1)} - \beta^{(m)})/\beta^{(m)}\right| < \varepsilon$, onde ε é um valor arbitrário.

A convergência de (5.2) 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 não-linearidade acentuada (vide Seção 5.3) e/ou mal condicionamento da matriz $\tilde{X}$, prejudiquem a convergência do processo iterativo. Em Souza (1986) há uma discussão detalhada do método de Newton-Raphson e de outros métodos iterativos usuais em regressão normal não-linear. Ratkowsky (1983) sugere algumas

técnicas para se obter valores iniciais para os parâmetros de β, as quais serão aplicadas a seguir para alguns dos modelos descritos na Seção 5.1.

1 - 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 drogra A. Denotando tais estimativas por α_0 e δ_0, os valores inciais 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 (5.2) à convergência deve-se tentar novas estimativas para ρ e k e repetir o procedimento.

2 - Modelo de Von-Bertalanffy

O primeiro passo nesse caso é obter um valor inicial para α. Como esse 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 α_0, γ_0 e δ_0 não levarem (5.2) à convergência, deve-se tentar uma nova estimativa para α e repetir o procedimento.

3 - Modelos sigmoidais

Para os modelos de Gompertz e logístico, os valores iniciais são obtidos de maneira muito parecida. Para ambos, deve-se inicialmente atribuir um valor inicial para α, por exemplo observando o gráfico de y contra x ou tomando $\alpha_0 = y_{\max}$, já que α representa a assíntota. Conhecendo-se α_0 tem-se, respectivamente, as relações lineares

$$\log\{-\log(\mu/\alpha_0)\} = \beta - \gamma x$$

e

$$\log\{(\alpha_0/\mu) - 1\} = \beta - \gamma x.$$

Logo, os valores iniciais β_0 e γ_0 saem respectivamente, das regressões lineares de $\log\{-\log(y/\alpha_0)\}$ e de $\log\{(\alpha_0/y) - 1\}$ sobre x.

Para os demais modelos, Richards, MMF e Weibull, a estimativa inicial α_0 pode ser obtida da mesma forma acima. Entretanto, aparece agora o parâmetro adicional δ. Em particular para o modelo de Richards tem-se a relação linear

$$\log\left\{\left(\frac{\alpha_0}{\mu}\right)^{\delta} - 1\right\} = \beta - \gamma x;$$

logo, conhecendo-se uma estimativa para δ, os valores iniciais β_0 e γ_0 serão obtidos da regressão linear de $\log\{(\frac{\alpha_0}{y})^{\delta_0} - 1\}$ sobre x. Ratkowsky (1983) sugere que δ_0 seja obtido através do ponto de inflexão da curva, isto é, do ponto (x_F, μ_F) tal que $d^2\mu/dx^2$ seja igual a zero.

Diferenciando a parte sistemática do modelo de Richards duas vezes em relação a x e igualando a zero, obtém-se

$$x_F = (\beta - \log\delta)/\gamma \text{ e } \mu_F = \alpha(1+\delta)^{-1/\delta}$$

Portanto, obtendo-se uma estimativa para μ_F, por exemplo através do gráfico de y contra x, extrai-se δ_0 da equação para μ_F.

No modelo MMF, o parâmetro β pode ser estimado inicialmente pelo gráfico de y contra x, por exemplo atribuindo a β_0 o valor de μ quando $x = 0$. Para esse modelo, δ_0 pode ser obtido através das equações para o ponto de inflexão (x_F, μ_F), onde

$$x_F = \{\frac{\gamma(\delta-1)}{\delta+1}\}^{1/\delta} \text{ e } \mu_F = \{\beta(\delta+1) + \alpha(\delta-1)\}/2\delta.$$

Logo, conhecendo-se as estimativas para μ_F e x_F, os valores iniciais δ_0 e γ_0 saem, respectivamente, das equações para μ_F e x_F.

Para o modelo de Weibull, β_0 pode ser obtido analogamente ao modelo MMF. Denotando por y_{INI} a estimativa para μ tal que $x = 0$, obtém-se

$$\beta_0 = \alpha_0 - y_{INI}.$$

Substituindo β_0 e α_0 na componente sistemática do modelo chega-se à seguinte relação linear:

$$z_0 = \log\gamma + \delta \log x,$$

onde $z_0 = \log\{-\log(\frac{\alpha_0-\mu}{\beta_0})\}$, sugerindo que γ_0 e δ_0 sejam obtidos da regressão linear simples de $\log\{-\log(\frac{\alpha_0-y}{\beta_0})\}$ sobre $\log x$.

4 - Modelo assintótico de regressão

Para o modelo assintótico, dado por

$$y = \alpha - \beta\gamma^x + \varepsilon,$$

os valores iniciais são facilmente obtidos. Inicialmente deve-se estimar o parâmetro α, a assíntota, graficamente ou por $y_{\max}$. Substituindo α_0 na parte sistemática do modelo obtém-se a relação linear

$$z_0 = \log\beta + x\log\gamma,$$

onde $z_0 = \log(\alpha_0 - \mu)$. Logo, β_0 e γ_0 saem da regressão linear simples de $\log(\alpha_0 - y)$ sobre x.

Pelos exemplos acima pode-se notar a importância da interpretabilidade dos parâmetros da componente sistemática de um modelo normal não-linear, na estimação desses mesmos parâmetros.

5.2.1 - Resultados Assintóticos

Nesta seção serão apresentados os resultados assintóticos mais relevantes relacionados com a estimação e testes de hipóteses para o parâmetro $\beta = (\beta_1, \ldots, \beta_p)^T$ do modelo normal não-linear.

A log-verossimilhança do modelo (5.1), como função de β, é expressa na forma

$$L(\beta) = (2\pi\sigma^2)^{-n/2} \exp\{-S(\beta)/2\sigma^2\}.$$

A estimativa de máxima verossimilhança $\hat{\beta}$, cuja estimativa é obtida pelo processo iterativo dado em (5.2) sendo consistente e tendo 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 $100(1-\alpha)\%$ para β_j, será formado pelos limites

$$\hat{\beta}_j \pm t_{\alpha/2} \times (-\hat{k}^{jj})^{1/2},$$

onde $t_{\alpha/2}$ é o quantil $(1-\alpha/2)$ de uma t de Student com $(n-p)$ graus de liberdade e $-\hat{k}^{jj}$ é a estimativa do elemento (j,j) de K^{-1}.

Uma região de aproximadamente $100(1-\alpha)\%$ de confiança para 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 $(1-\alpha)$ de uma distribuição F e a matriz $\tilde{X}$ é avaliada em $\hat{\beta}$. Essa última expressão é uma adaptação da região de confiança da regressão normal-linear.

Para testar a hipótese $H: \beta \in B$, onde B é um subconjunto do espaço paramétrico, utiliza-se usualmente a estatística da razão de máxima verossimilhança, dada por

$$-2\log\lambda = n\log\{S(\tilde{\beta}) - S(\hat{\beta})\},$$

onde $S(\tilde{\beta})$ é a soma de quadrados de resíduos para o modelo ajustado em H. Sob essa hipótese, a estatística acima tem assintoticamente distribuição χ^2 com $(p-m)$ graus de liberdade, onde $m = dim(B)$. Johansen (1983) mostra que a estatística $-2\log\lambda$ é assintoticamente equivalente à estatística

$$n\sum_{i=1}^{n}\{\eta_i(\tilde{\beta}) - \eta_i(\hat{\beta})\}^2 / S(\hat{\beta}),$$

que é mais fácil de ser calculada.

Uma estatística alternativa para testar H é dada por

$$F = \frac{(n-p)}{(p-m)} \quad \frac{S(\tilde{\beta}) - S(\hat{\beta})}{S(\hat{\beta})},$$

que sob essa hipótese tem assintoticamente distribuição F com $(p-m)$ e $(n-p)$ graus de liberdade. Logo, deve-se rejeitar H, para um nível

de significância α, se $F \geq F_{(p-m),(n-p)}(\alpha)$. Esse resultado vale também quando a variável de resposta não é normal, havendo contudo algumas condições adicionais de regularidade.

5.2.2 - Exemplos

1 - Modelo de Gompertz para explicar o comprimento de um certo tipo de feijoeiro

A versão do modelo de Gompertz, dada por $\log \mu = \alpha - \exp(\delta - \gamma x)$, onde $\log \alpha$ representa a assíntota, é freqüentemente utilizado para explicar o comprimento de diversos tipos de feijoeiros pela quantidade de água na raiz dos mesmos. Para ilustrar, será utilizado o conjunto de dados

y:	1.3	1.3	1.9	3.4	5.3	7.1	10.6	16.0
	16.4	18.3	20.9	20.5	21.3	21.2	20.9	

e $x = 0.5 + \ell$, $\ell = 0, 1, \ldots, 14$.

Iniciando o processo iterativo (5.2) com os valores $\alpha_0 = 3.0, \delta_0 = 2.1$ e $\gamma_0 = 0.40$ chega-se à convergência após 7 iterações com as estimativas $\hat{\alpha} = 3.114(0.037)$, $\hat{\delta} = 2.106(0.235)$ e $\hat{\gamma} = 0.388(0.046)$, as quais indicam que os parâmetros são bem determinados. Em particular, o tamanho máximo esperado para esse tipo de feijoeiro será aproximadamente $\exp(\hat{\alpha}) = 22.55$ cms com desvio padrão de 1.038.

A Figura 5.1 exibe o gráfico dos valores ajustados e observados contra a quantidade de água na raiz da planta, mostrando que o modelo é adequado para ajustar esse conjunto de dados.

2 - Testando a Interação entre Duas Drogas

Considere o conjunto de dados apresentado na Tabela 5.1 e o modelo descrito na Seção 5.1, para avaliar a atividade conjunta de duas drogas.

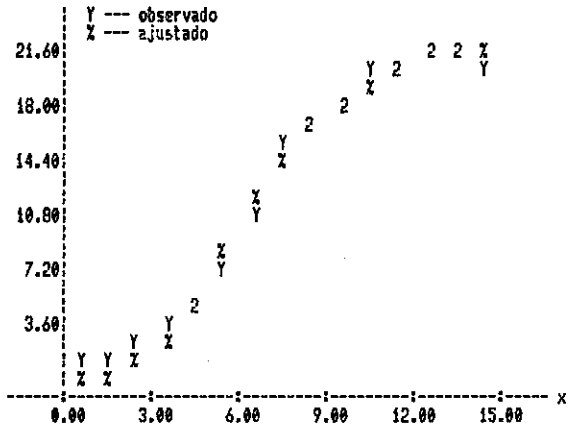

Figura 5.1: *Valores observados e ajustados pelo modelo de Gompertz contra a covariável* x.

Após o ajuste desse modelo, usualmente testa-se a significância da interação, isto é, a hipótese $H: k = 0$.

As estimativas obtidas sob a hipótese alternativa são apresentadas na Tabela 5.2 e a soma de quadrados de resíduos correspondente vale 220.18. Sob H, inicializando o processo iterativo (5.2) com os valores $\alpha_0 = 20.0$, $\delta_0 = 14.0$ e $\rho_0 = 0.10$ obtém-se na convergência $\hat{\alpha} = -18.30(2.83)$, $\hat{\delta} = 10.56(0.83)$, $\hat{\rho} = 0.046(0.0035)$ e soma de quadrados de resíduos dada por 243.97.

Em ambos os ajustes, a comparação das estimativas obtidas com os respectivos desvios padrões indica que os parâmetros são bem determinados. Em particular, a significância da interação k pode ser avaliada através da estatística F, obtendo-se o valor

$$F = \frac{(56-4)}{(4-3)} \quad \frac{243.97 - 220.18}{220.18} \cong 5.62.$$

que é significativo a 5%. Como $\hat{k} < 0$, há uma indicação de antagonismo entre as duas drogas, isto é, que a mistura de ambas produz um efeito menor que a soma dos efeitos individuais das duas drogas.

Será mostrado na Seção 5.4.5, através de algumas técnicas de diagnóstico, que a significância de k deve-se essencialmente a algumas "misturas" extremas que incluem apenas uma das drogas. Esse resultado mostra a necessidade de uma análise de diagnóstico após o ajuste do modelo. O valor de $\hat{\rho} = 0,046$ indica que a insulina padrão é aproximadamente 22 vezes mais eficaz que a insulina na forma suberoyl A1-B29.

§5.3 Medidas de Não-Linearidade

O principal objetivo das medidas de não-linearidade é verificar se o grau de não-linearidade de um problema de estimação não-linear é suficientemente pequeno para que as técnicas usuais de estimação, desenvolvidas para a regressão linear, sejam utilizadas como uma boa aproximação para o caso não-linear.

A primeira tentativa relevante no sentido de desenvolver uma medida de não- linearidade foi de Beale (1960). Contudo, Guttman e Meeter (1965) mostraram que a medida proposta por ele tende a subestimar o verdadeiro grau de não- linearidade do modelo. Uma outra contribuição importante foi a de Box (1971), que obteve a aproximação de ordem n^{-1} para o viés do estimador de máxima verossimilhança do vetor β de um modelo normal não-linear. Entretanto, foi somente no início desta década que surgiu o trabalho mais relevante nesta área. Bates e Watts (1980) utilizando alguns conceitos de geometria diferencial, desenvolveram duas medidas de curvatura para os modelos normais não-lineares. Essas medidas indicam, respectivamente, o

grau de não-linearidade intrínseca de um modelo e o grau de não-linearidade aparente ou devida à parametrização utilizada.

Ratkowsky (1983) comparou algumas formas paramétricas para diversos modelos normais não-lineares através de simulações e utilizou as medidas de Box e de Bates e Watts.

Para que o leitor tenha uma idéia mais clara dos conceitos de não-linearidade intrínseca e de não-linearidade aparente, serão comparados a seguir um modelo linear e um modelo não-linear para o caso de $n = 2$ e $p = 1$.

Considere inicialmente o modelo linear simples dado por $y_i = \beta x_i + \varepsilon_i$, $i = 1, 2$, onde x denota uma covariável qualquer e β um parâmetro desconhecido. Nesse caso, o espaço de estimação (espaço de todos os valores ajustados possíveis) tem dimensão igual a um e é formado pelos pontos

$$X\beta = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \beta, \ \beta \in \mathbf{R},$$

ou seja, é uma reta no $\mathbf{R}^2$. Além disso, para qualquer conjunto de soluções $\beta^{(1)}, \beta^{(2)}, \ldots$, tais que $\beta^{(i+1)} - \beta^{(i)} = \triangle$, onde $\triangle$ é uma constante arbitrária, as soluções possíveis para $X\beta$ serão tais que

$$X\beta^{(i+1)} - X\beta^{(i)} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \triangle, \ i = 1, 2, \ldots,$$

ou seja, se as soluções para β forem igualmente espaçadas então os valores ajustados correspondentes também serão igualmente espaçados.

Considere agora o modelo normal não-linear $y_i = x_i^{\beta} + \varepsilon_i$, $i = 1, 2$ e os dados apresentados em Ratkowsky (1983, pg. 7)

$$y = (2.5 \ 10)^T \text{ e } X = (2 \ 3)^T.$$

Nesse caso o espaço de estimação não é mais uma reta, e sim uma curva em torno da estimativa de máxima verossimilhança $\hat{\beta} = 2.05$. A curva correspondente aos pontos $(2^{\beta} \quad 3^{\beta})^T$ com β variando em espaçamentos iguais a 0.5 é exibida pela Figura 5.2. Note que os pontos do espaço de estimação não são igualmente espaçados como ocorreu no caso linear.

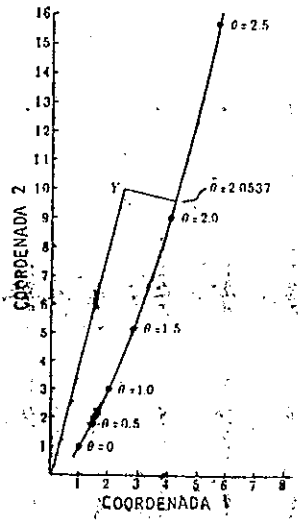

Figura 5.2 - *Representação da curva* $(2^{\beta} \quad 3^{\beta})^T$ *com* β *variando em espaçamentos iguais. (Extraída do livro de Ratkowsky, (1983).*

Assim, quanto mais essa curva se afasta da reta tangente em $\hat{\beta}$ maior será o que Bates e Watts (1980) chamam de "não-linearidade intrínseca" do modelo, e quanto mais desiguais forem os espaçamentos entre os pontos do espaço de estimação maior será o que ambos chamam de "não-linearidade aparente causada pela parametrização do modelo".

Portanto, a não-linearidade de um modelo pode ser devida a duas causas. A primeira é a curvatura real do modelo ou intrínseca como definem Bates e Watts, que é invariante com qualquer tipo de reparametrização. A

segunda é a curvatura devida à forma como os parâmetros aparecem no modelo. Essa última pode ser eliminada ou pelo menos reduzida através de reparametrizações. Para ilustrar isso, considere o modelo normal não-linear descrito anteriormente com a seguinte reparametrização:

$$y_i = x_i^{\log \phi} + \varepsilon_i, \quad i = 1, 2,$$

onde $\phi = \exp(\beta)$. A Figura 5.3 exibe os pontos da curva $(2^{\log \phi} \quad 3^{\log \phi})^T$ com espaçamentos iguais a 1.0 para ϕ. Nota-se que os espaçamentos entre os pontos correspondentes são praticamente iguais, indicando que o grau de não-linearidade aparente foi substancialmente reduzido com essa reparametrização. Entretanto, a curvatura do espaço de estimação continua com a mesma forma anterior, como era de se esperar.

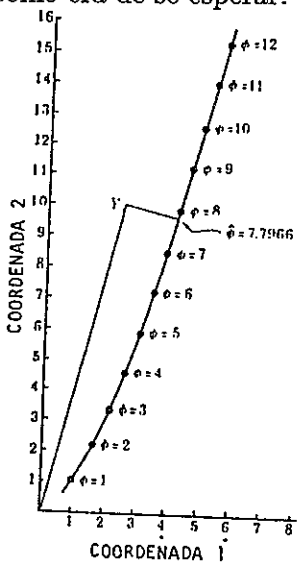

Figura 5.3: *Representação da curva* $(2^{\log \phi} \; 3^{\log \phi})^T$ *com* ϕ *variando em espaçamentos iguais a 1.0. (Extraída do livro de Ratkowsky, 1983).*

5.3.1 - Medidas de Curvatura de Bates e Watts

Considere o modelo de regressão normal não-linear definido na Seção 5.1. Uma reta no espaço paramétrico passando por $\hat{\beta}$, pode ser expressa, usando um parâmetro b por

$$\beta(b) = \hat{\beta} + bh,$$

onde $h = (h_1, \ldots, h_p)^T$ é um vetor de valores não-nulos. Essa reta gera uma curva sobre o espaço de estimação, definida por

$$\eta_h(b) = \eta(\hat{\beta} + bh).$$

A tangente a essa curva no ponto $b = 0$ é expressa na forma

$$\dot{\eta}_h = \tilde{X}h, \tag{5.3}$$

onde $\tilde{X}$ é aqui a matriz Jacobiana da transformação de $\eta(\beta)$ em $\beta = \hat{\beta}$. O conjunto de todas as combinações lineares da forma (5.3) é também chamado de plano tangente em $\eta(\hat{\beta})$.

A aceleração da curva η_h é definida por

$$\ddot{\eta}_h = h^T \hat{W} h,$$

onde W é um "vetor" de dimensão $n \times p \times p$ com i-ésima face dada por $W_i = (\partial^2 \eta_i / \partial \beta_r \partial \beta_s)$, $i = 1, \ldots, n$ e $r, s = 1, \ldots, p$. Portanto, cada elemento do vetor $n \times 1$ $\ddot{\eta}_h$ é da forma $h^T \hat{W}_i h$, $i = 1, \ldots, n$.

O vetor de aceleração $\ddot{\eta}_h$ pode ser decomposto em três componentes. A primeira componente $\ddot{\eta}_h^{IN}$ determina a variação na direção do vetor de velocidade instantânea $\dot{\eta}_h$ normal ao plano tangente, enquanto a segunda e

a terceira componentes, cuja norma será denotada $\ddot{\eta}^{PE}$, determinam, respectivamente, a variação na direção de $\ddot{\eta}_h$ paralela ao plano tangente e a variação na velocidade do ponto móvel. Essas componentes foram transformadas por Bates e Watts (1980) nas seguintes curvaturas:

A - Curvatura intrínseca

Definida por

$$K_h^{IN} = \left\| \ddot{\eta}_h^{IN} \right\| / \left\| \dot{\eta}_h \right\|^2 .$$

B - Curvatura devida à parametrização

Definida por

$$K_h^{PE} = \left\| \ddot{\eta}_h^{PE} \right\| / \left\| \dot{\eta}_h \right\|^2 .$$

Essas curvaturas podem ser padronizadas de tal modo que fiquem invariantes com mudanças de escala. Isso é obtido multiplicando K_h^{IN} e K_h^{PE} por $s\sqrt{p}$ com $s = \{S(\hat{\beta})/(n-p)\}^{1/2}$. Tem-se portanto as curvaturas padronizadas

$$\gamma_h^{IN} = s\sqrt{p}\, K_h^{IN} \quad \text{e} \quad \gamma_h^{PE} = s\sqrt{p} K_h^{PE}.$$

As medidas relativas acima podem ser usadas não somente para comparar diferentes parametrizações de um determinado problema, mas também diferentes conjuntos de dados para o mesmo modelo ou para modelos diferentes.

As medidas de não-linearidade de Bates e Watts (1980) são definidas como sendo as curvaturas máximas

$$\gamma^{IN} = \max_h\{K_h^{IN}\} \text{ e } \gamma^{PE} = \max_h\{K_h^{PE}\}.$$

Bates e Watts sugerem o critério

$$\gamma^{IN} \geq 2F^{-1/2} \text{ e } \gamma^{PE} \geq 2F^{-1/2},$$

como guia para indicar se o modelo ajustado tem, respectivamente, curvatura intrínseca e curvatura aparente acentuada, onde F é o quantil $(1-\alpha)$ de uma distribuição F com p e $(n-p)$ graus de liberdade.

Para o cálculo das medidas acima é preciso inicialmente decompor a matriz $\tilde{X}$ num produto de duas matrizes Q e R, isto é, $\tilde{X} = QR$, onde Q é uma matriz $n \times n$ ortogonal e R é uma matriz $n \times p$ definida por

$$R = \begin{bmatrix} \tilde{R} \\ 0 \end{bmatrix},$$

com $\tilde{R}$ sendo uma matriz $p \times p$ triangular superior e inversível. As matrizes Q e R podem ser obtidas a partir da decomposição de Businger e Golub (1965).

A seguir deve-se obter o "vetor" $U = L^T \hat{W} L$, onde $L = \tilde{R}^{-1}$. Os elementos de U são vetores $n \times 1$ denotados por U_{kj}, $k, j = 1, \ldots, p$. Define-se então o que Bates e Watts chamam de "vetor" de aceleração $A = Q^T U$ de dimensão $n \times p \times p$. O (k, j)- ésimo elemento desse "vetor" é um vetor $n \times 1$ expresso na forma $Q^T U_{kj}$. O "vetor" A é portanto dado por

$$Q = \begin{bmatrix} A^T U_{11} \cdots Q^T U_{1p} \\ Q^T U_{p1} \cdots Q^T U_{pp} \end{bmatrix},$$

onde $Q^T U_{kj1} = (a_{kj1}, \ldots, a_{kjn})^T$. A i-ésima face de A é expressa na forma

$$A_i = \begin{bmatrix} a_{11i} & \cdots & a_{1pi} \\ & \vdots & \\ a_{p11} & \cdots & a_{ppi} \end{bmatrix}, \quad i = 1, \cdots, n.$$

Seja A^{IN} o "vetor" constituído das p primeiras faces de A e A^{PE} o "vetor" constituído das últimas $(n - p)$ faces de A. Então as medidas de não-linearidade serão dadas por

$$\gamma^{IN} = \max_h \left\| h^T A^{IN} h \right\| \text{ e } \gamma^{PE} = \max_h \left\| h^T A^{PE} h \right\|,$$

onde $\|h\| = 1$. Para efetuar os cálculos acima não há, em geral, fórmulas explícitas, sendo necessário recorrer a algum processo iterativo. Souza (1986) discute a obtenção de γ^{IN} e γ^{PE} através de um processo iterativo proposto por Bates e Watts (1980).

5.3.2 - Viés de Ordem n^{-1} de Box

Cox e Snell (1968) deduziram uma aproximação de ordem n^{-1} para o viés do estimador de máxima verossimilhança do parâmetro β em uma classe geral de modelos que inclui o modelo normal não-linear como um caso particular. Box (1971) utilizando esse trabalho, obteve uma aproximação $b \cong E\{\hat{\beta} - \beta\}$ em forma matricial, dada por

$$b = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T d, \tag{5.4}$$

onde d é um vetor $n \times 1$ com elementos $d_i = -\frac{1}{2}\sigma^2 tr\{(\tilde{X}^T\tilde{X})^{-1}W_i\}$, $i = 1, \ldots, n$. Portanto, o viés é simplesmente a estimativa de mínimos quadrados para o conjunto de coeficientes da regressão normal linear de d sobre as colunas de $\tilde{X}$. Aqui $\tilde{X}$ é avaliada no parâmetro verdadeiro β.

Cook et al. (1986) mostraram que d é essencialmente a diferença entre os valores esperados das aproximações linear e quadrática para $\eta(\beta)$. Logo, o viés será pequeno se todos os elementos de d forem suficientemente próximos de zero, o que indica que o modelo é essencialmente linear, ou se d é ortogonal às colunas de $\tilde{X}$.

Battes e Watts (1980) mostraram que o viés de Box está relacionado com a medida de não-linearidade γ^{PE}. Portanto, o viés pode ser reduzido através de reparametrizações no modelo e a expressão (5.4) pode indicar quais parâmetros são os maiores responsáveis por um valor alto de não-linearidade.

Em particular para os modelos normais parcialmente não-lineares os termos de (5.4) valem

$$\tilde{X} = [X, f(\gamma), \delta f'(\gamma)]$$

e

$$d_i = 2f_i'(\gamma)Cov(\hat{\delta}, \hat{\gamma}) + \delta f_i''(\gamma)\,\mathrm{Var}(\hat{\gamma}), \quad i = 1, \ldots, n.$$

Logo, o viés fica dado por

$$b = -\delta^{-1} Cov(\hat{\delta}, \hat{\gamma}) I_p - \frac{1}{2}\delta \operatorname{Var}(\hat{\gamma})(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T f''(\gamma),$$

onde I_p é um vetor $p \times 1$ de zeros com o valor um na última posição e $(\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T f''(\gamma)$ são as estimativas dos coeficientes da regressão normal linear de $f''(\gamma)$ sobre $\tilde{X}$. Note que $Cov(\hat{\delta}, \hat{\gamma})$ contribui somente para o viés de $\hat{\gamma}$.

Box (1971) também desenvolveu uma fórmula para avaliar o viés dos estimadores de uma nova reparametrização, mostrando que o novo viés pode ser obtido através do viés da parametrização anterior.

Considere a reparametrização

$$\phi = g(\beta),$$

onde ϕ é um escalar, $g(\cdot)$ é uma função diferenciável e $\beta = (\beta_1, \ldots, \beta_p)^T$. Seja $b_{\hat{\phi}}$ o viés de ordem n^{-1} de ϕ. Box mostrou que

$$b_{\hat{\phi}} = G^T b + \frac{1}{2} tr\{M \operatorname{Var}(\hat{\beta})\},$$

onde G é um vetor $p \times 1$ com as derivadas de $g(\beta)$ com relação a β e M é uma matriz $p \times p$ de derivadas $\partial^2 g(\beta)/\partial\beta_r \partial\beta_s, \quad r, s = 1, \ldots, p$. Ambos, G e M, são avaliados em $\hat{\beta}$.

A variância de $\hat{\phi}$ pode também ser expressa em função da variância de $\hat{\beta}$.

$$\operatorname{Var}(\hat{\phi}) = tr\{(GG^T) \operatorname{Var}(\hat{\beta})\}.$$

Em particular para $p = 1$

$$b_{\hat{\phi}} = b\frac{\partial g(\beta)}{\partial \beta} + \frac{1}{2} \operatorname{Var}(\hat{\beta}) \frac{\partial^2 g(\beta)}{\partial \beta^2}$$

e

$$\mathrm{Var}(\hat{\phi}) = \mathrm{Var}(\hat{\beta})\{\frac{\partial^2 g(\beta)}{\partial \beta^2}\}^2,$$

com todas as derivadas sendo avaliadas em $\hat{\beta}$.

5.3.3 - Aperfeiçoamento da Estatística da Razão de Máxima Verossimilhança

Cordeiro e Paula (1989a) utilizando o critério de correção de Bartlett (1937) e as expansões de Lawley (1956) corrigiram a estatística da razão de máxima verossimilhança $-2\log\lambda$, até ordem n^{-1}, para a classe dos modelos exponenciais não-lineares (vide Seção 7.1), que engloba como caso particular os modelos normais não-lineares. Esse fator de correção, denotado por c, faz com que a estatística corrigida $-2c^{-1}\log\lambda$ se aproxime melhor da qui-quadrado de referência do que a estatística usual $-2\log\lambda$.

Para ilustrar, suponha a partição $\beta = (\beta_q^T \beta_{p-q}^T)^T$, $p > q$ e a hipótese H: $\beta_{p-q} = 0$. Nesse caso a estatística da razão de máxima verossimilhança é simplesmente dada por $-2\log\lambda = 2(\hat{L}_p - \hat{L}_q)$, onde $\hat{L}_q$ e $\hat{L}_p$ são, respectivamente, a log-verossimilhança maximizada sob H e para o modelo sob investigação. A correção de Bartlett é dada por

$$c = 1 + (\varepsilon_p - \varepsilon_q)/(p-q), \tag{5.5}$$

onde ε_p é um termo bastante complicado envolvendo valores esperados de derivadas da log-verossimilhança. Particularmente para os modelos normais não-lineares, tem-se

$$\varepsilon_p = \frac{\sigma^2}{4}\psi = \frac{\sigma^2}{4}\{2tr(B_d - BZ) - 1^T DMDI\}, \tag{5.6}$$

onde $Z = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}$, $D = diag\{d_1,\ldots,d_n\}$, $B = \{b_{ij}\}$, $b_{ij} = tr\{W_i(\tilde{X}^T\tilde{X})^{-1}W_j(\tilde{X}^T\tilde{X})^{-1}\}$, 1 é um vetor $n \times 1$ de 1's e $M = I - Z$

é o operador de projeção ortogonal de vetores do $\mathbf{R}^n$ no subespaço gerado pelas colunas da matriz $\tilde{X}$.

Mostra-se, utilizando (5.6), que

$$E\left(\frac{SQ\text{Res}}{\sigma^2}\right) \cong n - p - \frac{\sigma^2}{4}\psi,$$

isto é, o valor esperado da soma de quadrados de resíduos dividida por σ^2 é aproximadamente igual a $(n-p)$, que é o valor esperado no caso linear, mais uma contribuição devida à não-linearidade em $\eta(\beta)$, multiplicada por $-\sigma^2/4$. Johansen (1983) relacionou ψ com a medida de não-linearidade de Beale (1960).

Restringindo-se à subclasse dos modelos parcialmente não-lineares, ε_p reduz-se para

$$\varepsilon_p = \frac{\sigma^2}{4}\,\text{Var}^2(\hat{\gamma})1^T\Gamma M\Gamma 1,$$

onde $\sigma^2\,\text{Var}(\hat{\gamma})$ é a variância assintótica de $\hat{\gamma}$, $\Gamma = \delta\text{diag}\{f_1''(\gamma),\ldots,f_n''(\gamma)\}$ e $1^T\Gamma M\Gamma 1$ é a soma de quadrados de resíduos após a regressão linear de $\Gamma 1$ sobre as colunas de $\tilde{X}$.

Na prática o fator c deve ser estimado sob o menor modelo, isto é, tanto ε_p quanto ε_q em (5.5) devem ser computados sob $H\colon \beta_{p-q} = 0$. Para ilustrar, suponha que num modelo parcialmente não-linear o interesse é testar $H\colon \gamma = \gamma^{(0)}$. Logo, o fator de correção fica dado por

$$c = 1 + \frac{\sigma^2}{4}\,\text{Var}^2(\hat{\gamma})1^T\Gamma M\Gamma 1;$$

onde as quantidades $\text{Var}(\hat{\gamma}), \Gamma$ e M devem ser computadas sob H. Aqui, em particular, $\text{Var}(\hat{\gamma})$ é o elemento que ocupa a posição (p,p) da matriz $(\tilde{X}^T\tilde{X})^{-1}$.

5.3.4 - Exemplos

1 - **Reparametrização do Modelo MMF através das Medidas de Bates e Watts e de Box.**

Considere o modelo MMF, dado por

$$\mu = (\beta\gamma + \alpha x^{\delta})/(\gamma + x^{\delta}),$$

e o conjunto de dados abaixo (Ratkowsky, 1983, pg. 88).

y:	8.93	10.80	18.59	22.33	39.35	56.11
	61.73	64.62	67.08			

e

x:	9	14	21	28	42	57	63	70	79

onde y representa a produção e x o tempo de cultivo de uma certa cultura.

