Modelos Logísticos para Dados Binários

Giovani Loiola da Silva

Dissertação Apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para Obtenção do Grau de Mestre
em
Estatística

Área de Concentração: **Estatística**
Orientador: **Prof. Dr. Gilberto Alvarenga Paula**

- SÃO PAULO. Junho de 1992 -

Aos meus pais

AGRADECIMENTOS

Ao Prof. Dr. Gilberto Alvarenga Paula pela orientação desta dissertação e pelo apoio presente ao longo da mesma.

Ao Prof. Dr. Júlio da Motta Singer pelo seu incentivo sui generis durante a primeira fase do mestrado. Ao Prof. Dr. Carlos Daniel Mimoso Paulino pelas correções e sugestões para uma melhor clareza na apresentação deste texto. À Marcia Branco pela gentileza de providenciar a entrega do trabalho.

Aos companheiros que pela sua amizade me apoiaram ao longo do interminável período para conclusão do mestrado, entre outros, Paula e Fábio (USP), Rosa e Maurício (UFCE), Estela e Pedro (UFSCar), Delfina e Paulo (UTLisboa) e Cláudia e Ricardo (CRUSP).

Ao CNPQ e à CAPES pelo apoio financeiro para efetivação desta tarefa.

Abstract

The aim of this work is to present the principal regression models for binary data. Particular emphasis is given for the logistic regression models.

In Chapter 1 the classical statistical methods for binary data in contingency tables are presented. The linear logistic model, which has been largely applied in the Biomedical Sciences, is discussed in Chapter 2.

In Chapter 3 several logistic regression models are presented. These models were in general developed for the purpose of improving the extreme probability fittings. Special emphasis is given for the nonlinear logistic model proposed by Stukel(1988) and for the estimation of letal doses in quantal essays.

Several illustrative examples are given in the work and a particular program was developed for fitting the Stukel's model.

Resumo

O objetivo deste trabalho é apresentar os principais modelos de regressão para a análise de dados binários, com ênfase especial para os modelos logísticos de regressão.

No Capítulo 1 são apresentados os métodos tradicionais para tabelas de contingência com respostas binárias. O modelo logístico linear, que tem sido largamente utilizado na área de Biociências, é discutido no Capítulo 2.

No Capítulo 3 são apresentados diversos modelos logísticos generalizados, os quais foram desenvolvidos com o intuito de aperfeiçoar o ajuste de probabilidades extremas. Um destaque especial é dado ao modelo logístico não-linear proposto por Stukel(1988) e a estimação de doses letais em experimentos binomiais de dose-resposta.

Diversos exemplos são apresentados no decorrer do trabalho e um programa foi desenvolvido no "Software" SOC para o ajuste do modelo de Stuckel.

ÍNDICE

SUMÁRIO

Em várias áreas do conhecimento científico se coloca a questão do estudo estatístico do efeito, numa determinada variável de interesse, de várias outras variáveis que se julga terem um papel importante na explicação da variabilidade daquela. Para esse fim, existem vários métodos estatísticos cuja relevância e aplicação dependem da natureza das variáveis envolvidas. Esta dissertação confinar-se-á a situação onde a variável de interesse (resposta) é do tipo binário.

A análise de dados binários será aqui descrita numa forma sequencial através de 3 blocos de modelos estruturais cuja diferenciação ocorre na natureza das variáveis explicativas (covariáveis) incluidas e na complexidade de sua estrutura matemática. Todos eles se enquadram dentro do mesmo modelo probabilístico, o modelo produto de Binomiais independentes, cuja validade é portanto assumida ao longo da dissertação. A análise estatística de todos estes modelos basear-se-á na aplicação da metodologia de máxima verossimilhança.

O objetivo central deste trabalho é apresentar uma panorâmica dos vários métodos de análise de dados binários e contribuir para a definição de instrumentos que permitam a implementação computacional de métodos ainda pouco familiares ao analista de dados deste tipo.

O Capítulo 1 retrata a análise de dados binários por tabelas de contingência centrado no uso de certos parâmetros de interesse e na aplicação a algumas situações específicas. Estes métodos são considerados tradicionais na análise de dados binários, tendo sido muito utilizados no campo de Epidemiologia. Daí a relevância que é atribuida neste capítulo aos estudos prospectivos, com base nos quais se podem analisar os dados retros-

pectivos.

A análise descrita neste capítulo não se esgota em métodos assintóticos, contrariamente ao que se passará nos capítulos seguintes, já que uma parte desta é dirigida às inferências condicionais exatas frequentemente usadas (não sem controvérsia, dadas as vozes discordantes). Dois exemplos foram selecionados para apresentar a aplicação de parte destes métodos. O Apêndice 1 contém um programa na linguagem SOC por nós elaborado para efeitos de computação das estimativas de máxima verossimilhança em k tabelas 2x2.

A adaptação dos modelos de regressão a dados binários possibilita uma nova abordagem deste tipo de dados, sem a limitação da natureza categorizada das covariáveis presentes na abordagem tradicional. Esta, nova abordagem, a que o Capítulo 2 é dirigida, fica quase que circunscrita aos modelos de regressão logística, pela sua capacidade de interpretação e simplificação de cálculos inferênciais, face aos seus competidores (probit, complementar log-log, etc). A descrição das inferências paramétricas baseadas nestes modelos, do ajustamento destes e do problema de seleção de modelos ocupa uma parte substancial do conteúdo desse capítulo. A aplicação destes métodos é ilustrada através da análise de dois exemplos. O conjunto de dados para um destes exemplos encontra-se registrado no Apêndice 2.

O Capítulo 3 tem como objetivo apresentar várias generalizações do modelo de regressão logística, construidas na base da necessidade de melhoria do ajuste, eventualmente fraco, deste. Estes modelos têm a particularidade de, ao englobarem parâmetros de forma, permitirem uma maior flexibilidade da curva de resposta, propiciando assim uma maior capacidade de explicação dos dados.

Dentro desta extensa classe de modelos de regressão logística generalizados, dirigimos a nossa atenção a uma subclasse (os modelos de Stukel) pelas suas características mais englobadoras. A inspeção de sua estrutura funcional e a descrição do seu ajuste e do seu uso na "predição" de doses letais em dados de dose-resposta constituem uma parte significativa do conteúdo desse capítulo. A implemtentação deste tipo de inferências é possibilitada pela elaboração de programas computacionais via SOC que reproduziremos no Apêndice 3.

Por fim, a aplicação destes modelos generalizados e, em particular, dos modelos de Stukel é ilustrada com base em três exemplos, dos quais 2 são relativos a estudos de dose-resposta.

CAPÍTULO 1

MÉTODOS TRADICIONAIS PARA ANÁLISE DE DADOS BINÁRIOS

Os estudos com dados binários vêm se ampliando nas últimas décadas, principalmente na área epidemiológica com importante participação dos estudos em Cancerologia (Oncologia). Porém, só após a publicação de dois artigos, Cornfield (1951) e Mantel and Haenszel (1959), é que o uso da metodologia estatística na análise deste tipo de dados foi generalizado no campo epidemiólogico.

Gart (1971) elaborou um excelente artigo de revisão dos métodos estatísticos, aqui considerados tradicionais, para análise de dados binários. Breslow and Day (1980) apresentam um livro bastante didático direcionado para o estudo de caso-controle, que introduziremos no decorrer deste capítulo, com ênfase em pesquisa de câncer. Além destes, outros textos referem-se parcialmente a estes métodos e serão citados no decorrer desta apresentação.

Mas o que são dados binários ? Para esta caracterização nos referiremos primeiramente ao conjunto das variáveis envolvidas num experimento a ser analisado pelos métodos tradicionais em citação. Há uma variável de interesse (Y), habitualmente designada por variável resposta, possuindo somente dois valores possíveis. Em face da dicotomização dessa variável atribui-se por conveniência o código 1 a um dos valores, definindo geralmente o acontecimento de interesse ("sucesso"), e 0 ao outro valor ligado ao acontecimento

complementar ("insucesso"). Voltaremos a nos referir a esta escolha no próximo capítulo. Assim, Y é uma variável binária, fato que justifica a designação de dados binários para os dados resultantes de um estudo com uma variável resposta desse tipo. As demais variáveis envolvidas no experimento são chamadas covariáveis ou variáveis explicativas, podendo ser ou não binárias.

Como ilustração de algumas situações envolvendo variáveis respostas binárias, temos :

i) A ocorrência ou não de câncer pulmonar em indivíduos com hábito de fumar;

ii) O resultado positivo ou negativo de um exame médico;

iii) A quebra ou não de um componente resistente após um determinado teste de impacto;

iv) A sobrevivência ou não, após um período fixo, de um animal que recebeu uma dose de uma certa droga.

Nestes estudos o interesse concentra-se fundamentalmente nas probabilidades de ocorrência dos valores da variável Y, que denotaremos por

$$\pi_1 = P(Y = 1) \quad e \quad \pi_2 = P(Y = 0).$$

Um fator importante é a maneira de fazer a amostragem dos dados binários da qual depende o modelo probabilístico que se considera gerador dos dados. Tal amostragem pode conduzir, entre outros, aos estudos prospectivos (coorte ou seguimento) nos quais são fixos os totais correspondentes aos níveis das covariáveis, e aos estudos retrospectivos onde os totais fixos, identificados como os casos e os controles correspondem agora aos níveis da variável resposta.

Neste trabalho nos dedicaremos a análise de dados prospectivos. Os métodos considerados apenas necessitam de simples adaptações para se tornarem aplicáveis aos dados retrospectivos, aliás. bastante frequentes em Epidemiologia. A análise tradicional dos estudos retrospectivos com emparelhamento dos casos e controles pode ser vista em Breslow and Day (1980, ch.5).

Neste capítulo as covariáveis envolvidas serão consideradas categorizadas, razão pela qual os dados binários serão agrupados de maneira a formarem tabelas de contingência. Dessa maneira os estudos englobar-se-ão nas análises tradicionais de dados categorizados

(vide Bishop et al. (1975), Agresti (1990), etc.).

Apresentaremos em seguida três situações importantes de dados binários descritas por tabelas bi ou tridimensionais, acompanhadas de métodos tradicionais de análise. Propriedades das distribuições estatísticas mencionadas neste texto poderão ser encontradas em qualquer texto introdutório de estatística (e.g., Armitage, 1971).

1.1. Uma Única Tabela 2x2.

Consideremos inicialmente no estudo de dados binários uma única covariável também do tipo binário. Denotando-a por X, ela formará com a variável resposta Y uma tabela de contingência 2x2. Para ilustração considere-se o seguinte exemplo :

Exemplo 1.1(Estudo Prospectivo) : Em um estudo envolvendo câncer pulmonar e hábito de fumar é usual considerar a primeira variável como a variável resposta Y, onde a ocorrência ou não de câncer pulmonar em um indivíduo são designados, respectivamente, pelos acontecimentos $(Y = 1)$ e $(Y = 0)$. A população estudada é dividida em duas subpopulações, fumante e não fumante (níveis da covariável X) e, de cada uma, é extraída uma amostra aleatória simples. Os indivíduos amostrados ficam em observação por um período fixo de tempo - daí a denominação de estudo de seguimento - e ao fim desse período verifica-se quantos foram diagnosticados como portadores de câncer pulmonar entre os fumantes e os não fumantes. Esses dados formam uma tabela de contingência 2x2, ponto de partida para análise estatística. $\diamondsuit$

Geralmente no estudo de tabela 2x2 há interesse em verificar se há ou não associação entre as duas variáveis. As medidas de associação são as quantidades usuais que visam avaliar a intensidade de uma possível associação. No Exemplo 1.1 temos uma indagação imediata: Existe associação entre o hábito de fumar e a ocorrência de câncer pulmonar ? Em caso afirmativo, é possivel quantificá-la ?

No quadro da Tabela 1.1 podem ser definidas várias medidas de associação, através das proporções populacionais correspondentes aos quatro grupos (caselas) formados pelos níveis da variável resposta Y e da covariável X.

O risco relativo é uma medida de associação bastante conhecida na área epidemiológica, sendo definido como a razão entre a proporção de indivíduos com $(Y = 1)$ na subpopulação 1 (nível 1 da covariável X) e a proporção de indivíduos com $(Y = 1)$ na

outra subpopulação, ou seja,

$$\frac{p_1}{p_1 + p_2} \Big/ \frac{p_3}{p_3 + p_4}.$$

No Exemplo 1.1 o risco relativo é definido como o risco de um fumante contrair câncer pulmonar, $\{p_1/(p_1 + p_2)\}$, dividido pelo risco de um não fumante contrair essa doença, $\{p_3/(p_3 + p_4)\}$.

Tabela 1.1 : Proporções populacionais no estudo de uma única tabela 2x2.

Covariável	Variável resposta Y	
X	$Y = 1$	$Y = 0$
nível 1	p_1	p_2
nível 2	p_3	p_4

Se p_1 e p_3 são bem pequenos em relação a p_2 e p_4, respectivamente (como acontece com as doenças raras), o risco relativo é aproximado por uma outra medida de associação. denominada razão de chances ou razão de produtos cruzados (*Odds Ratio*), definida por

$$\psi = \frac{p_1 p_4}{p_2 p_3}. \tag{1.1}$$

A interpretação desse parâmetro quanto a associação entre a variável Y e a covariável X reflete-se em três situações : $\psi = 1$ indica inexistência de associação, $0 < \psi < 1$ representa associação negativa e $\psi > 1$ associação positiva. No Exemplo 1.1 a razão de chances pode ser traduzida como a chance de um fumante contrair câncer pulmonar (p_1/p_2) dividida pela chance de um não fumante adquirir tal doença (p_3/p_4).

Existem outras medidas de associação que são comumente baseadas na razão de chances, das quais a mais comum é $ln\psi$. Mais detalhes sobre medidas de associação podem ser encontrados em Goodman and Kruskal (1979) e em textos de análise de dados categorizados.

1.1.1. Modelo Probabilístico Não Condicional.

Uma vez definida a razão de chances como o principal parâmetro de interesse, passemos à consideração de um modelo probabilístico que sirva de base à realização das inferências pretendidas.

Como já foi referido, denotaremos a ocorrência de sucesso (insucesso) por $Y = 1$ ($Y = 0$). Os dois níveis de X serão codificados igualmente por $X = 1$ e $X = 0$. As probabilidades de sucesso para qualquer indivíduo de cada subpopulação são designadas, respectivamente, por $\pi_1 = P(Y = 1 \mid X = 1)$ e $\pi_2 = P(Y = 1 \mid X = 0)$. Supõe-se assim que estas probabilidades positivas e constantes para todos os indivíduos de cada subpopulação.

Sejam n_1 e $n_2 = n - n_1$ as dimensões das duas amostras aleatórias simples extraídas independentemente das subpopulações identificadas pelos níveis 1 ($X = 1$) e 2 ($X = 0$) da covariável, cujos tamanhos são considerados suficientemente grandes. Denotando por Y_1 e Y_2 a frequência (absoluta) de sucessos em cada uma das amostras referidas, fica claro que essas duas variáveis aleatórias são independentemente distribuídas (dado (n_i, π_i), $i = 1, 2$) segundo distribuições binomiais de parâmetros (n_1, π_1) e (n_2, π_2), respectivamente.

Os dados resultantes da classificação dos n indivíduos estão descritos na Tabela 1.2.

Tabela 1.2 : Distribuição de n indivíduos no estudo de uma única tabela 2x2.

Covariável	Variável resposta Y		
X	$Y = 1$	$Y = 0$	Total
nível $1(X = 1)$	y_1	$n_1 - y_1$	n_1
nível $2(X = 0)$	y_2	$n_2 - y_2$	n_2
Total	m	$n - m$	n

Pelo argumento desenvolvido, estes dados são considerados gerados pelo modelo produto de Binomiais (independentes), definido pela função de probabilidade conjunta

$$P(y_1, y_2) = \prod_{i=1}^{2} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} . \tag{1.2}$$

A função de verossimilhança será identificada pelo núcleo da função de probabilidade da amostra, e o seu logaritmo (natural) será denominado função log-verossimilhança e denotado por $L(\bullet)$. Para o modelo probabilístico (1.2) tem-se então

$$L(\pi_1, \pi_2) = \sum_{i=1}^{2} \{y_i ln\pi_i + (n_i - y_i) ln(1 - \pi_i)\} .$$

Esta função desempenhará um papel importante pelo fato de conduzir aos estimadores de máxima verossimilhança, denotados neste texto por **MV**, aos quais nos restringiremos pelas "boas" propriedades de que desfrutam. Neste caso, maximizando a função acima em relação aos parâmetros π_1 e π_2, encontramos os estimadores de **MV** desses parâmetros, dados, respectivamente, por y_1/n_1 e y_2/n_2. Logo, usando a propriedade invariante desses estimadores, temos para a razão de chances, definida em (1.1), o seguinte estimador de **MV**

$$\widetilde{\psi} = \frac{y_1(n_2 - y_2)}{y_2(n_1 - y_1)} \,. \tag{1.3}$$

Este estimador de ψ é viciado, ou mais concretamente, seu valor esperado é infinito (Gart , 1971); porém assintoticamente é não viciado com variância dada por

$$\psi^2 \left\{ \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)} \right\}.$$

O estimador $\widetilde{\psi}$ será chamado de estimador de **MV** não condicional de ψ, porque se baseia no modelo (1.2) dito não condicional para diferenciação com os outros modelos que consideraremos posteriormente. Observe que o modelo não condicional pode ser também considerado "condicional", já que temos fixos os totais marginais da covariável X. Neste texto tais totais serão sempre tomados como fixos, e, desse modo, as distribuições em estudo nunca serão entendidas como distribuições condicionais.

Como já foi referido, uma das funções da razão de chances mais usadas é $ln\psi$, cujo estimador de **MV** não condicional é $ln\widetilde{\psi}$ com variância assintótica dada por

$$\left\{ \frac{1}{n_1\pi_1(1-\pi_1)} + \frac{1}{n_2\pi_2(1-\pi_2)} \right\}. \tag{1.4}$$

Em estudos prospectivos os totais marginais n_1 e n_2 da Tabela 1.2 são frequentemente fixos, e o modelo (1.2) encaixa-se perfeitamente a esses estudos. Já os estudos retrospectivos têm como fixos não os totais marginais da covariável X, mas sim os totais marginais da variável resposta Y. Assim, para usar o modelo probabilístico (1.2) nos estudos retrospectivos devemos efetuar uma troca nas variáveis, do tipo : Y passa a funcionar como X, e vice-versa. Uma vantagem da razão de chances, sobre outras medidas de associação consiste precisamente na sua invariância face a várias modalidades de amostragem.

1.1.2. Modelo Probabilístico Condicional.

O carácter assintótico das inferências de interesse sob o modelo (1.2) e o confronto com situações de pequenas amostras levou à procura de modelos alternativos capazes de viabilizar inferências exatas.

Uma reparametrização do tipo $\pi_1 = \psi\pi_2/\{\psi\pi_2 + (1-\pi_2)\}$ na distribuição (1.2) produz uma nova apresentação dessa distribuição, ou seja,

$$P(y_1, y_2) = exp\left\{ y_1 ln\psi + (y_1 + y_2) ln\left(\frac{\pi_2}{1-\pi_2}\right)\right\} \frac{(1-\pi_2)^n}{(\psi\pi_2 + 1 - \pi_2)^{n_1}} .$$

Agora usando o critério da fatorização, temos o conjunto de estatísticas suficientes para $[ln\psi \ , \ ln\{\pi_2/(1-\pi_2)\}]$ dado por $(y_1 \ , \ y_1 + y_2)$. O fato de a distribuição (1.2) condicionada na estatística $y_1 + y_2 = m$ não depender do parâmetro perturbador $ln\{\pi_2/(1-\pi_2)\}$ (vide Lehmann, 1986, pg. 58), tem justificado a sua escolha como modelo base para a realização de inferências sobre o parâmetro de interesse ψ (vide, por exemplo, Cox and Snell, 1989, pg.27-28).

A eliminação do parâmetro perturbador π_2, através do condicionamento em $y_1 + y_2$ não é um procedimento pacífico dado que essa estatística não é ancilar para ψ. Os trabalhos de Basu (1977), Berkson (1978a,1978b) - com a discussão subsequente de Barnard, Basu, Corsten and de Kroon e Kempthorne (1979) - Upton (1982), Yates (1984) e Haber (1989) ilustram bem a polêmica existente na comunidade estatística.

O condicionamento em $y_1 + y_2 = m$ de (1.2) produz o modelo caracterizado pela família de distribuições Hipergeométricas não-centrais definidas por

$$P(y_1 \mid m; \psi) = \frac{\binom{n_1}{y_1}\binom{n_2}{m-y_1}\psi^{y_1}}{\sum_t \binom{n_1}{t}\binom{n_2}{m-t}\psi^t} \quad , \quad 0 < \psi < +\infty \ , \tag{1.5}$$

onde t varia do $max(0, m - n_2)$ ao $min(n_1, m)$.

Quando $\psi = 1$, a expressão (1.5) traduz a conhecida distribuição Hipergeométrica central, ou seja,

$$P(y_1 \mid m; \psi = 1) = \frac{\binom{n_1}{y_1}\binom{n_2}{m-y_1}}{\binom{n_1+n_2}{m}} , \tag{1.6}$$

cuja média e variância são, respectivamente,

$$E(1) = E(Y_1; \psi = 1) = \frac{mn_1}{n} \tag{1.7}$$

e

$$V(1) = Var(Y_1; \psi = 1) = \frac{n_1 n_2 (n-m)m}{n^2(n-1)} . \quad (1.8)$$

Para este modelo condicional a função log-verossimilhança é definida por

$$L(\psi) = y_1 ln\psi - ln \left\{ \sum_t \binom{n_1}{t} \binom{n_2}{m-t} \psi^t \right\} .$$

O estimador de **MV** condicional de ψ, $\widehat{\psi}$, é obtido como solução positiva da equação

$$y_1 = E(Y_1; \widehat{\psi}).$$

Atendendo a que o momento de ordem r da distribuição condicional, $E({Y_1}^r; \psi)$, pode ser expresso por $P_r(\psi) = E({Y_1}^r; \psi)P_0(\psi)$ onde $P_r(\psi) = \sum_t t^r \binom{n_1}{t}\binom{n_2}{m-t}\psi^t, r = 1, 2, \ldots$ e $P_0(\psi) = \sum_t \binom{n_1}{t}\binom{n_2}{m-t}\psi^t$, a equação de verossimilhança pode reescrever-se como

$$y_1 = \frac{P_1(\widehat{\psi})}{P_0(\widehat{\psi})}. \quad (1.9)$$

Com o aumento das marginais da Tabela 1.2, o procedimento para encontrar $\widehat{\psi}$ em (1.9) torna-se impraticável, pois a equação que dá origem ao estimador $\widehat{\psi}$ contém polinômios em ψ de grau bastante elevado. Nesse caso, uma saída é o uso de métodos numéricos para estimar iterativamente ψ sem precisar extrair as raízes dos polinômios $P_0(\psi)$ e $P_1(\psi)$. Este problema será discutido na próxima seção.

1.1.3. Testes de Hipóteses e Estimação Intervalar.

Uma vez apresentada a estimação pontual para o parâmetro de interesse ψ passaremos agora a testes de hipóteses e intervalos de confiança para ψ. Tais testes e intervalos podem ser construídos com base em ambos os modelos, não condicional e condicional.

Iniciando a análise pelas inferências condicionais exatas de ψ, seja a hipótese $H_0 : \psi = \psi_0$ versus $H_{11} : \psi < \psi_0$. O nível descritivo do teste dessa hipótese - probabilidade sob H_0 de obtenção de valores da estatística do teste tão ou mais desfavoráveis a H_0 (no sentido de H_{11}) do que o valor observado - é aqui definido pela probabilidade "nula" de cauda inferior da distribuição condicional, i. e.,

$$P_I = \sum_{t \leq y_1} P(t \mid m; \psi_0).$$

Note-se que P_I se mantém com a modificação de H_0 para $\psi \geq \psi_0$. Analogamente para testarmos a hipótese H_0 versus $H_{12} : \psi > \psi_0$ temos o nível descritivo do teste representado pela probabilidade "nula" da cauda superior dada por

$$P_S = \sum_{t \geq y_1} P(t \mid m; \psi_0).$$

Para o teste bilateral da hipótese H_0 versus $H_1 : \psi \neq \psi_0$ (vide Cox and Snell, 1989, pg.30), o nível descritivo do teste será $2\{min(P_I, P_S)\}$.

Quando nestes testes fazemos $\psi = 1$, estamos objetivamente testando se não há associação entre a variável resposta Y e a covariável X, sendo o teste resultante o conhecido teste exato de Fisher (Fisher, 1934).

Tabelas para facilitar os cálculos do teste exato de Fisher podem ser encontradas em Pearson and Hartley (1976, tab.38). Outros autores inspiraram-se no teste exato de Fisher para proporem testes exatos alternativos para tabelas 2x2, como por exemplo Davis (1986).

Para a estimação intervalar exata do parâmetro de interesse ψ, usando o modelo condicional, os respectivos limites de confiança são baseados nas probabilidades P_I e P_S. Denotando por ψ_I e ψ_S os limites de confiança inferior e superior de ψ, respectivamente, associados ao intervalo de confiança a $100(1-\alpha)\%$, eles ficam determinados pelas equações

$$\frac{\alpha}{2} = \sum_{t \leq y_1} P(t \mid m; \psi_I) \quad e \quad \frac{\alpha}{2} = \sum_{t \geq y_1} P(t \mid m; \psi_S).$$

Com o aumento dos totais marginais da Tabela 1.2, ou seja, quando n_1, n_2, m e $n - m$ são grandes, encontramos dificuldades em fazer inferências exatas. Neste caso usaremos duas distribuições assintóticas para inferir sobre ψ. A primeira distribuição assintótica é baseada na distribuição condicional, cuja aproximação será feita pela distribuição Normal (vide Hannan and Harkness, 1963), com média, $E_A(\psi)$, obtida pela solução da equação

$$\frac{E_A(\psi)\{n_2 - m + E_A(\psi)\}}{\{n_1 - E_A(\psi)\}\{m - E_A(\psi)\}} = \psi \ . \tag{1.10}$$

que para ψ fixo resulta numa equação quadrática em $E_A(\psi)$, e variância dada por

$$V_A(\psi) = \left[\frac{1}{E_A(\psi)} + \frac{1}{\{n_1 - E_A(\psi)\}} + \frac{1}{\{m - E_A(\psi)\}} + \frac{1}{\{n_2 - m + E_A(\psi)\}}\right]^{-1}. \tag{1.11}$$

Somente quando $\psi = 1$ a expressão (1.10) não resulta numa equação quadrática, verificando-se que a média assintótica é igual ao valor médio exato da distribuição condicional, expresso em (1.7). Consequentemente, $V_A(1) = \{n_1 n_2 m(n-m)\}/n^3$, em (1.11), sendo pois aproximadamente igual à variância exata da mesma distribuição, expressa em (1.8).

A equação (1.10) quando $\psi \neq 1$ apresenta duas raízes. Somente a raiz (que é única) que for maior do que o $max(0, m - n_2)$ e menor do que o $min(n_1, m)$ será aceita. Essa raiz (vide Paula, 1982) pode ser definida por

$$E_A(\psi) = || \, r \, | - s \, |,$$

onde $r = 0.5[\{n/(\psi - 1)\} + m + n_1]$ e $s = [r^2 - \{mn_1\psi/(\psi - 1)\}]^{1/2}$.

Após a estimação de $E_A(\psi)$ para um valor fixo de ψ, podemos também construir um teste de hipóteses do tipo $H_0 : \psi = \psi_0$, de modo semelhante ao feito no procedimento exato. Assim, as aproximações para as probabilidades das caudas inferior e superior do teste dessa hipótese nula, mediante a consideração de correções de continuidade, são, respectivamente, dadas por

$$P_I \simeq \Phi\left\{\frac{y_1 - E_A(\psi_0) + 1/2}{\sqrt{V_A(\psi_0)}}\right\} \; e \; P_S \simeq 1 - \Phi\left\{\frac{y_1 - E_A(\psi_0) - 1/2}{\sqrt{V_A(\psi_0)}}\right\},$$

onde $\Phi(\bullet)$ é a função de distribuição da Normal padrão.

Outra maneira de testar a hipótese nula, $H_0 : \psi = \psi_0$, em amostras grandes, é através da seguinte estatística

$$X^2 = \frac{\{| \, y_1 - E_A(\psi) \, | - 1/2\}^2}{V_A(\psi)}, \tag{1.12}$$

tendo, aproximadamente. distribuição Qui-Quadrado com 1 grau de liberdade, sob H_0. Quando a hipótese nula é $\psi = 1$, a estatística (1.12) torna-se

$$X^2 = \frac{\{| \, y_1(n_2 - m + y_1) - (m - y_1)(n_1 - y_1) \, | - n/2\}^2 (n-1)}{n_1 n_2 m(n-m)}. \tag{1.13}$$

No contexto de amostras grandes, consideremos agora o modelo não condicional como base de inferências aproximadas para o parâmetro de interesse ψ. Tem-se em Woolf

(1955) a distribuição assintótica de $ln\widetilde{\psi}$, que é uma Normal com média $ln\psi$ e variância $Var_A(ln\widetilde{\psi})$, expressa em (1.4). Assim, dado um estimador consistente para $Var_A(ln\widetilde{\psi})$, $\widehat{Var}_A(ln\widetilde{\psi})$, uma estatística alternativa a (1.13), sem correção de continuidade, para o teste da hipótese $H_0 : \psi = 1$, é dada por

$$X^2 = \left\{ \frac{ln\widetilde{\psi}}{\widehat{Var}_A(ln\widetilde{\psi})} \right\}^2 , \tag{1.14}$$

que sob H_0 tem igualmente uma distribuição assintótica Qui-Quadrado com 1 de liberdade. Ainda com relação a esta distribuição, temos o intervalo de $100(1-\alpha)\%$ de confiança para $ln\psi$, dado por

$$(ln\widehat{\psi}_I, ln\widehat{\psi}_S) = (ln\widetilde{\psi} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}_A(ln\widetilde{\psi})}) , \tag{1.15}$$

onde $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})\%$ da distribuição Normal padrão. Consequentemente, aplicando a função exponencial aos limites de confiança de $ln\psi$ acima, chega-se aos limites de $100(1-\alpha)\%$ de confiança para ψ, $(\widehat{\psi}_I, \widehat{\psi}_S)$.

Em algumas situações $Var_A(ln\widetilde{\psi})$ pode ser super-estimada, fato que tem servido para a apresentação de outros intervalos de confiança para ψ. Um destes procedimentos baseia-se no cálculo do valor da estatística (1.13) e na sua substituição em (1.14), de modo a obter-se uma expressão para $\widehat{Var}_A(ln\widetilde{\psi})$. Uma vez substituída em (1.15), obtém-se os seguintes limites de $100(1-\alpha)\%$ de confiança para ψ,

$$(\widehat{\psi}_I, \widehat{\psi}_S) \simeq (\widetilde{\psi}^{(1 \pm z_{1-\frac{\alpha}{2}}/x)}) , \tag{1.16}$$

onde x é a raiz quadrada do valor observado da estatística (1.13), sem correção de continuidade.

Elementos adicionais sobre inferências em uma única tabela 2x2 podem ser encontrados em Agresti (1990, ch.3).

1.2. Combinação de Tabelas 2x2.

Os métodos apresentados na Seção 1.1 para análise de dados binários agrupados em uma única tabela 2x2, podem sofrer reduções na eficiência de sua análise. Na maior parte das vezes tais reduções ocorrem devido ao fato de haver outras variáveis interferindo nas variáveis principais, variável resposta Y e covariável X, associadas à tabela 2x2. Essa

interferência pode vir através de associações existentes entre todas as variáveis citadas (confundimentos).

Nesta seção consideraremos somente uma variável causadora de confundimentos. Essa variável será denominada fator de confundimento. Um método de controle do fator de confundimento consiste em categorizar o mesmo, em digamos, k níveis, e na formação de uma tabela tridimensional kx2x2. Neste conjunto de k tabelas 2x2, podem evidenciar-se diversos padrões de associação entre as variáveis principais nos diferentes níveis do fator de confundimento, geralmente denominados de estratos ou blocos.

No Exemplo 1.1, onde a variável resposta (Y) é o diagnóstico de câncer pulmonar e a covariável X é o hábito de fumar, podemos considerar razoável que a variável idade seja introduzida no estudo como um fator de confundimento. Assim, os estratos serão as faixas etárias em que é categorizada tal variável.

As probabilidades de ocorrência da variável resposta, para os níveis 1 e 2 da covariável X, referentes ao i-ésimo estrato, serão denotadas por π_{i1} e π_{i2}, respectivamente. Destas probabilidades, formaremos o parâmetro de interesse em cada estrato, ou seja, a razão de chances do estrato i, dada por

$$\psi_i = \frac{\pi_{i1}(1-\pi_{i2})}{\pi_{i2}(1-\pi_{i1})} \ , \quad i = 1, \ldots, k \ . \tag{1.17}$$

A classificação dos n_i indivíduos, supostos amostrados aleatoriamente do estrato i, $i = 1, \ldots, k$, com relação aos níveis das variáveis principais, está exposta na Tabela 1.3.

Tabela 1.3: Distribuição de n_i indivíduos do estrato i.

Covariável	Variável resposta Y		
X	$Y=1$	$Y=0$	Total
nível $1(X=1)$	y_{i1}	$n_{i1}-y_{i1}$	n_{i1}
nível $2(X=0)$	y_{i2}	$n_{i2}-y_{i2}$	n_{i2}
Total	m_i	n_i-m_i	n_i

1.2.1. Inferência sobre as razões de Chances.

Como já vimos na seção anterior, necessitamos de modelos probabilísticos para fundamentar as inferências sobre os parâmetros de interesse. Assim, apresentaremos duas famílias de distribuições de probabilidades para a combinação das tabelas 2x2.

O primeiro modelo generaliza o modelo não condicional da Seção 1.1 no sentido em que as observações (Y_{i1}, Y_{i2}) em cada estrato são modeladas por um produto de duas Binomiais independentes, e consideradas independentes de estrato para estrato. Sendo assim, este produto de kx2 Binomiais continuará sendo chamado de modelo não condicional e definido pela função de probabilidade conjunta

$$P(\mathbf{y}_1, \mathbf{y}_2) = \prod_{i=1}^{k} \prod_{j=1}^{2} \binom{n_{ij}}{y_{ij}} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{n_{ij} - y_{ij}} \quad , \tag{1.18}$$

onde $\mathbf{y}_j = (y_{1j}, \ldots, y_{kj})^T$, j = 1,2.

