UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

INTRODUÇÃO A REGRESSÃO LOGÍSTICA EXATA

AUTOR: Mathias Azevedo Bastian Bressel

ORIENTADOR: Álvaro Vigo

MONOGRAFIA APRESENTADA PARA OBTENÇÃO
DO TÍTULO DE BACHAREL EM ESTATÍSTICA

PORTO ALEGRE, MAIO DE 2002.

AGRADECIMENTOS

Ao professor e amigo Álvaro Vigo pelos seus ensinamentos, orientação e incentivo ao logo do curso.

A professora Jándyra Maria Guimarães Fachel pelo intenso apoio, incentivo e amizade durante o curso.

Ao professor Carlos Augusto Crusius pelos ensinamentos estatísticos e por sua visão crítica da estatística e do mundo.

Aos demais professores do Departamento de Estatística da UFRGS que contribuíram na minha formação com seus conhecimentos.

Aos meus colegas de curso pela convivência e pelas ajudas no decorrer do curso.

A minha família e namorada pelo apoio que me deram.

A Cytel software corporation por ter cedido o Proc-LogXact com o manual para realizar as regressões logísticas.

ÍNDICE

1. INTRODUÇÃO

Tomar decisões sempre fez parte da rotina humana. Certos tipos de questões são essencialmente determinísticas (o que comer, vestir, etc.) e usualmente podem ser solucionadas de acordo com as preferências individuais (ou de um grupo, se for o caso) e/ou por restrições e imposições do contexto. Outros tipos, no entanto, têm uma natureza mais complexa e freqüentemente podem ser descritas por modelos probabilísticos. Por exemplo, para o lançamento de um novo modelo de automóvel seria importante para o fabricante conhecer as preferências dos potenciais consumidores, para poder decidir quais devem ser as características deste novo veículo (potência do motor, cores disponíveis, etc.), facilitando o processo de decisão. Como exemplo da área médica, poderia ser a aprovação ou não de uma nova vacina (para imunização contra o dengue, por exemplo) envolvendo pesquisadores de diversas áreas e a necessidade de estudos observacionais e/ou experimentais cuidadosamente planejados.

Em ambos exemplos, as decisões não podem ser tomadas através de julgamento estritamente pessoal, sendo necessário o uso de métodos que permitam extrair eficientemente as informações contidas nos dados. Essas informações são vitais e compõem a base do critério científico, utilizado para tomar as decisões.

O tipo de problema descrito no segundo exemplo é muito comum nas áreas biomédicas, bem como no mundo dos negócios, na engenharia de qualidade, etc. Em linhas gerais, o problema é caracterizado essencialmente por envolver a necessidade de decidir por uma dentre duas alternativas (autorizar ou não a vacina, produzir ou não um novo modelo de veículo, etc.), com base em um conjunto de informações adicionais, chamados de fatores explanatórios ou experimentais. Nesta situação, diz-se que a variável resposta é binária (ou dicotômica) e geralmente é rotulada por Y. Por sua vez, as variáveis explanatórias geralmente são rotuladas por $x_1, x_2, ..., x_p$, as quais podem ser discretas ou contínuas.

Freqüentemente deseja-se descrever os efeitos das variáveis explanatórias sobre desfechos binários e a modelagem desses efeitos pode ser uma estratégia extremamente poderosa. Um modelo que apresenta um bom ajuste usualmente permite avaliar os efeitos

das variáveis explicativas, podendo descrever associações e interações, bem como gerar boas estimativas das probabilidades dos eventos associados à variável resposta.

Diversos modelos estão disponíveis para uma situação como esta. Nelder & Wedderburn (1972) introduziram uma classe de modelos de regressão chamada de *modelos lineares generalizados*, os quais são essencialmente caracterizados por possuírem três componentes: a *componente aleatória*, que identifica a distribuição de probabilidade da variável resposta; a *componente sistemática*, que especifica a função linear das variáveis explanatórias, que é utilizada como preditor; e a *função de ligação*, que descreve a relação funcional entre a componente sistemática e o valor esperado da componente aleatória. Esta classe de modelos está amplamente descrita na literatura, cabendo destacar, ainda, McCullagh & Nelder (1989) e Cordeiro (1986). Diversos modelos para resposta dicotômica pertencem à classe de modelos lineares generalizados e o *modelo de regressão logística* é, possivelmente, o mais importante deles. Neste modelo, a função de ligação baseia-se na transformação *logito* de uma proporção.

Para ilustrar, considere o desfecho dicotômico Y, tal que $Y=1$ se ocorre "sucesso" e $Y=0$, em caso contrário e, também, o vetor de variáveis explicativas $\mathbf{x}=(x_1,x_2,...,x_p)'$. Quando o vetor de covariáveis assume o valor $\mathbf{x}$, a probabilidade condicional de ocorrer "sucesso" pode ser escrita como $\pi(\mathbf{x})=P(Y=1|\mathbf{x})$ e, naturalmente, $1-\pi(\mathbf{x})=P(Y=0|\mathbf{x})$. Assim, o modelo logístico pode ser escrito na forma

$$log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}=\alpha+\sum_{j=1}^{p}x_j\beta_j.$$

A função $g(\pi)=log\frac{\pi}{1-\pi}$ é chamada de *função de ligação logito* ou *função de ligação logística* e o termo do $\alpha+\sum_{j=1}^{p}x_j\beta_j$ é a componente sistemática do modelo. É importante perceber que $\pi(\mathbf{x})$ pode ser reescrita como

$$\pi(\mathbf{x})=\frac{exp\left\{\alpha+\sum_{j=1}^{p}x_j\beta_j\right\}}{1+exp\left\{\alpha+\sum_{j=1}^{p}x_j\beta_j\right\}}$$

e corresponde à *função de distribuição da densidade logística*. Fica evidente, portanto, que a componente aleatória do modelo de regressão logística é a função densidade de probabilidade logística. Veja, por exemplo, Agresti (1990, p.85), McCullagh & Nelder (1989, p.107) e Harrel (2001, p. 216).

Além da transformação logito, diversas funções de ligação poderiam ser utilizadas. Por exemplo, a *transformação probito*, denotada por $g(\pi)=\Phi^{-1}\{\pi\}$, utiliza a inversa da função de distribuição da densidade normal padrão como função de ligação, conduzindo ao denominado *modelo de probitos*. Outros exemplos são as funções de ligação $g(\pi)=log(-log(1-\pi))$ e $g(\pi)=-log(-log\,\pi)$, denominadas de *função de ligação complementar log-log* e *log-log*, respectivamente. Uma descrição detalhada destes modelos e de suas propriedades pode ser encontrada, por exemplo, em McCullagh & Nelder (1989, p.107) e Agresti (1990, p.102).

Um procedimento alternativo de modelagem é o modelo para respostas binárias da classe dos *modelos aditivos generalizados* proposta por Hastie & Tibshirani (1990, p.95) e também descrito por Venables & Ripley (1999, p.288). No entanto, uma discussão detalhada dos modelos lineares generalizados e dos modelos aditivos generalizados está além dos objetivos desta monografia, mas as referências listadas dão uma boa orientação inicial ao leitor interessado.

Neste trabalho são discutidos alguns aspectos do modelo logístico para resposta binária, que, geralmente, não recebem um tratamento apropriado. Em particular, deseja-se apresentar e discutir aspectos da estimação de parâmetros e testes de hipóteses a eles associados. No modelo de regressão logística os parâmetros geralmente são estimados pelo *método da máxima verossimilhança*, mas é vital salientar que uma característica importante destes estimadores é que suas propriedades ótimas dependem criticamente do tamanho da amostra.

Em diversas situações práticas, no entanto, o tamanho da amostra é pequeno e, conseqüentemente, as estimativas dos parâmetros derivados pela máxima verossimilhança não gozam das propriedades desejadas, ou então não podem ser determinadas pelos procedimentos computacionais usuais. Na primeira situação, a utilização de tais estimativas poderia produzir enormes distorções nos resultados.

Um procedimento alternativo ao método da máxima verossimilhança foi apresentado por Cox (1970), mas tornou-se factível apenas após o desenvolvimento de algoritmos mais rápidos e eficientes. Neste método, as inferências são baseadas na distribuição exata da estatística suficiente correspondente ao parâmetro de regressão de interesse, condicional aos valores observados das estatísticas suficientes associadas aos demais parâmetros. Graças aos avanços computacionais recentes, diversos procedimentos computacionais têm implementado métodos exatos, permitindo o ajuste do *modelo de regressão logística exata.* Ao contrário das propriedades assintóticas do método da máxima verossimilhança, as propriedades dos estimadores derivados pelos métodos exatos não dependem criticamente do tamanho da amostra. Devido ao grande esforço computacional necessário, a regressão logística exata é aconselhável quando o tamanho da amostra é pequeno ou quando o delineamento possui uma estrutura desbalanceada ou altamente estratificada. Veja Mehta & Patel (1995).

O ajuste do modelo de regressão logística exata e as inferências baseadas nos testes sobre os parâmetros do modelo podem ser vistas como uma extensão do teste exato de Fisher para tabelas de contingência 2 x 2. Em particular, o ajuste do modelo de regressão logística exata, quando o desfecho e a variável explicativa são dicotômicos, produz um resultado equivalente ao do teste exato de Fisher. Veja Hosmer & Lemeshow (2000, p. 331).

Os principais objetivos deste trabalho são investigar os aspectos metodológicos da regressão logística exata e do modelo logístico por máxima verossimilhança, comparando as estimativas dos parâmetros através de dados empíricos. Também se deseja identificar vantagens e desvantagens dos métodos e apresentar rotinas computacionais para o ajuste dos modelos.

Este trabalho está estruturado da seguinte forma: o Capítulo 2 descreve sucintamente as principais técnicas de análise estatística de dados com resposta dicotômica. Também são apresentados os procedimentos de estimação e testes de hipóteses derivados pelo princípio da máxima verossimilhança. O Capítulo 3 apresenta o método de regressão logística exata para o caso estratificado e não estratificado. Em especial, são discutidos aspectos da estimação de parâmetros e testes de hipóteses, baseados na distribuição condicional da estatística suficiente e pelo método da verossimilhança condicional. O

Capítulo 4 descreve os procedimentos computacionais disponíveis que permitem ajustar o modelo de regressão logística exata. Como o desenvolvimento desta metodologia está diretamente ligado à evolução dos computadores e dos algoritmos, este capítulo pode ser usado como um tutorial para auxiliar um iniciante no assunto. No Capítulo 5 serão discutidos alguns exemplos que ilustram de forma bem clara as diferenças, vantagens e desvantagem de se utilizar a regressão logística exata em diferentes situações.

As considerações finais sobre o estudo, apresentadas no Capítulo 6, resumem as lições aprendidas durante o desenvolvimento do trabalho e apontam possíveis direções para continuidade da pesquisa. Especial atenção é dada à discussão sobre vantagens e desvantagens dos métodos de inferência condicional e não condicional. As referências citadas ao longo do texto e anexos estão dispostos nos capítulos 7 e 8, respectivamente.

2. MODELOS PARA RESPOSTA BINÁRIA

Os procedimentos de modelagem são extremamente úteis para descrever a relação entre o desfecho dicotômico Y e o conjunto de variáveis explicativas $x_1, x_2, ..., x_p$, pois permite estimar a magnitude e a direção dos efeitos. Os métodos encontram-se amplamente discutidos na literatura, cabendo destacar Hosmer & Lemeshow (2000), Harrel (2001), McCullagh & Nelder (1989), Agresti (1990), Cox & Snell (1989) e Breslow & Day (1980), entre outros.

Considere o desfecho dicotômico Y cujos valores podem ser genericamente representados por "fracasso" $(Y=0)$ e "sucesso" $(Y=1)$, e as variáveis explanatórias representadas pelo vetor $\mathbf{x}=(x_1, x_2, ..., x_p)'$. Seja $P(Y=1)=\pi$ e $P(Y=0)=1-\pi$ e, presumivelmente, existe uma relação de dependência entre a probabilidade de "sucesso" e as covariáveis.

Neste trabalho, a relação matemática entre π e as covariáveis $x_1, x_2, ..., x_p$ é motivada pela possibilidade de existir uma variável subjacente, latente e contínua, usualmente não observável diretamente. Em outras palavras, os dados observados surgem da dicotomização de uma variável aleatória contínua subjacente. Outras alternativas, como a *transformação logito empírica*, são discutidas por Cox & Snell (1989, p.13,24), mas estão além dos objetivos desta monografia.

De maneira similar ao modelo de regressão linear, a relação de dependência entre π e o vetor de variáveis explanatórias $\mathbf{x}=(x_1, x_2, ..., x_p)'$ pode ser expressa através da esperança condicional de Y, dado que as covariáveis assumem o valor $\mathbf{x}$, denotada por $E(Y|\mathbf{x})$. Pela definição de esperança de variáveis aleatórias discretas obtemos

$$E(Y|\mathbf{x}) = 0 \times P(Y=0|\mathbf{x}) + 1 \times P(Y=1|\mathbf{x}) = P(Y=1|\mathbf{x}) = \pi(\mathbf{x})$$

e, portanto, $0 \leq E(Y|\mathbf{x}) \leq 1$.

Admita que, para um dado valor do vetor de covariáveis $\mathbf{x}$, a função de distribuição acumulada da variável latente e contínua U é representada por $F(u|\mathbf{x})$ e que a resposta binária assume valor $Y=1$ se, e somente se, $U>0$. Portanto,

$$\pi(\mathbf{x})=P(Y=1|\mathbf{x})=1-F(0|\mathbf{x}). \tag{2.1}$$

Como a variável aleatória U não é diretamente observável, não há perda de generalidade em adotar o ponto de corte igual a 0 e, também, se o desvio padrão é constante, poderia ser igual a 1. Assim, se a variável latente U pode ser descrita pela distribuição normal com média $\alpha+\sum_{j=1}^{p}x_j\beta_j$, segue que

$$\pi(\mathbf{x})=\Phi\left(\alpha+\sum_{j=1}^{p}x_j\beta_j\right), \tag{2.2}$$

onde $\Phi(.)$ representa a função de distribuição acumulada de uma variável aleatória com *distribuição de probabilidade normal padrão.* Note que a relação pode ser linearizada mediante a transformação inversa

$$\Phi^{-1}(\pi(\mathbf{x}))=\alpha+\sum_{j=1}^{p}x_j\beta_j. \tag{2.3}$$

É importante observar que este é o modelo de probitos proposto por D. J. Finney em meados do século passado, e está extensamente descrito na literatura, particularmente no contexto dos *bioensaios*. Veja Cox & Snell (1989, p.15).