Na convergência chega-se às estimativas $\hat{\alpha} = 80.96$, $\hat{\beta} = 8.895$, $\hat{\gamma} = 49.577$ e $\hat{\delta} = 2.828$. As medidas de não- linearidade são estimadas por $\hat{\gamma}^{IN} = 0.180$ e $\hat{\gamma}^{PE} = 90.970$ e o valor crítico para um nível de significância de 5% vale $1/2\sqrt{F} = 0.229$, onde F é o quantil de 95% de uma distribuição F com 3 e 6 graus de liberdade. Portanto, a não-linearidade devida à parametrização do modelo é altamente significativa, enquanto a não-linearidade intrínseca é não-significativa.

Para determinar quais os parâmetros que possivelmente estão causando essa não-linearidade acentuada, utiliza-se a medida e Box par o viés. É usual expressar o viés como uma porcentagem da estimativa correspondente. Para as estimativas acima essas porcentagens valem respectivamente, 1.525%, -1.643%, 119.478% e 0.921%. Nota-se portanto, um valor muito elevado para o viés de $\hat{\gamma}$, indicando que possivelmente uma reparametrização nesse parâmetro possa reduzir o viés.

Ratkowsky sugere a simulação das distribuições das estimativas com viés acentuado, para se ter uma idéia da reparametrização a ser aplicada. No exemplo acima, a transformação $\phi = g(\gamma) = \log \gamma$ é a mais recomendada, obtendo-se $\hat{\phi} = \log \hat{\gamma} = 10.81$ e $b_{\hat{\phi}} = 0.1087$. Logo, a porcentagem do viés vale agora $0.1087 \times 100/10.81 \cong 1.00\%$, uma redução substancial em relação à porcentagem inicial.

2 - Modelo para Explicar a Resistência de um Termostato

Como foi mencionado na Seção 5.1, o modelo $\mu = -\alpha + \delta/(\gamma + z)$ é frequentemente utilizado para explicar a resistência y de um termostato pela temperatura z. Na versão acima do modelo, a variável de resposta é expressa na forma $\log y$.

Esse modelo será utilizado para ajustar o conjunto de dados abaixo (Ratkowsky, 1983, pg. 120)

y:	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

$$z = 50 + 5 \times \ell, \quad \ell = 0, 1, \ldots, 15.$$

As estimativas dos parâmetros são dadas por $\hat{\alpha} = 5.145$, $\hat{\delta} = 6,14 \times 10^5$ e $\hat{\gamma} = 3.44 \times 10^4$, e as medidas de não- linearidade são estimadas por $\hat{\gamma}^{IN} = 0.0002$ e $\hat{\gamma}^{PE} = 1.6056$.

Essa última é significativa a 5%, já que $1/2\sqrt{F} = 0.271$, onde F é o quantil de 95% de uma distribuição F com 3 e 13 graus de liberdade.

As porcentagens do viés de Box são desprezíveis para cada estimativa (menores que 0.001%) e as simulações não indicam afastamentos da normalidade. Esse resultado, um tanto contraditório, pode ser explicado pelo fato das medidas de não-linearidade de Bates e Watts serem globais enquanto

a medida de Box é individual, assim como as simulações, que indicam as distribuições marginais das estimativas. Logo, pode ocorrer que a medida de curvatura γ^{PE} seja significativa e nenhum parâmetro esteja influindo de forma acentuada na não-linearidade do modelo.

§5.4 Técnicas de Diagnóstico

Exceto com relação aos resíduos, as técnicas mais usuais de diagnóstico em regressão normal não-linear são simples adaptações da regressão linear. Algumas dessas técnicas serão apresentadas nesta seção.

5.4.1 - Matriz de Projeção

No caso normal não-linear utiliza-se na detecção de pontos mais afastados dos demais, possivelmente pontos influentes, a matriz de projeção ortogonal no subespaço tangente a $\eta(\hat{\beta})$, dada por

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

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 $\hat{h}_{ii} \geq 2p/n$ continua sendo adotado como guia para detectar pontos suspeitos de serem influentes.

5.4.2 - Resíduo Projetado

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 instratável, principalmente para pequenas amostras. Além disso, os mesmos

em geral têm esperança diferente de zero e distribuição dependendo fortemente dos valores ajustados, o que pode levá-los a não refletirem exatamente a distribuição dos erros. Logo, nesses casos, os critérios de diagnóstico da regressão normal-linear podem falhar. Por exemplo, um resíduo muito diferente de zero, que segundo os critérios da regressão linear seria um ponto aberrante, pode agora não ser, caso o valor esperado desse seja também substancialmente diferente de zero.

Será definido a seguir um novo resíduo, que apesar de algebricamente ser mais complexo, tem propriedades mais próximas das do resíduo ordinário da regressão normal-linear.

Ao desenvolverem $\eta'(\hat{\beta})$ e $\eta(\hat{\beta})$ por série de Taylor em torno de β até primeira e segunda ordem, respectivamente, Cook e Tsai (1985) encontraram a seguinte aproximação para r:

$$r \cong (I-H)r - \tilde{X}\sum_{i=1}^{n} r_i W_i \Delta - \frac{1}{2}(I-H)\Delta^T W \Delta, \tag{5.7}$$

onde H é o projetor ortogonal em $C(\tilde{X})$ (subespaço gerado pelas colunas de $\tilde{X}$) e $\Delta = \hat{\beta} - \beta$.

Uma aproximação quadrática para r é obtida substituindo a primeira aproximação linear para r e Δ, respectivamente, na expressão (5.7) mostrando que

$$E(r) \cong (I-H)f$$

e

$$Cov(r, \eta(\hat{\beta})) \cong NN^T\sigma^2 - \mathrm{Var}(r),$$

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 de (5.7) 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 $\tilde{X}_{jk} = (\partial^2\eta_1/\partial\beta_k\partial\beta_j, \ldots, \partial^2\eta_n/\partial\beta_k\partial\beta_j)^T$ 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 (5.7), Cook e Tsai (1985) definiram o resíduo projetado

$$(I-H_2)r = (I-H)\varepsilon - (I-H_1)\varepsilon. \tag{5.8}$$

O primeiro termo de (5.8) 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 $q = \text{posto}(H_1)$ for pequeno em relação a $(n-p)$, então essa perda também será pequena. Independente disso, se a medida de não-linearidade intrínseca γ^{IN} for significativa, isto é, $\gamma^{IN} > 2F^{-1/2}$, o ganho será substancial.

De (5.8) mostra-se facilmente que

$$E\{(I-H_2)r\} = 0, \quad Var\{(I-H_2)r\} = \sigma^2(I-H_2)$$

e

$$E\{r^T(I-H_2)r\} = \sigma^2 tr(I-H_2).$$

Logo, uma estimativa alternativa para σ^2 é dada por

$$\tilde{\sigma}^2 = \frac{r^T(I-\hat{H}_2)r}{tr(H_2)}$$

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 de r sobre $(I - \hat{H})\tilde{X}_{kj}, \quad k, j = 1, \ldots, p$.

Na regressão linear, como foi visto no Capítulo 1, mesmo para erros não-correlacionados e de variância constante, os resíduos são correlacionados e com variância diferentes. São definidos então os resíduos studentizados que mesmo correlacionados, apresentam média zero e variância constante e igual a 1.

Similarmente, define-se agora $s = s\{(I - \hat{H}_1)r\}$ como sendo o vetor de resíduos projetados studentizados, cuja i-ésima componente será dada por

$$s_i = \frac{\{(I - \hat{H}_1)r\}_i}{\tilde{\sigma}\{(I - \hat{H}_2)\}_{ii}^{1/2}}, \quad i = 1, \ldots, n. \tag{5.9}$$

Cook e Tsai (1985) exibem para um exemplo particular o gráfico de $(t_i - s_i)$ contra os valores de uma única covariável e mostram os diferentes diagnósticos que são obtidos se os critérios utilizados para s_i forem também adotados para os resíduos ordinários studentizados t_i, $i = 1, \ldots, n$. Paula (1987) mostra como obter os s_i's pelo sistema *GLIM*.

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 $s_{(i)}$ contra

$\Phi^{-1}\left(\frac{i-3/8}{n+1/4}\right)$ pode ser útil, onde $\Phi(\cdot)$ é a função acumulativa da normal padrão.

5.4.3 - Medidas de Influência

As medidas de influência para o modelo normal não-linear são baseadas na regressão linear. A única diferença, que pode ser relevante, é a substituição da estimativa $\hat{\beta}(i)$ pela estimativa correspondente $\hat{\beta}^1_{(i)}$, que é obtida inicializando o processo iterativo (5.2) em $\hat{\beta}$ sem a i-ésima observação e tomando a estimativa de um passo. Como o método de Newton-Raphson utiliza em cada passo uma aproximação quadrática para $L(\beta)$, a estimativa $\hat{\beta}^1_{(i)}$ 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

$$\hat{\beta}^1_{(i)} = \hat{\beta} - \frac{(\tilde{X}^T\tilde{X})^{-1}}{(1-\hat{h}_{ii})}\tilde{x}_i r_i, \tag{5.10}$$

onde $\tilde{X}$ e $\tilde{x}_i$ são avaliados em $\hat{\beta}$ e $\tilde{x}_i$ é a i-ésima linha 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 agora dada 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 na Seção 5.2. Usando (5.10) na expressão acima, obtém-se a forma aproximada

$$D^1_i = \frac{\tilde{t}^2_i}{p}\frac{\hat{h}_{ii}}{(1-\hat{h}_{ii})},$$

onde $\hat{t}_i = r_i/\{s(1-\hat{h}_{ii})^{1/2}\}$ 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 contra a ordem das observações é usual, de vendo-se dar atenção àqueles pontos com o D_i^1 correspondente mais afastado dos demais. Se o interesse é detectar pontos influentes nas estimativas individuais $\hat{\beta}_j$, $j = 1,\ldots,p$, sugere-se o gráfico de $\Delta_i\hat{\beta}_j = (\hat{\beta}_j - \hat{\beta}_{(i)j})/DP(\hat{\beta}_j)$ contra a ordem das observações.

5.4.4 - Gráfico da Variável Adicionada

Como foi visto na Seção 2.7 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 mais sentido incluir um novo parâmetro na parte sistemática, que em muitos casos pode significar uma interação, do que uma nova covariável.

Suponha então o preditor não-linear $\eta(\beta)$ para o modelo reduzido e o preditor não-linear $\eta(\beta,\gamma)$ com o 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 $H\colon \gamma = 0$. A estimativa $\hat{\gamma}$ corresponde à estimativa do parâmetro da regressão linear simples, passando pela origem, de $y - \eta(\hat{\beta})$ sobre $(I - H)\tilde{X}_{\hat{\gamma}}$. Logo, o gráfico proposto pode revelar como as observações estão contribuindo nessa relação e como estão se afastando dela.

5.4.5 - Exemplos

1 - Obtenção do Resíduo Projetado

Suponha um modelo normal não-linear com ligação entre μ e η dada por

$$\log \mu = \eta, \text{ onde } \eta = \beta_1 + \beta_2 x_2.$$

Portanto tem-se

$$\tilde{X} = \left\{ \frac{d\mu}{d\eta} \; \frac{d\mu}{d\eta} x_2 \right\} = \{\exp(\eta) \; x_2 \exp(\eta)\}$$

e

$$W = \left\{ \frac{d^2\mu}{d\eta^2} \; \frac{d^2\mu}{d\eta^2} x_2 \; \frac{d^2\mu}{d\eta^2} x_2^2 \right\} = \left\{ \exp(\eta) \; x_2 \exp(\eta) \; x_2^2 \exp(\eta) \right\}.$$

Serão utilizados para ajustar esse modelo os dados abaixo (Draper e Smith, 1981), em que y é a fração média de cloro disponível num produto manufaturado e x_2 o tempo de fabricação do mesmo (em semanas):

y:	0.490	0.475	0.450	0.437	0.433	0.455
	0.423	0.407	0.407	0.407	0.405	0.393
	0.405	0.400	0.395	0.400	0.390	0.390

e

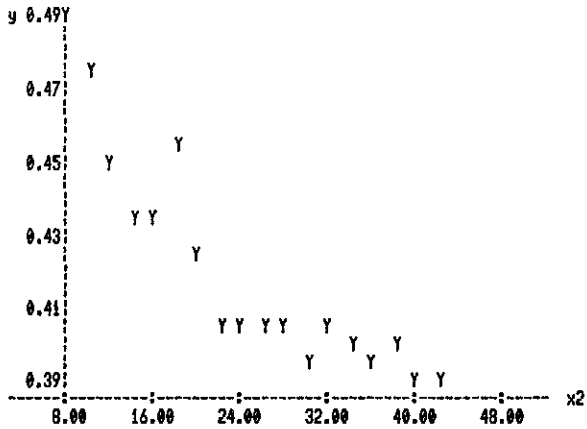

Figura 5.4: *Gráfico da fração média de cloro* (y) *contra o tempo de fabricação* (x_2).

$x_2 = 2(3+\ell)$, $\ell = 1, 2, \ldots, 18$. A Figura 5.4 exibe o gráfico de y contra x_2.

Utilizando o processo iterativo dado em (5.2) chega-se à convergência após 3 iterações, com as estimativas $\hat{\beta}_1 = -0.710(0.0186)$ e $\hat{\beta}_2 = -0.006(0.0007)$.

A partir desse ajuste foram obtidos os resíduos ordinários r_i's e a matriz de projeção $\hat{H}$. Note que as duas primeiras colunas de W pertencem a $C(\tilde{X})$. Logo, $(I - \hat{H})\tilde{X}_{11} = (I - \hat{H})\tilde{X}_{12} = 0$ e o vetor $\hat{H}_1 r$ corresponderá aos valores ajustados da regressão linear de r sobre $(I - \hat{H})\tilde{X}_{22}$, onde $\tilde{X}_{22}$ é a terceira coluna da matriz W. O vetor de resíduos projetados será então dado por $r - \hat{H}_1 r$. Como $q =$ posto $(H_2) = 1$, a perda de informação quando se passa do subespaço dos resíduos ordinários para o subespaço dos resíduos projetados, será pequena. Os resíduos projetados studentizados são obtidos diretamente de (5.9).

As Figuras 5.5 e 5.6 exibem, respectivamente, os gráficos de t_i contra x_{i2} e s_i contra x_{i2}, $i = 1, \ldots, n$.

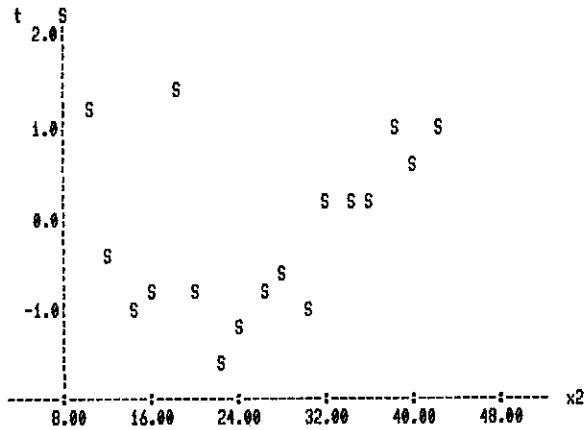

Figura 5.5: *Gráfico dos resíduos ordinários studentizados t_i's contra os valores x_2.*

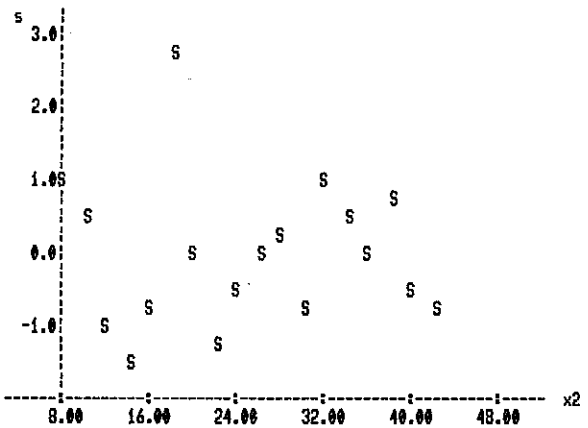

Figura 5.6: *Gráfico dos resíduos projetados s_i's contra os valores de x_2.*

Comparando essas figuras nota-se algumas divergências entre os diagnósticos produzidos pelos gráficos individuais, se forem adotados os mes-

mos critérios em cada um. Particularmente a observação ♯6 destaca-se como aberrante na Figura 5.6 o que parece estar em concordância com o posicionamento desse ponto na Figura 5.4.

2 - Técnicas de Diagnóstico para avaliar a interação **entre** duas drogas

Considere o modelo abaixo, que foi apresentado em (1) da Seção 5.1, usualmente utilizado para avaliar a mistura de duas drogas A e B, separadamente ativas,

$$y = \eta(\beta, k) = \alpha + \delta \log\{x_1 + \rho x_2 + k(\rho x_1 x_2)^{1/2}\} + \varepsilon, \tag{5.11}$$

onde $\varepsilon \sim N(0, \sigma^2)$ e $\beta = (\alpha, \delta, \rho)^T$. Para ilustrar esse modelo serão utilizados os dados apresentados em Darby e Ellis (1976), onde y é a conversão da glicose $(3 - 3H)$ para um certo tipo de lípide em células de ratos, e x_1 e x_2 são, respectivamente, as doses (p mol $^{-1}$) de dois tipos de insulina: (A) insulina na forma padrão e (B) insulina na forma suberoyl $A1 - B29$. Esses dados, Tabela 5.1, são apresentados em sete misturas diferentes, cada uma com duas doses totais, havendo quatro repetições para cada um dos 14 tratamentos formados. Para ajustar o modelo a esse conjunto de dados, foi utilizado o processo iterativo dado na Seção 5.2. As estimativas dos parâmetros estão na Tabela 5.2, e a correspondente tabela de análise da variância, Tabela 5.3, cujos resultados indicam que o modelo proposto é adequado para explicar a variável de resposta.

Tabela 5.1: *Dados do experimento com dois tipos de insulina.*

mistura	Razão da insulina padrão para a insulina A1-B29	Dose total ($pmol^{-1}$)	Repetições para cada tratamento
1	1:0	20.9	14.0 14.4 14.3, 15.2
		41.9	24.6 22.4 22.4 26.7
2	1:1.85	52.9	11.7 15.0 12.9 8.3
		106	20.6 18.0 19.6 20.5
3	1:5.56	101	10.6 13.9 11.5 15.5
		202	23.4 19.6 20.0 17.8
4	1:16.7	181	13.8 12.6 12.3 14.0
		362	15.8 17.4 18.0 17.0
5	1:50.0	261	8.5 9.0 13.4 13.5
		522	20.6 17.5 17.9 1.8
6	1:150	309	12.7 9.5 12.1 8.9
		617	18.6 10.0 19.0 21.1
7	0:1	340	12.3 15.0 10.1 8.8
		681	10.9 17.1 17.2 17.4

Fonte: Darby e Ellis (1976).

O principal interesse neste tipo de experimento é saber se há interação entre as duas drogas. Darby e Ellis (1976), McCullagh e Nelder (1983, Cap. 10) e recentemente Giltinan et al. (1988) analisaram os dados da Tabela 5.1 utilizando métodos diferentes, contudo, chegaram a conclusões muito parecidas. Particularmente, concluíram que as "misturas" extremas têm um peso disproporcional na estimativa da interação entre as drogas.

Tabela 5.2: *Estimativas do modelo (5.11).*

Efeito	Parâmetro	Estimativa	Desvio Padrão
Constante	α	-17.340	2.740
Inclinação	δ	10.540	0.790
Potência	ρ	0.046	0.003
Interação	κ	-0.301	0.120

Tabela 5.3: *Tabela de Análise de variância.*

Fonte	g.l.	Soma de quadrados	Quadrado médio	Estatística
Regressão	3	1061.44	353.81	83.64
Resíduo	52	220.18	4.23	
Falta de ajuste	10	65.40	6.54	1.77
Erro puro	42	154.78	3.68	

Será confirmada a seguir a conclusão acima através de três técnicas diferentes de diagnóstico, apresentadas neste capítulo. A primeira técnica, consiste na análise do gráfico dos resíduos ordinários studentizados t_i's contra a ordem das observações, que é exibido pela Figura 5.7. Pode-se notar nesse gráfico, duas observações (♯8 e ♯12) mal ajustadas, as quais poderão ser consideradas aberrantes se a distribuição dos resíduos for aproximadamente normal. A primeira dessas observações refere-se a uma "mistura" extrema.

A segunda técnica é uma análise da variação de primeiro passo da estimativa $\hat{k}$, quando uma observação é excluída. O gráfico de $\Delta_i^1\hat{k}$ contra

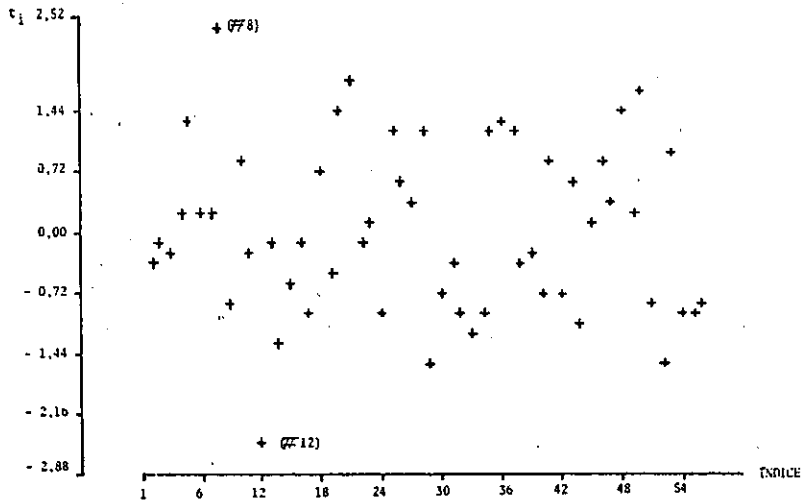

Figura 5.7: *Gráfico dos resíduos studentizados versus a ordem das observações.*

a ordem das observações é dado pela Figura 5.8. Nota-se no mesmo que três observações (♯8, ♯50 e ♯52) são mais influentes, sendo que cada uma individualmente, quando excluída, causa uma variação de aproximadamente 15% na estimativa $\hat{k}$. Agora as três observações são de "misturas" extremas.

Finalmente, na Figura 5.9 tem-se o gráfico de $y - \hat{\mu}$ contra $\hat{H}\tilde{X}_{\hat{k}}$ (gráfico da variável adicionada) para avaliar a influência conjunta das observações na estimativa $\hat{k}$, onde $\tilde{X}_{\hat{k}}$ é um vetor $n \times 1$ com as derivadas de $\eta(\beta, k)$ com relação a k, computado juntamente com H sob a hipótese $H\colon k = 0$.

Observando a Figura 5.9, nota-se que três observações (♯8, ♯50 e ♯5) têm uma influência disproporcional na inclinação da reta com coeficiente $\hat{k}$. De fato, a retirada desses pontos conduz à estimativa $\hat{k} = 0.160 (DP(\hat{k}) = 0.127)$, que significa um aumento de aproximadamente 47% apontando para uma interação nula entre as duas drogas contradizendo a conclusão da Seção 5.2.2. É provável que esse tipo de influência seja devida a doses extremas exageradas. Provavelmente a relação linear entre a log-dose e a resposta,

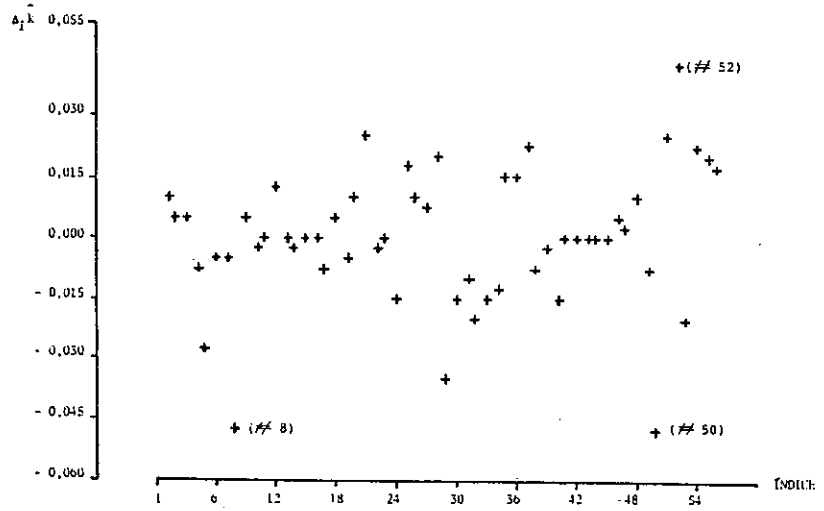

Figura 5.8: *Gráfico da variação em $\hat{k}$ versus a ordem das observações.*

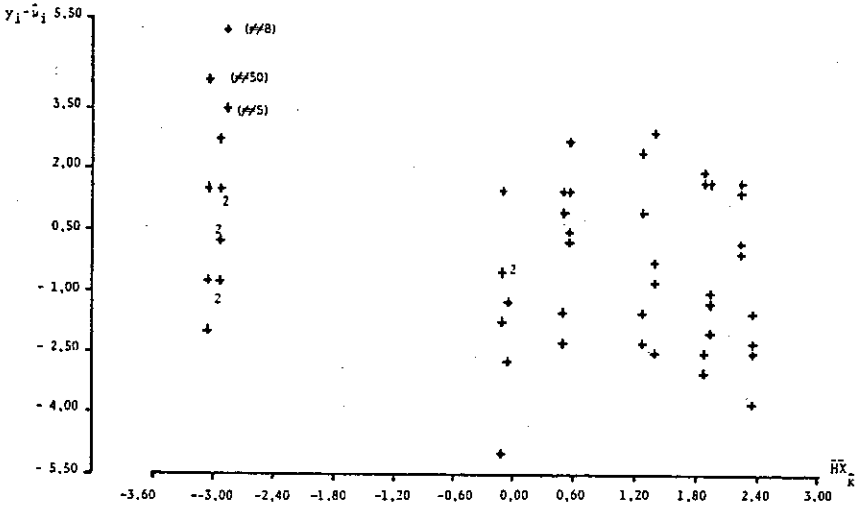

Figura 5.9: *Gráfico da variável adicionada correspondente à estimativa $\hat{k}$.*

que é assumida quando as drogas são analisadas separadamente, não deve continuar valendo nesses casos.

§5.5 Exercícios

1 - Os dados abaixo referem-se à idade aproximada (em anos) e ao comprimento (em cm) de peixes de 3 espécies segundo o sexo.

Idade (em anos)	Comprimento(cm) Espécie A		Espécie B		Espécie C	
	M	F	M	F	M	F
0	21.5	22.9	19.4	19.3	20.0	20.0
1	33.0	32.2	29.5	29.0	34.0	34.0
2	38.8	38.2	36.8	35.0	38.5	39.5
3	43.1	43.0	42.2	40.4	43.0	44.0
4	45.3	46.6	42.9	44.6	46.0	48.0
5	46.0	47.5	47.4	49.5		
6		50.2		51.6		
7		52.8		51.6		
8				56.9		
9				58.3		

M: macho, F: fêmea.

Resolver as seguintes questões:

(i) - Supondo o modelo de Von-Bertalanffy obter valores iniciais para os parâmetros α, δ e γ para cada um dos conjuntos de dados acima.

(ii) - Ajustar o modelo de Von-Bertalanffy a cada um desses conjuntos de dados, utilizando o processo iterativo dado na Seção 5.2 e os valores iniciais obtidos em (i).

(iii) - Após cada ajuste obter um intervalo de confiança de 95% para o comprimento máximo esperado e para a taxa média de crescimento.

(iv) - Verificar para cada espécie se o comprimento máximo e se a taxa média de crescimento diferem entre machos e fêmeas. Utilize nível de significância de 5%.

(v) - Explicite o cálculo do resíduo projetado para o modelo de Von-Bertalanffy.

2 - Mostre que o ponto de inflexão da curva MMF é dado por (x_F, μ_F), onde $x_F = \left\{\frac{\gamma(\delta-1)}{\delta+1}\right\}^{1/\delta}$ e $\mu_F = \{\beta(\delta+1) + \alpha(\delta-1)\}/2\delta$.

3 - Considere o seguinte conjunto de dados:

x	0	1	2	3	4
y	44.4	54.6	63.8	65.7	68.9

(i) - Obtenha os valores iniciais para o processo iterativo (5.2), supondo o modelo assintótico.

(ii) - Ajuste o modelo assintótico aos dados acima e encontre uma região de confiança de aproximadamente 95% para os parâmetros.

4 - Considere o seguinte modelo normal não-linear:

$$y = \delta\{1 - \exp(-\gamma x)\} + \varepsilon.$$

(i) - Obtenha valores iniciais para δ e γ.

(ii) - Expressar a região de confiança para (δ, γ) numa forma fechada e de fácil computação.

(iii) - Como fica o fator de correção de Bartlett para testar $H: \gamma = 0$?

(iv) - Ajuste esse modelo ao seguinte conjunto de dados:

x	1	2	3	4	5	7
y	4,3	8.2	9.5	10.4	12.1	13.1

(v) - Teste a hipótese $H: \gamma = 0$. Use $\alpha = 0.05$

5 - Ajuste o modelo logístico ao seguinte conjunto de dados:

x	0	1	2	3	4	5	6	8	10
y	1.23	1.52	2.95	4.34	5.26	5.84	6.21	6.50	6.83

Utilize a estatística F para testar a hipótese $H: \gamma = 1$. Use $\alpha = 0.05$

6 - Construir uma região de confiança para o modelo

$$y = \theta_1 x^{\theta_2} + \varepsilon,$$

e ajustar esse modelo ao seguinte conjunto de dados:

x	4	10	17	22	25
y	5	20	45	66	85

Testes a hipótese $H: \theta_2 = 1$ e adote um nível de significância de 5%.

7 - Ajuste o modelo apresentado por Ratkowsky (1983) para explicar, numa determinada reação química, a razão da reação y pelas concentrações x_1 e x_2 de dois reagentes (Seção 5.1) ao seguinte conjunto de dados:

x_1	x_2	y	x_1	x_2	y
1,0	1.0	0.126	3.0	0.0	0.614
2.0	1.0	0.219	0.3	0.0	0.318
1.0	2.0	0.076	3.0	0.8	0.298
2.0	2.0	0.126	3.0	0.0	0.509
1.0	0.0	0.186	0.2	0.0	0.247
3.0	0.0	0.606	3.0	0.8	0.319
0.2	0.0	0.268			

Faça uma análise completa de diagnóstico.

CAPÍTULO 6

MODELOS LINEARES GENERALIZADOS

§6.1 Definição

A teoria dos MLGs (Modelos Lineares Generalizados) vem desempenhando um papel importante na Estatística moderna devido ao grande número de métodos estatísticos que engloba. Ela tem sido objeto de estudo por especialistas e não-especialistas interessados em análise de dados univariados. Os MLGs representam ainda um meio unificado de ensino da Estatística, em qualquer curso de graduação ou pós-graduação. Esses modelos foram definidos por Nelder e Wedderburn (1972) e apresentam duas componentes: uma aleatória e outra sistemática.

Admite-se um vetor de observações $y = (y_1, \cdots, y_n)^T$ independentemente distribuídas com médias $\mu = (\mu_1, \cdots, \mu_n)^T$ e densidade na família exponencial de distribuições

$$f(y;\theta,\phi) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)], \tag{6.1}$$

onde $E(y) = \mu = b'(\theta)$ e Var $(y) = a(\phi)V$ com $V = b''(\theta) = d\mu/d\theta$ denominada *função de variância*, sendo expressa em termos da média μ. O

parâmetro natural $\theta = \int V^{-1} d\mu$ é função de μ, $\theta = q(\mu)$. Geralmente, $a(\phi) = \phi$ sendo ϕ interpretado como um parâmetro de dispersão. O parâmetro θ caracteriza a distribuição em (6.1).

As funções $a(\cdot)$, $b(\cdot)$ e $c(\cdot, \cdot)$ valem respectivamente: ϕ, $1/2\ \theta^2$ e $-1/2[y^2/\phi + \log(2\pi\phi)]$ para a distribuição normal com variância ϕ; 1, e^θ, $-\log y$ para a distribuição de Poisson; ϕ, $-\log(-\theta)$, $(\phi - 1)\log(y\phi) + \log\phi - \log\Gamma(\phi)$ para a distribuição gama com o quadrado do coeficiente de variação igual a ϕ; $1/\phi$, $\log(1 + e^\theta)$, $\log\binom{\phi}{\phi y}$ para a distribuição binomial com índice ϕ. Aqui $\Gamma(\cdot)$ representa a função gama. Entre outras distribuições que estão na forma (6.1) citam-se a binomial negativa e a normal inversa.

