Denise Pimenta Nacle

Existência de Estimadores de Máxima Verossimilhança em Modelos de Regressão Logística

Dissertação apresentada ao Curso de Mestrado em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Estatística.

Orientador: Prof. Enrico Antônio Colosimo

Universidade Federal de Minas Gerais
Departamento de Estatística
Belo Horizonte
Novembro de 2004

Agradecimentos

A Deus,

por ter me colocado de pé diante de tantos tropeços e ter me proporcionado tantas alegrias no decorrer dessa caminhada.

Ao mestre Enrico,

pelos conhecimentos adquiridos e por ter contribuído para que me tornasse uma pessoa melhor.

Aos meus Pais,

pela constante dedicação, amor e carinho. Sem vocês eu não teria aprendido a valorizar o ser humano independente de suas origens, a lutar sempre em quaisquer circunstâncias.

Aos meus Alunos,

por serem a inspiração para o exercício do magistério.

Ao meu Marido,

pela tolerância, compreensão e amor dedicados neste período.

Aos meus irmãos, sobrinhos e cunhados,

pelo constante apoio e incentivo.

Sumário

Resumo

O modelo de regressão logística é o método estatístico freqüentemente utilizado para tratar respostas binárias, e a estimação dos seus coeficientes é geralmente feita usando o método de máxima verossimilhança. Mas, como tal método baseia-se em propriedades assintóticas dos estimadores, necessitando de amostras geralmente grandes, os resultados obtidos da teoria assintótica podem não ser adequados ou podem não existir, mesmo quando se dispõe de amostras grandes, mas cujos dados são esparsos.

Os dados logísticos podem ser classificados em três categorias mutuamente exclusivas e exaustivas, segundo Albert e Anderson (1984): Separação Completa, Separação Quase-Completa e *Overlap*. Para as duas primeiras categorias, os estimadores de máxima verossimilhança não existem.

Este trabalho foi motivado por dois bancos de dados reais que estão classificados na Categoria de Separação Quase-Completa e, portanto, os estimadores de máxima verossimilhança não existem. São apresentadas duas propostas de solução da literatura (regressão logística exata e adição de uma pequena constante aos dados) e, ainda, uma nova solução que consiste simplesmente em retirar, aleatoriamente, uma observação de uma das caselas não-nulas (com mesmo valor da covariável ou mesmo valor da resposta) e adicioná-la à casela nula.

Através de Simulações de Monte Carlo, foram comparadas as três propostas de solução quanto ao Erro Quadrático Médio, em que os melhores resultados foram obtidos pela adição de uma pequena constante aos dados e pela eficácia da nova proposta.

Abstract

The logistic regression model is the statistical method frequently used to deal with binary responses and the estimation of their coefficients is usually done using the method of maximum likelihood. But as this method is based on the asymptotic properties of the estimators, it needs sample sizes generally large, so the theory asymptotic's results cannot be appropriate or cannot exist, even when we have the use of large samples, but their data are sparse.

The logistic data can be classified into three mutually exclusive and exhaustive categories, according to Albert and Anderson (1984): complete separation, quasicomplete separation and overlap. For the first two categories, the maximum likelihood estimators do not exist.

This researche has been motivated for two real data sets that are classified on the category of separation quasicomplete and, consequently, there are no maximum likelihood estimators. Then, two proposals from the literature (exact logistic regression and addition of a small constant in data) were discussed and it was presented the new proposal, that consists in taking away randomly the results of any of the non-null cell (with same value from covariable or same response value) and add it to the null cell.

The comparison of the proposals is done by using simulations of Monte Carlo. The criterion used for this comparison was the mean-square error. The best results obtained were based on the addition of a small constant in data and the effectiveness of the new proposal.

Capítulo 1

Introdução

Variáveis categóricas aparecem, com grande freqüência, em estudos da área da saúde. Elas podem ser do tipo nominal (como é o caso da variável sexo) ou ordinal (quando há uma ordenação entre as categorias - como é o caso da variável nível de desnutrição classificado como leve, moderado ou grave). A variável que apresenta apenas duas categorias é chamada de binária ou dicotômica, cujos resultados são codificados, usualmente, como 0 (ausência) ou 1 (presença) de determinada característica. Por exemplo, um paciente pode ter (ou não) uma determinada doença.

O modelo de regressão logística é o método estatístico freqüentemente utilizado para tratar respostas binárias (Hosmer e Lemeshow,1989). A estimação dos coeficientes de um modelo de regressão logística é geralmente feita usando o método de máxima verossimilhança. Mas, como tal método baseia-se em propriedades assintóticas dos estimadores, necessitando de amostras geralmente grandes, os resultados obtidos da teoria assintótica podem não ser adequados ou podem não existir, mesmo quando dispõe-se de amostras grandes, mas os dados são esparsos.

Albert e Anderson (1984) mostraram que os conjuntos de dados logísticos podem ser classificados em três categorias mutuamente exclusivas e exaustivas : Separação Completa, Separação Quase-Completa e *Overlap*. Eles também mostraram que os estimadores de máxima verossimilhança não existem para as duas primeiras categorias.

Este trabalho foi motivado por dois bancos de dados reais que estão classificados na Categoria de Separação Quase-Completa e, portanto, os estimadores de máxima verossimilhança não existem. O primeiro banco de dados refere-se a um grupo de

pacientes submetidos à craniotomia, em que a resposta de interesse é a ocorrência (ou não) de meningite nos primeiros 30 dias após a cirurgia e duas covariáveis são consideradas na explicação dessa variável resposta: Gravidade do caso (0: baixa; 1: alta) e duração da cirurgia (em horas). Este banco de dados foi fornecido pelo Sr. Bráulio Roberto Gonçalves Marinho Couto, sendo proveniente do Hospital São Francisco de Assis, localizado em Belo Horizonte, MG.
O segundo banco de dados refere-se ao estudo de possíveis associações de hemorragia peri-intraventricular no parto e várias covariáveis de interesse. Este banco de dados foi fornecido pelo Professor Marcos Borato Viana (Departamento de Pediatria, UFMG).

Encontra-se também neste trabalho um terceiro banco de dados (Minitab, versão 12.2) referente à taxa de pulsação (0: baixa; 1: alta) de um grupo de indivíduos e duas covariáveis são consideradas na explicação dessa variável resposta: peso (em quilos) e fumante (0: não; 1: sim). Este banco de dados está classificado na categoria de separação *overlap* e, portanto, existem estimadores de máxima verossimilhança.

Nas Seções 1.1 e 1.2, deste capítulo, apresentam-se os modelos lineares generalizados e, o modelo de regressão logística, respectivamente. A Seção 1.3 apresenta os objetivos e a organização desta dissertação.

1.1- Modelos Lineares Generalizados

Os modelos lineares generalizados (MLG), introduzidos por Nelder & Wedderburn (1972), cuja referência clássica é McCullagh e Nelder (1989), são especificados por três componentes: um *componente aleatório*, que identifica a distribuição de probabilidades da variável resposta; um *componente sistemático*, que especifica uma função linear de variáveis explicativas e parâmetros; e uma *função de ligação*, que descreve a relação funcional entre o componente sistemático e o valor esperado do componente aleatório. Ou melhor :

- O componente aleatório de um MLG fica explicitamente identificado pelas realizações independentes da variável resposta $Y\ (Y_1, \cdots, Y_n)$ e por sua distribuição de probabilidades, que deve pertencer à família exponencial. Esta família inclui várias distribuições importantes, como a Poisson e a Binomial. Por exemplo :

  - Quando cada Y_i assumir apenas dois resultados possíveis (classificados como sucesso ou fracasso) tem-se a distribuição Bernoulli. Uma generalização desta situação seria considerar cada Y_i como sendo o número de sucessos em um número fixo de realizações de um experimento. Assim, é assumida a distribuição Binomial;
  - Quando cada Y_i é o resultado de uma contagem não-negativa, a distribuição assumida é a de Poisson.