Entretanto, a distribuição normal é apenas uma possibilidade para descrever a forma da variável latente U. De fato, qualquer distribuição de probabilidade unimodal e simétrica pode ser postulada para a variável latente, geralmente produzindo resultados similares. Contudo, a suposição de uma *distribuição de probabilidade logística* para a variável latente tem se mostrado bastante adequada, principalmente pela facilidade de cálculo. Veja, por exemplo, Vigo (1994, p.59). Assim, se U é descrita pela densidade logística com parâmetro de locação $\alpha+\sum_{j=1}^{p}x_j\beta_j$ e parâmetro de escala igual a 1, então sua função de distribuição acumulada é dada por

$$\frac{exp\left\{u-\alpha-\sum_{j=1}^{p}x_j\beta_j\right\}}{1+exp\left\{u-\alpha-\sum_{j=1}^{p}x_j\beta_j\right\}}, \tag{2.4}$$

tal que

$$\pi(\mathbf{x})=P(Y=1|\mathbf{x})=\frac{exp\left\{\alpha+\sum_{j=1}^{p}x_j\beta_j\right\}}{1+exp\left\{\alpha+\sum_{j=1}^{p}x_j\beta_j\right\}}, \tag{2.5}$$

e, naturalmente,

$$1-\pi(\mathbf{x})=P(Y=0|\mathbf{x})=\frac{1}{1+exp\left\{\alpha+\sum_{j=1}^{p}x_j\beta_j\right\}}. \tag{2.6}$$

Conseqüentemente, o modelo torna-se linear nos parâmetros mediante a transformação logito, ou seja,

$$log\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}=\alpha+\sum_{j=1}^{p}x_j\beta_j, \tag{2.7}$$

sendo denominado de modelo logístico ou modelo de regressão logística. Antes de discutir aspectos da estimação de parâmetros deste modelo, convém mencionar, mais uma vez, que outras distribuições de probabilidade podem ser postuladas para a variável subjacente U. Na existência de razões que levem a acreditar que a distribuição subjacente é assimétrica, uma possibilidade é utilizar a *distribuição de valor extremo*, também chamada de *distribuição de Gumbel*. Neste caso, usando a linguagem dos modelos lineares generalizados, a função de ligação complementar log-log conduz ao modelo

$$log\{-log\{1-\pi(\mathbf{x})\}\}=\alpha+\sum_{j=1}^{p}x_j\beta_j, \tag{2.8}$$

enquanto que a transformação log-log define o modelo

$$-log\{-log\,\pi(\mathbf{x})\}=\alpha+\sum_{j=1}^{p}x_j\beta_j. \tag{2.9}$$

Embora seja uma contrapartida natural para o modelo complementar log-log, o modelo (2.9) raramente é usado devido ao seu comportamento inapropriado quando $\pi < 1/2$, que usualmente é a região de interesse. Uma discussão aprofundada destes modelos pode ser encontrada, por exemplo, em Cox & Snell (1989), McCullagh & Nelder (1989) e Agresti (1990).

O modelo logístico definido pela equação (2.5) é descrito com maior freqüência na literatura para o caso no qual as respostas binárias $Y_1, Y_2, \ldots, Y_n$ são geradas a partir de uma amostra aleatória simples. É importante mencionar, no entanto, que nos últimos anos o modelo tem sido estendido para outros contextos, ou seja, para planos amostrais mais complexos. Veja Hosmer & Lemeshow (2000, p.203).

Existem vários métodos que podem ser úteis para construir estimadores pontuais para os parâmetros. De uma maneira geral, os mais usuais são o *método da máxima verossimilhança*, o *método de mínimos quadrados*, *métodos bayesianos*, *método do qui-quadrado mínimo*, etc. Veja, por exemplo, Mood, Graybill e Boes (1974), Rohatgi (1976), Bickel & Doksum (1976), Roussas (1973), Wilks (1962).

Os parâmetros do modelo logístico usualmente são estimados pelo método da máxima verossimilhança. A seguir, são mostradas algumas etapas do procedimento de estimação por máxima verossimilhança para o modelo definido na equação (2.5), as quais estão amplamente descritas na literatura. Aos leitores interessados sugere-se consultar Hosmer & Lemeshow (2000, p.8-9), Agresti (1990, p.112), McCullagh & Nelder (1989, p. 114) ou Cox & Snell (1989, p.36, 179), entre outros.

Admita que as n respostas binárias podem ser consideradas variáveis aleatórias e independentes com distribuição de Bernoulli. Observe que não é uma suposição utópica, pois usualmente os ensaios são gerados por uma amostra aleatória. Para $i = 1, \ldots, I$, seja $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ o vetor que representa o i-ésimo conjunto de k variáveis explanatórias, onde $x_{i0} = 1$. Quando existem fatores explanatórios contínuos, podem existir um conjunto $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})$ para cada indivíduo e, neste caso, $I = n$. Com esta nova notação e fazendo $\beta_0 = \alpha$, o modelo da equação (2.5) pode ser reescrito como

$$\pi(\mathbf{x}_i) = \frac{exp\left\{\sum_{j=0}^{p} \beta_j x_{ij}\right\}}{1 + exp\left\{\sum_{j=0}^{p} \beta_j x_{ij}\right\}}. \qquad (2.10)$$

Se ocorre mais de uma observação da resposta Y em um determinado valor fixo $\mathbf{x}_i$, então é suficiente conhecer os correspondentes números de observações n_i e de sucessos Y_i, para todo $i = 1,\ldots,I$. Assim, $Y_1, Y_2, \ldots, Y_I$ são variáveis aleatórias independentes e com distribuição *Binomial*, tais que $Y_i \sim B(n_i; \pi(\mathbf{x}_i))$; $E(Y_i) = n_i \pi(\mathbf{x}_i)$ e $n_1 + n_2 + \ldots + n_I = n$.

A função massa de probabilidade conjunta de $(Y_1,\ldots,Y_I)$ é proporcional ao produto de I funções binomiais, ou seja,

$$\prod_{i=1}^{I} \pi(x_i)^{y_i} \{1 - \pi(x_i)\}^{n_i - y_i} = \left\{\prod_{i=1}^{I} (1 - \pi(\mathbf{x}_i))^{n_i}\right\} \left\{\prod_{i=1}^{I} \exp\left\{\log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right)^{y_i}\right\}\right\} \qquad (2.11)$$

$$= \left\{\prod_{i=1}^{I} (1 - \pi(\mathbf{x}_i))^{n_i}\right\} \exp\left\{\sum y_i \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right)\right\}.$$

Um procedimento usual para determinar os estimadores de máxima verossimilhança consiste essencialmente em utilizar o método do cálculo para maximizar a função de verossimilhança, mas usualmente é mais fácil maximizar o logaritmo desta função, produzindo os mesmos resultados. No presente contexto, o método pode ser resumido em três etapas: a) obter as derivadas parciais de primeira ordem do logaritmo da função de verossimilhança, em relação aos parâmetros desconhecidos; b) montar as *equações de verossimilhança*, igualando a zero as derivadas parciais de primeira ordem; e, c) resolver o sistema de equações pra obter os estimadores de máxima verossimilhança.

Para o modelo definido em (2.10), o i-ésimo logito é igual a $\sum_{j=1}^{p} \beta_j x_{ij}$, de tal forma que a última expressão da equação (2.11) pode ser escrita como

$$exp\left\{\sum_{i=1}^{I} y_i \sum_{j=1}^{p} \beta_j x_{ij}\right\} = exp\left\{\sum_{j=1}^{p}\left(\sum_{i=1}^{I} y_i x_{ij}\right)\beta_j\right\}$$

e, como $(1-\pi(\mathbf{x}_i)) = \left(1+exp\left\{\sum_{j=1}^{p}\beta_j x_{ij}\right\}\right)^{-1}$, o logaritmo da função de verossimilhança é proporcional a

$$L(\boldsymbol{\beta}) = \sum_j \left(\sum_i y_i x_{ij}\right)\beta_j - \sum_i n_i \log\left\{1+exp\left(\sum_j \beta_j x_{ij}\right)\right\}. \qquad (2.12)$$

Observe que a equação (2.12) depende da distribuição Binomial somente através da estatística suficiente $\left\{\sum_{i=1}^{I} y_i x_{ij};\ j=1,2,\ldots,p\right\}$. As derivadas parciais de $L(\boldsymbol{\beta})$ em relação aos elementos de $\boldsymbol{\beta}$ são dadas por

$$\frac{\partial L}{\partial \beta_a} = \sum_i y_i x_{ia} - \sum_i n_i x_{ia}\left\{\frac{exp\left(\sum_j \beta_j x_{ij}\right)}{1+exp\left(\sum_j \beta_j x_{ij}\right)}\right\}.$$

Igualando a zero as derivadas parciais, obtêm-se as equações de verossimilhança

$$\sum_i y_i x_{ia} - \sum_i n_i \hat{\pi}_i x_{ia} = 0, \quad a = 0,\ldots,p \qquad (2.13)$$

onde $\hat{\pi}_i = exp\left(\Sigma_j \hat{\beta}_j x_{ij}\right)\Big/\left\{1+exp\left(\Sigma_j \hat{\beta}_j x_{ij}\right)\right\}$ denota a estimativa de máxima verossimilhança de $\pi(\mathbf{x}_i)$.

É importante mencionar que, exceto para casos especiais extremos, o logaritmo da função de verossimilhança para os modelos logístico e de probitos é sempre côncavo e, portanto, os estimadores de máxima verossimilhança existem e são únicos. Veja Wedderburn (1976). Particularmente para o modelo logístico, os conjuntos de dados podem ser classificados em três categorias mutuamente exclusivas e exaustivas: separação completa, separação quase complèta e sobreposição. Para as duas primeiras categorias os estimadores de máxima verossimilhança não existem. Veja Albert & Anderson (1984).

Outro aspecto importante é que as equações de verossimilhança definidas em (2.13) são funções não lineares dos estimadores de máxima verossimilhança e, assim, devem ser resolvidas através de procedimentos iterativos. Detalhes do procedimento de estimação,

inclusive do *método iterativo de Newton-Raphson* pode ser encontrado, por exemplo, em Agresti (1990, p.112-117).

Uma característica bem conhecida dos estimadores de máxima verossimilhança é que suas propriedades são ótimas para amostras grandes. No entanto, quando o tamanho da amostra é pequeno, para dados desbalanceados ou altamente estratificados, os estimadores de máxima verossimilhança podem não existir ou produzem resultados pouco confiáveis. Um método alternativo para fazer inferências sobre os parâmetros do modelo logístico é o proposto por Cox (1970). Este método baseia-se na distribuição exata da estatística suficiente e será descrito no próximo capítulo.

3 REGRESSÃO LOGÍSTICA EXATA

A regressão logística exata foi originalmente sugerida por Cox (1970), mas só começou a ser utilizada a partir de um eficiente algoritmo desenvolvido por Tritchler (1984). Nos anos seguintes, novos algoritmos foram desenvolvidos por Hirji, Mehta e Patel (1987, 1988) e Hirji (1992), que aliados à disponibilidade de computadores mais potentes, permitiram aplicar a metodologia em problemas mais sofisticados.

Apesar do método existir a cerca de trinta anos, apenas na última década foram disponibilizados programas computacionais com algoritmos para o modelo de regressão logística exata, como, por exemplo, o programa estatístico LogXact 4 for Windows produzido pela Cytel Software. Estes recursos também foram disponibilizados no procedimento PROC LOGISTIC do programa estatístico SAS – Statistical Analysis System, a partir da versão 8.1. Uma versão do programa LogXact 4, denominada PROC LogXact, também está disponível para ser utilizada em conjunto com programa SAS, versão 8.1 ou superior. Detalhes destes procedimentos computacionais serão ilustrados e discutidos no próximo capítulo, bem como nos estudos de caso apresentados no Capítulo 5.

Apesar dos avanços para estender o modelo logístico para estruturas mais complexas, nesta seção são considerados aspectos metodológicos da estimação e testes de hipóteses sobre os parâmetros do modelo logístico apenas para o caso *não estratificado* e *estratificado*, os quais acredita-se serem suficientes para chamar a atenção das propriedades mais importantes.

3.1 Regressão logística não estratificada

Considere o conjunto de variáveis aleatórias dicotômicas e independentes $Y_1, Y_2, ..., Y_n$. Para cada Y_i está associado o vetor p-dimensional que corresponde aos valores observados das variáveis explanatórias no i-ésimo indivíduo, usualmente denotado por

$\mathbf{x}_i = (x_{1i}, x_{2i}, ..., x_{pi})'$. Com esta nova notação, o modelo especificado na equação (2.7) pode ser escrito como

$$\log \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = \alpha + \mathbf{x}_i'\boldsymbol{\beta}, \tag{3.1}$$

onde α e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ são parâmetros desconhecidos e $\pi(\mathbf{x}_i) = P(Y = 1|\mathbf{x}_i)$. A probabilidade conjunta de observar os valores $y_1, y_2, ..., y_n$, chamada de *função de verossimilhança*, é dada por

$$P(Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n) = \frac{\exp\left\{\sum_{i=1}^{n} y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta})\right\}}{\prod_{i=1}^{n} \exp\left\{1 + \sum_{i=1}^{n} y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta})\right\}}, \tag{3.2}$$

onde $\mathbf{y} = (y_1, y_2, ..., y_n)'$.

Uma maneira usual de estimar os parâmetros do modelo é maximizar a *função de verossimilhança não condicional* (3.2) com respeito aos parâmetros $\alpha, \beta_1, \beta_2, ..., \beta_p$. Este método de estimação é denominado de *método da máxima verossimilhança* e foi proposto por R. A. Fisher em 1912 e desenvolvido posteriormente em Fisher (1922) e Fisher (1925) – Apud Azzalini (1996, p.52)[1]. Formas análogas de escrever e maximizar a função de verossimilhança associada ao modelo logístico podem ser encontradas também em Azzalini (1996, p.88), McCullagh & Nelder (1989, p.114), Cox & Snell (1989, p.27), Agresti (1990, p.112), Harrell (2001, p.228) ou Hosmer & Lemeshow (2000, p.8), entre outras referências.

No entanto, por conveniência e sem perda de generalidade, neste texto está sendo adotada a abordagem descrita por Hosmer & Lemeshow (2000). Admita, então, que o objetivo básico é fazer inferências sobre os coeficientes de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ e que o intercepto α pode ser considerado um parâmetro de ruído (*nuisance*, em inglês). Assim, o parâmetro α pode ser eliminado condicionando a função de verossimilhança (3.2) no valor observado da sua *estatística suficiente*, $\sum_{i=1}^{n} Y_i = m$. Embora não seja vital

---

[1] FISHER, R.A. (1922). **On The Mathematical Foundations Of Theoretical Statistics.** *Philos. Trans. Roy. Soc.* London, Ser. A **222**, 309-368.
FISHER, R.A. (1925). **Theory of statistical estimation.** *Proc. Cambridge Philos. Soc.* **22**, Pt. 5, 700-725.

neste momento, uma definição formal de estatística suficiente pode ser encontrada, por exemplo, em Azzalini (1996, p.30-32). Particularmente para o modelo logístico, veja Azzalini (1996, p.41), Cox & Snell (1989, p.27-28) ou McCullagh & Nelder (1989, p.115-116).

Assim, a *função de verossimilhança condicional* é dada por

$$P\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \middle| \sum_{i=1}^{n} Y_i = m\right) = \frac{exp\left\{\sum_{i=1}^{n} y_i \mathbf{x}_i' \boldsymbol{\beta}\right\}}{\sum_{R} exp\left\{\sum_{i=1}^{n} y_i \mathbf{x}_i' \boldsymbol{\beta}\right\}}, \tag{3.3}$$

onde o somatório em R no denominador da equação (3.3) indica a soma sobre o conjunto

$$R = \left\{(y_1, y_2, \ldots, y_n) : \sum_{i=1}^{n} y_j = m\right\}.$$

É importante notar que o vetor p-dimensional $\mathbf{t} = (t_1, t_2, \ldots t_p)' = \sum_{i=1}^{n} y_i \mathbf{x}_i$, onde $t_j = \sum_{i=1}^{n} y_i x_{ji}$, é suficiente para $\boldsymbol{\beta}$ e sua distribuição de probabilidade exata é dada por

$$P(T_1 = t_1, T_2 = t_2, \ldots, T_p = t_p) = \frac{c(\mathbf{t}) exp\{\mathbf{t}'\boldsymbol{\beta}\}}{\sum_{\mathbf{u}} c(\mathbf{u}) exp\{\mathbf{u}'\boldsymbol{\beta}\}}, \tag{3.4}$$

onde $c(\mathbf{t}) = |S(\mathbf{t})|$, $S(\mathbf{t}) = \left\{(y_1, y_2, \ldots, y_n) : \sum_{i=1}^{n} y_i = m, \sum_{i=1}^{n} y_i x_{ji} = t_j, j = 1,2,\ldots,p\right\}$, $|S|$ representa o número de elementos distintos do conjunto S e a soma no denominador da equação (3.4) é sobre todo $\mathbf{u}$ que satisfaz $c(\mathbf{u}) \geq 1$. Em outras palavras, $c(\mathbf{t})$ é o número de seqüências de 0's e 1's na forma $(y_1, y_2, \ldots, y_n)$, tais que $\sum_{i=1}^{n} y_i = m$ e $\sum_{i=1}^{n} y_i x_{ji} = t_j$, para $j = 1,2,\ldots,p$. A distribuição exata (3.4) é usada para obter estimativas pontuais e intervalos de confiança, bem como testar hipóteses sobre os coeficientes de regressão. Veja Mehta & Patel (1985), Cox & Snell (1989, p.28-29) e Hosmer & Lemeshow (2000, p.332).