Essas distribuições são associadas aos seguintes tipos de dados: Poisson ($V = \mu$) e binomial negativa ($V = \mu + k\mu^2$) - contagens, binomial ($V = \mu(1 - \mu)$) - proporções, normal ($V = 1$) - dados contínuos, gama ($V = \mu^2$) e normal inversa ($V = \mu^3$) - dados contínuos positivos. Além de incluir várias distribuições a família exponencial (6.1) apresenta propriedades interessantes para estimação, testes de hipóteses e outros problemas de inferência. Assim, esta família permite incorporar dados que exibem assimetria, dados de natureza discreta ou contínua e dados que são restritos a um intervalo do conjunto dos reais. As distribuições gama e normal inversa são associadas a dados contínuos assimétricos. Se os dados exibem simetria e o intervalo de variação é o conjunto dos reais, a normal deve ser a escolhida. Quando os dados apresentam coeficiente de variação constante, deve ser preferida a distribuição gama. A distribuição de Poisson aplica-se aos dados na forma de contagens, mas pode também ser usada para análise de dados contínuos que apresentam variância aproximadamente igual à média. A distribuição binomial serve para análise de dados na forma de proporções, podendo ainda ser útil na análise de dados contínuos ou discretos apresentando subdispersão.

A componente sistemática admite a existência de uma função de ligação $g(\cdot)$ entre as médias das observações e a estrutura linear do modelo, através de:

$$g(\mu) = \eta = X\beta \tag{6.2}$$

onde $\eta = (\eta_1, \cdots, \eta_n)^T$, chamado preditor linear, é uma função linear dos parâmetros desconhecidos $\beta = (\beta_1, \cdots, \beta_p)^T$ e $g(\cdot)$, suposta conhecida, é diferenciável. Em geral $g(\cdot)$ é não-linear e X é uma matriz modelo, totalmente conhecida, e de posto completo. Possíveis termos em um preditor linear são dados a seguir:

Termo	Forma Algébrica	Forma no Modelo
contínuo (covariável)	λx	X
quantitativo (fator)	α_i	A
interação	$(\alpha\beta)_{ij}$	$A \cdot B$
cruzado	$\lambda x_1 x_2$	$X12$
misto	$\alpha_i x$	$A \cdot X$

Assim, na matriz modelo $X = \{x_{ir}\}$, de ordem $n \times p$, x_{ir} pode representar a presença ou ausência de um nível de um fator classificado em categorias, ou pode ser o valor de uma covariável quantitativa. A forma da matriz modelo representa matemáticamente o desenho do experimento. Uma covariável contínua x geralmente corresponde a um único parâmetro β, contribuindo com o termo βx para o modelo, enquanto uma variável qualitativa A, denominada frequentemente de fator, inclui na estrutura linear um conjunto de parâmetros α_i, onde i representa os níveis do fator. Então, na estrutura linear $\eta_i = \alpha_i + \beta x$, representando grupos distintos de um fator A mais uma covariável contínua x, a ordenada varia com o nível

do fator mas a declividade é a mesma. Entretanto, em alguns casos, a declividade deve variar com o nível do fator e, portanto, o termo βx deve ser substituído pelo mais geral $\beta_i x$, produzindo $\eta = \alpha_i + \beta_i x$. O termo $\beta_i x$ é denominado misto, pois a declividade associada com a covariável é suposta diferente para cada nível do fator.

A função de ligação relaciona o preditor linear η ao valor esperado μ do vetor y. A compatibilidade com a estrutura do erro e a interpretação do modelo devem ser levadas em consideração ao se escolher a ligação. As ligações usuais são: potência $\eta = \mu^\lambda$, onde λ é um número real, logística $\eta = \log[\mu/(1-\mu)]$, "probit" $\eta = \phi^{-1}(\mu)$, sendo $\phi(\cdot)$ a função de distribuição acumulada da normal reduzida, e o complemento $\log - \log$ $\eta = \log[-\log(1-\mu)]$. As três últimas funções são apropriadas para o modelo binominal, pois transformam o intervalo $(0,1)$ em $(-\infty, +\infty)$. Casos importantes da ligação potência são identidade, recíproco, raiz quadrada e logaritmo, correspondentes a $\lambda = 1$, -1, $1/2$ e 0, respectivamente.

Para o modelo clássico de regressão a ligação é a identidade no sentido de que valores esperados dos dados e preditores lineares podem ter qualquer valor real. Entretanto, quando os dados são contagens e a distribuição é de Poisson, a ligação identidade é menos atrativa pois, não restringe os valores esperados ao intervalo $(0, \infty)$.

Quando efeitos sistemáticos multiplicativos contribuem para as médias dos dados, uma ligação logaritmo torna os efeitos aditivos contribuindo para os preditores lineares e, portanto, pode ser a mais apropriada. A escolha de uma ligação compatível com a distribuição do erro proposta para os dados, deve resultar de considerações a priori, exame intensivo dos dados, facilidade de interpretação do modelo e, mais usualmente, uma mistura de tudo isso. Um preditor linear adequado com uma matriz modelo que forme um modelo

parcimonioso é, na prática, algo difícil de se obter, devido aos problemas de ordem combinatória e de ordem estatística. Entretanto, uma maneira de amenizar tal situação é a aplicação de técnicas de seleção de covariáveis. O analista, em geral, planeja um modelo intermediário entre o minimal que contém o menor número de termos necessários para o ajustamento e o modelo maximal que inclui o maior número de termos que se pode considerar.

Um grande número de casos especiais dos MLGs podem ser definidos por (6.1) e (6.2): o modelo clássico de regressão discutido no Capítulo 1 corresponde a $y \sim N(\mu, \sigma^2)$ e $\mu = \eta$; o modelo log-linear visto no Capítulo 3 é dado por $y \sim P(\mu)$ e $\log \mu = \eta$; o modelo logístico para análise de proporções estudado no Capítulo 4 pode ser obtido de $y \sim B(1, \mu)$ e $\log[\mu/(1-\mu)] = \eta$; e outros modelos familiares. Torna-se clara agora a palavra "generalizado", significando uma distribuição mais ampla que a normal para a variável resposta e uma função não-linear conectando a média desta variável com a parte determinística do modelo.

Na Seção 6.2 apresentam-se as etapas de trabalho com os modelos lineares generalizados. Nas Seções 6.3, 6.4 e 6.5 consideram-se a estimação, as medidas da qualidade do ajuste do modelo e a análise do desvio respectivamente. A Seção 6.6 trata de distribuições assintóticas e regiões de confiança, a 6.7 das técnicas de diagnóstico e a 6.8 do método das covariáveis adicionadas. Finalmente, na Seção 6.9 apresenta-se a análise dos dados da Tabela 2.1 através de um modelo gama.

§6.2 Etapas de Trabalho com os Modelos Lineares Generalizados

O processo de trabalho com os MLGs pode ser dividido em três etapas; (i) formulação dos modelos; (ii) ajustamento dos modelos; (iii) inferência. Maiores detalhes podem ser vistos em Cordeiro (1986). É difícil de propor uma estratégia geral para o processo de escolha do MLG a ser ajustado aos dados que se dispõe. Um ponto fundamental no processo de ajustamento é que não se deve ficar restrito a um único modelo, achando que ele é o mais importante e excluir outros alternativos. É prudente considerar a escolha restrita a um conjunto amplo de modelos estabelecidos por princípios como: facilidade de interpretação, boas previsões anteriores e conhecimento profundo da estrutura dos dados. Algumas características nos dados podem não ser descobertas, mesmo por um modelo muito bom e, portanto, um conjunto razoável de modelos adequados aumenta a possibilidade de se detectar essas características.

A etapa de ajustamento representa o processo de estimação dos parâmetros lineares dos modelos e de determinadas funções destes parâmetros, que representam medidas de adequação dos valores estimados. Vários métodos podem ser usados para estimar os parâmetros do MLG. Como o método de máxima verassimilhança conduz a uma estimação bastante simples, este método é o preferido. O algoritmo para solução das equações de máxima verossimilhança é similar a um processo iterativo de Newton-Raphson, tendo como característica principal o uso da matriz de valores esperados de derivadas segundas do logaritmo da verossimilhança (in-

formação), no lugar da matriz correspondente de valores observados. A implementação deste algoritmo está feita no software GLIM ("Generalized Linear Interactive Modelling").

Em geral, o algoritmo de ajustamento deve ser aplicado não a um MLG isolado, mas a vários modelos de um conjunto bem amplo. Este conjunto amplo deve ser realmente relevante para o tipo de dados que se pretende analisar e pode ser formulado de várias maneiras: (a) definindo uma família de ligações; (b) considerando diferentes opções para a escala de medição; (c) adicionando (ou retirando) vetores colunas independentes a partir de uma matriz básica original.

A etapa de inferência tem o objetivo principal de verificar a adequação do modelo como um todo e realizar um estudo detalhado quanto a discrepâncias locais. Estas discrepâncias, quando significativas, podem implicar na escolha de outro modelo ou em aceitar a existência de dados aberrantes. Em qualquer caso, toda a metodologia de trabalho deverá ser repetida. O analista deve, nesta etapa, verificar a precisão e interdependência das estimativas, construir regiões de confiança e testes sobre os parâmetros de interesse, analisar estatísticamente os resíduos e realizar previsões.

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

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

ses gráficos. Dados com erros grosseiros podem ser detectados como tendo grandes resíduos, ou o modelo ajustado deve requerer mais covariáveis, por exemplo, interações de ordem superior. A inspeção gráfica é um meio poderoso de inferência nos MLGs.

Para o teste do MLG utiliza-se o critério da razão de máxima verossimilhança em relação ao modelo saturado (desvio) e a estatística de Pearson generalizada. Toda a parte de inferência é baseada em resultados assintóticos e pouco se sabe sobre a validade desses resultados em amostras muito pequenas.

§6.3 Estimação

A idéia central dos MLGs é transformar as médias dos dados, no lugar de transformar as observações como nos modelos de Box e Cox (Capítulo 2) para se obter um modelo de regressão linear. Nelder e Wedderburn (1972) demonstraram que a solução das equações de máxima verossimilhança de um MLG equivale a calcular repetidamente uma regressão linear ponderada de uma variável dependente modificada sobre a matriz modelo X usando uma função de peso que se modifica no algoritmo iterativo. Esse processo converge rapidamente na maioria dos casos, exceto se n for pequeno.

O método de máxima verossimilhança é usado para estimar os parâmetros lineares $\beta_1, \cdots, \beta_p$ e, portanto, os preditores lineares $\eta_1, \cdots, \eta_n$ e os valores médios $\mu_1, \cdots, \mu_n$. Seja $L(\beta)$ o logaritmo da função de verossimilhança para um dado MLG e sejam $\hat{\beta}$, $\hat{\eta} = X\hat{\beta}$ e $\hat{\mu} = g^{-1}(\hat{\eta})$ as estimativas de máxima verossimilhança de β, η e μ, respectivamente. De agora por diante admite-se $a(\phi) = \phi$ com o parâmetro de dispersão ϕ constante para

todas as observações, embora possivelmente desconhecido. As equações de máxima verossimilhança $\partial L(\beta)/\partial\beta_r = 0$, $r = 1, \cdots, p$, são não- lineares exceto para erro normal ($V = 1$) com ligação identidade ($\eta = \mu$) e, portanto, não podem ser resolvidas explicitamente.

Seja K a matriz de informação de Fisher para β. Tem-se

$$K = \left\{-E\left(\frac{\partial^2 L(\beta)}{\partial\beta_r \partial\beta_s}\right)\right\} = \frac{1}{\phi} X^T W X, \tag{6.3}$$

onde $W = \text{diag}\left\{\frac{1}{V}\left(\frac{d\mu}{d\eta}\right)^2\right\}$. Expandindo a função escore

$$\partial L/\partial\beta = (\partial L(\beta)/\partial\beta_1, \cdots, \partial L(\beta)/\partial\beta_p)^T, \text{ onde}$$

$$\partial L(\beta)/\partial\beta_r = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)V_i^{-1}\frac{d\mu_i}{d\eta_i}x_{ir},$$

em série de Taylor e usando K no lugar da matriz de derivadas de 2ªordem $\left\{\frac{\partial^2 L(\beta)}{\partial\beta_r\partial\beta_s}\right\}$, obtém-se o processo iterativo

$$\frac{\partial L}{\partial\beta}^{(m)} = K^{(m)}(\beta^{(m+1)} - \beta^{(m)}),$$

supondo que $\beta^{(m)}$ é a m-ésima aproximação para $\hat{\beta}$ e usando uma notação similar para outras quantidades que variam no processo.

Demonstra-se que a nova aproximação $\beta^{(m+1)}$ para $\hat{\beta}$ satisfaz

$$K^{(m)}\beta^{(m+1)} = X^T W^{(m)} y^{*(m)}, \tag{6.4}$$

sendo K definido aqui sem o parâmetro ϕ e

$$y^* = \eta + H(y - \mu) \tag{6.5}$$

com $H = \text{diag}\ \{d\eta/d\mu\}$.

Logo, a solução das equações de máxima verossimilhança equivale a calcular repetidamente uma regressão linear ponderada de y^* sobre X usando W como uma função de peso. Observar que $\text{Cov}\ (y^*) = \phi W^{-1}$ e que W e y^* são modificados a cada etapa de (6.4). O GLIM utiliza (6.4) na determinação de $\hat{\beta}$. Quando a função de ligação é linear, $y^* = y$ e $W = \text{diag}\ \{d\theta/d\eta\}$. Observar que ϕ afeta a estrutura de covariância dos $\hat{\beta}'s$ mas não entra nas equações de determinação dos $\hat{\beta}'s$.

Para o modelo normal-linear W reduz-se à matriz identidade e a estimativa $\hat{\beta}$ é calculada exatamente. Como cada observação y representa uma estimativa da média μ correspondente, o meio natural de inicializar (6.4) é tomar como primeira aproximação $\mu^{(1)} = y$ e então calcular $y^{*(1)} = \eta^{(1)} = g(y)$, $W^{(1)}$ e $K^{(1)}$, obtendo-se uma segunda aproximação $\beta^{(2)}$ para $\hat{\beta}$, e continuando as iterações até o processo convergir. Em geral, o processo converge antes de 10 iterações. Além de $\hat{\beta}$, o GLIM fornece a estrutura de covariância assintótica $\hat{K}^{-1}$, os valores ajustados $\hat{\mu}_1, \ldots, \hat{\mu}_n$, os erros padrões das estimativas dos parâmetros, resíduos, etc.

Também o GLIM pode ser usado para obter transformações produzindo aproximadamente linearidade entre duas variáveis, geração de números aleatórios, análise gráfica e outras manipulações de dados. O leitor pode consultar o manual deste software disponível para microcomputadores de 16 bits (Baker e Nelder, 1978).

§6.4 Medidas da Qualidade do Ajuste

A etapa de inferência tem o objetivo principal de verificar a adequação do modelo como um todo e realizar um estudo detalhado quanto às discrepâncias locais. Estas discrepâncias, quando significativas, podem implicar na escolha de outro modelo. Neste caso, toda metodologia empregada deverá ser repetida. O analista deve, nesta etapa, verificar a precisão e interdependência das estimativas, construir regiões de confiança e testes sobre os parâmetros de interesse, analisar estatísticamente os resíduos e realizar previsões. Os métodos de inferência nos MLGs baseiam-se, fundamentalmente, na teoria assintótica de máxima verossimilhança (MV), pois, em geral, não é possível a obtenção de distribuições exatas para estimativas e estatísticas-testes. As condições de regularidade que garantem estes resultados são satisfeitas para os MLGs.

Ajustar um modelo a um conjunto de dados pode ser considerado como uma forma de substituir um conjunto de valores observados y por um conjunto de valores ajustados $\hat{\mu}$ deduzidos do modelo, envolvendo relativamente um número menor de parâmetros. Claro que os $\hat{\mu}$'s não são exatamente iguais aos y's e, então, a questão passa a ser o quanto discrepante são estes valores. Duas medidas desta discrepância são, usualmente, consideradas: a razão de máxima verossimilhança, conhecida como *desvio* e uma forma generalizada da estatística de adequação de Pearson.

Dois modelos são casos limites do processo de ajustamento: o modelo *nulo* e o *saturado*. O modelo nulo tem um único parâmetro μ representativo para todos os y's e toda a variação nos dados é devida à compo-

nente aleatória. O modelo saturado tem n parâmetros, um parâmetro para cada observação, ajusta-se exatamente aos dados, isto é, as estimativas das médias são iguais às próprias observações e toda a variação é devida à componente sistemática. Na prática procura-se um modelo com p parâmetros situado entre esses modelos limites. O modelo saturado não resume os dados e simplesmente os repete. Entretanto, este serve como um limite de discrepância para o modelo em investigação através da função desvio $D(y;\hat{\mu})$ definida por

$$D(y;\hat{\mu}) = 2\{L(y;y) - L(y;\hat{\mu})\}, \tag{6.6}$$

onde $L(y;\hat{\mu})$ é o máximo da log-verossimilhança para o modelo em investigação com p parâmetros e $L(y;y)$ é o máximo da log- verossimilhança para o modelo saturado. O parâmetro de dispersão nas expressões do desvio para os modelos normal, gama e normal inverso, aparece como um divisor.

Um modelo mal ajustado aos dados tem um grande desvio e um modelo bem adequado aos dados um pequeno desvio. Porém, um grande número de parâmetros significa um grau de complexidade na interpretação do modelo. Procura-se um modelo com poucas variáveis independentes tendo uma interpretação fácil e de desvio moderado. Os *graus de liberdade* associadas ao desvio são definidos por $v = n - p$. O desvio sem o parâmetro ϕ pode ser computado a partir dos dados e das estimativas de máxima verossimilhança $\hat{\mu}_1, \cdots, \hat{\mu}_n$. Como este é medido em relação ao modelo saturado, termos envolvendo constantes e os dados sozinhos não aparecem na expressão de $D(y;\hat{\mu})$.

Para os modelos normal com variância $\phi = \sigma^2$ e Poisson, o desvio iguala, respectivamente,

$$\sigma^{-2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \quad \text{e} \quad 2\left\{\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + \sum_{i=1}^{n}(\hat{\mu}_i - y_i)\right\}.$$

Logo, para o modelo normal o desvio é, simplesmente, a soma dos quadrados dos resíduos dividido pela variância.

O desvio funciona como critério de parada do algoritmo de ajustamento descrito em (6.4) e, após a convergência, o GLIM fornece diretamente o seu valor sem o parâmetro de dispersão ϕ e os seus graus de liberdade.

Para o teste de um MLG compara-se $D(y;\hat{\mu})$ e seus graus de liberdade v com alguma distribuição teórica de probabilidade. Geralmente, adota-se a distribuição qui-quadrado. A dificuldade em realizar este teste é que para os modelos normal, gama e normal inverso, o parâmetro ϕ, usualmente, é desconhecido. Quando o modelo em investigação é verdadeiro, o desvio não é, em geral, distribuído como $\mathcal{X}^2_{n-p}$, nem mesmo assintoticamente, e pouco se sabe sobre a adequação de uma aproximação do tipo $\mathcal{X}^2_{n-p}$ para pequenas amostras. Apesar disso, na prática, se contenta em testar um MLG sem muito rigor, comparando o desvio com o valor crítico $\mathcal{X}^2_{n-p}(\alpha)$ da distribuição qui-quadrado à um nível de significância igual a α. Se este for maior que $\mathcal{X}^2_{n-p}(\alpha)$, o modelo será rejeitado e, caso seja menor ou igual, aceito. Este teste pode ser aperfeiçoado através de um fator de correção e maiores detalhes podem ser vistos em Cordeiro (1983).

A estimação do parâmetro de dispersão por máxima verossimilhança é um processo simples. Entretanto, pode-se estimá-lo da maneira descrita a seguir. De agora por diante seja D_p o desvio de um modelo M_p com p parâmetros sem o parâmetro de dispersão ϕ, suposto constante. Considera-se o ajustamento de um *modelo maximal* aos dados com $m < n$ parâmetros e com desvio D_m. Deve-se esperar que um modelo bem ajustado aos dados tenha um desvio próximo dos seus graus de liberdade. Logo, $D_m/(n-m)$ deverá ser uma estimativa razoável de ϕ no teste de um modelo com $p < m$ parâmetros. Para os modelos normal e normal inverso esta estimativa coincide com a de máxima verossimilhança obtida resolvendo a

equação

$$\phi^2 \sum_{i=1}^{n} c'(y_i, \phi) = \sum_{i=1}^{n} [y_i q(y_i) - b(q(y_i))] - D_p/2. \tag{6.7}$$

A estimativa de máxima verossimilhança inserida em (6.6) possibilitará testar o modelo M_p em investigação comparando o desvio com χ^2_{n-p}. O teste sobre o parâmetro ϕ poderá ser visto em Cordeiro (1987a).

Uma outra maneira de testar a adequação do modelo M_p é através de uma estatística expressa pelo quociente

$$R = \frac{(D_p - D_m)/(m-p)}{D_m/(n-m)},$$

e que não envolve o parâmetro ϕ. Esta estatística pode ser aproximada por uma distribuição F com $(m-p)$ e $(n-m)$ graus de liberdade e o MLG poderá ser testado de maneira semelhante ao teste usual do modelo normal-linear. Como $(D_p - D_m)/\phi$ tem distribuição assintótica $\chi^2_{(m-p)}$, pois representa uma autêntica razão de verossimilhança entre hipóteses de dimensões finitas, apesar de cada desvio isoladamente não ter distribuição assintótica qui-quadrado, pode-se substituir o denominador $D_m/(n-m)$ na estatística R pela estimativa de ϕ obtida de (6.7) e considerar a nova razão como, aproximadamente, tendo distribuição F com $(m-p)$ e $(n-p)$ graus de liberdade.

§6.5 Análise do Desvio

A análise do desvio tem como objetivo a construção de uma seqüência de modelos encaixados e a verificação da significância dos termos adicionados. Considera-se agora dois modelos encaixados M_{p_1} e M_{p_2}, sendo o modelo M_{p_1} com matrix X_1, parâmetros $\beta_1, \cdots, \beta_{p_1}$ e desvio D_{p_1} e M_{p_2} com matrix X_2, contendo os mesmos parâmetros que M_{p_1} e parâmetros adicionais $\beta_{p_1+1}, \beta_{p_1+2}, \cdots, \beta_{p_2}$ e desvio D_{p_2}, ambos com a mesma ligação e distribuição. O teste da hipótese

$$H_0: \beta_{p_1} = \cdots = \beta_{p_2} = 0,$$

ou equivalentemente M_{p_1} versus M_{p_2}, é baseado no fato da razão de máxima verossimilhança $(D_{p_1} - D_{p_2})/\phi$ ter distribuição assintótica $\mathcal{X}^2_{p_2-p_1}$.

A partir de uma seqüência de modelos encaixados, obtidos pela adição de termos um a um, obtém-se uma correspondente seqüência de desvios decrescentes com os seus associados graus de liberdade. Assim, supondo a seqüência de modelos encaixados $M_{p_1}, M_{p_2}, M_{p_3}, \cdots, M_{p_k}$ com dimensões respectivas $p_1 < p_2 < p_3 < \cdots < p_k$ e desvios $D_{p_1} > D_{p_2} > D_{p_3} > \cdots > D_{p_k}$, todos os modelos com a mesma ligação e distribuição, constrói-se uma tabela de análise do desvio, conhecida como *ANODEV*. Esta tabela é uma generalização da tabela ANOVA usual. As diferenças de desvios e graus de liberdade fornecem a redução no desvio para a adição de cada termo sucedido e generaliza a análise de variância sequencial para modelos lineares. O modelo nulo contém somente a média, representada por 1, e o modelo saturado ou completo tem desvio zero.

TABELA ANODEV

Modelo	Desvio	G.l.	Desvio	G.l.	Termo
1	D_{p_1}	p_1			
			$D_{p_1} - D_{p_2}$	$p_2 - p_1$	A
$1 + A$	D_{p_2}	p_2			
			$D_{p_2} - D_{p_3}$	$p_3 - p_2$	A/B
$1 + A + B$	D_{p_3}	p_3			
			$D_{p_3} - D_{p_4}$	$p_4 - p_3$	$C/(A,B)$
$1 + A + B + C$	D_{p_4}	p_4			
...	...	...	...	...	...
saturado	0	0			

A estatística $(D_{p_i} - D_{p_j})$, $p_j > p_i$, é interpretada como uma medida da variação dos dados explicada pelos termos que estão em M_{p_j} e não estão em M_{p_i}, incluídos os efeitos dos termos em M_{p_i} e ignorando quaisquer efeitos dos termos que não estão em M_{p_j}. Se

$$(D_{p_i} - D_{p_j}) > \phi \ \chi^2_{p_j - p_i}(\alpha)$$

os efeitos dos termos que estão em M_{p_j} e não estão em M_{p_i} são significativos. Cada seqüência de modelos corresponde a uma tabela ANODEV diferente, sendo a seqüência determinada pelo interesse de obter modelos parcimoniosos e de pesquisar os efeitos de alguns termos na variação dos dados, quando outros termos já foram incluídos no modelo.

Numa análise de desvio dois termos A e B são *ortogonais* se a redução que A (ou B) causar no desvio de um modelo M_{p_i} for a mesma, quer B (ou A) esteja incluído ou não em M_{p_i}. Em geral, ocorre a não-ortogonalidade dos termos de um MLG e a interpretação da tabela ANODEV é mais complicada que a da tabela ANOVA usual.

§6.6 Distribuições Assintóticas e Regiões de Confiança

Pode-se demonstrar a normalidade assintótica de $\hat{\beta}$, com média igual ao parâmetro verdadeiro β desconhecido e matriz de covariância consistentemente estimada por

$$\hat{K}^{-1} = \phi(X^T \hat{W} X)^{-1}. \tag{6.8}$$

Os erros padrões das estimativas $\hat{\beta}_1, \cdots, \hat{\beta}_p$ são iguais às raízes quadradas dos elementos da diagonal de $\hat{K}^{-1}$ e fornecem informações valiosas sobre a exatidão dessas estimativas. Usando a notação $K^{-1} = \{-k^{rs}\}$ para a inversa da matriz de informação, os coeficientes calculados por

$$\hat{r}_{ij} = \frac{-\hat{k}^{ij}}{[\hat{k}^{ii}\hat{k}^{jj}]^{1/2}} \tag{6.9}$$

permitem verificar, pelo menos aproximadamente, a interdependência dos $\hat{\beta}_j$'s. Um intervalo de 100(1-α)% de confiança, aproximado, para um particular β_i, pode ser obtido de

$$\left[\hat{\beta}_i - z_{\alpha/2}(-\hat{k}^{ii})^{1/2}, \hat{\beta}_i + z_{\alpha/2}(-\hat{k}^{ii})^{1/2}\right], \tag{6.10}$$

onde $-k^{ii}$ é o elemento (i,i) de K^{-1} e $\Phi(-z_{\alpha/2}) = \alpha/2$. Com ϕ estimado deve-se usar a distribuição t de student com $(n-p)$ graus de liberdade no lugar da normal para computar (6.10). O teste de hipótese sobre um particular β_i pode ser baseado neste intervalo.

Desejando-se uma região de confiança aproximada para um conjunto particular de parâmetros $\beta_{i_1}, \cdots, \beta_{i_j}$ deve-se computar o desvio D_p do modelo M_p com todos os p parâmetros e o desvio D_{p-q} do modelo M_{p-q} com $p-q$ parâmetros linearmente independentes supondo que os q parâmetros de interesse têm valores fixados: $\beta_{ij} = \beta_{ij}^*$, $j = 1, \cdots, q$. Estes valores fixados formam a parte linear conhecida na estrutura do modelo dada por $\sum_{j=1}^{q} x_{ij}\beta_{ij}^*$, sendo x_{ij} a coluna correspondente a β_{ij}. Uma região de 100(1-α)% de confiança aproximada para $\beta_{i_1}, \cdots, \beta_{i_q}$ é dada pelo conjunto de pontos β_{ij}^*, $j = 1, \cdots, q$, não rejeitados pela estatística $\phi^{-1}(D_{p-q} - D_p)$

$$\{\beta_{ij}^*, j = 1, \cdots, q; \phi^{-1}(D_{p-q} - D_p) \leq \mathcal{X}_q^2(\alpha)\}, \tag{6.11}$$

onde $\mathcal{X}_q^2(\alpha)$ é o ponto da $\mathcal{X}_q^2$ correspondente a um nível de significância igual a α. Testes sobre um subconjunto de parâmetros pode ser feito com base em (6.11) desde que ϕ seja estimado.

A matrix de covariância assintótica das estimativas dos preditores lineares é dada, aproximadamente, por

$$\text{Cov}\,(\hat{\eta}) = X\ \text{Cov}\,(\hat{\beta})X^T = \phi X(X^T W X)^{-1}X^T. \tag{6.12}$$

O erro da aproximação é de ordem n^{-2}. Esta matriz, denotada por $Z = \{z_{ij}\}$, onde z_{ij} é o termo de ordem n^{-1} em $\text{Cov}\,(\hat{\eta}_i, \hat{\eta}_j)$ sem o parâmetro escalar ϕ, desempenha um papel importante na expansão assintótica do desvio e na determinação das tendenciosidades das estimativas de máxima verossimilhança dos parâmetros β, η e μ. A matriz de covariância assintótica dos valores ajustados $\hat{\mu}$ pode ser calculada da expansão em série de Taylor da inversa da função de ligação $\mu = g^{-1}(\eta)$, resultando

$$\text{Cov}\,(\hat{\mu}) = H^{-1}\ \text{Cov}\,(\hat{\eta})H^{-1} = \phi H^{-1} Z H^{-1}.$$

As estruturas de covariância de $\hat{\beta}$, $\hat{\eta}$ e $\hat{\mu}$ devem ser avaliadas no ponto $\hat{\beta}$. Consideram-se as distribuições assintóticas $N_n(X\beta, \phi Z)$ e $N_n(\mu, \phi H^{-1} Z H^{-1})$ como aproximações para as distribuições exatas de $\hat{\eta}$ e $\hat{\mu}$, respectivamente. Testes de hipóteses e regiões de confiança para os componentes de η e μ podem ser deduzidos dessas distribuições, mas, em geral, são obtidos de testes e regiões para os componentes de β. Observar que todas as covariâncias assintóticas dependem de ϕ, que se for desconhecido deverá ser estimado.

Os erros padrões $(\phi z_{ii})^{1/2}$ e os coeficientes de correlação estimados, $\hat{r}_{ij} = \hat{z}_{ij}/(\hat{z}_{ii}\hat{z}_{jj})^{1/2}$, de $\hat{\eta}_1, \cdots, \hat{\eta}_n$, são resultados aproximados, que dependem fortemente do tamanho da amostra. Entretanto, são guias úteis de informações sobre a confiabilidade e interdependência das estimativas dos preditores lineares e podem, também, ser usados para obter intervalos de confiança aproximados para esses parâmetros.

Uma estatística que também é usada para testar a adequação de um MLG e para construir regiões de confiança análogas à (6.11) é a versão da estatística de Pearson dada por

$$X_p^2 = \phi^{-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 / \hat{V}_i. \tag{6.13}$$

§6.7 Técnicas de Diagnóstico

Nesta seção são discutidas as principais técnicas de diagnóstico nos MLGs, tais como: análise de resíduos, análise global de influência e diagnóstico local de um único ponto influente.

§6.7.1 Análise de Resíduos

Por técnicas de diagnóstico entende-se a análise dos resíduos para detectar observações aberrantes e o estudo da influência de observações sobre o ajustamento global do modelo. Para a análise dos resíduos várias estatísticas podem ser propostas. Cox e Snell (1968) propuzeram a expressão do resíduo generalizado correspondente à observação y_i na forma

$$r_i = h_i(y_i, \hat{\mu}_i), \tag{6.14}$$

onde $h_i(y_i, \hat{\mu}_i)$ é uma função escolhida de modo a estabilizar a variância ou induzir simetria na distribuição amostral de r_i e $\hat{\mu}_i$ é o respectivo valor ajustado.

Várias formas para a função $h(\,\cdot\,,\,\cdot\,)$ têm sido propostas. O resíduo de Pearson é o mais simples definido como componente de (6.13) sem o escalar ϕ

$$p_i = (y_i - \hat{\mu}_i)/\hat{V}_i^{1/2}. \tag{6.15}$$

A desvantagem deste resíduo é que sua distribuição é, geralmente, bastante assimétrica para modelos não-normais. Da distribuição assintótica de $\hat{\beta}$ com estrutura de covariância $\phi(X^TWX)^{-1}$ pode-se concluir que Var $(\hat{\mu}_i)$ e Var $(y_i - \hat{\mu}_i)$ são, aproximadamente, iguais a $V_i\, h_{ii}$ e $V_i(1 - h_{ii})$, respectivamente, onde h_{ii} é o i-ésimo elemento da diagonal da matriz de projeção expressa por

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}. \tag{6.16}$$

Esta matriz desempenha na teoria dos MLGs o mesmo papel da matriz "hat" do modelo clássico de regressão. Notar que esta matriz independe de ϕ quando este é constante para as observações.

As considerações anteriores conduzem à definição do resíduo de Pearson studentizado com expressão

$$t_i = (y_i - \hat{\mu}_i)/\{\hat{V}_i(1 - \hat{h}_{ii})\}^{1/2}, \tag{6.17}$$

onde todas as quantidades estão estimadas em $\hat{\beta}$. O resíduo t_i é mais informativo que p_i. O gráfico dos resíduos t_i ordenados crescentemente versus os quantis da distribuição proposta para os dados permite verificar a adequação desta distribuição. Ainda, o gráfico de p_i ou t_i versus $\hat{\mu}_i$ serve para detectar pontos aberrantes no ajustamento enquanto que versus o índice i possibilita achar as observações que se mostram dependentes ou exibindo alguma forma de correlação serial. Para saber se algum parâmetro importante está sendo omitido do modelo, usa-se o gráfico dos resíduos versus a covariável ou os níveis do fator correspondente ao parâmetro omitido.