O segundo modelo para as k tabelas 2x2, será formado a partir de cada modelo não condicional para cada tabela 2x2 por condicionamento no número total de sucessos em cada estrato. Como vimos anteriormente, o condicionamento de y_{i1} em $y_{i1} + y_{i2} = m_i$ resuta na distribuição Hipergeométrica não central,

$$P(y_{i1} \mid m_i) = \frac{\binom{n_{i1}}{y_{i1}} \binom{n_{i2}}{m_i - y_{i1}} \psi_i^{y_{i1}}}{\sum_{t_i} \binom{n_{i1}}{t_i} \binom{n_{i2}}{m_i - t_i} \psi_i^{t_i}} \quad , \tag{1.19}$$

onde t_i varia do $max(0, m_i - n_{i2})$ ao $min(n_{i1}, m_i)$, $i = 1, \ldots, k$. A conjunção de (1.19) para todos os estratos leva ao chamado modelo condicional definido pela função de probabilidade conjunta

$$P(\mathbf{y}_1 \mid \mathbf{m}) = \prod_{i=1}^{k} P(y_{i1} \mid m_i) \ ,$$

onde $\mathbf{m} = (m_1, \ldots, m_k)^T$.

A partir destes dois tipos de distribuições conjuntas, descreveremos a estimação pontual dos ψ_i's, definidos em (1.17). Usando inicialmente a distribuição conjunta não condicional de $(\mathbf{y}_1, \mathbf{y}_2)^T$, expressa em (1.18), deduz-se imediatamente que os estimadores de **MV** não condicional para ψ_i's são dados por

$$\widetilde{\psi}_i = \frac{y_{i1}(n_{i2} - y_{i2})}{y_{i2}(n_{i1} - y_{i1})} \ , \quad i = 1, \ldots, k. \tag{1.20}$$

A função log-verossimilhança para $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)^T$ correspondente ao modelo condicional é

$$L(\boldsymbol{\psi}) = \sum_{i=1}^{k} y_{i1} ln\psi_i - \sum_{i=1}^{k} lnP_0^{(i)}(\psi_i), \tag{1.21}$$

onde $P_0^{(i)}(\psi_i) = \sum_{t_i} \binom{n_{i1}}{t_i}\binom{n_{i2}}{m_i - t_i}\psi_i^{t_i}$. Consequentemente, os estimadores de MV condicional dos ψ_i's , $\widehat{\psi}_i$'s, são obtidos como solução das seguintes equações não lineares

$$y_{i1} = \frac{P_1^{(i)}(\widehat{\psi}_i)}{P_0^{(i)}(\widehat{\psi}_i)}, \quad i = 1, \ldots, k \quad , \tag{1.22}$$

onde $P_r^{(i)}(\psi_i) = \sum_{t_i} {t_i}^r \binom{n_{i1}}{t_i}\binom{n_{i2}}{m_i - t_i}\psi_i^{t_i}$, $r = 1, 2, \ldots$. E mais uma vez ficamos diante de dificuldades nos cálculos dos polinômios $P_0^{(i)}(\psi_i)$ e $P_1^{(i)}(\psi_i)$, $i = 1, \ldots, k$, para encontrarmos os $\widehat{\psi}_i$'s. Novamente, por conveniência, deixaremos para a próxima subseção a apresentação de um processo iterativo para solução das equações dadas em (1.22).

Para a construção de intervalos de confiança e testes de hipóteses referentes à razão de chances em cada estrato, podem-se usar os resultados apresentados na Seção 1.1.

1.2.2. Estimação Pontual da Razão de Chances Comum.

No estudo de k tabelas 2x2, uma nova questão surge naturalmente. As razões de chances são homogêneas ? Ou seja, entre os k estratos do fator de confundimento os ψ_i's permanecem constantes ? A hipótese referida é descrita por

$$H_{01} : \psi_1 = \psi_2 = \ldots = \psi_k (= \psi) \ . \tag{1.23}$$

onde ψ é o valor comum (desconhecido) dos ψ_i's. Embora a sua interpretação seja discutida no próximo capítulo, adiantamos já que a hipótese H_{01} corresponde a inexistência de interação entre o fator de confundimento e a covariável X.

Inicialmente suporemos verdadeira a hipótese (1.23). Neste contexto, apresentaremos quatro estimativas pontuais para ψ. A primeira é a estimativa de Mantel-Haenszel. bastante conhecida em Epidemiologia, devido à simplicidade do seu cálculo, tendo a forma

$$\widehat{\psi}_{MH} = \frac{\sum_{i=1}^{k} \{y_{i1}(n_{i2} - y_{i2}\}/n_i}{\sum_{i=1}^{k} \{y_{i2}(n_{i1} - y_{i1})\}/n_i} \ . \tag{1.24}$$

Vista de outra maneira, a estimativa de Mantel-Haenszel nada mais é do que uma média aritmética ponderada de todos $\widetilde{\psi}_i$'s, definidos em (1.20), i. e.,

$$\widehat{\psi}_{MH} = \frac{\sum_{i=1}^{k} v_i \widetilde{\psi}_i}{\sum_{i=1}^{k} v_i},$$

onde $v_i = y_{i2}(n_{i1} - y_{i1})/n_i$.

A segunda estimativa de ψ, sugerida em Woolf (1955), é uma média geométrica ponderada dos $\widetilde{\psi}_i$'s definida por

$$\widehat{\psi}_W = exp\left(\frac{\sum_{i=1}^{k} u_i ln\widetilde{\psi}_i}{\sum_{i=1}^{k} u_i}\right), \tag{1.25}$$

com os pesos $u_i = \{\frac{1}{y_{i1}} + \frac{1}{n_{i1}-y_{i1}} + \frac{1}{y_{i2}} + \frac{1}{n_{i2}-y_{i2}}\}^{-1}$, $i = 1, \ldots, k$. Observe que para compormos (1.25), devemos ter as frequências das caselas da Tabela 1.3 diferentes de zero, em cada estrato. Caso isso não ocorra, usualmente soma-se à frequência de cada casela uma constante a, $a > 0$, sendo frequente tomar-se $a = 0,5$. Feita essa adaptação o estimador é recomposto com base nas frequências modificadas.

A terceira estimativa de ψ é a estimativa de **MV** não condicional baseada no modelo (1.18) com a reparametrização, $\pi_{i1} = \pi_{i2}/\{\psi\pi_{i2} + (1 - \pi_{i2})\}$. Tal estimativa, denotada por $\widetilde{\psi}$, é obtida como solução não trivial de um sistema não linear de $(k+1)$ equações com $(k+1)$ incógnitas. O cálculo desta estimativa será deixado para o próximo capítulo, onde descreveremos como as dificuldades computacionais momentâneas na determinação de $\widetilde{\psi}$ são contornadas a partir de outros parâmetros mais fáceis de estimar.

A última estimativa de ψ que apresentaremos é a chamada estimativa de **MV** condicional, $\widehat{\psi}$. Tal estimativa é obtida diretamente da distribuição (1.21), sob a hipótese (1.23). De novo, $\widehat{\psi}$ é encontrado através da solução de uma equação não linear,

$$\sum_{i=1}^{k} y_{i1} = \sum_{i=1}^{k} E(Y_{i1}; \widehat{\psi}).$$

Este cálculo pode ser efetuado através do método iterativo de Newton-Raphson composto por

$$ln\psi^{(t+1)} = ln\psi^{(t)} + \frac{1}{V(\psi^{(t)})}\{y_{\bullet 1} - E(\psi^{(t)})\}, \quad t = 0, 1, \ldots, \tag{1.26}$$

onde $E(\psi) = \sum_i E_i(\psi)$, $V(\psi) = \sum_i V_i(\psi)$, $E_i(\psi) = E(Y_{i1}; \psi)$, $V_i(\psi) = Var(Y_{i1}; \psi)$ e $y_{\bullet 1} = \sum_i y_{i1}$, $i = 1, \ldots, k$.

O algoritmo (1.26) exige a determinação de $E_i(\psi)$ e $V_i(\psi)$, média e variância da distribuição condicional do estrato i, $i = 1, \ldots, k$, que, por sua vez, requer as raízes dos polinômios $P_0^{(i)}(\psi)$, $P_1^{(i)}(\psi)$ e $P_2^{(i)}(\psi)$. A determinação destas em amostras grandes confronta uma complexidade computacional ao processo. O procedimento proposto por McCullagh and Nelder (1983, pg.89-90) permite obter $E_i(\psi)$ e $V_i(\psi)$, $i = 1, \ldots, k$, em cada passo do processo iterativo, sem precisar encontrar as raízes dos polinômios envolvidos, como descreveremos em seguida.

A etapa inicial do processo (1.26) usa como valor inicial de ψ a estimativa de Mantel-Haenszel, dada em (1.24), i. e., $\psi^{(0)} = \widehat{\psi}_{MH}$. A variância inicial para cada estrato $V_i(\psi^{(0)})$, é a variância da distribuição (1.19), quando $\psi_i = 1$, dada em (1.8), adaptada ao estrato i, $i = 1, \ldots . k$. Daí decorre, naturalmente, $V(\psi^{(0)}) = \sum_i V_i(\psi^{(0)})$.

A média inicial do i-ésimo estrato, será então deduzida pela seguinte relação, apresentada por Mantel and Hankey (1975)

$$\frac{E_i(\psi)\{n_{i2} - m_i + E_i(\psi)\} + V_i(\psi)}{\{n_{i1} - E_i(\psi)\}\{m_i - E_i(\psi)\} + V_i(\psi)} = \psi,$$

que, para ψ e $V_i(\psi)$ fixos, resulta numa equação quadrática em $E_i(\psi)$, ou seja,

$$(1-\psi)E_i(\psi)^2 + \{n_{i2} - m_i + (n_{i1} + m_i)\psi\}E_i(\psi) + [V_i(\psi) - \{n_{i1}m_i + V_i(\psi)\}\psi] = 0. \tag{1.27}$$

A solução positiva dessa equação, para $V_i(\psi^{(0)})$ e $\psi^{(0)}$ fixos, define $E_i(\psi^{(0)})$, e consequentemente, $E(\psi^{(0)}) = \sum_i E_i(\psi^{(0)})$, $i = 1, \ldots, k$.

Substituindo $\psi^{(0)}$, $E(\psi^{(0)})$ e $V(\psi^{(0)})$ em (1.26), obtemos $ln\psi^{(1)}$ e consequentemente, $\psi^{(1)}$. As etapas seguintes do processo (1.26) usam o ψ estimado na etapa imediatamente anterior na equação (1.27) modificada, fazendo $V_i(\psi) = 0$, para a obtenção de $E_i(\psi)$ e, consequentemente, de $E(\psi)$. Utilizando $E_i(\psi)$ na expressão da variância assintótica, dada em (1.11), relativa ao estrato i, obtém-se $V_i(\psi)$ e, portanto, $V(\psi) = \sum_i V_i(\psi)$. Repetindo este procedimento até atingirmos uma convergência razoável encontraremos a estimativa de máxima verossimilhança condicional $\widehat{\psi}$ procurada.

Esta aproximação numérica produz resultados satisfatórios quando os totais marginais da Tabela 1.3 forem pelo menos iguais a 5. No caso contrário é aconselhável empregar

os cálculos exatos para estimar ψ. O processo (1.26) é também adequado para uma única tabela 2x2; logo podemos usá-lo para solução das equações não lineares (1.9) e (1.22).

No Apêndice 1 encontra-se um programa para a estimação de ψ, usando os procedimentos citados para a implementação do algoritmo iterativo (1.26). Um algoritmo alternativo a este, mais apropriado para pequenas amostras e de custo computacional relativamente pequeno, encontra-se em Gaill et al. (1981).

Os quatro estimadores de ψ referidos são consistentes e assintoticamente possuem distribuição Normal. Para amostras grandes mostra-se (vide Gart, 1962) que as variâncias assintóticas de $\widehat{\psi}_W$ e $\widetilde{\psi}$ são

$$Var_A(\widehat{\psi}_W) \;=\; Var_A(\widetilde{\psi}) \;=\; \psi^2 w^{-1} \quad ,$$

onde $w = \sum_i w_i$,

$$w_i = \left\{ \frac{1}{n_{i1}\pi_{i1}(1-\pi_{i1})} + \frac{1}{n_{i2}\pi_{i2}(1-\pi_{i2})} \right\}^{-1}, \quad i = 1, \ldots, k. \tag{1.28}$$

Quando $n = \sum_i n_i$ cresce, ou para n fixo, os totais n_{i1}, n_{i2}, m_i e $(n_i - m_i)$ crescem, $i = 1, \ldots, k$, o estimador $\widehat{\psi}$ concretizado pelo processo (1.26) converge em distribuição para uma Normal (vide McCullagh and Nelder, 1983, pg.90), cuja variância assintótica é dada por

$$Var_A(\widehat{\psi}) \;=\; \frac{\psi^2}{\sum_{i=1}^{k} Var(Y_{i1}; \psi)} \;,$$

onde $Var(Y_{i1}; \psi)$ é a variância da distribuição (1.19). Por último, ainda com suposição de amostras grandes, temos a variância assintótica do estimador de Mantel-Haenszel, dada por

$$Var_A(\widehat{\psi}_{MH}) \;=\; \psi^2 \frac{\sum_{i=1}^{k} z_i w_i^{-1}}{(\sum_{i=1}^{k} z_i)^2} \;,$$

onde w_i está expresso em (1.28) e $z_i = (n_{i1}n_{i2}/n_i)(1-\pi_{i1})\pi_{i2}$.

Mckinlay (1975) fez uma comparação, em termos do erro quadrático médio, entre dois dos estimadores há pouco apresentadas para ψ, $\widehat{\psi}_W$ e $\widehat{\psi}_{MH}$, e outro sugerido em Birch

(1964). Desse estudo concluiu a favor da estimativa de Woolf, $\widehat{\psi}_W$, por possuir na maioria das situações estudadas o menor erro quadrático médio. Breslow (1981) também discute os quatro estimadores de ψ em grandes amostras, destacando que as variâncias assintóticas dos estimadores $\widetilde{\psi}$, $\widehat{\psi}_W$ e $\widehat{\psi}$, são assintoticamente equivalentes. Contudo, devido a facilidade de cálculo, usa-se geralmente a estimativa de Mantel-Hanszel, $\widehat{\psi}_{MH}$.

1.2.3. Testes e Intervalos de Confiança para a Razão de Chances Comum.

Há pouco consideramos a hipótese de homogeneidade dos ψ_i's nos estratos, H_{01}, dada em (1.23), para estimação pontual do ψ_i comum. Agora apresentaremos três estatísticas para avaliação dessa hipótese num contexto de grandes amostras, e, em seguida, um intervalo de confiança para ψ na mesma situação.

A hipótese H_{01} pode ser testada, a partir do modelo condicional, pelo teste da razão de verossimilhanças de Wilks (1962), cuja estatística é

$$X^2_{MV} = -2\left[\sum_{i=1}^{k} y_{1i} ln(\widehat{\psi}/\widehat{\psi}_i) + \sum_{i=1}^{k} ln\{P_0^{(i)}(\widehat{\psi}_i)/P_0^{(i)}(\widehat{\psi})\}\right] ,$$

onde $\widehat{\psi}$ e $\widehat{\psi}_i$ são as estimativas de **MV** condicional de ψ_i, respectivamente restrita a H_{01} e irrestrita. Quando $n \to +\infty$, onde $n = \sum_i n_i$, esta estatística, sob H_{01}, tem distribuição assintótica Qui-Quadrado com $(k-1)$ graus de liberdade.

A segunda estatística para teste da hipótese H_{01} é mais simples do que a anterior, já que é baseada na soma ponderada dos quadrados dos desvios do logaritmo dos $\widetilde{\psi}_i$'s, em relação ao logaritmo de $\widehat{\psi}_W$ (vide Hosmer and Lemeshow, 1989, pg.74), ou seja,

$$X^2_H = \sum_{i=1}^{k} w_i (ln\widetilde{\psi}_i - ln\widehat{\psi}_W)^2 ,$$

onde $w_i, \widetilde{\psi}_i$ e $\widehat{\psi}_W$ estão expressos em (1.28), (1.20) e (1.25), respectivamente. De novo, sob H_{01}, esta estatística tem distribuição assintótica Qui-Quadrado com $(k-1)$ graus de liberdade.

A terceira e última estatística para testar a hipótese nula de homogeneidade dos ψ_i's (vide Breslow and Day, 1980, pg.142), é definida por

$$X^2_{BD} = \sum_{i=1}^{k} \frac{\{y_{i1} - E_i(\widehat{\psi})\}^2}{V_i(\widehat{\psi})} . \qquad (1.29)$$

onde as estimativas $E_i(\widehat{\psi})$ e $V_i(\widehat{\psi})$ são obtidas da solução positiva da equação (1.27), quando $V_i(\psi) = 0$, e da expressão (1.11), restrita ao estrato i, $i = 1, \ldots, k$, respectivamente. Se n é bem maior do que o número de estratos, a estatística (1.29), sob H_{01}, tem aproximadamente distribuição Qui-Quadrado com $(k-1)$ graus de liberdade.

Se a hipótese nula H_{01} não for rejeitada, a suposição da homogeneidade dos ψ_i's permite construir testes de hipóteses e intervalos de confiança para o ψ_i comum a partir das distribuições conjuntas citadas. No entanto, restringir-nos-emos a apresentação de um só teste de hipóteses e intervalo de confiança para ψ, baseado na distribuição conjunta condicional, sob H_{01}, e em amostras grandes. Outros testes e intervalos para ψ podem ser encontrados particularmente em Gart (1962, 1971 e 1985), Breslow and Day (1980, pg.141-143), Hosmer and Lemeshow (1989, pg.74).

Sob a validade de H_{01} a hipótese de não associação das variáveis Y e X corresponde a $H_{02} : \psi = 1$. O teste desta hipótese, citado em Mantel and Haenszel (1959), usa a estatística (com correção de continuidade)

$$X^2_{MH} = \frac{\{|\sum_{i=1}^{k} y_{i1} - \sum_{i=1}^{k} E_i(1)| - 1/2\}^2}{\sum_{i=1}^{k} V_i(1)} , \qquad (1.30)$$

onde $E_i(1)$ e $V_i(1)$ são a média e variância da distribuição condicional referente ao estrato i, expressas, respectivamente, em (1.7) e (1.8), uma vez adaptadas ao estrato i, $i = 1, \ldots, k$. Sob H_{02}, a estatística (1.30) tem distribuição assintótica Qui-Quadrado com 1 grau de liberdade. Mostra-se que este teste é equivalente ao teste de escore da hipótese H_{02} (vide Day and Byar (1979) e Pregibon (1982)), num modelo logístico linear.

Um intervalo de confiança para ψ pode ser construído de forma análoga ao exposto em (1.16), tomando como base a distribuição assintótica de $ln\widehat{\psi}$ e a estatística (1.30), sem correção de continuidade. A referida distribuição para $ln\widehat{\psi}$ é Normal, com média $ln\psi$ e variância estimada por $V_A(\widehat{\psi})^{-1}$, resultante da expressão (1.11), após o cálculo de $E(\widehat{\psi})$ na relação (1.10) (vide McCullagh and Nelder, 1983, pg.90). Os limites do intervalo de $100(1-\alpha)\%$ de confiança para ψ são dados por

$$(\widehat{\psi}_I, \widehat{\psi}_S) \simeq (\widehat{\psi}^{(1 \pm z_{1-\frac{\alpha}{2}}/x)}) , \qquad (1.31)$$

onde x é a raiz quadrada do valor observado da estatística X^2_{MH}, sem correção de continuidade, dada em (1.30) e $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})\%$ da distribuição Normal padrão.

referência, o qual formará com os demais níveis, as razões de chances. Aqui escolheremos o nível 1 como referência, sendo então as razões de chances

$$\psi_1 = 1 \quad e \quad \psi_j = \frac{\pi_j(1-\pi_1)}{\pi_1(1-\pi_j)} , \quad j = 2, \ldots, h , \tag{1.32}$$

onde π_j é a probabilidade de sucesso, $(Y = 1)$, de um indivíduo pertencente ao nível j da covariável X.

Para a situação estratificada as razões de chances passam a ser representadas por

$$\psi_{ij} = \frac{\pi_{ij}(1-\pi_{i1})}{\pi_{i1}(1-\pi_{ij})} , \quad i = 1, \ldots, k \; e \quad j = 1, \ldots, h , \tag{1.33}$$

onde π_{ij} é a probabilidade de sucesso para um indivíduo pertencente ao j-ésimo nível de X e ao estrato i do fator de confundimento $(\psi_{i1} = 1, \forall i)$.

A hipótese de ausência de associação completa entre as três variáveis envolvidas, será denotada por

$$H_0 : \psi_{ij} = 1 , \quad \forall \quad i = 1, \ldots, k \quad e \quad j = 1, \ldots, h . \tag{1.34}$$

O estudo da influência conjunta na resposta da covariável e do fator de confundimento sob o modelo não condicional (produto de kxh Binomiais independentes) é tratado nos textos correntes de análise de dados categorizados (uma boa referência é Agresti, 1990). Aqui nos concentraremos no modelo condicional, restringido a H_0, que é considerada como a única hipótese para a qual se dirige o nosso interesse.

É fácil constatar que o referido modelo é traduzido pelo produto de k distribuições Hipergeométricas (multivariadas) centrais, definido pela função de probabilidade conjunta

$$P(\mathbf{y} \mid \mathbf{m}) = \prod_{i=1}^{k} \frac{\prod_{j=1}^{h} \binom{n_{ij}}{y_{ij}}}{\binom{n_i}{m_i}} , \tag{1.35}$$

onde $\mathbf{y} = (y_{11}, \ldots, y_{1h}, y_{21}, \ldots, y_{kh})^T$, $\mathbf{m} = (m_1, \ldots, m_k)^T$ e $m_i = \sum_j y_{ij}$, $i = 1, \ldots, k$. A média, a variância e a covariância desta distribuição são, respectivamente, dadas por

$$E_{ij}(1) = E(Y_{ij} \mid m_i) = \frac{m_i n_{ij}}{n_i} . \tag{1.36}$$

$$Var(Y_{ij} \mid m_i) = \frac{n_{ij}(n_i - n_{ij})(n_i - m_i)m_i}{n_i^2(n_i - 1)} \tag{1.37}$$

e

$$Cov(Y_{ij}, Y_{il} \mid m_i) = \frac{-n_{ij}n_{il}m_i(n_i - m_i)}{n_i^2(n_i - 1)} , \tag{1.38}$$

onde $i = 1, \ldots, k$ e $j, l = 1, \ldots, h$.

Uma estatística para testar a hipótese global de não associação entre as variáveis Y e X nos vários estratos, expressa em (1.34), é dada em uma forma quadrática, ou seja,

$$X_Q^2 = \mathbf{q}^T \mathbf{A}^{-1} \mathbf{q} , \tag{1.39}$$

onde o vetor $\mathbf{q}$ é formado pelos desvios $y_{\bullet j} - E_{\bullet j}(1)$, $y_{\bullet j} = \sum_i y_{ij}$ e $E_{\bullet j}(1) = \sum_i E_{ij}(1)$, $j = 1, \ldots, h-1$, e a matriz de covariância $\mathbf{A} = [a_{jl}]$ pelos elementos

$$a_{jj} = \sum_{i=1}^{k} Var(Y_{ij} \mid m_i) , \quad j = 1, \ldots h-1 ,$$

$$a_{jl} = \sum_{i=1}^{k} Cov(Y_{ij}, Y_{il} \mid m_i) , \quad j \neq l = 1, \ldots, h-1 ,$$

cujas parcelas são definidas em (1.37) e (1.38). Quando os totais marginais da Tabela 1.4 em todos os estratos são grandes, a estatística (1.39), sob H_0, tem aproximadamente uma distribuição Qui-Quadrado com $h-1$ graus de liberdade.

Na situação não estratificada, a estatística (1.39) reduz-se à estatística usual do teste de homogeneidade de h proporções (vide Armitage, 1971), dada por

$$X_A^2 = \left(\frac{n_1 - 1}{n_1}\right) \sum_{j=1}^{h} \{y_{1j} - E_{1j}(1)\}^2 \left\{ \frac{1}{E_{1j}(1)} + \frac{1}{n_{1j} - E_{1j}(1)} \right\} . \tag{1.40}$$

Sob H_0, a distribuição da estatística (1.40), pode ser aproximada pela distribuição Qui-Quadrado com $h-1$ graus de liberdade.

Há situações em que a covariável X possui níveis quantitativos ou qualitativos ordinais. Um exemplo é dado pelo estudo da associação entre a dose de uma determinada droga e o percentual de cura de um tratamento específico, onde cura é a resposta com sucesso, $(Y = 1)$, e as doses, os níveis de X. Nestas condições, a estatística (1.40) não

leva em conta o valor real da dose obtida e sim a categoria que ela pertence. Daí a necessidade de recorrer a estatísticas alternativas que levem em conta a informação contida nos níveis quantitativos. Breslow and Day (1980, pg.149) sugerem uma estatística baseada na regressão dos desvios $\{y_{1j} - E_{1j}(1)\}$ em relação aos níveis quantitativos $x_j, j = 1, \ldots, h$, da covariável X. Essa estatística também tem origem na distribuição (1.35), quando $k = 1$, e é obtida pelo quociente da variável $\sum_j x_j\{y_{1j} - E_{1j}(1)\}$ e sua variância, resultando em

$$X_D^2 = \frac{n_1^2(n_1 - 1)\{\sum_{j=1}^{h} x_j(y_{1j} - E_{1j}(1))\}^2}{m_1(n_1 - m_1)\{n_1 \sum_{j=1}^{h} x_j^2 n_{1j} - \sum_{j=1}^{h} x_j^2 n_{1j}^2\}} . \tag{1.41}$$

Quando os n_{1j} crescem, $j = 1, \ldots, h$, a distribuição da estatística tende para a ditribuição Qui-Quadrado com 1 grau de liberdade. Uma generalização da estatística (1.41), para k estratos do fator de confundimento, encontra-se na mesma referência. A incorporação do carácter quantitativo da covariável na análise encontra um quadro favorável na metodologia que abordaremos no próximo capítulo, onde inclusive a consideração de várias covariáveis não provoca problemas maiores.

1.4. Exemplos.

Encerrada a apresentação de alguns métodos estatísticos tradicionais para análise de dados binários, passamos à ilustração dos mesmos através de dois exemplos. O primeiro exemplo serve-nos para aplicar a teoria exata, em pequenas amostras, e o outro a teoria apropriada para grandes amostras.

Exemplo 1.2 : Em Fisher (1935) descreve-se um experimento envolvendo apreciadores de chá. Uma mulher britânica se diz capaz de distinguir se o leite ou o chá é o primeiro a ser introduzido nas xícaras. Para testar sua afirmação lhe foram dadas 8 xícaras de chá com leite, onde somente em 4 delas o leite foi o primeiro a ser introduzido. Em seguida, as xícaras foram aleatorizadas, e apresentadas individualmente à mulher que deveria responder a favor do leite em 4 xícaras, e do chá nas restantes. Os resultados deste experimento encontram-se na Tabela 1.5.

Os totais marginais da Tabela 1.5 são pequenos, e, de acordo com o planejamento, fixos. Logo, é razoável adaptarmos à situação descrita, a teoria exata de uma tabela 2x2, Subseção (1.1.2). A distribuição (1.5) fundamentará o estudo da associação entre as

variáveis : verdadeira ordem de colocação do leite nas xícaras de chá, covariável X, e a resposta da britânica quanto à essa ordem, variável resposta Y.

Tabela 1.5 :Experimento da distinção do sabor de chá, quanto a colocação inicial do leite ou chá na xícara.

Colocação inicial na xícara	Resposta quanto a colocação inicial		
	Leite	Chá	Total
Leite	3	1	$4_{(n_1)}$
Chá	1	3	$4_{(n_2)}$
Total	$4_{(m)}$	$4_{(n-m)}$	$8_{(n)}$

O teste exato de Fisher será usado para analisar a possível existência de associação positiva, afirmação da britânica, entre a variável resposta Y e a covariável X. As hipóteses associadas ao teste são : $H_0 : \psi = 1$ (não existência de associação) versus $H_A : \psi > 1$ (associação positiva). Sob H_0, o nível descritivo deste teste é

$$P_S = \sum_{t \geq 3}^{4} P(t \mid m; \psi = 1) = \frac{\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}}{\binom{8}{4}} \simeq 0,243.$$

Devido ao valor de P_S, razoavelmente grande para um percentual de erro, concluise pela "aceitação" de H_0. Ou seja, os dados não favorecem a afirmação da britânica. Nesta base, crê-se ser inadequado prosseguir com a estimação do parâmetro de interesse ψ. $\Diamond$

Exemplo 1.3 : Innes et al. (1969) expõem alguns conjuntos de dados, entre os quais escolhemos um subconjunto desse estudo prospectivo. O objetivo do experimento consiste em verificar o possível efeito cancerígeno do fungicida Avadex (2.3 - Dichloroallyl diisoprophyl triolcarbamate) .

O planejamento deste experimento inicia-se com a observação de 403 camundongos. Destes, alguns foram alimentados com 560 ppm do fungicida Avadex, constituindo o grupo dos tratados. Os camundongos restantes, que não receberam este tratamento, formam o grupo controle.

Para eliminar os fatores de confundimentos estabeleceram-se estratos. Os estratos foram formados pela combinação de duas raças de camundongos, raça X e raça Y, e das

categorias de sexo. Após 85 semanas, os animais foram diagnosticados como portadores ou não de tumores pulmonares. O resultado deste experimento encontra-se na Tabela 1.6.

Tabela 1.6 :Distribuição dos camundongos, quanto a ocorrência de tumores pulmonares, causados pelo fungicida Avadex (Innes et al., 1969).

Estrato	Grupo	com tumor	sem tumor	Total
X-Machos	Tratados	4	12	16
	Controle	5	74	79
	Total	9	86	95
X-Fêmeas	Tratados	2	14	16
	Controle	3	84	87
	Total	5	98	103
Y-Machos	Tratados	4	14	18
	Controle	10	80	90
	Total	14	94	108
Y-Fêmeas	Tratados	1	14	15
	Controle	3	79	82
	Total	4	93	97

A ocorrência ou não de tumores pulmonares nos camundongos pode ser considerada como uma variável resposta binária Y. Os grupos de animais tratados e de controle representam os níveis da covariável X e os estratos correspondem aos níveis do cruzamento da raça com o sexo. A análise destes dados binários podem ser efetuada pelos métodos apresentados na Seção 1.2.

Diante do fato de termos totais marginais razoavelmente elevados para os 4 estratos (ver Tabela 1.6), torna-se impraticável analizarmos estes dados com base na distribuição exata. Assim, apoiaremos as inferências em métodos assintóticos. Por questões didáticas, dividiremos a análise dos dados em três partes. A primeira parte será baseada na distribuição assintótica não condicional e as partes restantes na distribuição assintótica condicional.

Tabela 1.7 :Análise estatística dos dados da Tabela 1.6.

Parte 1 (Estimação dos ψ_i's) :

Estrato		Int. Confiança (95%), (1.15)
X-Macho(1)	$\widehat{\psi}_1 = 4,93$	$1,16 < \psi_1 < 21,01$
X-Fêmea(2)	$\widehat{\psi}_2 = 4,00$	$0,61 < \psi_2 < 26,12$
Y-Macho(3)	$\widehat{\psi}_3 = 2,29$	$0,62 < \psi_3 < 8,33$
Y-Fêmea(4)	$\widehat{\psi}_4 = 1,88$	$0,18 < \psi_4 < 19,39$

Parte 2 (Testes de hipóteses) :

Hipótese	Estatística	G. liberdade	Nível descritivo
$H_{01} : \psi_1 = \ldots = \psi_4 (= \psi)$	$X^2_{BD} = 0,867$	3	0,8334
$H_{02} : \psi = 1$	$X^2_{MH} = 6,908$	1	0,0086

Parte 3 (Estimação de ψ) :

$\widehat{\psi}_{MH} = 3,079$; $\widehat{\psi}_W = 3,109$; $\widehat{\psi} = 3,093$;

Intervalo Confiança (95%), (1.31) : $1,33 < \psi < 7,18$.

Na parte 1 da Tabela 1.7 observa-se que as razões de chances nos 4 estratos variam, aproximadamente, entre 2 e 5. Isto possibilita a especulação de uma associação positiva entre as variáveis Y e X, dentro dos estratos. Por exemplo, no estrato 1, a razão de chances é $\widehat{\psi}_1 = (4 \cdot 74)/(5 \cdot 12) = 4,93$ (vide (1.20)), cuja interpretação é que a chance de ocorrência de tumor pulmonar no grupo tratado, é aproximadamente 5 vezes maior do que a chance no grupo controle. Os intervalos de confiança para os ψ_i's, foram determinados através de (1.15) com coeficiente de confiança de 95%. Dos intervalos produzidos, somente o primeiro não contém a unidade. Logo, concluímos que somente dentro do primeiro estrato existe evidência de associação entre o fungicida Avadex e a ocorrência de tumores pulmonares ao nível de 5% de significância.

Na parte 2 da Tabela 1.7 evidencia-se a "aceitação" da hipótese de homogeneidade dos ψ_i's, H_{01}, já que a estatística usada, expressa em (1.29), produziu o nível descritivo de 83,34%. Uma vez considerado o ψ comum nos 4 estratos, o teste de Mantel-Haenszel, dado em (1.30), gerou o nível descritivo de 0,86%. Logo, conclui-se a favor da rejeição de $\psi = 1$. Esta conclusão é confirmada pelas estimativas de ψ expostas na parte 3 da Tabela, em especial pelo intervalo de confiança nela referido. As estimativas pontuais sugerem que a chance de ocorrência de tumores pulmonares nos camundongos tratados com Avadex,

é aproximadamente 3 vezes maior do que a chance para o grupo controle. Em resumo, a análise aponta para a existência de uma associação positiva entre a ocorrência de tumores pulmonares e o tratamento com o agente fungicida. $\Diamond$

CAPÍTULO 2

MODELOS DE REGRESSÃO LOGÍSTICA

A análise de dados binários, que até agora foi baseada em tabelas de contingência, será a partir deste capítulo abordada numa perspectiva de análise de regressão. Se, por um lado, o estudo via dados agrupados pode facilitar a interpretação de certos resultados, por outro, este novo ponto de vista permite esclarecer certas questões que seriam de difícil análise no anterior.

Como os modelos tradicionais de regressão não podem ser aplicados diretamente aos dados binários, há necessidade de construir modelos especiais. Este capítulo preocupa-se basicamente com o modelo de regressão logística, sendo suas seções dedicadas a sua análise, ajustamento e interpretação. Na Seção 2.8 e no Capítulo 3 serão apresentados outros modelos para dados binários.

Referências ao modelo de regressão logística podem ser encontradas em textos já citados no Capítulo 1, entre outros, Cox and Snell (1989), Breslow and Day (1980) e Agresti (1990). O livro de Hosmer and Lemeshow (1989) é dedicado exclusivamente à análise do modelo em referência. Kleinbaum et al. (1982) concentram-se na aplicação deste modelo à área epidemiológica.