A vantagem de escrever a função de verossimilhança na forma (3.3) é que agora as inferências sobre os coeficientes de regressão $\boldsymbol{\beta}$ podem ser realizadas tanto para o método assintótico quanto para o método exato. O primeiro caso consiste essencialmente em maximizar a função de verossimilhança condicional especificada na equação (3.3). Por

outro lado, as inferências exatas sobre $\boldsymbol{\beta}$ requerem o cálculo dos coeficientes $c(\mathbf{u})$, para os quais algumas estatísticas suficientes são fixadas em seus valores observados, enquanto que outras devem variar sobre todos os valores possíveis. Veja Mehta & Patel (1995).

O exemplo descrito a seguir é útil para ilustrar alguns aspectos básicos da regressão logística exata, tais como a identificação do vetor de estatística suficiente e sua distribuição condicional exata.

EXEMPLO: (Dados hipotéticos): O presente exemplo é de caráter estritamente pedagógico e foi apresentado e discutido por Derr (2000). Considere a amostra aleatória $(Y_1, Y_2, Y_3, Y_4)'$ que gerou as respostas binárias y_i mostradas no Quadro 3.1, onde x_0 é o indicador do intercepto e x_1 variável explicativa.

Quadro 3.1 – Dados observados

Observação	Resposta (y)	x_0	x_1
1	0	1	1
2	1	1	1
3	0	1	2
4	1	1	0

Fonte: Derr (2000)

Utilizando a notação matricial usual, os dados observados podem ser escritos através dos vetores $\mathbf{y}_0 = (0,1,0,1)'$, $\mathbf{x}_0 = (1,1,1,1)'$ e $\mathbf{x}_1 = (1,1,2,0)'$. O valor observado do vetor de estatística suficiente para $\boldsymbol{\beta}$, escrito como $\mathbf{t} = (t_1, t_2, \ldots t_p)' = \sum_{i=1}^{n} y_i \mathbf{x}_i$, é

$$\mathbf{t} = (t_0, t_1)' = 0 \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 1 \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0 \times \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 1 \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

O objetivo básico da análise condicional exata é determinar a probabilidade de se observar o vetor de respostas binárias $\mathbf{y}_0$, em relação ao total de 2^n combinações possíveis

da forma $\mathbf{y} = (y_1, y_2, ..., y_n)'$. O Quadro 3.1 mostra as $2^4 = 16$ combinações possíveis de respostas binárias $\mathbf{y} = (y_1, y_2, y_3, y_4)'$ associadas ao Exemplo. Admita, por exemplo, que o objetivo é determinar a distribuição da estatística suficiente associada a covariável x_1, condicional ao valor observado da estatística suficiente associada a x_0. Essa distribuição pode ser derivada da distribuição conjunta de $\mathbf{t} = (t_0, t_1)'$, considerando apenas os vetores para os quais a primeira componente assume o valor $t_0 = 2$, como mostra o Quadro 3.3.

Quadro 3.2 – Combinações de respostas binárias possíveis para o Exemplo ($2^n = 16$) e correspondentes valores das componentes do vetor de estatística suficiente.

i	Respostas binarias				$\mathbf{t} = (t_0, t_1)'$	
	y_1	y_2	y_3	y_4	t_0	t_1
1	0	0	0	0	0	0
2	0	0	0	1	1	0
3	0	0	1	0	1	2
4	0	0	1	1	2	2
5	0	1	0	0	1	1
6	0	1	0	1	2	1
7	0	1	1	0	2	3
8	0	1	1	1	3	3
9	1	0	0	0	1	1
10	1	0	0	1	2	1
11	1	0	1	0	2	3
12	1	0	1	1	3	3
13	1	1	0	0	2	2
14	1	1	0	1	3	2
15	1	1	1	0	3	4
16	1	1	1	1	4	4

Quadro 3.3 – Distribuição condicional de t_1 dado t_0.

t_0	t_1	Freqüência	Probabilidade
2	1	2	2/6
2	2	2	2/6
2	3	2	2/6
Total		6	1

É importante observar que, conceitualmente, é simples gerar a distribuição conjunta condicional da estatística suficiente mediante a enumeração completa da distribuição conjunta de $\mathbf{y} = (y_1, y_2, ..., y_n)'$, mas o método pode não ser factível computacionalmente quando o tamanho da amostra n aumenta. No Exemplo 1, $n = 4$ e, assim, existem apenas $2^4 = 16$ combinações distintas de seqüências de resposta binárias $\mathbf{y} = (y_1, y_2, y_3, y_4)'$. Por sua vez, aumentando o tamanho de amostra para apenas $n = 30$, o número de combinações de vetores $\mathbf{y} = (y_1, y_2, ..., y_{30})'$ cresce para cerca de 1,074 bilhões, fato que ilustra as dificuldades computacionais do método. No entanto, o algoritmo proposto por Hirji, Mehta e Patel (1987), denominado de *algoritmo de mudança multivariada* (*multivariate shift algorithm*, em inglês), permite gerar e contar com grande rapidez os vetores $\mathbf{y} = (y_1, y_2, ..., y_n)'$, mesmo para tamanhos de amostra grande, possibilitando um crescimento na utilização dos métodos exatos, em particular da regressão logística exata. Veja Derr (2000).

3.2 Regressão logística estratificada

Suponha que existam N estratos, cada um com variável resposta binária. Considere que o i-ésimo estrato tenha m_i respostas e $n_i - m_i$ não respostas. Para todo $1 \leq i \leq$ N, e $1 \leq j \leq n_i$, seja $Y_{ij} = 1$ se o j-ésimo indivíduo no i-ésimo estrato respondeu; $Y_{ij} = 0$ em caso contrário. Defina, também, a probabilidade $\pi_{ij} = P(Y_{ij} = 1 | \mathbf{x}_{ij})$, onde $\mathbf{x}_{ij}$ é o vetor de

covariáveis p-dimensional para o j-ésimo indivíduo no i-ésimo estrato. Nesta situação, o modelo de regressão logística pode ser escrito como

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \alpha_i + \mathbf{x}_{ij}'\boldsymbol{\beta}, \tag{3.5}$$

onde α_i é um parâmetro específico do estrato i e $\boldsymbol{\beta}$ é um vetor de parâmetros $(p \times 1)$, comuns em todos os N estratos. O interesse usual é em fazer inferências sobre $\boldsymbol{\beta}$, considerando os α_i como parâmetros de ruído. Naturalmente, os parâmetros de ruído poderiam ser estimados pelo método da máxima verossimilhança.

Entretanto, a teoria assintótica da máxima verossimilhança exige que a dimensão do espaço de parâmetros seja fixa, enquanto que o número de observações deve aumentar. Segundo Cox & Hinkley (1974, p.292) salientam que, em geral, quando a dimensão do espaço de parâmetros é grande em relação ao número de observações, os estimadores de máxima verossimilhança podem apresentar um forte viés. Um exemplo clássico deste tipo de problema, descrito por Breslow & Day (1980, p.249) e (Andersen (1973, p.69) Apud Hosmer & Lemeshow (2000))[2] é o caso da estimação da *razão de chances* (*odds ratio*, em inglês) em um estudo com dados pareados. Neste caso, o modelo logístico deveria conter um conjunto de parâmetro de ruído para cada estrato (um para cada par) e um único parâmetro para a razão de chances, comum para todos os pares. Se os parâmetros de ruído forem estimados, a estimativa da razão de chances tende a convergir para o quadrado do seu real valor.

Ao invés de estimar todos os parâmetros específicos de cada estrato, um procedimento alternativo é eliminar os parâmetros de ruído condicionando no valor observado das suas estatísticas suficientes, que nestes casos é número de respostas m_i, em cada estrato. Veja Breslow & Day (1980) e Mehta & Patel (1995).

A função de verossimilhança condicional, ou probabilidade condicional de observar $Y_{ij} = y_{ij},\ j = 1,2,\ldots,n_i,\ i = 1,2,\ldots,N$ é

$$P\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n \middle| m_1, m_2, \ldots, m_N\right) = \frac{exp\left(\sum_{i=1}^{n}\sum_{j=1}^{n_i} y_{ij}\left(\mathbf{x}_{ij}'\boldsymbol{\beta}\right)\right)}{\sum_{i=1}^{N}\sum_{R_i} exp\left(\sum_{i=1}^{N}\sum_{j=1}^{n_i} y_{ij}\left(\mathbf{x}_{ij}'\boldsymbol{\beta}\right)\right)}, \tag{3.6}$$

---

[2] ANDERSEN, E. B. (1973) **Conditional Inference and Models for Measuring.** Mental Hygienisk Forlag, Copenhagen

onde a soma no denominador é sobre o conjunto $R_i = \left\{ (y_{i1}, y_{i2}, \ldots, y_{in_i}) : \sum_{j=1}^{n_i} Y_{ij} = m_i \right\}$, para $i = 1,2,\ldots,N$. Note que o parâmetro de ruído α_i foi eliminado da função de verossimilhança condicional. Este procedimento de estimação, popularizado por Breslow & Day (1980) consiste em fazer inferências assintóticas sobre $\boldsymbol{\beta}$, mediante a maximização da equação (3.6).

Por outro lado, inferências exatas são baseadas na distribuição da estatística suficiente para $\boldsymbol{\beta}$. O vetor de estatísticas suficientes para $\boldsymbol{\beta}$ é dado por

$$\mathbf{t} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij}, \tag{3.7}$$

e a sua distribuição condicional exata é

$$P(T_1 = t_1, T_2 = t_2, \ldots, T_p = t_p) = \frac{c(\mathbf{t}) exp\{\boldsymbol{\beta}' \mathbf{t}\}}{\sum_{\mathbf{u}} c(\mathbf{u}) exp\{\boldsymbol{\beta}' \mathbf{u}\}}, \tag{3.8}$$

onde $c(\mathbf{t}) = |S_N(\mathbf{t})|$, $S_N(\mathbf{t}) = \left\{ (y_{ij}, j = 1,\ldots,n_i, i = 1,\ldots,N) : \sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij} = \mathbf{t}, \sum_{j=1}^{n_i} y_{ij} = m_i \right\}$, $|S_N|$ representa o número de elementos distintos no conjunto S_N e a soma do denominador está definida para todo $\mathbf{u}$ que satisfaz $c(\mathbf{u}) \geq 1$. Em outras palavras, $c(\mathbf{t})$ representa o número de maneiras de selecionar as seqüências binárias $\{y_{ij}, i = 1,\ldots,N, j = 1,\ldots,n_i\}$ que satisfazem as condições

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij} = \mathbf{t}, \tag{3.9}$$

e

$$\sum_{j=1}^{n_i} y_{ij} = m_i. \tag{3.10}$$

É importante observar que tanto na regressão logística não estratificada quanto na estratificada, a distribuição de T tem a mesma forma, portanto pode-se desenvolver um único algoritmo numérico para ambos os casos.

Agora, que já foram expostas às regressões logísticas estratificada e não estratificada, poder-se-á ver como é feita a inferência condicional exata para um único parâmetro β_p, para vários parâmetros e para combinações lineares entre os parâmetros.

3.3 Inferência condicional exata

As inferências realizadas através de um modelo logístico usualmente envolvem estimação ou testes de hipóteses sobre um ou mais coeficientes de regressão e podem ser úteis para avaliar a importância do(s) preditor(es). Esta seção resume os principais aspectos dos procedimentos para fazer inferências no modelo de regressão logística exata e baseia-se essencialmente em Mehta & Patel (1995). Outros detalhes também podem ser encontrados, por exemplo, em Derr (2000) e Hosmer e Lemeshow (2000, p.333).

3.3.1 Para um único parâmetro β_p

Sem perda de generalidade, admita que deseja-se fazer inferências sobre um único coeficiente de regressão β_p. Pelo princípio da suficiência, a distribuição condicional de T_p, dado $t_1, t_2, \ldots, t_{p-1}$, depende somente de β_p e seja $f(t_p|\beta_p)$ a distribuição de probabilidade condicional $P(T_p = t_p | T_1 = t_1, T_2 = t_2, \ldots, T_{p-1} = t_{p-1})$. Então,

$$f(t_p|\beta_p) = \frac{c(t_1, t_2, \ldots, t_p) exp\{\beta_p t_p\}}{\sum_u c(t_1, t_2, \ldots, t_{p-1}, u) exp\{\beta_p u\}}, \qquad (3.11)$$

onde a soma no denominador da equação (3.11) é relativa a todos os valores de u que satisfazem a condição $c(t_1, t_2, \ldots, t_{p-1}, u) \geq 1$. Como esta probabilidade condicional não envolve os parâmetros de ruído $(\beta_1, \beta_2, \ldots . \beta_{p-1})$, pode ser usada para fazer inferência sobre

β_p. É conveniente observar, também, que a distribuição condicional definida na equação (3.11) tem a mesma forma, independentemente da distribuição conjunta do vetor **t** ser definida pelas equações (3.4) ou (3.8), produzindo, assim, um procedimento unificado para as inferências sobre um único parâmetro β_p no caso estratificado e não estratificado.

Para ilustrar, considere o teste de hipóteses que especifica que $H_0: \beta_p = 0$ contra a alternativa $H_0: \beta_p \neq 0$, mediante o qual é possível avaliar a importância da variável explanatória rotulada como x_p. Note que a rejeição da hipótese nula sugere que a variável explanatória x_p possivelmente está linearmente relacionada com o logito da probabilidade de "sucesso". O *p-valor* exato associado pode ser obtido pelas probabilidades condicionais associadas à uma especificada região crítica *E*, ou seja,

$$p = \sum_{v \in E} f\left(v \middle| \beta_p = 0\right). \tag{3.12}$$

A região E pode ser especificada de diferentes maneiras, usualmente conduzindo a teste distintos. Dois testes bem populares são os denominados de *teste das probabilidades condicionais* (*conditional probabilities test*, em inglês) e *teste de escores condicionais* (*conditional scores test*, em inglês). No primeiro, a região crítica, denotada por E_{cp}, é definida por todos os valores da estatística de teste cuja probabilidade condicional não é superior ao valor observado t_p, podendo ser escrita por

$$E_{cp} = \left\{ v: f\left(v \middle| \beta_p = 0\right) \leq f\left(t_p \middle| \beta_p = 0\right) \right\}. \tag{3.13}$$

Por sua vez, a região crítica do teste baseado em escores contém todos os valores da estatística de teste para os quais os escores condicionais são maiores ou iguais ao escore condicional associado ao valor observado da estatística de teste. Assim, esta região crítica é definida por

$$E_{cs} = \left\{ v: \ \left(v - \mu_p\right)^2 \sigma_p^{-2} \geq \left(t_p - \mu_p\right)^2 \sigma_p^{-2} \ \right\}, \tag{3.14}$$

onde μ_p e σ_p^2 representam a média e a variância de T_p, respectivamente, baseada na distribuição condicional especificada em (3.11), sob $\beta_p = 0$.

Para ambos os tipos de testes exatos é preciso utilizar um algoritmo que forneça todos os coeficientes $c\left(t_1, t_2, \ldots, t_{p-1}, v\right)$, para $t_1, t_2, \ldots, t_{p-1}$ fixados em seus valores

observados e v variando em todo domínio de T_p. Após gerar estes coeficientes, o cálculo do *p-valor* exato fica apenas restrito às operações de ordenamento e de soma apropriadas.