Outros tipos de resíduos têm interesses específicos. O resíduo que visa a normalização e estabilização da variância tem expressão

$$a_i = \{N(y_i) - N(\hat{\mu}_i)\}/N'(\hat{\mu}_i)\hat{V}_i^{1/2}, \tag{6.18}$$

onde $N(\,\cdot\,)$ é uma função escolhida de tal forma a aproximar a distribuição do resíduo à normal e $N'(\mu)V^{1/2}$ é uma aproximação para o desvio padrão de $N(y)$. O resíduo dado pela componente do desvio, que tem também o objetivo de normalização, é utilizado usualmente na forma $\delta_i d_i^{1/2}$, sendo δ_i o sinal de $y_i - \hat{\mu}_i$ e d_i a i-ésima componente do desvio dado pela expressão (6.6)

$$d_i = 2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}, \tag{6.19}$$

onde $\tilde{\theta}_i = q(y_i)$, $\hat{\theta}_i = q(\hat{\mu}_i)$ e $\theta = q(\mu)$ é a função que relaciona o parâmetro natural com a média. A vantagem deste resíduo é que ele não requer o conhecimento da função normalizadora. Observa-se que para modelos bem ajustados as diferenças entre $\delta_i d_i^{1/2}$ e t_i são pequenas enquanto que, de uma forma geral, os resultados com o uso de $\delta_i d_i^{1/2}$ e a_i são similares.

Todos esses resíduos são facilmente calculados através do GLIM ou GENSTAT.

§6.7.2 Análise Global de Influência

A idéia básica de influência é verificar a dependência do modelo estatístico sobre as várias observações ajustadas. Se uma pequena perturbação em algumas observações conduzir a uma mudança apreciável nas estimativas dos parâmetros β's, estas observações poderão ser consideradas influentes. As medidas usuais de influência geralmente correspondem às variações das estimativas decorrentes da eliminação de observações. Para exemplificar, a influência da observação y_i sobre o modelo é obtida pela distância de $\hat{\beta}$ a $\hat{\beta}_{(i)}$, onde $\hat{\beta}_{(i)}$ é a estimativa de β quando y_i é eliminada do modelo. Nos modelos lineares generalizados podemos usar as seguintes medidas desta distância:

$$(\hat{\beta}_{(i)} - \hat{\beta})^T [\text{ Cov } (\hat{\beta})]^{-1} (\hat{\beta}_{(i)} - \hat{\beta}), \tag{6.20}$$

$$2\{L(\hat{\beta}) - L(\hat{\beta}_{(i)})\}, \tag{6.21}$$

ambas as estatísticas tendo distribuição assintótica qui-quadrado com um grau de liberdade.

Pode-se demonstrar que estas medidas são, aproximadamente, iguais à estatísitca de Cook (1977) do modelo clássico de regressão (vide expressão (1.19)) dada por, sem a divisão por p,

$$C_i = \frac{(y_i - \hat{\mu}_i)^2 \hat{h}_{ii}}{\hat{V}_i(1-\hat{h}_{ii})^2} = \left(\frac{\hat{h}_{ii}}{1-\hat{h}_{ii}}\right) t_i^2. \tag{6.22}$$

As medidas de diagnóstico global da influência dos dados sobre o ajustamento são funções de resíduos e das estimativas dos elementos da diagonal da matriz de projeção.

Um gráfico de C_i versus i indicará as observações influentes sobre o ajustamento como um todo. Observar que C_i será grande quando $1 - \hat{h}_{ii}$ for pequeno. Assim, os valores das estimativas dos elementos da diagonal de $I - H$ menores do que $1 - 2p/n$, pois $tr(I - H) = n - p$, indicarão, provavelmente, quais os pontos influentes do modelo. Esses elementos dependem fortemente da matriz modelo e, em geral, muito pouco da estimativa de μ. Os gráficos de $1 - \hat{h}_{ii}$ versus i ou $\hat{\mu}_i$ são, frequentemente, usados para visualizar os pontos influentes.

§6.7.3 Diagnóstico Local de um Único Ponto Influente

As quantidades $\hat{h}_{ii}$ e C_i da seção anterior, embora possibilitem verificar quais as observações que dominam, em grande parte o ajustamento, não podem medir o *efeito local* sobre cada uma das estimativas $\hat{\beta}$'s do modelo. Nesta seção, apresenta-se uma fórmula aproximada para determinar o efeito local de cada observação sobre as estimativas dos parâmetros. Claro está

que a diferença $\hat{\beta}_{(i)} - \hat{\beta}$ proporciona uma medida do efeito local da i-ésima observação sobre as estimativas dos β's. Entretanto, para calcular esta diferença teríamos de ajustar $n+1$ modelos caso hajam n pontos a pesquisar as suas influências.

Através de expansões em série de Taylor até 1ªordem pode-se demonstrar (Pregibon, 1979; Cordeiro, 1986) que

$$\hat{\beta}_{(i)} - \hat{\beta} \doteq \frac{\hat{w}_i^{1/2}(y_i - \hat{\mu}_i)}{\hat{V}_i^{1/2}(1 - \hat{h}_{ii})}(X^T\hat{W}X)^{-1}x_i, \tag{6.23}$$

onde x_i^T é o vetor linha de X correspondente à i-ésima observação e $\hat{w}_i$ é o i-ésimo elemento da diagonal de W, com todas as quantidades sendo estimadas no ponto $\hat{\beta}$. A vantagem no uso desta expressão como uma medida de diagnóstico da influência local da observação i sobre as diversas estimativas dos parâmetros β's é que é necessário apenas o ajustamento do modelo em investigação.

Seja $\hat{\beta}_r$ uma estimativa de interesse e $\rho_r\hat{K}^{-1}$ a linha correspondente da matriz $\hat{K}^{-1}$, onde ρ_r é um vetor $1 \times p$ de zeros com 1 na r-ésima componente. O gráfico de $(\hat{\beta}_{(i)r} - \hat{\beta}_r)/\text{ Var }(\hat{\beta}_r)^{1/2}$ versus i, denominado *curva empírica de influência* sobre a estimativa $\hat{\beta}_r$, é o mais usado para detectar as observações que causam instabilidade local nas estimativas de interesse. Geralmente, esses gráficos discriminam que os mesmos dados causam as maiores instabilidades em todas as estimativas. Entretanto, as alterações nos valores das estimativas $\hat{\beta}$'s podem se compensar de tal maneira que os valores ajustados variam muito pouco. Neste caso, outros diagnósticos mais globais devem ser usados e os memos são baseados na razão de máxima verossimilhança, no desvio (6.6) ou na estatística de Pearson generalizada (6.13).

Os resultados de diagnóstico global (Seção 6.7.2) e de diagnóstico local (Seção 6.7.3) de influência de uma observação sobre o ajustamento podem ser generalizados para avaliar a influência de um conjunto de pontos. O leitor poderá consultar Cordeiro (1986) para extensões das fórmulas (6.22) e (6.23).

§6.8 Método das Covariáveis Adicionadas

O método das covariáveis adicionadas consiste em aumentar a estrutura linear do modelo através de covariáveis bastante adequadas para representar anomalias específicas no MLG. A forma mais comum do método teve origem no trabalho de Box e Tidwell (1962), que consideraram uma regressão com parâmetros não-lineares nas covariáveis. No preditor η aparecendo uma função $h(x;\gamma)$, onde γ é não-linear em x, expande-se a função em série de Taylor ao redor de um valor próximo conhecido $\gamma^{(0)}$, tornando γ um parâmetro linear na covariável adicional

$$\frac{\partial h(x;\gamma)}{\partial \gamma} \mid \gamma = \gamma^{(0)}.$$

No método, a estrutura linear do modelo aumentado é do tipo

$$g(\mu) = X\beta + Z\gamma, \tag{6.24}$$

onde $Z = (z_1, \cdots, z_q)$, z_i é um vetor coluna $n \times 1$ conhecido e $\gamma = (\gamma_1, \cdots, \gamma_q)^T$. Em casos especiais, as colunas z_i podem ser funções do ajuste do modelo usual, isto é, $Z = Z(X\hat{\beta})$, ou funções específicas das

covariáveis originais $z_i = z_i(x)$. A importância das covariáveis adicionadas é expressa pela diferença dos desvios dos modelos $g(\mu) = X\beta$ e (6.24). Se a adição das covariáveis $Z\gamma$ alterar substancialmente o ajustamento, as anomalias em questão afetarão seriamente o modelo original. Em geral, quando isto ocorre, as formas das covariáveis adicionais produzem uma ação corretiva.

Um bom exemplo do uso de uma covariável adicional está no teste para verificar a não-aditividade de um MLG, onde considera-se $(X\hat{\beta}) \otimes (X\hat{\beta})$, $\otimes$ é o produto direto, como uma covariável adicional. Se no ajuste do modelo aumentado o coeficiente desta covariável for significativamente diferente de zero, aceita-se a não- aditividade no modelo original. A transformação potência da variável resposta, pode ser uma medida corretiva para eliminar a não-aditividade. Para verificar se a escala de uma covariável isolada x_i está correta considera-se $\hat{\beta}_i^2(x_i \otimes x_i)$, onde $\hat{\beta}_i$ é o coeficiente estimado de x_i, como uma covariável adicional. Quando o coeficiente associado a esta covariável no ajuste do modelo aumentado é estatísticamente zero, aceita-se a linearidade de η em x_i. Um outro método gráfico, alternativo, é baseado na estatística $v_i = \hat{\beta}_i x_i + \hat{y}^* - \hat{\eta} = \hat{\beta}_i x_i + \hat{H}(y - \hat{\mu})$, que representa uma medida da linearidade da covariável x_i (McCullagh e Nelder, 1983, Capítulo 11). A estatística v_i é, simplesmente, um resíduo parcial generalizado para a covariável x_i expresso na escala da variável dependente modificada y^*. A escala de x_i será considerada correta, se o gráfico de v_i versus x_i for aproximadamente linear. Caso contrário, a forma do gráfico deve sugerir a ação corretiva.

Inferência sobre γ pode ser realizada a partir da redução no desvio do modelo com a inclusão de $Z\gamma$, ou através da distribuição normal assintótica de $\hat{\gamma}$, de média igual ao parâmetro verdadeiro γ e covariância $\phi\left[(Z^TWZ)^{-1} + L\{X^TWX - X^TWZL\}^{-1}L^T\right]$, onde $L = (Z^TWZ)^{-1}$

$Z^T W X$.

O método das covariáveis adicionadas é bastante usado para estimar a ligação e para identificar observações não-explicadas pelo modelo. No 1ºcaso, considera-se um MLG com função de ligação dependendo de um conjunto de parâmetros $\alpha^T = (\alpha_1, \cdots, \alpha_r)$ desconhecidos $\eta = g(\mu; \alpha) = X\beta$. Para o teste de $H_0: \alpha = \alpha_0$, onde α_0 é um vetor especificado, pode-se formar o modelo com ligação conhecida

$$g(\mu; \alpha_0) = X\beta + \hat{D}(\alpha_0 - \alpha), \tag{6.25}$$

onde agora o parâmetro desconhecido α aparece explicitamente como um conjunto adicional de parâmetros lineares,

$$D = \left\{ \frac{\partial}{\partial \alpha_i} g(\mu_i; \alpha) \right\}_{\alpha = \alpha_0}$$

é uma matriz $n \times r$ que depende basicamente do parâmetro β desconhecido e $\hat{D}$ é o seu valor correspondente à estimativa de máxima verossimilhança de β supondo H_0 verdadeira.

Assim, testar H_0 equivale a testar o modelo X contra o modelo alternativo $(X\hat{D})$, ambos os modelos com a mesma função de ligação $\eta = g(\mu; \alpha_0)$. Uma alternativa para verificar se a ligação é adequada, seria computar a redução no desvio após a inclusão da covariável $\hat{\eta} \otimes \hat{\eta}$. Se isto causar uma redução significativa no desvio a ligação não será satisfatória. Um outro método seria traçar o gráfico da variável dependente modificada estimada $\hat{y}^* = \hat{\eta} + \hat{H}(y - \hat{\mu})$ versus $\hat{\eta}$, onde $H = \text{diag}\{d\eta/d\mu\}$. Se o gráfico for aproximadamente linear, a ligação estará correta.

Faz-se agora a identificação de observações não explicadas pelo modelo usual $g(\mu) = X\beta$. Admite-se, então, que as observações $y_q = (y_{i_1}, \cdots, y_{i_q})^T$ não foram bem ajustadas. Define-se y_{n-q} como o vetor de observações

complementar de y_q e supõe-se a partição X_q, X_{n-q} para X, onde X_{n-q} é a submatriz obtida de X eliminando as linhas $i_1, \cdots, i_q$. Sejam ainda $\mu_q = E(y_q)$ e $\mu_{n-q} = E(y_{n-q})$ e usa-se uma notação similar para W e η. As equações de MV baseadas apenas nos dados y_{n-q} são

$$X_{n-q}^T \tilde{W}_{n-q}(y_{n-q} - \tilde{\mu}_{n-q}) = 0. \tag{6.26}$$

Seja $g(\mu) = X\beta + Z\gamma$ um modelo aumentado com todas as observações y, sendo $Z = (z_{i_1}, \cdots, z_{i_q})$ uma matriz de covariáveis adicionais e z_{i_j} um vetor coluna $n \times 1$ formado por zeros, exceto 1 na componente correspondente ao dado mal ajustado y_{i_j}, $j = 1, \cdots, q$. As equações de MV para o modelo aumentado, são

$$\begin{bmatrix} X^T \\ Z^T \end{bmatrix} \hat{W}(y - \hat{\mu}) = 0. \tag{6.27}$$

Considerando as q últimas equações vem, $\hat{\mu}_{i_j} = y_{i_j}$, $j = 1, \cdots, q$, isto é, a EMV da média da observação mal ajustada é igual à própria observação. Substituindo nas p primeiras equações, constata-se que as estimativas em $\hat{\mu}_{n-q}$ satisfazem equações idênticas à (6.26). Como $Z\gamma$ não contribui para os preditores lineares em η_{n-q}, as estimativas $\tilde{\mu}_{n-q}$ e $\hat{\mu}_{n-q}$ são iguais. Assim, a eliminação de um conjunto de pontos mal ajustados equivale a aumentar a matriz modelo com uma matriz adequada. Tem-se $\hat{\eta}_{n-q} = X_{n-q}\hat{\beta}$, $\hat{\mu}_{n-q} = g^{-1}(\hat{\eta}_{n-q})$, $\hat{\mu}_q = y_q$, $\hat{\eta}_q = g(y_q)$ e $\hat{\gamma} = g(y_q) - X_q\hat{\beta}$. A contribuição na log-verossimilhança de cada observação y_{i_j} mal ajustada, pode ser computada como o aumento no desvio resultante da eliminação da covariável z_{i_j} no modelo ampliado. Os vetores z_{i_j} que produzem aumentos significantes no desvio, correspondem às observações aberrantes do modelo original.

§6.9 Análise dos Dados da Tabela 2.1 Através de um Modelo Gama

Nesta seção apresenta-se uma análise alternativa dos dados da Tabela 2.1. Na Seção 2.8 foi proposto um modelo normal-linear para $\log V$. Aqui, considera-se V tendo distribuição gama. Para encontrar uma ligação adequada constrói-se gráficos da variável dependente modificada estimada ($\hat{y}^*$) versus preditores lineares estimados ($\hat{\eta}$). Entre as ligações identidade, recíproco e logaritmo, esta última apresenta gráfico de y versus $\hat{\eta}$ (Figura 6.1) mais próximo da 1ªbissetriz, sendo portanto a preferida. Ainda a ligação logaritmo produz menor desvio (6.109) em relação às demais ligações. Poder- se-ia aqui usar a expressão (6.25) para estimar a melhor ligação em uma família paramétrica de ligações.

As médias estimadas de V do modelo gama com ligação logaritmo são dadas por

$$\hat{\mu}_i = 0.449 \quad 1.015^{T_i} \quad 0.999^{P_i} \quad 1.059^{C_i} \quad 0.889^{N_i}$$
$$\times 1.123^{V_i} \quad 0.627^{R_{1i}} \quad 0.771^{R_{2i}}. \tag{6.28}$$

A expressão (6.28) é bastante próxima de (2.29) que corresponde ao modelo log-normal para V. Os valores ajustados segundo os dois modelos são, aproximadamente, os mesmos. O gráfico de V versus $\hat{\mu}$ é dado na Figura 6.2 e mostra uma linearidade ao longo da 1ªbissetriz, exceto para os maiores valores dos dados, como deveria ser esperado, pois Var (V) é proporcional a μ^2. Como visto na Seção 6.4, uma estimativa de ϕ é obtida dividindo o desvio pelos graus de liberdade; no caso, $\tilde{\phi} = 0.146$. Assim, o coeficiente de variação dos dados é estimado em $\tilde{\phi}^{1/2} = 0.38$ e, portanto, os dados estão 38% dispersos em relação às suas médias.

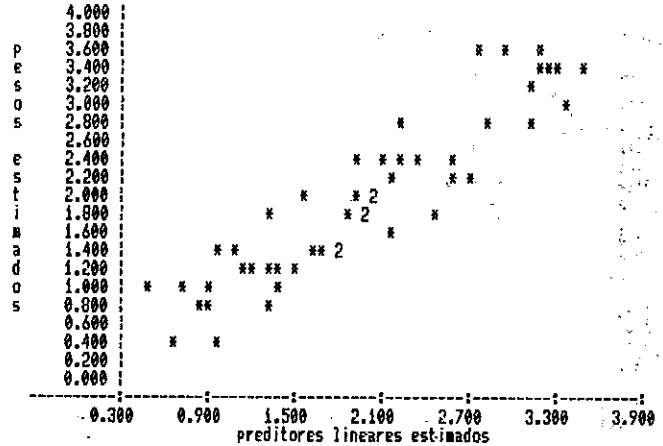

Figura 6.1 - *Gráfico de $\hat{y}^*$ versus $\hat{\mu}$ segundo um modelo gama para V.*

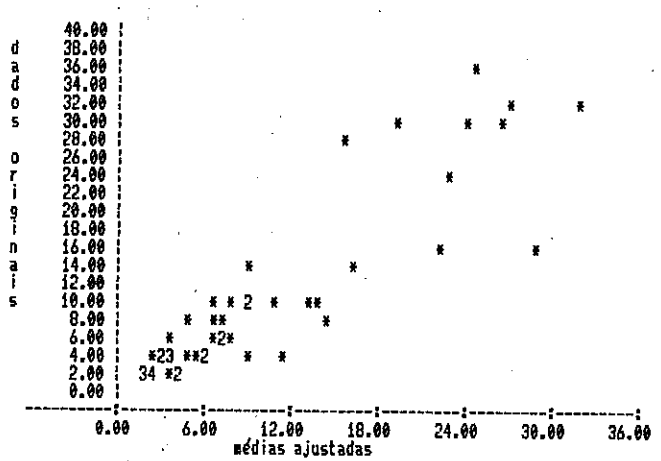

Figura 6.2 - *Gráfico dos dados V_i versus médias ajustadas $\hat{\mu}_i$ segundo um modelo gama para V.*

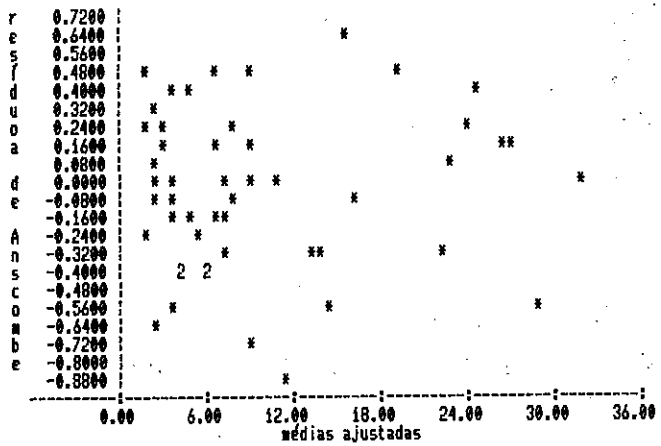

Figura 6.3 - *Gráfico dos resíduos de Anscombe versus médias ajustadas segundo um modelo gama para V.*

Adota-se aqui os resíduos de Anscombe (1953), deduzidos de (6.18) para o modelo gama, com expressão

$$a_i = 3(V_i^{1/3} - \hat{\mu}_i^{1/3})/\hat{\mu}_i^{1/3}. \tag{6.29}$$

O gráfico de a_i versus $\hat{\mu}_i$ mostrado na Figura 6.3 não revela dados aberrantes e os pontos se apresentam aleatóriamente distribuídos, sem nenhum padrão definido, podendo-se concluir que os erros são independentes. Os resíduos ordenados versus os quantis da $N(0,1)$ mostrados na Figura 6.4 suportam a distribuição gama para V.

As Figuras 6.5 e 6.6 apresentam os gráficos dos elementos estimados da diagonal da matriz projeção, definida em (6.16), versus as médias ajustadas e da estatística de Cook, expressão (6.22), versus o índice das observações, respectivamente. Os elementos estimados da diagonal de H são bem inferiores a $2p/n = 0.32$ e, portanto, não há indícios de observações influentes. A mesma conlusão é tirada da análise da Figura 6.6.

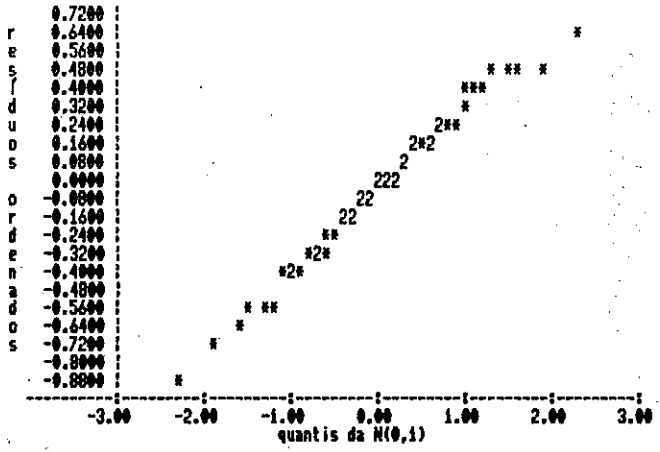

Figura 6.4 - *Gráfico dos resíduos ordenados versus os quantis da* $N(0,1)$ *segundo um modelo gama para* V.

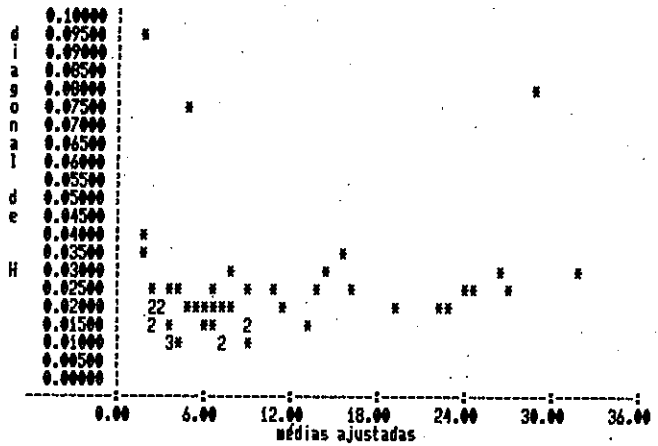

Figura 6.5 - *Elementos estimados da diagonal da matriz de projeção versus médias estimadas segundo um modelo gama para* V.

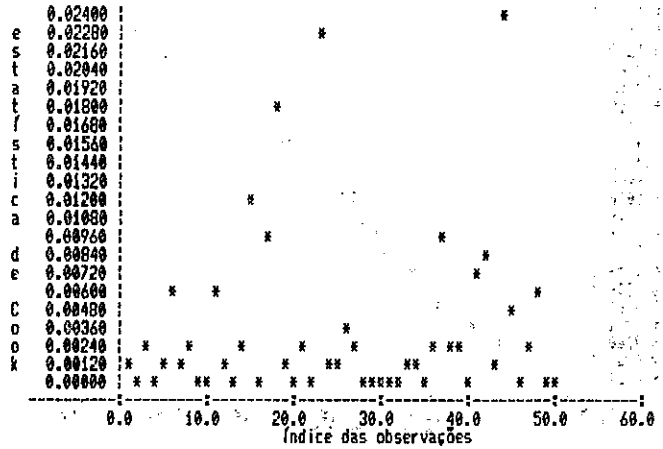

Figura 6.6 - *Estatística de Cook versus o índice das observações segundo um modelo gama para V.*

§6.10 Exercícios

1. Ajustar um modelo linear generalizado com erro gama ou normal inverso aos dados dos Exercícios 2 e 3 da Seção 2.9.
2. Usar a expressão (6.25) para estimar a ligação do modelo gama ajustado aos dados da Tabela 2.1.
3. Aplicar o algoritmo (6.4) aos seguintes modelos, calculando todas as quantidades que variam no processo iterativo:
   (a) modelo log-linear;
   (b) modelo logístico-linear;
   (c) modelo probit para estudo de proporções.

4. Considere um MLG supondo o parâmetro de dispersão ϕ constante para todas as observações. Determinar, através da função desvio, critérios para os seguintes testes:
   (a) $\phi = \phi^{(0)}$ versus $\phi \neq \phi^{(0)}$;
   (b) $\beta = \beta^{(0)}$ versus $\beta \neq \beta^{(0)}$ (Cordeiro, 1987a).
5. A Tabela abaixo representa os tempos de duração em dias de um produto alimentar perecível. Analisar estes dados no contexto dos MLGs.

$$24, 24, 26, 26, 32, 32, 33, 33, 33, 35, 41, 42, 43$$
$$47, 48, 48, 50, 52, 54, 55, 57, 57, 57, 60, 62, 70$$

6. Usando a expressão (6.18) calcular os resíduos de Anscombe para os modelos de Poisson e normal inverso.
7. Analisar os dados seguintes relativos aos números de moças da Nova Zelândia por faixa etária e estágio de desenvolvimento do busto (1=imaturo, 5=completamente desenvolvido) apresentados por Goldstein (1979).

		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	3	50	214	337	397	273
busto	4	7	72	160	333	501
	5	0	5	39	132	289

8. Comparar os valores do preditor linear para as ligações logística, "probit" e complemento log-log variando μ em (0,1).
9. Comparar os gráficos de $\eta = \log[\{(1-\mu)^{-\lambda} - 1\}/\lambda]$ versus μ para λ =-1, -0.5, 0, 0.5, 1 e 2.

10. Explicar como um modelo de Box-Cox poderia ser formulado no contexto dos MLGs.
11. Além das distribuições citadas neste capítulo, dar exemplo de mais 5 distribuições que pertencem à (6.1), definindo $q(\mu)$, V e ϕ.
12. Para os modelos normal, gama, normal inverso e Poisson com componente sistemática $\eta_\ell = \mu_\ell^\lambda = \alpha + \beta x_\ell$, e para o modelo binomial com $\eta_\ell = \log[\{(1-\mu_\ell)^{-\lambda} - 1\}\lambda^{-1}] = \alpha + \beta x_\ell$, sendo λ conhecido, calcular: (a) as estruturas de covariância assintótica de $\hat{\eta}$ e $\hat{\mu}$; (b) as estatísticas escore, de Wald e da razão de MV nos testes: $H_1: \beta = 0$ versus $H_1': \beta \neq 0$ e $H_2: \alpha = 0$ versus $H_2': \alpha \neq 0$; (c) intervalos de confiança para os parâmetros α e β.
13. Sejam $y_1, \cdots, y_n$ variáveis binárias independentes e identicamente distribuídas com $P(y_\ell = 1) = 1 - P(y_\ell = 0) = \mu$, $0 < \mu < 1$. A distribuição de y_ℓ pertence à família (6.1), com parâmetro natural θ. Demonstrar que a estatística de Wald para testar $H_0: \theta = 0$ versus $H: \theta \neq 0$ é $W = n\hat{\theta}^2 \exp(\hat{\theta})/\{1 + \exp(\hat{\theta})\}^2$, sendo os valores possíveis de $\hat{\theta}$ iguais a $\log\{t/(n-t)\}$, $t = 1, 2, \cdots, n-1$. Quais as formas das estatísticas escore e da razão de MV?
14. Demonstrar que em todo modelo linear generalizado com ligação $\eta = \mu^\lambda$ $(\lambda \neq 0)$ ou $\eta = \log \mu (\lambda = 0)$, neste último caso X com uma coluna de 1's, é satisfeita a relação

$$\sum_{i=1}^{n} \hat{V}_i^{-1}(y_i - \hat{\mu}_i) = 0.$$

15. Demonstrar que para os modelos normal e normal inverso supondo $\mu_1 = \cdots = \mu_n$, isto é, observações independentes e identicamente distribuídas, o desvio tem distribuição $\mathcal{X}^2_{n-1}$ segundo o modelo.

16. Demonstrar que para o modelo gama simples, em que todas as médias são iguais, o desvio reduz-se à estatística clássica $D_1 = 2n\phi^{-1}\log(\overline{y}/\tilde{y})$, onde $\overline{y}$ e $\tilde{y}$ são, as médias aritmética e geométrica dos dados $y_1, \cdots, y_n$, respectivamente.

CAPÍTULO 7

MODELOS NÃO-EXPONENCIAIS NÃO-LINEARES

§7.1 Uma Classificação dos Modelos de Regressão

Em todos os modelos estudados nos seis capítulos anteriores a variável de resposta tinha distribuição pertencente à família exponencial (6.1).

Introduz-se, nesta seção, modelos fora desta família. Admite-se um vetor de observações $y = (y_1, \ldots, y_n)^T$ independentemente distribuídas com y_i tendo densidade

$$f(y; \theta_i, \phi_i) = \exp\{t(y, \theta_i)/a(\phi_i) + c(y, \phi_i)\}, \tag{7.1}$$

onde $a(\cdot), t(\cdot,\cdot)$ são funções bem comportadas, conhecidas, com $a(\phi_i) > 0$ para $i = 1, \ldots, n$, suposto conhecido para cada observação. Considera-se uma estrutura determinística especificada por

$$g(\theta_i) = \eta_i = h(x_i; \beta), \tag{7.2}$$

onde $x_i = (x_{i1}, \ldots, x_{ip})^T$ é um vetor de valores conhecidos relativos a p covariáveis, $i = 1, \ldots, n$, $\beta = (\beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros desconhecidos e $h(\cdot;\cdot)$ é uma função contínua diferenciável arbitrária.

Em (7.1) θ é o parâmetro de principal interesse, não necessariamente igual à média de y. Qualquer distribuição de probabilidade que depende de um único parâmetro desconhecido pode, sem perda de generalidade, ser escrita na forma (7.1), bastando considerar $a(\phi) = \phi = 1$. Na grande maioria das aplicações $a(\phi) = \phi$ sendo o parâmetro ϕ o mesmo para todas as observações, embora possivelmente desconhecido, ou então $a(\phi_i) = \phi_i = \sigma^2/v_i$ sendo os v_i's números distintos conhecidos e σ^2 desconhecido. No contexto dos MLGs os v_i's são denominados de pesos a priori.

No caso de $t(y,\theta) = y\theta - b(\theta)$, (7.1) reduz-se à família exponencial (6.1), denominando este caso de *forma padrão* para $t(y,\theta)$. Define-se o modelo *não-exponencial não-linear* por (7.1) e (7.2) supondo que $t(y,\theta)$ é não-padrão e que $h(x;\beta)$ é não-linear em pelo menos um parâmetro β. Jørgensen (1983) chamou este modelo de classe estendida de modelos lineares generalizados, o qual foi, posteriormente, denominado por ele de *modelos de dispersão* (Jørgensen 1987a,b,1989). A palavra dispersão vem do papel análogo desempenhado por ϕ à variância do modelo normal.

A partir do modelo não-exponencial não-linear adota-se aqui a seguinte terminologia de modelos (Cordeiro e Paula, 1988, 1989a,b,c): (a) o *modelo exponencial não-linear* é definido supondo $t(y,\theta) = y\theta - b(\theta)$ e $h(x;\beta)$ não-linear; (b) o *modelo exponencial linear* (mais conhecido como *modelo linear generalizado*) é especificado por $t(y,\theta) = y\theta - b(\theta)$ e $g(\theta_i) = \eta_i = x_i^T\beta$. Neste caso, a média $E(y_i) = \mu_i$ pode ser escrita como função de η_i produzindo $F(\mu_i) = \eta_i = x_i^T\beta$, onde $F(\cdot) = g(b'^{-1}(\cdot))$ é a função de ligação (Seção 6.1); (c) o *modelo não-exponencial linear* é definido supondo que a função $t(y,\theta)$ não está na forma padrão e que $h(x_i;\beta) = x_i^T\beta$. Por extensão, o *modelo exponencial parcialmente não-linear* é definido por $t(y,\theta) = y\theta - b(\theta)$ sendo (7.2) reduzido à soma de uma componente linear com uma componente

não-linear, isto é,

$$g(\theta_i) = \eta_i = x_i^T \beta + h_i(z_i; \gamma).$$

Em muitos problemas reais, o analista não tem como eliminar a não-linearidade da componente sistemática do modelo e, portanto, torna-se importante o ajustamento dos modelos não-lineares. Mesmo a eliminação de não-linearidade por expansão de (7.2) em série de Taylor, pode implicar numa forte inadequação do modelo ajustado.