- O componente sistemático de um MLG relativo a um vetor $\eta = (\eta_1, \cdots, \eta_n)$, chamado *preditor linear,* é composto por um conjunto de variáveis explicativas X (matriz de planejamento) e pelo vetor de $(p+1)$ parâmetros β do modelo. Assim,

$$\eta = X\beta \ .$$

- O terceiro componente de um MLG é a ligação entre os componentes sistemático e aleatório, isto é, a função de ligação monótona g dada por

$$g(\mu) = X\beta \quad em\ que\ \ \mu = E(Y) .$$

A função g é a maneira como a média da resposta Y está associada ao preditor linear η .

Um caso clássico do modelo linear generalizado é o modelo linear normal

$$Y_i = X_i\beta + e_i \qquad i = 1, \cdots, n,$$

em que os elementos e_i são independentes e todos têm distribuição Normal, com média 0 e variância σ^2. Este é um modelo linear generalizado porque os elementos de Y são variáveis aleatórias independentes Y_i com distribuição $N(\mu_i, \sigma^2)$, que pertence à família exponencial. Neste caso, a função de ligação g é a identidade, isto é, $g(\mu_i) = \mu_i$.

1.2- O Modelo de Regressão Logística

Considere agora modelos lineares generalizados que usam como ligação a função $\log it$, dada por

$$g(\mu_i) = \log it(\mu_i) = \log\left(\frac{\mu_i}{(1-\mu_i)}\right) = X_i'\beta \quad i = 1, \cdots, n, \qquad (1.1)$$

em que:

- cada Y_i tem distribuição Bernoulli, pertencente à família exponencial. As respostas $Y_i's$ são independentes e o modelo é conhecido como o de regressão logística;
- $\log\left(\frac{\mu_i}{1-\mu_i}\right)$ é chamado de função $\log it$ e sua interpretação natural consiste no logaritmo da chance.
- $\mu_i = \Pr[Y_i = 1 \mid X_i] = \frac{e^{X_i'\beta}}{1+e^{X_i'\beta}} = \frac{1}{1+e^{-X_i'\beta}}$;
- $(1-\mu_i) = \Pr[Y_i = 0 \mid X_i] = \frac{1}{1+e^{X_i'\beta}}$;

- o vetor X_i é a i-ésima linha da matriz de planejamento X ;
- β é um vetor de $(p+1)$ parâmetros a serem estimados.

Os modelos de regressão logística são usados para determinar os fatores que estão associados, independentemente, à ocorrência do evento de interesse e para estimar a probabilidade de um indivíduo, caracterizado pelos valores das covariáveis, desenvolver o evento de interesse.

A escolha da função de ligação $\log it$ na construção do modelo para respostas binárias tem várias razões, como, por exemplo: facilidade computacional, comportamento sigmoidal de sua distribuição acumulada e, principalmente, pela facilidade na interpretação dos parâmetros do modelo.

Apresenta-se, a seguir, dois exemplos para ilustrar a utilização deste modelo.

Exemplo 1: Considere um grupo de 102 pacientes submetidos à craniotomia em que a resposta de interesse Y é a ocorrência (ou não) de meningite nos primeiros 30 dias após a cirurgia. Duas covariáveis são consideradas na explicação desta variável resposta: X_1 : Gravidade do caso (0: baixa e 1: alta) e X_2 : duração da cirurgia (horas). Estes dados foram fornecidos pelo Sr. Bráulio Roberto Gonçalves Marinho Couto e é proveniente do Hospital São Francisco de Assis, localizado em Belo Horizonte, MG, Brasil, de julho de 1991 a junho de 1992. O modelo logístico é dado por :

$$\log\left(\frac{\mu_i}{(1-\mu_i)}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \qquad i = 1, 2, \cdots, 102 \ .$$

As estimativas dos parâmetros são os valores que maximizam a verossimilhança do modelo. Supondo que as observações vêm de pacientes diferentes sendo, portanto, independentes, a verossimilhança é apenas o produto da probabilidade de ocorrência de

cada resposta observada. Chamando de L a função de verossimilhança, pode-se escrever

$$L(\beta) = \prod_{i=1}^{n} \Pr\{Y_i = y_i \mid X_i = x_i, \beta\}. \qquad (1.2)$$

Considere os dados do **Exemplo 1** na Tabela 1.1.

Tabela 1.1- Dados referentes a um grupo de 102 pacientes submetidos à craniotomia.

X_1	X_2	N_1	N_0	X_1	X_2	N_1	N_0	X_1	X_2	N_1	N_0
0	2,50	0	1	1	2,00	0	6	0	1,00	0	3
1	1,33	0	1	0	1,25	0	1	0	6,00	0	3
1	6,00	0	2	0	2,17	0	1	1	1,50	1	1
0	4,50	0	1	1	6,50	0	1	1	10,00	1	0
0	1,50	0	3	1	1,00	0	3				
0	1,33	0	4	1	4,00	0	4				
0	5,00	0	3	1	3,00	0	8				
1	0,75	0	1	0	4,00	0	8				
0	2,00	0	8	0	4,75	0	1				
0	3,50	0	3	0	3,00	0	13				
1	3,25	0	1	1	8,00	0	1				
0	1,83	0	4	0	5,50	0	1				
1	7,00	0	1	0	2,67	0	1				
0	1,67	0	1	0	2,25	0	1				
0	8,00	0	1	0	7,00	0	2				
1	3,50	0	1	0	3,67	0	1				
0	3,17	0	1	0	2,33	0	1				
1	5,50	0	1	0	6,50	0	1				

em que:

X_1: gravidade do caso (0: baixa e 1: alta)

X_2: duração da cirurgia (em horas)

N_1 : número de ocorrências de meningite

N_0: número de casos em que não ocorreu a meningite

em que cada uma das K possíveis combinações dos valores das covariáveis apresentam N_{1i} sucessos para N_i observações, a expressão (1.2) pode ser escrita da seguinte forma

$$L(\beta) = \prod_{i=1}^{K} \left(\frac{e^{X_i'\beta}}{1 + e^{X_i'\beta}} \right)^{N_{1i}} \times \left(\frac{1}{1 + e^{X_i'\beta}} \right)^{N_i - N_{1i}}. \tag{1.3}$$

A expressão do logaritmo da função verossimilhança para β, isto é, $\ln L(\beta) = l(\beta)$, para $\mu_i = \frac{e^{X_i'\beta}}{1 + e^{X_i'\beta}}$, $(1 - \mu_i) = \frac{1}{1 + e^{X_i'\beta}}$ fica

$$l(\beta) = \sum_{i=1}^{k} [N_{1i} \log \mu_i + (N_i - N_{1i}) \log (1 - \mu_i)]. \tag{1.4}$$

Por definição, os estimadores de máxima verossimilhança são os valores de β que maximizam a expressão (1.4) ou, equivalentemente, os valores de β que são a solução para o sistema de $(p + 1)$ equações obtido por fazer o vetor escore igual a zero. Isto é,

$$\frac{\partial l(\beta)}{\partial \beta} = U_j(\beta) = \sum_{i=1}^{k} x_{ij} (N_{1i} - N_i \mu_i) = 0; \qquad j = 1, \cdots, p + 1,$$

em que x_{ij} é o $j - ésimo$ elemento de X_i. Em geral, o método iterativo de Newton-Raphson é usado para resolver esse sistema de equações.