Também é possível derivar uma versão assintótica do teste baseado em escores condicionais, onde *p-valor* é obtido comparando o escore observado $(t_p - \mu_p)^2 \sigma_p^{-2}$ com uma distribuição de probabilidade qui-quadrado com 1 grau de liberdade. É importante observar, no entanto, que mesmo na versão assintótica é necessário calcular a média condicional μ_p e a variância condicional σ_p^2. Aproximações assintóticas para os momentos condicionais em discussão foram descritas por Zelen (1991) apud Metha, C. R. & Patel, N. R. (1995)[3].

Um intervalo de confiança de nível $(1-\alpha)\times 100\%$ para β_p, denotado genericamente por (β_-, β_+), pode ser obtido mediante a inversão dos testes discutidos acima. Assim, sejam

$$F_1(t_p|\beta) = \sum_{v \geq t_p} f(v|\beta)$$

e

$$F_2(t_p|\beta) = \sum_{v \leq t_p} f(v|\beta).$$

Defina, também, $t_{\min}$ e $t_{\max}$ como sendo, respectivamente o menor e o maior valor possível para t_p na distribuição (3.11). O limite inferior do intervalo de confiança, β_-, é tal que

$$F_1(t_p|\beta_-) = \alpha/2, \text{ se } t_{min} < t_p \leq t_{max},$$

$$\beta_- = -\infty, \text{ se } t_p = t_{min}.$$

Similarmente, o limite superior para o intervalo, β_+, é dado por

$$F_2(t_p|\beta_+) = \alpha/2, \text{ se } t_{min} \leq t_p < t_{max},$$

$$\beta_+ = +\infty, \text{ se } t_p = t_{max}.$$

---

[3] ZELEN, M. (1991) **Multinomial Response Models.** *Computational Statistics and Data Analysis,* **12**, 249-254.

É possível mostrar que este intervalo contém o verdadeiro valor β_p, com $(1-\alpha)\times 100\%$ de confiança. Veja Mehta & Patel (1995).

Uma estimativa pontual para β_p pode ser obtida de duas formas distintas. Uma maneira é derivar o *estimador de máxima verossimilhança condicional*, rotulado como β_{emvc}, mediante a maximização de $f(t_p|\beta)$, para um dado β. Contudo, se $t_p = t_{min}$ ou se $t_p = t_{max}$, o estimador β_{emvc} não está definido, pois a função de verossimilhança não pode ser maximizada. Em outras palavras, em tais casos não existe estimador de máxima verossimilhança para β_p e uma alternativa é usar a mediana definida por

$$\beta_{emrv} = (\beta_+ + \beta_-)/2, \tag{3.15}$$

onde β_- e β_+ são determinados ao nível de confiança $\alpha = 0{,}5$. Se $\beta_- = -\infty$, então $\beta_{emrv} = \beta_+$ e, se $\beta_+ = +\infty$, então $\beta_{emrv} = \beta_-$. Ao contrário do estimador de máxima verossimilhança, o estimador β_{emrv} está sempre definido e é imparcial para β_p. Este estimador também goza de diversas propriedades úteis. Veja Mehta & Patel (1995) e Hirji, Tsiatis e Mehta (1989).

3.3.2 Inferência para vários parâmetros

Para fazer inferência sobre vários parâmetros simultaneamente, é preciso calcular a distribuição conjunta das correspondentes estatísticas suficientes, condicionada nos valores observados das demais estatísticas suficientes. Por exemplo, o vetor $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ que contém os coeficientes de regressão $\boldsymbol{\beta}$, pode ser dividido em duas partes. Uma partição tem dimensão $(p_1 \times 1)$ e é denotada por $\boldsymbol{\beta}_1$, enquanto que a outra partição tem dimensão $(p_2 \times 1)$ e é rotulada como $\boldsymbol{\beta}_2$. Sejam $\mathbf{t}_1$ e $\mathbf{t}_2$ os correspondentes vetores de estatísticas suficientes.

Admita que o objetivo é testar a hipótese nula que especifica que todos os coeficientes de regressão da segunda partição são iguais a zero, ou seja,

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0},$$

contra a hipótese alternativa que pelo menos um dos elementos de $\boldsymbol{\beta}_2$ é diferente de zero. Pelo princípio da suficiência, a distribuição condicional de $\mathbf{T}_2$, dado $\mathbf{T}_1 = \mathbf{t}_1$, não depende dos parâmetros de ruído $\boldsymbol{\beta}_1$ e, assim, pode-se escrever a probabilidade condicional $P(\mathbf{T}_2 = \mathbf{t}_2 | \mathbf{T}_1 = \mathbf{t}_1)$ por $f(\mathbf{t}_2 | \boldsymbol{\beta}_2)$, onde

$$f(\mathbf{t}_2 | \boldsymbol{\beta}_2) = \frac{c(\mathbf{t}_1, \mathbf{t}_2) exp\{\boldsymbol{\beta}_2' \mathbf{t}_2\}}{\sum_{\mathbf{u}} c(\mathbf{t}_1, \mathbf{u}) exp\{\boldsymbol{\beta}_2' \mathbf{u}\}}, \tag{3.16}$$

e a soma no denominador da equação (3.16) é sobre todos os valores de $\mathbf{u}$ que satisfazem a condição $c(\mathbf{t}_1, \mathbf{u}) \geq 1$. O *p-valor* exato e bilateral para testar $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ pode ser obtido mediante a soma das probabilidades condicionais definidas na equação (3.16) que pertencem à região crítica E, definida por

$$p = \sum_{\mathbf{v} \in E} f(\mathbf{v} | \boldsymbol{\beta}_2 = \mathbf{0}). \tag{3.17}$$

Como na seção anterior, existem dois tipos de regiões críticas que levam, respectivamente, ao *teste das probabilidades condicionais* e ao *teste de escores condicionais*, cujas correspondentes regiões críticas são dadas por

$$E_{cp} = \{\mathbf{v} : f(\mathbf{v} | \boldsymbol{\beta}_2 = \mathbf{0}) \leq f(\mathbf{t}_2 | \boldsymbol{\beta}_2 = \mathbf{0})\}, \tag{3.18}$$

e

$$E_{cs} = \{\mathbf{v} : (\mathbf{v} - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-2} (\mathbf{v} - \boldsymbol{\mu}_2) \geq (\mathbf{t}_2 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}_2^{-2} (\mathbf{t}_2 - \boldsymbol{\mu}_2)\}, \tag{3.19}$$

onde $\boldsymbol{\mu}_2$ é o vetor de médias e $\boldsymbol{\Sigma}_2$ é a matriz de variância e covariância associada à distribuição condicional $f(\mathbf{t}_2 | \boldsymbol{\beta}_2 = \mathbf{0})$. Em ambos os testes é necessário um algoritmo que produza os coeficientes $c(\mathbf{t}_1 | \mathbf{v})$ para $\mathbf{t}_1$ fixo e $\mathbf{v}$ variando por toda a amplitude de $\mathbf{T}_2$.

3.3.3 Inferência simultânea para combinações lineares de parâmetros

Por conveniência de notação, sem perda de generalidade, nesta seção o modelo logístico é apresentado de uma maneira unificada para os casos estratificado e não estratificado. Assim, o modelo pode ser escrito como

$$\text{logit}(\Pi) = \mathbf{X}\boldsymbol{\beta}, \tag{3.20}$$

onde $\text{logit}(\Pi)$ é um vetor $n \times 1$ de logitos das probabilidades de "sucesso", cuja j-ésima componente é $log\left(\pi_j/(1-\pi_j)\right)$, $\mathbf{X}$ é a matriz de dados $n \times p$ e os termo constante devem ser incorporados no vetor $p \times 1$ de parâmetros $\boldsymbol{\beta}$. Admita que se deseja testar a hipótese

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0},$$

onde $\mathbf{C}$ é uma matriz $(r \times p)$ de posto completo. A hipótese nula pode ser testada reescrevendo o modelo (3.20) na foram

$$\text{logit}(\Pi) = \mathbf{X}\boldsymbol{\beta} + \mathbf{XGC}\boldsymbol{\beta}, \tag{3.21}$$

onde $\mathbf{G}'$ é o complemento ortogonal de $\mathbf{C}$, ou seja, $\mathbf{GC} = \mathbf{0}$. O modelo (3.21) pode ser reparametrizado como

$$\text{logit}(\Pi) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \tag{3.22}$$

onde $\mathbf{X}_1 = \mathbf{X}$, $\mathbf{X}_2 = \mathbf{XG}$, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$, $\boldsymbol{\beta}_2 = \mathbf{C}\boldsymbol{\beta}$ e a hipótese nula $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ pode ser testada derivando a distribuição condicional exata de $\mathbf{T}_2 | \mathbf{T}_1 = \mathbf{t}_1$, de maneira similar ao caso descrito na Seção 3.3.2.

4. PROCEDIMENTOS COMPUTACIONAIS

O imenso avanço na área da informática dos últimos anos possibilitou a implementação das idéias sugeridas por Cox (1970) relativas aos métodos exatos para inferências sobre parâmetros do modelo de regressão logística. De fato, além de computadores com maior capacidade de processamento, também foi necessária a criação de algoritmos mais rápidos e eficientes para enumerar exaustivamente as possíveis seqüências de respostas binárias $(y_1, y_2, ..., y_n)$.

Dentre os principais algoritmos, é importante destacar os trabalhos desenvolvidos por Tritchler (1984), Hirji, Metha e Patel (1987, 1988), Hirji (1992) e Metha, Patel e Senchaudhuri (2000). O algoritmo proposto por Tritchler (1984) contempla apenas o caso de uma única variável explanatória, com a possibilidade de estratificação para dados pareados, mas apresenta um substancial melhoramento na etapa de enumeração exaustiva, mediante uma aplicação específica do algoritmo de Pagano & Tritchler (1983). Por sua vez, nos trabalhos de Hirji, Metha e Patel (1987, 1988) foi proposto um algoritmo mais geral e eficiente para avaliar as permutações da distribuição condicional de $T_2|T_1 = t_1$ definida na equação (3.16) para o caso não estratificado e, subseqüentemente, estendido pra o caso estratificado. Uma extensão destes algoritmos permitiu, também, o ajuste do modelo de regressão logística exata para respostas politômicas. Veja Hirji (1992).

Teoricamente, a regressão logística exata pode ser utilizada para qualquer tamanho de amostra e quaisquer quantidades de variáveis explanatórias. Na prática, entretanto, seria necessário um longo tempo de espera, mesmo com computadores com grande capacidade de processamento. Uma exemplificação destas dificuldades foi descrita por Hosmer & Lemeshow (2000, p.338).

Existem muitos programas estatísticos que permitem o ajuste de modelos logísticos, com estimação de parâmetros e testes de hipóteses através de métodos assintóticos. No entanto, talvez devido às dificuldades computacionais mencionadas, poucos disponibilizam os procedimentos para inferência exata descritos no Capítulo 3. Dentre eles, cabe destacar os pacotes estatísticos SAS (Statistical Analysis System), versão 8.1 ou posterior e LogXact 4 for Windows. O programa LogXact foi desenvolvido pela empresa Cytel Software e

também é disponibilizado através do procedimento PROC LogXact, para ser utilizado em conjunto com o programa SAS.

Para o programa estatístico SPSS (Statistical Software for the Social Science) está em fase experimental uma rotina (macro) para ajustar o modelo de regressão logística exata, mas não está disponível para uso comercial.

No presente trabalho, os modelos foram ajustados mediante a utilização do PROC LOGISTIC do programa SAS, versão 8.2, bem como do procedimento PROC LogXact. Nas próximas seções serão descritas as principais características destas rotinas.

4.1 Procedimento PROC LOGISTIC do SAS

O SAS é um dos programas mais utilizados para análise estatística de dados nas mais variadas áreas. Esta grande aceitação seguramente deve-se à grande quantidade de métodos de análise implementados e, naturalmente, a sua constante atualização e inclusão de novas metodologias. Estes aspectos tornam o programa SAS uma ferramenta para análise estatística de dados extremamente versátil e poderosa.

A partir da versão 8.1, o procedimento PROC LOGISTIC do programa SAS permite o ajuste do modelo de regressão logística exata apenas para respostas binárias. Nesta seção serão descritos alguns aspectos básicos deste procedimento, os quais estão detalhadamente documentados no manual do programa. Veja SAS (1999). Outra restrição, além de possuir um desfecho dicotômico, é o tamanho da amostra. A rotina incorporada ao SAS só comporta cálculos exatos para amostras não muito grandes. Como, em geral, busca-se a metodologia exata quando a amostra é pequena, normalmente isto não se torna um grande empecilho.

De fato, o procedimento PROC LOGISTIC permite o ajuste de diversos tipos de modelos, para resposta dicotômica ou ordinal. Particularmente os modelos para resposta dicotômica, permitem fazer inferência exata sobre os parâmetros de regressão, mediante a utilização dos algoritmos desenvolvidos por Hirji, Mehta e Patel (1987) e Hirji (1992). Veja SAS (What's New). A sintaxe básica para ajustar o modelo de regressão logística exata é

PROC LOGISTIC *<EXACTONLY>*

<EXACTOPTIONS *(opções)* **>;**

MODEL *desfecho <(desfecho_opções)> = <efeitos> < /opções* **>;**

EXACT *<'rótulo'> efeitos <\opções* **>;**

Os itens no interior das marcas <> são opcionais e, portanto, podem ser omitidos. A opção *EXACTONLY* suprime os resultados da análise pela máxima verossimilhança não condicional, de tal forma que apenas a análise exata é executada. As opções especificadas entre os parênteses do item *EXACTOPTIONS* são aplicadas em toda a análise exata requisitada. As opções disponíveis são *MAXTIME* e *STATUSTIME*. A primeira especifica o tempo máximo (em segundos) que o procedimento PROC LOGISTIC pode usar pra determinar as permutações da distribuição da estatística suficiente. O valor atribuído como padrão (*default*, em inglês) é de 7 dias e, no caso de exceder o limite de tempo, o processamento é interrompido, sendo exibida uma mensagem no SAS LOG. Por sua vez, a opção *STATUSTIME* especifica o intervalo de tempo (em segundos) para imprimir a linha de status no SAS LOG. Estas informações podem ser úteis para acompanhar o progresso dos cálculos das distribuições condicionais exatas. O intervalo de tempo especificado é de forma aproximada e não tem valor padrão definido a priori.

O subcomando MODEL é obrigatório e define o modelo que se deseja ajustar. Existem dois conjuntos de opções. O primeiro está associado à variável resposta e é rotulado como *<(desfecho_opções)>*. O outro conjunto, especificado em *</opções>* após os efeitos do modelo, é chamado de opções do modelo.

Por sua vez, as opções do subcomando EXACT são aplicadas apenas aos efeitos especificados. As opções disponíveis são as seguintes:

ALPHA = valor

ESTIMATE < =palavra-chave >

JOINT

JOINTONLY

ONESIDE

OUTDIST = Nome do arquivo de dados

A opção *ALPHA* permite modificar o nível de significância usado para construir os limites de confiança para os parâmetros. Se não for especificado, assume o valor padrão 0,05, produzindo limites com 95% de confiança.

Mediante a opção *ESTIMATE=PARAM* pode-se requisitar as estimativas pontuais e os intervalos de confiança, bem como os testes de hipóteses, individualmente para cada parâmetro (condicional aos outros parâmetros) especificados na opção EXACT. Estes resultados são gerados automaticamente, ou seja, não é necessário especificar a opção *ESTIMATE=PARAM*, mas também é possível solicitar apenas as estimativas pontuais e intervalos de confiança para a razão de chances associada aos efeitos, mediante *ESTIMATE=ODDS*. Por outro lado, a opção *ESTIMATE=BOTH* produz ambos os resultados.