A Tabela 7.1 ilustra a classificação dos modelos de regressão quanto à distribuição pertencente ou não à família exponencial e a linearidade ou não-linearidade de $h(x;\beta)$. Assim, os modelos normais não-lineares com ligação identidade, vistos no Capítulo 5, são casos especiais, de grande importância prática, do modelo exponencial não-linear.

Os modelos exponenciais lineares (ou MLGs) foram estudados no Capítulo 6. Jørgensen (1987a,b, 1989) prefere chamar estes modelos de *modelos exponenciais de dispersão* e apresenta uma teoria elegante para estimação e testes de hipóteses, com propriedades assintóticas de interesse. Convém salientar que as pesquisas na área dos MLGs e extensões têm crescido a passos agigantados, principalmente após a publicação do livro de McCullagh e Nelder (1983). Recentemente, Cordeiro e Davison (1989) coletaram mais de 350 artigos publicados nesta área, entre técnicas e com aplicações.

Tabela 7.1: *Uma classificação dos modelos de regressão.*

$h(x;\beta)/t(y,\theta)$	Padrão	não-padrão
	(a) modelo exponencial não-linear	
	Alguns casos especiais:	modelo não-exponencial
não-linear	(a1) modelo exponencial parcialmente	não-linear
	não-linear	
	(a2) modelo normal não-linear	
	(b) modelo exponencial linear	(c) modelo não-exponencial
		linear
	Alguns casos especiais:	linear Alguns casos especiais:
	(b1) modelo normal linear	modelos normal,
	(b2) modelo log-linear	normal inverso
linear	(b3) modelos probit e logístico	e log-normal com
	(b4) modelo normal inverso linear	coeficientes de variação
	(b5) modelo gama com ligação potência	conhecido e estrutura
	(b6) modelo binomial negativo	linear para as médias

As seguintes propriedades dos modelos exponenciais de dispersão estão entre as mais importantes:

(i) Se Y tem distribuição na família exponencial (6.1) com média μ e variância $\phi V(\mu)$, então $(Y-\mu)/\phi^{1/2}$ converge em distribuição para $N(0, V(\mu))$ quando $\phi \to 0$.

(ii) Sejam $Y_1, \ldots, Y_n$ independentes em (6.1) com a mesma média μ e parâmetros de dispersão distintos $\phi_i = \sigma^2/v_i$, $i = 1, \ldots, n$; então $\sum_{i=1}^{n} v_i Y_i / v_+$ tem densidade em (6.1) com média μ e parâmetro de dispersão $\phi = \sigma^2/v_+$. Notar que se a densidade em (6.1) for normal, este

resultado será um caso particular da propriedade de convolução da normal: se $Y_1, \ldots, Y_n$ são independentes e $Y_i \sim N(\mu_i, \sigma^2/v_i)$, $i = 1, \ldots, n$, então $\sum_{i=1}^n v_i Y_i / v_+$ é normal de média $\sum_{i=1}^n v_i \mu_i / v_+$ e variância σ^2 / v_+.

Na Seção 7.2 alguns modelos não-exponenciais são apresentados. O algoritmo de ajustamento dos modelos não-exponenciais não-lineares é deduzido na Seção 7.3. A Seção 7.4 trata do teste de adequação do modelo. Em 7.5 a seleção de covariáveis é apresentada. As Seções 7.6 e 7.7 apresentam, respectivamente, algumas distribuições assintóticas e as principais medidas de diagnóstico. Na Seção 7.8 apresenta-se uma análise de dados através do modelo log-gama. Finalmente, na Seção 7.9, são discutidos testes de hipóteses e regiões de confiança para os modelos não-exponenciais não-lineares.

§7.2 Alguns Modelos Não-exponenciais

Os modelos não-exponenciais de regressão, embora não tão amplamente usados como os modelos exponencias, têm despertado grande interesse nos últimos anos, principalmente pelo desenvolvimento dos softwares estatísticos de ajustamento. A seguir apresentam-se vários modelos não- exponenciais definidos na forma (7.1):

(1) modelo log-gama com densidade $f(y; \theta, \phi) = c(\phi) \exp[\phi\{y - \theta - \exp(y - \theta)\}]$, $y > 0$, onde $c(\phi)$ é uma função normalizadora;

(2) modelo hiperbólico com densidade

$$f(y; \theta) = y^{-1} \exp[-\frac{1}{2}\{y^{-1} \exp(\theta) + y \exp(-\theta)\}]/2K_0(1),$$

onde $K_\nu(\cdot)$ é a função de Bessel com índice ν;

(3) modelo log normal inverso generalizado com

$$t(y,\theta) = \alpha(y-\theta) - \beta cosh(y-\theta),$$

supondo α e β conhecidos;

(4) família de modelos da forma $f(y;\theta,\phi) = \delta\phi^{1/\delta}\exp\{-\phi\,|y-\theta|^{\delta}\}/2\Gamma(\delta^{-1})$, sendo $\delta \neq 2$, onde $\Gamma(\delta) = \int_0^\infty x^{\delta-1}e^{-\delta x}dx$. Para $\delta = 1$ tem-se o modelo de Laplace;

(5) modelo logarítmico com densidade definida por

$$t(y,\theta) = y\log\theta - \log\{-\log(1-\theta)\},\ y = 1,2,\ldots \text{ e } 0 < \theta < 1;$$

(6) modelo em série de potências com densidade

$$f(y;\theta) = \exp\{\log a_y + y\log\theta - \log b(\theta)\},\ y = 0,1,2,\ldots,\theta > 0,\ a_y \geq 0$$

e

$$b(\theta) = \sum_{y=0}^{\infty} a_y\theta^y;$$

(7) modelo beta-$\beta(\phi\theta,\phi(1-\theta))$ com média θ e ϕ um parâmetro de dispersão;

(8) modelo de von Mises definido por $t(y,\theta) = \cos(y-\theta)$.

Supondo agora que $t(y,\theta)$ envolve um parâmetro c conhecido para cada observação, sendo $t(y,\theta) = t(y,\theta,c), p(\phi) = \phi = 1$ e $q(y,\phi) = q(y,c)$, vem ainda os seguintes modelos especiais:

(9) normal $-N(\theta,c^2\theta^2)$, log-normal $-LN(\theta,c^2\theta^2)$, normal inverso $-N^-(\theta,c^2\theta^2)$ com média θ e coeficiente de variação c;

(10) gama $-G(\theta,c)$ com média θ e parâmetro de escala c;

(11) Weibull $-W(\theta, c)$ com média θ e parâmetro de forma c;

(12) todos os modelos que dependem de um único parâmetro θ desde que $t(y, \theta)$ tenha qualquer forma diferente de $y\theta - b(\theta)$.

O leitor deve notar que os modelos definidos em (9) e (10) não pertencem à classe dos MLGs, pois a cada passo usa-se uma parametrização diferente.

Todos os modelos não-exponenciais (1) a (12) poderão ter componente linear ou não-linear para os parâmetros β's e, portanto, estarão classificados no lado direito da Tabela 7.1, nos cantos inferior e superior, respectivamente.

§7.3 Algoritmo de Ajustamento

Seja $y = (y_1, \ldots, y_n)^T$ o vetor de observações, supostas independentes, seguindo o modelo não-exponencial não-linear (7.1), (7.2). A log-verossimilhança para os parâmetros β's, considerada regular, é dada por

$$L(\beta) = \sum_{i=1}^{n} t(y_i, \theta_i)/a(\phi_i) + \sum_{i=1}^{n} c(y_i, \phi_i), \tag{7.3}$$

sendo θ uma função de β através de (7.2). Como as observações são independentes $L(\beta)$ é *aditiva* podendo ser escrita como a soma de n contribuições: $L(\beta) = \sum_{i=1}^{n} L\{\theta_i(\beta); y_i\}$.

A função escore $U(\beta) = \partial L(\beta)/\partial\beta$ pode ser expressa como

$$U(\beta) = \tilde{X}^T \Phi W H \dot{t}, \qquad \textit{onde} \tag{7.4}$$

$\Phi = \text{diag}\{a(\phi)^{-1}\}, W = \text{diag}\{-D_2(\theta)(d\theta/d\eta)^2\}, H = \text{diag}\{-D_2(\theta)^{-1} d\eta/d\theta\}$ com $D_2(\theta) = E\{t''(y, \theta)\}$, $\dot{t} = (t'(y_1, \theta_1), \ldots, t'(y_n, \theta_n))^T$ e $\tilde{X} = \frac{\partial \eta}{\partial \beta}$

sendo uma matriz $n \times p$ com elementos $\partial\eta_i/\partial\beta_s$. No caso de $h(\cdot;\cdot)$ ser linear $\tilde{X}$ reduz-se à matriz de planejamento, denominada aqui de X. Quando $t(y,\theta) = y\theta - b(\theta)$, W e H reduzem-se às definições dadas na Seção 6.3 e $\dot{t} = y - \mu$.

Seja $\hat{\beta}$ a estimativa de máxima verossimilhança de β. As equações $U(\hat{\beta}) = 0$ são, em geral, não-lineares e a solução $\hat{\beta}$ deve ser obtida por processos iterativos. O método de Newton-Raphson para resolver $U(\hat{\beta}) = 0$ é definido por

$$J(\beta^{(m)})(\beta^{(m+1)} - \beta^{(m)}) = U(\beta^{(m)}), \tag{7.5}$$

onde $J(\beta) = -\frac{\partial^2 L(\beta)}{\partial\beta^T \partial\beta}$ representa a matriz de derivadas de 2ª ordem da verossimilhança com sinal menos (matriz de informação observada de Fisher). O processo iterativo (7.5) é obtido por expansão de $U(\beta)$ em série de Taylor até 1ª ordem. Pode-se escrever

$$J(\beta) = -\sum_{i=1}^{n} \frac{\partial L(\beta)}{\partial\eta_i} \frac{\partial^2\eta_i}{\partial\beta\partial\beta^T} - \left(\frac{\partial\eta}{\partial\beta}\right)^T \frac{\partial^2 L(\beta)}{\partial\eta\partial\eta^T} \frac{\partial\eta}{\partial\beta}$$

e tomando valor esperado obtém-se a matriz de informação $K(\beta)$ de Fisher

$$E\{J(\beta)\} = K(\beta) = \tilde{X}^T W \Phi \tilde{X}, \tag{7.6}$$

pois $E\left\{\frac{-\partial^2 L(\beta)}{\partial\eta^T\partial\eta}\right\} = W\Phi$. Substituindo $J(\beta)$ por $K(\beta)$ em (7.5) visando a robustecer o algoritmo iterativo e usando (7.4) vem

$$K(\beta^{(m)})\beta^{(m+1)} = \tilde{X}^{(m)^T} \Phi W^{(m)} y^{*(m)}, \tag{7.7}$$

onde

$$y^* = \tilde{X}\beta + H\dot{t}. \tag{7.8}$$

A solução das equações de máxima verossimilhança para o modelo não-exponencial não-linear equivale, portanto, a calcular repetidamente uma regressão linear ponderada de uma variável dependente modificada y^* sobre a matriz $\tilde{X}$, com esta matriz e a *função de peso* $W\phi$ se modificando no processo iterativo. Nota-se que $\mathrm{Cov}(y^*) = W^{-1}\phi^{-1}$. A inicialização do processo iterativo pode ser feita a partir das estimativas $\hat{\hat{\theta}}_i$, $i = 1, \ldots, n$ segundo o modelo saturado, isto é, sem a componente sistemática (7.2). Observa-se que $\hat{\hat{\theta}}_i$ é a solução da equação $t'(y_i, \hat{\hat{\theta}}) = 0$, $i = 1, \ldots, n$. Entretanto, para modelos não-lineares, torna-se necessário uma escolha adicional para β a fim de inicializar $\tilde{X}$. Para os modelos exponenciais lineares, (7.7) e (7.8) reduzem-se à (6.4) e (6.5), respectivamente.

No caso do parâmetro ϕ ser constante este não afeta as estimativas dos β's. A matriz de informação particionada em β e ϕ é bloco- diagonal e, portanto, estes parâmetros são ortogonais.

Jørgensen (1984) apresenta uma classe de algoritmos, chamada *algoritmos delta*, que possibilita escolhas alternativas para a função de peso $W\Phi$, como as seguintes:

(i) pesos iguais aos elementos da matriz de informação observada $-J(\beta)$;

(ii) pesos iguais aos elementos da função escore $\partial L(\beta)/\partial\eta = W\Phi H\dot{t}$ divididos por $(\hat{\hat{\eta}}_i - \eta_i)$, sendo $\hat{\hat{\eta}} = g(\hat{\hat{\theta}})$;

(iii) pesos iguais aos quadrados dos elementos da função escore $\partial L(\beta)/\partial\eta$ divididos pelas componentes do desvio

$$2\{t(y_i, \hat{\hat{\theta}}_i) - t(y_i, \hat{\theta}_i)\}/a(\phi_i).$$

Ele discute critérios de escolha destes pesos incluindo extensões para o caso da log-verossimilhança não ser aditiva e sugere ainda incluir um escalar λ em (7.8), definindo $y^* = \tilde{X}\beta + \lambda H\dot{t}$, de forma a controlar o gradiente do

algoritmo e garantir $L(\beta^{m+1}) \geq L(\beta^{(m)})$.

A implementação do algoritmo delta pode ser feita em softwares com procedimentos de otimização do tipo (7.7), tais como, GLIM, BMDP (Dixon, 1981), GENSTAT e SAS. Green (1984) e Jǫrgensen (1984) apresentam sugestões para isto. Mais recentemente, Cordeiro (1987b) e Cordeiro e Paula (1986,1988,1989b,c) discutem uma maneira muito simples de implementar (7.7) no GLIM usando a diretiva OFFSET. Vários exemplos têm mostrado as potencialidades desta implementação.

Apesar do procedimento (7.7) ter sido deduzido para modelos definidos por (7.1) e (7.2), o mesmo se aplica ao ajustamento de vários outros modelos com pequenas modificações (Jǫrgensen, 1984; Green, 1984):

(a) modelos com parte linear do tipo $Y = X\beta + \sigma\varepsilon$, onde os ε_i's são variáveis aleatórias i.i.d. com qualquer distribuição de probabilidade;
(b) modelos com função de ligação composta (Seção 8.2);
(c) modelos de regressão para dados ordinais (McCullagh, 1980);
(d) estimação robusta e resistente;
(e) modelo de Box e Cox (Capítulo 2);
(f) estimação marginal e condicional da verossimilhança;
(g) modelos heterocedásticos (Seção 1.16);
(h) modelos envolvendo parâmetros não-lineares extras que não aparecem naturalmente na função de regressão (Nelder, 1985);
(i) modelos de quase-verossimilhança (Seção 8.7);
(j) modelos autoregressivos (Seção 8.9).

§7.4 Teste de Adequação

Considere um modelo M_p não-exponencial não-linear definido por (7.1) e

(7.2) com p parâmetros.

Sejam $\hat{L}_p$ o máximo da log-verossimilhança segundo M_p cujas estimativas são $\hat{\theta}_1, \ldots, \hat{\theta}_n$, e $\hat{L}_n$ a log-verossimilhança maximizada segundo o modelo saturado, cujas estimativas são $\hat{\hat{\theta}}_1, \ldots, \hat{\hat{\theta}}_n$. Para o teste de adequação do modelo ajustado usa-se uma das estatísticas

$$D_p = 2\sum_{i=1}^{n}\{t(y_i, \hat{\hat{\theta}}_i) - t(y_i, \hat{\theta}_i)\}/a(\phi_i) \tag{7.9}$$

ou

$$X_p^2 = \sum_{i=1}^{n}\{y_i - \hat{E}(y_i)\}/\hat{V}\mathrm{ar}(y_i), \tag{7.10}$$

onde $\hat{E}(y_i)$ e $\hat{V}\mathrm{ar}(y_i)$ são as estimativas de máxima verossimilhança, segundo M_p, da média e da variância de y_i. Qualquer uma dessas estatísticas mede a discrepância entre os dados $y_1 \ldots, y_n$ e os seus valores ajustados, sendo D_p uma óbvia extensão do desvio do MLG, denominada aqui de desvio do modelo M_p.

Para usar (7.9) ou (7.10) é necessário conhecer alguma estimativa consistente do parâmetro ϕ, como por exemplo a sua estimativa de máxima verossimilhança calculada de

$$a(\hat{\phi})^2 \sum_{i=1}^{n} c'(y_i, \hat{\phi}) - a'(\hat{\phi}) \sum_{i=1}^{n} t(y_i, \hat{\theta}_i) = 0. \tag{7.11}$$

Esta equação, em geral não-linear, pode ser expressa como uma função de ϕ e dos dados igual ao desvio do modelo M_p, reduzindo-se à (6.7) para os modelos exponenciais. Uma solução explícita para $\hat{\phi}$ em (7.11)

raramente é possível, como no modelo (4) discutido na Seção 7.2 quando $\hat{\phi} = -\delta^{-1} \sum_{i=1}^{n} \left| y_i - \hat{\theta}_i \right|^{\delta}$.

Tabela 7.2: *Estimativas de MV segundo alguns modelos saturados* ($\Gamma(\cdot)$ *é a função log-gama e* $\psi(\cdot)$ *a função digama)*

Modelo	Estimativa $\hat{\hat{\theta}}$	Modelo	Estimativa $\hat{\hat{\theta}}$
$N(\theta, c^2\theta^2)$	$\{(1+4c^2)^{1/2}-1\}y/2c^2$	$G(\theta, c)$	$\varphi(c\hat{\hat{\theta}}) = log(cy)$
$N^-(\theta, c^2\theta^2)$	$\{(4+c^2)^{1/2}+c^2\}y/2$	$W(\theta, c)$	$\Gamma(1+c^{-1})y$
$LN(\theta, c^2\theta^2)$	$(1+c^2)^{1/2}y$	$log-gama$	y

A Tabela 7.2 apresenta as estimativas $\hat{\hat{\theta}}$ para alguns modelos discutidos na Seção 7.2, mais especificamente, (1), (9), (10) e (11).

Para o teste do modelo M_p, definem-se os graus de liberdade por $\nu = n - p$.

O modelo será rejeitado se os valores das estatísticas (7.9), (7.10) forem superiores ao ponto $\chi^2(\alpha)$ da distribuição χ^2 correspondente ao nível de significância α. Este teste é aproximado e pouco se sabe sobre a sua adequação em pequenas amostras. Entretanto, o teste baseado no desvio pode, em princípio, ser aperfeiçoado através do cálculo do fator de correção de Bartlett. O valor esperado de D_p até termos de ordem n^{-1} para o cálculo do fator de Bartlett é conhecido para os modelos exponenciais lineares (Cordeiro, 1983) e não- lineares (Cordeiro e Paula, 1989a) e para os modelos não-exponenciais lineares (Cordeiro, 1985). Entretanto, ainda não se tem a expressão geral de $E(D_p)$ para os modelos não-exponenciais não-lineares, objeto de pesquisa atual dos autores deste texto. A demonstração do aperfeiçoamento dos testes pode ser visto em Cordeiro (1987a).

Quando ϕ for estimado uma aproximação do tipo χ^2_{n-p} para D_p poderá não ser adequada. Sendo assim, pode-se igualar o valor observado do desvio D_m de um modelo maximal M_m, bem ajustado aos dados, ao 1º momento da aproximação χ^2_{n-m}. Isto implicará na estimativa $\tilde{\phi}$ obtida de

$$a(\tilde{\phi}) = \frac{2}{(n-m)} \sum_{i=1}^{n} \{t(y_i, \hat{\hat{\theta}}_i) - t(y_i, \hat{\theta}_i)\}, \tag{7.12}$$

que poderá ser inconsistente, e na estatística-teste $R = \frac{(D_p - D_m)/(m-p)}{D_m/(n-m)}$ para o modelo M_p, onde aqui os D's são computados sem o escalar ϕ. O teste aproximado compara R com o ponto $F_{m-p,n-m}(\alpha)$ da distribuição F com $(m-p)$ e $(n-m)$ graus.

Toda a teoria de análise do desvio desenvolvida na Seção 6.5, em princípio, poderá ser usada aqui para testar efeitos de termos incluídos no preditor $\eta = h(x;\beta)$.

§7.5 Seleção de Covariáveis

A seleção de um conjunto de covariáveis para formar um modelo parcimonioso é, na prática, difícil devido a problemas de ordem combinatória e estatística. O problema combinatório é enumerar todas as combinações possíveis de covariáveis que podem ser selecionadas e o estatístico é ponderar, com a inclusão de mais parâmetros na componente sistemática, o efeito da redução do desvio do modelo com a complexidade dos parâmetros introduzidos no modelo. Para modelos não- lineares torna-se muito mais difícil medir o grau de complexidade da não- linearidade do modelo. Esta mesma

dificuldade existe na interpretação da não-exponencialidade da distribuição da variável resposta.

Como uma extensão da relação de covariáveis do modelo normal-linear, pode-se usar, como 1ª aproximação, o critério de informação de Akaile (1974) expresso por

$$AIC_p = D_p + 2p - 2\sum_{i=1}^{n} t(y_i, \hat{\hat{\theta}}_i)/a(\phi_i),$$

como medida de ajuste do modelo e o seu grau de complexidade. Entretanto, o uso deste critério deverá ficar restrito a modelos com medidas de não- linearidade (Beale, 1960) aproximadamente iguais, estando fixadas a distribuição em (7.1) e a ligação em (7.2).

A estatística AIC_p ajuda o analista na seleção de modelos complexos e tem demonstrado, pelo menos para modelos lineares, que produz soluções razoáveis para muitos problemas de seleção de modelos que não podem ser tratados pela teoria convencional de máxima verossimilhança. O gráfico de AIC_p versus p fornece uma boa indicação para a comparação de modelos. Considerando dois modelos encaixados M_{p_i} e M_{p_j}, $p_j > p_i$, o valor esperado de $AIC_{p_j} - AIC_{p_i}$, supondo M_{p_i} verdadeiro, é próximo de $p_j - p_i$, sendo o erro de ordem n^{-1}. Assim, na comparação de modelos sucessivamente mais ricos, o gráfico de AIC_p versus p revelará os pares de modelos (M_{p_i}, M_{p_j}) com declividade observada maior que 1 e, portanto, que o modelo maior (M_{p_j}) não é significativamente melhor que o menor (M_{p_i}).

Até termos de 1ª ordem deduz-se que

$$U(\hat{\beta}) \doteq U(\beta) - \tilde{K}(\hat{\beta} - \beta) = 0,$$

e

$$\hat{\beta} - \beta \doteq \tilde{K}^{-1} U(\beta) = (\tilde{X}^T W \Phi \tilde{X})^{-1} \tilde{X}^T \Phi W H \dot{t}, \tag{7.13}$$

sendo o erro desta aproximação de ordem $O_p(n^{-1})$. Propriedades de ordem superior relativas à estimativa $\hat{\beta}$, como por exemplo o vício de $\hat{\beta}$ até $1/n$, são deduzidas de (7.13). A distribuição assintótica $N_p(\beta, K(\beta)^{-1})$ de $\hat{\beta}$ pode ser usada para a construção de testes e intervalos de confiança, para os parâmetros $\beta_1, \ldots, \beta_p$. Alternativamente, testes e regiões de confiança podem ser baseados na função desvio.

A matriz de covariância assintótica de $\hat{\beta}$ estimada no ponto $\hat{\beta}$, isto é $K(\hat{\beta})^{-1}$, fornece informações valiosas sobre a interdependência dessas estimativas. Apesar de $\hat{\beta}$ não depender de ϕ quando este for constante para todas as observações, necessita-se de uma estimativa deste parâmetro para obter $\hat{K}(\hat{\beta}) = a(\phi)^{-1} \hat{\tilde{X}}^T \hat{W} \hat{\tilde{X}}^T$, onde os circunflexos indicam estimativas em $\hat{\beta}$. Sendo $-\hat{k}^{jj}$ o j-ésimo elemento da diagonal da matriz $K(\hat{\beta})^{-1}$ avaliada em $\phi = \hat{\phi}$ pode-se demonstrar que $(\hat{\beta}_j - \beta_j)/(-\hat{k}^{jj})^{1/2}$ converge em distribuição para a t de Student com $n - p$ graus quando $n \to \infty$, supondo o modelo verdadeiro. Este resultado permite testar hipóteses sobre o parâmetro β_j.

Representando agora por $L = L(\phi, \beta)$ a log-verossimilhança como função de ϕ e β, é fácil ver que a matriz de informação particionada

nesses parâmetros tem estrutura bloco-diagonal e, portanto, as estimativas $\hat{\phi}$ e $\hat{\beta}$ têm distribuições normais assintóticas independentes. Como $K_\phi = E(-\partial^2 L/\partial\phi^2) = m(\phi)\sum_{i=1}^n E\{t(y_i,\theta_i)\} + \sum_{i=1}^n E\{c''(y_i,\phi_i)\}$ com $m(\phi) = \{2a'(\phi)^2 - a''(\phi)^2 a(\phi)\}/a(\phi)^3$, tem-se a convergência assintótica de $\hat{\phi}$ para $N(\phi, K_\phi^{-1})$, devendo K_ϕ ser estimado em $\hat{\beta}, \hat{\phi}$.

A matriz de covariância assintótica dos preditores $\eta_1, \ldots, \eta_n$ é dada por $C_{\hat{\eta}} = a(\phi)\tilde{X}(\tilde{X}^T W \tilde{X})^{-1}\tilde{X}^T$, sendo estimada no ponto $\hat{\phi}, \hat{\beta}$. Também a matriz $C_{\hat{\theta}} = a(\phi)G\tilde{X}(\tilde{X}^T W \tilde{X})^{-1}\tilde{X}^T G$ com $G = \text{diag}\{\frac{d\theta}{d\eta}\}$, representa a estrutura de covariância assintótica das estimativas $\hat{\theta}_1, \ldots, \hat{\theta}_n$.

As distribuições assintóticas $N_n(\eta, C_{\hat{\eta}})$ e $N_n(\theta, C_{\hat{\theta}})$ para $\hat{\eta}$ e $\hat{\theta}$, respectivamente, possibilitam fazer inferência sobre os parâmetros η e θ. Embora, usualmente, a inferência fique restrita ao parâmetro β.

§7.7 Medidas de Diagnóstico

As chamadas medidas de diagnóstico dos modelos lineares generalizados (Seção 6.7) são agora estendidas para os modelos não-exponenciais não-lineares. Inicialmente, os resíduos para estes modelos podem ser definidos como iguais à raízes quadradas $d_i^{1/2}$ das componentes do desvio com o sinal dado por $y_i - \hat{E}(y_i)$,

$$d_i = 2\{t(y_i, \hat{\hat{\theta}}_i) - t(y_i, \hat{\theta}_i)\}, \tag{7.14}$$

o qual se reduz a (6.19) para os modelos exponenciais.

A expressão (7.14) representa uma distância da observação y_i ao seu valor ajustado $\hat{E}(y_i)$ medida na escala da log-verossimilhança. Um valor

grande para d_i indica que a i-ésima observação está mal ajustada pelo modelo. Claramente $D_p = \sum_{i=1}^{n} d_i$. Os resíduos $\pm d_i^{1/2}$ têm, aproximadamente, distribuição normal reduzida. Para calcular (7.14) a única dificuldade é obter as estimativas dos θ's segundo o modelo saturado.

Considere que $\hat{E}(y_i) = \hat{\rho}_i y_i$ sendo $\hat{\rho}_i$ obtido do ajustamento do modelo em investigação com p parâmetros. A seguir apresentam-se expressões para d_i em alguns modelos não-exponenciais discutidos na Seção 7.2:

(a) modelo normal $-N(\theta, c^2\theta^2)$ com média θ e coeficiente de variação c, constante para todas as observações e conhecido. Tem-se $t(y,\theta) = -y^2/2c^2\theta^2 + y/\theta c^2 - \log\theta$ e $\hat{\hat{\theta}} = y\alpha$, onde $\alpha = [(1+4c^2)^{1/2} - 1]/2c^2$. Assim, (7.14) implica em

$$d_i = (\alpha^2 - \hat{\rho}_i^2)/2c^2\alpha^2\hat{\rho}_i^2 + (\hat{\rho}_i - \alpha)/2\hat{\rho}_i c^2 + \log(\hat{\rho}_i/\alpha).$$

(b) modelo log-normal $-LN(\theta, c^2\theta^2)$ com média θ e coeficiente de variação c, constante para todas as observações e conhecido. Tem-se $t(y,\theta) = -\{\log(y/\theta) + \frac{1}{2}\log(1+c^2)\}^{1/2}\log(1+c^2)$ e então $\hat{\hat{\theta}} = y(1+c^2)^{1/2}$. Assim,

$$d_i = \log\{(1+c^2)^{1/2}/\hat{\rho}_i\}/\log(1+c^2).$$

(c) modelo gama $-G(\theta, c)$ com média θ e parâmetro de escala c constante e conhecido para todas as observações. Sabe-se que $t(y,\theta) = c\theta\log y + c\theta\log c - \log\Gamma(c\theta)$. Assim, $\hat{\hat{\theta}}$ vem de $\psi(c\hat{\hat{\theta}}) = \log(cy)$, onde $\psi(\cdot)$ é a função digama. Para c grande $\hat{\hat{\theta}} \doteq y + 1/2$ e, portanto, (7.14) reduz-se a

$$d_i = c\log(y_i c)\{(1-\hat{\rho}_i)y_i + 1/2c\} + \log\left\{\frac{\Gamma(c\hat{\rho}_i y_i)}{\Gamma(cy_i + 1/2)}\right\}.$$

(d) modelo de Weibull $-W(\theta, c)$ com média θ e parâmetro de forma c constante e conhecido para todas as observações. Aqui $t(y,\theta) =$

$-\{\Gamma(1+c^{-1})\}^c(y/\theta)^c - c\log\theta$ e $\hat{\hat{\theta}} = \Gamma(1+c^{-1})y$. Logo,

$$d_i = c\log\left\{\frac{\hat{\rho}_i}{\Gamma(1+c^{-1})}\right\} + \left\{\frac{\Gamma(1+c^{-1})}{\hat{\rho}_i}\right\}^c - 1.$$

(e) modelo log-gama definido por $t(y,\theta) = y-\theta-\exp(y-\theta)$. Tem-se $\hat{\hat{\theta}} = y$ e então,

$$d_i = [\exp\{1-\hat{\rho}_i)y_i\} - (1-\hat{\rho}_i)y_i].$$

Em todos esses exemplos as expressões dos resíduos são facilmente calculadas nos programas GLIM e GENSTAT.

A análise gráfica dos resíduos $\pm d_i^{1/2}$ permite detectar anomalias locais no modelo, tais como, uma falsa distribuição para Y, uma ou mais observações que não pertencem à distribuição proposta para os dados, omissão de covariáveis ou fatores importantes e correlação serial entre observações (Seção 6.7.1).

Os resíduos de Pearson em (6.15) e de Pearson studentizado em (6.17) podem ser estendidos para os modelos não-exponenciais não-lineares como $p_i = \{y_i - \hat{E}(y_i)\}/\hat{V}\text{ar}(y_i)^{1/2}$ e $t_i = p_i/(1-\hat{h}_{ii})^{1/2}$, sendo $\hat{h}_{ii}$ o i-ésimo elemento de diagonal da matriz de projeção local $\hat{H} = \hat{W}^{1/2}\hat{\tilde{X}}(\hat{\tilde{X}}^T\hat{W}\hat{\tilde{X}})^{-1}\hat{\tilde{X}}^T\hat{W}^{1/2}$ estimada em $\hat{\beta}$. Os resíduos t_i têm aproximadamente, média zero e variância um. Para os modelos exponenciais Pregibon (1982) mostra que t_i^2 é uma estatística escore para testar a hipótese de que a observação i é um ponto aberrante.

Para os modelos não-exponenciais o cálculo dos resíduos p_i e t_i, dependendo da parametrização do modelo, pode ser mais trabalhoso do que o cálculo de d_i, que é diretamente obtido na 4ª macro de ajustamento do modelo do usuário (Baker e Nelder, 1978; Cordeiro e Paula, 1988, 1989b,c). A

extensão do resíduo (6.18) para os modelos não- exponenciais é mais difícil devido ao problema de encontrar a função normalizadora $N(\cdot)$.

Entre todos esses resíduos aqueles definidos pela expressão (7.14) devem estar mais próximos da normalidade e têm uma interpretação simples em termos de contribuições para o desvio, sendo portanto os preferidos para uso na prática.