Exemplo 2: Considere um grupo de 92 indivíduos em que a resposta de interesse Y é a Taxa de Pulsação (0: baixa ; 1: alta). Duas covariáveis são consideradas na explicação dessa variável resposta: X_1: peso; X_2: fumante (0:não ; 1:sim) e N_0: número respostas 0; N_1: número respostas 1. Os dados estão apresentados na Tabela 1.2 .

Tabela 1.2- Dados referentes a um grupo de 92 indivíduos em que a resposta de interesse é a Taxa de Pulsação.

X_1	X_2	N_0	N_1	X_1	X_2	N_0	N_1	X_1	X_2	N_0	N_1
215	0	1	0	153	1	0	1	125	0	4	0
195	0	1	0	150	1	2	2	123	0	1	0
190	1	2	0	150	0	4	2	122	0	1	0
190	0	2	0	148	0	1	0	121	1	0	1
185	1	1	0	145	1	1	0	120	0	3	0
180	1	1	1	145	0	3	1	118	0	1	1
180	0	1	0	142	0	1	0	116	0	0	2
175	1	1	0	140	1	0	1	115	0	2	0
175	0	1	0	140	0	2	1	112	1	1	0
170	1	2	0	138	0	1	1	110	0	2	0
170	0	2	0	136	0	0	1	108	1	1	0
165	0	1	0	135	1	0	1	108	0	1	0
164	1	1	0	135	0	2	0	102	0	1	0
160	1	1	1	133	0	1	0	95	0	0	1
160	0	2	0	131	1	0	1				
157	0	1	0	130	1	2	0				
155	1	2	0	130	0	2	1				
155	0	7	1	125	1	0	1				

O modelo logístico é dado por :

$$\log\left(\frac{\mu_i}{(1-\mu_i)}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \qquad i = 1, 2, \cdots, 92 \ .$$

O modelo ajustado é dado por :

$$\log\left(\frac{\hat{\mu}_i}{(1-\hat{\mu}_i)}\right) = 1{,}987 \ - \ 0{,}0250 \ \ \textit{Peso} \ + 1{,}1930 \ \ \textit{Fumante}$$

Existem situações em que a função $l(\beta)$ não tem valor máximo para valores finitos dos coeficientes de regressão. Isto significa dizer que não existem estimativas de máxima verossimilhança. O **Exemplo 1** é um exemplo de uma situação sem estimativas de máxima verossimilhança. Nesse caso, a função de verossimilhança é monótona.

Albert e Anderson (1984) mostraram que os conjuntos de dados logísticos podem ser classificados em três categorias mutuamente exclusivas e exaustivas: Separação Completa, Separação Quase-Completa e *Overlap*, e que os estimadores de máxima verossimilhança não existem para as duas primeiras categorias. O **Exemplo 2** ilustra a situação em que os dados encontram-se na categoria *Overlap* mostrando, portanto, as estimativas de máxima verossimilhança. Albert e Anderson (1984) e Santner e Duffy (1986) forneceram técnicas complicadas para classificar um conjunto de dados logístico em uma dessas três categorias. Tais técnicas tornam-se simples quando se têm duas covariáveis sendo, uma delas, binária. Essa é a situação do **Exemplo 1**, já mencionado anteriormente. O **Exemplo 3**, apresentado a seguir, também ilustra a situação de inexistência de estimativas de máxima verossimilhança.

Exemplo 3: Considere um grupo de 104 pacientes onde a resposta de interesse Y é a presença (ou não) de hemorragia peri-intraventricular no parto. Esse banco de dados foi fornecido pelo Professor Marcos Borato Viana (Departamento de Pediatria, UFMG). Várias covariáveis foram consideradas para a possível explicação dessa variável resposta:

X_1 : uso de corticóide (0: Não; 1: Sim)

X_2 : tipo de parto (0: cesariano; 1: vaginal)

X_3 : peso ao nascimento (0: ≥ 1000 gramas; 1: <1000 gramas)

X_4 : IGC (Idade Gestacional Calculada : em semanas)

X_5 : presença de infecção (0: não; 1: provável; 2: confirmada)

X_6 : número de dias em que a criança permaneceu sob ventilação mecânica assistida (VM)

X_7 : ventilação por pressão positiva (VPP) – ambu – ressuscitador auto inflável (0: Não; 1: Sim)

Os dados estão apresentados no ANEXO A1.

Esta classificação de um conjunto de dados logístico, segundo Albert e Anderson (1984), é descrita com cuidado no próximo capítulo.

1.3- Objetivos e Organização desta Dissertação

O objetivo deste trabalho é propor soluções alternativas ao ajuste usual do modelo de regressão logística, quando não existem estimadores de máxima verossimilhança. O Capítulo 2 apresenta a classificação de dados logísticos segundo Albert e Anderson (1984) definindo as três categorias, mutuamente exclusivas e exaustivas, que indicam a existência (ou não) de estimativas de máxima verossimilhança. Mostra-se ainda que os bancos de dados deste trabalho estão na categoria de Separação Quase-Completa e, portanto, não apresentam estimadores de máxima verossimilhança. As propostas alternativas de solução aparecem no Capítulo 3 e, no Capítulo 4, apresenta-se os resultados de simulações para as três propostas de solução. Este estudo é finalizado com os resultados e conclusões no Capítulo 5.

Capítulo 2

Existência (ou não) de Estimadores de Máxima Verossimilhança: Definição e Exemplos

O problema de maximização do logaritmo da função de verossimilhança (1.4) será examinado considerando as configurações possíveis de n pontos amostrais no espaço de observação $\Re^p$. Como já dito anteriormente, as configurações possíveis caem em três categorias mutuamente exclusivas e exaustivas: separação completa, separação quase-completa e *overlap*. Em cada caso, concentra-se na solução finita de máxima verossimilhança, e no valor do máximo da função de verossimilhança (1.4). Considera-se, ainda, que a não existência da estimativa de máxima verossimilhança significa ausência de um máximo finito. Apresenta-se a seguir a definição destas categorias segundo Albert e Anderson (1984) e ilustra-se para uma situação em que $p = 2$.

2.1- Definição das Categorias de Separação

A definição destas categorias segundo Albert e Anderson (1984) segue-se.

2.1.1- Separação Completa

Pode-se dizer que existe separação completa nos pontos amostrais se existe um vetor $\beta \in \Re^{p+1}$ que aloca corretamente todas as observações aos seus grupos $(j = 0, 1)$, tal que, para todo $i \in E_j$

$$X_i'\beta > 0 \qquad i \in E_0 \quad \text{e}$$

$$X_i'\beta < 0 \qquad i \in E_1,$$

em que E_j é o conjunto de linhas identificadas da matriz X com valor de $Y = j \quad (j = 0, 1)$. A Figura 2.1(a) ilustra esta categoria de separação para $\Re^2$.

2.1.2- Separação Quase-Completa

Quando o conjunto de dados X não é completamente separado, é preciso fornecer outro conceito de separação. Existe separação quase-completa dos pontos amostrais se existe um vetor $\beta \in \Re^{p+1}$ tal que, para todo $i \in E_j$ e para $j = 0, 1$

$$X_i'\beta \geq 0 \qquad i \in E_0 \quad \text{e}$$
$$X_i'\beta \leq 0 \qquad i \in E_1 \tag{2.1}$$

com igualdade para, pelo menos, um valor de i. A Figura 2.1(b) ilustra esta categoria de separação para $\Re^2$. Os dados dos **Exemplos 1 e 3**, apresentados no Capítulo 1, pertencem a essa categoria, como será mostrado.