A opção *JOINT* permite testar a hipótese nula que especifica que os parâmetros associados aos efeitos definidos no subcomando EXACT são simultaneamente iguais a zero. Além deste teste, automaticamente são apresentados os resultados dos testes individuais para estes parâmetros. No entanto, é possível suprimir os resultados dos testes individuais, mediante a opção *JOINTONLY*. A opção *ONESIDE* faz com que todos os p-valores calculados para os parâmetros e para as razões de chances sejam para os testes unilaterais. A opção *OUTDIST* mostra todas as distribuições condicionais para as variáveis definidas na opção *EXACT*. Este conjunto de dados contém todos os possíveis valores das estatísticas suficientes, a probabilidade de ocorrência de cada possível valor e o escore das estatísticas suficientes.

Para ilustrar, é conveniente utilizar o exemplo hipotético descrito por Derr (2000), relativo a um pequeno estudo de dose-resposta. Inicialmente são selecionados 18 indivíduos ao acaso. Para cada indivíduo é alocada, também ao acaso, uma de seis doses da droga, as quais variam de 0 até 5, de tal forma que cada dose é testada em três pacientes. Ao término do experimento é registrado o número de indivíduos não sobreviventes em cada grupo da dose. A Figura 4.1 mostra a rotina computacional usada para ajustar o modelo logístico através do procedimento PROC LOGISTIC do programa SAS. Note que os dados observados são informados nas linhas que antecedem o comando PROC LOGISTIC.

```
data dose;
   input Dose Deaths Total @@;
   datalines;
0 0 3   1 0 3   2 0 3   3 0 3
4 1 3   5 2 3
;
proc logistic data=dose descending;
    model Deaths/Total = Dose;
    exact Dose / estimate=both;
run;
```

Figura 4.1 – Rotina do programa SAS para ajustar o modelo logístico aos dados do estudo de dose-resposta descrito por Derr (2000).

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.1478	1	0.0043 [1]
Score	5.7943	1	0.0161 [1]
Wald	2.7249	1	0.0988 [1]

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-9.4745	5.5677	2.8958	0.0888
Dose	1	2.0804	1.2603	2.7249	0.0988 [2]

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Dose	8.007	0.677	94.679 [3]

Exact Conditional Analysis
Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Dose	Score	5.4724	0.0245	0.0190 [4]
	Probability	0.0110	0.0245	0.0190

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
Dose	1.7999	0.1157	5.8665	0.0245 [5]

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
Dose	6.049	1.123	353.000 [6]	0.0245

Figura 4.2 – Resultados do estudo de dose-resposta descrito por Derr (2000) utilizando o procedimento PROC LOGISTIC do SAS.

A Figura 4.2 apresenta os resultados das análises assintótica e exata, sendo que alguns aspectos importantes assinalados são descritos abaixo:

(1) – Testes de hipóteses assintóticos para avaliar a veracidade da hipótese nula que especifica que todos coeficientes de regressão são simultaneamente nulos, isto é, $H_0 : \beta = 0$;

(2) – Estimativas pontuais e testes de hipótese individuais para os parâmetros do modelo, derivados pelo método da máxima verossimilhança;

(3) – Estimativa pontual e intervalo de confiança para a razão de chances, derivados pelo método da máxima verossimilhança;

(4) – Estatística do teste exato baseado em escores e do teste de probabilidade para avaliar a hipótese nula que especifica que todos os parâmetros dos efeitos especificados são simultaneamente nulos; o *mid-p-value* é calculado da mesma forma que o p-valor exato, mas como os valores da estatística são discretos, o *mid-p-value* inclui somente a metade da probabilidade de ocorrência do ponto crítico;

(5) – Estimativa pontual exata e correspondente intervalo de confiança para um parâmetro individual (condicional aos demais parâmetros), bem como o *p-valor* associado ao teste exato para avaliar a hipótese nula que especifica que o parâmetro é igual a zero;

(6) – Estimativa pontual exata e correspondente intervalo de confiança para a razão de chances, bem como o *p-valor* associado ao teste para avaliar se a razão de chances é igual a 1.

É conveniente mencionar que uma descrição detalhada dos recursos disponíveis no procedimento PROC LOGISTIC pode ser vista em SAS (1999). No exemplo em discussão, os testes baseado em escores (*p=0,0043*) e da razão de verossimilhança (*p=0,0161*) da análise assintótica corriqueira, conduzem à rejeição da hipótese nula que especifica que o coeficiente de regressão associado ao efeito da dose é nulo. No entanto, o *p-valor* associado ao teste de Wald é *p=0,0988*, sugerindo a não rejeição da hipótese nula. Note, também, que o intervalo de confiança para a razão de chances contém o valor 1, sugerindo que o

aumento da dose da droga não produz impacto sobre a mortalidade. Este aparente conflito nas conclusões é um sinal revelador que as aproximações da análise assintótica não são confiáveis.

De fato, a análise assintótica não parece razoável para este problema, haja vista que o tamanho da amostra é realmente pequeno, ou seja, foram alocados apenas 3 indivíduos para cada dose da droga. Um procedimento alternativo é usar os recursos dos testes exatos disponíveis no PROC LOGISTIC, mediante a instrução "exact dose / estimate=both" na rotina computacional apresentada na Figura 4.1. Os resultados da análise exata são mais coerentes, pois ambos testes (probabilidade e de escores) sugerem que o aumento da dose da droga induz a um aumento da mortalidade (*p=0,0190*). Ainda, a estimativa pontual para o coeficiente de regressão é $\hat{\beta} = 1{,}7999$, cujo *p-valor* associado ao teste condicional exato para testar $H_0{:}\beta = 0$ é *p=0,0245*, também sugerindo a associação entre a dose e a mortalidade conforme descrito anteriormente. A estimativa exata para a razão de chances é igual a 6,049 e o correspondente intervalo com 95% de confiança é (1,1; 353,0). A interpretação geral do estudo indica que para cada aumento de 1 unidade da droga administrada aos indivíduos, corresponde a um aumento de aproximadamente 6 vezes na chance de não sobrevivência. Entretanto, é importante observar a amplitude excessivamente grande do intervalo derivado pelo método exato. Uma discussão detalhada deste exemplo pode ser vista em Derr (2000), onde o autor conclui que, face ao tamanho da amostra pequeno e aos resultados conflitantes dos testes assintóticos, a análise exata parece ser mais apropriada.

Este exemplo mostra claramente como as conclusões da análise assintótica e da análise condicional exata podem ser conflitantes. Stokes, Davis e Koch (1995) apud Derr (2000, p. 2)[4] observaram que, de forma geral, o método exato tente a produzir resultados mais conservadores, sendo recomendado quando o tamanho da amostra é pequeno e os *p-valores* são inferiores a 0,1.

---

[4] STROKES, M. E.; DAVIS, C. S.; KOCH, G. G. (1995) **Categorical Data Analysis Using the SAS System.** Cary, NC: SAS Institute Inc, Second edition.

4.2 Procedimento Proc-LogXact 4

O Proc-LogXact 4 é uma versão especial do programa LogXact 4, desenvolvido especialmente para usuários do SAS, versão 8.1 ou posterior, mas com os mesmos recursos disponíveis no LogXact 4. O Proc-LogXact permite fazer inferências através dos procedimentos de máxima verossimilhança não condicional, máxima verossimilhança condicional e inferência condicional exata para o modelo de regressão logística e de Poisson.

A estimação de parâmetros pelo método da máxima verossimilhança não condicional consiste essencialmente em maximizar a função de verossimilhança. Sob este enfoque, os testes de hipóteses podem ser realizados através da estatística de Wald, da razão de verossimilhança ou pela estatística baseada em escores, comparando o valor observado com a distribuição assintótica de Qui-quadrado. Na maioria dos programas estatísticos, estes são os únicos métodos disponíveis. Veja, por exemplo, SAS (1999).

O procedimento da máxima verossimilhança condicional está implementado no Proc-LogXact através dos métodos de *máxima verossimilhança condicional parcial* e *completa*. O primeiro é um método assintótico de inferência que, na atual versão, está implementado apenas para o ajuste do modelo logístico, sendo aplicável quando existe uma variável de estratificação, tal como estudos de casos e controles ou estudos multi-cêntricos. Nestas situações, o modelo logístico especifica que cada estrato possui um termo constante próprio, enquanto os demais parâmetros do modelo não variam com os estratos. Usualmente os termos constantes não são importantes e, portanto, podem ser eliminados da função de verossimilhança através do condicionamento em suas estatísticas suficientes. Os demais parâmetros são estimados mediante a maximização desta função de verossimilhança condicional. Os testes de hipóteses são realizados através da estatística de Wald, da razão de verossimilhança ou de escores eficientes, comparando os valores observados com a distribuição assintótica de Qui-quadrado. Veja Cytel (2001, p.4).

O método assintótico da máxima verossimilhança condicional completa, por sua vez, consiste em maximizar a função de verossimilhança condicional em relação a todas as

estatísticas suficientes, exceto àquelas associadas aos parâmetros envolvidos na inferência corrente. Veja Cytel (2001, p.4).

Finalmente, o método de inferência condicional exata é uma extensão lógica do procedimento de máxima verossimilhança condicional completa. Inicialmente obtém-se a função de verossimilhança condicional completa, que não depende de nenhum parâmetro de ruído. As inferências exatas baseadas nas permutações da distribuição da estatística de teste apropriada (por exemplo, as estatísticas de Wald condicional, de razão de verossimilhança ou de escores eficientes). Este procedimento foi sugerido por Cox (1970) e pode ser aplicado tanto para dados estratificados quanto não estratificados. Veja Cytel (2001).

É importante salientar a qualidade da documentação do programa Proc LogXact 4. De fato, o manual apresenta a metodologia com extrema precisão e clareza, incluindo o desenvolvimento matemático formal dos procedimentos de estimação e testes de hipóteses. Além disso, inúmeros exemplos reais são usados para ilustrar as rotinas necessárias para ajustar os modelos e, principalmente, discutir a interpretação dos resultados. Alguns aspectos básicos sobre a utilização do programa Proc LogXact 4 são apresentados a seguir, mas sugere-se ao leitor a consulta do manual.

Usuários do programa SAS não encontrarão maiores dificuldades para usar o procedimento Proc LogXact. De fato, o Proc LogXact deve ser utilizado através do programa SAS e invocado como se fosse um procedimento do SAS. Mesmo correndo o risco da redundância, convém mencionar que entrada dos dados é feita através das maneiras convencionais do programa SAS. A Figura 4.3 ilustra a sintaxe básica do procedimento Proc LogXact 4, sendo brevemente comentada a seguir. Os elementos que aparecem em letras maiúsculas e em negrito são palavras-chave. Aqueles em *itálico* são nomes ou valores numéricos que devem ser informados pelo usuário. Os elementos dispostos entre colchetes [] são opcionais, enquanto que aqueles dentro dos símbolos {} indicam parâmetros separados que devem ser escolhidos entre opções separadas por uma barra vertical. Uma descrição detalhada das facilidades disponíveis e exemplos pode ser vista em Cytel (2001).

```
PROC LOGXACT [<opções>];
        {CLASS_HI | CLASS_LO} lista das variáveis;
        {STRATUM | ST} nome da variável;
        {FREQ} nome da variável;
        {RATE | MLTRATE} nome da variável;
        MODEL desfecho = <efeitos> [/<m1-opções>];
        MODEL eventos/n° ensaios = <efeitos> [/<m2-opções>];
        [<rótulo:>] {TEST | TE} [/<opções dos testes>] efeitos;
        {ESTIMATE | ES} [/<opções de estimação>] efeitos;
        BY lista de variáveis;
```

Figura 4.3 – Sintaxe básica do procedimento Proc LogXact

A instrução PROC LOGXACT é usada para invocar, no programa SAS, o procedimento Proc LogXact. A lista completa de opções é mostrada no Quadro 4.1 e descrita no manual do usuário. Veja Cytel (2001, p.28).

As instruções CLASS_HI e CLASS_LO são usadas para definir como referência o nível mais alto e mais baixo, respectivamente, para as variáveis especificadas na lista de variáveis. A instrução STRATUM especifica a variável que identifica os estratos na análise estratificada, enquanto que FREQ é usada para definir a variável com as freqüências na regressão logística binária. As opções RATE e MLTRATE são usadas apenas no modelo de regressão de Poisson e identificam a taxa multiplicadora.

Para dados não agrupados, o modelo que se deseja ajustar deve ser especificado através da instrução MODEL, onde a palavra *desfecho* identifica a variável resposta. Os efeitos principais e as interações devem ser separados por um espaço em branco. O evento que se deseja modelar pode ser especificado através da opção EVENTCODE, CODE ou simplesmente CO. O valor padrão utilizado pelo programa é 1. A opção LINK pode ser usada para especificar o tipo de modelo desejado, cujas alternativas são {PREG|POISSON}

para o modelo de regressão de Poisson e {LREG|LOGISTIC} para o modelo logístico, sendo este último o modelo assumido como padrão.

A instrução [*rótulo*:] é usada para especificar os rótulos que identificam os resultados dos testes, quando mais de um comando TEST é utilizado. Os tipos de testes disponíveis são os seguintes:

a) {AS | EX | MO | MC}. A opção AS é usada para requisitar os testes assintóticos; EX para testes exatos e MO e MC para os métodos Monte Carlo e Monte Carlo através de cadeias de Markov (*Monte Carlo Markov Chain*, em inglês). Se o tipo de teste não é especificado, o programa gera os testes assintóticos, ou seja o valor padrão é AS. O diagrama disposto na Figura 4.4 pode ser útil para comparar os testes não assintóticos;
b) a opção TESTDISTFILE = *nome da saída do SAS* é opcional e pode ser usada para os tipos de testes EX, MO, MC e especifica a base de dados onde será armazenada a distribuição da estatística teste;
c) a opção TEST_TYPE = {"SCORES" | "PROB" | "LHRATIO"} especifica o tipo de teste utilizado, sendo que "SCORES" requisita o teste baseado nos escores eficientes, "PROB" é usado para solicitar o teste de probabilidade, e "LHRATIO" requisita o teste da razão de verossimilhança. Quando a opção TEST_TYPE não é usada, o programa assume "SCORES" como padrão, quando EX é especificado. Se MO ou MC é especificado, o único teste disponível é o da razão de verossimilhança, identificado pela opção "LHRATIO".

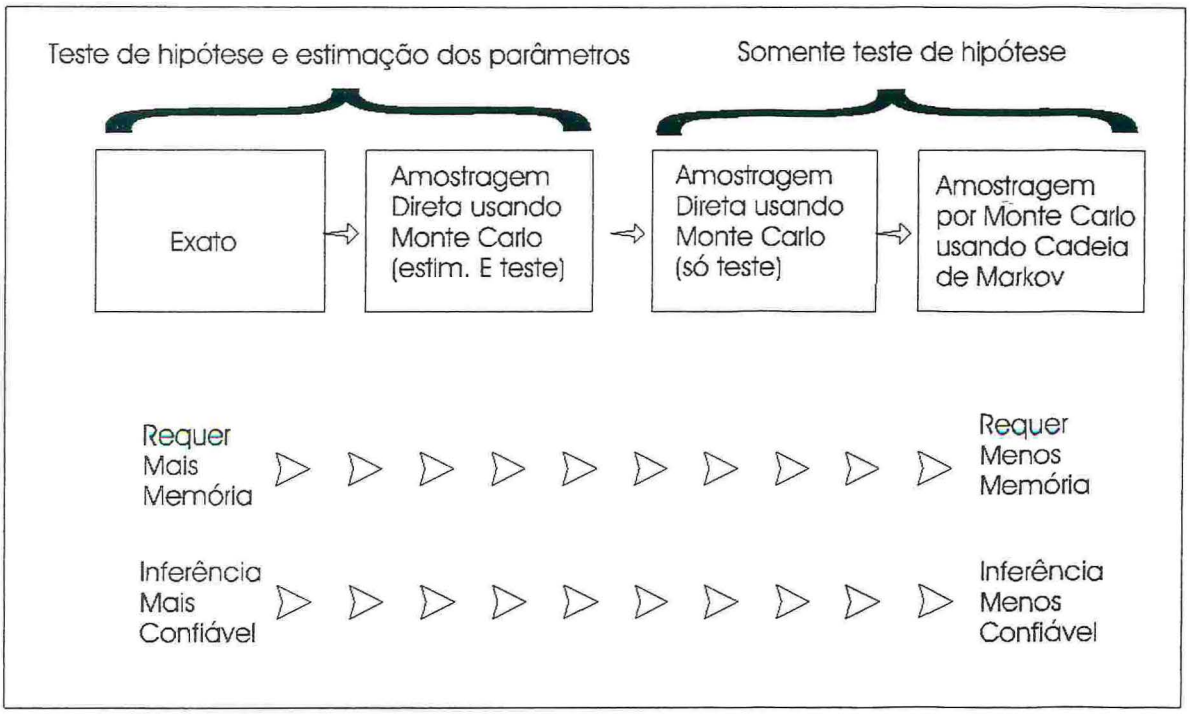

Figura 4.4 – Comparação entre os tipos de testes.