As medidas mais comuns de diagnóstico são definidas a partir da eliminação de algumas observações do conjunto de dados que se analisa. A eliminação de um único dado é mais fácil de se analisar, embora pesquisas recentes tenham sido dirigidas para o problema da eliminação de duas ou mais observações (Atkinson, 1986; Cook, 1986). Como visto na Seção 6.7.2 as expressões (6.20) e (6.21) são usadas para medir a influência global da observação i sobre o ajustamento do modelo exponencial. Estas expressões poderão ser aplicadas ao estudo da influência global no modelo não-exponencial, apesar da estatística (6.22) não ser usada com este objetivo.

Ainda com a finalidade de aperfeiçoar o diagnóstico do modelo, pode-se usar gráficos de variáveis adicionais para encontrar observações, se estas existirem, que impliquem fortemente na necessidade de covariáveis extras ou em transformações no modelo. Para se ter um gráfico deste tipo faz-se a regressão da variável adicional z sobre as colunas da matriz modelo local $\tilde{X}$ usando a matriz de pesos W, isto é, calcula-se desta regressão os resíduos parciais

$$u = (I - \hat{H})\hat{W}^{1/2}z, \tag{7.15}$$

onde H é a matriz "hat" que entra na expressão dos resíduos t_i's, e traça-se o gráfico dos resíduos de Pearson p_i versus u_i.

§7.8 Uma Análise de Dados Reais

Considera-se aqui os 24 dados analisados por Young e Bakir (1987) e apresentados na Tabela 7.3, onde y é o logaritmo do tempo de ruptura em horas para ensaios em peças de aço inoxidável com x representando o logaritmo do esforço a que a peça está submetida. Admite-se a distribuição log-gama para a resposta e uma componente sistemática do tipo

Tabela 7.3: *Tempos de ruptura* (y) *em peças de aço inoxidável e logaritmos dos esforços submetidos* (x)

y	7.144	7.401	7.413	7.444	7.487	7.799
x	28.84	28.84	28.84	28.84	28.84	28.84
y	5.136	5.549	5.580	6.346	6.387	6.658
x	31.63	31.63	31.63	31.63	31.63	31.63
y	4.331	4.466	4.564	4.745	4.804	4.883
x	34.68	34.68	34.68	34.68	34.68	34.68
y	3.091	3.611	3.664	3.714	3.738	3.761
x	38.02	38.02	38.02	38.02	38.02	38.02

$$\theta = \beta_0 + \beta_1(x - \overline{x}).$$

Verifica-se, facilmente, que H é a matriz identidade e $D_2(\theta) = E\{t''(y,\theta)\} = -1$, sendo as componentes do desvio iguais à $2\{\exp(y-\hat{\theta}) - (y-\hat{\theta}) - 1\}$, onde $\hat{\theta}$ é a estimativa de MV de θ. Cordeiro e Paula (1989b,c) apresentam o programa GLIM para o ajustamento deste modelo utilizando a diretiva OWN que define o modelo do usuário.

A partir das estimativas iniciais $\beta_0 = 4$ e $\beta_1 = 10$ chega-se à convergência após 5 iterações, sendo o critério de parada: a soma de quadrados de diferenças relativas entre estimativas sucessivas dos parâmetros lineares menor do que 10^{-8}. O desvio do modelo iguala 2.459 e as estimativas obtidas são (erros padrões entre parênteses): $\hat{\beta}_0 = 5.461(0.068)$ e

$\hat{\beta}_1 = -14.060(0.663)$. O parâmetro de dispersão ϕ foi estimado em 0.112. As Figuras 7.1 e 7.2 mostram os gráficos dos dados y versus valores ajustados $\hat{\theta}$ e dos resíduos $y-\hat{\theta}$ versus $\hat{\theta}$, respectivamente, que indicam a adequação do modelo. Ainda, a Figura 7.3 apresenta o gráfico dos resíduos ordenados versus os quantis da distribuição log-gama $\log\log\{24/(24.5-i)\}$, $i = 1, \ldots, 24$ que mostra evidência de que a log-gama é satisfatória para a variável y.

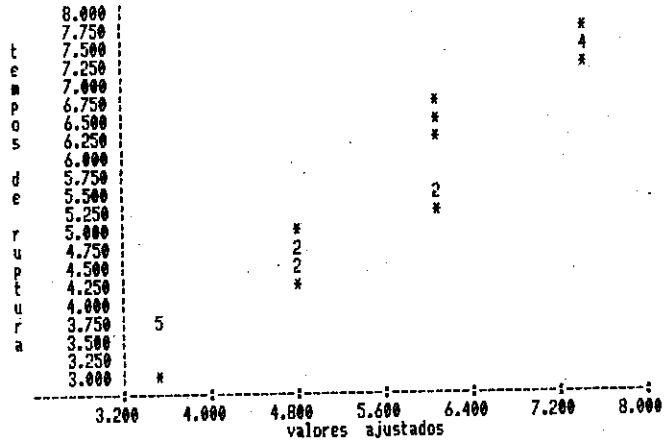

Figura 7.1: *Gráfico dos valores observados versus ajustados*

§7.9 Testes de Hipóteses e Regiões de Confiança

Considera-se um modelo definido por (7.1) e (7.2) que foi ajustado aos dados e cujo interesse é testar hipóteses e construir regiões de confiança para os parâmetros β's.

Na teoria de testes de hipóteses três estatísticas são, usualmente, consideradas: a razão de máxima verossimilhança $(-2\log\lambda)$, a estatística de

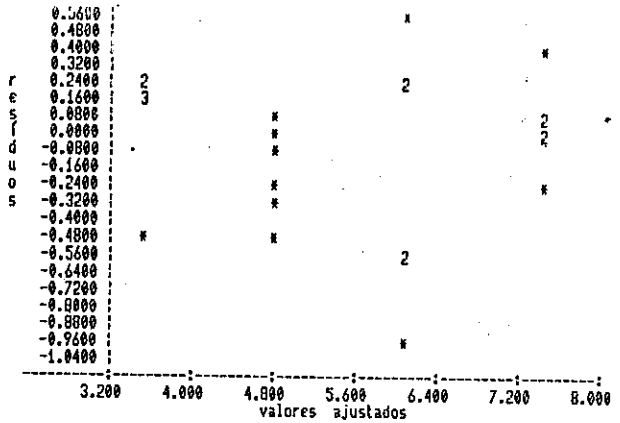

Figura 7.2: *Gráfico dos resíduos versus valores ajustados*

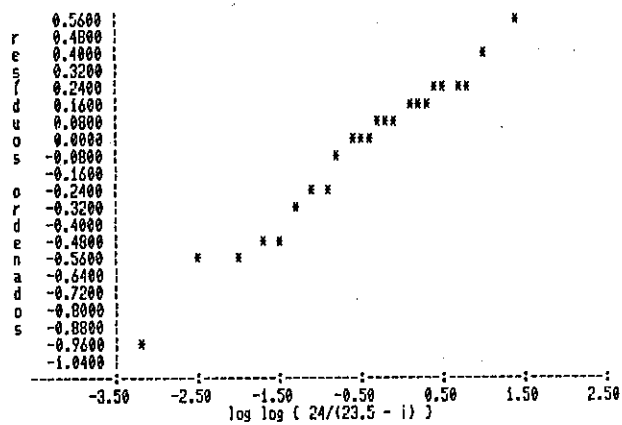

Figura 7.3: *Gráfico dos resíduos ordenados versus* $\log\log\{24/(24.5-i)\}$

Wald (W) e a estatística escore (E). O leitor poderá consultar o Capítulo 9 de Cox e Hinkley (1979) para ver as definições dessas estatísticas.

Inicialmente apresenta-se o teste da hipótese nula simples $H_1: \beta = \beta^{(0)}$, onde $\beta^{(0)}$ é um valor especificado para β, versus a hipótese alternativa $A_1: \beta \neq \beta^{(0)}$. Admitem-se, nesta seção, testes condicionais ao parâmetro

ϕ, supondo $a(\phi) = \phi$ constante para todos os dados. As três estatísticas citadas anteriormente reduzem-se, no teste de H_1 versus A_1, à

$$-2\log\lambda = 2\phi^{-1}\sum_{i=1}^{n}\{t(y_i,\hat{\theta}_i) - t(y_i,\theta_i^{(0)})\}, \tag{7.16}$$

$$W = \phi^{-1}(\hat{\beta}-\beta^{(0)})^T\hat{\tilde{X}}^T\hat{W}\hat{\tilde{X}}(\hat{\beta}-\beta^{(0)}), \tag{7.17}$$

$$E = \phi^{-1}\dot{t}^{(0)T}H^{(0)}W^{(0)}C_{\hat{\eta}}^{(0)}W^{(0)}H^{(0)}\dot{t}^{(0)}, \tag{7.18}$$

onde $C_{\hat{\eta}} = \tilde{X}(\tilde{X}^TW\tilde{X})^{-1}\tilde{X}^T$ é, com a exceção do multiplicador ϕ, a estrutura de covariância assintótica de $\hat{\eta}$, com os sobrescritos $\hat{}$ e (0) indicando estimativas nos pontos $\hat{\beta}$ e $\beta^{(0)}$, respectivamente.

As estatísticas (7.16), (7.17) e (7.18) são assintoticamente equivalentes e, segundo H_1, têm aproximadamente a distribuição χ^2_p.

Admite-se agora a partição $(\beta_q^T\beta_{p-q}^T)^T$ para β e a hipótese nula composta $H_2{:}\,\beta_q = \beta_q^{(0)}$ a ser testada contra $A_2{:}\,\beta_q \neq \beta_q^{(0)}$. Sejam $\overline{\hat{\beta}} = (\beta_q^{(0)T}\overline{\hat{\beta}}_{p-q}^T)^T$ e $\hat{\beta} = (\hat{\beta}_q^T\hat{\beta}_{p-q}^T)^T$ as estimativas de máxima verossimilhança de β segundo H_2 e A_2, respectivamente. Particionando-se $\tilde{X} = (\tilde{X}_q\tilde{X}_{p-q})$ da mesma maneira que β, obtém-se as matrizes de covariância assintótica de $\hat{\beta}$ e $\overline{\hat{\beta}}$, respectivamente, como

$$Cov(\hat{\beta}) = \begin{bmatrix} K_{qq} & K_{q(p-q)} \\ K_{(p-q)q} & K_{(p-q)(p-q)} \end{bmatrix}^{-1}$$

e

$$\mathrm{Cov}(\overline{\hat{\beta}}) = \begin{bmatrix} 0 & 0 \\ 0 & K^{-1}_{(p-q)(p-q)} \end{bmatrix},$$

onde $K_{qq} = \phi^{-1}\tilde{X}_q^T W \tilde{X}_q$, $K_{(p-q)q} = K_{q(p-q)}^T = \phi^{-1}\tilde{X}_{p-q}^T W \tilde{X}_q$ e $K_{(p-q)(p-q)} = \phi^{-1}\tilde{X}_{(p-q)(p-q)}^T W \tilde{X}_{(p-q)(p-q)}$.

Nota-se ainda que

$$\text{Cov}(\hat{\beta}_q) = \{K_{qq} - K_{q(p-q)}K_{(p-q)(p-q)}^{-1}K_{q(p-q)}^T\}^{-1}, \tag{7.19}$$

sendo o erro em (7.19) de ordem n^{-2}. Aqui $-K_{(p-q)(p-q)}^{-1}K_{(q-p)q}$ representa a matriz dos coeficientes da regressão de $\hat{\beta}_{(p-q)}$ sobre β_q. Seja $C(\beta_q, \beta_{p-q})$ a expressão do lado direito de (7.19) que iguala à estrutura de covariância assintótica da distribuição marginal de $\hat{\beta}_q$. Nesta expressão ϕ aparece como um multiplicador. Demonstra-se que

$$C(\beta_q, \beta_{p-q}) = \phi[\tilde{X}_q^T W^{1/2}\{I_n - H_{p-q}\}W^{1/2}\tilde{X}_q]^{-1}, \tag{7.20}$$

onde

$$H_{p-q} = W^{1/2}\tilde{X}_{p-q}(\tilde{X}_{p-q}^T W \tilde{X}_{p-q})^{-1}\tilde{X}_{p-q}^T W^{1/2}. \tag{7.21}$$

A matriz H_{p-q} de ordem n, simétrica e idempotente, representa uma matriz de projeção local segundo o modelo com componente sistemática envolvendo somente os parâmetros em β_{p-q}.

Para o teste de H_2 versus A_2 tem-se $-2\log\lambda = 2\phi^{-1}\sum_{i=1}^n\{t(y_i, \hat{\theta}_i) - t(y_i, \overline{\hat{\theta}_i})\}$, onde $\hat{\theta}_i$ e $\overline{\hat{\theta}_i}$ são os valores de θ nos pontos $\hat{\beta}$ e $\overline{\hat{\beta}}$, respectivamente. As estatísticas de Wald e escore reduzem-se à

$$W = (\hat{\beta}_q - \beta_q^{(0)})^T C(\hat{\beta}_q, \hat{\beta}_{p-q})^{-1}(\hat{\beta}_q - \beta_q^{(0)}) \tag{7.22}$$

e

$$E = \frac{1}{\phi^2}\overline{\hat{t}}\,\overline{\hat{H}}\,\overline{\hat{X}}_q C(\beta_q^{(0)}, \overline{\hat{\beta}}_{p-q})\overline{\hat{X}}_q^T\,\overline{W}\,\overline{\hat{H}}\,\overline{\hat{t}}, \tag{7.23}$$

as quantidades estimadas no ponto $\overline{\hat{\beta}}$ sendo indicadas com sobrescritos $\hat{}$. Estas três estatísticas têm, aproximadamente, distribuição χ^2_q. Observar que ϕ aparecerá como um divisor em todas as três estatísticas. As estatísticas W e E requerem apenas a estimação do modelo segundo A_2 e H_2, respectivamente, mas a estatística $-2\log\lambda$ exige o ajustamento do modelo segundo as duas hipóteses. Para os modelos exponenciais lineares todas as expressões deduzidas nesta seção são, claramente, válidas, embora não ocorram simplificações consideráveis.

Havendo interesse de determinar uma região de confiabilidade especificada para um conjunto de parâmetros $\beta_{i_1},\ldots,\beta_{i_q}$ pode-se usar a mesma metodologia descrita na Seção 6.6 para os modelos exponenciais lineares.

§7.10 Exercícios

1. Dar exemplos de modelos não-exponenciais, determinando as quantidades básicas para ajustamento aos dados, a expressão do desvio e as medidas de diagnóstico apresentadas na Seção 7.7.
2. Caracterizar as seguintes distribuições como modelos não- exponenciais calculando as fórmulas dos desvios respectivos:
   (i) distribuição de von Mises;
   (ii) distribuição de valor extremo;
   (iii) distribuição de Cauchy com parâmetro de escala conhecido;
   (iv) distribuição tipo III do sistema de Pearson com parâmetro de locação desconhecido e demais parâmetros conhecidos;
   (v) distribuição logística com locação desconhecida e variância conhecida;

(vi) distribuição obtida da composição de uma Poisson de parâmetro θ com um retângular em (a, b), a e b conhecidos;

(vii) distribuição deduzida da composição de uma Poisson com uma gama que resulta na binomial negativa expressa por $\binom{\alpha+y-1}{\alpha-1}\left(\frac{\theta}{\theta+1}\right)^y(1+\theta)^{-\alpha}$, supondo α conhecido. Quando esta distribuição produziria um modelo exponencial?

3. Estimar ϕ no modelo (7) da Seção 7.2.
4. Aplicar as equações (7.7) de ajustamento do modelo não-exponencial supondo uma componente sistemática linear equivalente:
   (a) à classificação de um único fator;
   (b) à classificação cruzada de dois fatores sem interação e uma única observação por cela.
5. Na Seção 7.7 foi definido $\rho_i = E(y_i)/y_i$. Calcular ρ_i para os modelos (1), (2), (3), (5), (6) e (8) discutidos na Seção 7.2 obtendo as expressões dos resíduos como componentes do desvio.
6. Demonstrar as expressões das estimativas segundo os modelos saturados dados na Tabela 7.2.
7. Ajustar um modelo de Weibull aos dados da Tabela 7.3.
8. Considere um modelo não-exponencial não-linear com um único parâmetro β na componente sistemática. Calcular de (7.16), (7.17) e (7.18) as expressões dessas estatísticas para testar $H: \beta = \beta^{(0)}$ versus $A: \beta \neq \beta^{(0)}$. Supondo que este modelo tenha um parâmetro de dispersão ϕ, obter as expressões das estatísticas de Wald, escore e da razão de MV para os seguintes testes:
   (a) $H_1: \phi = \phi^{(0)}$ versus $A_1: \phi \neq \phi^{(0)}$;
   (b) $H_2: \beta = \beta^{(0)}$ versus $A_2: \beta \neq \beta^{(0)}$.
9. Analisar os dados da Tabela 2.1 com um modelo log-normal com coe-

ficiente de variação constante c. Como estimaria c?

10. Aplicar as expressões (7.22) e (7.23) para testar no modelo da Seção 7.8 as seguintes hipóteses: $H_1: \beta_0 = 4$ versus $A_1: \beta_0 \neq 4$ e $H_2: \beta_1 = -12$ versus $A_2: \beta_1 \neq -12$.

CAPÍTULO 8

MODELOS DE REGRESSÃO MAIS COMPLEXOS

§8.1 Modelo Linear Generalizado com um Parâmetro Não-Linear Extra

Um parâmetro não-linear extra α aparece nos modelos lineares generalizados, mais frequentemente, nas seguintes situações:

(1) na função de ligação visando a definir uma família paramétrica de ligações (Seções 6.8 e 8.6);

(2) como parâmetro de transformação da variável resposta ou de variáveis explicativas (Seções 2.1, 6.8 e 8.6);

(3) na função de variância dos modelos de quase-verossimilhança (Seção 8.7.1) 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;

(4) no modelo logístico com probabilidade de sucesso da forma (Cox, 1984);

$$\mu = \alpha + |1 - \alpha| \exp(\eta)/[1 + \exp(\eta)];$$

(5) em distribuições especiais como o parâmetro de forma da Weibull;

(6) no modelo logístico generalizado (Seção 4.6.3), onde α_1 e α_2 são parâmetros não-lineares extras.

A estimação conjunta de α e dos β'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 β'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 $\hat{\alpha}$.

Neste capítulo são abordados alguns modelos de regressão mais complexos que aqueles estudados nos capítulos anteriores. Os modelos lineares generalizados com ligação composta são apresentados na Seção 8.2. Os modelos aditivos generalizados introduzidos por Hastie e Tibshirani (1986, 1987) e os modelos semi-paramétriscos de Green e Yandell (1985) são vistos nas Seções 8.3 e 8.4, respectivamente. Na Seção 8.5 são detalhados os modelos para análise de dados de sobrevivência. Uma classe de modelos definida por duas transformações é introduzida na Seção 8.6, constituindo uma área de pesquisa a ser mais explorada. Os modelos de quase- verossimilhança, objeto de várias pesquisas recentes, são discutidos na Seção 8.7, com a apresentação de duas análises de dados reais. A Seção 8.8 considera os modelos com estrutura de autocorrelação interna. Finalmente, na Seção 8.9, são citados outros modelos específicos.

§8.2 Modelos Lineares Generalizados com Ligação Composta

Considere um modelo com distribuição (6.1), mas com componente sistemática definida por

$$E(y) = \mu = C\gamma$$

$$f(\gamma) = \eta = X\beta \tag{8.1}$$

onde μ e y são vetores $n \times 1$, C e X são matrizes conhecidas $n \times m$ e $m \times p$, respectivamente, $\gamma = (\gamma_1, \cdots, \gamma_m)^T$, $\eta = (\eta_1, \cdots, \eta_m)^T$ e $\beta = (\beta_1, \cdots, \beta_p)^T$. 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 (8.1) considera uma estrutura não- linear $\mu_i = c_i(\gamma)$ entre μ e γ. O ajustamento do modelo $\mu_i = c_i(\gamma)$, $f(\gamma) = \eta \doteq X\beta$, pode ser feito via o algoritmo descrito em (6.4) com pequenas modificações. Sem perda de generalidade trabalha-se sem o escalar ϕ. Tem-se $\partial L(\beta)/\partial\beta = \tilde{X}^T V^{-1}(y - \mu)$, onde $V = \text{diag}\,\{V_1, \cdots, V_n\}$, $L = \{d\mu_i/d\eta_k\}$ 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

$$X^T L^{(m)T} V^{(m)^{-1}} L^{(m)} X\beta^{(m+1)} = X^T L^{(m)T} V^{(m)^{-1}} y^{*(m)}, \tag{8.2}$$

onde $y^* = L\eta + y - \mu$. A variável dependente y^*, a matriz modelo LX e os pesos V^{-1} se modificam no processo (8.2) . O GLIM não pode ser

usado diretamente e o usário deve trabalhar com um programa especial (Thompson e Baker, 1981).

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 $H = \text{diag}\{d\eta_1/d\gamma_1, \cdots, d\eta_m/d\gamma_m\}$ e, então, $\tilde{X} = CH^{-1}X$ e $y^* = CH^{-1}\eta + y - \mu$.

§8.3 Modelos Aditivos Generalizados

Os modelos aditivos generalizados são definidos pela componente aleatória dos MLGs e uma componente sistemática da forma

$$g(\mu) = \eta = \beta + \sum_{j=1}^{p} f_j(x_j), \tag{8.3}$$

com as restrições $E\{f_j(x_j)\} = 0$ para $j = 1, \cdots, p$, onde os $f_j(x_j)$ são funções não-paramétricas a serem estimadas.

Assim, a estrutra 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 $S(y \mid x_{ij})$, $j = 1, \cdots, p$, $i = 1, \cdots, n$.

O suavizador mais usado tem a forma linear $S(y \mid x_{ij}) = \hat{a}_{ij} + \hat{b}_{ij}x_{ij}$, 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 $N_{ij} = \{x_{(i-r)j}, \cdots, x_{ij}, \cdots, x_{(i+r)j}\}$, 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 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 $\hat{f}_j(x_{ij}) = 0 \ \forall \ i, j$ e $\hat{\beta} = \overline{y}$;
2. Fazer $j = 1, \cdots, p$ e $i = 1, \cdots, n$ 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 $\hat{f}_j(x_{ij}) = S(r_j \mid x_{ij})$ 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 para-se; caso contrário, ir para 2.

Observar 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 6.3). O algoritmo fica sendo:

1. Inicializar $\hat{f}_j(x_{ij}) = 0$, $j = 1, \cdots, p$, $\hat{\beta} = g(\overline{y})$, $\hat{\eta} = \hat{\beta}1$, $\hat{W} = W(\overline{y})$ 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 $j = 1, \cdots, p$;
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}, $i = 1, \cdots, n$ e $j = 1, \cdots, p$;
4. Atualizar $\hat{\beta} = g\left(\frac{1^T \hat{W} \hat{y}^* 1}{n}\right)$, $\hat{\eta} = \hat{\beta} + \sum_{j=1}^{p} \hat{f}_j(x_j)$, $\hat{\mu} = g^{-1}(\hat{\eta})$, $\hat{H} = H(\hat{\mu})$, $\hat{W} = W(\hat{\mu})$ e $\hat{y}^* = \hat{\eta} + \hat{H}(y - \hat{\mu})$;
5. Calcular o desvio $D(y; \hat{\mu})$ do modelo da Expressão (6.6) como função de y e $\hat{\mu}$. Quando $D(y; \hat{\mu})$ convergir para-se; caso contrário, ir para 2.

§8.4 Modelos Semi-Paramétricos

Os modelos semi-paramétriscos 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 $s(\cdot)$ é especificada por uma soma $s(t) = \sum_{i=1}^{q} \gamma_i\, g_i(t)$ de q funções básicas $g_1, \cdots, g_q$ sendo os γ's parâmetros desconhecidos. O problema de maximização consiste em definir uma log-verossimilhança penalizada como

função dos parâmetros β e γ e maximizá-la

$$\max_{\beta,\gamma}[L\{\eta(\beta,\gamma)\} - \lambda J\{s(\gamma)\}/2] \tag{8.4}$$

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 $L\{\eta(\beta,\gamma)\}$. 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 γ'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 $(1/n, 2)$.

§8.5 Modelos para Análise de Dados de Sobrevivência

Nesta seção serão apresentados alguns modelos usuais para análise de dados em que a variável de 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.

§8.5.1 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 frequente 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

$$h(t;x) = \lambda(t)\exp(x^T\beta), \tag{8.5}$$

onde $\beta = (\beta_1, \cdots, \beta_p)^T$ é um vetor de parâmetros desconhecidos associados às covariáveis de $x = (x_1, \cdots, x_p)^T$, $\lambda(t)$ é uma função não-negativa do tempo e $\eta = x^T\beta$ é o preditor linear.

O modelo (8.5) 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

$$S(t;x) = \exp\{-\Lambda(t)\exp(x^T\beta)\}, \tag{8.6}$$

onde $\Lambda(t) = \int_{-\infty}^{t} \lambda(u)du$. Similarmente a densidade de probabilidade de t fica expressa na forma

$$f(t;x) = \Lambda'(t)\exp\{\eta - \Lambda(t)\exp(\eta)\}. \tag{8.7}$$

A distribuição do tempo de sobrevivência t do modelo (8.5), pertence à família exponencial não-linear, mas não à (6.1). Em particular, $E\{\Lambda(t)\} = \exp(-\eta)$ e Var $\{\Lambda(t)\} = \exp(-2\eta)$.

A estimação dos β'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 $L(\beta)$, enquanto um indivíduo censurado em t_i contribui com $\log S(t_i;x_i)$. A função $L(\beta)$ reduz-se à

$$L(\beta) = \sum_{j=1}^{n}\{y_i \log f(t_i;x_i) + (1-y_i)\log S(t_i;x_i)\},$$

que pode ser expressa numa forma mais conveniente usando (8.6) e (8.7),

$$L(\beta) = \sum_{j=1}^{n}(y_i \log \mu_i - \mu_i) + \sum_{j=1}^{n} \log\{\lambda(t_i)/\Lambda(t_i)\}, \tag{8.8}$$

onde $\mu_i = \Lambda(t_i)\exp(\eta_i)$. A segunda soma de (8.8) não depende dos β's e, portanto, (8.8) tem a mesma forma da log-verossimilhança de um modelo de Poisson com n observações independentes $y_1, \cdots, y_n$, médias $\mu_1, \cdots, \mu_n$, e preditores lineares que são dados por $\eta_i = \log \mu_i - \log \Lambda(t_i)$, $i = 1, \cdots, n$.

As estimativas de máxima verossimilhança para os β's podem ser obtidas pelo sistema GLIM, 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 β's é feita da maneira usual.

A Tabela 8.1 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 β'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

$$\hat{\alpha} = (n-m)/\sum_{i=1}^{n}(\hat{\mu}_i - y_i)\log t_i \tag{8.9}$$

Admite-se uma estimativa inicial para α e ajusta-se a y, um modelo log-linear com "offset" $\alpha \log t$. De (8.9) reestima-se α, continuando o processo até a convergência.

Tabela 8.1 - *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	$\alpha \exp(\alpha t)$	$\alpha \exp\{\alpha t + \eta - \exp(\alpha t + \eta)\}$	αt

O modelo de valor extremo pode ser transformado no de Weibull com a transformação $\exp(t)$, no lugar de t.

§8.5.2 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 (8.5). 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 (8.8). 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 $t_1, t_2, \cdots, t_k$ 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 (8.5), 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 (6.1), 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

$$\log P_j = y_i^T\beta - \log\{\sum_{s\in R(t_j)} \exp(x_s^T\beta)\},$$

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 $b''(\beta)$. Contudo, 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 6.3 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 (6.4) e (6.5).

Em McCullagh e Nelder (1983, pg. 191) há uma discussão sobre métodos iterativos para a estimação 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. Notar que se $R(t_j)$ tem $M+1$ elementos, para todo j, então $L(\beta)$ coincide com a log- verossimilhança definida em (4.35), 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 $L(\beta)$ 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_1 ter falhaddo antes do indivíduo x_2, mais essa mesma probabilidade no sentido inverso, isto é,

$$P_{j\ (\text{ Real })} = \frac{\exp(x_1^T\beta)}{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta)} \cdot \frac{\exp(x_2^T\beta)}{\left\{ \sum\limits_{s\in R(t_j)} \exp(x_s^T\beta) - \exp(x_1^T\beta) \right\}}$$

$$+ \frac{\exp(x_2^T\beta)}{\sum\limits_{s\in R(t_j)} \exp(x_s^T\beta)} \cdot \frac{\exp(x_1^T\beta)}{\left\{ \sum\limits_{s\in R(t_j)} \exp(x_s^T\beta) - \exp(x_2^T\beta) \right\}}. \tag{8.10}$$

Denotando por m_j o número de empates no tempo t_j, o número de termos em $P_{j(\text{ Real })}$ é $m_j!$, o que poderá aumentar a complexidade de $L(\beta)$ se m_j for grande. Algumas expressões aproximadas para $P_{j(\text{ Real })}$ foram propostas. Em particular, Cox (1972) propõe a seguinte aproximação para (8.10):

$$P_{j\,(\text{ Cox })} = \frac{\exp(x_1^T\beta)\exp(x_2^T\beta)}{\sum\limits_{(s<k)\in R(t_j)} \exp(x_s^T\beta)\exp(x_k^T\beta)};$$

entretanto, uma aproximação mais usual foi sugerida por Peto (1972), e é dada por

$$P_{j\,(\text{ Peto })} = 2\frac{\exp(x_1^T\beta)}{N\left[\left\{\sum\limits_{s\in R(t_j)}\exp(x_s^T\beta)\right\}/N\right]} \times \frac{\exp(x_2^T\beta)}{(N-1)\left[\left\{\sum\limits_{s\in R(t_j)}\exp(x_s^T\beta)\right\}/N\right]}, \tag{8.11}$$

onde N é o número de indivíduos no conjunto de risco $R(t_j)$. Logo, supondo que há r_j indivíduos em $R(t_j)$ e que ocorreram m_j empates, a aproximação de Peto generaliza para

$$P_{j\,(\text{ Peto })} = \prod_{i=1}^{m_j} \frac{\exp(x_i^T\beta)}{\binom{r_j}{m_j}\left\{\sum\limits_{s\in R(t_j)}\exp(x_s^T\beta)\right\}^{m_j}}. \tag{8.12}$$

Essa aproximação tem sido sugerida (McCullagh e Nelder, 1983, pg. 190) quando há poucos empates. Também em Whitehead (1980) há uma discussão sobre a estimação dos parâmetros β's do modelo de Cox, quando há empates e a aproximação (8.12) é utilizada.

Cox (1975) mostra que toda a teoria usual para a estatísitica da razão de máxima verossimilhança continua valendo para os modelos de riscos proporcionais.

§8.5.3 Riscos Não-Proporcionais

Em algumas situações práticas a suposição de riscos proporcionais pode não ser verificada ao longo de todo o tempo. Suponha, como exemplo, que dois tratamentos (A e B) são aplicados separadamente em indivíduos portadores de uma doença grave, com o intuito de prolongar a vida dos mesmos. Pode ocorrer de o tratamento A ser superior ao tratamento B a curto prazo, entretanto, a partir de um certo tempo o tratamento B pode tornar-se mais eficaz.

Nesses casos, o modelo proposto em (8.5) não é o mais adequado, devendo ser modificado para uma forma mais geral, em que o tempo é também tratado como uma covariável. Essa forma é dada por

$$h(t;x) = \lambda(t)\exp\{g(t;x)\},$$

onde $g(t;x)$ depende agora do tempo t, das covariáveis de x e dos parâmetros desconhecidos $\beta_1, \cdots, \beta_p$.

As Figuras 8.1 e 8.2 ilustram este tipo de problema. Na primeira figura tem-se o gráfico da função de risco relativo $\exp\{g(t;x)\}$ contra o tempo t. Notar que os riscos entre os dois tratamentos não são proporcionais ao longo do tempo, com o tratamento A sendo sempre superior ao tratamento B.

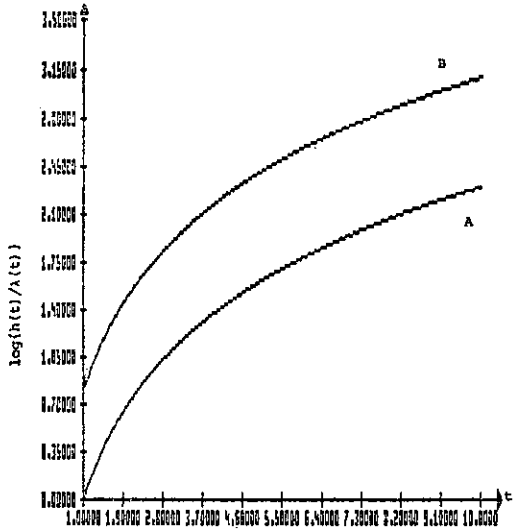

Figura 8.1 - *Exemplo de um modelo com riscos não-proporcionais e crescentes ao longo do tempo.*

Na Figura 8.2, pelo gráfico de $\exp\{g(t;x)\}$ contra o tempo, nota-se que o tratamento A é inicialmente mais eficiente, entretanto a partir de um certo tempo o tratamento B torna-se superior.

Do ponto de vista de estimação dos parâmetros β's, muda muito pouco em relação ao que foi apresentado na seção anterior. A principal mudança consiste na inclusão de termos adicionais envolvendo t na componente sistemática $\exp(x^T\beta)$ do modelo de Cox.