2.1. Introdução ao Modelo Logístico.

Como já foi referido no Capítulo 1, os estudos mais simples de dados binários envolvem uma variável resposta binária, denotada por Y, e uma covariável que denotaremos daqui em diante por x, em vez da correspondente letra maiúscula usada no capítulo anterior. A investigação de um "bom" modelo de regressão para descrever a relação entre Y e x deve basear-se, pelo tipo de modelo probabilístico envolvido, na probabilidade de sucesso, dada por

$$P(Y = 1 \mid x) = \pi(x). \tag{2.1}$$

Notemos que $\pi(x)$ representa o valor esperado de Y dado x, quantidade básica no modelo de regressão linear, definido por

$$E(Y \mid x) = \beta_0 + \beta_1 x \quad , \tag{2.2}$$

onde β_0 e β_1 são constantes reais desconhecidas. Todavia a expressão (2.2) é pouco apropriada para dados binários já que $\beta_0 + \beta_1 x$ varia ao longo da reta enquanto $E(Y \mid x)$ está necessariamente contida em (0,1).

Uma maneira de resolver a limitação citada acima é encontrar uma função matemática que se adapte à situação. A função logística, Kotz and Johnson (1985, pg.122-123), dada por

$$f(w) = \frac{e^w}{1+e^w} = (1+e^{-w})^{-1} \ , \quad -\infty < w < +\infty \ , \tag{2.3}$$

tem contradomínio em (0,1), e é monótona crescente em w. Assim, adaptando a equação de regressão (2.2) à função (2.3) temos, para $w = \beta_0 + \beta_1 x$, a probabilidade de sucesso

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \quad . \tag{2.4}$$

A expressão (2.4) é equivalente a

$$ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \ , \tag{2.5}$$

na qual se destaca a linearidade em termos do logaritmo da razão de chances de sucesso, geralmente denominado de *logit*. As expressões (2.4) e (2.5) são assim formas alternativas

de definição do modelo de regressão logística linear simples, ou simplesmente, modelo logístico simples.

Este modelo apresenta as seguintes características :

i) $\pi(x)$ fica restrito ao intervalo (0,1) para qualquer valor de x ;

ii) É essencialmente linear em x para $0,2 \leq \pi(x) \leq 0,8$;

iii) Os coeficientes de regressão são facilmente interpretáveis, em termos da razão de chances e do tipo de amostragem dos dados, como veremos posteriormente.

A análise de regressão logística é regida pelas mesmas linhas da análise de regressão linear, apesar de seus modelos não terem as suposições de normalidade e variância constante. Notemos que a variável resposta possui distribuição de Bernoulli com variância dependente de x e dada por

$$Var(Y) = \pi(x)[1 - \pi(x)] .$$

Os modelos logísticos serão , por enquanto, ajustados para estudos prospectivos, em concordância com a sua origem. A sua adaptação a estudos retrospectivos será referida numa seção futura deste capítulo.

É importante salientar que alguns modelos logísticos constituem casos especiais dos modelos log-lineares, relevantes na análise de dados categorizados (vide, e.g., Bishop et al (1975), Fienberg (1980) e Agresti (1990)).

2.2. Modelo Logístico Múltiplo.

A explicação da probabilidade de sucesso de uma variável resposta binária, em muitas situações, é feita através de um conjunto de covariáveis. Essas covariáveis podem ser fatores de confundimento, como vimos no Capítulo 1, interações entre variáveis explicativas ou até mesmo variáveis indicadoras de estratos.

Ao considerarmos várias covariáveis no modelo logístico temos o vetor de covariáveis $\mathbf{x} = (x_0, \ldots, x_{p-1})^T$, onde cada componente x_j, $j = 1, \ldots, p-1$, corresponde a um valor fixo de cada variável ($x_0 \equiv 1$) associado à probabilidade de sucesso do seguinte modo :

$$\pi(\mathbf{x}) = P(Y = 1 \mid \mathbf{x}) = \frac{e^{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p-1} \beta_j x_j}} , \qquad (2.6)$$

onde β_j é o parâmetro desconhecido associado à covariável x_j, $j = 1, \ldots, p-1$. A expressão (2.6), que se particulariza em (2.4) quando $p = 2$, define o chamado modelo de regressão logística múltipla que, por conveniência, denominaremos somente de modelo logístico. Outra forma de apresentá-lo é

$$logit[\pi(\mathbf{x})] = ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \sum_{j=1}^{p-1}\beta_j x_j \ . \tag{2.7}$$

Deste modo, β_j pode ser interpretado como variação do logaritmo das chances, provocada pelo acréscimo de uma unidade de x_j, $j = 1, \ldots, p-1$, com os valores das outras covariáveis permanecendo inalterados.

A formulação (2.7) permite verificar facilmente que a razão de chances de sucesso para dois valores distintos do vetor de covariáveis, $\mathbf{x}$ e $\mathbf{x}^*$, é

$$\psi_{\mathbf{x},\mathbf{x}^*} = \frac{\pi(\mathbf{x})[1-\pi(\mathbf{x}^*)]}{\pi(\mathbf{x}^*)[1-\pi(\mathbf{x})]} = exp\left[\sum_{j=1}^{p-1}\beta_j(x_j - x_j^*)\right]. \tag{2.8}$$

Para a situação apresentada, os dados observáveis para uma amostra aleatória de n indivíduos são formados pelas variáveis respostas Y_j (sucesso ou insucesso) e pelos valores $\mathbf{x}_j$ das covariáveis explicativas, $j = 1, \ldots, n$. Considerando as variáveis aleatórias Y_j independentes (condicionalmente aos $\mathbf{x}_j$) o modelo probabilístico ficará definido por um produto de n distribuições de Bernoulli.

Contudo, em muitas situações, o número de combinações possíveis dos valores das covariáveis é limitado (digamos k) e cada configuração i é apresentada por n_i indivíduos, $i = 1, \ldots, k$, de modo que $\sum_{i=1}^{k} n_i = n$.

Designando por π_i, $i = 1, \ldots, k$, a probabilidade de sucesso em cada configuração, suposta constante para todos os n_i (supostos fixos) indivíduos, podemos associar uma distribuição Binomial para o número de sucessos em cada configuração com parâmetros n_i e π_i. Deste modo, tais dados podem ser considerados gerados pelo modelo produto de Binomiais

$$P(\mathbf{y};\boldsymbol{\pi}) = \prod_{i=1}^{k}\binom{n_i}{y_i}\pi_i{}^{y_i}(1-\pi_i)^{n_i-y_i}, \tag{2.9}$$

onde $\mathbf{y} = (y_1, \ldots, y_k)^T$ e $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)^T$.

Os dados não agrupados por configurações, listados individualmente, podem ser considerados um caso particular dos agrupados. Em virtude disso tomaremos preferência por esta apresentação na análise de regressão logística.

2.3. Interpretação dos Parâmetros da Regressão Logística.

Uma das principais vantagens do modelo logístico é a facilidade de interpretação dos seus parâmetros devido ao fato de haver uma relação entre os coeficientes de regressão e a razão de chances, o principal parâmetro de interesse conforme foi visto no Capítulo 1.

A caracterização dos modelos será feita com os dados agrupados em configurações de covariáveis. Para cada configuração designaremos o preditor linear, o último membro da expressão (2.7), como

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \quad , \tag{2.10}$$

onde $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{i(p-1)})^T$ é o vetor de valores das covariáveis para a configuração i ($x_{i0} \equiv 1$), $i = 1, \ldots, k$, e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$. A função *logit*, primeiro membro da expressão (2.7) será denotada por

$$logit[\pi(\mathbf{x}_i)] = ln\left[\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right] \ , \ i = 1, \ldots, k \quad . \tag{2.11}$$

A matriz de especificação do modelo, contendo os k valores de todas as covariáveis, será designada por $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)^T$, usualmente chamada de matriz de planejamento.

2.3.1. Parâmetros no Modelo Logístico Simples.

O modelo logístico simples, expresso em (2.4) ou (2.5), será ilustrado para algumas situações particulares.

A primeira dessas situações é justamente a descrita na Seção 1.1, onde a covariável também é binária. Consequentemente, as probabilidades de sucesso para as duas configurações de x, definidas por $x = 1$ e $x = 0$, são dadas, respectivamente, por

$$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \quad e \quad \pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}} \ .$$

Consequentemente, a razão das chances de sucesso é traduzida por

$$\psi = \frac{\pi(1)[1-\pi(0)]}{\pi(0)[1-\pi(1)]} = e^{\beta_1} \ , \tag{2.12}$$

tornando evidente a interpretação de β_1 através de ψ.

Se a codificação da covariável binária x for em termos dos valores genéricos a e b temos

$$\psi = \frac{\pi(a)[1-\pi(b)]}{\pi(a)[1-\pi(b)]} = e^{\beta_1(a-b)} . \tag{2.13}$$

Se $a - b = 1$, a expressão (2.13) é igual a (2.12). Caso contrário a relação entre ψ e β_1 depende da codificação adotada. Por exemplo, usando $a = 1$ e $b = -1$, a codificação em termos dos desvios da média das caselas usada na análise de regressão *dummy*, na análise de variância usual e na modelagem log-linear tradicional, temos $\psi = e^{2\beta_1}$.

Deste modo, a interpretação correta dos parâmetros através de ψ não pode ser processada sem antes conhecermos a codificação adotada. De agora em diante adotaremos $a = 1$ e $b = 0$, usualmente conhecidos como codificação de casela de referência (a configuração padrão $x = 0$), por permitir uma interpretação simples e automática dos parâmetros.

A segunda situação que consideraremos é aquela onde a covariável é contínua. Neste caso o coeficiente β_1 pode ser interpretado como a variação ocorrida no logaritmo da razão de chances quando há um acréscimo de "1" unidade em x , i.e., $\beta_1 = h(x+1) - h(x)$, para todo valor de x, onde $h(x)$ é a função definida em (2.5).

Algumas vezes o acréscimo de "1" unidade na covariável pode não fazer muito sentido no contexto do problema. Por exemplo, em alguns estudos biológicos a verificação do efeito da idade dos indivíduos numa escala de 1 ano (ao invés, e.g., de 10 anos) não tem significado real. Assim, ao mudarmos de "1" para "c" a unidade de variação de x, o logaritmo da razão de chances muda em $h(x + c) - h(x) = c\beta_1$.

Caso haja suspeita que $h(x)$ não seja linear em x devemos agrupar a covariável contínua para análise, ou mesmo ajustar relações de outro tipo (quadrático, cúbico, etc); para maiores detalhes vide Hosmer and Lemeshow (1989, ch.4).

Um exemplo típico deste caso é a situação enunciada no final da Seção 1.3 e reexposta no Exemplo 2.1.

Exemplo 2.1 : Suponha que a covariável contínua x seja dosagem de uma determinada droga e o interesse do estudo é verificar a possível relação entre x e a probabilidade de cura. Admitindo que a probabilidade de cura para uma dosagem x, $\pi(x)$, é dada por (2.4), o logaritmo da razão de chances de cura para as duas doses, x_1 e x_2, é proporcional à

diferença entre essas doses, i.e., $ln(\psi) = \beta_1(x_1 - x_2)$. Deste modo, para $\beta_1 > 0$ a razão de chances de cura entre duas doses cresce com o aumento da dosagem e para $\beta_1 < 0$ ocorre o contrário. $\Diamond$

Consideraremos agora uma terceira situação com uma covariável politômica com p níveis, $p \geq 3$. Na adoção da parametrização de casela de referência (nível 1 de x) devemos considerar um vetor de dimensão $p-1$, cujos valores identificam os p níveis de x do seguinte modo : O nível 1 (de referência) é dado pelo vetor nulo e os restantes $p-1$ níveis pelos vetores $(1,0,0,\ldots,0)$, $(0,1,0,\ldots,0),\ldots,$ $(0,0,\ldots,1)$. O nível i, $i = 2,\ldots,p$ é assim representado pelo vetor resultante do vetor nulo substituindo por 1 a componente $i-1$. Por exemplo, no caso $p = 4$ temos a seguinte representação

$$\begin{array}{lcl} \text{nível } 1 & : & (\ 0\ \ 0\ \ 0\ ) \\ \text{nível } 2 & : & (\ 1\ \ 0\ \ 0\ ) \\ \text{nível } 3 & : & (\ 0\ \ 1\ \ 0\ ) \\ \text{nível } 4 & : & (\ 0\ \ 0\ \ 1\ )\ . \end{array}$$

Notemos que esta representação é equivalente a considerar $p-1$ variáveis *dummy* D_j, $j = 1,\ldots,p-1$, tal que o nível 1 é indicado por $D_j = 0$, $j = 1,\ldots,p-1$, e o nível i, $i = 2,\ldots,p$, por $D_j = 1$, $j = i-1$, e $D_{j'} = 0$, $j' \neq j = 1,\ldots,p-1$. Desta forma, cada nível corresponde a cada configuração possível destas $p-1$ variáveis de planejamento. Este conjunto de $p-1$ configurações pode ser representado por uma matriz $\mathbf{X}^* = [x_{ij}]$ de dimensão $px(p-1)$, onde cada coluna representa os valores x_{ij}, $i = 1,\ldots,p$, da j-ésima variável de planejamento e cada linha $\mathbf{x}_i^* = (x_{i1},\ldots,x_{i(p-1)})^T$ representa a configuração de valores das $p-1$ variáveis *dummy* identificadoras do nível i (notemos que $\mathbf{x}_1^* = \mathbf{0}^T$).

Denotando os parâmetros associados a cada variável *dummy* por β_{1j}, $j = 1,\ldots,$ $p-1$, o modelo logístico fica representado por

$$logit[\pi(\mathbf{x}_i)] \ = \ \beta_0 + \sum_{j=1}^{p-1} \beta_{1j} x_{ij}\ , \quad i = 1,\ldots,p\ . \tag{2.14}$$

No caso particular referido

$$logit[\pi(\mathbf{x}_i)] \ = \ \beta_0 + \mathbf{x}_i^{*T} \begin{pmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{13} \end{pmatrix}\ , \quad i = 1,\ldots,4\ ,$$

onde $\mathbf{x}_i^{*T}$ é a linha i de

$$\mathbf{X}^* = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \tag{2.15}$$

As razões das chances das diferentes configurações com relação à primeira,

$$\psi_{i1} = \frac{\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_1)]}{\pi(\mathbf{x}_1)[1-\pi(\mathbf{x}_i)]} , \ i = 1, \ldots, p$$

são assim traduzidas por

$$\psi_{11} = 1 \quad e \quad \psi_{i1} = e^{\beta_{1(i-1)}} , \quad i = 2, \ldots, p . \tag{2.16}$$

Devemos relembrar que o novo modelo poderá ser especificado por uma matriz de planejamento formada por uma outra codificação. Por exemplo. a codificação do tipo da utilizada em análise de variância, expressando os desvios da média dos vários *logits*. No caso de $p = 4$, a matriz (2.15) ficaria só com a primeira linha alterada, por substituição do vetor $(0,0,0)$ por $(-1,-1,-1)$. com esta nova reparametrização já não há uma relação direta entre os parâmetros e a razão de chances. Por exemplo, $\psi_{i1} = exp(2\beta_{1i} + \sum_{j \neq i} \beta_{1j})$, $i = 2,3,4$ e $\psi_{11} = 1$.

Se a covariável politômica em referência for considerada ordinal devemos obter melhores resultados na análise se usarmos em lugar da matriz do tipo de (2.15) uma matriz de polinômios ortogonais. Este método é bastante comum na análise de regressão linear, pois ele possibilita o estudo da tendência linear, quadrática,..., entre a variável resposta e os níveis crescentes da covariável. Hosmer and Lemeshow (1989, ch.4) ilustram a aplicação deste tipo de método na análise do modelo logístico.

2.3.2. Parâmetros no Modelo Logístico Múltiplo.

Ao considerarmos a interpretação dos parâmetros no modelo logístico com $p-1$ ($p > 2$) covariáveis dados em (2.6) ou (2.7), devemos ter cuidado com a possível existência de interações e (ou) confundimentos. Recordando, o termo confundimento é usado frequentemente na área epidemiológica para descrever a possível interferência (associação) das covariáveis na variável resposta. O termo interação, já conhecido da análise de regressão e da análise de variância, é usado quando nas configurações de covariáveis a suposição de linearidade da função *logit* é aceita, mas a inclinação não é constante.

Para fins ilustrativos, a primeira situação que consideraremos envolve duas covariáveis binárias, x_1 e x_2, sem interação. O modelo resultante de (2.7) tem a forma

$$logit[\pi(\mathbf{x})] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ , \tag{2.17}$$

onde $\mathbf{x} = (x_1, x_2)$ e x_i, $i = 1, 2$ são codificadas como 0 ou 1. Deste modo

$$\begin{aligned} \pi(0,0) &= (1 + e^{-\beta_0})^{-1} \ , \\ \pi(1,0) &= (1 + e^{-\beta_0 - \beta_1})^{-1} \ , \\ \pi(0,1) &= (1 + e^{-\beta_0 - \beta_2})^{-1} \text{ e} \\ \pi(1,1) &= (1 + e^{-\beta_0 - \beta_1 - \beta_2})^{-1} \ . \end{aligned}$$

Destas podemos construir razões de chances entre cada par de configurações de covariáveis; tomando por exemplo a primeira configuração como casela de referência, i.e.,

$$\psi_{ij} = \frac{\pi(i,j)[1 - \pi(0,0)]}{\pi(0,0)[1 - \pi(i,j)]} \ , \quad i, j = 0, 1 \ ,$$

consequentemente teremos

$$\psi_{00} = 1, \ \psi_{10} = e^{\beta_1 J}, \ \psi_{01} = e^{\beta_2 J} \ e \ \psi_{11} = e^{\beta_1 + \beta_2 J} \ . \tag{2.18}$$

Observe que ao assumirmos o modelo (2.17) encontramos em (2.18) a seguinte relação multiplicativa

$$\psi_{11} = \psi_{10} . \psi_{01} \ . \tag{2.19}$$

Ao acrescentarmos o termo interação no modelo (2.17) o tornamos num modelo saturado com a seguinte forma

$$logit[\pi(\mathbf{x})] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \ . \tag{2.20}$$

As razões de chances permanecem as mesmas, com exceção de ψ_{11} que passa a ser dada por $\psi_{11} = e^{\beta_1 + \beta_2 + \beta_3}$. Logo, $\beta_3 = ln[\psi_{11}/(\psi_{10} . \psi_{01})]$, o que torna evidente que a ausência de interação entre x_1 e x_2 ($\beta_3 = 0$) equivale à verificação da relação multiplicativa (2.19).

Outra situação em dados binários foi descrita na Seção 1.2, onde foram consideradas uma covariável binária e um fator de confundimento com k estratos. Esta situação pode ser traduzida pelo modelo logístico

$$\pi_i(x) = \frac{e^{\beta_{0i} + \beta_{1i} x}}{1 + e^{\beta_{0i} + \beta_{1i} x}} \ , \quad i = 1, \ldots, k \ , \tag{2.21}$$

onde x assume os valores 1 ou 0. A razão de chances dentro de cada estrato fica dada por

$$\psi_i = \frac{\pi_i(1)[1-\pi_i(0)]}{\pi_i(0)[1-\pi_i(1)]} = e^{\beta_{1i}} , \quad i = 1, \ldots, k .$$

Notemos que o modelo (2.21) representa uma estrutura saturada; logo há interesse em considerar casos especiais que reflitam uma estrutura simplificada nas relações entre as variáveis envolvidas. Por exemplo, a imposição da condição $\beta_{1i} = \beta_1$, $i = 1, \ldots, k$, leva à homogeneidade da associação entre a variável resposta e a covariável para todos os estratos.

A generalização de (2.21) a uma situação com dois ou mais fatores de confundimento torna-se inadequada no sentido em que não detalha os efeitos desses fatores quando tomados individualmente. Por este motivo, torna-se mais eficiente partir do modelo logístico (2.7) com a inclusão dos efeitos principais dos fatores de confundimento e das suas interações com as covariáveis principais, julgadas importantes, e tentar a sua simplificação. Por exemplo, no caso particular considerado, modelado segundo (2.20), a ausência de interação entre x_2, tomado como fator de confundimento, e x_1 ($\beta_3 = 0$) corresponde à homogeneidade das razões das chances para os vários níveis de x_2 - situação já englobada na expressão (1.23).

2.4. Estimação no Modelo Logístico.

Uma vez apresentada a interpretação dos parâmetros de alguns modelos logísticos partimos agora para a sua estimação no caso do modelo geral, i.e.,

$$\pi(\mathbf{x}_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}} , \quad i = 1, \ldots, k , \tag{2.22}$$

onde $\eta_i = \mathbf{x}_i^T\beta = \sum_{j=0}^{p-1} x_{ij}\beta_j$ encontra-se definido detalhadamente em (2.10).

Para o modelo probabilístico (2.9) temos a seguinte função log-verossimilhança

$$L(\boldsymbol{\pi}) = \sum_{i=1}^{k}[y_i ln(\pi_i) + (n_i - y_i)ln(1-\pi_i)] . \tag{2.23}$$

Como $\pi_i \equiv \pi(\mathbf{x}_i)$, dado em (2.22), é função de $\boldsymbol{\beta}$, reparametrizando (2.23) obtemos

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{k} y_i \mathbf{x}_i^T\boldsymbol{\beta} - \left[\sum_{i=1}^{k} n_i ln\left(1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}\right)\right] . \tag{2.24}$$

A função *score*, $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, tem como componente j , $j = 0, \ldots, p-1$,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{k} y_i x_{ij} - \sum_{i=1}^{k} n_i x_{ij} \left(\frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right) ,$$

e em forma matricial tem a seguinte expressão

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \boldsymbol{\mu}(\boldsymbol{\beta}) , \tag{2.25}$$

onde $\mathbf{X}$ é a matriz de planejamento e $\boldsymbol{\mu}(\boldsymbol{\beta}) = (\mu_1, \ldots, \mu_k)^T$ com $\mu_i = n_i \pi(\mathbf{x}_i)$, $i = 1, \ldots, k$.

Por conseguinte, o estimador de **MV** de $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, é obtido da solução do sistema de equações não lineares

$$\mathbf{X}^T [\mathbf{y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})] = \mathbf{0} . \tag{2.26}$$

Como, em geral, o sistema (2.26) não tem uma solução analítica há necessidade de um procedimento iterativo. O método de Newton-Raphson tem base no seguinte algoritmo:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(t)}) \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}^{(t)}) \mathbf{y}^* , \quad t = 0, 1, \ldots , \tag{2.27}$$

onde $\mathbf{W}(\boldsymbol{\beta}) = diag\{n_i \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)],\ i = 1, \ldots, k\}$ e $\mathbf{y}^* = (y_1^* \ldots, y_k^*)^T$ com componentes $y_i^* = \{y_i - \mu_i\} / \{n_i \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]\}$.

O processo (2.27) pode ter início com $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ $[\pi(\mathbf{x}_i) = 0,5]$ prosseguindo até que as diferenças das estimativas de duas iterações sucessivas sejam pequenas. Condições suficientes para a existência e unicidade de $\boldsymbol{\beta}$ e considerações sobre a convergência do processo podem ser vistas em McCullagh and Nelder (1989, ch.4).

Observemos que a matriz Hessiana da função (2.24) é

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X} ,$$

logo sendo identificada com o simétrico da matriz de informação de Fisher

$$\mathbf{I}(\boldsymbol{\beta}) \equiv E\left[-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = \mathbf{X}^T \mathbf{W}(\boldsymbol{\beta}) \mathbf{X} . \tag{2.28}$$

A inversa de $\mathbf{I}(\boldsymbol{\beta})$ é a matriz de covariância assintótica dos estimadores de **MV** dos parâmetros de regressão logística, cuja estimativa de **MV** será dada por $[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1}$.

Quando $min(n_1, \ldots, n_k) \to +\infty$, para k fixo, ou quando $k \to +\infty$ e os n_i's são fixos, podemos dizer que $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tem distribuição assintótica Normal p-variada com vetor média nulo e matriz de covariância estimada por $[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1}$. Do exposto podemos construir intervalos de $100(1-\alpha)\%$ de confiança para β_j através da expressão

$$\widehat{\beta}_j \pm t_{(k-p,1-\frac{\alpha}{2})} \sqrt{I^{jj}(\widehat{\boldsymbol{\beta}})}\ , \tag{2.29}$$

onde $I^{jj}(\widehat{\boldsymbol{\beta}})$, $j = 1, \ldots, p$ são os elementos da diagonal principal de $[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1}$ e $t_{(k-p,1-\frac{\alpha}{2})}$ é o percentil $100(1-\frac{\alpha}{2})\%$ da distribuição t-Student com $k-p$ graus de liberdade. Logo, levando em conta a relação entre os coeficientes de regressão e as razões de chances, exposta na Seção 2.3, obtemos sem dificuldade intervalos de confiança para estes parâmetros de interesse. Por exemplo, no modelo logístico simples com covariável binária, o intervalo de confiança $100(1-\alpha)\%$ para a razão de chances (2.12) é obtido de (2.29) via aplicação da função exponencial aos seus limites de confiança.

A estimação dos parâmetros do modelo logístico não tem que ser restrita ao método de máxima verossimilhança. Existem outras abordagens ao problema das quais citamos o método de mínimos quadrados ponderados (não iterativo), introduzido por Grizzle et al. (1969) e o método da análise discriminante devido a Cornfield (1962) - para uma comparação entre este último método e o método da **MV** (vide Efron (1975), Press and Wilson (1978) e Hosmer et al. (1983)).

Notemos que os estimadores de **MV** são frequentemente calculados usando um algoritmo de mínimos quadrados iterativamente reponderados, sendo também considerados estimadores de "mínimos quadrados". A abordagem sugerida por Grizzle et al. usa somente uma iteração no processo.

2.5. Testes de Hipóteses no Modelo Logístico.

Uma vez considerado um modelo logístico para a explicação de dados binários o passo seguinte será a análise da sua adequação e simplificação. Este objetivo será concretizado através da construção de testes de ajustamento do modelo e testes sobre os seus parâmetros.

Se o vetor de parâmetros do modelo, $\boldsymbol{\beta}$, tiver dimensão igual ao número de observações o modelo é dito saturado, ajusta-se completamente aos dados, sendo os valores preditos os próprios valores observados. Assim, podemos formar um teste de ajustamento de um modelo logístico não saturado se o compararmos com o modelo saturado. Esta comparação pode ser feita através da estatística da razão de verossimilhanças de Wilks, cuja expressão é

$$2\sum_{i=1}^{k}\left[y_i ln\left(\frac{y_i}{\widehat{\mu}_i}\right)+(n_i-y_i)ln\left(\frac{n_i-y_i}{n_i-\widehat{\mu}_i}\right)\right]. \tag{2.30}$$

onde $\widehat{\mu}_i = n_i\widehat{\pi}(\mathbf{x}_i) = n_i(1+e^{-\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}})^{-1}$, $i=1,\ldots,k$, e $\widehat{\boldsymbol{\beta}}$ é o estimador de **MV** do vetor de parâmetros do modelo logístico referido. Quando temos grandes amostras, no mesmo sentido apresentado na seção anterior, a estatística (2.30) tem, sob a hipótese de validade do modelo logístico, aproximadamente distribuição Qui-Quadrado com $k-p$ graus de liberdade, onde p é a quantidade de parâmetros. A estatística (2.30) na literatura de modelos lineares generalizados é chamada *deviance*, e será denotada por $D(\mathbf{y},\widehat{\boldsymbol{\mu}})$, onde $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1,\ldots,\widehat{\mu}_k)^T$.

Outra estatística para o teste de ajustamento do modelo logístico é a estatística do Qui-Quadrado de Pearson, dada por

$$X_P^2 = \sum_{i=1}^{k}\left[\frac{y_i-n_i\widehat{\pi}(\mathbf{x}_i)}{n_i\widehat{\pi}(\mathbf{x}_i)}\right]^2, \tag{2.31}$$

que sob a hipótese nula é equivalente assintóticamente a $D(\mathbf{y},\widehat{\boldsymbol{\mu}})$. Para os problemas derivados de k poder aumentar com n, vide Hosmer and Lemeshow (1989, ch.5).

Consideremos agora os testes de significância dos parâmetros do modelo. Para isso, seja $\boldsymbol{\beta} = (\boldsymbol{\beta}_q^T,\boldsymbol{\beta}_{p-q}^T)^T$, $q<p$, uma partição do vetor de parâmetros, e consideremos a hipótese nula definida por $H_0 : \boldsymbol{\beta}_q = \boldsymbol{\beta}_q^{(0)}$. Um caso particularmente importante desta hipótese é $\boldsymbol{\beta}_q^{(0)} = \mathbf{0}$, cuja validade proporciona uma simplificação do modelo em questão.

A hipótese H_0 pode ser testada através de várias estatísticas de teste. Uma delas é a estatística da razão de verossimilhanças de Wilks, dada por

$$E_{RV} = -2[L(\boldsymbol{\beta}_q^{(0)}, \widehat{\boldsymbol{\beta}}_{p-q}^{(0)}) - L(\widehat{\boldsymbol{\beta}}_q, \widehat{\boldsymbol{\beta}}_{p-q})] , \tag{2.32}$$

onde $L(\bullet)$ é a função log-verossimilhança, expressa em (2.24), e $\widehat{\boldsymbol{\beta}}_{p-q}^{(0)}$ e $(\widehat{\boldsymbol{\beta}}_q, \widehat{\boldsymbol{\beta}}_{p-q})$ são os estimadores de **MV** de $\boldsymbol{\beta}_{p-q}$, sob H_0, e de $\boldsymbol{\beta}$, respectivamente. Uma expressão equivalente da estatística (2.32) é dada através da diferença entre as *deviances*, i.e.,

$$D_{p-q}(\mathbf{y}, \widehat{\boldsymbol{\mu}}^{(0)}) - D_p(\mathbf{y}, \widehat{\boldsymbol{\mu}}) , \tag{2.33}$$

onde $\widehat{\boldsymbol{\mu}}^{(0)}$ e $\widehat{\boldsymbol{\mu}}$ são os estimadores de **MV** de $\boldsymbol{\mu}$, sob H_0, e $\boldsymbol{\mu}$, respectivamente.

Outra estatística para o teste de H_0, baseada no vetor das q restrições impostas por H_0 a $\boldsymbol{\beta}$, $\boldsymbol{\beta}_q - \boldsymbol{\beta}_q^{(0)} = \mathbf{0}$, é a chamada estatística de Wald

$$E_W = (\widehat{\boldsymbol{\beta}}_q - \widehat{\boldsymbol{\beta}}_q^{(0)})^T [\mathbf{I}_{qq}(\widehat{\boldsymbol{\beta}})]^{-1} (\widehat{\boldsymbol{\beta}}_q - \widehat{\boldsymbol{\beta}}_q^{(0)}) , \tag{2.34}$$

onde $\mathbf{I}_{qq}(\boldsymbol{\beta})$ é a sub-matriz de $\mathbf{I}(\boldsymbol{\beta})$ associada do sub-vetor $\widehat{\boldsymbol{\beta}}_q$ de $\widehat{\boldsymbol{\beta}}$. Caso o interesse esteja em testar a nulidade de somente um parâmetro do modelo logístico, $\beta_j = 0$, $j = 1, \ldots, p$, a estatística (2.34) toma a forma

$$E_W = \frac{\widehat{\beta}_j^2}{I^{jj}(\widehat{\boldsymbol{\beta}})} , \tag{2.35}$$

onde $I^{jj}(\widehat{\boldsymbol{\beta}})$ foi definido em (2.29).

Uma terceira estatística para testar H_0 está ligada à função *score* $\mathbf{U}(\boldsymbol{\beta})$. Por definição de $\widehat{\boldsymbol{\beta}}$, $\mathbf{U}(\widehat{\boldsymbol{\beta}}) = 0$ e, sob a validade de H_0, o estimador restrito $\widehat{\boldsymbol{\beta}}^{(0)} = (\boldsymbol{\beta}_q^{(0)T}, \widehat{\boldsymbol{\beta}}_{p-q}^{(0)T})^T$ deve estar próximo de $\widehat{\boldsymbol{\beta}}$. Deste modo, valores pequenos de $\mathbf{U}(\widehat{\boldsymbol{\beta}}^{(0)})$ sugerem a consistência de H_0. Uma forma de medir a distância entre $\mathbf{U}(\widehat{\boldsymbol{\beta}}^{(0)})$ e o vetor nulo é através da forma quadrática

$$E_S = \mathbf{U}(\widehat{\boldsymbol{\beta}}^{(0)})^T [\mathbf{I}(\widehat{\boldsymbol{\beta}}^{(0)})]^{-1} \mathbf{U}(\widehat{\boldsymbol{\beta}}^{(0)}) , \tag{2.36}$$

denominada estatística *score* eficiente de Rao. Este teste, chamado teste de *score* ou teste dos multiplicadores de Lagrange, torna-se atrativo pelo fato de exigir apenas os estimadores

de **MV** de β, sob H_0, o que só por si simplifica a parte computacional do teste. Quando, no contexto do modelo (2.21), queremos testar $H_0 : \beta_{1j} = 0$ através do teste de *score*, a expressão (2.36) corresponde à estatística de Mantel-Haenszel, apresentada em (1.30).

As três estatísticas de teste E_{RV}, E_W e E_S mencionadas são todas assintoticamente equivalentes, sendo a distribuição aproximada, para amostras grandes e sob H_0, à distribuição Qui-Quadrado com q graus de liberdade. A sua comparação em termos assintóticos é discutida particularmente em Cox and Hinkley (1974, sec.9.3), Rao (1973, sec.6e) e Buse (1982).

2.6. Seleção de Covariáveis.

Uma vez descrita a estimação e a construção de testes sobre os parâmetros de um determinado modelo logístico, interessa saber como ligar esses procedimentos inferenciais com vista a encontrar o modelo que "melhor" descreva os dados, i.e., um modelo reduzido que inclua as covariáveis mais importantes para a explicação das probabilidades de sucesso $\pi(\mathbf{x})$.

Este problema pode ser essencialmente resolvido pelos métodos de seleção de variáveis usados em análise de regressão. Contudo, a questão de interpretação do modelo é aqui destacada, implicando particularmente que a inclusão de certas interações impõem a inclusão dos seus efeitos associados de ordem inferior, na ótica do princípio hierárquico.

Outro aspecto que interessa chamar atenção é que o processo de eliminação e (ou) inclusão de covariáveis não deve ser aplicado de uma forma mecânica e sim conjugado com o bom senso. Por exemplo as variáveis consideradas "biologicamente" importantes não devem ser deixadas de mão pela sua possível falta de significância estatística.

Concentraremos aqui em dois métodos que são frequentemente usados na seleção de modelos logísticos. Outros métodos podem ser encontrados em Hosmer and Lemeshow (1989, ch.3) e Cordeiro e Paula (1989, cap.4).