2.1.3- *Overlap*

Se não existem, nos pontos amostrais, nenhuma das duas separações (completa ou quase-completa), esses estão, necessariamente, na categoria de separação *overlap*. Existe separação *overlap* dos pontos amostrais, se existe um vetor $\beta \in \Re^{p+1}$ tal que, para todo $i \in E_j$ e para $j = 0, 1$

$$X_i'\beta < 0 \qquad i \in E_0 \quad \text{e}$$
$$X_i'\beta > 0 \qquad i \in E_1.$$

A Figura 2.1(c) ilustra esta categoria de separação para $\Re^2$.

Albert e Anderson (1984) também mostraram que os estimadores de máxima verossimilhança não existem para as categorias de Separação Completa e Quase-Completa. O **Exemplo 2**, mencionado no Capítulo 1, ilustra dados classificados na categoria de Separação *Overlap*, apresentando as estimativas de máxima verossimilhança.

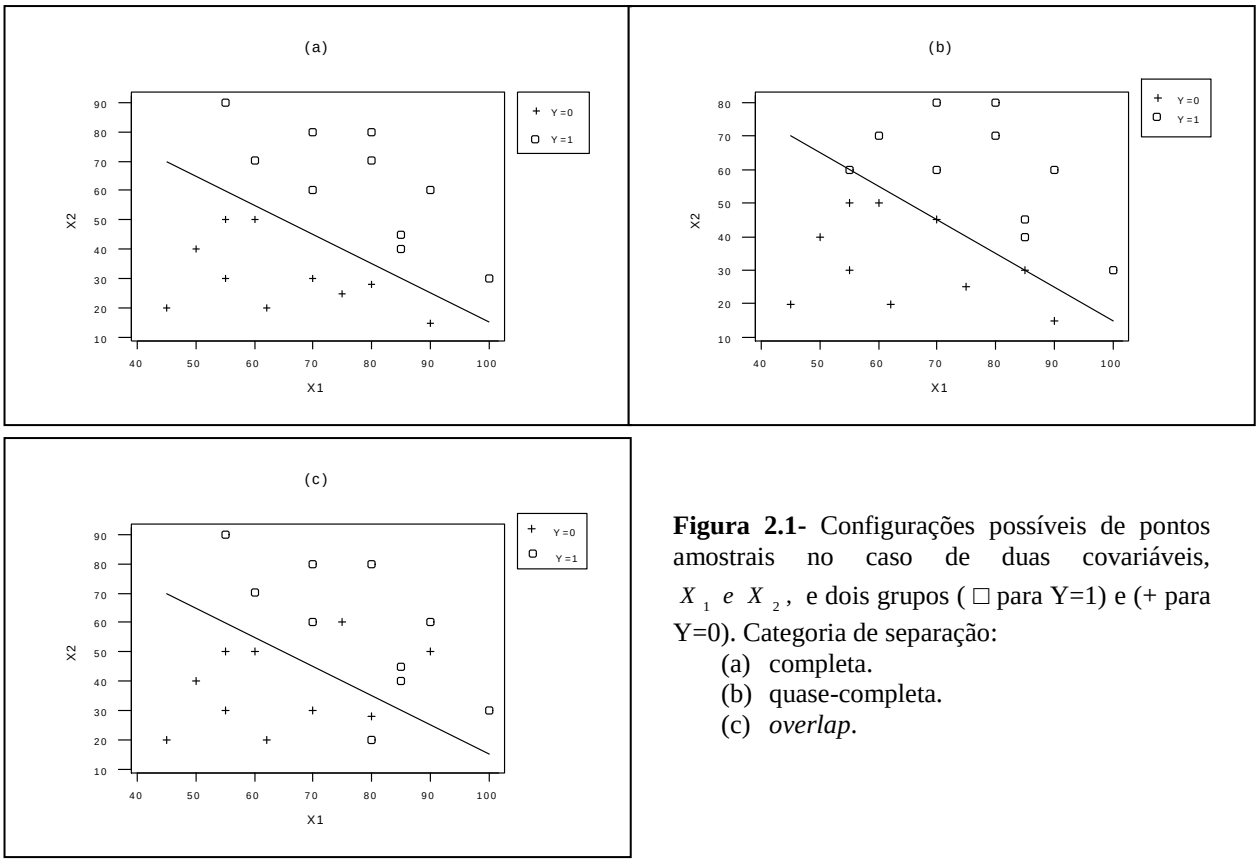

Figura 2.1- Configurações possíveis de pontos amostrais no caso de duas covariáveis, X_1 *e* X_2, e dois grupos (□ para Y=1) e (+ para Y=0). Categoria de separação:

(a) completa.
(b) quase-completa.
(c) *overlap*.

2.2- Situações Reais : Separação Quase-Completa

A análise de respostas binárias utilizando o modelo de regressão logística, mostra que não são raras as situações reais na categoria de separação quase-completa. Isso gera um grave problema, pois a análise estatística fica comprometida pela inexistência de estimadores de máxima verossimilhança. Essa situação está freqüentemente associada à existência de uma variável explicativa categórica. A situação de quase-separação é identificada com uma casela nula cruzando essa covariável categórica com a resposta

binária. O **Exemplo 1,** apresentado no Capítulo 1, ilustra essa condição. Nesta seção vamos provar esse fato.

2.2.1- Caso 1 – Uma Única Covariável Binária

Considere a representação da Tabela 2.1 resultante do cruzamento de Y com X_1 binária.

Tabela 2.1- Cruzamento da covariável binária X_1 com a resposta Y.

Y	X_1		***Total***
	0	***1***	
0	a	b	a+b
1	c	d	c+d
Total	a+c	b+d	a+b+c+d=N

Teorema 1: Se qualquer uma, mas somente uma, das caselas da Tabela 2.1 for nula, o conjunto de dados está na categoria de separação quase-completa.

Demonstração:

- Se $a = 0$, $\beta' = (-1 \;\; 1)$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1), pois:
  - $X_i'\beta = \beta_0 + \beta_1 X_{i1} \Rightarrow$ fazendo $\beta' = (-1 \;\; 1) \Rightarrow \quad X_i'\beta = -1 + X_{i1}$
  - para $Y = 0 \quad (i \in E_0)$: $X_{i1} = 1 \quad \Rightarrow \quad X_i'\beta = 0 \quad \Rightarrow \quad X_i'\beta \geq 0$
  - para $Y = 1 \quad (i \in E_1)$:
    - $X_{i1} = 0 \quad \Rightarrow \quad X_i'\beta = -1$

      $\Rightarrow \quad X_i'\beta \leq 0$
    - $X_{i1} = 1 \quad \Rightarrow \quad X_i'\beta = 0$

- Se $b = 0$, $\beta' = (0 \quad -1)$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1), pois:
  - $X_i'\beta = \beta_0 + \beta_1 X_{i1} \Rightarrow$ fazendo $\beta' = (0 \quad -1) \Rightarrow \quad X_i'\beta = -X_{i1}$
  - para $Y = 0 \quad (i \in E_0)$: $X_{i1} = 0 \quad \Rightarrow \quad X_i'\beta = 0 \quad \Rightarrow \quad X_i'\beta \geq 0$
  - para $Y = 1 \quad (i \in E_1)$:
    - $X_{i1} = 0 \quad \Rightarrow \quad X_i'\beta = 0$

      $\Rightarrow \quad X_i'\beta \leq 0$
    - $X_{i1} = 1 \quad \Rightarrow \quad X_i'\beta = -1$

- Se $c = 0$, $\beta' = (1 \quad -1)$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1), pois:
  - $X_i'\beta = \beta_0 + \beta_1 X_{i1} \Rightarrow$ fazendo $\beta' = (1 \quad -1) \Rightarrow \quad X_i'\beta = 1 - X_{i1}$
  - para $Y = 0 \quad (i \in E_0)$:
    - $X_{i1} = 0 \quad \Rightarrow \quad X_i'\beta = 1$