Uma mudança importante ocorre na estimação da função $\lambda(t)$, que é usualmente assumida constante para cada intervalo $(t_{j-1}, t_j), j = 1, \cdots, k$. Conhecendo a estimativa dessa função tem-se diretamente a estimativa da função de sobrevivência $S(t)$, dada por

$$\hat{S}(t) = \exp\left[-\int_0^t \exp\{\hat{g}(u;x)\}\hat{\lambda}(u)du\right].$$

que simplifica-se bastante quando $\hat{g}(t;x) = x^T\hat{\beta}$.

Para os modelos de riscos proporcionais a estimativa de $\lambda(t)$ no intervalo (t_{j-1}, t_j) é expressa na forma

$$\hat{\lambda}(t) = m_j / \left\{ (t_j - t_{j-1}) \sum_{s \in R(t_j)} \exp(x_s^T \hat{\beta}) \right\}.$$

Quando os riscos são não-proporcionais, $\hat{\lambda}(t)$ é obtida de uma expressão um pouco mais complexa, que envolve a solução de uma integral,

$$\hat{\lambda}(t) = m_j / \left\{ \int_{t_{j-1}}^{t_j} \sum_{s \in R(t_j)} \exp\{g(t;u)\} du \right\}.$$

Em Carter et al. (1983) são apresentados diversos exemplos de dados de sobrevivência com riscos não-proporcionais.

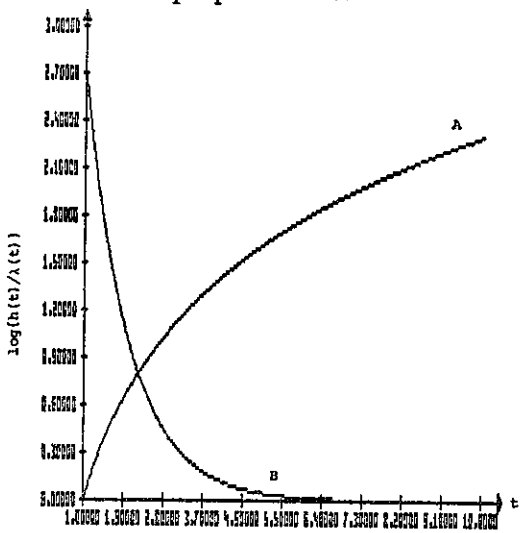

Figura 8.2 - *Exemplo de um modelo com riscos não-proporcionais e com tendências opostas ao longo do tempo.*

§8.6 Uma Classe de Modelos Definida por Duas Transformações

Define-se agora uma classe de modelos indexada por duas transformações do tipo Box-Cox:

$$(8.13) \qquad y^* = (y-1)^{\lambda_1}/\lambda_1 \sim FE(\theta,\phi),\ (\mu^{\lambda_2}-1)/\lambda_2 = \eta$$

onde $E(y^*) = \mu$ and $\eta = X\beta$. Assim, os dados brutos $y_1, \cdots, y_n$ são transformados por um parâmetro λ_1 produzindo dados modificados $y_1^*, \cdots, y_n^*$ que devem seguir alguma distribuição na família exponencial (6.1), com as médias dos y_i^*'s sendo ainda transformadas por um parâmetro λ_2 de tal maneira a produzir o preditor linear.

Esta classe de modelos engloba como casos especiais vários modelos importantes. Os próprios MLGs são definidos por $\lambda_1 = 1$ e os modelos de Box e Cox por $\lambda_2 = 1$ com $F(\theta,\phi)$ sendo a distribuição normal. Quando $\lambda_1 = 0$ e $\lambda_2 = -1$ admite-se erros da família exponencial na escala logarítmica e linearidade na escala inversa. A grande desvantagem prática dos modelos de Box e Cox em relação a esta classe é exigir um único λ produzindo dois efeitos: normalidade do erro e linearidade dos efeitos sistemáticos.

Faz-se agora uma análise de 25 dados de mobilidade social (Bishop, Fienberg e Holland, 1975), apresentados na Tabela 8.2, onde y_{ij} representa a frequência observada de famílias tendo o pai a profissão i e o filho a profissão j, que ilustra as potencialidades da classe de modelos aqui proposta.

A estrutura linear do modelo é definida por $\eta_{ij} = \beta +$ pai $(i) +$ filho (j), $i,j = 1, \cdots, 5$, onde pai (i) e filho (j) são os efeitos das profissões i e j

do pai e filho, respectivamente, e β é uma média geral. Algum tipo de restrição sobre os efeitos das profissões i e j do pai e do filho é necessária para que a matriz modelo tenha inversa. Apesar da restrição $\sum_{i=1}^{5}$ pai $(i) =$ $\sum_{j=1}^{5}$ filho $(j) = 0$ ser a mais usual, outras equivalentes podem ser propostas como pai $(1) =$ filho $(1) = 0$, adotada pelo sistema GLIM. Ao estimativas dos η_{ij}'s independem do tipo de vínculo dos efeitos desses parâmetros.

Tabela 8.2 - Mobilidade Social na Grã-Bretanha

Frequências de profissões pai/filho com os 5 níveis seguintes: 1-executivo (mais alto), 2-subordinado ao executivo, 3-administrativo, 4-profissional habilitado e 5-sem habilitação (mais baixo)

		Filho				
		1	2	3	4	5
	1	50	45	8	18	8
	2	28	174	84	154	55
Pai	3	11	78	110	223	96
	4	14	150	185	714	447
	5	3	42	72	320	411

Um modelo log-linear de independência de linhas e colunas seria fortemente rejeitado, isto é, y_{ij} tendo distribuição $P(\mu_{ij})$ com $\log \mu_{ij} = \eta_{ij}$. Propõe-se então um modelo normal definido em (8.13) com a escolha de λ_1 e λ_2 visando a obter a maior log- verossimilhança maximizada $\hat{L}(\lambda_1, \lambda_2)$ em relação aos parâmetros lineares supondo os parâmetros de transformação fixos. A Tabela 8.2 apresenta valores para esses parâmetros de transformação com as correspondentes log-verossimilhanças maximais obtidas de expressão equivalente a (2.3), onde a soma dos quadrados dos resíduos

corresponde ao desvio do modelo. Desta tabela se conclui que o modelo com maior log- verossimilhança maximizada corresponde à transformação raiz quadrada ($\lambda_1 = 0.5$) dos dados para produzir normalidade com ligação inversa ($\lambda_2 = -1$). As estimativas na estrutura linear (os erros padrões entre parantêses), considerando uma parametrização para os parâmetros lineares com pai (1) = filho (1) = 0, são: β = 0.29 (0.10), pai (2) = -0.09 (0.06), pai (3) = -0.10 (0.06), pai (4) = -0.12 (0.06), pai (5) = -0.11 (0.06), filho (2) = -0.12 (0.08), filho (3) = -0.12 (0.08), filho (4) = -0.14 (0.08) e filho (5) = -0.14 (0.08).

Observar que quando λ_1 é negativo a log-verossimilhança maximizada praticamente independe de λ_2. A adequação global do modelo ($\lambda_1 = 0.5$, $\lambda_2 = -1$) é evidenciada nos gráficos das Figuras 8.3 (y^* versus $\hat{\mu}$) e 8.4 (y versus $\hat{E}(y)$), apesar dos pontos correspondentes às observações (pai,filho) iguais à (1,1), (2,2), (2,5), (5,2) e (5,5) não estarem bem ajustados. Aqui $E(y)$ é deduzido de expansão em série de Taylor dada por

$$E(y) = (\lambda_1\mu + 1)^{1/\lambda_1} + \frac{1}{2}\sigma^2(1-\lambda_1)(\lambda_1\mu+1)^{(1-2\lambda_1)/\lambda_1}. \tag{8.14}$$

Torna-se importante frisar que um modelo melhor poderia ser encontrado através de um estudo detalhado dos resíduos, implicando em examinar distribuições alternativas dentro de $F(\theta, \phi)$. As médias ajustadas aos dados originais obtidas de (8.14) permitirão estimar as probabilidades de transição do tipo P { filho (j) | pai (i)} e consequentemente conhecer a mobilidade social em questão.

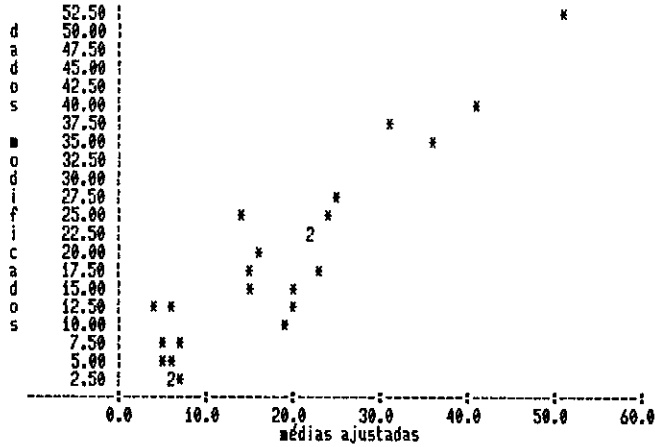

Figura 8.3 - *Gráfico dos dados modificados versus valores ajustados.*

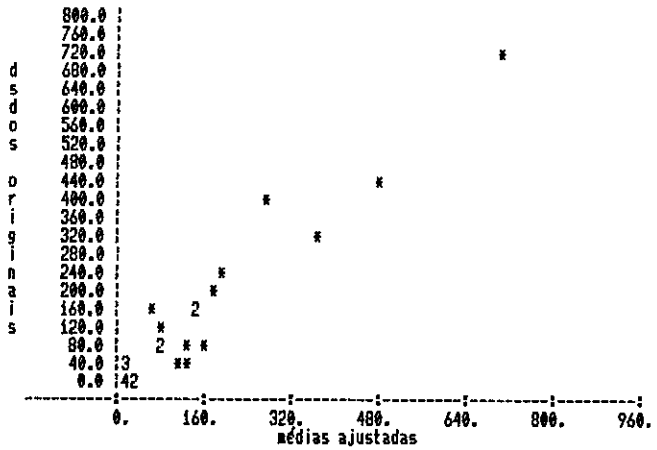

Figura 8.4 - *Gráfico dos dados originais versus médias ajustadas.*

Tabela 8.3 - *Verossimilhanças Maximizadas para Vários Modelos Normais Definidos em (8.13).*

$\lambda_1 =$	-1.5	-1.0	-0.5	0	0.5	1.0	1.5
-1.0	-164.8	-136.4	-112.8	-96.11	-89.85	-95.13	-135.8
-0.5	-164.7	-136.4	-112.9	-96.89	-90.97	-95.95	-109.8
$\lambda_2 =$ 0	-164.6	-136.3	-113.0	-97.91	-93.05	-98.01	-111.7
0.5	-164.6	-136.2	-113.3	-99.24	-93.48	-101.6	-117.5
1.0	-164.5	-136.1	-113.4	-100.9	-102.6	-116.8	-135.8

§8.7 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:

$$E(y_i) = \mu_i(\beta), \quad \text{Var}\,(y_i) = \phi V_i(\mu_i). \tag{8.15}$$

Aqui os μ_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, possivelmente desconhecido, podendo ainda ser uma função de regressores adicionais. Usualmente $\mu_i(\beta)$ equivale à componente sistemática do MLG.

Define-se a (log)-*quase-verossimilhança* para uma única observação apenas com a suposição de existência de sua média e de sua variância, por

$$Q = Q(y;\mu) = \frac{1}{\phi}\int (y-\mu)V(\mu)^{-1}d\mu. \tag{8.16}$$

Para $V(\mu) = k$, μ, μ^2, $\mu(1-\mu)$, $\mu + \mu^2/k$ e μ^3, com k uma constante, e integrando (8.16), 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. Observar 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 quase-verossimilhança tem propriedades semelhantes à verossimilhança

$$(8.17) \quad E\{\partial Q/\partial\mu\} = 0, \; E[\{\partial Q/\partial\mu\}^2] = -E\{\partial^2 Q/\partial\mu^2\} = 1/[\phi\, V(\mu)].$$

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

$$(8.18) \qquad -E\{\partial^2 Q/\partial\mu^2\} \leq -E\{\partial^2 L/\partial\mu^2\}.$$

Se y seguir a família exponencial de distribuições (Expressão (6.1)) 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 (8.18) somente ocorre no caso de L ser a log-verossimilhança da família exponencial. O lado esquerdo de (8.8) 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 $E\{\partial^2(Q-L)/\partial\mu^2\}$ 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 quase-verossimilhança para n observações é igual à soma de n contribuições definidas por (8.16). As estimativas de máxima quase- verossimilhança $\tilde{\beta}_1, \cdots, \tilde{\beta}_p$ são obtidas maximizando esta soma. Supondo que ϕ seja constante para os n dados $y_1, \cdots, y_n$ obtém-se o sistema de equações para os $\tilde{\beta}$'s, que não dependem de ϕ,

$$\sum_{i=1}^{n}(y_i - \mu_i) \cdot (\partial\mu_i/\partial\beta_i)/V_i(\mu_i) = 0. \tag{8.19}$$

A maximização da quase-verossimilhança generaliza o método de mínimos quadrados, o qual 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 GENSTAT, GLIM, BMDP ou SAS, na pior das hipóteses utilizando subprogramas especiais.

Na análise de dados na forma de contagens trabalha-se com o erro de Poisson supondo que Var $(y_i) = \phi\mu_i$. O parâmetro ϕ é estimado igualando a razão de quase-verossimilhança $2\{Q(y;y) - Q(y;\tilde{\mu})\}$ aos graus de liberdade $(n-p)$ da χ^2 de referência ou então usando a expressão mais simples

$$\tilde{\phi} = (n-p)^{-1}\sum_{i=1}^{n}(y_i - \tilde{\mu}_i)^2/\tilde{\mu}_i. \tag{8.20}$$

Os dados apresentarão sub-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 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) (vide Seção 8.7.1) 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 quase-verossimilhança (8.16) 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. Recentemente, 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, com expressão

$$Q^+ = -1/2 \sum_i \log\{2\pi\phi_i V(y_i)\} - 1/2 \sum_i D(y_i;\mu_i)/\phi_i, \tag{8.21}$$

sendo o somatório sobre todas as observações e a função $D(y;\mu)$, denominada de (*quase-desvio*), sendo uma simples extensão do desvio do MLG, definida para uma observação por

$$D(y;\mu) = -2\int_y^\mu (y-x)V(x)^{-1}dx, \tag{8.22}$$

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 8.4 apresenta quase-verossimilhanças para algumas funções de variância, com a exceção do escalar ϕ, deduzidas de (8.16). Desta tabela os desvios são facilmente obtidos.

Tabela 8.4 - *Quase-verossimilhanças associadas à quatro funções de variância.*

Função de Variância $V(\mu)$	Quase-Verossimilhança $Q(y;\mu)$
$\mu^\lambda (\lambda \neq 0, 1, 2)$	$\mu^{-\lambda}\left(\frac{y\mu}{1-\lambda} - \frac{\mu^2}{2-\lambda}\right)$
$\mu(1-\mu)$	$y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)$
$\mu^2(1-\mu)^2$	$(2y-1)\log\left(\frac{\mu}{1-\mu}\right) - \frac{y}{\mu} - \frac{1-y}{1-\mu}$
$\mu + \mu^2/\alpha$	$y\log\left(\frac{\mu}{\alpha+\mu}\right) + \alpha\log\left(\frac{\alpha}{\alpha+\mu}\right)$

§8.7.1 Modelo de Quase-Verossimilhança com Função de Variância Paramétrica

Agora admite-se o seguinte modelo de quase-verossimilhança com função de variância paramétrica:

$$E(y_i) = \mu_i(\beta), \quad \text{Var}\,(y_i) = \phi V_\lambda(\mu_i), \tag{8.23}$$

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 1ª ordem obtém-se a média e a variância modificadas $E(y)^* = \mu + \varepsilon\, d\mu/d\eta$ e $\text{Var}\,(y)^* = \phi V(\mu) + \lambda(d\mu/d\eta)^2$ e, portanto, a função de variância torna-se parametrizada por 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 (8.19) para obter as estimativas de máxima quase-verossimilhança dos β's. A estimativa de λ corresponderá ao maior valor da quase- verossimilhança estendida maximizada tratada como função de λ, obtida da expressão (8.21), ou ainda ao menor valor do *desvio estendido* -2 $Q^+(\lambda)$ 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 como exemplo uma função de variância contendo um parâmetro desconhecido que está multiplicando o quadrado da média na análise dos números de falhas da Tabela 8.5 em 32 peças de comprimentos variáveis de uma fábrica (Hinde, 1982). O ajustamento do modelo de regressão de Poisson para o número de falhas, com covariável igual ao logaritmo do tamanho da peça (log(com)), conduz a uma estimativa $\tilde{\phi}$, obtida de (8.20), igual a 2.267, o que indica que os dados podem estar superdispersos. Neste ajustamento o preditor linear estimado (erros padrões entre parentêses) corresponde a $\tilde{\eta} = \log \tilde{\mu} = -4.17\ (1.14) + 1.00\ (0.18) \log(\text{com})$.

Sendo assim, considera-se o ajustamento do modelo com função de variância $V(\mu_i) = \mu_i + \lambda\mu_i^2$, escolhendo para λ o valor que minimiza o desvio estendido. A função desvio estendida para este modelo vem de (8.21)

$$-2Q^+(\lambda) = \sum \{\log[2\pi(y_i + \lambda y_i^2)] + D(y_i; \tilde{\mu}_i)\} \tag{8.24}$$

sendo $D(y_i;\mu_i)$ obtido de (8.22) (vide também Tabela 8.4) e $\alpha = 1/\lambda$

$$D(y;\mu) = 2y\log\left\{\frac{(\mu+\alpha)y}{(y+\alpha)\mu}\right\} + 2\alpha\log\left\{\frac{\mu+\alpha}{y+\alpha}\right\}. \tag{8.25}$$

Tabela 8.5 - *Números de falhas em 32 peças de tecidos de comprimentos variáveis.*

Comprimento da peça em metros / número de falhas

551/6	651/4	832/17	375/9	715/14	868/8	271/5	630/7	491/7
372/7	645/6	441/8	895/28	458/4	642/10	492/4	543/8	842/9
905/23	542/9	522/6	122/1	657/9	170/4	738/9	371/14	735/17
749/10	495/7	716/3	952/9	417/2				

A Figura 8.5 apresenta o gráfico de $-2Q^{+}(\lambda)$ computado de (8.21) versus λ, implicando na estimativa $\tilde{\lambda}$ de λ igual à, aproximadamente, 0,13. Para este valor de λ (melhor modelo em $V_\lambda(\mu)$) o preditor linear estimado corresponde a $\tilde{\eta}$ =-3.78 (3.59) + 0.94 (0.56) log(com). A Tabela 8.6 apresenta os valores ajustados e resíduos segundo este modelo mostrando uma concordância razoável entre y e $\tilde{\mu}$. As Figuras 8.6 e 8.7 fornecem os gráficos de y versus $\tilde{\mu}$ e dos resíduos padronizados iguais a $(y-\tilde{\mu})/V_{0.13}(\tilde{\mu})^{1/2}$ versus $\tilde{\mu}$, respectivamente. Esses gráficos suportam a adequação do modelo. Este exemplo mostra a grande utilidade de incorporar um parâmetro desconhecido na função de variância do modelo.

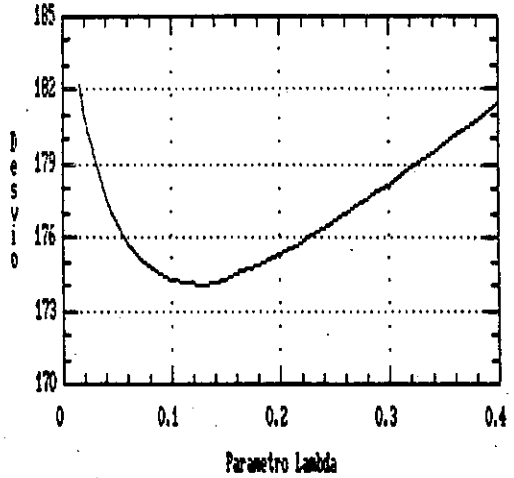

Figura 8.5 - *Gráfico de* $-2Q^{+}(\lambda)$ *(vertical) versus* λ *(horizontal).*

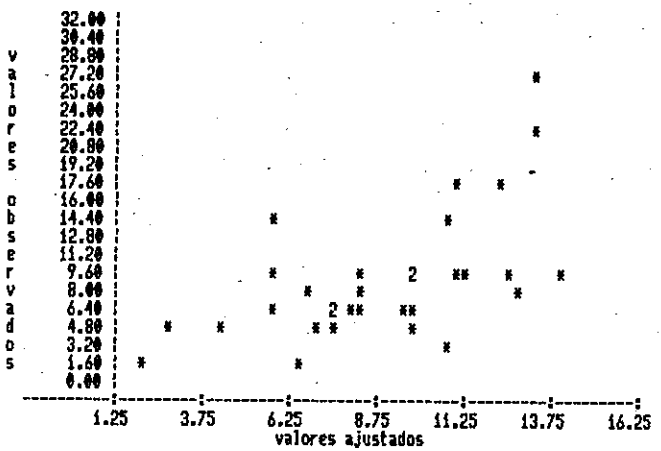

Figura 8.6 - *Gráfico de* y *versus* $\tilde{\mu}$ *no modelo de quase-verossimilhança correspondente à* $\lambda = 0.13$.

Tabela 8.6 - *Freqüências de falhas, médias ajustadas e resíduos segundo um modelo de quase-verossimilhança com função de variância paramétrica, correspondentes à estimativa* $\hat{\lambda}$*=0.13.*

dados	médias ajustadas	resíduos	dados	médias ajustadas	resíduos
6.000	8.364	-0.56581	8.000	8.251	-0.06059
4.000	9.776	-1.22588	9.000	12.435	-0.60222
17.000	12.297	0.83196	23.000	13.303	1.60921
9.000	5.836	0.98751	9.000	8.236	0.18491
14.000	10.672	0.65933	6.000	7.952	-0.48534
8.000	12.794	-0.82127	1.000	2.042	-0.64821
5.000	4.307	0.26721	9.000	9.860	-0.18135
7.000	9.481	-0.53922	4.000	2.785	0.62389
7.000	7.509	-0.13220	9.000	10.993	-0.38564
7.000	5.793	0.37892	14.000	5.778	2.58482
6.000	9.692	-0.78882	17.000	10.951	1.17417
8.000	6.792	0.33789	10.000	11.146	-0.21933
28.000	13.166	2.48276	7.000	7.566	-0.14622
4.000	7.036	-0.82720	3.000	10.686	-1.52114
10.000	9.650	0.07514	9.000	13.948	-0.78993
4.000	7.524	-0.91338	2.000	6.445	-1.29159

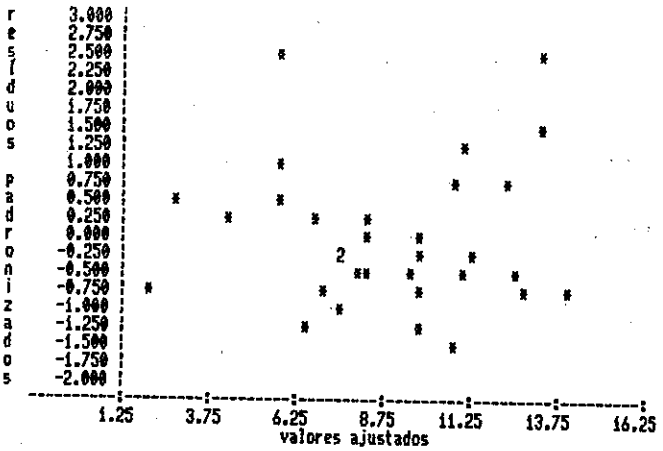

Figura 8.7 - *Gráfico dos resíduos padronizados versus médias ajustadas* $\tilde{\mu}$.

§8.7.2 Modelo de Quase-Verossimilhança com Parâmetro de Dispersão Não-Constante

Admite-se agora uma classe de modelos de quase-verossimilhança com parâmetro de dispersão não-constante

$$\eta = g(\mu) = X\beta, \quad \tau = h(\phi) = Z\gamma, \tag{8.26}$$

onde $\mu_i = E(y_i)$, 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 (8.19) para obter as estimativas de quase-verossimilhança dos β's, e então γ será escolhido visando a

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

$$(8.27)\quad U_\beta^+ = \partial Q^+/\partial\beta = X^T W H(y-\mu),\ U_\gamma^+ = \partial Q^+/\partial\gamma = \frac{1}{2} Z^T L(D-\phi),$$

onde $W = \text{diag}\,\{\phi^{-1}V(\mu)^{-1}g'(\mu)^{-2}\}$, $H = \text{diag}\,\{g'(\mu)\}$, $L = \text{diag}\,\{\phi^{-2}h'(\mu)^{-1}\}$ e $D = (D(y_1;\mu_1),\cdots,D(y_n;\mu_n))^T$. 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

$$(8.28)\qquad \tilde{X}^T\tilde{W}^{(m)}\tilde{X}\rho^{(m+1)} = \tilde{X}^T\tilde{W}^{(m)}\tilde{y}^{*(m)}.$$

onde $\tilde{X} = \begin{pmatrix} X & 0 \\ 0 & Z \end{pmatrix}$, $\tilde{W} = \begin{pmatrix} W & 0 \\ 0 & 1/2C \end{pmatrix}$, $\tilde{H} = \begin{pmatrix} H & 0 \\ 0 & C^{-1}L \end{pmatrix}$, $\tilde{y}^* = \begin{pmatrix} \eta \\ \tau \end{pmatrix} + \tilde{H}\begin{pmatrix} y-\mu \\ D-\phi \end{pmatrix}$, $C = \text{diag}\,\{\phi^{-2}h'(\phi)^{-2}\}$. A matriz C tem elementos obtidos da aproximação de 1ªordem $E\{D(y;\mu)\} = \phi$.

Assim, ajustar o modelo de quase-verossimilhança (8.26) 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 (8.28) pode ser feita usando os softwares já citados nesta seção.

Analisa-se os dados de toxoplasmose que constam da Tabela 8.7 onde y é o número de indivíduos com teste positivo para esta doença na amostra total de m indivíduos em 34 cidades de El Salvador caracterizadas pela precipitação anual de chuva x em mm. Admite-se que os dados (y) seguem um

modelo binomial com média μ e índice m, com uma componente sistemática logística dupla definida por:

$$\log[\mu/(y-\mu)] = \eta = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3,$$

$$\phi = (1+e^{-\tau})/1.25, \quad \tau = \gamma_1 + \gamma_2 M + \gamma_3 M^2,$$

Tabela 8.7 - *Dados de Toxoplasmoses em 34 Cidades de El Salvador e Valores Ajustados.*

y/m	0.500	0.300	0.200	0.300	1.000	0.600	0.250	0.368	0.500	0.800
x	1735	1936	2000	1973	1750	1800	1750	2077	1920	1800
m	4	10	5	10	2	5	8	19	8	10
$\tilde{\mu}/m$	0.539	0.446	0.384	0.408	0.547	0.550	0.547	0.337	0.462	0.550
y/m	0.292	0.000	0.500	0.182	0.000	0.545	0.000	0.611	0.444	0.278
x	2050	1830	1650	2200	2000	1770	1920	1770	2240	1620
m	24	1	30	22	1	11	1	54	9	18
$\tilde{\mu}/m$	0.348	0.540	0.420	0.395	0.384	0.552	0.462	0.552	0.472	0.346
y/m	0.167	0.000	0.727	0.532	0.471	0.438	0.561	0.692	0.535	0.707
x	1756	1650	2259	1796	1890	1871	2063	2100	1918	1834
m	12	1	11	77	51	16	82	13	43	75
$\tilde{\mu}/m$	0.549	0.420	0.497	0.551	0.492	0.509	0.342	0.332	0.464	0.535
y/m	0.615	0.300	0.167	0.622						
x	1780	1900	1976	2292						
m	13	10	6	37						
$\tilde{\mu}/m$	0.552	0.482	0.406	0.628						

onde $X_1 = (x_1 - \overline{x})/[\Sigma(x_1 - \overline{x})^2/33]^{1/2}$, $M_i = (m_i - \overline{m})/[\Sigma(m_i - \overline{m})^2/33]$ são valores padronizados. A regressão quadrática sobre o tamanho padronizado

da amostra para o parâmetro de dispersão ϕ na variância de y, Var $(y) = \phi\mu(m-\mu)/m$, foi proposta por Efron (1986).

O programa de ajustamento deste modelo através do GLIM é bastante simples (Cordeiro e Demétrio, 1989) ocorrendo a convergência após 14 iterações para satisfazer o critério: soma de quadrados da diferença entre estimativas sucessivas menor do que 0.0001. As estimativas e erros padrões entre parentêses obtidos são: $\tilde{\beta}_i$ =-0.1023 (0.1300), $\tilde{\beta}_2$ =-0.700 (0.2138), $\tilde{\beta}_3$ = -0.1653 (0.1009), $\tilde{\beta}_4$ =0.2819 (0.07985), $\tilde{\gamma}_1 = 2.151$ (3.131), $\tilde{\gamma}_2$ =0.4489 (3.397) e $\tilde{\gamma}_3$ =-0.7811 (1.684). As probabilidades ajustadas $(\tilde{\mu}/m)$ estão apresentadas também na Tabela 8.7. As Figuras 8.8 e 8.9 correspondem aos gráficos das freqüências observadas (y) versus as médias ajustadas $(\tilde{\mu})$ e dos resíduos padronizados $(y-\tilde{\mu})/$ Var $(y)^{1/2}$ com a variância sendo estimada em $\tilde{\beta}$, $\tilde{\gamma}$ versus $\tilde{\mu}$, respectivamente. Estes gráficos revelam que o modelo ajustado parece adequado.

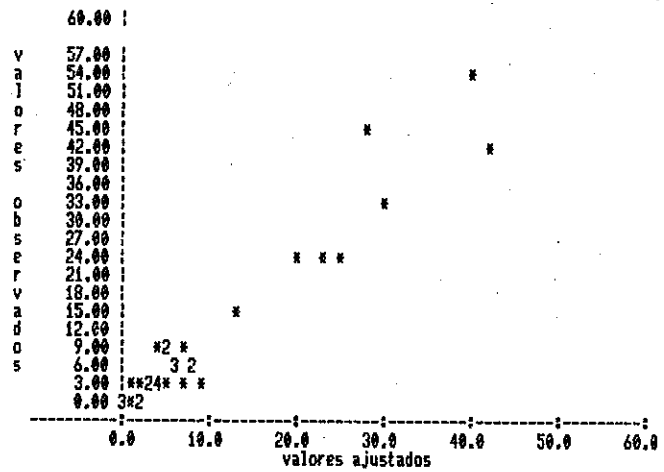

Figura 8.8 - *Gráfico dos números de indivíduos com toxoplasmoses versus médias ajustadas.*

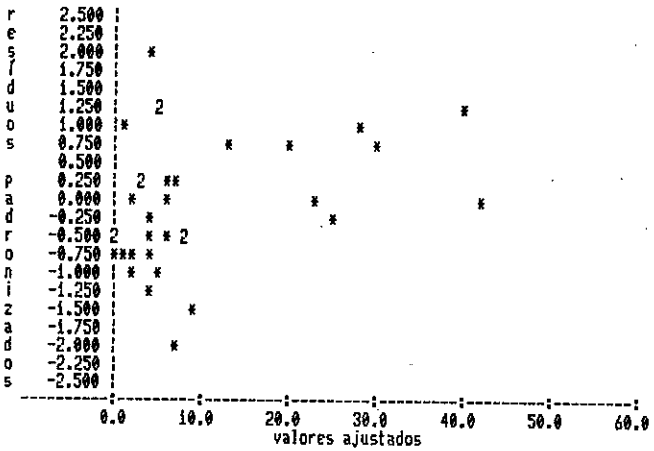

Figura 8.9 - *Gráfico dos resíduos padronizados versus médias ajustadas.*

§8.8 Modelos de Regressão com Estrutura de Autocorrelação Interna

Os modelos mais comuns com estrutura de autocorrelação interna são os modelos de séries temporais (Seção 1.17). Os modelos ARMA de Box e Jenkins poderão ser colocados na forma de ajustamento dos MLGs a partir da decomposição da matriz de informação K desses modelos (Expressão (1.62)), suposta particionada em relação à variância σ^2 e aos parâmetros lineares (parâmetros da parte autoregressiva e de médias móveis), na forma $L^T L$. Esta decomposição poderá sempre ser feita numericamente, por

programas especiais, acoplando à mesma o algoritmo de ajustamento dos MLGs.