Finalmente, a opção ESTIMATE (ou ES) é usada para especificar o método de estimação, cujos métodos disponíveis são {AS | EX | MO}, sendo que o método assintótico é o padrão. Outros detalhes desta opção são descritos no manual do usuário.

Quadro 4.1 – Lista de opções para o procedimento PROC LOGXACT.

Nome do item	Valores	Valor padrão	Descrição
DATA = *nome do conjunto de dados*			Nome conjunto de dados
OUTPUT = *nome do conjunto de dados*			Nome do arquivo de saída
ALPHA=	0,001 - 0,999	0,95	Alfa
SCORE_ACC | SC =	0 – 8	3	Precisão do escore
AS_ITER | AS_IT =	0 – 100	20	Iterações assintóticas
MAX_MEM | ME =	1-2048	4	Memória máxima em MB
MAX_TIME | TI =	1 - 10^6	5	Tempo máximo em min
OUT_TYPE | OU =	"BETA", "ODDS"	"BETA"	Tipo de saída
PVALUE_TYPE | PV =	"ONE_SIDE", "TWO_SIDE"	"TWO_SIDE"	Tipo de p-valor
DISP_ACC | DI =	-1 a 6	4	Precisão da saída
MC | CL =	0,001 - 0,999	0,95	IC para MC & MCMC
RA | SEED	0 - 10^9	0	Semente, gerada ao acaso, para MC & MCMC
DI_FREQ | FR =	0 - 10^9	1000	Freqüência Para MC & MCMC
MO_SAMPLE | MO_CR =	1 - 10^9	10000	Tam. Amostra para MC
MC_SAMPLE | MC_CR =	1 - 10^9	10000	Tam. Amostra para MCMC
MC_ITER | MC_IT =	1 – 10^6	100	Iterações por cadeias de MCMC
NOPRINT		*Print*	Não mostra saída

O estudo de dose-resposta apresentado na Seção 4.1 e descrito por Derr (2000), é útil para exemplificar a utilização do procedimento Proc LogXact. O modelo logístico pode ser ajustado mediante a rotina abaixo. Observe que as instruções são similares às do procedimento PROC LOGISTIC.

```
data dose;
input Dose Deaths Total @@;
datalines;
0 0 3   1 0 3   2 0 3   3 0 3
4 1 3   5 2 3
;
PROC LOGXACT DATA=dose;
MODEL Deaths/Total= Dose;
ES/EX INTERCEPT;
RUN;
```

Figura 4.5 – Rotina do Proc LogXact para ajustar o modelo logístico aos dados do estudo de dose-resposta descrito por Derr (2000).

Output created by Proc-Logxact 4 for Windows from Cytel Software Corporation (c)

Binary Logistic Regression

Basic Information

Data file name	DOSE
Model	Deaths=Intercept+Dose[1]
Groupsize variable	Total
Stratum variable	<Unstratified>[2]
Analysis type	Estimate : Exact[3]
Number of terms	2
Total observations	18
Observations rejected	0
Number of groups	6

Summary statistics

Statistic	Value	DF	P-value
Deviance	0.4343	4	0.9796
Likelihood Ratio[4]	16.8809	2	0.0002

Parameter Estimates

	Point Estimate			Confidence interval and P-value for Beta			
	Type	Beta	SE(Beta)	Type	95.0% C.I. Lower	Upper	Pvalue 2*1-sided
Dose	MLE	2.0804[6]	1.2603	Asymptotic	-0.3897	4.5505	0.0988[8]
	CMLE	1.8000[5]	1.0784	Exact	0.1157	5.8665	0.0245[7]
Intercept	MLE	-9.4746[6]	5.5677	Asymptotic	-20.3872	1.4380	0.0888[8]
	MUE	-3.2208[5]	NA	Exact	-INF	-0.9861	0.0026[7]

Figura 4.6 – Resultados do estudo de dose-resposta descrito por Derr (2000) utilizando o PROC LOGXACT.

A Figura 4.6 apresenta os resultados da análise, cabendo destacar os seguintes aspectos:

(1) – Identifica o modelo ajustado;

(2) – Identifica se os dados são estratificados ou não estratificados, bem como a variável de estratificação;

(3) – Indica qual o tipo de análise utilizado (assintótico ou exato);

(4) – Testes de hipóteses exatos para avaliar a hipótese nula de que todos os parâmetros são simultaneamente iguais a zero, ou seja, $H_0:\beta = 0$.

(5) – Estimativas dos parâmetros mediante os métodos exatos;

(6) – Estimativas assintóticas dos parâmetros;

(7) – *p-valores* associados aos testes exatos para avaliar a hipótese nula que especifica que o parâmetro é nulo;

(8) – *p-valores* associados aos testes assintóticos para avaliar a hipótese nula que especifica que o parâmetro é nulo;

Os resultados gerados pelo procedimento Proc-LogXact estão divididos em basicamente três partes. A primeira disponibiliza informações básicas, tais como o modelo, tipo de análise, presença de estratificação, número de observações, etc. A segunda apresenta os testes para avaliar a hipótese nula que especifica que os coeficientes de regressão são simultaneamente nulos. No exemplo em discussão, o teste da razão de verossimilhança conduz à rejeição da hipótese nula, sugerindo que a taxa de mortalidade se modifica com o aumento das doses da droga.

Finalmente, a terceira parte apresenta as estimativas, testes de hipóteses e intervalos de confiança para os parâmetros do modelo, pelos métodos exatos e assintóticos. Nos resultados do exemplo, é importante observar a discrepância entre os *p-valores* associados aos testes exato (*p=0,0245*) e assintótico (*0,0988*) para avaliar a hipótese nula em que o coeficiente de regressão é igual a zero, conduzindo a conclusões divergentes. Convém lembrar que os resultados são similares àqueles gerados pelo procedimento PROC LOGISTIC, e que este conflito nas conclusões já havia sido percebido. Entretanto, face ao pequeno tamanho da amostra, a análise exata parece mais apropriada para este caso.

Outros aspectos dos procedimentos PROC LOGISTIC e PROC LOGXACT serão explorados no próximo capítulo, que trata da aplicação da metodologia em estudo.

Para maiores detalhes sobre as rotinas dos programas utilizados para realizar estes exemplos e sobre suas saídas, consulte o Anexo C.

5. APLICAÇÕES

A finalidade deste capítulo é ilustrar aplicações da metodologia de regressão logística exata em situações reais. Também se deseja abordar aspectos computacionais para o ajuste dos modelos, bem como da interpretação dos resultados. Para tanto, foram selecionados dois conjuntos de dados, que são úteis para ilustrar distintas peculiaridades que podem surgir durante a utilização da metodologia. Embora os exemplos utilizados estejam descritos na literatura, são importantes para clarificar as potencialidades e limitações dos métodos de inferência exata e assintótica para o ajuste do modelo logístico.

EXEMPLO A: Os dados deste exemplo foram gerados em estudo com 46 pacientes com sarcoma osteogênico não-metastático, conduzido por Goorin et al. (1987) e, posteriormente, descrito por Mehta & Patel (1995) e Cytel (2001, p.60). O objetivo básico é determinar preditores para estimar a probabilidade de não haver recidiva da doença em um intervalo de três anos. Este desfecho é representado pela variável rotulada como DFI3, onde os valores 0 e 1 indicam, respectivamente, a não ocorrência e a ocorrência da recidiva no período de três anos. As variáveis explanatórias de interesse são SEXO (0=Feminino, 1=Masculino), presença de alguma patologia óssea, rotulada por AOP (acrônimo do nome original "*any osteoid pathology*", onde 0=Não e 1=Sim) e presença de infiltração linfocítica, rotulada por LYINF (acrônimo do nome original "lymphocytic infiltration", onde 0=Não e 1=Sim). Os dados são mostrados na Tabela 5.1, onde a coluna denominada FREQ representa o número de indivíduos em cada contexto.

Em uma etapa de triagem inicial, o *teste exato de Fisher* pode ser usado para avaliar se a ocorrência ou não de recidiva em três anos está associada com as variáveis explanatórias. As tabelas de contingência 2x2 resultantes do cruzamento do desfecho DFI3 podem ser úteis para investigar e interpretar as associações, bem como para requisitar o correspondente teste exato de Fisher nos programas estatísticos usuais. Embora não seja difícil gerar estas tabelas, elas são mostradas em Cytel (2001, p.60).

Tabela 5.1 – Conjunto de dados relativos ao estudo sobre sarcoma osteogênico.

DFI3	LYINF	SEXO	AOP	FREQ
1	0	0	0	3
1	0	0	1	2
1	0	1	0	4
1	0	1	1	1
1	1	0	0	5
1	1	0	1	3
1	1	1	0	5
1	1	1	1	6
0	1	0	1	2
0	1	1	0	4
0	1	1	1	11

Fonte: Mehta & Patel (1995).

Os *p-valores* associados ao teste exato de Fisher bilateral nas tabelas de contingência resultantes dos cruzamentos (DFI3 × LYINF), (DFI3 × SEXO) e (DFI3 × AOP) são, respectivamente, *p=0,0075*, *p=0,0259* e *p=0,0322*. Isto sugere que a ocorrência ou não da recidiva está associada com as variáveis explanatórias sob investigação. Entretanto, o estudo tem como finalidade avaliar o impacto simultâneo das variáveis explanatórias sobre o desfecho. Usualmente isto pode ser realizado mediante o modelo

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \alpha + \sum_{i=1}^{3} \beta_i x_{ij},$$

onde os rótulos x_{1j}, x_{2j} e x_{3j} representam, respectivamente, os valores assumidos pelos preditores LYINF, SEXO e AOP, enquanto que π_j é a probabilidade de não haver recidiva da doença no período de três anos, para o j-ésimo indivíduo. Contudo, não é possível ajustar o modelo logístico através do método da máxima verossimilhança, pois a variável x_{1j} (LYINF) é um preditor perfeito (não há recidiva em 3 anos para indivíduos sem

infiltração linfática). Este fato pode ser constatado na Tabela 5.2, que dispõe o cruzamento das variáveis LYINF e DFI3, onde uma das células de contingência contém freqüência zero. Como conseqüência, o logaritmo da função de verossimilhança não pode ser maximizado, pois não é possível avaliar as derivadas de primeira e de segunda ordem no estimador de máxima verossimilhança (EMV). Portanto, o coeficiente de regressão β_1 e correspondente intervalo de confiança não podem ser obtidos pelo método convencional da máxima verossimilhança.

Tabela 5.2 – Tabela de contingência resultante do cruzamento entre a infiltração linfocítica e o desfecho (3 anos livres da doença).

DFI3	Infiltração linfocítica		Total
	Não	Sim	
Não	0	17	17
Sim	10	19	29
Total	10	36	46

Contudo, parece ser possível analisar os dados através do método de inferência condicional exata, cujos aspectos metodológicos foram descritos na Seção 3.1. Para tanto, considere (t_0, t_1, t_2, t_3) o vetor que contém as estatísticas suficientes associadas aos parâmetros do modelo logístico, representados pelo vetor $(\alpha, \beta_1, \beta_2, \beta_3)$, onde $t_j = \sum_{i=1}^{46} y_i x_{ji}$. Note que os valores observados das estatísticas suficientes são $t_0 = 29$, $t_1 = 19$, $t_2 = 16$ e $t_3 = 12$.

A Tabela 5.3 apresenta a distribuição do número de permutações $c(t_0 = 29,\ t_1,\ t_2 = 16,\ t_3 = 12)$, para todos os possíveis valores de t_1. É importante observar que apesar de existir aproximadamente 800 milhões de seqüências binárias $(y_1, y_2, ..., y_{46})$, existem apenas 8 vetores distintos das estatísticas suficientes da forma $(t_0 = 29,\ t_1,\ t_2 = 16,\ t_3 = 12)$, pois muitas das seqüências binárias produzem os mesmos valores das estatísticas suficientes.

Convém observar, também, que o valor observado $t_1 = 19$ é o menor valor possível para t_1, sendo esta a razão pela qual não é possível estimar os parâmetros do modelo logístico pelo método assintótico da máxima verossimilhança.

Tabela 5.3 – Distribuição condicional exata de t_1.

t_1	$c(29, t_1, 16, 12)$
19	29.445.360
20	147.312.480
21	271.271.448
22	231.819.344
23	95.325.644
24	17.473.144
25	1.204.008
26	19.448
Total	793.870.896

Ao contrário do método assintótico, agora é possível fazer inferências exatas sobre o coeficiente de regressão associado à variável LYINF. Por exemplo, o *p-valor* associado ao teste para avaliar a hipótese nula $H_0 : \beta_1 = 0$, contra a alternativa $H_0 : \beta_1 \neq 0$, é obtido por

$$p = 2 \times \frac{29.445.360}{793.870.896} = 0{,}0742,$$

sugerindo uma possível associação entre a variável LYINF com o desfecho. Note que esta informação seria perdida pelo método assintótico da máxima verossimilhança, haja vista que o correspondente estimador para β_1 não existe.

Por sua vez, como o valor observado $t_1 = 19$ é o menor valor possível para a estatística suficiente t_1, o estimador pontual para β_1 pelo método condicional exato é o estimador mediano não viciado definido pela equação (3.19), na Seção 3.3.1 .

Na Tabela 5.4 e 5.5 é apresentado um resumo dos resultados obtidos pelo método assintótico da máxima verossimilhança e pelo método condicional exato, respectivamente,

os quais foram gerados pelos procedimentos PROC LOGISTIC e PROC LOGXACT. No Anexo A1 consta a rotina computacional utilizada para ajustar o modelo logístico. O Anexo A2 contém a listagem completa dos resultados gerados pelo procedimento PROC LOGISTIC, enquanto que aqueles gerados pelo PROC LOGXACT estão disponíveis no Anexo A3, não havendo, porém, discrepâncias significativas entre os resultados destes dois procedimentos, apenas os sinais dos coeficientes são inversos devido à parametrização. A outra diferença, mas nada significativa, é que na saída do PROC LOGISTIC do SAS aparecem os resultados por máxima verossimilhança, mas juntamente com estes resultados, o SAS exibe um aviso de que as estimativas por máxima verossimilhança podem não existir e a validade dos resultados é questionável, enquanto que o PROC LOGXACT prefere não calcular as estimativas.

Tabela 5.4 – Estimativas pontuais, por intervalo e testes de hipóteses para os parâmetros do modelo logístico ajustados aos dados do Exemplo A, pelo PROC LOGISTIC.