      $\Rightarrow \quad X_i'\beta \geq 0$
    - $X_{i1} = 1 \quad \Rightarrow \quad X_i'\beta = 0$
  - para $Y = 1 \quad (i \in E_1)$: $X_{i1} = 1 \quad \Rightarrow \quad X_i'\beta = 0 \quad \Rightarrow \quad X_i'\beta \leq 0$

- Se $d = 0$, $\beta' = (0 \quad 1)$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1), pois:
  - $X_i'\beta = \beta_0 + \beta_1 X_{i1} \Rightarrow$ fazendo $\beta' = (0 \quad 1) \Rightarrow \quad X_i'\beta = X_{i1}$
  - para $Y = 0 \quad (i \in E_0)$:
    - $X_{i1} = 0 \quad \Rightarrow \quad X_i'\beta = 0$

      $\Rightarrow \quad X_i'\beta \geq 0$
    - $X_{i1} = 1 \quad \Rightarrow \quad X_i'\beta = 1$
  - para $Y = 1 \quad (i \in E_1)$: $X_{i1} = 0 \quad \Rightarrow \quad X_i'\beta = 0 \quad \Rightarrow \quad X_i'\beta \leq 0$

Desta forma, o Teorema 1 fica provado.

2.2.2- <u>Caso 2</u> – p Covariáveis, incluindo uma Binária

Suponha que seja considerada uma seqüência de covariáveis quaisquer $X_1, X_2, \cdots, X_p$, e que, sem perda de generalidade, X_1 é binária. Considere a representação da Tabela 2.1 resultante do cruzamento de Y com X_1: binária. O vetor $\beta \in \Re^{p+1}$ que satisfaz às condições para que estes dados estejam na categoria de separação quase-completa segue diretamente daquele obtido no Teorema 1. Isto é:

- Se $a = 0$, $\beta' = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1).
- Se $b = 0$, $\beta' = \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1).
- Se $c = 0$, $\beta' = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1).
- Se $d = 0$, $\beta' = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1).

Isto mostra que, independente do número de covariáveis que estejam sendo consideradas, com X_1 binária e na situação da Tabela 2.1, existe um vetor β (*com* $p+1$ *parâmetros*) que satisfaz as condições para que os dados estejam na categoria de separação quase-completa se existir uma, mas somente uma, casela nula.

Exemplo 1 revisado: Considerando ainda os dados do **Exemplo 1**, a Tabela 2.2 mostra o cruzamento da variável resposta, *ocorrência de meningite durante os primeiros 30 dias após a cirurgia* e a covariável *gravidade do caso* (X_1). Como pode ser observado, existe uma casela nula nesta tabela e, de acordo com o Teorema 1, e Albert & Anderson (1984), não existe estimador de máxima verossimilhança para estes dados. A Figura 2.2 também mostra graficamente que estes dados estão na categoria de separação quase-completa.

Tabela 2.2- Cruzamento da covariável binária X_1 com a resposta N_1.

	X₁: gravidade do caso		
N₁: ocorrência de meningite	***0: baixa***	***1: alta***	***Total***
0: não	68	32	100
1: sim	0	2	2
Total	68	34	102

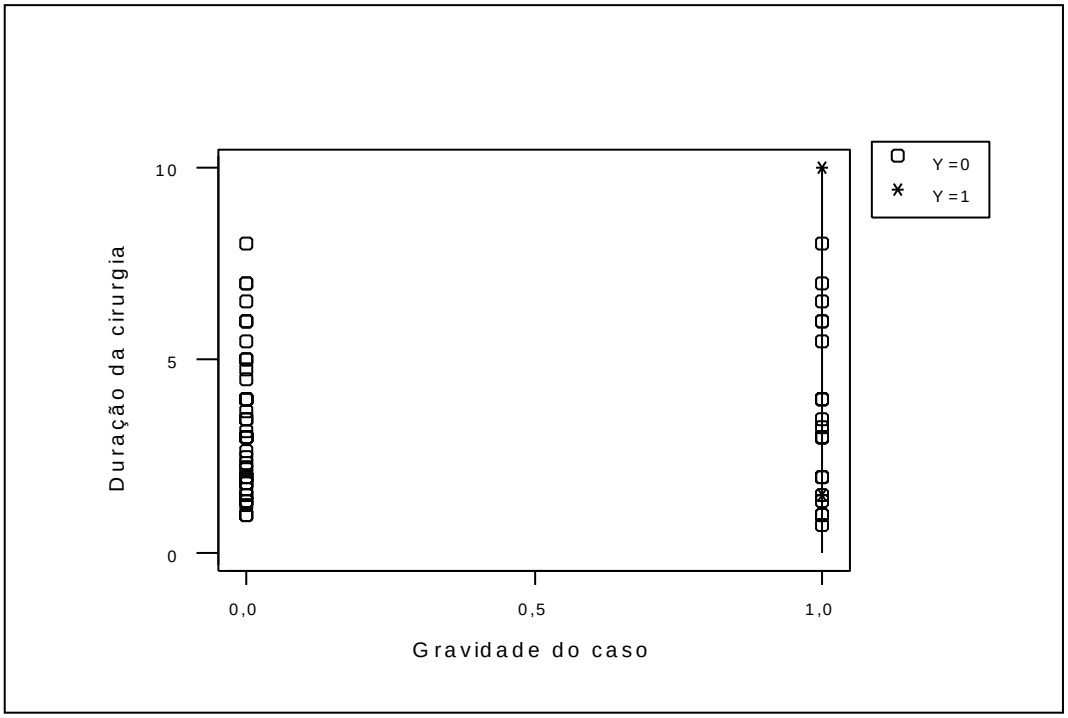

Figura 2.2- Pacientes submetidos à craniotomia.

2.2.3- <u>Caso 3</u> – Uma Única Covariável Categórica, com 3 Categorias

Este caso refere-se à situação em que a covariável X_1 é categórica, com mais de duas categorias ou classificações. Suponha que a variável categórica X_1 tenha três categorias, isto é, $X_1 : 0, 1, 2$. O cruzamento resultante desta covariável com Y está apresentado na Tabela 2.3.

Tabela 2.3- Cruzamento da covariável categórica X_1 com a resposta Y.

	X1			
Y	***0***	***1***	***2***	***Total***
0	a	b	c	a+b+c
1	d	e	f	d+e+f
Total	A+d	b+e	c+f	a+b+c+d+e+f=N

Essa covariável, em situações reais, é representada por variáveis indicadoras. Isto é, usa-se duas variáveis indicadoras para representar X_1 que tem três categorias. Uma forma usual de representação está apresentada na Tabela 2.4.

Tabela 2.4- Variáveis indicadoras X_{11} e X_{12} representando a covariável X_1 com três categorias.

		X11	***X12***
	0	0	0
X1	***1***	1	0
	2	0	1

Será mostrado que se uma, e somente uma, das caselas da Tabela 2.3 for nula, o conjunto de dados estará na categoria de separação quase-completa.

Teorema 2: Se qualquer uma, mas somente uma, das caselas da Tabela 2.3 for nula, o conjunto de dados está na categoria de separação quase-completa.

Demonstração:

A Tabela 2.5 mostra o cruzamento das duas variáveis indicadoras de X_1 com a resposta.

Tabela 2.5- Cruzamento das variáveis indicadoras com Y .