2.6.1. Seleção Usual.

Há dois tipos de procedimentos para a investigação do "melhor" modelo : seleção *forward* ou eliminação *backward*. O primeiro procedimento consiste em partir do modelo mais simples em direção aos modelos mais complexos enquanto o segundo segue no sentido contrário.

Dado o menor custo computacional do primeiro procedimento, utilizaremos a idéia básica da seleção *forward* para enunciarmos um algoritmo de busca do "melhor" modelo. Este será formado de etapas, onde a etapa 1 diz respeito à seleção dos efeitos principais do modelo, a etapa 2 à seleção das interações de primeira ordem das covariáveis selecionadas na etapa 1 e assim sucessivamente em relação às demais ordens de interações.

Nos preocuparemos em descrever somente a etapa 1, já que as restantes se processam de forma completamente análoga. Assim, partindo do modelo mais simples, que possui somente o intercepto como parâmetro, o modelo inicial, seguiremos os seguintes passos na etapa 1 :

(1) Cada covariável forma um modelo logístico simples que comparamos com o modelo inicial através do teste da razão de verossimilhanças, dado em (2.32). O nível descritivo P desses testes determina a inclusão ou não da respectiva covariável conforme P seja, respectivamente, inferior ou não a 25% - critério de Mickey and Greeland (1989);

(2) Consideramos o modelo com todas as covariáveis selecionadas no passo (1) e, com base nele, testamos a nulidade de cada um dos seus coeficientes, através da estatística de Wald, expressa em (2.35). A exclusão ou não de cada covariável é determinada pelo nível descritivo do respectivo teste de Wald : se for superior à 5% ou 10% (níveis convêncionais) a variável é excluída;

(3) Comparamos através da razão de verossimilhanças o ajustamento do modelo sem as covariáveis excluídas no passo anterior com o modelo inicial do passo (2), e de acordo com o resultado, tomamos um ou outro como o modelo base para as etapas seguintes.

Como já adiantamos, estes passos são agora essencialmente repetidos nas etapas seguintes, envolvendo as interações de primeira ordem, segunda ordem, e assim sucessivamente, associadas às covariáveis presentes no modelo final da etapa 1.

2.6.2. Seleção Stepwise.

Outro método de seleção de covariáveis "importantes" para explicar a variável resposta é o método *stepwise*. Este método baseia-se num algoritmo misto de seleção *forward* e de eliminação *backward*, que inclui ou exclui as covariáveis conforme a sua importância de acordo com algum critério.

O grau de importância de uma covariável é medido pelo nível descritivo do teste da razão de verossimilhanças entre os modelos que a incluem e a excluem. Quanto menor for este nível tanto mais importante será considerada a covariável. Como a covariável mais importante por este critério não é necessariamente significativa do ponto de vista estatístico, há que impor um limite superior P_E (os valores usuais estão no intervalo $[0,15 - 0,25]$) para estes níveis descritivos, a fim de atrair candidatos importantes em princípio à entrada.

Dado que a presença de várias covariáveis num modelo pode tornar uma ou outra dispensáveis, faremos a verificação da importância da presença de cada covariável confrontando o seu respectivo nível descritivo com um limite inferior P_S, superior a P_E. As covariáveis com um nível descritivo associado superior a P_S serão assim candidatas à remoção.

Descreveremos agora uma variante deste algoritmo usada por Hosmer and Lemeshow (1989, ch.3). A etapa inicial começa com o ajustamento do modelo só com intercepto e é constituída pelos seguintes passos :

(1) Construímos testes da razão de verossimilhanças entre o modelo inicial e os modelos logísticos simples formados com cada uma das covariáveis do estudo. O mínimo dos níveis descritivos associados a cada teste será comparado com $P_E = 0,15$. Se P_E for maior incluímos a covariável referente àquele nível mínimo e passamos ao passo seguinte; caso contrário, paramos a seleção e selecionamos o último modelo;

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

(3) Comparamos o ajuste do modelo logístico contendo as covariáveis selecionadas nos passos anteriores com os modelos que dele resultam por exclusão individual de cada uma das covariáveis. Se o máximo dos níveis descritivos destes testes da razão de verossimilhanças for menor do que $P_S = 0,20$, a covariável associada a este nível permanece no modelo. Caso contrário, ela é removida. Em qualquer circunstância, o algoritmo segue para o passo seguinte.

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

(5) Uma vez selecionadas as covariáveis "importantes", i.e., os seus efeitos principais, na etapa 1, damos entrada na etapa 2 através do passo (1) com o objetivo de selecionar as interações que envolvem aquelas covariáveis, e assim por diante.

Uma desvantagem deste procedimento é a de exigir as estimativas de **MV** em cada passo, o que encarece o trabalho computacional, particularmente em grandes amostras. Alguns autores optaram por aproximações para este método de seleção. O pacote científico **BMDP** (Dixon, 1987) usa aproximações lineares nos testes da razão de verossimilhanças. Peduzzi et al. (1980) apresentam uma variante deste método baseada no uso da estatística de Wald para testar as covariáveis não inclusas no modelo.

2.7. Análise de Resíduos.

Um modelo logístico pode ser avaliado de várias maneiras. Por exemplo, através das medidas que comparam os valores observados com os valores preditos pelo modelo, já apresentadas na Seção 2.5. Aqui faremos essa avaliação através do estudo dos resíduos, uma técnica de diagnóstico bastante usada na regressão linear.

A diferença entre o valor observado da variável resposta (y) e o valor ajustado pelo modelo corrente ($\widehat{\mu}$) é chamada de resíduo. O resíduo da i-ésima configuração de covariáveis é

$$y_i - \widehat{\mu}_i \;=\; y_i - n_i\widehat{\pi}(\mathbf{x}_i) \;=\; y_i - n_i\left[\frac{e^{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i^T\widehat{\boldsymbol{\beta}}}}\right], \tag{2.37}$$

onde $\widehat{\boldsymbol{\beta}}$ é a estimativa de **MV** do vetor $\boldsymbol{\beta}$ no modelo corrente, $i = 1, \ldots, k$.

O resíduo mais comum é o resíduo de Pearson

$$r_i \;=\; \frac{y_i - n_i\widehat{\pi}(\mathbf{x}_i)}{\sqrt{n_i\widehat{\pi}(\mathbf{x}_i)[1-\widehat{\pi}(\mathbf{x}_i)]}}\,, \quad i = 1, \ldots, k\,, \tag{2.38}$$

que é parte integrante da estatística (2.31).

Quando os n_i's são pequenos ou $\pi(\mathbf{x}_i)$ está próximo de 0 ou 1 a aproximação Normal para os resíduos de Pearson torna-se precária. Resíduos mais apropriados para esta situação são baseados na *deviance*, apresentada em (2.30), e definidos por

$$d_i = \pm\left[2y_i ln\left(\frac{y_i}{\widehat{\mu}_i}\right) + 2(n_i - y_i) ln\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right)\right]^{\frac{1}{2}}, \quad i = 1, \ldots, k, \tag{2.39}$$

onde o sinal é o mesmo de $(y_i - \widehat{\mu}_i)$. Para as situações particulares $y_i = 0$ e $y_i = n_i$ os resíduos (2.39), chamados de resíduos *deviance*, são respectivamente,

$$d_i = -\sqrt{2n_i \mid ln[1 - \widehat{\pi}(\mathbf{x}_i)] \mid} \quad e \quad d_i = \sqrt{2n_i \mid ln[\widehat{\pi}(\mathbf{x}_i)] \mid}.$$

De acordo com Cox and Snell (1968) pode-se obter uma outra padronização para os resíduos, que leva em consideração a variabilidade de $\widehat{\pi}(\mathbf{x}_i)$. Esta idéia pode ser concretizada através da divisão dos resíduos (2.38) e (2.39) por $(1 - \widehat{h}_{ii})^{\frac{1}{2}}$, onde $\widehat{h}_{ii}$ é a estimativa do i-ésimo elemento da diagonal principal da matriz de projeção

$$\mathbf{H}(\boldsymbol{\beta}) = \mathbf{W}(\boldsymbol{\beta})^{\frac{1}{2}}\mathbf{X}[\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}(\boldsymbol{\beta})^{\frac{1}{2}}, \tag{2.40}$$

onde as matrizes apresentadas estão definidas no processo (2.27).

Além das medidas expostas para análise dos resíduos, existem os métodos gráficos. O gráfico de resíduos de r_i ou d_i versus $\widehat{\mu}_i$ serve para detectar pontos aberrantes (*outliers*) no ajustamento do modelo. Landwehr et al. (1984) apresentam outros métodos gráficos para o modelo logístico. Williams (1987) através de gráficos estuda a adequação de suposições distribucionais do modelo.

Outras técnicas de diagnóstico para o modelo logístico podem ser encontradas com detalhes em Pregibon (1981), Pierce and Schafer (1986), Williams (1987) e Hosmer and Lemeshow (1989, ch.5).

A falta de ajustamento do modelo logístico pode ser atribuída a várias causas. Uma delas tem a ver com a inadequação do modelo Binomial na explicação da variabilidade da variável resposta. O caso de *Overdispersion* é um exemplo dessa inadequação. O tratamento deste problema pode ser encontrado particularmente em Williams (1982), Anderson (1982), McCullagh and Nelder (1989, pg.124-128) e Follmann and Lambert (1989).

2.8. Tópicos Especiais.

Esta seção tem como objetivo apresentar questões específicas relativas aos modelos logísticos para dados binários que temos vindo a considerar. No entanto, queremos chamar a atenção que tais modelos podem ser generalizados para dados politômicos (i.e., com uma variável resposta com distribuição Multinomial) usando vários tipos possíveis de *logits*, associados ou não ao caráter nominal ou ordinal das categorias da variável resposta. Dado que o tratamento destes modelos sai fora do âmbito deste trabalho remetemos o leitor eventualmente interessado para McCullagh and Nelder (1989, ch.5), Hosmer and Lemeshow (1989, ch.8) e Agresti (1990, ch.9), entre outros.

2.8.1. Modelo Logístico para Estudos Retrospectivos.

Em várias situações, os dados sobre a variável resposta binária Y e o vetor de covariáveis $\mathbf{x}$ não são obtidos de maneira prospectiva. Em vez desse planejamento, um conjunto de n_1 casos (indivíduos com $Y = 1$) e $n_0 = n - n_1$ controles (indivíduos com $Y = 0$) é selecionado e posteriormente identificado segundo os valores de $\mathbf{x}$. Este tipo de estudo, chamado de retrospectivo, é muitas vezes motivado por questões econômicas ligadas ao custo e a duração do experimento, particularmente quando Y é incidência de uma doença rara.

Os dados sobre casos e controles resultam assim de uma amostragem direta de um modelo para $P(\mathbf{x} \mid y)$, $y = 0, 1$, contrariamente aos dados prospectivos que estão associados ao modelo $\pi(\mathbf{x}) = P(y|\, \mathbf{x})$. A sua análise pode ser processada de modo análogo àquele que foi visto para os dados de um estudo de seguimento, i.e., através da especificação e ajuste de um modelo estatístico para $P(\mathbf{x} \mid y)$. Este procedimento pode tornar-se complicado quando $\mathbf{x}$ envolve um grande número de variáveis explicativas, particularmente contínuas.

Uma abordagem alternativa consiste em especificar um modelo para $P(y \mid \mathbf{x})$ de modo a induzir um modelo para $P(\mathbf{x} \mid y)$, que é então usado para as inferências de interesse. Neste situação, o uso de um modelo prospectivo logístico revela-se conveniente pelo fato da metodologia de **MV** para os dados retrospectivos envolver ainda um modelo numa forma logística. Descreveremos, em seguida, uma justificativa em traços gerais desta afirmação, baseada no argumento de Farewell (1979).

Para isso, suponhamos então que $P(Y = 1 \mid \mathbf{x})$ é representada pelo modelo logístico

$$\pi(\mathbf{x}) = \frac{e^{\mathbf{x}^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T \boldsymbol{\beta}}},$$

onde $\mathbf{x} = (x_0, x_1, \ldots, x_{p-1})^T$ com $x_0 \equiv 1$ é um valor genérico para o vetor de $p-1$ variáveis explicativas e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$.

A função de verossimilhança para a amostra de n_1 casos e $n_0 = n - n_1$ controles é definida por

$$L(\boldsymbol{\beta}) = \prod_{i=0}^{1} \prod_{j=1}^{n_i} P(\mathbf{x}_{ij} \mid y_i) , \tag{2.41}$$

onde $y_i = i$ e $\mathbf{x}_{ij}$ representa os n_i valores de $\mathbf{x}$ observados no grupo i, $i = 0, 1$.

Nos restringiremos ao caso em que $\mathbf{x}$ é discreto. A função de probabilidade $P(\mathbf{x} \mid y)$ para um indivíduo da população global com resposta y coincide com a mesma probabilidade para um indivíduo da amostra se a seleção desta não depender de $\mathbf{x}$. Isto é, se z representar a variável indicadora da seleção amostral para qualquer indivíduo, com distribuição condicional em cada grupo independente de $\mathbf{x}$ e denotada por

$$P(z = 1 \mid y = 1) = \phi_1 \quad e \quad P(z = 1 \mid y = 0) = \phi_0 ,$$

temos

$$P(\mathbf{x} \mid y; z = 1) = \frac{P(\mathbf{x} \mid y) P(z = 1 \mid y; \mathbf{x})}{P(z = 1 \mid y)} = P(\mathbf{x} \mid y) . \tag{2.42}$$

Por outro lado, para qualquer indivíduo selecionado com vetor de covariáveis $\mathbf{x}$.

$$P(y \mid \mathbf{x}; z = 1) = \frac{P(y \mid \mathbf{x}) P(z = 1 \mid y)}{\sum\limits_{y=0,1} P(y \mid \mathbf{x}) P(z = 1 \mid y)} ,$$

e assim

$$\pi^*(\mathbf{x}) \equiv P(y = 1 \mid \mathbf{x}; z = 1) = \frac{\pi(\mathbf{x})\phi_1}{\pi(\mathbf{x})\phi_1 + [1 - \pi(\mathbf{x})]\phi_0} = \frac{\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right]\frac{\phi_1}{\phi_0}}{1 + \left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right]\frac{\phi_1}{\phi_0}} .$$

ou melhor,

$$\pi^*(\mathbf{x}) = \frac{e^{\mathbf{x}^T \boldsymbol{\alpha}}}{1 + e^{\mathbf{x}^T \boldsymbol{\alpha}}} , \tag{2.43}$$

onde $\boldsymbol{\alpha} = (\beta_0^*, \beta_1, \ldots, \beta_{p-1})^T$ com $\beta_0^* = \beta_0 + ln\frac{\phi_1}{\phi_0}$.

Tendo em conta (2.42), a aplicação do teorema de Bayes permite reescrever (2.41) como

$$L(\boldsymbol{\beta};\phi_1;\phi_0) = \prod_{i=0}^{1}\prod_{j=1}^{n_i}\left[\frac{P(y_i \mid \mathbf{x}_{ij}; z=1)P(\mathbf{x}_{ij} \mid z=1)}{P(y_i \mid z=1)}\right], \tag{2.44}$$

onde, por definição, $P(y_i \mid z=1) = \frac{n_i}{n}$, $i = 0, 1$. A estimação por MV dos parâmetros é processada assim por maximização de (2.44) sujeita à restrição de

$$\sum_{\{\mathbf{x}_{1j}\}} \pi^*(\mathbf{x}_{1j})P(\mathbf{x}_{1j} \mid z=1) = \frac{n_1}{n}\ .$$

Anderson (1972) e Prentice and Pyke (1979) mostram com argumentos distintos que o estimador de MV de $\boldsymbol{\alpha}$ é aquele que resulta da maximização irrestrita de

$$L^*(\boldsymbol{\alpha}) = \prod_{i=0}^{1}\prod_{j=1}^{n_i} P(y_i \mid \mathbf{x}_{ij}; z=1) = \prod_{j=1}^{n_1}\pi^*(\mathbf{x}_{1j})\prod_{j=1}^{n_0}[1-\pi^*(\mathbf{x}_{1j})] \tag{2.45}$$

e que a matriz de covariância assintótica do estimador de $(\beta_1, \ldots, \beta_{p-1})^T$ é idêntica à que resulta de (2.45).

A análise da expressão (2.45) revela que a função de verossimilhança relevante é a verossimilhança de um modelo logístico, definido em (2.43). Assim, as inferências sobre $\boldsymbol{\alpha}$, que inclui o parâmetro de interesse $(\beta_1, \ldots, \beta_{p-1})^T$ em estudos de caso-controle podem ser processadas de modo idêntico aquele que foi discutido para dados prospectivos. Porém, devemos ter cuidado que, sem informação adicional, não é possível fazer inferências sobre o parâmetro β_0 de $\boldsymbol{\beta}$ e os parâmetros perturbadores ϕ_1 e ϕ_0 devido à sua não identificabilidade no modelo (2.44), visto que $\boldsymbol{\alpha} = \boldsymbol{\beta} + \boldsymbol{\phi}$, onde $\boldsymbol{\phi} = (ln\frac{\phi_1}{\phi_0}, \mathbf{0}^T_{(p-1)})^T$.

Para mais detalhes sobre a análise de dados retrospectivos via regressão logística, incluindo a consideração de covariáveis contínuas, vide Anderson (1972) e Prentice and Pyke (1979).

Outra abordagem para a função de verossimilhança (2.44) é construirmos uma distribuição condicional para eliminação dos parâmetros perturbadores com base nas estatísticas suficientes do modelo, da mesma forma que referimos no Capítulo 1 para tabelas de contingência. A distribuição resultante serviria para inferências exatas dos parâmetros β_j, $j = 1, \ldots, p-1$. Esse modelo é chamado de modelo logístico condicional, e pode ser visto com detalhes em Breslow and Day (1980, ch.7), Cox and Snell (1989), Prentice and Breslow (1978), Farewell (1979) e Prentice and Pyke (1979).

Um caso importante para o uso do modelo logístico condicional é o de dados retrospectivos emparelhados. Vide, e.g., Breslow and Day (1980, ch.7), Kleinbaum et al. (1982) e Hosmer and Lemeshow (1989, ch.7).

2.8.2. Modelos Lineares Generalizados.

O modelo logístico e outros modelos para dados binários são casos especiais dos Modelos Lineares Generalizados (**MLG**'s) introduzidos por Nelder e Wedderburn (1972). A especificação desses modelos faz-se via três componentes :

i) A componente aleatória, que identifica a distribuição de probabilidade da variável resposta, consiste de observações independentes de membros da família exponencial;

ii) A componente sistemática, que especifica uma função linear das variáveis explicativas que é usada como um preditor;

iii) A função de ligação $g(\bullet)$, onde g é diferenciável e monótona, descrevendo a relação funcional entre a componente sistemática e o valor esperado da componente aleatória.

Para uma longa exemplificação de um **MLG** vide McCullagh and Nelder (1989). Relativamente ao modelo logístico, a distribuição (2.9) pode ser representada por

$$P(y_i;\pi_i) \;=\; (1-\pi_i)^{n_i} ln\binom{n_i}{y_i} exp\left[y_i ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] \;, \quad i=1,\ldots,k \;. \tag{2.46}$$

chamando atenção que cada observação da variável resposta em cada configuração pertence à família exponencial. Por outro lado, a função de ligação é o *logit* de π_i, $i=1,\ldots,k$, por sinal o parâmetro natural dessa sub-família, estabelecendo a relação com a componente sistemática, dada em (2.10). Deste modo, fica claro como o modelo logístico se encaixa na definição de um **MLG**.

Ainda em relação ao modelo logístico é sabido que o valor esperado da componente aleatória referente a cada observação da i-ésima configuração é $\mu_i = n_i\pi_i$; daí o uso frequente de $g(\pi_i)$ em vez de $g(\mu_i)$ para a função de ligação ao preditor η_i. A função *logit*, dada em (2.11), é aqui a função de ligação canônica, pois transforma μ_i (ou π_i) no parâmetro natural. Outra função de ligação comumente usada em outros contextos é a função $g(\mu_i) = \mu_i$, conhecida como função de ligação identidade. Porém, no modelo

logístico ela pode produzir estimativas de π_i fora do intervalo $[0-1]$, como já referimos no início deste capítulo.

A Tabela 2.1 descreve alguns exemplos de **MLG**'s.

Tabela 2.1 : Alguns Modelos Lineares Generalizados.

Componente Aleatória	Componente Sistemática	Ligação	Modelo
Normal	contínuas	identidade	Regressão Linear
Normal	categorizadas	identidade	Análise de Variância
Normal	mistas	identidade	Análise de Covariância
Binomial	mistas	*logit*	Regressão Logística
Poisson	mistas	logaritmica	Log-linear

O modelo logístico não é o único **MLG** capaz de descrever dados binários. Três outros modelos apresentam-se na Tabela 2.2. O modelo *probit* ficou popularizado em experimentos toxicológicos (Bliss, 1935) - para maiores detalhes vide Finney (1971). O modelo complementar log-log (*extremit*) tem sido usado em análise de sobrevivência (vide Lee (1980) e Cox and Oakes (1984)) - o uso nesta do modelo logístico pode ser visto nos recentes artigos de Efron (1988) e Abbott (1985).

Tabela 2.2 : Modelos para dados binários a partir de função de ligação.

Função de Ligação	Modelo
$g_1(\pi) = ln[\pi/(1-\pi)]$	Logístico
$g_2(\pi) = \Phi^{-1}(\pi)^{(*)}$	*Probit*
$g_3(\pi) = ln[-ln(1-\pi)]$	Complementar Log-log
$g_4(\pi) = ln[-ln(\pi)]$	Log-log

$^{(*)}$ $\Phi(\bullet)$: função de distribuição da Normal padrão.

De acordo com McCullagh and Nelder (1989, ch.4) é difícil distinguir as funções *logit* e *probit* no intervalo $0.01 \leq \pi \leq 0.99$; a função complementar log-log está próxima

da função logística para valores pequenos de π e apresenta um crescimento mais lento do que as *logit* e *probit* para valores de π próximos de 1 ; as funções $g_1(\pi)$ e $g_2(\pi)$ são funções simétricas, i.e.,

$$g(\pi) = -g(1-\pi) , \tag{2.47}$$

o mesmo não ocorrendo com relação a $g_3(\pi)$ e $g_4(\pi)$. Além disso, destacam as vantagens do modelo logístico pela facilidade de estimação e interpretação dos parâmetros nos vários tipos de amostragem dos dados.

2.8.3. Distribuição de Tolerância e Dose Letal.

O modelo logístico é frequentemente usado em Toxicologia, onde se pretende geralmente descrever o efeito de um medicamento tóxico na morte dos indivíduos em estudo. Este caso envolve uma covariável contínua e uma variável resposta binária e a relação entre elas é frequentemente denominada de modelo de dose-resposta.

A relação entre a probabilidade de sucesso no modelo de dose-resposta e o preditor linear, $\eta = \beta_0 + \beta_1 x$, é feita por uma função injetora $g(\pi) = \eta$. Notando que π está restrito ao intervalo $(0,1)$ para valores de η em $(-\infty, +\infty)$, é razoável modelarmos π como uma função de distribuição acumulada, ou seja,

$$\pi = g^{-1}(\eta) = F(\eta) = \int_{-\infty}^{\eta} f(w)dw , \tag{2.48}$$

onde $f(\bullet)$ representa uma função de densidade de probabilidade, denominada de **função de tolerância**. Isto porque nos estudos de toxicologia podemos interpretar π como a probabilidade de morte para uma dosagem η.

Todos os quatro modelos para dados binários apresentados na Tabela 2.2 podem ser obtidos da expressão (2.48) através das funções de tolerância Logística, Normal padrão, Gumbel de mínimos e Gumbel de máximos, respectivamente. Deste modo, as probabilidades de sucesso em (2.48) para os modelos *probit*, complementar log-log e log-log são, respectivamente,

$$\pi(\eta) = \Phi(\eta) , \quad \pi(\eta) = 1 - e^{-e^{\eta}} \quad e \quad \pi(\eta) = e^{-e^{\eta}} , \tag{2.49}$$

onde $\Phi(\bullet)$ é a função de distribuição da Normal padrão. Para o modelo logístico a probabilidade de sucesso, dada em (2.4), corresponde à função de distribuição Logística, se

$\beta_1 > 0$. No caso $\beta_1 < 0$, é $1 - \pi(x)$ que representa essa função de distribuição. A função de distribuição logística é definida por

$$F(x) = \frac{e^{\frac{x-\mu}{\tau}}}{1 + e^{\frac{x-\mu}{\tau}}}, \quad -\infty < x < +\infty, \tag{2.50}$$

onde μ e τ são parâmetros de locação e escala, respectivamente. Logo, o preditor linear $\eta = \beta_0 + \beta_1 x$ tem os parâmetros $\beta_0 = -\mu/\tau$ e $\beta_1 = 1/\tau$, $\tau > 0$.

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

$$P = F(\beta_0 + \beta_1 DL_{100P}), \quad 0 < P < 1, \tag{2.51}$$

onde $F(\bullet)$ é uma função de distribuição acumulada.

A dose letal mais comum em Toxicologia é a dose mediana (DL_{50}), embora em certos casos sejam as doses extremas, por exemplo DL_1 ou DL_{99}, o centro das atenções. Convém lembrar que hoje os modelos de dose-resposta estão difundidos em várias áreas do conhecimento logo a covariável pode ser idade, peso, resistência de um material, etc.

Sob o modelo logístico. o estimador de **MV** de DL_{100P} é, pela propriedade invariante,

$$\widehat{DL}_{100P} = \frac{1}{\widehat{\beta}_1}\left[ln\left(\frac{P}{1-P}\right) - \widehat{\beta}_0\right] \equiv d(\widehat{\boldsymbol{\beta}}), \tag{2.52}$$

onde $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^T$ é o estimador de **MV** do parâmetro do modelo.

A variância assintótica de $\widehat{DL}_{100P}$ será construída com base na série de Taylor até primeira ordem de $d(\widehat{\boldsymbol{\beta}})$ em torno de $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$, i.e.,

$$d(\widehat{\boldsymbol{\beta}}) - d(\boldsymbol{\beta}) \simeq \left[\frac{\partial d(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\right]^T (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Assim, fazendo

$$\mathbf{D}(\boldsymbol{\beta}) \equiv \frac{\partial d(\widehat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \left\{-\frac{1}{\beta_1}, \frac{1}{\beta_1^2}\left[\beta_0 - ln\left(\frac{P}{1-P}\right)\right]\right\}^T$$

temos

$$Var_A[d(\widehat{\boldsymbol{\beta}})] \;=\; \mathbf{D}(\boldsymbol{\beta})^T[\mathbf{I}(\boldsymbol{\beta})]^{-1}\mathbf{D}(\boldsymbol{\beta})\;. \tag{2.53}$$

Com base em (2.53), um intervalo de $100(1-\alpha)\%$ de confiança para DL_{100P}, em grandes amostras, é

$$\widehat{DL}_{100p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}_A[d(\widehat{\boldsymbol{\beta}})]}\;, \tag{2.54}$$

onde $\widehat{Var}_A[d(\widehat{\boldsymbol{\beta}})]$ é a variância (2.53), avaliada em $\widehat{\boldsymbol{\beta}}$, e $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})\%$ da distribuição Normal padrão.

2.9. Exemplos.

Nesta seção ilustraremos alguns dos métodos descritos neste capítulo para análise de modelos logísticos. Dois exemplos serão considerados retratando duas situações específicas. O primeiro exemplo aborda dados retrospectivos numa aplicação direta da Subseção 2.8.1, e o segundo trata do modelo de dose-resposta voltado principalmente para a estimação das doses letais .

Exemplo 2.2 : No setor de Anatomia e Patologia do Hospital Heliópolis, São Paulo, entre 1970 e 1982, foi realizado um estudo retrospectivo, cujos dados, apresentados no Apêndice 2, foram obtidos de Paula et al. (1984). O objetivo principal desse estudo era avaliar a associação entre algumas variáveis histológicas e o tipo, maligmo ou benigno, do Processo Infeccioso Pulmonar (PIP).

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

A observação de cada um dos 175 pacientes fez-se através de variáveis histológicas nos fragmentos de tecidos retirados da região pulmonar. Dessas variáveis somente as intensidades de histiócitos-linfócitos e de fibrose-frouxa foram consideradas importantes na discriminação dos dois tipos de PIP. Além destas, o conjunto de covariáveis será formado por dois fatores potenciais de confundimento, sexo e idade. A descrição da codificação das variáveis encontra-se na Tabela 2.3.

Tabela 2.3 : Codificação dos dados do Processo Infeccioso Pulmonar, Paula et al. (1984).

Variável	Nome	Código	Notação do Apêndice 2
y	Tipo de Processo Infeccioso Pulmonar	1=maligno 0=benigno	PIP
x_1	Idade	anos	IDA
x_2	Sexo	1=masculino 0=feminino	SEX
x_3	Intensidade de Histiócitos-Linfócitos	1=alta 0=baixa	HL
x_4	Intensidade de Fibrose Frouxa	1=alta 0=baixa	FF

Como discutimos na Subseção 2.8.1, o modelo logístico (2.43) será ajustado aos dados retrospectivos. Os parâmetros β_j, $j \neq 0$, são estimados diretamente deste modelo, enquanto β_0, se ϕ_1 e ϕ_0 forem conhecidos, pode ser introduzido posteriormente, visto que as razões de chances são invariantes com a introdução deste.

Para seleção de covariáveis que formem o "melhor" modelo logístico usaremos o método de seleção *stepwise*. Os testes de razão de verossimilhanças de Wilks entre modelos com inclusão ou exclusão de variáveis, ou mesmo interações, darão origem aos níveis descritivos apresentados nas tabelas que se seguem. Por exemplo, na etapa 1, seleção dos efeitos principais, o valor observado da estatística do teste que compara o modelo só com intercepto (modelo inicial) com o modelo com inclusão da covariável IDA a este, é $E_{MV} = 236,34 - 190,92 = 45,42$. O uso da distribuição Qui-Quadrado com 1 grau de liberdade produz o nível descritivo $P = 0,000$.

O nível descritivo acima faz parte da Tabela 2.4, onde encontram-se os outros níveis descritivos para inclusão ou exclusão de covariáveis em cada passo de decisão da etapa 1 do método de seleção. O passo 1 inclui a covariável IDA, pois o seu nível descritivo, que é o mínimo neste passo, é inferior a $P_E = 0,20$ (nível padrão para inclusão de variáveis). O passo seguinte nesta etapa inclui a variável HL, e agora com duas variáveis incluídas no modelo serão testadas as exclusões individuais destas variáveis. Os níveis descritivos associados a esses testes encontram-se na Tabela 2.4, na linha de referência do passo 3 e

abaixo da curva em forma de escada. O máximo desses níveis estará identificado por um asterisco e, sendo inferior a $P_S = 0,25$ (nível padrão de exclusão), a variável associada a este nível não é retirada do modelo. Seguindo esta lógica, encontramos os níveis descritivos mínimos em cada passo de decisão como o primeiro elemento acima da curva em "escada". Sendo todos inferiores a P_E concluímos pela entrada no modelo de todas covariáveis. Relativamente às exclusões observamos que os níveis com asterisco são inferiores a P_S, e assim nenhuma das covariáveis sai do modelo. Em resumo, o modelo resultante da etapa 1 da seleção *stepwise* é o modelo com todos os efeitos principais do conjunto das covariáveis.

Tabela 2.4 : Níveis descritivos usados na seleção *stepwise* - Etapa 1.

Passo de decisão	Idade	HL	Sexo	FF
1	0,000	0,000	0,288	0,001
2	0,000	0,000	0,100	0,003
3	0,000	0,000*	0,050	0.124
4	0,000	0,000	0,050*	0,182
5	0,000	0,000	0,050	0,182*

De forma análoga processar-se-á a etapa 2, cujos níveis descritivos para tomada de decisão em cada passo encontram-se na Tabela 2.5. Concluímos então que só três interações de primeira ordem serão incluídas no modelo, e nenhuma delas foi excluída posteriormente. Estas interações são $IDA.HL$, $HL.FF$ e $SEX.FF$.

Tabela 2.5 : Níveis descritivos usados na seleção *stepwise* - Etapa 2.

Passo de decisão	IDA.HL	HL.FF	SEX.FF	IDA.FF	IDA.SEX	HL.SEX
1	0,012	0,014	0,059	0,056	0,663	0,063
2	0,012	0,027	0,060	0,232	0,218	0,099
3	0,023	0,027*	0,012	0.233	0,275	0,176
4	0,028*	0.005	0,012	0,207	0,403	0,791

Na etapa 3 nenhuma interação de segunda ordem foi incluída, já que o mínimo dos níveis descritivos dos testes de inclusão foi inferior a P_E. Assim, o modelo resultante da seleção *stepwise* possui todos os efeitos principais do conjunto de covariáveis e as interações de primeira ordem $IDA.HL$, $HL.FF$ e $SEX.FF$. As estimativas dos parâmetros e os respectivos desvios padrões assintóticos deste modelo encontram-se na Tabela 2.6.

Tabela 2.6 : Estimativas dos parâmetros e desvios padrões associados ao modelo logístico resultante da seleção *stepwise*.

Efeito	Parâmetro	Estimativa	Desvio padrão
constante	β_0	-1,409	0,937
IDA	β_1	0,039	0,017
HL	β_2	-5,521	1,682
SEX	β_3	1,402	0,583
FF	β_4	-1,978	0,887
IDA.HL	β_5	0,062	0,029
HL.FF	β_6	2,908	1,103
SEX.FF	β_7	-3,349	1,476

O teste de ajustamento do modelo logístico selecionado produz para a estatística (2.30) o valor $146,22$ correspondente a um nível descritivo $0,8751$ (calculado de uma distribuição Qui-Quadrado com 167 graus de liberdade). Logo, conclui-se que há adequação do modelo. Outras técnicas de avaliação do modelo encontram-se na Seção 2.7.

Como o interesse principal é estudar a associação entre o tipo de PIP e as variáveis histológicas no conjunto de covariáveis, formaremos as razões de chances para os níveis dessas variáveis.

A razão de chances de um paciente com nível alto de intensidade de HL, em relação ao nível baixo de HL, estar com PIP do tipo maligno é denotada por ψ_{HL}. Supondo que os pacientes tenham o mesmo sexo, idade (x_1) e nível de intensidade de FF (x_4), a razão de chances acima pode ser estimada por

$$\widehat{\psi}_{HL} = exp\{-5,521 + 0,062x_1 + 2,908x_4\} . \qquad (2.55)$$

Da expressão (2.55) podemos concluir que a chance de um PIP maligno é menor para os pacientes com alta intensidade de HL que para os pacientes com baixa intensidade

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

Para ilustramos a aplicação da expressão (2.55), suponhamos que dois pacientes de 60 anos e do mesmo sexo tenham sido submetidos a exames no hospital referido a fim de ser diagnosticado o tipo de PIP. Após os exames, admitamos que se constatou para ambos o nível baixo de intensidade FF, enquanto apenas um apresentou alta intensidade de HL. Deste modo, a chance estimada do paciente, cujo exame não detectou alta intensidade de HL, estar com PIP maligno, em relação ao outro, é $\widehat{\psi}_{HL} = exp(-1,801) = 0,165$.