Parâmetro	**Máxima verossimilhança**		**Exato**	
	Estimativa	*p-valor*	Estimativa	*p-valor*
Intercepto	14,7426	0,9489	3,5350	0,0002
LYINF	-12,6349	0,9562	-1,8859	0,0742
SEX0	-1,6362	0,0729	-1,5479	0,1392
AOP	-1,2204	0,1135	-1,1561	0,2182

Tabela 5.5 – Estimativas pontuais, por intervalo e testes de hipóteses para os parâmetros do modelo logístico ajustados aos dados do Exemplo A, pelo PROC LOGXACT (somente método exato).

Parâmetro	**Estimativa**	**IC 95%**	***p-valor***
Intercepto	-3,5351	($-\infty$ a -1,4765)	0,0002
LYINF	1,8860	(-0,1615 a ∞)	0,0742
SEXO	1,5479	(-0,3627 a 4,0238)	0,1392
AOP	1,1561	(-0,5114 a 2,9972)	0,2182

Uma inspeção na Tabela 5.4 revela as enormes discrepâncias entre os resultados produzidos pelos métodos assintótico e exato, tanto nas magnitudes das estimativas, quanto nos *p-valores*. Em particular, observe que estes métodos poderiam conduzir a decisões e interpretações opostas sobre a importância da variável LYINF. De fato, é uma situação artificial, pois o estimador de máxima verossimilhança para o coeficiente de regressão associado à variável LYINF não existe, conforme mostra a saída do PROC LOGXACT, que fornece apenas as estimativas pelo método exato.

Para reforçar os problemas que tem a máxima verossimilhança, pode-se observar a tabela 5.6 que mostra as estimativas para a razão de chances de acordo com o método da máxima verossimilhança e o método exato. É completamente não-informativo o intervalo gerado pela máxima verossimilhança para a infiltração linfocítica, pois cobre toda a amplitude de valores possíveis para a razão de chances. Já o intervalo de confiança pelo método exato diz que, com 95% de confiança, a razão de chance de que um paciente, que tenha infiltração linfocítica em relação a um que não tenha, manifeste a não ocorrência da recidiva no período de três anos não excede 1,175.

Tabela 5.6 – Estimativas pontuais e por intervalo para a razão de chances do modelo logístico ajustados aos dados do Exemplo A.

	Máxima verossimilhança		**Exato**	
Efeito	estimativa pontual	IC 95%	estimativa pontual	IC 95%
LI	<0,001	(<0,001 a >999.999)	0,152	(0 a 1,175)
SEX	0,195	(0,033 a 1,164)	0,213	(0,018 a 1,437)
AOP	0,295	(0,065 a 1,338)	0,315	(0,050 a 1,668)

Um aspecto importante a ser considerado é o chamado *supercondicionamento* (overconditioning, em inglês), descrito por Cox & Snell (1989, p.185) e Cytel (2001, p.195). O problema do supercondicionamento pode produzir testes de hipóteses muito conservadores e, algumas vezes, perda de informação. Em resumo, testes conservadores podem surgir devido ao fato de que o nível de significância fixado pode não ser atingível em testes de hipóteses em que a distribuição de probabilidade da estatística de teste é discreta.

Uma maneira prática para evitar eventuais perturbações decorrentes de supercondicionamento é utilizar um nível de significância menos exigente para os testes de hipóteses na inferência exata. Embora a razão seja a mesma, Stokes, Davis e Koch (1995) recomendam considerar um nível de significância de 0,10 para as inferências exatas.

EXEMPLO B: A incomum doença conhecida como *clostridium difficile* é basicamente caracterizada por ser uma forma aguda de diarréia e está associada a pacientes hospitalizados por longos períodos e usuários de antibióticos. Os dados mostrados na Tabela 5.7, descritos por Cytel (2001, p.139), foram gerados na investigação de um conjunto de 2493 pacientes, 60 dos quais apresentaram a doença. Inicialmente, cinco características podem estar associadas a ocorrência da doença: idade, sexo, tempo de internação e exposição ao antibiótico clindamicina (Clindomicyn, em inglês). Por sua vez, a exposição ao antibiótico cefalexina (Cephalexin, em inglês) parece ser um fator extremamente importante. Para as variáveis cefalexina e clindamicina, os valores "0" e "1" indicam, respectivamente, a não exposição e exposição aos correspondentes antibióticos. Similarmente, os valores "0" e "1" representam indivíduos do sexo masculino e feminino, respectivamente. Para as categorias da variável idade, indivíduos com 50 anos ou mais foram codificados através do valor "1", e os demais com o valor "0". Finalmente, o tempo de internação é codificado com o valor "1", se o período de hospitalização é superior a uma semana, e "0" em caso contrário.

Tabela 5.7 – Conjunto de dados relativos ao estudo sobre *clostridium difficile*.

Taxa de presença da diarréia	Variáveis explanatórias				
	cefalexina	Clindamicina	Sexo	Idade	Tempo
0/174 (0%)	0	0	0	0	0
1/113 (0,88%)	0	0	0	0	1
0/349 (0%)	0	0	0	1	0
16/451 (3,55%)	0	0	0	1	1
0/213 (0%)	0	0	1	0	0
3/108 (2,78%)	0	0	1	0	1
2/409 (0,49%)	0	0	1	1	0
15/558 (2,69%)	0	0	1	1	1
0/5 (0%)	0	1	0	0	0
0/5 (0%)	0	1	0	0	1
0/8 (0%)	0	1	0	1	0
10/31 (32,26%)	0	1	0	1	1
0/10 (0%)	0	1	1	0	0
1/9 (11,11%)	0	1	1	0	1
1/6 (16,76%)	0	1	1	1	0
6/39 (15,38%)	0	1	1	1	1
1/1 (100%)	1	0	0	1	1
4/4 (100%)	1	0	1	1	1

Fonte: Cytel (2001, p.139)

Inicialmente, é importante observar que apesar do conjunto de dados ser relativamente grande, apenas 2,4% dos indivíduos (60/2493) apresentam a doença. Note, ainda, que apenas cinco indivíduos foram expostos ao antibiótico cefalexina e todos apresentaram a diarréia, fato que pode criar mais dificuldades para a análise.

O objetivo básico é utilizar o modelo de regressão logística para avaliar e estimar o impacto da exposição ao antibiótico cefalexina sobre a probabilidade de ocorrência da doença, ajustado pelos efeitos das covariáveis idade, sexo, tempo de hospitalização e exposição ao antibiótico clindamicina. Para estes dados, contudo, não é possível fazer inferências pelo método assintótico da máxima verossimilhança, pois não há convergência

no processo iterativo da estimação. Conseqüentemente, estes resultados gerados não são válidos, como, por exemplo, a estimativa infinita para a razão de chances associada ao efeito da exposição ao antibiótico cefalexina.

Como a exposição a cefalexina é uma informação potencialmente importante, não parece razoável simplesmente descartar esta variável, até mesmo porque o objetivo principal do estudo é avaliar o seu efeito. Este exemplo ilustra uma situação em que, mesmo com um tamanho de amostra grande, o método assintótico da máxima verossimilhança não é adequado. Uma alternativa é usar os procedimentos de inferência condicional exata descritos na Seção 3.3.

Os resultados gerados pelos procedimentos PROC LOGISTIC e PROC LOGXACT, relativo às estimativas dos parâmetros (pontual e por intervalo) e aos *p*-valores associados aos parâmetros de regressão, são mostrados nas tabelas 5.8 e 5.9. A rotina computacional usada para o ajuste dos modelos está disposta no Anexo B1, enquanto que os Anexos B2 e B3 apresentam as listagens completas dos resultados gerados pelos procedimentos PROC LOGISTIC e PROC LOGXACT, respectivamente. Neste exemplo, infelizmente, não poderemos comparar os resultados dos procedimentos PORC LOGISTC e PROC LOGXACT, pois o primeiro não suporta tamanhos de grupos muito grandes, conseqüentemente não calculando as estimativas para o método exato.

Tabela 5.8 – Comparação ente os métodos da máxima verossimilhança e exato para as estimativas dos parâmetros e respectivos *p-valores*.

	Máxima verossimilhança		**EXATO**	
Parâmetro	Estimativa	*p-valor*	Estimativa	*p-valor*
Cefalexina	34,2982	1,0000	5,3347	<0,0001
Clindamicina	2,2191	<0,0001	2,2093	<0,0001
Sexo	-0,1907	0,5000	-0,1901	0,5923
Idade	0,8922	0,0641	0,8882	0,0710
TEMPHOSP	2,4718	<0,0001	2,4670	<0,0001

Examinando a tabela 5.8, percebe-se a grande diferença na estimativa do parâmetro para cefalexina, e, mais ainda, nos p-valores que estão um em cada extremo. Cabe ressaltar

que estes resultados assintóticos foram obtidos pelo procedimento PROC LOGISTIC, pois o PROC LOGXACT não produziu resultados por máxima verossimilhança por motivos explicados anteriormente no exemplo A.

Tabela 5.8 – Comparação ente os métodos da máxima verossimilhança e exato para as estimativas das razões de chance e seus respectivos intervalos de confiança.

	Máxima verossimilhança		**EXATO**	
Efeito	Estimativa pontual	IC 95%	Estimativa pontual	IC 95%
Cefalexina	>999,999	(<0,001 a >999,999)	207,4128	(27,5208 a ∞)
Clindamicina	9,199	(4,964 a 17,046)	9,1095	(4,6223 a 17,4572)
Sexo	0,826	(0,475 a 1,438)	0,8269	(0,4569 a 1,4977)
Idade	2,441	(0,949 a 6,277)	2,4307	(0,9427 a 8,0035)
TEMPHOSP	11,844	(3,650 a 38,429)	11,7869	(3,7368 a 59,7702)

Uma inspeção da(s) tabelas(s) revela com clareza a enorme divergência dos resultados relativos ao efeito da cefalexina. Note que o intervalo com 95% de confiança derivado pelo método assintótico da máxima verossimilhança para a razão de chances associada à variável cefalexina não é informativo, pois é dado por $(0,+\infty)$. Contudo, pelo método exato o referido intervalo é dado por $(27{,}5;+\infty)$ e significa que, com 95% de confiança, ele contém a verdadeira razão de chances de indivíduos expostos ao antibiótico cefalexina apresentem a doença, em relação aos não expostos, ajustado para as demais covariáveis.

Os *p-valores* dispostos na Tabela 5.8 são correspondentes ao teste para avaliar a hipótese nula que especifica que a razão de chances é igual 1. Note que para a variável cefalexina os valores associados aos métodos assintótico ($p=1{,}000$) e exato ($p<0{,}0001$) são opostos. Naturalmente, o método da máxima verossimilhança não pode ser usado aqui, mas eventualmente poderia conduzir um usuário desatento a concluir equivocadamente em direção contrária aos resultados do método exato.

6. CONSIDERAÇÕES FINAIS

Neste trabalho foram discutidos aspectos básicos de procedimentos exatos para inferências do modelo de regressão logística com resposta binária. Em contraposição ao procedimento assintótico da máxima verossimilhança descrito no Capítulo 2, os métodos exatos podem ser extremamente úteis quando o tamanho da amostra é pequeno, os dados são desbalanceados ou altamente estratificados. Nestes casos, os resultados assintóticos podem não ser confiáveis e é bem conhecido que as propriedades dos estimadores de máxima verossimilhança são ótimas para amostras grandes. Além disso, no Capítulo 2 foram mencionadas as condições para a existência dos estimadores de máxima verossimilhança dos parâmetros do modelo logístico. Veja Albert & Anderson (1984).

O procedimento de inferência estatística exata abordado neste trabalho, geralmente é chamado de regressão logística exata. Apesar de ter sido originalmente proposto por Cox (1970), somente tornou-se factível após o desenvolvimento de computadores com maior capacidade de processamento e algoritmos mais eficientes. Aspectos básicos da metodologia foram considerados no Capítulo 3, que posteriormente é aplicada a dois conjuntos de dados reais apresentados na literatura.

Os Exemplos A e B ilustram a potência do método de regressão logística exata em duas situações distintas. Note que, em ambos os casos, os estimadores de máxima verossimilhança para os parâmetros do modelo logístico não existem. Por outro lado, as inferências exatas permitem extrair e interpretar razoavelmente as informações contidas nos dados.

O ajuste dos modelos foi realizado através dos procedimentos PROC LOGISTIC e PROC LOGXACT discutidos no Capítulo 4. Em linhas gerais, os resultados são bastante similares, porém o programa PROG LOGXACT parece ser potencialmente mais eficiente para amostras grandes. Este fato é ilustrado no Exemplo B, onde o procedimento PROC LOGISTIC não gerou os resultados da inferência exata.

Entretanto, é prudente observar que em algumas situações os métodos exatos apresentados podem não funcionar adequadamente. Por exemplo, nos casos em que não houver memória suficiente disponível para usar os recursos de inferência exata disponíveis no PROC LOGXACT, uma alternativa é usar os métodos de Monte Carlo ou de Monte

Carlo Markov Chain. Estes tópicos não foram abordados no trabalho e são objetos de continuidade da pesquisa. Veja Cytel (2001).

Embora esteja além dos objetivos do trabalho, também é interessante confrontar a metodologia apresentada com as alternativas bayesianas. Como continuidade da pesquisa, é importante explorar os procedimentos de análise de resíduos e os métodos de diagnóstico da regressão. Outra continuação natural da pesquisa é a extensão da metodologia para a situação em que o desfecho é politômico. Veja Hirji (1992).

REFERÊNCIAS BILIOGRÁFICAS:

1. AGRESTI, A. (1990) **Categorical Data Analysis.** John Wiley, New York.
2. ALBERT, A. & ANDERSON, J.A. (1984). **On the existence of maximum likelihood estimates in logistic regression models.** *Biometrika*, 71, 1-10.
3. AZZALINI, A. (1996). **Statistical inference: Based on the likelihood.** Chapman and Hall. London.
4. BICKEL, P.J. & DOKSUM, K.A. (1976). **Mathematical Statistics: Basic Ideas and Selected Topics.** San Francisco, Holden-Day, Inc.
5. BRESLOW, N. E. & DAY, N. E. (1980) **Statistical Methods in Cancer Research – The Analysis of Case-Control Studies.** Lyon, IARC Scientific Publications.
6. CORDEIRO, G. M. (1986). **Modelos lineares generalizados.** VII Simpósio Nacional de Probabilidade e Estatística, Campinas, SP.
7. COX, D.R. (1970) **The Analysis of Binary Data.** Methuen & Co Ltd, London.
8. COX, D. R. & HINKLEY, D. V. (1974) **Theoretical Statistics.** Chapman e Hall, London.
9. COX, D.R. & SNELL, E.J. (1989) **Analysis of Binary Data, Second Edition.** Chapman and Hall, London
10. CURNOW, R.N.; HODGE A.; WILLESMITH, J.W. (1997) **Analysis of the Bovine Spongiform Encephalopathy Maernal Cohort Study: the Discordant Case-Control Pairs.** *Journal of Applied Statistics*, 46, 345-349.
11. CYTEL SOFTWARE CORPORATION (2001). **Proc-LogXact 4 for SAS Users.** Cambridge, MA, USA.
12. DERR, R. E. (2000) **Performing exact logistic regression with the SAS® System.** *Proceeding of the Twenty-Fifth Annual SAS User Group International Conference*, Cary, NC: SAS Institute Inc.
13. FINNEY, D. J. (1971) **Statistical Mehod in Biological Assay.** Griffin London, second edition.
14. HARREL JR., F. E. (2001). **Regression Modeling Strategies: with applications to linear models, logistic regression, and survival analysis.** Springer-Verlag, New York
15. HASTIE, T. J. & TIBSHIRANI, R. J. (1990). **Generalized Additive Models.** Chapman and Hall/CRC, London.
16. HIJRI, K. F. (1992) **Exact Distributions for Polytomous Data.** *Journal of the American Statistical Association*, 87, 487-792.
17. HIJRI, K. F.; METHA, C. R.; PATEL, N. R. (1987) **Computing distributions for Exact Logistic Regression.** *Journal of the American Statistical Association*, 82, 1110-1117.
18. HIJRI, K. F.; METHA, C. R.; PATEL, N. R. (1988) **Exact Inference for Matched Case-Control Studies.** Biometrics, 44, 803-814.
19. HOSMER, D. W. & LEMESHOW, S. (2000). **Applied logistic regression.** Second Edition, John Wiley, New York.
20. MCCULLAGH, P. & NELDER, J. A. (1989) **Generalized linear models.** Second Edition, Chapman and Hall.