		X_{11}		X_{12}	
		0	***1***	***0***	***1***
Y	***0***	a+c	b	a+b	c
	1	d+f	e	d+e	f

- Se $a = 0$, tem-se a seguinte configuração:

Y	X_{11}	X_{12}	
0	1	0	
0	1	0	
⋮	⋮	⋮	
0	1	0	
0	0	1	E_0
0	0	1	
⋮	⋮	⋮	
0	0	1	

1	0	0	
1	0	0	
⋮	⋮	⋮	
1	0	0	
1	1	0	
1	1	0	E_1
⋮	⋮	⋮	
1	1	0	
1	0	1	
1	0	1	
⋮	⋮	⋮	
1	0	1	

- $\beta' = (-1 \;\; 1 \;\; 1)$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1), pois:
- $X_i'\beta = \beta_0 + \beta_{11} X_{i11} + \beta_{12} X_{i12} \Rightarrow$ fazendo $\beta' = (-1 \;\; 1 \;\; 1) \Rightarrow$

  $X_i'\beta = -1 + X_{i11} + X_{i12}$
- para $Y = 0 \quad (i \in E_0)$: $X_i'\beta = 0$ para todo $i \in E_0$ e, portanto, $X_i'\beta = 0$
- para $Y = 1 \quad (i \in E_1)$: $X_i'\beta \leq 0$

➢ Se $b = 0$ ou $e = 0$ ou $c = 0$ ou $f = 0$ segue imediatamente, pelo Teorema 1, que o conjunto de dados está na categoria de separação quase-completa.

➢ Se $d = 0$, tem-se a seguinte configuração:

Y	X_{11}	X_{12}	
0	0	0	
0	0	0	
⋮	⋮	⋮	
0	0	0	
0	1	0	
0	1	0	E_0
⋮	⋮	⋮	
0	1	0	
0	0	1	
0	0	1	
⋮	⋮	⋮	
0	0	1	
1	1	0	
1	1	0	
⋮	⋮	⋮	
1	1	0	
1	0	1	E_1
1	0	1	
⋮	⋮	⋮	
1	0	1	

- $\beta' = (1 \ -1 \ -1)$ é um vetor que satisfaz às condições para que os dados estejam na categoria de separação quase-completa (2.1), pois:
- $X_i'\beta = \beta_0 + \beta_{11} X_{i11} + \beta_{12} X_{i12} \Rightarrow$ fazendo $\beta' = (1 \ -1 \ -1) \Rightarrow$

  $X_i'\beta = 1 - X_{i11} - X_{i12}$
- para $Y = 0 \quad (i \in E_0)$: $X_i'\beta \geq 0$
- para $Y = 1 \quad (i \in E_1)$: $X_i'\beta \leq 0$

Segue que o Teorema 2 está provado.

Duas observações devem ser registradas com relação à extensão deste resultado :

1- Existem outras formas de criar variáveis indicadoras para representar a covariável X_1. No entanto, o Teorema 2 continua valendo com prova similar.
2- A prova do Teorema 2 para covariáveis categóricas com mais de três classificações segue de forma similar.

2.2.4- <u>Caso 4</u> – p Covariáveis, incluindo uma Categórica, com 3 Categorias

Suponha que seja considerado um conjunto de covariáveis quaisquer $X_1, X_2, \cdots, X_p$ e que, sem perda de generalidade, X_1 é categórica. Suponha ainda que a variável categórica tenha 3 classificações, isto é, $X_1 : 0, 1, 2$, e que suas indicadoras sejam X_{11} (indicadora da categoria 1) e X_{12} (indicadora da categoria 2), como apresentado na Tabela 2.4.

O vetor $\beta \in \Re^{p+1}$ que satisfaz às condições para que esses dados estejam na categoria de separação quase-completa segue diretamente daquele obtido no Teorema 2. Isto é:

- Se $a = 0$, tem-se que $\beta' = \begin{pmatrix} -1 & 1 & 1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$
- Se $b = 0$, tem-se que $\beta' = \begin{pmatrix} 0 & -1 & 0 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$
- Se $c = 0$, tem-se que $\beta' = \begin{pmatrix} 0 & 0 & -1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$

- Se $d = 0$, tem-se que $\beta' = \begin{pmatrix} 1 & -1 & -1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$

- Se $e = 0$, tem-se que $\beta' = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$

- Se $f = 0$, tem-se que $\beta' = \begin{pmatrix} 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}_{p+1}$

A seguinte situação real ilustra essa seção.

Exemplo 3 revisado: Considere os dados do **Exemplo 3.** O cruzamento da variável resposta de interesse *Y: presença (ou não) de hemorragia peri-intraventricular no parto* com a covariável X_5: *presença de infecção (0: não; 1: provável; 2: confirmada)* está apresentado na Tabela 2.6. Nota-se a presença de uma casela nula no cruzamento da resposta Y com essa variável categórica.

Tabela 2.6- Cruzamento da resposta Y com a variável categórica X_5

Y	X_5 **: Infecção**			
Hemorragia	***0: não***	***1: provável***	***2: confirmada***	***Total***
0: grau 1	11	22	58	91
1: graus 2 a 4	0	5	8	13
Total	11	27	66	104

Isso significa que os estimadores de máxima verossimilhança não existem para esse conjunto de dados.

Capítulo 3

Propostas de Solução

Como foi visto no Capítulo 2, podem existir situações em que as estimativas de máxima verossimilhança (ML) não existem (Albert & Anderson,1984). A literatura apresenta algumas propostas de solução para o caso de Separação Completa e Quase-Completa (Colosimo, Franco e Couto, 1995; King e Ryan, 2002). Neste capítulo serão apresentadas duas propostas da literatura: regressão logística exata e adicionar uma pequena constante aos dados e, ainda, propor uma nova solução, que é mais simples e fácil de ser implementada.

3.1- Regressão Logística Exata

A idéia do método de Regressão Logística Exata (RLE), proposto por Cox (1970), era basear a inferência em distribuições permutacionais exatas das estatísticas suficientes correspondentes aos parâmetros de interesse da regressão, condicionada nas estatísticas suficientes dos parâmetros restantes (Inferência Condicional Exata).

A regressão logística exata foi vista, durante muito tempo, como uma abordagem computacionalmente inviável, o que resultou num método raramente usado. Foi Tritchler (1984) quem deu o primeiro passo que permitiu que a abordagem pudesse ser usada através da produção de um método numérico. Vale mencionar também a contribuição neste sentido de Hirji, Mehta e Patel (1987, 1988), Hirji (1992) e Mehta e Patel e Senchaudhuri (1998, 2000).

O *LogXact* é o único *software* disponível comercialmente que suporta a inferência condicional exata para a regressão logística. Isso porque, a cada nova versão, vem

aumentando enormemente o poder dos algoritmos numéricos. Assim, esse *software* será usado para obter os resultados.

Esta seção descreve a teoria para a Inferência Condicional Exata e utiliza um exemplo, apresentado por Souza (2.000), que mostra claramente a utilização dessa teoria. Os outros dois exemplos seguintes são os já mencionados neste texto.

3.1.1- O Modelo, a Verossimilhança e as Estatísticas Suficientes

Sejam $Y_1, Y_2, \cdots, Y_n$ variáveis aleatórias binárias independentes em que, para cada Y_i, existe um vetor $(p \times 1)$ de variáveis explicativas $X_i = \left(X_{i1}, X_{i2}, \cdots, X_{ip}\right)'$. Seja μ_i a probabilidade de que $Y_i = 1$. A regressão logística, conforme (1.1), modela a dependência de μ_i em X_i através da relação

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = X_i'\beta, \quad i = 1, \cdots, n \tag{3.1}$$

em que $\beta \equiv \left(\beta_0, \beta_1, \beta_2, \cdots, \beta_p\right)'$ apresenta $(p+1)$ parâmetros desconhecidos. A função de verossimilhança, conforme (1.3), ou a probabilidade de um conjunto de valores observados $y_1, y_2, \cdots, y_n$ é

$$\Pr\left(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n\right) = \frac{\exp\left[\sum_{i=1}^{n} y_i\left(X_i'\beta\right)\right]}{\prod_{i=1}^{n}\left[1 + \exp\left(X_i'\beta\right)\right]}$$

ou, equivalentemente,

$$\Pr\left(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n\right) = \frac{\exp\left[\sum_{i=1}^{n} y_i\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip}\right)\right]}{\prod_{i=1}^{n}\left[1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip}\right)\right]} \quad (3.2)$$

A forma usual de fazer inferências sobre β_0 *e* β^*, em que $\beta^* = (\beta - \beta_0 1)$ consiste em maximizar (3.2) com respeito a esses coeficientes de regressão.