Um intervalo de $100(1-\alpha)\%$ de confiança para ψ_{HL}, em grandes amostras, é formado pelos limites

$$\widehat{\psi}^{I}_{HL}, \widehat{\psi}^{S}_{HL} = exp\left\{ ln\widehat{\psi}_{HL} \pm z_{1-\frac{\alpha}{2}} . \sqrt{\widehat{Var}_A(ln\widehat{\psi}_{HL})} \right\} , \tag{2.56}$$

onde $\widehat{\psi}_{HL}$ está expresso em (2.55), $z_{1-\frac{\alpha}{2}}$ é o percentil de ordem $100(1-\frac{\alpha}{2})\%$ da distribuição Normal padrão e

$$Var_A(\widehat{\psi}_{HL}) = \sum_{i=1,2,4} x_i Var_A(\widehat{\beta}_i) + \sum_{i \neq j=1,2,4} 2x_i x_j Cov_A(\widehat{\beta}_i, \widehat{\beta}_j) . \quad x_1 \equiv 1 .$$

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

$$\widehat{\psi}_{FF} = exp\{-1,978 - 3,349x_2 + 2,908x_3\} . \tag{2.57}$$

Desta expressão podemos deduzir que a chance de um PIP maligno é menor para os pacientes com alta intensidade de FF que para os pacientes com baixa intensidade de FF, isto entre os homens independentemente do nível de intensidade de HL e para as mulheres com baixa intensidade de HL. Para as mulheres com alta intensidade de HL ocorre o contrário nesta chance. Em ambos níveis de HL a razão de chances ψ_{FF} é maior para as mulheres do que para os homens.

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

$$\widehat{\beta}_0^* = \widehat{\beta}_0 + ln\left(\frac{71/71}{104/270}\right) \Leftrightarrow \widehat{\beta}_0 = -1,409 - (0,954) = -2,363 .$$

Deste modo, ficamos aptos a estimar $\pi(\mathbf{x})$ para qualquer valor de $\mathbf{x}$, como ilustra-se na Tabela 2.7.

Tabela 2.7 : Estimativas de $\pi(\mathbf{x})$ para várias configurações.

Idade	Sexo	HL	FF	$\pi(\mathbf{x})$
51	masculino	alto	alto	0,022
29	feminino	baixo	alto	0,038
62	feminino	alto	baixo	0,159
29	feminino	baixo	baixo	0,224
50	feminino	baixo	baixo	0,395
44	masculino	baixo	baixo	0,677

◇

Exemplo 2.3 : Em Bliss (1935) encontramos uma situação típica para ajuste do modelo dose-resposta. O estudo baseia-se no comportamento de besouros adultos à exposição a cinco horas de disulfeto de carbono gasoso (CS_2). A curva de dose-resposta da mortalidade dos besouros foi formada a partir de oito dosagens (valores de x). Os resultados obtidos dos 481 besouros investigados encontram-se nas três primeiras colunas da Tabela 2.8.

O modelo logístico simples foi ajustado aos dados, e a equação de regressão é estimada por

$$logit[\widehat{\pi}(x)] = -60,72 + 34,27x .$$

Os valores ajustados por este modelo e pelo modelo *probit* para o número de besouros mortos apresentam-se nas colunas 4 e 5, respectivamente, da Tabela 2.8. Observamos assim uma grande concordância entre os valores preditos pelos dois modelos.

Tabela 2.8 : Mortalidade de besouros (Bliss,1935).

Dosagem $log_{10}CS_2(\frac{mg}{litre})$	Besouros	Besouros mortos	Ajuste *logit*	Ajuste *probit*
1,6907	59	6	3,45	3,27
1,7242	60	13	9,84	10,89
1,7552	62	18	22,45	23,65
1,7842	56	28	33,89	33,88
1,8113	63	52	50,10	49,60
1,8369	59	53	53,29	53,63
1,8610	62	61	59,22	59,63
1,8839	60	60	58,74	59,21

No modelo logístico as estimações, pontual e intervalar, das doses letais constituem aplicações das expressões (2.52) e (2.54), respectivamente, avaliadas no vetor $\widehat{\boldsymbol{\beta}} = (-60,72 \quad 34,27)^T$ e na sua matriz de covariância assintótica estimada

$$[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1} = \begin{pmatrix} 26,03 & -15,08 \\ -15,08 & 8.48 \end{pmatrix} . \tag{2.58}$$

Por exemplo, o intervalo de 95% de confiança para a dose letal mediana é

$$1,772 \pm 1,96\sqrt{(-0,029 \quad -0,052)\,[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1}\begin{pmatrix} -0,029 \\ -0,052 \end{pmatrix}} .$$

Esta e outras estimações para a dose letal encontram-se na Tabela 2.9.

Tabela 2.9 : Estimação de algumas doses letais no modelo logístico para mortalidade de besouros.

Dose Letal	Estimação Pontual	Intervalo de 95% de Confiança
DL_{50}	1,772	(1,764;1,779)
DL_{99}	1,906	(1,885;1,927)
DL_1	1,889	(1,615;1,663)

O teste de ajustamento do modelo logístico de dose-resposta produz para a estatística *deviance* o valor de $11,23$ com o correspondente nível descritivo de $0,0815$ (6 graus de liberdade). O modelo logístico não constitui, assim, um instrumento capaz de uma descrição satisfatória dos dados, o mesmo acontecendo ao modelo *probit* já que conduz a um nível descritivo de aproximadamente 10%.

Deste modo, interessa prosseguir na via da procura de modelos que se mostrem mais apropriados para a explicação destes dados. Tal será feito no próximo capítulo. $\Diamond$

CAPÍTULO 3

MODELOS DE REGRESSÃO LOGÍSTICA GENERALIZADOS

No Capítulo 2 ficou evidente a importância do modelo logístico para análise de dados binários. Porém, nas últimas duas décadas foram propostos novos modelos com o objetivo de aperfeiçoar o ajuste do modelo logístico. Todos esses modelos têm a particularidade de envolverem parâmetros adicionais de forma, introduzidos na função de ligação ou na função de tolerância ou ainda no preditor linear, e de englobarem o modelo logístico como caso especial. Alguns desses modelos logísticos generalizados serão apresentados neste capítulo. Um destaque será dado ao modelo de Stukel (1988) por possuir uma maior capacidade generalizadora.

Os problemas de dose-resposta serão aqui enfatizados por serem suscetíveis à inadequação do modelo logístico. Com efeito, a estimação de doses letais extremas é bastante sensível ao grau de ajuste do modelo, sendo particularmente beneficiada por modelos de dose-resposta mais flexíveis. Os exemplos selecionados neste capítulo são uma prova desse fato.

Este capítulo será estruturado de modo a privilegiar a análise dos modelos de regressão logística generalizados (**MRLG**'s) mais comentados na literatura, referindo algumas das dificuldades que lhe estão associadas. Em seguida, proceder-se-à à descrição das inferências para o modelo de Stukel, com as necessárias adaptações para os demais

modelos. Haverá destaque, nesse contexto, para o teste de ajuste do modelo logístico padrão.

É importante observar que todos estes modelos estruturais são enquadrados no modelo probabilístico produto de Binomiais, expresso em (2.9).

3.1. Modelo de Prentice.

Prentice (1976) propôs a seguinte função de tolerância

$$f(w) = \frac{1}{B(m_1, m_2)} e^{wm_1}(1+e^w)^{-(m_1+m_2)}, \tag{3.1}$$

representando a função de densidade de probabilidade do logaritmo de uma distribuição F-Snedecor com $2m_1$ e $2m_2$ graus de liberdade. Ela pode ser vista como uma generalização da função de tolerância do modelo logístico através da inclusão dos dois novos parâmetros m_1 e m_2.

Note que quando $m_1 = m_2 = 1$, a expressão (3.1) define o modelo logístico. No caso de $m_1, m_2 \to +\infty$ obtêm-se o modelo *probit*. O modelo complementar log-log corresponde ao caso de $m_1 \to +\infty$ e $m_2 = 1$. Quando $m_1 = m_2$ têm-se os chamados modelos simétricos, e quando $m_1 < m_2$ e $m_1 > m_2$ a função (3.1) é dita ser assimétrica negativa e positiva, respectivamente.

A probabilidade de sucesso $\pi(\eta)$, onde $\eta = \mathbf{x}^T\boldsymbol{\beta}$, fica então definida por

$$\pi(\eta) = \int_{-\infty}^{\eta} f(w)dw.$$

O Gráfico 3.1 descreve as formas para a probabilidade de sucesso $\pi(\eta)$ com base em alguns valores de m_1 e m_2 na função (3.1). Pode-se notar que $\pi(\eta)$, onde $\eta \in (-6;6)$, cresce mais lentamente (rapidamente) do que a função $\pi(\eta)$ do modelo logístico ($m_1 = m_2 = 1$) quando ($m_1 > 1, m_2 > 1$) (($m_1 < 1, m_2 < 1$)) até o ponto mediano. A partir deste ponto, o crescimento dá-se de forma contrária. Além disso, nas situações ($m_1 > 1, m_2 < 1$) e ($m_1 < 1, m_2 > 1$) a curva $\pi(\eta)$ evolui, respectivamente, de forma mais lenta e mais rápida, do que a curva logística, em todo o seu domínio.

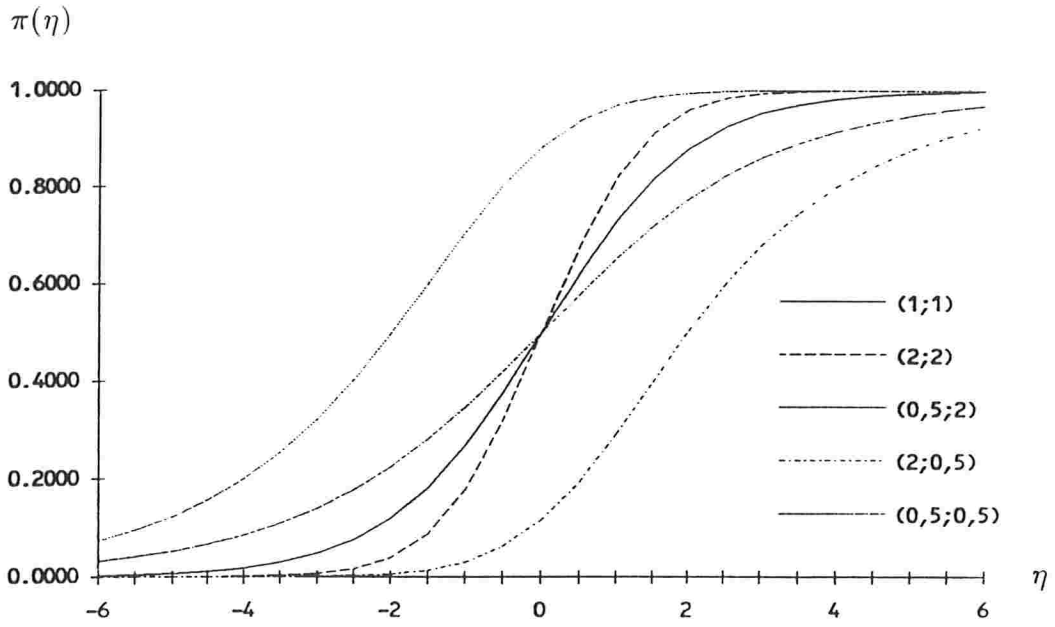

Gráfico 3.1 : Curva de $\pi(\eta)$ no modelo de Prentice para alguns valores de $(m_1 : m_2)$: modelo logístico - (1;1).

Portanto, o modelo de Prentice pode proporcionar um bom ajuste para $\pi(\eta)$, quando esta função nao é simétrica, ou mesmo sendo simétrica, quando tem uma inclinação mais pronunciada ou mais suave, do que a logística. Porém, partilha da desvantagem de não possuir uma expressão fechada para $\pi(\eta)$. Uma consequência imediata desta limitação ocorre na estimação do novo vetor de parâmetros do modelo, contituído por β e $\mathbf{m} = (m_1, m_2)^T$. De acordo com a função log-verossimilhança, dada em (2.23), a probabilidade de sucesso depende agora de β e $\mathbf{m}$: logo devemos usar também as derivadas parciais $\partial\pi_i/\partial m_l$, $i = 1, \ldots, k$ e $l = 1, 2$, para obtenção dos estimadores de MV dos referidos parâmetros. Estas derivadas, que são definidas abaixo, podem não admitir uma expressão em forma fechada

$$\frac{\partial \pi_i}{\partial m_l} = \int_{-\infty}^{\eta_i} \left(\frac{\partial f(w)}{\partial m_l} \right) dw \ , \quad i = 1, \ldots, k \ \epsilon \ l = 1, 2. \tag{3.2}$$

Para amenizar esta dificuldade, Prentice (1976) lembra que para m_1 e m_2 fixos (digamos, $\mathbf{m} = \mathbf{m}_0$) pode-se construir um processo iterativo para estimação de β. Desta forma, o procedimento de Newton-Raphson requer somente as submatrizes da matriz de informação de Fisher e da função *score* referentes a β, avaliadas em cada etapa do processo em $\mathbf{m}_0$ e $\beta_0^{(t)}$, $t \geq 0$.

Uma outra possibilidade de ajuste para o modelo de Prentice consiste em fixar o valor de um dos parâmetros adicionais. Os modelos mais usados, nesse caso, são obtidos fazendo $m_2 = 1$ ou $m_1 = 1$, e são denominados, respectivamente, de modelos Prentice-m_1 e Prentice-m_2. Esses modelos podem ser expressos em formas fechadas, dadas por

$$\pi_i = (1 + e^{-\eta_i})^{m_1} \quad e \quad \pi_i = 1 - (1 + e^{\eta_i})^{-m_2}, \quad i = 1, \ldots, k. \tag{3.3}$$

Pelas expressões apresentadas em (3.3) nota-se que as respectivas derivadas parciais, definidas em (3.2), também possuem uma forma fechada. Consequentemente, os estimadores de **MV** dos parâmetros envolvidos nos 2 modelos convenientes são facilmente encontrados através de um processo iterativo para solução do sistema de equações não lineares. Nota-se ainda que o modelo logístico é obtido de (3.3) quando $m_1 = 1$ e $m_2 = 1$, respectivamente. Com a fixação de outros conjuntos de valores para m_1 ou m_2 pode-se obter uma variedade de modelos.

Com o intuito de avaliar se a troca do modelo logístico pelos modelos de Prentice melhora substancialmente a qualidadade do ajuste, o teste de *score* tem sido bastante utilizado pelo fato de os cálculos serem avaliados sob a hipótese nula, i.e., com as estimativas dos parâmetros da regressão logística.

Métodos assintóticos para inferências simultâneas de $\boldsymbol{\beta}$ e $\mathbf{m}$ no modelo de Prentice têm sido inibidos pela dificuldade dos cálculos das derivadas parciais, apresentadas em (3.2). Entretanto, Prentice (1976) sugere que inferências sobre $(m_1; m_2)$ podem ser baseadas na distribuição assintótica da estatística da razão de verossimilhanças de Wilks, associada à maximização da função log-verossimilhança para um conjunto de valores de $(m_1; m_2)$ previamente fixados.

El-Saidi and George (1990) utilizam o modelo Prentice-m_1 para detalhar a estimação de **MV** e o ajuste do modelo em quatro problemas particulares de dose-resposta.

3.2. Modelo de Pregibon.

Como já foi mencionado na Subseção 2.8.2, a identificação dos modelos para dados binários pode ser feita também através da função de ligação. A definição ou escolha da função de ligação torna-se assim um instrumento importante para a investigação da adequação do modelo. Pregibon (1980), no contexto dos modelos lineares generalizados. preocupa-se com a função de ligação assumida na modelagem dos dados, definindo com isso um método para testar a adequação da mesma.

Pregibon (1980) parte do princípio que a função de ligação assumida num modelo particular, denotada por $g_0(\pi)$, e a função de ligação correta, porém desconhecida, denotada por $g_*(\pi)$, pertencem a uma dada família de funções de ligação, $g(\pi)$. Em particular, Pregibon assume que a função de ligação depende de dois parâmetros desconhecidos, a serem estimados, i.e., $g(\pi) \equiv g(\pi; \lambda_1, \lambda_2)$.

Usando para $g_*(\pi)$ a expansão de Taylor de 1ª ordem em torno de $g_0(\pi)$, temos

$$g_*(\pi) \simeq g_0(\pi) + (\lambda_1^* - \lambda_1^0)z_1 + (\lambda_2^* - \lambda_2^0)z_2 \ , \tag{3.4}$$

onde $z_l = [\frac{\partial g(\pi;\lambda_1,\lambda_2)}{\partial \lambda_l}] \mid_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^0}$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$, $\boldsymbol{\lambda}^0 = (\lambda_1^0, \lambda_2^0)^T$ e λ_l^* e λ_l^0, $l = 1, 2$, são os parâmetros associados às funções de ligação $g_*(\pi)$ e $g_0(\pi)$, respectivamente. Sendo assim, podemos aproximar a função de ligação $g_*(\pi) = \mathbf{x}^T\boldsymbol{\beta}$ por

$$g_0(\pi) \simeq \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T(\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*) \ . \tag{3.5}$$

onde $\mathbf{z} = (z_1, z_2)^T$ e $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*)^T$.

Uma vantagem prática deste desenvolvimento é que podemos encontrar uma aproximação para a função de ligação $g_*(\pi)$, e a partir desta, elaborar o seguinte procedimento para testar a adequação da função de ligação $g_0(\pi)$:

i) Ajustar o modelo com função de ligação $g_0(\pi)$, i.e., $g_0(\pi) = \eta = \mathbf{x}^T\boldsymbol{\beta}$;

ii) Construir as variáveis adicionais $z_l \equiv z_l(\widehat{\boldsymbol{\beta}}), l = 1, 2$;

iii) Adicionar $\mathbf{z}$ obtido em ii) às variáveis explicativas e ajustar o novo modelo com função de ligação $g_0(\pi) = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T(\boldsymbol{\lambda}^0 - \boldsymbol{\lambda})$.

Uma redução significativa no resultado do ajuste feito na etapa iii) indica a inadequação da função de ligação $g_0(\pi)$. Essa verificação pode ser feita comparando a variação da função *deviance*, ou da estatística do teste de *score*, com os níveis nominais da distribuição Qui-Quadrado com 2 graus de liberdade.

As estimativas obtidas para $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$ no final do procedimeto descrito acima não coincidem necessariamente com aquelas resultantes do método de **MV** completo. Pregibon define um processo iterativo para obtenção dessas últimas estimativas, usando as primeiras como valores iniciais. Com base em Bickel (1975), Pregibon argumenta que o seu método com uma única iteração produz um estimador assintoticamente equivalente ao estimador de **MV**.

A parada na primeira iteração deste processo, visando conseguir as estimativas de MV para os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$, reduz na maioria dos casos as dificuldades para obtenção da inversa da função de ligação. Consideremos, por exemplo, a família de funções de ligação proposta por Pregibon

$$g(\pi;\lambda_1,\lambda_2) = \frac{\pi^{\lambda_1-\lambda_2}-1}{\lambda_1-\lambda_2} - \frac{(1-\pi)^{\lambda_1+\lambda_2}-1}{\lambda_1+\lambda_2}\ , \tag{3.6}$$

onde $\lambda_l \in \Re$, $l = 1,2$, e que tem como caso particular a função de ligação *logit*, definida por

$$g_0(\pi) = \lim_{\lambda_1,\lambda_2 \to 0} g(\pi;\lambda_1,\lambda_2)\ .$$

Nota-se que a inversa da função de ligação (3.6) não possui uma forma fechada, impossibilitando assim a computação da probabilidade de sucesso π.

Em particular, quando $\lambda_2 = 0$ em (3.6), temos uma família de ligação simétrica que gera uma classe de modelos denominados modelos Pregibon-λ_1. No caso do modelo ser de dose-resposta a taxa de crescimento da curva é controlado por λ_1 verificando-se que, quando λ_1 descresce, a taxa de crescimento aumenta. Analogamente quando $\lambda_1 = 0$ em (3.6) temos uma família de ligação assimétrica que forma os modelos Pregibon-λ_2.

As etapas para o teste de adequação da função de ligação *logit* ($g_0(\pi) = ln\{\frac{\pi}{1-\pi}\}$) no modelo de Pregibon são dadas por :

i) Ajuste do modelo logístico para obtênção de $\widehat{\boldsymbol{\beta}}$, o estimador de MV de $\boldsymbol{\beta}$;

ii) Cálculo de $\widehat{\eta}$ e, consequentemente de $\widehat{\pi} = g_0^{-1}(\widehat{\eta})$ para composição das variáveis adicionais no modelo de Pregibon, i.e.,

$$z_r = \frac{1}{2}\left\{(-1)^{r+1}[ln(\widehat{\pi})]^2 - [ln(1-\widehat{\pi})]^2\right\}\ , \quad r = 1,2\ ; \tag{3.7}$$

iii) Ajuste do modelo logístico com $g_0(\pi) = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T(-\boldsymbol{\lambda})$.

Um fato importante no procedimento de Pregibon para teste da adequação da função de ligação $g_0(\pi)$ é que o teste de *score* da hipótese $H_0 : \boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^* = \mathbf{0}$ é idêntico ao teste de *score* da hipótese $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}^0$ na família que abrange todas as funções de ligação $g(\pi)$ (Pregibon, 1985).

3.3. Modelo de Aranda-Ordaz.

Ainda com o objetivo de melhorar a qualidade do ajuste da probabilidade de sucesso de um modelo Binomial, Aranda-Ordaz (1981) apresenta duas famílias de transformações (ou famílias de funções de ligação) que generalizam a função *logit*. Ele preocupa-se em atribuir a cada uma das famílias de transformações da probabilidade de sucesso π a característica simétrica ou assimétrica.

A família de transformações simétricas é definida por

$$g_{\delta_1}(\pi) = \frac{2}{\delta_1}\left[\frac{\pi^{\delta_1} - (1-\pi)^{\delta_1}}{\pi^{\delta_1} + (1-\pi)^{\delta_1}}\right], \tag{3.8}$$

onde δ_1 é chamado de parâmetro de transformação. A função (3.8), ao satisfazer a expressão (2.47), é uma função simétrica, possuindo ainda a seguinte característica : $g_{\delta_1}(\pi) = g_{-\delta_1}(\pi)$. A transformação logística é obtida de (3.8) quando $\delta_1 \to 0$. A probabilidade de sucesso, obtida pela inversão de (3.8), caracterizará o chamado modelo de Aranda-Ordaz simétrico,

$$\pi(\eta) = \begin{cases} 0, & \frac{\delta_1\eta}{2} \leq -1\ ; \\ \frac{(1+\frac{\delta_1\eta}{2})^{1/\delta_1}}{(1+\frac{\delta_1\eta}{2})^{1/\delta_1}+(1-\frac{\delta_1\eta}{2})^{1/\delta_1}}, & |\ \frac{\delta_1\eta}{2}\ | < 1\ ; \\ 1, & \frac{\delta_1\eta}{2} \geq 1\ . \end{cases} \tag{3.9}$$

Já a família de transformações assimétricas, bastante oportuna em situações com problemas no ajuste de valores extremos, tem a forma

$$g_{\delta_2}(\pi) = ln\left[\frac{(1-\pi)^{-\delta_2} - 1}{\delta_2}\right], \tag{3.10}$$

onde o parâmetro de transformação é δ_2. Quando $\delta_2 = 1$ e $\delta_2 \to 0$ a função (3.10) dá origem aos modelos logístico e complementar log-log, respectivamente. A inversa desta função compõe o denominado modelo de Aranda-Ordaz assimétrico,

$$\pi(\eta) = \begin{cases} 1 - (1+\delta_2 e^{\eta})^{-1/\delta_2}, & \delta_2 e^{\eta} > -1\ ; \\ 1, & \delta_2 e^{\eta} \leq -1\ . \end{cases} \tag{3.11}$$

Os dois modelos mencionados podem ser vistos como modelos lineares generalizados, diferindo do modelo logístico através da função de ligação. As respectivas funções

de ligação são definidas em termos da inversa das transformações (3.9) e (3.11) para evitar situações desagradáveis, como, por exemplo, a estimação de $\pi(\eta)$ fora do intervalo (0,1) ou pesos não finitos no método de estimação por mínimos quadrados ponderados.

Os dois modelos de Aranda-Ordaz podem assim ser ajustados facilmente no pacote estatístico **GLIM** (Baker and Nelder, 1978) para cada valor do parâmetro de transformação. Fixar valores para δ_l, $l = 1,2$, e observar a função *deviance* é a maneira apresentada por Aranda-Ordaz para encontrar os estimadores de **MV** dos parâmetros e para testar o ajuste dos modelos.

Esse procedimento, embora bastante simples, exige o ajuste de vários modelos e não garante que as estimativas correspondentes ao modelo de menor *deviance* sejam as estimativas de **MV**. A obtênção das estimativas de **MV** num único processo iterativo exigiria a implementação de um algoritmo mais complexo que o utilizado no pacote **GLIM**.

3.4. Modelo de Guerrero/Johnson.

Na mesma linha dos modelos discutidos na última seção apresenta-se agora uma outra família de transformações. Esta também é discutida em McCullagh and Nelder (1989, ch.11) por introduzir um parâmetro na função de ligação do modelo.

O uso de transformações nos dados antes do ajuste do modelo é uma técnica usual na análise de modelos lineares. Entre outras utilidades, estas transformações corrigem a não linearidade da equação de regressão. Uma família de transformações bastante abrangente foi desenvolvida por Box and Cox (1964), denominada de família de transformações potência. Guerrero an Johnson (1982) aplicam este tipo de transformação à chance de sucesso $(\frac{\pi}{1-\pi})$ como forma de generalização do modelo logístico.

Assim, a transformação potência da chance de sucesso deve satisfazer a seguinte relação linear

$$\left[\frac{\pi}{1-\pi}\right]^{\tau} = \eta = \mathbf{x}^T\boldsymbol{\beta} \,. \tag{3.12}$$

onde a função de ligação é dada por

$$\left[\frac{\pi}{1-\pi}\right]^{\tau} = \begin{cases} ln(\frac{\pi}{1-\pi}), & \tau = 0 \,; \\ \frac{1}{\tau}[(\frac{\pi}{1-\pi})^{\tau} - 1], & \tau \neq 0 \,. \end{cases}$$

Observe que o modelo (3.12), que denominaremos de modelo Guerrero/Johnson,

inclui o modelo logístico como caso especial, firmando-se assim como um modelo alternativo ao modelo logístico.

Ao supormos o modelo (3.12) aceitamos a existência de algum valor de τ que o torne verdadeiro. Então para alguns valores de $\boldsymbol{\beta}$ e τ podemos obter a probabilidade de sucesso como

$$\pi(\eta) = \begin{cases} (1+e^{-\eta})^{-1}, & \tau = 0\ ; \\ 0, & \tau \neq 0 \text{ e } \tau\eta < -1\ ; \\ [1+(1+\tau\eta)^{-1/\tau}]^{-1}, & \tau \neq 0 \text{ e } \mid \tau\eta \mid < 1\ ; \\ 1, & \tau \neq 0 \text{ e } \tau\eta \geq 1\ . \end{cases} \tag{3.13}$$

Salientamos desde já que a determinação dos estimadores de **MV** de $\boldsymbol{\beta}$ e τ não requer um procedimento muito complicado, como será evidenciado na Seção 3.6.

Os modelos de dose-resposta usam frequentemente a transformação logaritmica nos níveis de dosagens. Guerrero and Johnson (1982) seguem a sugestão de Cox and Snell (1989) de aplicar a transformação potência nestes níveis em lugar da logaritmica. Detalhes sobre o uso de transformações nas variáveis independentes podem ser encontrados em Box and Tidwell (1962).

Uma desvantagem desta família de transformações é a necessidade de controle de τ e η, a fim de evitar estimações da probabilidade de sucesso fora do intervalo (0,1).

3.5. Modelo de Morgan.

Morgan (1985), preocupado com a complexidade dos **MRLG**'s, apresentados até então, se baseia num modelo intitulado *quantit*, proposto por Copenhaver and Mielke (1977), para mostrar com base em certos conjuntos de dados de dose-resposta que modelos menos complicados podem melhorar significativamente o ajuste do modelo logístico.

O modelo proposto por Morgan, também chamado de modelo logístico cúbico, é um desses modelos menos complicados para a análise de dados binários, tendo a forma

$$\pi(\eta) = \frac{e^{\eta+\nu\eta^3}}{1+e^{\eta+\nu\eta^3}}\ . \tag{3.14}$$

Quando $\nu = 0$ obtemos o modelo logístico, $logit(\pi) = \eta = \mathbf{x}^T\boldsymbol{\beta}$. Este modelo pode ser visto como uma aproximação de 1^{a} ordem do modelo de Aranda-Ordaz simétrico.

Morgan direciona o modelo (3.14) para dados de dose-resposta, objetivando assim uma melhoria na estimação das doses letais extremas. O interesse por tais doses vinha já

sendo encarado na literatura com certa expectativa. Veja-se, e.g., Wetherill (1963), onde as doses letais extremas são consideradas mais relevantes do que a dose letal mediana. Uma justificativa desta relevância é que, por exemplo DL_{90} é frequentemente mais dependente do ajuste do modelo usado, do que DL_{50}.

Uma dificuldade do modelo de Morgan é causada quando $\nu < 0$, pois nesse caso $\pi(\eta)$ não é uma função monótona em η. Isto implica que as doses letais extremas, há pouco referidas, não existam ou não sejam únicas.

3.6. Modelo de Stukel.

Uma nova classe de modelos é proposta agora na descrição da dependência da probabilidade de sucesso de um modelo Binomial em relação a um conjunto de variáveis independentes. Esta classe inclui-se nos **MRLG**'s, tendo a particularidade de superar as dificuldades associadas aos modelos anteriores.

Parâmetros de forma são introduzidos também com o objetivo de modificar a curva logística nas diferentes regiões de $\pi(\eta)$, principalmente nos extremos onde o ajuste pode ser inadequado.

Esta generalização foi apresentada por Stukel (1988), tendo a seguinte forma

$$\pi_\alpha(\eta) = \frac{e^{h_\alpha(\eta)}}{1+e^{h_\alpha(\eta)}} \tag{3.15}$$

ou

$$logit(\pi_\alpha) \equiv ln\left[\frac{\pi_\alpha}{1-\pi_\alpha}\right] = h_\alpha(\eta)\ ,$$

onde $h_\alpha(\eta)$ é uma função não linear estritamente crescente em η, indexada por dois parâmetros de forma $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$, definida em duas regiões $A_1 = \{\eta \geq 0\} \equiv \{\pi_\alpha \geq 1/2\}$ e $A_2 = \{\eta < 0\} \equiv \{\pi_\alpha < 1/2\}$, por

$$h_\alpha(\eta) = \begin{cases} (-1)^{r+1}\alpha_r^{-1}[e^{(-1)^{r+1}\alpha_r\eta} - 1], & (\alpha_r > 0) \cap A_r\ ; \\ \eta, & (\alpha_r = 0) \cap A_r\ ; \\ (-1)^r\alpha_r^{-1}ln[1+(-1)^r\alpha_r\eta], & (\alpha_r < 0) \cap A_r\ ,\ \text{r=1,2.} \end{cases} \tag{3.16}$$

Quando $\alpha_1 = \alpha_2 = 0$ em (3.15) temos o modelo logístico. Cada parâmetro de forma cuida, independentemente do outro, do comportamento de cada cauda das curvas $\pi_\alpha(\eta)$ ou $h_\alpha(\eta)$. Por exemplo, quando $\alpha_1 < 0$ a curva $\pi_\alpha(\eta)$ cresce mais lentamente do que

a curva logística na cauda superior, enquanto para $\alpha_1 > 0$ a curva torna-se mais íngreme (vide Gráfico 3.2). Se $\alpha_1 > 0$ e $\alpha_2 > 0$ as funções exponenciais correspondentes causam um crescimento mais rápido de $\pi_\alpha(\eta)$ relativamente ao da curva logística. As curvas de $\pi_\alpha(\eta)$ são simétricas ou assimétricas consoante $\alpha_1 = \alpha_2$ ou $\alpha_1 \neq \alpha_2$, respectivamente.

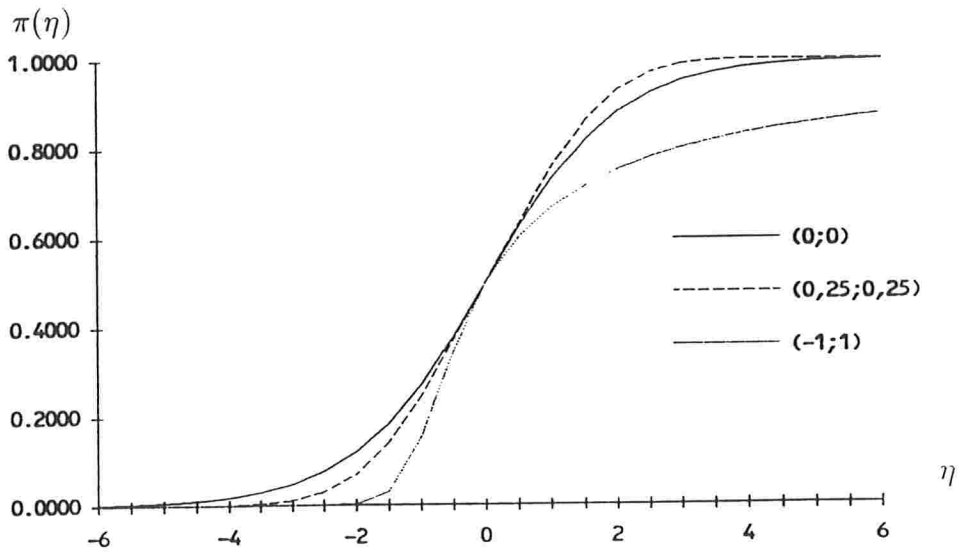

Gráfico 3.2 : Curva de $\pi(\eta)$ no modelo de Stukel para alguns valores de $(\alpha_1; \alpha_2)$: modelo logístico - (0;0).

O modelo (3.15) é aqui denominado de modelo de Stukel, sendo algumas de suas subclasses obtidas por delimitação do domínio de $\boldsymbol{\alpha}$. Entre outros, temos os modelos Stukel-α_1, Stukel-α_2, Stukel-α e Stukel-α_-, caracterizados, respectivamente, por $\alpha_2 = 0$, $\alpha_1 = 0$, $\alpha_1 = \alpha_2 = \alpha$ e $\alpha_1 = -\alpha_2 = \alpha_-$. Os dois primeiros modelos são úteis quando o objetivo de melhoria do comportamento de $\pi_\alpha(\eta)$ concentra-se nas caudas superior e inferior, respectivamente. Os dois últimos são atraentes quando os dados indicam um tratamento simétrico e assimétrico, respectivamente. Todos estes têm a vantagem de operar só com a inclusão de um parâmetro no modelo.