Outros modelos de séries temporais apresentam equações de máxima verossimilhança idênticas às equações de quase-verossimilhança (8.19). Por exemplo, considere a estimação da densidade espectral $f(w;\beta)$ de uma série temporal estacionária y_t, $t = 1, \cdots, n$. Pode- se usar o método de máxima verossimilhança supondo que as ordenadas do periodograma $I(w_i)$, $w_i = 2\pi i/n$, $0 < i \leq [n/2]$, são variáveis aleatórias exponenciais independentes. A log-verossimilhança a ser maximizada iguala

$$L(\beta) = -\sum_{i=1}^{n}\{\log f(w_i;\beta) + I(w_i)/f(w_i;\beta)\}, \tag{8.29}$$

o que implica nas equações de máxima verossimilhança ($r = 1, \cdots, p$) serem do mesmo tipo de (8.19)

$$\sum_{i=1}^{n}[\{I(w_i) - f(w_i;\beta)\}/\{f(w_i;\beta)^2\}]\partial f(w_i;\beta)/\partial\beta_r = 0. \tag{8.30}$$

Estas equações podem ser resolvidas sem grandes dificuldades no SAS, GLIM ou GENSTAT, particularmente, se a densidade espectral pertencer à família exponencial.

§8.9 Outros Modelos Especiais

Considera-se agora que os dados y_{it}, $i = 1, \cdots, n$, supostos independentes, de n indivíduos no tempo $t = 1, \cdots, k$, com variáveis explicativas

x_{itr} $(r = 1, \cdots, p)$, seguem a seguinte estrutura, bastante apropriada para experimentos com medidas repetidas e para dados longitudinais,

$$(8.31) \qquad E(y_{it}) = \mu_{it}, \quad \text{Var}\,(y_{it}) = \phi V(\mu_{it}), \quad g(\mu_{it}) = \sum_{r=1}^{p} x_{itr}\beta_r.$$

As equações de quase-verossimilhança dadas em (8.31) podem ser usadas para estimar os parâmetros β's. Essas estimativas são consistentes e assintóticamente normais, com matriz de covariância que depende da estrutura $V(\cdot)$ especificada para os y's. A estrutura $V(\cdot)$ deverá ser estimada inicialmente, embora esta estimativa inicial exerça, em geral, pouca influência nas estimativas finais dos β's. Uma forma conveniente para optar por $V(\cdot)$ é definí-la indexada por um pequeno número de parâmetros. Optando-se por $k(k-1)/2$ parâmetros tem-se eficiência assintótica global quando n tende para infinito.

Uma vasta lista de modelos de regressão com objetivos práticos pode ser encontrada entre as referências de Cordeiro e Davison (1989). Como ilustração citam-se alguns desses modelos: modelos para análise de dados na escala ordinal (McCullagh, 1980), que inclui como caso especial o modelo de riscos proporcionais do Cox (1972), modelos de regressão log- logística para dados de sobrevivência (Bennett, 1983), modelos de regressão baseados na família exponencial dupla (Efron, 1986), modelos mistos lineares de multi-níveis (Goldstein, 1986), modelos de regressão semi-paramétricos (Green, 1987), modelos exponenciais não-lineares para dados categorizados com erros de observação (Palmgren e Ekholm, 1987), modelos lineares generalizados com erros nas covariáveis (Schafer, 1987), modelos lineares generalizados dinâmicos (West, Harrison e Migon, 1985) e vários outros modelos mistos e compostos.

§8.10 Exercícios

1. Considere um modelo de quase-verossimilhança com função de variância $V(\mu) = \mu^\lambda$. Estimar conjuntamente os parâmetros β e λ apresentando o algoritmo iterativo. Fazer o mesmo para a função $V(\mu) = \mu + \lambda\mu^2$ e aplicar aos dados da Tabela 8.5.
2. Aplicar o algoritmo (8.19) na análise dos dados da Tabela 8.5 considerando a função de variância $V(\mu) = \phi\mu$, sendo ϕ expresso por uma regressão linear simples sobre o logaritmo do comprimento da peça de tecido.
3. Considere um modelo de quase-verossimilhança com função de variância $V(\mu) = \mu^\lambda$ e componente sistemática de uma única covariável: $\eta_e = g(\mu_e) = \beta x_e$, $e = 1, \cdots, n$ com $g(\,\cdot\,)$ conhecido. Calcular:
   (a) as estimativas de MV de β e λ;
   (b) a matriz de covariância dessas estimativas
   (c) as estatísticas escore, de Wald e da razão de MV nos seguintes testes: $H_1{:}\lambda = \lambda^{(0)}$ versus $A_1{:}\lambda \neq \lambda^{(0)}$, $H_2{:}\beta = \beta^{(0)}$ versus $A_2{:}\beta \neq \beta^{(0)}$;
   (d) intervalos de confiança para λ e β.
4. (a) Aplicar o algoritmo descrito na Seção 8.3 na análise dos dados da Tabela 8.5 supondo erro de Poisson com ligação logaritmo.
   (b) Ajustar um modelo aditivo generalizado aos dados da Tabela 2.1, supondo erro normal com ligação logaritmo.
5. Como estimar conjuntamente os parâmetros λ_1, λ_2, ϕ e β no modelo (8.13)? Seria possível apresentar um algoritmo iterativo para este ajustamento?
6. Analisar os dados da Tabela 8.2 supondo erro gama. Como se compara

o melhor modelo obtido com o modelo normal correspondente a $\lambda_1 = 0.5$, $\lambda_2 = -1$ e discutido na Seção 8.6?

7. Calcular o valor de λ que minimiza o desvio $D(y;\tilde{\mu})$ tratado como função de λ para $V(\mu) = \mu^\lambda$ e $V(\mu) = \mu + \lambda\mu^2$.
8. Aplicar o algoritmo (8.19) a alguns modelos heterocedásticos com erro normal e ligação identidade.

REFERÊNCIAS

Aitkin, M. and Clayton, D. (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Appl. Statist. *29*, 156-163.

Akaike, H. (1974) A new look at the statistical model identification. IEEE Trans.Auto.Catl. AC-19, nº 6, 716-723.

Anscombe, F.J. (1953) Contribution to the discussion of H. Hotelling's paper. J. R. Statist. Soc. B, *15*, 229-230.

Arnold, S.F. (1981) The theory of linear models and multivariate analysis. John Wiley, New York.

Atkinson, A.C. (1981) Two graphical displays for outlying and influential observations in regression. Biometrika *68*, 13-20.

Atkinson, A.C. (1985) Plots, transformations, and regressions. Clarendon Press, Oxford.

Atkinson, A.C. (1986) Masking unmasked. Biometrika *73*, 533-541.

Baker, R.J. and Nelder, J.A. (1978) The GLIM System, Release 3, Generalized Linear Interactive Modelling. Numerical Algorithms Group, Oxford.

Bartlett, M.S. (1937) Properties of sufficiency and statistical tests. Proc. R. Soc. A *160*, 268-82.

Bates, D.M. e Watts, D. G. (1980) Relative Curvature Measures of Nonlinearity. J. R. Statist. Soc. B *42*, 1-25.

Beale, E.M.L. (1960) Confidence region in nonlinear estimation. J. R. Statist. Soc. B *22*, 41-76.

Beaton, A. E., Rubin, D. B. e Barone, J.L. (1976) The acceptability of regression solutions: Another look at computational stability. J. Amer. Statist. Assoc. *71*, 158-168.

Belsley, D.A., Kuh, E. e Welsch, R.E. (1980) Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley, New York.

Bennett, S. (1983) Log-logistic regression models for survival data. Appl.Statist. *32*, 165-171.

Bickel, P.J. e Doksum, K.A. (1977) Mathematical Statistics. Holden-Day, San Francisco.

Bishop, Y.M.M., Fienberg, S. E. and Holland, P. W. (1975) Discrete Multivariate Analysis: Theory and Practice. MIT Press, Cambridge, MA.

Bowman, K.O. e Shenton, L.R. (1975) Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2. Biometrika *62*, 243-250.

Box, G.E.P. e Cox, D.R. (1964) An analysis of transformations (with discussion) J.R. Statist.Soc. B *26*, 211-246.

Box, G.E.P. e Jenkins, G.M. (1976) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Box, G.E.P. e Tidwell, P.W. (1962) Transformation of the independent variables. Technometrics *4*, 531-550.

Box, M.J. (1971) Bias in non-linear estimation (with discussion). J. R. Statist. Soc. B *33*, 171-201.

Breslow, N. E. e Day, N.E. (1980) Statistical methods in cancer research, Vol. I: The Analysis of case-control studies. IARC, Lyon.

Businger, P. e Golub, G.H. (1965) Least squares by householder transformation. Num. Math. *7*, 269-276.

Bussab, W. (1986) Análise de Variância e de Regressão. Atual Editora, São Paulo.

Butter, D.E. e Stokes, D. (1957) Political change in Britain, 2nd edition. Macmillan, London.

Carter Jr., W.H., Wampler, G.L. e Stablen, D.M. (1983) Regression analysis of survival data in cancer chemotherapy. Marcel Dekker, Inc., New York.

Cook, R.D. (1977) Detection of influential observations in linear regression. Technometrics *19*, 15-18.

Cook, R.D. (1986) Assessment of local influence (with discussion). J. R. Statist. Soc. B *48*, 133-169.

Cook, R.D. e Tsai, C.L. (1985) Residual in nonlinear regression. Biometrika *72*, 23-29.

Cook, R.D., Tsai, C.L. e Wei, B.C. (1986) Bias in non-linear regression. Biometrika *73*, 615-623.

Cook, R.D. e Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall, London.

Cook, R.D. e Weisberg, S. (1983) Diagnostics for heteroscedasticity in regression. Biometrika *70*, 1-10.

Cordeiro, G.M. (1983) Improved likelihood ratio statistics for generalized linear models. J.R. Statist.Soc. B *45*, 404-413.

Cordeiro, G.M. (1985) The null expected deviance for an extended class of generalized linear models. Lecture Notes in Statistcs *32*, 27-34. Springer-Verlag, Berlin.

Cordeiro, G.M. (1986) Modelos Lineares Generalizados. Livro texto do VII SINAPE, Campinas.

Cordeiro, G.M. (1987a) On the corrections to the likelihood ratio statistics. Biometrika *74*, 265-274.

Cordeiro, G.M. (1987b) Infereência em modelos não-exponenciais não-lineares. Atas do 16º Colóquio Brasileiro de Matemática, 448-470. IMPA, Rio de Janeiro.

Cordeiro, G.M. e Davison, A. (1989) Generalized linear models, A review: 1972-1988 and an annotated bibliography. Manuscrito não-publicado.

Cordeiro, G.M. e Demétrio, C.G.B. (1989) An algorithm for fitting a quasi-likelihood model with a non-constant dispersion parameter. A aparecer no Lecture Notes in Statistics, Proceedings of the GLIM 89 International Conference. Springer-Verlag, Berlin.

Cordeiro, G.M. e Klein, R. (1989) Bias correction in ARMA models. A ser submetido.

Cordeiro, G.M. e Paula, G.A. (1986) Alguns modelos não-lineares via o GLIM. A aparecer nas Atas do VII SINAPE. UNICAMP, Campinas.

Cordeiro, G.M. e Paula, G.A. (1988) Estimation, significance tests and diagnostic measures for the non-linear exponential family nonlinear model. A aparecer nos Anais do III Congresso Latino americano de Probabilidade Estatística Matemática, Montevidéu, Uruguai.

Cordeiro, G.M. e Paula, G.A. (1989a) Improved likelihood ratio statistics for exponential family nonlinear models. Biometrika *76*, 93-100.

Cordeiro, G.M. e Paula, G.A. (1989b) Um algoritmo de ajustamento dos modelos não-exponenciais não-lineares em GLIM, através de "offset". A aparecer na Revista Ciência e Cultura.

Cordeiro, G.M. e Paula, G.A. (1989c) Fitting non-exponential family non-linear models in GLIM by using the offset facilities. A aparecer no Lecture Notes in Statistics, Proceedings of the GLIM 89 International Conference. Springer-Verlag, Berlin.

Cornfield, J. (1951) A method of estimating comparative rates from clinical data, applications to cancer of the lung, breast and cervix. J.Nat. Cancer Inst. *11*, 1269-1275.

Cornfield, J. (1956) A statistical problem arising from retrospective studies. In the Proceedings of the Third Berkeley Symposium, IV, (ed. J. Neyman), 135-148. University of California Press, Berkeley.

Cox, C. (1984) Generalized linear models - the missing link. Appl. Statist. *33*, 18-24.

Cox, D.R. (1970) The analysis of binary data. Chapman and Hall, London.

Cox, D.R. (1972) Regression models and life tables (with discussion). J. R. Statist.Soc. B *74*, 187-220.

Cox, D.R.(1975) Partial likelihood. Biometrika *62*, 269-276.

Cox, D.R. and Hinkley, D.V. (1979) Theoretical Statistics. Chapman and Hall, London.

Cox, D.R. and Oakes, D. (1984) Analysis of Survival data. Chapman and Hall, London.

Cox, D.R. and Snell, E.J. (1968) A general definition of residuals. J. R. Statist. Soc. B *30*, 248-275.

Dachs, J.N.W. e Carvalho, J.F. (1984) Diagnóstico em regressão. Livro texto do VI SINAPE. UFRJ, Rio de Janeiro.

D'Agostino, R.B. (1971) An omnibus test of normality for moderate and large size samples. Biometrika *58*, 341-348.

Dantas, R.A. e Cordeiro, G.M. (1989) A análise de dados imobiliários através do sistema GLIM. A aparecer na Revista Brasileira de Estatística. IBGE, Rio de Janeiro.

Darby, S.C. e Ellis, M.J. (1976) A test for synergism between two drugs. Appl.Statist. *25*, 296-299.

Day, N.E. e Byar, D.P. (1979) Testing hypotheses in case-control studies - equivalence of Mantel-Haenszel statistics and logit score tests. Biometrics *35*, 623-630.

Dixon, W.J. (1981) BMDP Statistical software. Univeristy of California Press, Los Angeles.

Draper, N.R. e Cox. D.R. (1969) On distributions and their transformations to normality. J.R. Statist.Soc. B *31*, 472-476.

Draper, N.R. e Smith, H. (1981) Applied regression analysis. John Wiley, New York.

Durbin, J. e Watson, G.S. (1950) Testing for serial correlation in least squares regression, I. Biometrika *37*, 409-428.

Efron, B. (1986) Double exponential families and their use in generalized linear regression. J. Amer.Statist. Assoc. *81*, 709-721.

Fienberg, S.E. (1980) The analysis of cross-classified data. The MIT Press, Cambridge, Massachusetts.

Freedman, D. Pisani, R. e Purves, R. (1978) Statistics. W.W. Norton, New York.

Gail, M.H., Lubin, J.H. e Rubinstein, L.V. (1981) Likelihood calculations for matched case-control studies and survival studies with tied death times. Biometrika *68*, 703-707.

Gallant, A.R. (1975) Nonlinear regression. Am. Statist. 29, 73-81.

Gart, J.J. and Zweifel, J.R. (1967) On the bias of various estimators of the logit and its variance with application to quantal bioassay. Biometrika *54*, 181-187.

Giltinan, D.M., Capizzi, T.P. e Malani, H. (1988) Diagnostic tests for similar action of two compounds. Appl. Statist. *37*, 39-50.

Godfrey, L.G. (1978) Testing for multiplicative heteroscedasticity. J.Econ. *8*, 227-236.

Goldstein, H. (1979) Specifying a multivariate logit model using GLIM. The GLIM Newsletter 1, 23-26. Oxford.

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika *73*, 43-56.

Goodman, L.A. (1970) The multivariate analysis of qualitative data: interactions among multiple classifications. J. Amer.Statist.Assoc. *65*, 226-256.

Goodman, L.A. (1971a) The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classification. Technometrics *13*, 33-61.

Goodman, L.A. (1971b) Partitioning of X^2 analysis of marginal contingency tables and estimation of expected frequencies in multidimensional contingency tables. J. Amer. Statist. Assoc. *66*, 339-344.

Goodman, L.A. (1973) The analysis of multidimensional contigency tables when some variables are posterior to others: a modified path analysis approach. Biometrika *60*, 179-192.

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.

Green, P.J. (1987) Penalized likelihood for general semi-parametric regression models. Int.Statist.Rev. *55*, 245-259.

Green, P.J. e Yandell, B.S. (1985) Semi-parametric generalized linear models. Lecture Notes in Statistics *32*, 44-55. Springer-Verlag, Berlin.

Guttman, I. e Meeter, D.A. (1965) On Beale's measures of non-linearity. Technometrics *7*, 623-637.

Haberman, S.J. (1974) The Analysis of Frequency Data. Univ. of Chicago Press, Chicago, Illinois.

Haberman, S.J. (1978a) Analysis of qualitative data, Vol. 1. Academic Press, New York.

Haberman, S.J. (1978b) Analysis of qualitative data, Vol. 2. Academic Press, New York.

Hannan, J. W. e Harkness, W.L. (1963) Normal approximation to the distribution of two independent binomials conditional on fixed sum. Ann.Math.Stat. *34*, 1593-1595.

Hastie, T. e Tibshirani, R. (1986) Generalized additive models. Statistical Science *1*, 297-318.

Hastie, T. e Tibshirani, R. (1987) Generalized additive models. Some applications. J. Amer.Statist.Assoc. *82*, 371-386.

Hinde, J. (1982) Compound Poisson regression models. Lecture Notes in Statistics *14*, 109-121. Springer-Verlag, New York.

Hoaglin, D.C. e Welsch, R.E. (1978) The hat matrix in regression and ANOVA. Amer. Statistician *32*, 17-22.

Hodges, S.D. e Moore, P.G. (1972) Data uncertainties and least squares regression. Appl.Statist. *21*, 185-195.

Hoerl, A.E. e Kennard, R.W. (1970) Ridge regression: Biased estimation for non-orthogonal problems. Technometrics *12*, 55-67.

Huber, P. (1973) Robst regression: Asymptotics, conjectures, and Monte Carlo. Ann. Statist. *1*, 799-821.

Jennrich, R.I. (1969) Asymptotic properties of nonlinear least-squares estimation. Annals Math.Statist. *20*, 633-643.

Johansen, S. (1983) Some topics in regression. Scand. J. Statist. *10*, 161-194.

Jørgensen, B. (1983) Maximum; likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. Biometrika *70*, 19-28.

Jørgensen, B. (1984) The delta algorithm and GLIM. Int. Statist.Rev. *52*, 283-300.

Jørgensen, B. (1987a) Exponential dispersion models (with discussion). J.R. Statist.Soc. B *49*, 127-162.

Jørgensen, B. (1987b) Small-dispersion asymptotics. Rev. Bras. Prob. Estat. *1*, 59-90.

Jørgensen, B. (1989) The theory of exponential dispersion models and analysis of deviance. Livro Texto da 1ª Escola de Modelos Lineares. IME/USP, São Paulo.

Kihlberg, J.K., Narragon, E.A. and Campbell, B.J. (1964) Automobile crash injury in relation to size. Cornell Report NO VJ-1823-R11.

Kleinbaum, D.G., Kupper, L.L. e Chambless, L.E. (1982) Logistic regression analysis of epidemiologic data: Theory and practice. Commun. Statist. - Theor. Meth., *11*, 485-547.

Kupper, L.L., Stewart, J.R. e Williams, K.A. (1976) A note on controlling significance levels in stepwise regression. Amer.J.Epidem. *108*, 435-446.

Landwehr, J.M., Pregibon, D. e Shoemaker, A.C. (1984) Graphical methods for assessing logistic regression models. J. Amer. Statist. Assoc. *79*, 61-71.

Larntz, K. (1978) Samll sample comparisons of exact levels for chi-squared goodness of fit statistics. J.Amer.Statist.Assoc. *73*, 253-263.

Lawley, D.N. (1956) A general method for approximating to the distribution of likelihood ratio criteria. Biometrika *43*, 295-303.

Lee, K. (1977) On the asymptotic variances of $\hat{\mu}$ terms in log linear models of multidimensional contingency tables. J.Amer.Statist.Assoc. *72*, 412-419.

Mallows, C.L. (1973) Some comments on Technometrics *15*, 661-676.

Mantel, N. e Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J.Nat. Cancer Inst. *22*, 719-748.

Mantel, N. e Hankey, W. (1975) The odds ratio of a 2×2 contingency table. Amer. Statistician *29*,143-145.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). J.R. Statist.Soc. B *42*, 109-142.

McCullagh, P. (1983) Quasi-likelihood functions. Ann.Statist. *11*, 59-67.

McCullagh, P. e Nelder, J.A. (1983) Generalized linear models. Chapman and Hall, London.

Montgomery, D.C. e Peck, E.A. (1982) Introduction to linear regression analysis. John Wiley, New York.

Morettin, P.A. e Toloi C.M.C. (1981) Modelos para previsão de séries temporais. Livro texto (2 volumes) do 13º Colóquio Brasileiro de Matemática. Poços de Caldas.

Nelder, J.A. (1985) GLIM 3.77 - Macros for univariate optimization of an arbitrary function. GLIM Newsletter *11*, 12-13.

Nelder, J.A. e Mead, R. (1965) A simplex method for function minimization. Comp.J. *7*, 308-313.

Nelder, J.A. e Pregibon, D. (1987) An extended quasi-likelihood function. Biometrika *74*, 221-232.

Nelder, J.A. e Wedderburn, R.W.M. (1972) Generalized linear models. J.R. Statist.Soc. A *135*, 370-384.

Palmgren, J. e Ekholm, A. (1987) Exponential family non-linear models for categorical data with errors of observation. Applied Stochastic Models and Data Analysis *3*, 111-124.

Paula, G.A. (1982) Testes de hipóteses para o risco relativo em estudos epidemiológicos. Dissertação de Mestrado, IMECC, UNICAMP.

Paula, G.A. (1987) Projected residuals in GLIM. GLIM Newsletter *15*, 28-31.

Paula G.A., Fontes, L.R. e Imanaga, A.T. (1984) Associação entre o tipo de processo infeccioso pulmonar e algumas variáveis histológicas. Relatório de Análise Estatística, nº 8417. IME/USP, São Paulo.

Paula, G.A., Sevanes, M. e Ogando, M.A. (1988) Estudo de plantas brasileiras com efeito moluscicida em biomphalaria glabrata. Relatório de Análise Estatística, nº 8824. IME/USP, São Paulo.

Pearson, E.S. e Hartley, H.O. (1976) Biometrika tables for statisticians, Vol. 1. Cambridge University Press, Cambridge.

Peto, R. (1972) In discussion of regression models and life tables. J.R. Statist.Soc. B *34*, 187-220.

Plackett, R.L. (1981) The analysis of categorical data. Griffin, London.

Pregibon, D. (1979) Data analytic methods for generalized linear models. PhD Thesis, University of Toronto.

Pregibon, D. (1982) Score tests in GLIM with applications. Lecture Notes in Statistics *14*, 87-97. Springer-Verlag, New York.

Pregibon, D. (1984) Data analytic methods for matched case-control studies. Biometrics *40*, 639-651.

Prentice, R.L. e Breslow, N.E. (1978) Retrospective studies and faiture time models. Biometrika *65*, 153-158.

Ramsey, J.B. (1969) Tests for specification errors in classical linear least squares regression analysis. J.R. Statist.Soc. B *31*, 350-371.

Rao, C.R. (1973) Linear statistical inference and its applications. John Wiley, New York.

Ratkowsky, D.A. (1983) Nonlinear regression modelling. Marcel Dekker, New York.

Rothman, K.J. (1977) Epidemiologic methods in clinical trials. Cancer *39*, 1771-1775.

Rothman, K.J., Fyler, D.C., Goldblatt, A. e Kreidberg, M.B. (1979) Exogenous harmones and other drug exposures of children with congenital heart disease. Amer. J. Epidem. *109*, 433-439.

Ryan, B.F., Joiner, B.L. e Ryan, T.A. Jr. (1985) Minitab Handbook. Duxbury Press, Boston.

Schafer, D.W. (1987) Covariate measurement error in generalized linear models. Biometrika *74*, 385-391.

Scheffé, H. (1959) The analysis of variance. John Wiley, New York.

Searle, S.R. (1971) Linear models. John Wiley, New York.

Seber, G.A.F. (1977) Linear regression analysis. John Wiley, New York.

Souza, D.G. (1986) Algumas considerações sobre regressão não-linear. Dissertação de Mestrado, IME - USP, São Paulo.

Stone, M. (1980) Discussion of paper by D.M. Bates e D.G. Watts. J. R. Statist. Soc. B *42*, 17-19.

Stukel, T.A. (1985) Implementation of an algorithm for fitting a class of generalized logistic models. Lecture Notes in Statistics *32*, 160-167.

Stukel, T.A. (1988) Generalized logistic models. J. Amer. Statist. Assoc. *83*, 426-431.

Thompson, R. e Baker, R.J. (1981) Composite link functions in generalized linear models. Appl.Statist. *30*, 125-131.

Tukey, J.W. (1949) One degree of freedom for non-additivity. Biometrics *5*, 232-242.

Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika *61*, 439-447.

Weisberg, S. (1985) Applied linear regression. John Wiley, New York.

West, M., Harison, P.J. e Migon, H.S. (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion). J. Amer. Statist. Assoc. *80*, 73-95.

Wetherill, G.B., Duncombe, P., Kenward, M., Köllerström, J., Paul, S. R. e Vowden, B. J. (1986) Regression analysis with applications. Chapman and Hall, London.

Whitehead, J. (1980) Fitting Cox's regression model to survival data using GLIM. Appl.Statist. 29, 268-275.

Williams, D.A. (1987) Generalized linear model diagnostic using the deviance and single case deletions. Appl.Statist. *36*, 181-191.

Young, D.H. e Bakir, S.T. (1987) Bias correction for a generalized log-gamma model. Technometrics *29*, 183-191.

PALAVRAS CHAVES

Algoritmo delta 266

Análise
- de variância 6
- do desvio 236-237

Autocorrelação interna 319

Censura 291-292

Classe geradora 114

Classificação
- bidimensional 104
- de dois fatores 37
- de um fator 37
- dos modelos de regressão 258-261
- multidimensional 113
- tridimensional 110,120
- unidimensional 101

Coeficiente de correlação linear múltiplo 8, 35, 42

Combinação de tabelas 139

Componente sistemática 224

Critério de informação 271

Curva de influência 245

Curvatura 195-197

Dados categorizados 96

Desvio 118-121, 233-234

Desvio estendido 311

Desvios sinalizados 161, 242, 273-275

Distância de Cook 208, 244

Distribuição

- binomial 223
- binomial negativa 223, 285
- condicional 99, 103, 105-106, 132, 134, 140
- de formas quadráticas 12
- de Poisson 96-98, 100-101, 223
  - inferência 100-101
- de Weibull 181,186,264, 274-275, 294-295
- do valor extremo 295
- exponencial 294-295
- exponencial dupla 4
- hipergeométrica 106,135
- hipergeométrica generalizada 106, 134
- multinomial 99, 101-104
- normal 223, 263, 274
- normal inversa 223, 263, 274
- gama 223, 250, 263, 274

Distribuições assintóticas 238

Eliminação de observações 22, 76, 208, 276

Empates 298

Engenharia de Avaliações 82

Erros correlacionados 33, 52, 319-320

Erros nas covariáveis 34

Estatística

da razão de verossimilhança 73, 102, 107, 118, 143, 158, 188, 200, 233, 278, 280

de A'gostino 80

de Cook 23, 244

de Godfrey 82

de Mallows 42

de Mantel-Haenszel 141, 144-145

de Pearson 102, 118, 136, 143, 240, 268

de Wald 158-159, 280-281

escore 158-159, 280-281

F 9, 15-16, 21, 79, 188, 235

PRESS 43

Estatísticas suficientes 103, 105, 110-112, 114, 116

Estimação

condicional 106, 134, 142

da transformação de Box e Cox 70

de James-Stein 5

de máxima verossimilhança 2, 53-54, 70, 105, 117, 156, 183, 187, 229-231, 265, 268, 287, 294

de mínimos quadrados 4, 183

de mínimos quadrados ponderados 29, 148, 308

do parâmetro de dispersão 234-235, 268, 270, 308

ridge 46

Estudos

de caso e controle 137

de seguimento 133

emparelhados 171-172

Família exponencial 222

Fator de correção 200-201

Fatores ortogonais 40

Fórmula de Cornish-Fisher 80, 98

Função

escore (ou suporte) 54, 116, 156, 230, 264

estabilizadora de variância 31-32, 68

de ligação 224-225

de ligação composta 287

de quase-verossimilhança 306, 309

de risco 292

de sobrevivência 292

de variância 222

de variância paramétrica 32, 310

de verossimilhança parcial 296

desvio 118, 156, 233, 268

Lagrangeana 10

GLIM 228, 231, 234, 243

Gráfico da variável adicionada 25, 72, 209, 276

Graus de liberdade 119, 157, 188, 233

Heterocedasticidade 48-50

Homocedasticidade 32, 79

Independência condicional 110, 112

Influência 21, 208, 243-246, 276

Intervalos de confiança 26-28, 157, 187-188, 238-239

Inversa generalizada 3, 37, 39

Jacobiano 71, 74, 195

Log-verossimilhança 49, 53-54, 70, 100, 116, 156, 187, 229, 264, 293-294

Log-verossimilhança condicional 134

Matriz

de variância-covariância 12, 238

de covariância assintótica 117, 157, 187, 238-240, 272, 280

de informação 50, 55-56, 117, 230, 265

de projeção 5, 19, 162, 204, 209, 241-242, 276, 281

Medida

de curvatura de Bates e Watts 195-197

de curvatura aparente 196

de curvatura intrínseca 196

de influência 21, 208, 243-246

Medidas

de ajuste 232-235

de não-linearidade 191-194

de diagnóstico 204-209, 241-246

Método

das covariáveis adicionadas 246-249

de escore de Fisher 49-51, 54, 117, 230, 265

de Newton-Raphson 49, 141, 156, 183, 230, 265

Métodos

de estimação 2

de seleção de modelos 41

Mínimos quadrados ponderados 28

Modelo

assintótico de regressão 186

beta 263

clássico de regressão 1, 68
dificuldades 30, 69
eliminação de observações 76

complementar log-log 169

de Box e Cox 70

de dispersão 259

de Poisson 96-98

de regressão polinomial 44
de regressão ridge 46

de von Bertalanffy 180, 184
de von Mises 263

dinâmico 53
em série de potência 263
hiperbólico 262-263

log-normal 263, 274
log-normal inverso generalizado 263
log-gama 262, 274

logarítmico 263

logístico condicional 171
logístico duplo 317

logístico generalizado 175
logístico linear múltiplo 151
logístico linear simples 146
maximal 234
multinomial 99
normal-linear 1
adequação da componente sistemática 77
adição de uma covariável 71
normal não-linear 178
nulo 232
probit 169
saturado 232-233
Modelos
aditivos generalizados 288-290
ARMA 53
auto-regressivos 319-321
com estrutura de correlação 52, 319-321
de quase-verossimilhança 306-319
de riscos proporcionais 292-299
de riscos não-proporcionais 299-301
de séries temporais 52, 319-320
definidos por duas transformações 302-306
encaixados 119
exponenciais de dispersão 260

exponenciais lineares 259
exponenciais não-lineares 259
exponenciais parcialmente não-lineares 181, 198, 201, 259

heterocedásticos 48

lineares generalizados 222
com ligação composta 287-288
com parâmetros extras 285-286

log-lineares 104-121
algoritmo de ajustamento 115
encaixados 119
hierárquicos 108-115

logísticos
lineares 146-147, 151-155
não-lineares 173

não-exponenciais
lineares 259
não-lineares 259

parcialmente não-lineares 181, 198, 201

semi-paramétricos 290-291

sigmoidais 180, 185, 202

Multicolinearidade 34, 46

Não-linearidade 31, 191-194, 198
aparente 192-194
intrínseca 192-194

Não-normalidade 33
Offset 267, 295

Parâmetro
de dispersão 223, 234-235, 259, 268, 270
de dispersão não-constante 315-316
natural 223

Parâmetros ortogonais 237, 266

Polinôminos ortogonais 45

Predição da regressão 25

Preditor linear 224

Procedimento "stepwise" 42

Quase-Verossimilhança 306-319
estendida 309

Regiões de confiança 16, 26-27, 187-188, 238-240, 278, 282

Reparametrização 192-194, 196, 198-199

Resíduo
de Pearson 161, 241, 275
projetado 204, 207, 210

Resíduos
componentes do desvio 161, 242-243, 273-275
ordinários 19
padronizados 19, 32, 162, 228, 241
parciais 276, 289
preditivos 43

studentizados 19, 80, 162, 207, 242, 275

Restrições nos parâmetros 10

Risco relativo 133, 141

Seleção de covariáveis 41, 159-161, 270-271

Suavizador
de espalhamento 288-289
linear 288-289, 291

Tabela
ANODEV 236-237
ANOVA 9, 14, 39, 46
de contingência 104
2×2 132-133, 139-140

Técnicas
de diagnóstico 18, 204-210, 240-246, 273-276
gráficas 24

Teorema de Fisher-Cochran 13

Teste
da variável adicionada 72
de Durbin-Watson 34
de heterocedasticidade específica 81
de homocedasticidade 81
de normalidade 80
de Tukey 77-78

Testes

de adequação 118-121, 234
de hipóteses 157-159, 188, 236, 248, 278-282
de transformação 73, 77
para os riscos relativos 141-144

Transformações de variáveis 31, 68, 73-74, 77
dependentes 73
explicativas 77

Transformações padronizadas 74

Valores iniciais 183-187

Variância não-constante 31, 222

Viés de Box. 198-200

Impresso na Gráfica do