Suponha que tem-se interesse em fazer inferências sobre β, e considera-se β_0 como um parâmetro de perturbação. Assim, ao invés de estimar β_0 através da função de verossimilhança não-condicional (3.2), pode-se eliminá-lo condicionando nos valores observados de sua estatística suficiente

$$m = \sum_{i=1}^{n} y_i .$$

A função de verossimilhança condicional resultante está, assim, livre do parâmetro β_0, conforme segue

$$\Pr\left(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n / m\right) = \frac{\exp\left[\sum_{i=1}^{n} y_i X_i' \beta^*\right]}{\sum_R \exp\left[\sum_{i=1}^{n} y_i X_i' \beta^*\right]}, \quad (3.3)$$

em que o somatório externo no denominador de (3.3) é sobre o conjunto

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

Pode-se agora abordar a inferência sobre β^* de duas formas : assintótica e exata. Uma abordagem assintótica consiste em maximizar a função de verossimilhança condicional (3.3). A inferência exata sobre β^* é baseada na distribuição permutacional de suas estatísticas suficientes. Pode-se observar na forma de (3.3) que o vetor de estatísticas suficientes para β^* é

$$t = \sum_{i=1}^{n} y_i X_i ,$$

e sua distribuição é

$$\Pr\left(T_1 = t_1, T_2 = t_2, \cdots, T_p = t_p\right) = \frac{c(t)\, e^{t'\beta^*}}{\sum_u c(u)\, e^{u'\beta^*}},$$

em que $c(t) = \#\, S(t),$ com

$$S(t) = \left\{ (y_1, y_2, \cdots, y_n) : \sum_{i=1}^{n} y_i = m, \sum_{i=1}^{n} y_i x_{ij} = t_j,\ j = 1, \cdots, p \right\},$$

em que $\# S$ representa o número de elementos distintos em S, e o somatório no denominador é sobre todo u para o qual $c(u) \geq 1$. Em outras palavras, $c(t)$ é a contagem do número de seqüências binárias da forma $(y_1, y_2, \cdots, y_n)$ tal que $\sum_{i=1}^{n} y_i = m$ e $\sum_{i=1}^{n} y_i x_{ij} = t_j,\ j = 1, \cdots, p$. A inferência exata sobre β^* requer o cálculo dos coeficientes tal como $c(t)$ no qual algumas das estatísticas suficientes são fixadas nos seus valores observados e outras precisam variar sobre toda a sua extensão.

3.1.2- Inferência Condicional Exata para β_p

Suponha que, sem perda de generalidade, deseja-se fazer inferências sobre um único parâmetro β_p. Pelo princípio da suficiência, a distribuição condicional de T_p dado $t_1, t_2, \cdots, t_{p-1}$ depende apenas de β_p. Seja $f(t_p \mid \beta_p)$ representando a probabilidade condicional $\Pr(T_p = t_p \mid T_1 = t_1, \cdots, T_{p-1} = t_{p-1})$. Então

$$f(t_p \mid \beta_p) = \frac{c(t_1, t_2, \cdots, t_p)\, e^{\beta_p t_p}}{\sum_u c(t_1, t_2, \cdots, t_{p-1}, u)\, e^{\beta_p u}},$$

em que o somatório no denominador está sobre todos os valores de u para o qual $c(t_1, t_2, \cdots, t_{p-1}, u) \geq 1$. Como esta probabilidade não envolve os parâmetros de perturbação $(\beta_1, \beta_2, \cdots, \beta_{p-1})$, pode-se usá-la para fazer inferências sobre β_p.

3.1.3- Ilustração: Estudo de Dose-Resposta

A análise de dados para estudos do tipo dose-resposta, isto é, estudos em que as unidades experimentais (em geral, cobaias) são tratadas com dosagens diferentes de uma determinada droga, se baseia em um modelo que relaciona a probabilidade de uma resposta binária à dosagem usada da droga.

Considere K diferentes doses $x_1, x_2, \cdots, x_k$ de uma determinada droga e que, sem perda de generalidade, $x_1 < x_2 < \cdots < x_k$. Para $i = 1, 2, \cdots, k$, sejam :

n_i : número total de animais tratados com a dose x_i;

y_i : número de animais que responderam positivamente;

μ_i : probabilidade de o animal responder positivamente à dose x_i.

O modelo de regressão logística associado a esse caso é

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i\beta,$$

em que $\mu_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}}$. Assim, a distribuição conjunta de Y_i é dada por

$$\Pr\left[Y_i = y_i, i = 1,\cdots,k\right] = \frac{\left[\prod_{i=1}^{k}\binom{n_i}{y_i}\exp\left(\beta_0\sum_i y_i + \beta^*\sum_i y_i x_i\right)\right]}{\prod_{i=1}^{k}\left\{1+\exp\left(\beta_0+\beta^* x_i\right)\right\}^{n_i}}, \qquad (3.4)$$

em que $\beta^* = (\beta_1,\cdots,\beta_p)$.

As estatísticas suficientes para os parâmetros β_0 e β^* são, respectivamente,

$$m = \sum_i y_i \quad e \quad t = \sum_i y_i x_i .$$

Assim, substituindo m e t, a distribuição conjunta de M e T é dada por

$$\Pr\left[M = m, T = t\right] = \frac{c(m,t)\exp\left(\beta_0 m + \beta^* t\right)}{\sum_{(u,v)\in\Omega} c(u,v)\exp\left(\beta_0 u + \beta^* v\right)}$$

em que

$$\Omega = \left\{(u,v): \sum_i y_i = u, \sum_i y_i x_i = v, 0 \le y_i \le n_i, i = 1,\cdots,k\right\} \text{ e } c(m,t) = \sum\prod_{i=1}^{k}\binom{n_i}{y_i},$$

quando a soma é realizada sobre toda a seqüência $\{0 \le y_i \le n_i, i = 1, \cdots, k\}$ para a qual $m = \sum_i y_i$ e $t = \sum_i y_i x_i$.

No modelo logístico (3.1), testar a inexistência de efeito da dose é o mesmo que testar a hipótese nula $\beta^* = 0$. Para eliminar o efeito do parâmetro de perturbação β_0, usa-se a distribuição condicional de T dado M = m, dada por

$$\Pr[T = t \mid M = m] = \frac{c(m,t)\exp(\beta^* t)}{\sum_{v \in \Omega(m)} c(m,v)\exp(\beta^* v)} \quad em \quad que \quad \Omega(m) = \{v : (m,v) \in \Omega\}.$$

Sob H_0,

$$\Pr[T = t \mid M = m;\ \beta^* = 0] = \frac{c(m,t)}{\sum_{v \in \Omega(m)} c(m,v)}$$

A título de ilustração, considere um estudo simples do tipo dose-resposta realizado com somente 9 cobaias.

Tabela 3.1- Dados de um estudo do tipo dose-resposta

Resposta	Dose (x_i)			
(y_i)	***0***	***1***	***2***	***Total***
Presente (1)	0	0	2	2
Ausente (0)	3	3	1	7
Total	3	3	3	9

O modelo de regressão logística é $\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_i\beta$, em que $x_i = 0,1,2$ são as doses com $i = 1,2,3$.