As funções $\pi_\alpha(\eta)$ definem uma família de funções de distribuição indexada por $\boldsymbol{\alpha}$ e associada, pelas propriedades citadas de $h_\alpha(\eta)$, a uma correspondente família de funções de densidade. Vários membros desta família aproximam até os primeiros quatro momentos algumas distribuições importantes, como, por exemplo, a distribuição Normal (modelo *probit*) quando $\alpha_1 = \alpha_2 \simeq 0.165$, a distribuição Gumbel de mínimos (modelo complementar log-log) ou de máximos (modelo log-log) quando $\boldsymbol{\alpha} = (0,62; -0,037)^T$ e $\boldsymbol{\alpha} = (-0,037; 0,62)^T$, respectivamente, e a distribuição de Laplace reduzida quando $\alpha_1 = \alpha_2 \simeq -0,077$.

O modelo (3.15) tem como objetivo manter uma relativa simplicidade algébrica e numérica. A sua aproximação por expansão de Taylor até 1ª ordem é definida por

$$logit(\pi_\alpha) \;=\; \eta + \frac{1}{2}\alpha_1\eta^2 I_{A_1} - \frac{1}{2}\alpha_2\eta^2 I_{A_2} \;. \tag{3.17}$$

O uso desta aproximação simplificadora tem o inconveniente de não conduzir a uma função monótona para $\pi_\alpha(\eta)$ quando $\alpha_j < 0$, $j = 1, 2$. Se $\alpha_1 = -\alpha_2 = \alpha_-$ em (3.17) temos o modelo Stukel-α_- de 1ª ordem, que é idêntico ao modelo de Guerrero/Johnson, até 1ª ordem.

3.6.1. Estimação dos Parâmetros.

Fixados os parâmetros de forma, o modelo (3.15) é um modelo linear generalizado com função de ligação dada por

$$g(\pi) \;=\; h_\alpha^{-1}[logit(\pi_\alpha)] \;=\; \eta \;. \tag{3.18}$$

As funções $h_\alpha(\eta)$, expressas em (3.16), e suas derivadas de primeira ordem são contínuas em $\boldsymbol{\alpha}$ e em η. A garantia de continuidade de $\pi_\alpha(\eta)$ para qualquer valor de $\boldsymbol{\alpha}$ e da existência do estimador de **MV** de $\boldsymbol{\beta}$, dado $\boldsymbol{\alpha}$, podem ser verificadas utilizando-se os resultados de Wedderburn (1976).

Para encontrar a estimativa de **MV** de $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ Stukel propõe um método semelhante aos métodos mencionados para alguns dos modelos anteriores. Isto é, fixa $\boldsymbol{\alpha}$ e estuda o comportamento da função *deviance*. A implementação deste método no **GLIM** encontra-se em Stukel (1985).

Propomos aqui, alternativamente, o uso do algoritmo de Newton-Raphson para a estimação simultânea de $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, já que temos o respaldo de um *software* (**SOC**, 1988) que permite facilmente a sua implementação. Para a construção do método iterativo, note-se que a função log-verossimilhança, expressa em (2.23), com $\pi_i \equiv \pi_\alpha(\eta_i)$, dado em (3.15), tem a seguinte forma

$$L(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{k} y_i h_\alpha(\eta_i) - \left[\sum_{i=1}^{k} n_i ln\left(1 + e^{h_\alpha(\eta_i)}\right)\right] \;, \tag{3.19}$$

onde $h_\alpha(\eta_i)$ é dada pela função (3.16) referente a i-ésima configuração de covariáveis. Sendo θ_j o j-ésimo elemento do vetor paramétrico $\boldsymbol{\theta}$, o elemento j da função *score* é

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_j} \;=\; \sum_{i=1}^{k}\left[\frac{y_i - n_i\pi_i}{\pi_i(1-\pi_i)}\right]\frac{\partial \pi_i}{\partial \theta_j} \;, \quad j = 0, \ldots, p+1 \;, \tag{3.20}$$

e o elemento (j,l) da matriz de informação de Fisher é

$$E\left[-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_l}\right] = \sum_{i=1}^{k}\left\{\frac{n_i}{\pi_i(1-\pi_i)}\right\}\frac{\partial\pi_i}{\partial\theta_j}\,\frac{\partial\pi_i}{\partial\theta_l}\ , \quad j,l=0,\ldots,p+1\ . \tag{3.21}$$

Em forma matricial a função *score* e a matriz de informação de Fisher são, respectivamente, definidas por

$$\mathbf{U}(\boldsymbol{\theta}) = \mathbf{D}_1^T\mathbf{W}_1(\mathbf{y}-\boldsymbol{\mu}) \tag{3.22}$$

e

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{D}_1^T\mathbf{W}_2\mathbf{D}_1\ , \tag{3.23}$$

onde $\mathbf{W}_1 = diag\{1/[\pi_i(1-\pi_i)],\ i=1,\ldots,k\}$, $\mathbf{W}_2 = diag\{n_i\pi_i(1-\pi_i),\ i=1,\ldots,k\}$, $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_k)^T$ com $\mu_i = n_i\pi_i$, $i=1,\ldots,k$ e $\mathbf{D}_1 = [\partial\pi_i/\partial\theta_j\ ,\ i=1,\ldots,k$ e $j=0,\ldots,p+1]$.

De acordo com o modelo (3.15), os elementos da matriz $\mathbf{D}_1$ podem apresentar-se como

$$\frac{\partial\pi_i}{\partial\theta_j} = \frac{\partial\pi_i}{\partial h_\alpha(\eta_i)}\frac{\partial h_\alpha(\eta_i)}{\partial\theta_j} = \pi_i(1-\pi_i)\frac{\partial h_\alpha(\eta_i)}{\partial\theta_j}\ , i=1,\ldots,k\ e\ j=0,\ldots,p+1, \tag{3.24}$$

onde as derivadas parciais das funções $h_\alpha(\eta_i)$, definidas em (3.16), em relação a cada um dos parâmetros do modelo, são expressas por

$$\frac{\partial h_\alpha(\eta_i)}{\partial\beta_j} = \begin{cases} [e^{(-1)^{r+1}\alpha_r\eta_i}]x_{ij}, & (\alpha_r>0)\cap A_r\ ; \\ x_{ij}, & (\alpha_r=0)\cap A_r\ ; \\ [1+(-1)^{r+1}\alpha_r\eta_i]x_{ij}, & (\alpha_r<0)\cap A_r\ ,\ \text{j=0,...,p-1}\ ; \end{cases}$$

$$\frac{\partial h_\alpha(\eta_i)}{\partial\alpha_r} = \begin{cases} \alpha_r^{-2}\{e^{(-1)^{r+1}\alpha_r\eta_i}[\alpha_r\eta_i+(-1)^r]+(-1)^{r+1}\}, & (\alpha_r>0)\cap A_r\ ; \\ (-1)^{r+1}\eta_i^2/2, & (\alpha_r=0)\cap A_r\ ; \\ \alpha_r^{-2}\{(-1)^{r+1}ln[1+(-1)^r\alpha_r\eta_i]+\alpha_r\eta_i[1+(-1)^r\alpha_r\eta_i]^{-1}\}, & (\alpha_r<0)\cap A_r\ , \end{cases}$$

r=1,2. Fora dos domínios indicados, as derivadas são nulas.

Assim, a estimativa de MV de $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, é encontrada pela solução iterativa do sistema de equações não lineares, $\mathbf{U}(\widehat{\boldsymbol{\theta}}) = \mathbf{0}$. Para o efeito, o algoritmo de Newton-Raphson tem a forma geral

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + [\mathbf{I}(\boldsymbol{\theta}^{(t)})]^{-1}\mathbf{U}(\boldsymbol{\theta}^{(t)})\ , \quad t=0,1,\ldots\ , \tag{3.25}$$

onde $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ e $\mathbf{I}(\boldsymbol{\theta})$ e $\mathbf{U}(\boldsymbol{\theta})$ são a matriz de informação de Fisher e a função *score*, dadas em (3.22) e (3.23), respectivamente.

O processo iterativo (3.25) quando $\alpha_1 = \alpha_2 = 0$ é semelhante ao processo (2.27). Este processo terá sua convergência dependente do valor inicial para o vetor paramétrico. Parece razoável iniciá-lo com $\boldsymbol{\theta}^{(0)} = (\widetilde{\boldsymbol{\beta}}^T, 0, 0)^T$, onde $\widetilde{\boldsymbol{\beta}}$ é o estimador de **MV** de $\boldsymbol{\beta}$ no modelo logístico, e adotar como critério de parada $\mid (\theta_j^{(t+1)} - \theta_j^{(t)})/\theta_j^{(t)} \mid < \varepsilon$, $\forall\ j = 0, \ldots, p+1$, para algum ε. Sendo $\widehat{\boldsymbol{\theta}}$ o estimador de **MV** de $\boldsymbol{\theta}$ temos, em grandes amostras, $Var_A(\widehat{\boldsymbol{\theta}}) = [\mathbf{I}(\boldsymbol{\theta})]^{-1}$, onde $\mathbf{I}(\boldsymbol{\theta})$ encontra-se definida em (3.23).

Aproveitando a apresentação do processo iterativo (3.25) para o modelo de Stukel, tentaremos agora definir condensadamente as adaptações necessárias para que os demais modelos descritos neste capítulo possam ser expressos como casos particulares do mesmo. Observamos que isto é possível se tivermos os elementos da matriz $\mathbf{D}_1 = [\partial\pi_i/\partial\theta_j]$.

A Tabela 3.1 resume tais modificações para alguns **MRLG**'s escolhidos de forma a incluírem somente um parâmetro de forma por razões de simplicidade.

Tabela 3.1 : Derivadas parciais para adaptação do processo iterativo (3.25) dos diversos **MRLG**'s.

Modelo	$\theta_p(*)$	$\frac{\partial\pi_i}{\partial\beta_j}$, j=0,...,p-1	$\frac{\partial\pi_i}{\partial\theta_p}$
Prentice-m_1	m_1	$\pi_i(\frac{m_1 x_{ij}}{1+e^{\eta_i}})$	$\pi_i ln(\frac{e^{\eta_i}}{1+e^{\eta_i}})$
Prentice-m_2	m_2	$(1-\pi_i)(\frac{m_2 e^{\eta_i} x_{ij}}{1+e^{\eta_i}})$	$-(1-\pi_i) ln(1+e^{\eta_i})$
Aranda-O.	δ_2	$(1-\pi_i)(\frac{e^{\eta_i} x_{ij}}{1+\delta_2 e^{\eta_i}})$	$(1-\pi_i)[-\frac{ln(1+\delta_2 e^{\eta_i})}{\delta_2^2} +$
assimétrico			$\frac{e^{\eta_i}}{\delta_2(1+\delta_2 e^{\eta_i})}] I_{(\delta_2 e^{\eta_i} > -1)}$
Guerrero/	τ	$\pi_i(1-\pi_i)(\frac{x_{ij}}{1+\tau e^{\eta_i}})$	$\frac{\pi_i(1-\pi_i)}{\tau^2}[\frac{\tau\eta_i}{1+\tau\eta_i} - ln(1+\tau\eta_i)] I_B +$
Johnson			$\pi_i(1-\pi_i)(-\eta_i^2/2) I_{(\tau=0)}$
Morgan	ν	$\pi_i(1-\pi_i)[(1+3\nu\eta_i^2)x_{ij}]$	$\pi_i(1-\pi_i)\eta_i^3$

(*) Parâmetro de forma ; $B = \{(\tau \neq 0) \cap (\mid \tau\eta_i \mid < 1)\}$.

Uma outra maneira de se proceder à estimação de **MV** simultânea de $\boldsymbol{\beta}$ e $\boldsymbol{\alpha}$ é através do algoritmo Delta (Jorgensen, 1984), um tipo de algoritmo de Newton-Raphson equivalente ao usado no procedimento de mínimos quadrados reponderados iterativamente nos modelos lineares generalizados, que atualiza a matriz de planejamento após cada iteração. Na situação em estudo, a matriz de planejamento é acrescida de duas colunas

adicionais referentes ás variáveis

$$(z_1^{(t+1)}, z_2^{(t+1)}) = \left(-\frac{\partial g(\pi)}{\partial \alpha_1}, -\frac{\partial g(\pi)}{\partial \alpha_2} \right) \Big|_{(\widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\alpha}}^{(t)})} , \tag{3.26}$$

onde

$$z_r^{(t+1)} = \begin{cases} \frac{(\alpha_r|\eta| - 1 + e^{-\alpha_r|\eta|})}{\alpha_r^2} sinal(\eta), & (\alpha_r > 0) ; \\ \frac{1}{2}\eta_i^2 sinal(\eta), & (\alpha_r = 0) ; \\ \frac{\{\alpha_r|\eta| + [1-\alpha_r|\eta|]ln(1-\alpha_r|\eta|)\}}{\alpha_r^2} sinal(\eta), & (\alpha_r < 0), \text{ r=1,2} \end{cases}$$

com $\alpha_r = \widehat{\alpha}_r^{(t+1)}$, $\eta = \widehat{\eta}_i^{(t)} = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}^{(t)}$, $i = 1, \ldots, k$, e $(\widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\alpha}}^{(t)})$ é a estimação de $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ na t-ésima iteração.

Um caso particular da aplicação do algoritmo Delta já foi referido neste capítulo a propósito dos comentários feitos na Seção 3.2, ver Pregibon (1980).

Em muitas aplicações o interesse concentra-se na estimação da curva $\pi_\alpha(\eta)$, processada pontualmente por (3.15). avaliada em $(\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\alpha}}^T)^T$. Quanto aos intervalos de confiança para $\pi_\alpha(\eta)$, uma maneira de encontrá-los consiste em determinar os intervalos de confiança para $logit(\pi_\alpha)$ e efetuar a devida transformação. A expansão de Taylor até 1ª ordem de $h_{\widehat{\alpha}}(\widehat{\eta})$ ao redor do verdadeiro $h_\alpha(\eta)$ resulta em

$$logit(\widehat{\pi}_\alpha) = \widehat{h} \simeq h + \sum_{j=0}^{p-1} (\widehat{\beta}_j - \beta_j) \frac{\partial h}{\partial \beta_j} + \sum_{j=1}^{2} (\widehat{\alpha}_j - \alpha_j) \frac{\partial h}{\partial \alpha_j} , \tag{3.27}$$

onde $\widehat{\pi}_\alpha \equiv \widehat{\pi}_{\widehat{\boldsymbol{\alpha}}}(\widehat{\eta})$ e $\widehat{h} \equiv h_{\widehat{\boldsymbol{\alpha}}}(\widehat{\eta})$. Então $Var_A[logit(\widehat{\pi}_\alpha)] = E(\widehat{h} - h)^2 \simeq \mathbf{D}_2^T Var_A(\widehat{\boldsymbol{\theta}}) \mathbf{D}_2$, onde $\mathbf{D}_2$ é o vetor de derivadas parciais de $h_\alpha(\eta)$ com respeito aos parâmetros do modelo. Para amostras grandes, o intervalo de $100(1-\alpha)\%$ de confiança para $logit[\pi_\alpha(\eta)]$ é dado por

$$logit(\widehat{\pi}_\alpha) \pm z_{1-\frac{\alpha}{2}} \sqrt{\mathbf{D}_2^T \widehat{Var}_A(\widehat{\boldsymbol{\theta}}) \mathbf{D}_2} , \tag{3.28}$$

onde $\mathbf{D}_2$ e $\widehat{Var}_A(\widehat{\boldsymbol{\theta}})$, inversa da matriz de informação de Fisher, dada em (3.23), são avaliadas em $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\alpha}}^T)^T$, e $z_{1-\frac{\alpha}{2}}$ é o percentil de $100(1-\frac{\alpha}{2})\%$ da distribuição Normal padrão. Consequentemente, após a transformação, dada em (3.15), encontramos o intervalo de $100(1-\alpha)\%$ de confiança para $\pi_\alpha(\eta)$.

O teste de ajustamento dos **MRLG**'s pode ser efetuado, analogamente ao modelo logístico, pela função *deviance* definida em (2.30), que sob a hipótese do modelo verdadeiro tem distribuição nula Qui-Quadrado com $k-p-2$ graus de liberdade, em grandes amostras.

3.6.2. Teste de Score.

Uma maneira conveniente de efetuarmos o teste de melhoramento do modelo logístico por **MRLG**'s é através do teste de *score* pela economia computacional que resulta do uso exclusivo das estimativas sob o modelo logístico.

Vamos então construir o teste de *score*, no contexto do modelo de Stukel, para a hipótese $H_0 : \boldsymbol{\alpha} = \mathbf{0}$. Sendo $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, as partições correspondentes da função *score* e da matriz de informação de Fisher, dadas em (3.22) e (3.23), são definidas por

$$\mathbf{U}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}^T \quad com\ \mathbf{U}_1 = \left[\frac{\partial L(\boldsymbol{\theta})}{\partial \beta_j};\ j = 0, \ldots, p-1\right]\ e\ \mathbf{U}_2 = \left[\frac{\partial L(\boldsymbol{\theta})}{\partial \alpha_j};\ j = 1, 2\right]$$

e

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix},$$

onde $\mathbf{I}_{11} = \left[\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_l};\ j, l = 0, \ldots, p-1\right]$, $\mathbf{I}_{12} = \left[\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \beta_j \partial \alpha_l};\ j = 0, \ldots, p-1 \text{ e } l = 1, 2\right]$, $\mathbf{I}_{21} = \mathbf{I}_{12}^T$ e $\mathbf{I}_{22} = \left[\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \alpha_j \partial \alpha_l};\ j, l = 1, 2\right]$.

Seja $\widetilde{\mathbf{U}}_2$ o vetor *score* de $\boldsymbol{\alpha}$ avaliado sob H_0. Sob essa hipótese $\widetilde{\mathbf{U}}_2$ segue assintoticamente uma distribuição Normal bivariada, com vetor de média nulo e matriz de covariância $\widetilde{\mathbf{V}}_U = \mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12}$, onde as submatrizes $\mathbf{I}_{jh}$, $j, h = 1, 2$, são avaliadas em $(\widetilde{\boldsymbol{\beta}}^T, \mathbf{0}^T)^T$ com $\widetilde{\boldsymbol{\beta}}$ sendo o estimador de **MV** de $\boldsymbol{\beta}$ sob H_0. Assim, a estatística de *score*, sob H_0, fica dada por

$$\widetilde{\mathbf{U}}_2^T \widetilde{\mathbf{V}}_U^{-1} \widetilde{\mathbf{U}}_2\ , \tag{3.29}$$

onde $\widetilde{\mathbf{U}}_2 = (u_{21}, u_{22})^T$ com

$$u_{2j} = \lim_{\alpha_j \to 0} \frac{\partial^2 L(\boldsymbol{\theta})}{\partial \alpha_j} = \sum_{i \in A_j} \frac{1}{2} \widehat{\eta}_i^2 [y_i - n_i \pi_{\widehat{\boldsymbol{\alpha}}}(\widehat{\eta}_i)] sinal(\widehat{\eta}_i),\ j = 1, 2\ ,$$

seguindo aproximadamente uma distribuição Qui-Quadrado com 2 graus de liberdade para grandes amostras.

Uma alternativa para testar o melhoramento do modelo logístico pelo modelo de Stukel, ainda através do teste de *score*, é baseada na Seção 3.2. Pregibon (1985) adiciona novas variáveis no modelo logístico em quantidade igual aos parâmetros introduzidos e consegue disso um teste de *score* para a hipótese de anulamento dos correspondentes

parâmetros. No caso presente, estas variáveis adicionais encontram-se em (3.26), e sob H_0, têm a forma

$$(z_1, z_2) = \left(\frac{1}{2}\widehat{\eta}^2 I_{A_1}, -\frac{1}{2}\widehat{\eta}^2 I_{A_2}\right), \tag{3.30}$$

onde $A_1 = \{\widehat{\eta} \geq 0\}$ e $A_2 = \{\widehat{\eta} < 0\}$. Com as variáveis adicionais no modelo logístico, este torna-se $logit(\pi) = \eta + \alpha_1 z_1 + \alpha_2 z_2$, definindo o novo contexto de referência para o teste de *score* de $\alpha_1 = \alpha_2 = 0$.

A utilidade desta métodologia reside na sua praticidade, já que o teste da nulidade de coeficientes de regressão encontra-se bastante difundido. Além disso, outros **MRLG**'s podem ser igualmente beneficiados por esta opção. As respectivas derivadas de $g(\pi)$ com relação aos parâmetros de forma de alguns modelos definidos neste capítulo, avaliadas no modelo logístico (variáveis adicionais), encontram-se na Tabela 3.2.

Tabela 3.2 : Variáveis derivadas usadas no teste de *score* para os parâmetros de forma nos diversos **MRLG**'s.

Modelo	Parâmetro de forma	Variável adicional
Prentice	m_1	$(1-\widehat{\pi})^{-1} ln(\widehat{\pi})$
	m_2	$(\widehat{\pi})^{-1} ln(1-\widehat{\pi})$
Pregibon	λ_1	$\frac{1}{2}\{[ln(\widehat{\pi})]^2 - [ln(1-\widehat{\pi})]^2\}$
	λ_2	$-\frac{1}{2}\{[ln(\widehat{\pi})]^2 + [ln(1-\widehat{\pi})]^2\}$
Aranda-O. simétrico	δ_1	$\widehat{\eta}^3$
Aranda-O. assimétrico	δ_2	$[1 + \widehat{\pi}^{-1} ln(1-\widehat{\pi})]$
Guerrero/Johnson	τ	$-\frac{1}{2}\widehat{\eta}^2$
Morgan	ν	$\widehat{\eta}^3$
Stukel	α_1	$\frac{1}{2}\widehat{\eta}^2 I_{(\widehat{\eta}\geq 0)}$
	α_2	$-\frac{1}{2}\widehat{\eta}^2 I_{(\widehat{\eta}<0)}$
Stukel-α	$\alpha = \alpha_1 = \alpha_2$	$\frac{1}{2}\widehat{\eta}^2 sinal(\widehat{\eta})$
Stukel-α_-	$\alpha_- = \alpha_1 = -\alpha_2$	$\frac{1}{2}\widehat{\eta}^2$

3.6.3. Modelos de Dose-Resposta.

Os **MRLG**'s apresentados neste capítulo como alternativa a uma possível inadequação do modelo logístico, são utilizados frequentemente em dados de dose-resposta

devido ao fato da estimação de doses letais extremas ser bastante dependente do modelo ajustado.

Consideraremos apenas a classe de **MRLG**'s de Stukel para estimação das doses letais por tratar-se da classe mais geral. Lembramos ainda que os parâmetros de forma correspondentes conferem uma maior flexibilidade ao comportamento da curva de dose-resposta, pela facilidade de tratamento assimétrico e controle do peso das caudas.

Usando os mesmos passos da Subseção 2.8.3 e considerando $\pi_\alpha(\eta)$, expresso em (3.15), uma família de funções de distribuição acumulada indexada por $\boldsymbol{\alpha}$, passemos a estimação da dose letal de $100P\%$ de sucesso, $0 < P < 1$. O estimador de **MV** de DL_{100P}, dado pela inversão das funções $h_\alpha(\eta)$, expostas em (3.16), tem a seguinte forma

$$\widehat{DL}_{100P} \;=\; \frac{\widehat{\eta}_P - \widehat{\beta}_0}{\widehat{\beta}_1} \;\equiv\; d(\widehat{\boldsymbol{\theta}})\;, \tag{3.31}$$

onde $\widehat{\eta}_P$ é obtido das funções $h_\alpha(\eta)$, em (3.32) abaixo, por avaliação em $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^T, \widehat{\boldsymbol{\alpha}}^T)^T$:

$$\eta_P = \begin{cases} (-1)^{r+1}\alpha_r^{-1} ln[1 + (-1)^{r+1}\alpha_r logit(P)], & (\alpha_r > 0) \cap A_r\;; \\ logit(P), & (\alpha_r = 0) \cap A_r\;; \\ (-1)^{r+1}\alpha_r^{-1}[1 - e^{(-1)^r \alpha_r logit(P)}], & (\alpha_r < 0) \cap A_r\;,\ \text{r=1,2}\;. \end{cases} \tag{3.32}$$

onde $A_1 = \{P \geq 1/2\}$ e $A_2 = \{P < 1/2\}$.

Observamos que a expressão (3.31) quando $\alpha_1 = \alpha_2 = 0$ define o estimador de **MV** de DL_{100P} no modelo logístico, dado em (2.52).

A variância assintótica de $\widehat{DL}_{100P}$ obtida de uma expansão em série de Taylor de $d(\widehat{\boldsymbol{\theta}})$ até 1^a ordem em torno de $d(\boldsymbol{\theta})$, fica dada por

$$Var_A[d(\widehat{\boldsymbol{\theta}})] \;=\; \mathbf{D}(\boldsymbol{\theta})^T[\mathbf{I}(\boldsymbol{\theta})]^{-1}\mathbf{D}(\boldsymbol{\theta})\;. \tag{3.33}$$

onde $\mathbf{D}(\boldsymbol{\theta})$ é o vetor de derivadas parciais de DL_{100P} em relação aos parâmetros do modelo de dose-resposta, $\boldsymbol{\theta} = (\beta_0, \beta_1, \alpha_1, \alpha_2)^T$, i.e.,

$$\mathbf{D}(\boldsymbol{\theta}) \;=\; \left[-\frac{1}{\beta_1}, \frac{\beta_0 - \eta_P}{\beta_1^2}, \frac{1}{\beta_1}\left(\frac{\partial \eta_P}{\partial \alpha_1}\right), \frac{1}{\beta_1}\left(\frac{\partial \eta_P}{\partial \alpha_2}\right)\right]^T,$$

sendo, para $A_1 = \{P \geq 1/2\}$ e $A_2 = \{P < 1/2\}$,

$$\frac{\partial \eta_P}{\partial \alpha_r} = \begin{cases} \frac{1}{\alpha_r^2}\left[\frac{\alpha_r logit(P)}{1+(-1)^{r+1}\alpha_r logit(P)} + (-1)^r ln\left\{1+(-1)^{r+1}\alpha_r logit(P)\right\}\right], & (\alpha_r > 0) \cap A_r \ ; \\ \frac{1}{\alpha_r^2}\left[\left\{(-1)^{r+1} + \alpha_r logit(P)\right\} e^{(-1)^r \alpha_r logit(P)} + (-1)^r\right], & (\alpha_r < 0) \cap A_r \ ; \\ 0, & \text{c. c., r=1,2} \ . \end{cases} \tag{3.34}$$

Dados (3.31) e (3.33), um intervalo de $100(1-\alpha)\%$ de confiança para a dose letal de $100P\%$ de sucesso, em grandes amostras, tem a forma

$$\widehat{DL}_{100P} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}_A[d(\widehat{\boldsymbol{\theta}})]} \ , \tag{3.35}$$

onde $\widehat{Var}_A[d(\widehat{\boldsymbol{\theta}})]$ é a variância assintótica (3.33), avaliada em $\widehat{\boldsymbol{\theta}}$, e $z_{1-\frac{\alpha}{2}}$ é o percentil $100(1-\frac{\alpha}{2})\%$ da distribuição Normal padrão.

Este método de construção de intervalos de confiança assintóticos para as doses letais em **MRLG**'s, usado por Stukel (1990), é fundamentado na base de sua simplicidade e de sua performance em pequenas amostras, de acordo com os estudos de Schwenke and Milliken (1983) sobre intervalos de confiança para o valor inverso de uma equação de regressão não linear.

Contudo, a aplicação deste a doses letais extremas pode levar a intervalos de confiança inadequados por abrangerem valores impossíveis para as doses, ainda que as estimativas pontuais possam ser positivas. Como refere Stukel (1990), tal possibilidade resulta de uma inflação da variância assintótica provocada pelo aumento de $\mid \partial \eta_P / \partial \alpha_r \mid$, $r = 1, 2$, quando P tende a 0 ou 1 (note-se que esta quantidade é usada para cálculo de $\mathbf{D}(\boldsymbol{\theta})$). Este aumento é particularmente substancial quando $\alpha_r < 0$, $r = 1, 2$, devido à forma funcional daquelas derivadas - recorde-se (3.34). Daí a sugestão de Stukel em evitar modelos com $\widehat{\alpha}_r < 0$, $r = 1, 2$, quando a estimação de doses letais extremas for um objetivo de análise.

3.7. Exemplos.

Uma vez apresentados vários **MRLG**'s passamos agora à exemplificação de alguns destes modelos alternativos ao modelo logístico. Para esse fim, consideraremos três conjuntos de dados onde se verifica a necessidade de um melhoramento no ajuste do modelo logístico. Os dois primeiros são referentes à situação de dose-resposta, uma das mais sensíveis a ajustes inadequados do modelo logístico.

A análise descrita de todos esses exemplos será baseada na classe de modelos de Stukel. A razão está na diversidade desta classe e no fato dos outros **MRLG**'s, por nós também ajustados, não terem proporcionado resultados superiores.

A análise dos exemplos incidirá sobre a estimação de $\pi(\eta)$, o teste de *score* do modelo logístico em relação aos **MRLG**'s, o teste de ajustamento do modelo em referência e a estimação de algumas doses letais. Os cálculos são efetuados, em sua maioria, através de 2 programas por nós elaborados na linguagem **SOC** apresentados no Apêndice 3. O primeiro programa elabora o teste de *score* do modelo logístico no contexto dos modelos de Stukel e implementa o algoritmo para encontro da estimativa de **MV** do vetor paramétrico e para ajuste dos modelos de Stukel. O segundo estima doses letais, sendo assim usado só para dados de dose-resposta.

Os dois primeiros exemplos foram analisados na maioria dos artigos que originaram os **MRLG**'s discutidos aqui. Decidimos continuar a usá-los para exemplificar a aplicação dos nossos programas que se revelam distintos dos procedimentos computacionais usados naqueles artigos. O terceiro exemplo difere dos anteriores no aspecto em que introduz mais do que uma covariável, servindo então como ilustração da aplicação dos modelos de Stukel num contexto de regressão logística múltipla.

Exemplo 3.1 : Retomando o estudo da mortalidade de besouros devido à exposição a gás carbono (Bliss, 1935), verificamos no Exemplo 2.3 um mau ajuste do modelo logístico. Esta situação sugere a investigação de um novo modelo que melhore significativamente tal ajuste.

Alguns modelos de Stukel são os candidatos ao melhoramento do modelo logístico. Testes de *score* para avaliação destes encontram-se montados na Tabela 3.4, onde, de acordo com os níveis descritivos, todos os modelos apresentados são significativamente melhores do que o logístico. A preferência de melhor ajuste cairia sobre o modelo Stukel-α_2 se não fosse o fato de este possuir estimativa do parâmetro de forma negativa, como se registra na última coluna da Tabela 3.3. Tal escolha provocaria estimações inconvenientes para as doses letais extremas, já comentadas na Subseção 3.6.3.

Deste modo, o modelo escolhido para melhorar o ajuste do modelo logístico é o modelo Stukel-α_1, tendo o menor desvio padrão dos estimadores dos parâmetros de forma em relação ao modelo anterior. Do processo iterativo (3.25) encontramos a estimativa de **MV** do vetor paramétrico e da matriz de covariância assintótica do respectivo estimador,

dadas por

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} -47,386 \\ 26,649 \\ 0,532 \end{pmatrix} \quad e \quad [\mathbf{I}(\widehat{\boldsymbol{\theta}})]^{-1} = \begin{pmatrix} 41,863 & -23,678 & 1,487 \\ & 13,397 & -0,849 \\ & & 0,081 \end{pmatrix} .$$

Observe-se de os elementos de $[\mathbf{I}(\widehat{\boldsymbol{\theta}})]^{-1}$ acima, referente ao parâmetro $\boldsymbol{\beta}$, não se alteram muito quando comparados com $[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1}$ do modelo logístico, dada em (2.58).

Tabela 3.3 : Testes de *score* e estimação dos parâmetros referentes a alguns modelos de Stukel para mortalidade de besouros.

Modelo	Teste de *score*			Estimação
	Estatística	G.L.	Nível descritivo	$\widehat{\alpha}_j \pm \sqrt{\widehat{Var}_A(\widehat{\alpha}_j)}$
Stukel-(α_1, α_2)	7,731	2	0,021	0,157 $\pm$ 0,265
				-0,532 $\pm$ 0,444
Stukel-α_1	6,244	1	0,012	0,532 $\pm$ 0,285
Stukel-α_2	6,776	1	0,009	-0,771 $\pm$ 0,324

A estatística do teste de ajustamento, dada em (2.30), do modelo Stukel-α_1 tem como valor observado $\chi^2_{(5)} = 3,83$, apresentando-se assim com ajustamento bastante significativo ($P = 0,574$). A apresentação visual do melhoramento do ajuste do modelo escolhido face ao modelo logístico encontra-se no Gráfico 3.3.

Com base num melhor modelo para o estudo da mortalidade dos besouros (modelo Stukel-α_1), partimos para a estimação de doses letais neste experimento. Estimativas de algumas doses letais e seus respectivos desvios padrões e intervalos de 95% de confiança, encontram-se na Tabela 3.4. Desta concluímos que as doses estimadas referentes ao modelo Stukel-α_1 são de grandeza inferior às respectivas doses estimadas pelo modelo logístico, o que parece traduzir uma super-estimação das doses letais por este modelo.

Ainda com relação às estimativas das doses letais, encontramos nas estimativas dos desvios padrões das doses letais estimadas, para $P \geq 1/2$, no modelo Stukel-α_1 valores inferiores às respectivas estimativas no modelo logístico. Isto é um exemplo de que nem sempre a inclusão de parâmetros de forma inflaciona a variância.

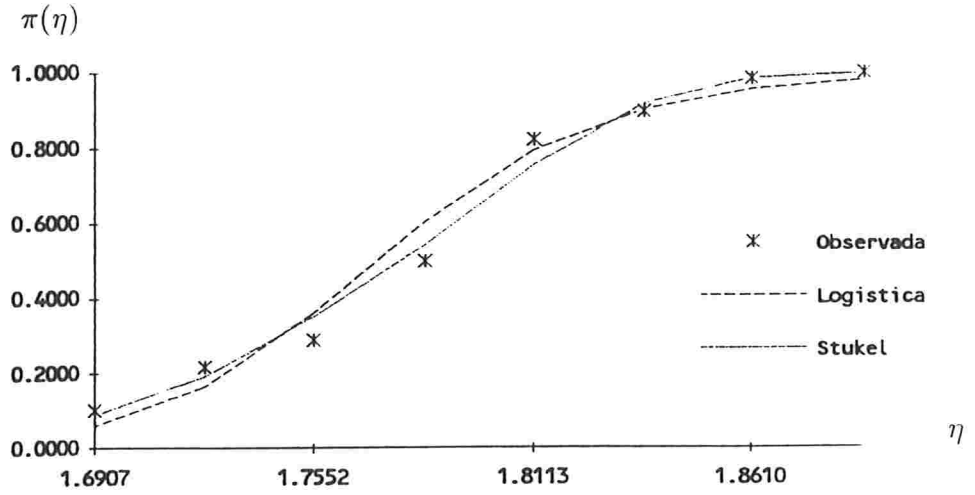

Gráfico 3.3 : Curva de $\pi(\eta)$ ajustada nos modelos logístico e Stukel-α_1 e os valores observados para mortalidade de besouros.