21. METHA, C. R. & PATEL, N. R. (1995) **Exact Logistic Regression: Theory and Examples.** Statistics in Medicine, **14**, 2143-2160.
22. METHA, C. R.; PATEL, N. R.; SENCHAUDHURI, P. (1988) **Efficient Monte Carlo Methods for Conditional Logistic Regression.** *Journal of the American Statistical Association*, **95**, 99-108.
23. MOOD, A. M.; GRAYBILL, F. A. e BOES, D. C. (1974) **Introduction to the Theory of Statistics.** Third Edition, Singapore, McGraw-Hill International Editions.
24. NELDER, J. A. & WEDDERBURN, R. W. M. (1972). **Generalized linear models.** *Journal of the Royal Statistical Society.* A 135, 370-84.
25. ROBERT, E. **Performing Exact Logistic Regression with the SAS System.** *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.2000.
26. ROHATGI, V. K. (1976) **An Introduction to Probability Theory and Mathematical Statistics.** New York, Wiley.
27. ROUSSAS, G. G. (1973) **A First Course in Mathematical Statistics.** Reading, Massachusetts, Addison-Wesley Publishing Company.
28. SAS INSTITUTE INC. (1999) **SAS OnlineDoc®.** Version 8, Cary, NC: SAS Institute Inc.
29. SNAPPIN, S. M. & SMALL, R. D. (1986) **Tests of Significance Using Regression Models for Ordered Categorical Data.** Biometrics, **42**, 583-592.
30. TANG, MAN-LAI. (2000) **On tests of linearity for Dose Response Data: Asymptotic, Exact Conditional and Exact Unconditional Tests.** *Journal of Applied Statistics*, **27**, 871-880.
31. TRICHLER, D. (1984) **An Algorithm for Exact Logistic Regression.** *Journal of the American Statistical Association,* **12**, 709-711.
32. VENABLES, W. N. & RIPLEY, B. D. (1999). **Modern Applied Statistical with S-Plus.** Third Edition. Springer-Verlag, New York.
33. VIGO, A. (1994) **Análise de Experimentos Industriais com Respostas Categóricas Ordenadas: Método de Taguchi e Modelo de McCullagh.** Dissertação de Mestrado, UNICAMP, Campinas, São Paulo.
34. WEDDERBURN, R.W.M. (1976). **On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models.** *Biometrika*, **63**, 27-32.
35. WILKS, S.S. (1962). **Mathematical Statistics.** New York, John Wiley & Sons, Inc.

ANEXOS

ANEXO A1

```
data ex51;
input LI SEX AOP DFI3;
DATALINES;
(observações dispostas no capítulo 5)
;
proc logistic;
      model DFI3= LI SEX AOP;
      exact INTERCEPT LI SEX AOP / estimate=both outdist=ex;
proc print data=ex;
run;
PROC LOGXACT DATA=EX51;
MODEL DFI3= LI SEX AOP;
ES/EX INTERCEPT;
RUN;
```

ANEXO A2

The LOGISTIC Procedure

Model Information

Data Set	WORK.EX51
Response Variable	DFI3
Number of Response Levels	2
Number of Observations	46
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	DFI3	Total Frequency
1	0	29
2	1	17

Model Convergence Status

Quasi-complete separation of data points detected.

WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	62.603	50.798
SC	64.431	58.112
-2 Log L	60.603	42.798

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	17.8049	3	0.0005
Score	13.1688	3	0.0043
Wald	5.5103	3	0.1380

The LOGISTIC Procedure

WARNING: The validity of the model fit is questionable.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	14.7426	230.1	0.0041	0.9489
LI	1	-12.6349	230.1	0.0030	0.9562
SEX	1	-1.6362	0.9123	3.2167	0.0729
AOP	1	-1.2204	0.7712	2.5042	0.1135

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
LI	<0.001	<0.001	>999.999
SEX	0.195	0.033	1.164
AOP	0.295	0.065	1.338

Association of Predicted Probabilities and Observed Responses

Percent Concordant	72.0	Somers' D	0.627
Percent Discordant	9.3	Gamma	0.771
Percent Tied	18.7	Tau-a	0.299
Pairs	493	c	0.813

The LOGISTIC Procedure
Exact Conditional Analysis
Sufficient Statistics

Parameter	Value
Intercept	29
LI	19
SEX	16
AOP	12

Conditional Exact Tests

			--- p-Value ---	
Effect	Test	Statistic	Exact	Mid
Intercept	Score	14.2546	<.0001	<.0001
	Probability	0.000079	<.0001	<.0001
LI	Score	4.5416	0.0606	0.0421
	Probability	0.0371	0.0606	0.0421
SEX	Score	3.3707	0.1169	0.0874
	Probability	0.0591	0.1169	0.0874
AOP	Score	2.4797	0.1535	0.1111
	Probability	0.0847	0.1535	0.1111

Exact Parameter Estimates

		95% Confidence Limits		
Parameter	Estimate			p-Value
Intercept	3.5350*	1.4765	Infinity	0.0002
LI	-1.8859*	-Infinity	0.1615	0.0742
SEX	-1.5479	-4.0238	0.3627	0.1392
AOP	-1.1561	-2.9972	0.5114	0.2182

NOTE: * indicates a median unbiased estimate.

Exact Odds Ratios

		95% Confidence Limits		
Parameter	Estimate			p-Value
Intercept	34.295*	4.378	Infinity	0.0002
LI	0.152*	0	1.175	0.0742
SEX	0.213	0.018	1.437	0.1392
AOP	0.315	0.050	1.668	0.2182

NOTE: * indicates a median unbiased estimate.

Obs	Intercept	LI	SEX	AOP	Count	Score	Prob
1	19	.	.	.	176198880	11.2139	0.00047
2	20	.	.	.	2922557040	6.9500	0.00780
3	21	.	.	.	17373912348	3.7012	0.04637
4	22	.	.	.	52711953788	1.4676	0.14070
5	23	.	.	.	93330521492	0.2490	0.24912
6	24	.	.	.	101973552828	0.0455	0.27218
7	25	.	.	.	69584589992	0.8572	0.18573
8	26	.	.	.	28931528336	2.6839	0.07722
9	27	.	.	.	6829338600	5.5257	0.01823
10	28	.	.	.	784414680	9.3826	0.00209
11	29	.	.	.	29445360	14.2546	0.00008
12	.	19	.	.	29445360	4.5416	0.03709
13	.	20	.	.	147312480	1.4937	0.18556
14	.	21	.	.	271271448	0.0981	0.34171
15	.	22	.	.	231819344	0.3548	0.29201
16	.	23	.	.	95325664	2.2638	0.12008
17	.	24	.	.	17473144	5.8251	0.02201
18	.	25	.	.	1204008	11.0387	0.00152
19	.	26	.	.	19448	17.9046	0.00002
20	.	.	14	.	299880	11.4024	0.00060
21	.	.	15	.	4898040	6.7930	0.00984
22	.	.	16	.	29445360	3.3707	0.05914
23	.	.	17	.	87836280	1.1354	0.17641
24	.	.	18	.	144347000	0.0871	0.28990
25	.	.	19	.	135419960	0.2259	0.27197
26	.	.	20	.	72090200	1.5516	0.14478
27	.	.	21	.	20663500	4.0644	0.04150
28	.	.	22	.	2795650	7.7642	0.00561
29	.	.	23	.	121550	12.6510	0.00024
30	.	.	.	8	1360	19.7833	0.00000
31	.	.	.	9	64600	13.9096	0.00019
32	.	.	.	10	1011160	9.0678	0.00291
33	.	.	.	11	7401120	5.2578	0.02129
34	.	.	.	12	29445360	2.4797	0.08469
35	.	.	.	13	68686800	0.7335	0.19756
36	.	.	.	14	97275360	0.0191	0.27979
37	.	.	.	15	84307080	0.3366	0.24249
38	.	.	.	16	44049720	1.6859	0.12670
39	.	.	.	17	13251160	4.0672	0.03811
40	.	.	.	18	2059720	7.4803	0.00592
41	.	.	.	19	123760	11.9253	0.00036

ANEXO A3

Output created by Proc-Logxact 4 for Windows from Cytel Software Corporation (c)

Binary Logistic Regression

Basic Information

Data file name	EX51
Model	DFI3=Intercept+LI+SEX+AOP
Frequency variable	Not specified
Stratum variable	<Unstratified>
Analysis type	Estimate : Exact
Number of terms	4
Total observations	46
Observations rejected	0
Number of groups	8

Summary statistics

Statistic	Value	DF	P-value
Deviance	NA	NA	NA
Likelihood Ratio	NA	NA	NA

Parameter Estimates

	Point Estimate			Confidence interval and P-value for Beta		
	Type	Beta	SE(Beta)	Type	95.0% C.I. Lower	Upper	Pvalue 2*1-sided
LI	MLE	?	?	Asymptotic	?	?	?
	MUE	1.8860	NA	Exact	-0.1615	+INF	0.0742
SEX	MLE	?	?	Asymptotic	?	?	?
	CMLE	1.5479	0.8884	Exact	-0.3627	4.0238	0.1392
AOP	MLE	?	?	Asymptotic	?	?	?
	CMLE	1.1561	0.7506	Exact	-0.5114	2.9972	0.2182
Intercept	MLE	?	?	Asymptotic	?	?	?
	MUE	-3.5351	NA	Exact	-INF	-1.4765	0.0002

ANEXO B1

```
data DIARREIA;
input diarreia totno cephalex clindomy sexo idade TEMPHOSP;
cards;
0  174  0  0  0  0  0
1  113  0  0  0  0  1
0  349  0  0  0  1  0
16  451  0  0  0  1  1
0  213  0  0  1  0  0
3  108  0  0  1  0  1
2  409  0  0  1  1  0
15  558  0  0  1  1  1
0  5  0  1  0  0  0
0  5  0  1  0  0  1
0  8  0  1  0  1  0
10  31  0  1  0  1  1
0  10  0  1  1  0  0
1  9  0  1  1  0  1
1  6  0  1  1  1  0
6  39  0  1  1  1  1
1  1  1  0  0  1  1
4  4  1  0  1  1  1
;
proc logistic;
      model diarreia/totno = cephalex clindomy sexo idade TEMPHOSP;
      exact intercept cephalex clindomy sexo idade TEMPHOSP;
run;
proc LOGXACT data=DIARREIA max_mem=350;
      model diarreia/totno = cephalex clindomy sexo idade TEMPHOSP;
es/ex;
run;
```

ANEXO B2

The LOGISTIC Procedure

Model Information

Data Set	WORK.DIARREIA
Response Variable (Events)	diarreia
Response Variable (Trials)	totno
Number of Observations	18
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	60
2	Nonevent	2433

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	567.772	448.601
SC	573.593	483.528
-2 Log L	565.772	436.601

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	129.1710	5	<.0001
Score	327.8258	5	<.0001
Wald	75.3388	5	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-6.6953	0.7123	88.3463	<.0001
cephalex	1	34.2982	2193623	0.0000	1.0000
clindomy	1	2.2191	0.3147	49.7065	<.0001
sexo	1	-0.1907	0.2828	0.4550	0.5000
idade	1	0.8922	0.4820	3.4271	0.0641
TEMPHOSP	1	2.4718	0.6005	16.9427	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
cephalex	>999.999	<0.001	>999.999
clindomy	9.199	4.964	17.046
sexo	0.826	0.475	1.438
idade	2.441	0.949	6.277
TEMPHOSP	11.844	3.650	38.429

Association of Predicted Probabilities and Observed Responses

Percent Concordant	76.4	Somers' D	0.648
Percent Discordant	11.6	Gamma	0.736
Percent Tied	12.0	Tau-a	0.030
Pairs	145980	c	0.824

ANEXO B3

Output created by Proc-Logxact 4 for Windows from Cytel Software Corporation (c)

Binary Logistic Regression

Basic Information

Data file name	DIARREIA
Model	diarreia=Intercept+cephalex+clindomy+sexo+idade+TEMPHOSP
Groupsize variable	totno
Stratum variable	<Unstratified>
Analysis type	Estimate : Exact
Number of terms	6
Total observations	2493
Observations rejected	0
Number of groups	18

Summary statistics

Statistic	Value	DF	P-value
Deviance	NA	NA	NA
Likelihood Ratio	NA	NA	NA

Parameter Estimates

	Point Estimate		Confidence interval and P-value for Beta			
	Type	Beta	SE(Beta)	Type	95.0% C.I. Lower	Upper	Pvalue 2*1-sided
cephalex	MLE	?	?	Asymptotic	?	?	?
	MUE	5.3347	NA	Exact	3.3149	+INF	8.002e-008
clindomy	MLE	?	?	Asymptotic	?	?	?
	CMLE	2.2093	0.3136	Exact	1.5309	2.8598	1.313e-009
sexo	MLE	?	?	Asymptotic	?	?	?
	CMLE	-0.1901	0.2821	Exact	-0.7832	0.4039	0.5923
idade	MLE	?	?	Asymptotic	?	?	?
	CMLE	0.8882	0.4811	Exact	-0.0590	2.0799	0.0710
TEMPHOSP	MLE	?	?	Asymptotic	?	?	?
	CMLE	2.4670	0.6003	Exact	1.3182	4.0905	3.077e-008
Intercept	MLE	?	?	Asymptotic	?	?	?

ANEXO C

The LOGISTIC Procedure

Model Information

Data Set	WORK.DOSE
Response Variable (Events)	Deaths
Response Variable (Trials)	Total
Number of Observations	6
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	3
2	Nonevent	15

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	18.220	12.072
SC	19.111	13.853
-2 Log L	16.220	8.072

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.1478	1	0.0043
Score	5.7943	1	0.0161
Wald	2.7249	1	0.0988

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-9.4745	5.5677	2.8958	0.0888
Dose	1	2.0804	1.2603	2.7249	0.0988

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Dose	8.007	0.677	94.679

Association of Predicted Probabilities and Observed Responses

Percent Concordant	88.9	Somers' D	0.867
Percent Discordant	2.2	Gamma	0.951
Percent Tied	8.9	Tau-a	0.255
Pairs	45	c	0.933

The LOGISTIC Procedure

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Dose	Score	5.4724	0.0245	0.0190
	Probability	0.0110	0.0245	0.0190

Exact Parameter Estimates

Parameter	Estimate	95% Confidence Limits		p-Value
Dose	1.7999	0.1157	5.8665	0.0245

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
Dose	6.049	1.123	353.000	0.0245

Output created by Proc-Logxact 4 for Windows from Cytel Software Corporation (c)

Binary Logistic Regression

Basic Information

Data file name	DOSE
Model	Deaths=Intercept+Dose
Groupsize variable	Total
Stratum variable	<Unstratified>
Analysis type	Estimate : Exact
Number of terms	2
Total observations	18
Observations rejected	0
Number of groups	6

Summary statistics

Statistic	Value	DF	P-value
Deviance	0.4343	4	0.9796
Likelihood Ratio	16.8809	2	0.0002

Parameter Estimates

	Point Estimate			Confidence interval and P-value for Beta			
	Type	Beta	SE(Beta)	Type	95.0% C.I. Lower	Upper	Pvalue 2*1-sided
Dose	MLE	2.0804	1.2603	Asymptotic	-0.3897	4.5505	0.0988
	CMLE	1.8000	1.0784	Exact	0.1157	5.8665	0.0245
Intercept	MLE	-9.4746	5.5677	Asymptotic	-20.3872	1.4380	0.0888