Tabela 3.4 : Estimação de algumas doses letais nos modelos logístico e Stukel-α_1 para mortalidade de besouros.

P	Modelo Logístico		Modelo Stukel-α_1	
	$\widehat{DL}_{100P}(*)$	Int. confiança $[\widehat{DL}_{100P}$:95%]	$\widehat{DL}_{100P}(*)$	Int. confiança $[\widehat{DL}_{100P}$:95%]
0,001	$1{,}570_{(0{,}018)}$	1,535-1,606	$1.519_{(0{,}034)}$	1.451-1.586
0,005	$1{,}617_{(0{,}014)}$	1.589-1.645	$1.579_{(0{,}026)}$	1.528-1.631
0,010	$1{,}638_{(0{,}013)}$	1.613-1.662	$1.606_{(0{,}023)}$	1.561-1.650
0,050	$1{,}686_{(0{,}009)}$	1.669-1,703	$1{,}668_{(0{,}014)}$	1.639-1.696
0,200	$1{,}731_{(0{,}006)}$	1,720-1,742	$1{,}726_{(0{,}007)}$	1.712-1,740
0,500	$1{,}772_{(0{,}004)}$	1.764-1,779	$1.778_{(0{,}005)}$	1.769-1.788
0,800	$1{,}812_{(0{,}005)}$	1.803-1.822	$1.817_{(0{,}005)}$	1.808-1.826
0,950	$1.858_{(0{,}008)}$	1.842-1.873	$1.845_{(0{,}006)}$	1.834-1.856
0,990	$1.906_{(0{,}011)}$	1.883-1,928	$1.865_{(0{,}010)}$	1.847-1.884
0,995	$1.926_{(0{,}013)}$	1.900-1,952	$1{,}873_{(0{,}011)}$	1.850-1.895
0,999	$1.973_{(0{,}017)}$	1.940-2.007	$1.887_{(0{,}016)}$	1.857-1.917

(*) desvio padrão assintóticos da dose letal estimada $\widehat{DL}_{100P}$.

Outros **MRLG**'s também foram ajustados a este conjunto de dados, sendo os mais significativos os que privilegiam a assimetria da curva $\pi(\eta)$, não se distanciando das conclusões obtidas pelo modelo Stukel-α_1. $\diamondsuit$

Exemplo 3.2 : Os problemas de dose-resposta não se esgotam em Toxicologia. Milecer and Szczotka (1966) investigam a idade do início de menstruação em 3918 garotas de Varsóvia. Para 25 médias de idade observou-se a ocorrência ou não do início de períodos de menstruação nas adolescentes. Os dados deste estudo encontram-se nas 3 primeiras colunas da Tabela 3.5.

A estatística do teste de ajustamento, dada em (2.30), para o modelo logístico tem o valor observado $\chi^2_{(23)} = 26,2$, logo observa-se um ajuste razoável deste modelo ao cunjunto de dados. Porém, pela análise de resíduos, ou mesmo pela observação do Gráfico 3.4, nota-se que as caudas da curva $\pi(\eta)$ podem ser melhor ajustadas.

Portanto, faremos uma investigação sobre os **MRLG**'s que poderão melhorar o ajuste nas caudas. De acordo com a Tabela 3.6, observamos que para os vários testes de *score*, onde o modelo logístico é posto em confronto com vários **MRLG**'s, o modelo Stukel-α_2 é o que produziu o melhor nível descritivo significativo, sendo assim o modelo escolhido para suprir a inadequação do modelo logístico nas caudas.

O ajuste do modelo Stukel-α_2 aos dados produziu como estimativa de **MV** do vetor paramétrico e da matriz de covariância assintótica do respectivo estimador os seguintes valores

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} -18,821 \\ 1,457 \\ 0,219 \end{pmatrix} \quad e \quad [\mathbf{I}(\widehat{\boldsymbol{\theta}})]^{-1} = \begin{pmatrix} 0,933 & -0,070 & 0.058 \\ & 0,005 & -0,004 \\ & & 0.006 \end{pmatrix} .$$

As estimativas pontuais de $n_i\pi_i$, $i = 1, \ldots, 25$, referente ao modelo adotado encontram-se nas 2 últimas colunas da Tabela 3.5. A sua comparação com os correspondentes valores para o modelo logístico refletem bem o fraco ajuste deste na cauda inferior da curva $\pi(\eta)$.

O teste de razão de verossimilhanças de Wilks para ajustamento do modelo Stukel-α_2 produz $\chi^2_{(22)} = 15,94$ com nível descritivo $P = 0,819$. O bom ajuste deste é fácilmente perceptível no Gráfico 3.4.

Tabela 3.5 : Ocorrência de período menstrual quanto à idade de garotas (Milicer and Szczotka, 1966).

Idade	Número de garotas		Estimação de $n\pi$	
	Menstruadas	Entrevistadas	M. Logístico	M. Stukel-α_2
9,21	0	376	0,76	0,01
10,21	0	200	2,06	0,37
10,58	0	93	1,74	0,58
10,83	2	120	3,34	1,56
11,08	2	90	3,72	2,28
11,33	5	88	5,36	4,10
11,58	10	105	9,32	8,44
11,83	17	111	14,19	14,45
12,08	16	100	18,06	19,72
12,33	29	93	23,15	26,02
12,58	39	100	33,26	37,21
12,83	51	108	46,27	50,42
13,08	47	99	52,46	55,25
13,33	67	106	66,67	68,38
13,58	81	105	75,41	75,96
13,83	88	117	92,79	92,45
14,08	79	98	83,51	82,73
14,33	90	97	86,97	85,98
14,58	113	120	111,45	110,19
14,83	95	102	97,05	96,06
15,08	117	122	118,00	116,97
15,33	107	111	118,55	107,78
15,58	92	94	92,61	92,09
15,83	112	114	112,87	112,38
17,53	1049	1049	1048,40	1047,82

Tabela 3.6 : Teste de *score* para o modelo logístico com relação à diferentes **MRLG**'s no estudo do início da idade menstrual.

Modelo	Estatística	Graus de liberdade	Nível descritivo
Prentice	8,089	2	0,018
Stukel-(α_1, α_2)	7,938	2	0,019
Pregibon	7,742	2	0,021
Stukel-α_2	7,232	1	0,007
Prentice-m_1	5,753	1	0,016
Guerrero/Johnson ou Stukel-α	3,726	1	0,054
Pregibon-λ_2	3,652	1	0,056
Aranda-O. assimétrico	3,267	1	0,071
Prentice-m_2	3,267	1	0,071
Stukel-α	3,097	1	0,078
Pregibon-λ_1	2,892	1	0,089
Aranda-O. simétrico ou Morgan	2,489	1	0,115
Stukel-α_1	0,706	1	0,401

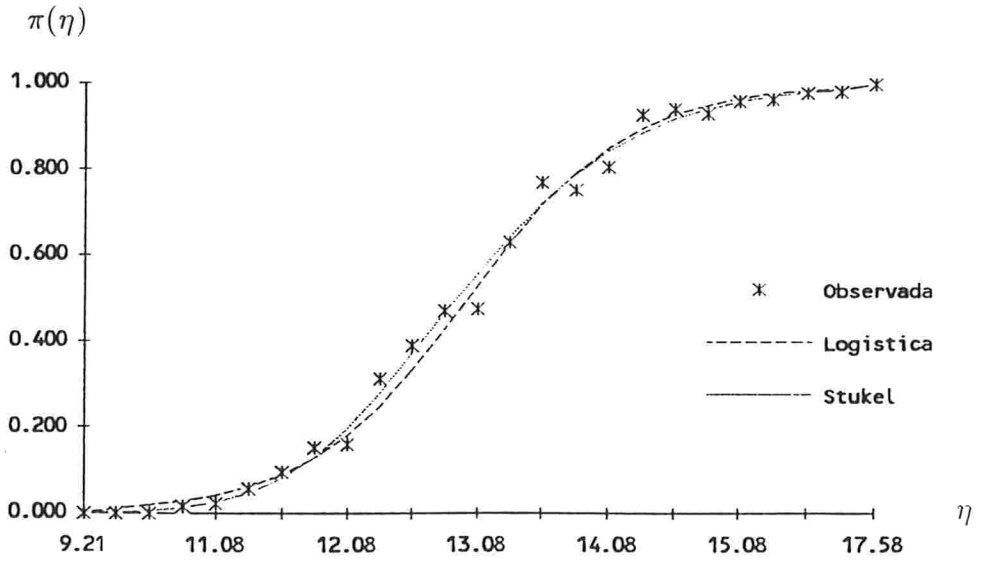

Gráfico 3.4 : Curva de $\pi(\eta)$ ajustada nos modelos logístico e Stukel-α_2 e os valores observados para início da idade de menstruação.

A Tabela 3.7 apresenta a estimação de algumas idades de início da menstruação permitindo uma comparação entre os modelos logístico e Stukel-α_2. Observa-se assim que as idades de início da menstruação extremas localizadas na cauda inferior foram aumentadas no modelo Stukel-α_2 em relação ao logístico, no resto da curva as idades de início da menstruação no modelo escolhido são mais parecidas entre si. Os intervalos de 95% de confiança para quase todas as idades de início da menstruação no modelo escolhido são maiores do que no logístico, devido logicamente ao aumento da estimativa de variância assintótica.

Tabela 3.7 : Estimação de algumas idades de início da menstruação nos modelos logístico e Stukel-α_2 no estudo da idade menstrual.

P	Modelo Logístico		Modelo Stukel-α_2	
	$\widehat{DL}_{100P}$	Int. confiança $[\widehat{DL}_{100P}$;95%]	$\widehat{DL}_{100P}$	Int. confiança $[\widehat{DL}_{100P}$;95%]
0,001	8,774	8,464-9,085	10,034	9,542-10,527
0,005	9,763	9,520-10,006	10,509	10,170-10,847
0,010	10,191	9,976-10,406	10,739	10,465-11,012
0,050	11,202	11,053-11,352	11,361	11,219-11,502
0,200	12,157	12,059-12,255	12,089	11,989-12,189
0,500	13,007	12,931-13,082	12,920	12,827-13,013
0,800	13,856	13,760-13,952	13,872	13,769-13,974
0,950	14,811	14,664-14,958	14,941	14,759-15,123
0,990	15,822	15,611-16,034	16,074	15,790-16,358
0,995	16,250	16,010-16,490	16,554	16,225-16,883
0,999	17,239	16,931-17,546	17,661	17,227-18,095

Outro modelos não referidos na Tabela 3.6 tiveram resultados semelhantes ao modelo Stukel-α_2, como é o caso do modelo *probit* ou do modelo considerado por Guerrero and Johnson (1982) com transformação potência na variável independente. $\Diamond$

Exemplo 3.3 : Um conjunto sobre o efeito convulsivo da insulina em ratos é encontrado em Finney (1978, sec.18.2). Nesse estudo consideram-se dois grupos caracterizados de acordo com o tipo de droga (padrão ou insulina) a aplicar nos ratos. Em cada grupo houve aplicação de algumas dosagens da respectiva droga, e para cada uma das 14 configurações

definidas pela combinação das duas covariáveis (droga e dosagem), observou-se o número de ratos que apresentaram convulsão ou sintomas de colapso. Os dads obtidos neste experimento apresentam-se na Tabela 3.8.

O teste de ajustamento do modelo logístico padrão com as duas covariáveis inclusas (sem interação) produz 26,61 como valor observado da estatística (2.30), cuja distribuição nula $\chi^2_{(11)}$ origina um nível descritivo $P = 0,005$. Se incluirmos o termo de interação entre as covariáveis no modelo anterior o teste resulta no nível descritivo $P = 0,013$ ($\chi^2_{(10)} = 22,38$). Logo, concluímos pela falta de ajustamento do modelo logístico padrão. Conclusões semelhantes são obtidas para os modelos *probit* e complementar log-log. Daí a necessidade de partirmos para a investigação de um dos **MRLG**'s como alternativa.

Tabela 3.8 : Convulsão em ratos (Finney, 1978).

Droga	Dosagem (0,001 IU)	Ratos	Ratos com convulsão
padrão	3,4	0	33
padrão	5,2	5	32
padrão	7,0	11	38
padrão	8,5	14	37
padrão	10,5	18	40
padrão	13,0	21	37
padrão	18,0	23	31
padrão	21,0	30	37
padrão	28,0	27	30
insulina	6,5	2	40
insulina	10,0	10	30
insulina	14,0	18	40
insulina	21,5	21	35
insulina	29,0	27	37

O modelo de Stukel-(α_1, α_2) sem interação entre as duas covariáveis melhora sensivelmente o modelo logístico padrão visto que o teste de *score* com $H_0 : \alpha_1 = \alpha_2 = 0$ (modelo logístico) resulta num valor observado para a sua estatística, expressa em (3.29). igual a $\chi^2_{(2)} = 9,83$ ($P = 0,007$). Os demais **MRLG**'s foram testados não originando

resultados significativamente superiores.

Portanto, diante da escolha do modelo de Stukel-(α_1, α_2) encontramos pelo processo iterativo (3.25) a estimativa de **MV** de $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \alpha_1, \alpha_2)^T$, onde β_1 e β_2 se referem, respectivamente, a droga e dosagem, e a estimativa da matriz de covariância assintótica de $\widehat{\boldsymbol{\theta}}$, dadas respectivamente por

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} -1,131 \\ 0,102 \\ -0,277 \\ -0,053 \\ 3,327 \end{pmatrix} \quad e \quad [\mathbf{I}(\widehat{\boldsymbol{\theta}})]^{-1} = \begin{pmatrix} 0,206 & -0,020 & 0,055 & 0,285 & 0,970 \\ & 0,002 & -0,005 & -0,029 & -0,091 \\ & & 0,021 & 0,080 & 0,266 \\ & & & 0,495 & 1,282 \\ & & & & 4,797 \end{pmatrix}.$$

Note-se, para efeitos comparativos, que as correspondentes estimativas para o modelo logístico sem interação são

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} -2,078 \\ 0,161 \\ -0.875 \end{pmatrix} \quad e \quad [\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1} = \begin{pmatrix} 0,051 & -0,003 & -0.002 \\ & 0.000 & -0.001 \\ & & 0.055 \end{pmatrix}.$$

O teste de ajustamento do modelo de Stukel-(α_1, α_2) produz $\chi^2_{(9)} = 8,11$ $(P = 0,523)$, o que revela uma vantagem significativa deste modelo relativamente ao modelo logístico. O ajuste pode ainda ser substancialmente melhorado com a introdução do termo de interação já que se obtem $\chi^2_{(8)} = 2,2$ $(P = 0,974)$. Na tabela 3.9 encontramos os valores observados e ajustados pelos modelos logístico padrão e Stukel-(α_1, α_2), ambas sem interação e com seus respectivos resíduos.

Ainda com relação a Tabela 3.9, observamos um melhoramento das frequências preditas pelo modelo de Stukel-(α_1, α_2) em cada grupo de droga, principalmente nas dosagens inferiores, o que evidencia a habilidade dos parâmetros de forma (fundamentalmente, α_2) na explicação das frequências observadas. O efeito da droga pode ser testado usando o teste da razão de verossimilhanças de Wilks, expresso em (2.32), adaptado ao modelo de Stukel-(α_1, α_2). Nesta situação o valor observado da estatística $\chi^2_{(1)} = 13,49$ $(P = 0,0002)$ sugere um efeito significativo da insulina na indução de convulsão em ratos.

Como observação final referimos que a introdução da transformação logarítmica na dosagem, como é comum no estudo de dados de dose-resposta, altera parte das conclusões acima citadas. Com efeito, o modelo logístico (sem interação) passa a ter um bom ajuste $(P \simeq 0,72)$, embora também fortemente superado pelo correspondente modelo de Stukel-(α_1, α_2) $(P \simeq 0,998)$. Neste contexto, a introdução de interação em ambos os modelos não se mostra relevante.

Tabela 3.9 : Estimação e resíduos dos modelos logístico e Stukel-(α_1, α_2) na convulsão em ratos.

Droga (*)	Dosagem (0,001 IU)	Observado	Logístico		Stukel-(α_1, α_2)	
			$n\widehat{\pi}$	Resíduo	$n\widehat{\pi}$	Resíduo
0	3,4	0	5,87	-5,87	0,75	-0,75
0	5,2	5	7,18	-2,18	4,15	0,85
0	7,0	11	10,60	0,40	11,04	-0,04
0	8,5	14	12,21	1,78	14,71	-0,71
0	10,5	18	16,20	1,80	19,38	-1,38
0	13,0	21	18,67	2,33	20,33	0,67
0	18,0	23	21,55	1,45	20,70	2,30
0	21,0	30	29,13	0,87	26,98	3,01
0	28,0	27	27,59	-0,59	25,20	1,80
1	6,5	2	5,18	-3,18	1,49	0,51
1	10,0	10	6,22	3,78	9,44	0,56
1	14,0	18	13,31	4,69	20,25	-2,25
1	21,5	21	21,90	-0,90	23,97	-2,97
1	29,0	27	31,39	-4,39	30,24	-3,24

(*) 0 - padrão , 1 - insulina . ◊

APÊNDICE 1

Programa ESTCOND.K22

ESTCOND.K22 é uma macro elaborada no módulo CM (Cálculo de Matrizes) do pacote científico SOC. Este programa tem como objetivo principal encontrar a estimativa de MV condicional da razão de chances (ψ) comum as k tabelas 2x2. A estimativa referida faz parte da teoria apresentada na Seção 1.2, sendo a solução do processo iterativo (1.26).

Para execução do ESTCOND.K22 só é necessário a definição da matriz **TC**. Essa matriz de ordem kx4 será formada pelos valores observados nas k tabelas 2x2. De acordo com os valores apresentados na Tabela 1.3, a i-ésima linha da matriz **TC** deverá respeitar a seguinte ordem : $(y_{1i},\ n_{1i} - y_{1i},\ y_{2i},\ n_{2i} - y_{2i})$, $i = 1, \ldots, k$.

A saída deste programa fornece, além da estimativa de MV de ψ, as seguintes estimativas : Mantel-Haenszel, Woolf e de MV não condicional dos ψ_i's, expressas, respectivamente, em (1.24), (1.25) e (1.20). Estas estimativas citadas no Exemplo 1.2, apresentadas na Tabela 1.7, podem ser obtidas atráves do seguinte programa :

```
cm
anote " ----- DADOS DE INNES ET AL. (1969) -----";
anote " ";
anote " Estudo Prospectivo - Exemplo 1.2";
anote " ";
anote " ";
TC = { 4 12 5 74 , 2 14 3 84 , 4 14 10 80 , 1 14 3 79 } ;
exec "a:estcond.k22";
fim ;
```

Listagem do Programa ESTCOND.K22 :

```
/* ------ Programa ESTCOND.K22 / Versao 1.0 - Nov.91 --------- */

/*     Este programa estima a razao de chances comum a k tabelas
   2x2 via estimacao de maxima verossimilhanca (MV) condicional
   pelo processo iterativo de Newton-Raphson com procedimentos
   sugeridos por McCullagh e Nelder (1983, pg.89-90), e ainda
   fornece as estimativas: Mantel-Haenszel, Woolf e MV nao
   condicional para o vetor das razoes de chances.              */

/*  ------------------------------------------------------------ */

/*  Matrizes auxiliares (uso interno).                           */

K = nlin(TC) ;  aux1 = vdiag(ident(k)) ;
N1 = TC[,1] + TC[,2] ;
N2 = TC[,3] + TC[,4] ;
M = TC[,1] + TC[,3] ;
N = N1 + N2 ;
YT = soma(TC[,1]) ;

/*     Estimativas inicias : media, variancia e P definidas no
   processo iterativo, denotadas, respectivamente, por  U0, V0
   e P0; Estimativas de Mantel-Haenszel (PMH), Woolf (PW) e  o
   vetor estimativa de MV nao condicional (PNC).                */

PNC = (TC[,1]*!TC[,4])/!(TC[,3]*!TC[,2]) ;
aux2 = (TC[,3]*!TC[,2])/!N ;
PMH = soma(aux2*!PNC)/soma(aux2) ;
P0 = PMH ;
aux3 = aux1/!( (aux1/!TC[,1]) + (aux1/!TC[,2]) + (aux1/!TC[,3])
       + (aux1/!TC[,4]) ) ;
PW = exp(soma(aux3*!log(PNC))/soma(aux3) ) ;
V0 = (N1*!N2*!M*!(N-M))/!(N*!N*!(N-aux1)) ;
aux4 = N2 - M + (N1 + M)*!(aux1*P0) ;
aux5 = V0 - (N1*!M + V0)*!(aux1*P0) ;
U0 = (-aux4+(aux4#2 -(aux1*4*(1-P0))*!aux5)#.5)/!(aux1*2*(1-P0));

/* Metodo iterativo para razao de chances comum (P).            */

INA = P0 ;
DELTA = (YT - soma(U0))/soma(V0) ;
P = exp(log(INA) + DELTA) ;
R=0 ;
U = aux1 ; V = aux1 ;
enquanto ( (abs(DELTA) > 0.0001)&&(R<=25) ) {
  aux4 = N2 - M + (N1 + M)*!(aux1*P) ;
  U = (-aux4 + (aux4#2 - (aux1*4*(1-P))*!(-N1*!M*P))#.5)/!(aux1*
      2*(1-P)) ;
  V = aux1/!( (aux1/!U) + (aux1/!(N1 - U)) + (aux1/!(M - U)) +
      (aux1/!(N2 - M - U)) ) ;
  DELTA = (YT - soma(U))/soma(V) ;
  P = exp(log(P) + DELTA) ;
  R=R+1 ;   }
```

```
/* ---------------------- Saida do Programa ------------------- */

anote "-------------------------------------------------------";
anote " ";
anote " ";
anote "                PROGRAMA ESTCOND.K22 ";
anote " ";
anote " Estimacao da Razao de Chances em k Tabelas 2x2";
anote " ";
anote " ";
se ( K == 1 )  {
  anote " ";
  anote "  Situacao : Uma unica tabela 2x2";
  anote "  Notacao de TC : (y\x)";
  anote " ";
  imprime TC $ ("(1\1),(0\1),(1\0),(0\0)") ;
  anote " ";
  anote "      Estimativa de maxima verossimilhanca condicional";
  anote "(PC) e nao condicional (PNC) e numero de iteracoes do ";
  anote "processo iterativo da razao de chances.";
  anote " ";
  S = { PNC P R };
  imprime S $ 15:3 ("PNC,PC,R");
}
se ( K > 2 )  {
  anote " ";
  anote " Situacao : K tabelas 2x2";
  anote " Notacao de TC : colunas (y\x) e linhas (estrato i)";
  anote " ";
  imprime TC $ ("(1\1),(0\1),(1\0),(0\0)") ;
  anote " ";
  anote "    Vetor de estimativa de maxima verossimilhanca (MV)";
  anote "nao condicional (PNC)";
  anote " ";
  S = PNC' ;
  imprime S $ 15:3 ;
  anote " ";
  anote "    Estimativas : Mantel-Haenszel (PMH), Woolf (PW) e";
  anote "MV condicional (PC) e numero de iteracoes do processo";
  anote "iterativo da razao de chances.";
  anote " ";
  S1 = { PMH PW P R };
  imprime S1 $ 15:3 ("PMH,PW,PC,R");
}
anote " ";
anote "       Fim do programa ESTCOND.K22";

/* ------------- Fim do programa ESTCOND.K22 --------------- */
```

APÊNDICE 2

Dados do Processo Infeccioso Pulmonar

PIP	IDA	SEX	HL	FF	PIP	IDA	SEX	HL	FF	PIP	IDA	SEX	HL	FF
0	26	1	1	0	0	21	1	1	0	0	45	1	1	1
0	19	0	1	1	0	16	0	1	1	0	43	1	1	0
0	72	0	1	1	0	53	1	1	0	0	33	1	1	0
0	39	1	1	0	0	41	1	1	0	0	49	1	1	0
0	26	0	1	1	0	27	0	1	1	0	46	1	1	0
0	27	1	1	1	0	65	1	1	0	0	55	1	1	0
0	27	1	1	0	0	32	0	1	0	0	22	0	1	0
0	23	1	1	0	0	42	0	1	0	0	34	1	1	0
0	82	0	1	0	0	23	0	1	0	0	18	1	1	0
0	20	0	0	0	0	23	1	1	0	0	28	1	1	0
0	22	1	1	0	0	50	0	1	1	0	64	1	0	0
0	29	0	1	0	0	24	0	1	1	0	59	1	1	1
0	50	1	0	1	0	38	1	0	1	0	20	0	1	1
0	43	1	1	1	0	27	1	1	1	0	20	1	0	0
0	24	1	1	0	0	46	1	1	1	0	44	1	1	1
0	40	1	0	1	0	21	0	1	1	0	21	1	1	0
0	42	1	1	1	0	23	1	1	1	0	43	1	1	0
0	38	0	1	1	0	53	1	1	1	0	53	1	1	0
0	21	1	0	0	0	57	1	1	0	0	36	1	0	0
0	63	0	1	0	0	21	1	1	0	0	45	1	1	0
0	77	1	0	0	0	58	1	1	0	0	18	0	1	1

PIP	IDA	SEX	HL	FF	PIP	IDA	SEX	HL	FF	PIP	IDA	SEX	HL	FF
0	22	1	1	1	0	30	1	1	1	0	46	1	0	0
0	78	0	1	1	0	23	1	1	1	0	64	1	0	1
0	56	1	0	0	0	56	1	1	1	0	44	1	0	0
0	23	1	0	0	0	62	1	0	0	0	53	1	0	0
0	23	1	1	0	0	23	1	1	0	0	63	1	1	0
0	49	1	1	0	0	21	1	1	1	0	17	1	1	1
0	41	0	1	1	0	45	1	1	1	0	49	1	0	0
0	51	1	0	1	0	62	0	1	0	0	48	1	0	0
0	27	1	1	0	0	18	0	1	0	0	33	0	1	0
0	67	1	1	1	0	75	1	0	1	0	67	1	0	0
0	87	1	0	0	0	53	0	0	0	0	18	1	0	1
0	30	1	1	1	0	48	1	1	1	1	58	1	1	1
0	31	0	1	1	0	56	1	0	0	0	48	1	1	0
1	76	1	1	1	1	64	1	0	1	1	44	1	0	0
1	34	1	0	0	1	51	1	1	1	1	45	1	0	0
1	60	1	0	0	1	73	1	1	0	1	72	1	1	1
1	62	1	0	0	1	60	0	0	0	1	58	0	0	0
1	43	1	0	0	1	62	1	0	0	1	55	1	0	0
1	58	1	0	0	1	15	1	0	0	1	61	0	0	0
1	60	1	0	0	1	61	0	1	0	1	56	0	1	0
1	56	0	0	0	1	78	1	0	0	1	21	1	0	0
1	75	1	0	0	1	56	0	0	0	1	56	1	0	0
1	73	0	0	0	1	56	1	0	0	1	62	0	1	0
1	52	0	0	0	1	57	0	0	0	1	29	0	0	0
1	51	1	0	0	1	77	1	1	1	1	73	1	0	0
1	40	0	1	0	1	65	1	0	0	1	60	1	0	0
1	69	0	1	0	1	67	1	1	1	1	57	0	0	0
1	58	1	0	0	1	72	1	1	0	1	51	0	0	0
1	57	1	1	0	1	36	0	0	0	1	70	1	0	0
1	58	1	1	1	1	59	1	1	0	1	59	1	0	0
1	69	1	0	1	1	61	1	1	0	1	67	1	0	0
1	69	1	1	1	1	52	1	0	0	1	59	1	1	0
1	50	0	0	0	1	48	1	0	0	1	64	1	0	0
1	49	1	0	0	1	78	1	0	0	1	66	1	0	0
1	74	1	0	0	1	50	1	0	0	1	57	0	1	0
1	75	0	0	0	1	55	0	1	1	1	50	0	0	0
1	70	0	0	0										

Obs : As siglas referentes as variáveis encontram-se definidas na Tabela 2.3

APÊNDICE 3

Programas : MRLG-ST.AJU e MRLG-ST.DOS

Os programas MRLG-ST.AJU e MRLG-ST.DOS também são macros elaboradas no módulo CM do SOC (1988), que objetivam implementar os **MRLG**'s de Stukel desenvolvidos na Seção 3.6.

O programa MRLG-ST.AJU ajusta o **MRLG** de Stukel e também apresenta como alternativa dois dos seus sub-modelos (Stukel-α_1 e Stukel-α_2) . Este está dividido em três partes : a primeira e a terceira são referentes ao ajuste do modelo e estimação dos parâmetros nos modelos logístico padrão e Stukel selecionado, respectivamente, e a segunda elabora um teste de *score* para avaliação do melhoramento do modelo de Stukel face ao modelo logístico padrão.

O programa MRLG-ST.DOS está confinado a dados de dose-resposta, objetivando a estimação de doses letais, e só poderá ser utilizado após a execução do MRLG-ST.AJU.

O método de estimação usado nos programas é o de máxima verossimilhança, e no primeiro deles os cálculos foram obtidos através do algoritmo de Newton-Raphson.

Para execução do MRLG-ST.AJU são necessários os seguintes dados :

Y : vetor de observações (nx1) ;

N : vetor dos totais amostrais (nx1) ;

X : matriz de especificação do modelo (nxp) ;

m = 1 : modelo Stukel-α_1 é selecionado ;

m = 2 : modelo Stukel-α_2 é selecionado ;

m $\neq$ 1 e 2 : modelo Stukel-(α_1, α_2) é selecionado ;

mvb = 1 : solicita impressão da estimativa da matriz de covariância dos parâmetros estimados no modelo logístico padrão; caso contrário mvb deverá ser diferente de 1 ;

re = 1 : os resíduos dos modelos selecionados serão apresentados na saída do programa ; caso contrário m $\neq$ 1 ;

a : coeficiente de confiança a usar nos intervalos de confiança apresentados na saída do programa;

mvt = 1 : solicita impressão da estimativa da matriz de covariância dos parâmetros estimados no modelo de Stukel selecionado; caso contrário mvt deverá ser diferente de 1 ;

mv = 1 : proporciona a impressão dos valores obtidos para a função log-verossimilhança nos modelos considerados e no saturado ;

ic = 1 : intervalos de confiança para nπ nos modelos considerados .

Para execução do MRLG-ST.DOS é necessário definir somente a probabilidade associada a dose letal a estimar, por exemplo, **pd = 0.5** para estimação da dose letal mediana.

As saídas dos programas oferecem uma descrição detalhada dos vetores e matrizes usados na análise dos modelos logístico e Stukel selecionado, evitando assim uma apresentação antecipada.

Como exemplo do uso destes programas apresentamos a seguir um programa de comandos, cuja saída proporcionou a análise do Exemplo 3.1 :

```
cm
/*          Arquivo de dados                              */

anote " ";
anote "       Os dados para analise sao de BLISS (1935). ";
anote " ";

/*          Conjunto de Dados .                           */

Y = {  6 13 18 28 52 53 61 60 }' ;
N = { 59 60 62 56 63 59 62 60 }' ;
X = { 1 1.6907 , 1 1.7242 , 1 1.7552 , 1 1.7842 , 1 1.8113 ,
      1 1.8369 , 1 1.861  , 1 1.8839 } ;

/*        Execucao dos programas de referencia.           */

m=1; mvb=1; re=1; a=95; mvt=1; mv=1; ic=1; exec "a:mrlg-st.aju" ;

pd=0.001; exec "a:mrlg-st.dos" ;
pd=0.005; exec "a:mrlg-st.dos" ;
pd=0.01; exec "a:mrlg-st.dos" ;
pd=0.05; exec "a:mrlg-st.dos" ;
pd=0.2; exec "a:mrlg-st.dos" ;
pd=0.5; exec "a:mrlg-st.dos" ;
pd=0.8; exec "a:mrlg-st.dos" ;
pd=0.95; exec "a:mrlg-st.dos" ;
pd=0.99; exec "a:mrlg-st.dos" ;
pd=0.995; exec "a:mrlg-st.dos" ;
pd=0.999; exec "a:mrlg-st.dos" ;

fim;
```

Listagem do Programa MRLG-ST.AJU :

```
/*  ------- Programa MRGL-ST.AJU - versao abril/92 ----------- */

/*      Este programa ajusta o modelo de regressao logistica
    generalizado de Stukel (1988) via estimacao de maxima
    verossimilhanca pelo processo iterativo de Newton-Raphson.

/*  ------------------------------------------------------------ */

/*  Matrizes auxiliares no programa (uso interno).              */

k=NLIN(Y) ; p=NCOL(X) ; UM = vdiag(ident(k)) ; aux01=UM;
aux02=UM; aux41=UM; aux42=UM ; Y1 = UM ; PI = UM ; H = UM ;
U0 = vdiag(ident(p)) ;
U1 = vdiag(ident(p+1)) ; U2 = vdiag(ident(p+2)) ; I0 = ident(p) ;
I1 = ident(p+1) ; I2 = ident(p+2) ;
D1 = UM*U1' ;
D2 = UM*U2' ;     /* notacao das derivadas da funcao h(eta) em
             relacao aos elementos do vetor parametrico theta. */

/* -------------------- PARTE 1 ------------------------------ */

/*   Metodo iterativo para estimativa do vetor parametrico no
   modelo logistico (B).                                        */

ina0 = U0*0 ;                                  /* valores iniciais */
delta0 = vdiag(ident(p)) ;
B = ina0 ;
r=0 ;
enquanto ( ( max(abs(delta0)) > 0.0001)&&(r<=25) ) {
                                              /* criterio de parada */
  P0 = UM/!(UM + exp(-X*B)) ;       /* probabilidade de sucesso */
  U0 = X'*(Y - N*!P0) ;                          /* funcao score */
  I0 = X'*diag(N*!P0*!(UM - P0))*X ;/* matriz de informacao de
                                                        Fisher */
  delta0 = inv(I0)*U0 ;
  B = B + delta0 ;                          /* vetor parametrico */
  r=r+1 ;   }
rl=r;              /* numero de iteracoes para estimacao de B */

/*   Estimativa do vetor das probabilidades de sucesso (P0) e
   da matriz de covariancia dos coeficientes estimados no
   modelo logistico (VB).                                       */

E = X*B ;                          /* preditor linear do logistico */
P0 = UM/!(UM + exp(-E)) ;
VB = inv(X'*diag(N*!P0*!(UM-P0))*X) ; /* matriz de covariancia
                                    de B estimado no logistico */
```

```
/*   Intervalo de 100(1-a)% de confianca para a equacao de
   regressao (n*pi_i) no modelo logistico.                    */

aux0 = ninv((1-(a/100))/2);
s=1;
enquanto ( s <= k ) {
  aux01[s,] = X[s,]*B - aux0*(sqrt(X[s,]*VB*X[s,]')) ;
  aux02[s,] = X[s,]*B + aux0*(sqrt(X[s,]*VB*X[s,]')) ;
  s=s+1 ; }
ICL1 = N*!(UM/!(UM+exp(-aux01)));          /* limite inferior */
ICL2 = N*!(UM/!(UM+exp(-aux02)));          /* limite superior */

/*  Teste de ajustamento do modelo logistico.                 */

L0 = Y'*log(P0) + (N-Y)'*log(UM - P0) ; /* funcao log-verossi-
                                  milhanca estimada no logistico */
s=1;        /* correcao para eliminar prob. observada = 1 ou 0 */
enquanto ( s <= k ) {
  se ( (Y[s,] == 0)||(Y[s,] == N[s,]) ) {
    se ( Y[s,] == 0 ) Y1[s,] = 1/N[s,] ;
    se ( Y[s,] == N[s,] ) Y1[s,] = Y[s,] - 1/N[s,] ;  }
  cc { Y1[s,] = Y[s,] ;   }
  s=s+1; }
LS = Y1'*log(Y1/!N) + (N-Y1)'*log(UM - (Y1/!N)) ;    /* funcao
                                  log-verossimilhanca no saturado */
EAL = 2*(LS - L0);      /* estatistica do teste de RV de Wilks */


/* -------------------- PARTE 2 ------------------------------ */

/*    Calculos do teste score do modelo logistico no modelo
   logistico generalizado de Stukel.                          */

s=1;
enquanto ( s<=k ) {  D2[s,1:p] = X[s,] ;
  se ( E[s,] >= 0 ) { D2[s,p+1] = 0.5*(E[s,]#2); D2[s,p+2]=0; }
  cc { D2[s,p+1]=0; D2[s,p+2] = -0.5*(E[s,]#2); }
  s=s+1 ;  }
U2 = { D2[,p+1] D2[,p+2] }'*(Y - N*!P0) ;  /* funcao score sob
               modelo logistico no modelo Stukel-alpha1,alpha2 */
U21 = D2[,p+1]'*(Y - N*!P0) ;       /* idem para Stukel-alpha1 */
U22 = D2[,p+2]'*(Y - N*!P0) ;       /* idem para Stukel-alpha2 */
aux1 = D2'*diag(N*!P0*!(UM-P0))*D2 ;
aux11 = {D2[,1:p] D2[,p+1]}'*diag(N*!P0*!(UM-P0))*{D2[,1:p]
        D2[,p+1]} ;
aux12 = {D2[,1:p] D2[,p+2]}'*diag(N*!P0*!(UM-P0))*{D2[,1:p]
        D2[,p+2]} ;
aux2 = aux1[p+1 p+2,1:p]*inv(aux1[1:p,1:p])*aux1[1:p,p+1 p+2] ;
aux21 = aux11[p+1,1:p]*inv(aux11[1:p,1:p])*aux11[1:p,p+1] ;
aux22 = aux12[p+1,1:p]*inv(aux12[1:p,1:p])*aux12[1:p,p+1] ;
aux3 = aux1[p+1 p+2 , p+1 p+2] - aux2 ;
aux31 = aux11[p+1,p+1] - aux21 ;
aux32 = aux12[p+1,p+1] - aux22 ;
E1 = U2'*inv(aux3)*U2 ;       /* estatistica do teste score no
                                    modelo Stukel-alpha1,alpha2 */
E11 = U21'*inv(aux31)*U21 ;         /* idem para Stukel-alpha1 */
E12 = U22'*inv(aux32)*U22 ;         /* idem para Stukel-alpha2 */
```

```
/* -------------------- PARTE 3 ------------------------------ */

/*   Metodo iterativo para estimativa do vetor parametrico (T)
   no modelo de regressao logistica generalizado de Stukel.    */

ina = {B,0,0} ;                                  /* valores iniciais */
delta = vdiag(ident(p+2)) ;
T = ina ;
r=0 ;
enquanto ( ( max(abs(delta)) > 0.0001)&&(r<=25) ) {
                                                 /* criterio de parada */
E=X*T[1:p,];
s=1;
enquanto ( s<=k ) {
  se ( E[s,] >= 0 ) {   D2[s,p+2] = 0 ;
    se ( T[p+1,] > 0 ) {
      H[s,] = (exp(E[s,]*T[p+1,])-1)/T[p+1,] ; /*funcao h(eta) */
      D2[s,1:p] = exp(T[p+1,]*E[s,])*X[s,] ;
      D2[s,p+1] = (T[p+1,]*E[s,]*exp(T[p+1,]*E[s,]) -
                  exp(T[p+1,]*E[s,]) + 1)/(T[p+1,]#2) ;   }
    se ( T[p+1,] == 0 ) {
      H[s,] = E[s,] ;
      D2[s,1:p] = X[s,] ;
      D2[s,p+1] = 0.5*(E[s,]#2) ;    }
    se ( T[p+1,] < 0 ) {
      H[s,] = -log(1-E[s,]*T[p+1,])/T[p+1,] ;
      D2[s,1:p] = (1/(1-T[p+1,]*E[s,]))*X[s,] ;
      D2[s,p+1] = (log(1-T[p+1,]*E[s,])+((T[p+1,]*E[s,])/
                  (1-T[p+1,]*E[s,])))/(T[p+1,]#2) ;   }
  }
  se ( E[s,] < 0 ) {   D2[s,p+1] = 0 ;
    se ( T[p+2,] > 0 ) {
      H[s,] = -(exp(-E[s,]*T[p+2,]) - 1)/T[p+2,] ;
      D2[s,1:p] = exp(-T[p+2,]*E[s,])*X[s,] ;
      D2[s,p+2] = (T[p+2,]*E[s,]*exp(-T[p+2,]*E[s,]) +
                  exp(-T[p+2,]*E[s,]) - 1)/(T[p+2,]#2) ;   }
    se ( T[p+2,] == 0 ) {
      H[s,] = E[s,] ;
      D2[s,1:p] = X[s,] ;
      D2[s,p+2] = -0.5*(E[s,]#2) ;    }
    se ( T[p+2,] < 0 ) {
      H[s,] = log(1+E[s,]*T[p+2,])/T[p+2,] ;
      D2[s,1:p] = (1/(1+T[p+2,]*E[s,]))*X[s,] ;
      D2[s,p+2] = (-log(1+T[p+2,]*E[s,])+((T[p+2,]*E[s,])/
                  (1+T[p+2,]*E[s,])))/(T[p+2,]#2) ;   }
   }
  s=s+1; }
PI = UM/!(UM+exp(-H)) ;              /* probabilidade de sucesso */
U2 = D2'*(Y-N*!PI) ;                                /* funcao score */
I2 = D2'*diag(N*!PI*!(UM-PI))*D2 ;  /* matriz de informacao de
                                                        Fisher */
delta = inv(I2)*U2 ;                          /* vetor parametrico */
T = T + delta ;
r=r+1 ; }
rs=r;                /* numero de iteracoes para estimacao de T */
```

```
/*   Metodo iterativo para estimativa do vetor parametrico (T)
   no modelo de regressao logistica generalizado Stukel-alpha1
   e Stukel-alpha2.                                            */

se ( (m == 1)||(m == 2)  ) {
  ina1 = { B , 0 } ;   /* notacao analoga ao modelo de Stukel-
                                                  alpha1,alpha2 */
  delta1 = vdiag(ident(p+1)) ;
  T1 = ina1 ;
  r=0 ;
  enquanto ( ( max(abs(delta1)) > 0.0001)&&(r<=25) ) {
    E=X*T1[1:p,];
    s=1;
    enquanto ( s<=k ) {
      se ( ((m==1)&&(E[s,]>=0))||((m==2)&&(E[s,]<0)) ) {
        se ( T1[p+1,] > 0 ) {
          H[s,] = ((-1)#(m+1))*(exp(((-1)#(m+1))*E[s,]*T1[p+1,])
                  - 1)/T1[p+1,] ;
          D1[s,1:p] = exp(((-1)#(m+1))*T1[p+1,]*E[s,])*X[s,] ;
          D1[s,p+1] = (T1[p+1,]*E[s,]*exp(((-1)#(m+1))*T1[p+1,]*
              E[s,]) + ((-1)#m)*exp(((-1)#(m+1))*T1[p+1,]*E[s,])
              + ((-1)#(m+1)))/(T1[p+1,]#2) ;   }
        se ( T1[p+1,] == 0 ) {
          H[s,] = E[s,] ;
          D1[s,1:p] = X[s,] ;
          D1[s,p+1] = ((-1)#(m+1))*0.5*(E[s,]#2) ;    }
        se ( T1[p+1,] < 0 ) {
          H[s,] = ((-1)#m)*log(1+((-1)#m)*E[s,]*T1[p+1,])/T1[p+1,];
          D1[s,1:p] = (1/(1+((-1)#m)*T1[p+1,]*E[s,]))*X[s,] ;
          D1[s,p+1] = (((-1)#(m+1))*log(1+((-1)#m)*T1[p+1,]*E[s,])
              +((T1[p+1,]*E[s,])/(1+((-1)#m)*T1[p+1,]*E[s,])))/
              (T1[p+1,]#2) ;   }   }
      cc {  H[s,] = E[s,] ; D1[s,1:p] = X[s,] ; D1[s,p+1] = 0 ; }
     s=s+1; }
    PI = UM/!(UM + exp(-H)) ;
    U1 = D1'*(Y-N*!PI) ;
    I1 = D1'*diag(N*!PI*!(UM-PI))*D1 ;
    delta1 = inv(I1)*U1 ;
    T1 = T1 + delta1 ;
  r=r+1 ; }
rs = r ; }

/*   Calculo da estimativa da matriz de covariancia de theta
                                                 estimado (VT).  */

se ( (m==1)||(m==2) ) {
  se (m==1) T2 = { T1 , 0 } ;
  se (m==2) T2 = { T1[1:p,] , 0 , T1[p+1,] } ;   }
cc { T2 = T ;  }
E=X*T2[1:p,];
s=1;
enquanto ( s<=k ) {
  se ( E[s,] >= 0 ) {   D2[s,p+2] = 0 ;
    se ( T2[p+1,] > 0 ) {
      H[s,] = (exp(E[s,]*T2[p+1,]) - 1)/T2[p+1,] ;
```

```
      D2[s,1:p] = exp(T2[p+1,]*E[s,])*X[s,] ;
      D2[s,p+1] = (T2[p+1,]*E[s,]*exp(T2[p+1,]*E[s,]) -
                  exp(T2[p+1,]*E[s,]) + 1)/(T2[p+1,]#2) ;   }
    se ( T2[p+1,] == 0 ) {
      H[s,] = E[s,] ;
      D2[s,1:p] = X[s,] ;
      D2[s,p+1] = 0.5*(E[s,]#2) ;    }
    se ( T2[p+1,] < 0 ) {
      H[s,] = -log(1-E[s,]*T2[p+1,])/T2[p+1,] ;
      D2[s,1:p] = (1/(1-T2[p+1,]*E[s,]))*X[s,] ;
      D2[s,p+1] = (log(1-T2[p+1,]*E[s,])+((T2[p+1,]*E[s,])/
                  (1-T2[p+1,]*E[s,])))/(T2[p+1,]#2) ;   }
      }
    se ( E[s,] < 0 ) {   D2[s,p+1] = 0 ;
      se ( T2[p+2,] > 0 ) {
        H[s,] = -(exp(-E[s,]*T2[p+2,]) - 1)/T2[p+2,] ;
        D2[s,1:p] = exp(-T2[p+2,]*E[s,])*X[s,] ;
        D2[s,p+2] = (T2[p+2,]*E[s,]*exp(-T2[p+2,]*E[s,]) +
            exp(-T2[p+2,]*E[s,]) - 1)/(T2[p+2,]#2) ;   }
      se ( T2[p+2,] == 0 ) {
        H[s,] = E[s,] ;
        D2[s,1:p] = X[s,] ;
        D2[s,p+2] = -0.5*(E[s,]#2) ;    }
      se ( T2[p+2,] < 0 ) {
        H[s,] = log(1+E[s,]*T2[p+2,])/T2[p+2,] ;
        D2[s,1:p] = (1/(1+T2[p+2,]*E[s,]))*X[s,] ;
        D2[s,p+2] = (-log(1+T2[p+2,]*E[s,])+((T2[p+2,]*E[s,])/
            (1+T2[p+2,]*E[s,])))/(T2[p+2,]#2) ;   }
      }
   s=s+1; }
PI = UM/!(UM + exp(-H)) ;
  se ( (m==1)||(m==2) ) {   /* calculo da matriz informacao de
                                                      Fisher */
    se (m==1) aux43 = { D2[,1:p] D2[,p+1] } ;
    se (m==2) aux43 = { D2[,1:p] D2[,p+2] } ; }
  cc { aux43 = D2 ;  }
I2 = aux43'*diag(N*!PI*!(UM-PI))*aux43 ;
VT = inv(I2) ;

/*   Intervalo de 100(1-a)% de confianca para a equacao de
   regressao (n*pi_i) no modelo de regressao logistica
   generalizado de Stukel.                                  */

s=1;
enquanto ( s <= k ) {
  aux41[s,] = H[s,] - aux0*(sqrt(aux43[s,]*VT*aux43[s,]')) ;
  aux42[s,] = H[s,] + aux0*(sqrt(aux43[s,]*VT*aux43[s,]')) ;
  s=s+1 ; }
ICS1 = N*!(UM/!(UM+exp(-aux41)));          /* limite inferior */
ICS2 = N*!(UM/!(UM+exp(-aux42)));          /* limite superior */
```

```
/*  Teste de ajustamento do modelo logistico generalizado de
   Stukel.                                                    */

L = Y'*log(PI) + (N-Y)'*log(UM - PI) ;
EAS = 2*(LS - L);       /* estatistica do teste de RV de Wilks  */

/* -------------- Saida do programa MRLG-ST.AJU -------------- */

anote "-----------------------------------------------------";
anote " ";
anote " ";
anote "                PROGRAMA MRLG-ST.AJU ";
anote " ";
anote " Modelo de Regressao Logistica Generalizado de Stukel";
anote " ";
anote " ";
anote "Parte 1 : Modelo de Regressao Logistica ";
anote " ";
anote "      Vetor Parametrico : Beta ";
anote "      Estimativa de MV de Beta :";
anote " ";
imprime B 10:3 ("Beta") ;
anote " ";
anote " ";
se ( mvb == 1 ) {
anote "      Estimativa da Matriz de Covariancia de Beta^";
anote " ";
imprime VB 10:3 ;
anote " ";
anote " "; }
anote "    Teste de Ajustamento do Modelo Logistico";
anote " ";
sai1 = { EAL (k-p) xprob(EAL,k-p) } ;
imprime sai1 $ 10:3 ("Estatist.,G.L.,Nivel Desc.") ;
anote " ";
anote " ";
se ( re == 1 ) {
  sai2 = { Y N (N*!P0) (Y-N*!P0) } ;
  imprime sai2 $ 10:3 ("Observado,Total,Ajustado,Residuo") ;
anote " ";
anote " "; }
anote "-----------------------------------------------------";
anote " ";
anote "Parte 2 : Testes score do modelo logistico em alguns";
anote "          modelos logistico generalizados de Stukel.";
anote " ";
anote " ";
anote "    Teste score no modelo Stukel-alpha1,alpha2";
anote " ";
```

```
anote " ";
anote "    Teste score no modelo Stukel-alpha1";
anote " ";
sai4 = { E11 1 xprob(E11,1) } ;
imprime sai4 $ 10:3 ("Estatist.,G.L.,Nivel Desc.") ;
anote " ";
anote " ";
anote "    Teste score no modelo Stukel-alpha2";
anote " ";
sai5 = { E12 1 xprob(E12,1) } ;
imprime sai5 $ 10:3 ("Estatist.,G.L.,Nivel Desc.") ;
anote " ";
anote " ";
anote "-------------------------------------------------------";
anote " ";
anote "Parte 3 : Modelo de Regressao Logistica Generalizado";
  se ( (m==1)||(m==2) ) {
     se (m==1) { anote "              Stukel-alpha1"; }
     se (m==2) { anote "              Stukel-alpha2"; } }
  cc { anote "             Stukel-alpha1,alpha2"; }
anote " ";
anote " ";
anote "     Vetor Parametrico : Theta ";
anote "     Estimativa de MV de Theta :";
anote " ";
  se ( (m==1)||(m==2) ) imprime T1 10:3 ("Theta") ;
  cc imprime T 10:3 ("Theta") ;
anote " ";
anote " ";
se ( mvt == 1 ) {
anote "     Estimativa da Matriz de Covariancia de Theta^";
anote " ";
imprime VT 10:3 ;
anote " ";
anote " "; }
anote "    Teste de Ajustamento do Modelo de Stukel";
anote " ";
sai6 = { EAS (k-nlin(VT)) xprob(EAS,k-nlin(VT)) } ;
imprime sai6 $ 10:3 ("Estatist.,G.L.,Nivel Desc.") ;
anote " ";
anote " ";
se ( re == 1 ) {
  sai7 = { Y N (N*!PI) (Y-N*!PI) } ;
  imprime sai7 $ 10:3 ("Observado,Total,Ajustado,Residuo") ;
anote " ";
anote " "; }
se ( mv == 1 ) {
anote " ";
```

```
anote "Valor da funcao log-verossimilhanca nos modelos :";
anote " ";
sai8={ LS L0 L };
imprime sai8 $ 10:3 ("Saturado,Logistico,Stukel");
anote " ";
anote " ";
anote "Numero de iteracoes nos metodos iterativos dos modelos :";
anote " ";
sai9 = { RL RS } ;
imprime sai9 $ 10:3 ("Logistico,Stukel");     }
se ( ic == 1 ) {
anote " ";
anote " ";
anote "  Intervalo de 100(1-a)% de Confianca para n*Pi_i no";
anote " Modelo Logistico e Logistico Generalizado de Stukel.";
anote " ";
sai10=a/100;
imprime sai10 $ ("Coef.Conf.");
anote " ";
anote "  Notacao : IC (limites de confianca do Modelo Logistico";
anote "             [L] e do modelo de R.L.G. de Stukel [S].";
anote " ";
sai11={ Y ICL1 ICL2 ICS1 ICS2 };
imprime sai11 $ 10:2 ("Observado,ICInf L,ICSup L,ICInf S,ICSup S");}
anote " ";
anote "      Fim do programa MRLG-ST.AJU ";

/*  ------------ Fim do programa MRLG-ST.AJU ----------------- */
```

Listagem do Programa MRLG-ST.DOS :

```
/* --------- Programa MRGL-ST.DOS - versao abril/92 ---------- */

/*   Este programa estimativa a dose letal de 100P% de sucesso
   [DL_(100P)] no modelo de regressao logistica generalizada de
   Stukel via estimacao de maxima verossimilhanca.             */

/* ------------------------------------------------------------ */

/*  Calculos auxiliares (uso interno).                          */

LP = log(pd/(1-pd)) ;                        /* logit de DL_(100P) */
DE = 1 ;      /* preditor linear indexado pela probabilidade P */
DP1 = 1 ;                              /* derivada de DE em alpha1 */
DP2 = 1 ;                              /* derivada de DE em alpha2 */
VS = 1 ;/* variancia do estimador da DL_(100) no modelo stukel */

/*  Calculos iniciais para estimacao de DL_(100P).              */

se ( pd >= 0.5 ) { DP2 = 0 ;
  se ( T2[3,] > 0 ) { DE = log(1+T2[3,]*LP)/T2[3,] ;
    DP1 = ((T2[3,]*LP)/(1+T2[3,]*LP)-log(1+T2[3,]*LP))/
          (T2[3,]#2);}
  se ( T2[3,] == 0 ) { DE = LP ; DP1 = 0 ; }
  se ( T2[3,] < 0 ) { DE = (1-exp(-T2[3,]*LP))/T2[3,] ;
    DP1 = ((1+T2[3,]*LP)*(exp(-T2[3,]*LP))-1)/(T2[3,]#2);}
}
se ( pd < 0.5 ) { DP1 = 0 ;
  se ( T2[4,] > 0 ) { DE = -(log(1-T2[4,]*LP))/T2[4,] ;
    DP2 = ((T2[4,]*LP)/(1-T2[4,]*LP)+log(1-T2[4,]*LP))/
          (T2[4,]#2);}
  se ( T2[4,] == 0 ) { DE = LP ; DP2 = 0 ; }
  se ( T2[4,] < 0 ) { DE = -(1-exp(T2[4,]*LP))/T2[4,] ;
    DP2 = ((T2[4,]*LP-1)*(exp(T2[4,]*LP)) + 1)/(T2[4,]#2) ; }
}

/*  Estimacao de DL(100p).                                      */

EPDL = (LP - B[1,])/B[2,] ;   /* estimativa de MV de DL_(100P)
                                                    no logistico */
EPDS = (DE - T2[1,])/T2[2,] ;            /* estimativa de MV de
                                  DL_(100P) no modelo Stukel */
DPDL = {(-1/B[2,]) , (-(LP-B[1,])/(B[2,]#2))} ;    /* vetor de
                       derivadas DL_(100P) em beta (logistico) */
DPDS = {(-1/T2[2,]), (-(DE-T2[1,])/(T2[2,]#2)), (DP1/T2[2,]),
       (DP2/T2[2,])}; /* vetor de derivadas DL_(100P) em theta
                                 (modelo de Stukel-alpha1,alpha2) */
VL = DPDL'*VB*DPDL ;     /* variancia do estimador da DL_(100) */
  se ( (m==1)||(m==2) ) {
    se (m==1) VS = DPDS[1:3,]'*VT*DPDS[1:3,] ;
    se (m==2) VS = {DPDS[1:2,] , DPDS[4,]}'*VT*{DPDS[1:2,] ,
                   DPDS[4,]};  }
cc { VS = DPDS'*VT*DPDS ;  }
```

```
/* -------------- Saida do programa MRLG-ST.DOS -------------- */

anote "-----------------------------------------------------";
anote " ";
anote " ";
anote "                PROGRAMA MRLG-ST.DOS ";
anote " ";
anote "Estimacao de Dose Letal no Modelo de Regressao Logistica";
se ( (m==1)||(m==2) ) {
  se (m==1) {
anote "                 e Logistica Generalizado de Stukel-alpha1";}
  se (m==2) {
anote "                 e Logistica Generalizado de Stukel-alpha2";}}
  cc {
anote "        e Logistica Generalizado de Stukel-alpha1,alpha2";}
anote "              e os respectivos desvios padroes DPDL DPDS";
anote " ";
sai1 = { pd*100 EPDL sqrt(VL) EPDS sqrt(VS) } ;
imprime sai1 $ 10:3 ("Dose (P),Logistico,DPDL,Stukel,DPDS") ;
anote " ";
anote " ";
anote "   Intervalos de 95% de Confianca para a Dose Letal ";
anote " ";
sai2 = { (EPDL-aux0*sqrt(VL)) (EPDS-aux0*sqrt(VS)) ,
       (EPDL+aux0*sqrt(VL)) (EPDS+aux0*sqrt(VS)) } ;
imprime sai2 10:3 ("Logistico,Stukel");
anote " ";
anote " Fim do programa MRLG-ST.DOS ";

/*  ------------ Fim do programa MRLG-ST.DOS ---------------- */
```

REFERÊNCIAS BIBLIOGRÁFICAS

Abbott, R.D. (1985). Logistic regression in survival analysis. *American Journal of Epidemiology.* **121**, 465-471.

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley, New York.

Anderson, D.A. (1988). Some models for overdispersed binomial data. *The Australian Journal of Statistics.* **30**, 125-148.

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika.* **59**, 19-35.

Aranda-Ordaz, F.J. (1981). On two families of transformations to additivity for binary response data. *Biometrika.* **68**, 357-363.

Armitage, P. (1971). *Statistical Methods in Medical Research.* Blackwell Scientific Publications, Oxford.

Baker, R.J. and Nelder, J.A. (1978). The GLIM System-Release 3.77. Distributed by Numerical Algorithms Group, Oxford.

Barnard, G.A. (1979). In contradiction to J. Berkson's dispraise : conditional tests can be more efficient. *Journal of Statistical Planning and Inference.* **3**, 181-187.

Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association.* **72**, 355-366.

Basu, D. (1979). Discussion of Joseph Berkson's paper "In dispraise of the exact test". *Journal of Statistical Planning and Inference.* **3**, 189-192.

Berkson, J. (1978a). In dispraise of the exact test. *Journal of Statistical Planning and Inference.* **2**, 27-42.

Berkson, J. (1978b). Do the marginal totals of the 2x2 table contain relevant information respecting the table proportions ? *Journal of Statistical Planning and Inference.* **2**, 43-44.

Bickel, P.J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association.* **70**, 428-434.

Birch, M.W. (1964). The detection of partial association, I. The 2x2 case. *Journal of the Royal Statistical Society.* **B 26**, 313-324.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice.* MIT Press, Cambridge, MA.

Bliss, C.I. (1935). The calculation of the dosage-mortality curve. *Annals of Applied Biology.* **22**, 134-167.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society.* **B 26**, 211-252.

Box, G.E.P. and Tidwell, P.W. (1962). Transformation of the independent variables. *Technometrics.* 4, 531-550.

Breslow, N.E. (1981). Odds ratio estimators when the data are sparse. *Biometrika.* **68**, 73-84.

Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research, Vol.1: The Analysis of Case-Control Studies.* IARC, Lyon.

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests : An expository note. *The American Statistician.* **36**, 153-157.

Copenhaver, T.W. and Mielke, P.W. (1977). Quantit analysis : A quantal assay refinement. *Biometrics.* **33**, 175-186.

Cordeiro, M.G. e Paula, G.A. (1989). *Modelos de Regressão para Análise de Dados Univariados.* 17º Colóquio Brasileiro de Matemática - IMPA, Rio de Janeiro.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and crevix. *Journal of the National Cancer*

Institute. **11**, 1269-1275.

Cornfield, J. (1962). Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure : A discriminant function analysis. *Federation Proceendings.* **21**, 58-61.

Corsten, L.C.A. and de Kroon, J.P.M. (1979). Comment on J. Berkson's paper "In dispraise of the exact test". *Journal of Statistical Planning and Inference.* **3**, 193-197.

Cox, D.R. and Hinkley, D.V. (1974). *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. *Journal of the Royal Statistical Society.* **B 30**, 248-275.

Cox, D.R. and Snell, E.J. (1989). *The Analysis of Binary Data.* 2nd ed., Chapman and Hall, London.

Davis, L.J. (1986). Exact tests for 2x2 contingency tables. *The American Statistician.* **40**, 139-141.

Day, N.E. and 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. (1987). BMDP Statistical Software. University of California Press, Berkeley.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association.* **70**, 892-898.

Efron, B. (1988). Logistic regression, survival analysis and the Kaplan-Mayer curve. *Journal of the American Statistical Association.* **83**, 414-425.

El-Saidi, M.A. and George, E.O. (1990). A generalized logistic model for quantal response bioassay. *Biometrical Journal.* **32**, 943-954.

Farewell, V.T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika.* **66**, 27-32.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data.* 2nd ed., MIT Press, Cambridge, MA.

Finney, D.J. (1971). *Probit Analysis.* 3rd ed., Cambridge University Press, Cambridge.

Finney, D.J. (1978). *Statistical Method in Biological Assay.* 3rd ed., Charles Griffin and Co Ltd, London.

Fisher, R.A. (1934). *Statistical Methods for Research Workers.* 14th ed. (1970), Oliver and Boyd, Edinburgh.

Fisher, R.A. (1935). *The Design of Experiments.* 8th ed. (1966), Oliver and Boyd, Edinburgh.

Follmann, D.A. and Lambert, D. (1989). Generalizing logistic regression by non-parametric mixing. *Journal of the American Statistical Association.* **84**, 295-300.

Fowlkes, E.B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika.* **74**, 503-515.

Gail, M.H., Lubin, J.H. and Rubinstein, L.V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika.* **68**, 703-707.

Gart, J.J. (1962). On the combination of relative risks. *Biometrics.* **18**, 601-610.

Gart, J.J. (1971). The comparison of proportions : A review of significance tests, confidence intervals and adjustments for stratification. *Review of the International Statistical Institute.* **39**, 148-169.

Gart, J.J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2x2 tables. *Biometrika.* **72**, 673-677.

Goodman, L.A. and Kruskal, W.H. (1979). *Measures of Association for Cross Classification.* Springer, New York.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969). Analysis of categorical data by linear models. *Biometrics.* **25**, 489-504.

Guerrero, V.M. and Johnson, R.A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika.* **69**, 309-314.

Haber, M. (1989). Do the marginal totals of a 2x2 contingency table contain information regarding the table proportions ? *Communications in Statistics.* **A 18**, 147-156.

Hannan, J. and Harkness, W. (1963). Normal approximation to the distribution of two independent binomials, conditional on fixed sum. *Annals of Mathematical Statistics.* **34**, 1593-1595.

Hosmer, T., Hosmer, D.W. and Fisher, L.L. (1983). A comparison of the maximum likelihood and discriminant function estimators of the coefficientes of the logistic regression model for mixed continuous and discrete variables. *Communications in Statistics.* **B 12**, 577-593.

Hosmer, D.W. and Lemeshow, S. (1989). *Applied Logistic Regression.* John Wiley, New York.

Innes, J.R.M., Ulland, B.M., Valerio, M.G., Petrucelli, L., Fishbein, L., Hart, E.R., Pallotta, A.J., Bates, R.R., Falk, H.L., Gart, J.J., Klein, M., Mitchell, I. and Peters, J. (1969). Biossay of pesticides and industrial chemicals for tumorigenicity in mice: A preliminary note. *Journal of the National Cancer Institute.* **42**, 1101-1114.

Jorgensen, B. (1984). The delta algorithm and GLIM . *International Statistical Review.* **52**, 283-300.

Kempthorne, O. (1979). In dispraise of the exact test : Reactions. *Journal of Statistical Planning and Inference.* **3**, 199-213.

Kleinbaum, D.G., Kupper, L.L. and Chambless, L.E. (1982). Logistic regression analysis of epidemiologic data : theory and practice. *Communications in Statistics - Theory and Methods.* **11**, 485-547.

Kotz, S., Johnson, N.L. and Read, C.B. (1985). *Encyclopedia of Statistical Sciences.* Vol. 5, John Wiley, New York.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association.* **79**, 61-71.

Lee, E.T. (1980). *Statistical Methods for Survival Data Analysis.* Wadsworth, Belmont, CA.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses.* 2nd ed., John Wiley, New York.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute.* **22**, 719-748.

Mantel, N. and Hankey, B.F. (1975). The odds ratios of a 2x2 contingency table. *The American Statistician.* **29**, 143-145.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* 2nd ed., Chapman and Hall, London.

Mckinlay, S.M. (1975). The effect of bias on estimators of relative risk for pair-matched and stratified samples. *Journal of the American Statistical Association.* **70**, 859-864.

Mehta, C.R., Patel, N.R. and Gray, R. (1985). Computing an exact confidence interval for the common odds ratio in several 2x2 contingency tables. *Journal of the American Statistical Association.* **80**, 969-973.

Mehta, C.R., Patel, N.R. and Senchaudhuri, P. (1988). Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association.* **83**, 999-1005.

Mickey, J. and Greenland, S. (1989). A study of the impact of confounder-selection criteria on effect estimation. *American Journal of Epidemiology.* **129**, 125-137.

Milicer, H. and Szczotka, F. (1966). Age at menarche in Warsaw girls in 1965. *Human Biology.* **38**, 199-203.

Morgan, B.J.T. (1985). The cubic logistic model for quantal assay data. *Applied Statistics.* **34**, 105-113.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society.* **A 135**, 370-384.

Paula, G.A. (1982). *Testes de hipóteses para o risco relativo em estudos epidemiológicos.* Dissertação de Mestrado, IMECC, UNICAMP.

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.

Pearson, E.S. and Hartley, H.O. (1976). *Biometrika Tables for Statisticians.* Vol. I. Cambridge University Press, Cambridge.

Peduzzi, P.N., Hardy, R.J. and Holford, T.R. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics.* **36**, 511-516.

Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association.* **81**, 977-986.

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics.* **29**, 15-24.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics.* **9**, 705-724.

Pregibon, D. (1982). Score tests in GLIM with applications. *Lecture Notes in Statistics.* **14**, 87-97. Springer-Verlag, New York.

Pregibon, D. (1985). Link tests. *Encyclopedia of Statistical Sciences*, eds. Kotz, S. and Johnson, N.L.. Vol. 5, pp. 82-85, John Wiley, New York.

Prentice, R.L. (1976). A generalization of the probit and logit methods for dose response curves. *Biometrics.* **32**, 761-768.

Prentice, R.L. and Breslow, N.E. (1978). Retrospective studies and faiture time models. *Biometrika.* **65**, 153-158.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika.* **66**, 403-411.

Press, S.J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association.* **73**, 699-705.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications.* 2nd ed., John Wiley, New York.

Schwenke, J.R. and Milliken, G.A. (1983). "On the calibration problem extended to nonlinear models", in Proceedings of the Biopharmaceutical Section. *American Statistical Association*, pp.68-72.

SOC (1988). Software Científico elaborado pelo Núcleo Tecnológico de Informática para

Agropecuária da Empresa Brasileira de Agropecuária (NTIA/EMBRAPA).

Stukel, T.A. (1985). "Implementation of an algorithm for fitting a class of generalized logistic models", in *Generalized Linear Models Conference Proceedings*, ed Gilchrist, R., pp. 160-167. Springer-Verlag, New York.

Stukel, T.A. (1988). Generalized logistic models. *Journal of the American Statistical Association.* **83**, 426-431.

Stukel, T.A. (1990). A general model for estimating ED_{100P} for binary response dose-response data. *The American Statistician.* **44**, 19-22.

Upton, G.J.G. (1982). A comparison of alternative tests for the 2x2 comparative trial. *Journal of the Royal Statistical Society.* **A 145**, 86-105.

Wedderburn, R.W.M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika.* **63**, 27-32.

Wetherill, G.B. (1963). Sequential estimation of quantal response curves. *Journal of the Royal Statistical Society.* **B 25**, 1-48.

Wilks, S.S. (1962). *Mathematical Statistics.* John Wiley, New York.

Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics.* **31**, 144-148.

Williams, D.A. (1987) . Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics.* **36**, 181-191.

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics.* **19**, 251-253.

Yates, F. (1984). Tests of significance for 2x2 contingency tables. *Journal of the Royal Statistical Society.* **A 147**, 426-463 (with discussion).