Os valores das estatísticas suficientes para os parâmetros β_0 e β^* são, respectivamente

$$m = \sum_i y_i = \sum_{i=1}^{3} y_i = y_1 + y_2 + y_3 = 0 + 0 + 2 = 2 \Rightarrow \quad m = 2 \qquad e$$

$$t = \sum_i y_i x_i = \sum_{i=1}^{3} y_i x_i = y_1 x_1 + y_2 x_2 + y_3 x_3 = (0.0) + (0.1) + (2.2) = 4 \Rightarrow \quad t = 4$$

As possíveis configurações para os valores das estatísticas suficientes para os parâmetros β_0 e β^* estão representadas na Tabela 3.2.

Agora, é preciso gerar $\sum \prod_{i=1}^{3} \binom{3}{y_i}$, pois existem valores repetidos de (M,T). Os resultados encontram-se na Tabela 3.3 onde estão realçados os valores para $C(M,T) = C(2,T)$, uma vez que o valor obtido para a estatística suficiente para β_0 foi $m = 2$. Desta forma,

$$\Pr\left[T = t \mid M = 2;\ \beta^*\right]$$

não depende de β_0.

A probabilidade condicional de T dado M, sob H_0, é dada por:

$$\Pr\left[T = t \mid M = m;\ \beta = 0\right] = \frac{c(m,t)}{\sum_{v \in \Omega(m)} c(m,v)} \quad .$$

Para $m = 2$, tem-se:

$$\Pr\left[T = t \mid M = 2;\ \beta = 0\right] = \frac{c(2,t)}{\sum_{v \in \Omega(2)} c(2,v)} \quad .$$

Como $t = 0,1,2,3,4$, tem-se:

$$\Pr\left[T = 0 \mid M = 2;\ \beta = 0\right] = \frac{c(2,0)}{\sum_{v \in \Omega(2)} c(2,v)} = \frac{3}{36} = 0{,}0833$$

$$\Pr\left[T = 1 \mid M = 2;\ \beta = 0\right] = \frac{c(2,1)}{\sum_{v \in \Omega(2)} c(2,v)} = \frac{9}{36} = 0{,}2500$$

$$\Pr\left[T = 2 \mid M = 2;\ \beta = 0\right] = \frac{c(2,2)}{\sum_{v \in \Omega(2)} c(2,v)} = \frac{12}{36} = 0{,}3333$$

$$\Pr\left[T = 3 \mid M = 2;\ \beta = 0\right] = \frac{c(2,3)}{\sum_{v \in \Omega(2)} c(2,v)} = \frac{9}{36} = 0{,}2500$$

$$\Pr\left[T = 4 \mid M = 2;\ \beta = 0\right] = \frac{c(2,4)}{\sum_{v \in \Omega(2)} c(2,v)} = \frac{3}{36} = 0{,}0833$$

Note que, para $t = 4$, o valor-p exato equivale a 2 x 0,0833 = 0,1667.

Tabela 3.2- Possíveis configurações para os valores das estatísticas suficientes para β_0 e β^*.

Y_1	Y_2	Y_3	$M = \sum_{i=1}^{3} y_i$	$T = \sum_{i=1}^{3} y_i x_i$	$c(m,t) = \prod_{i=1}^{3} \binom{n_i}{y_i} = \prod_{i=1}^{3} \binom{3}{y_i}$ com $0 \le y_i \le 3$
0	0	0	0	0	$c(0,0) = \binom{3}{0} \times \binom{3}{0} \times \binom{3}{0} = 1$
0	0	1	1	2	$c(1,2) = \binom{3}{0} \times \binom{3}{0} \times \binom{3}{1} = 3$
0	0	2	2	4	$c(2,4) = \binom{3}{0} \times \binom{3}{0} \times \binom{3}{2} = 3$
0	0	3	3	6	$c(3,6) = \binom{3}{0} \times \binom{3}{0} \times \binom{3}{3} = 1$
0	1	0	1	1	$c(1,1) = \binom{3}{0} \times \binom{3}{1} \times \binom{3}{0} = 3$
0	1	1	2	3	$c(2,3) = \binom{3}{0} \times \binom{3}{1} \times \binom{3}{1} = 9$
0	1	2	3	5	$c(3,5) = \binom{3}{0} \times \binom{3}{1} \times \binom{3}{2} = 9$
0	1	3	4	7	$c(4,7) = \binom{3}{0} \times \binom{3}{1} \times \binom{3}{3} = 3$
0	2	0	2	2	$c(2,2) = \binom{3}{0} \times \binom{3}{2} \times \binom{3}{0} = 3$
0	2	1	3	4	$c(3,4) = \binom{3}{0} \times \binom{3}{2} \times \binom{3}{1} = 9$
0	2	2	4	6	$c(4,6) = \binom{3}{0} \times \binom{3}{2} \times \binom{3}{2} = 9$
0	2	3	5	8	$c(5,8) = \binom{3}{0} \times \binom{3}{2} \times \binom{3}{3} = 3$
0	3	0	3	3	$c(3,3) = \binom{3}{0} \times \binom{3}{3} \times \binom{3}{0} = 1$
0	3	1	4	5	$c(4,5) = \binom{3}{0} \times \binom{3}{3} \times \binom{3}{1} = 3$
0	3	2	5	7	$c(5,7) = \binom{3}{0} \times \binom{3}{3} \times \binom{3}{2} = 3$
0	3	3	6	9	$c(6,9) = \binom{3}{0} \times \binom{3}{3} \times \binom{3}{3} = 1$
1	0	0	1	0	$c(1,0) = \binom{3}{1} \times \binom{3}{0} \times \binom{3}{0} = 3$

1	0	1	2	2	$c(2,2)=\binom{3}{1}\times\binom{3}{0}\times\binom{3}{1}=9$
1	0	2	3	4	$c(3,4)=\binom{3}{1}\times\binom{3}{0}\times\binom{3}{2}=9$
1	0	3	4	6	$c(4,6)=\binom{3}{1}\times\binom{3}{0}\times\binom{3}{3}=3$
1	1	0	2	1	$c(2,1)=\binom{3}{1}\times\binom{3}{1}\times\binom{3}{0}=9$
1	1	1	3	3	$c(3,3)=\binom{3}{1}\times\binom{3}{1}\times\binom{3}{1}=27$
1	1	2	4	5	$c(4,5)=\binom{3}{1}\times\binom{3}{1}\times\binom{3}{2}=27$
1	1	3	5	7	$c(5,7)=\binom{3}{1}\times\binom{3}{1}\times\binom{3}{3}=9$
1	2	0	3	2	$c(3,2)=\binom{3}{1}\times\binom{3}{2}\times\binom{3}{0}=9$
1	2	1	4	4	$c(4,4)=\binom{3}{1}\times\binom{3}{2}\times\binom{3}{1}=27$
1	2	2	5	6	$c(5,6)=\binom{3}{1}\times\binom{3}{2}\times\binom{3}{2}=27$
1	2	3	6	8	$c(6,8)=\binom{3}{1}\times\binom{3}{2}\times\binom{3}{3}=9$
1	3	0	4	3	$c(4,3)=\binom{3}{1}\times\binom{3}{3}\times\binom{3}{0}=3$
1	3	1	5	5	$c(5,5)=\binom{3}{1}\times\binom{3}{3}\times\binom{3}{1}=9$
1	3	2	6	7	$c(6,7)=\binom{3}{1}\times\binom{3}{3}\times\binom{3}{2}=9$
1	3	3	7	9	$c(7,9)=\binom{3}{1}\times\binom{3}{3}\times\binom{3}{3}=3$
2	0	0	2	0	$c(2,0)=\binom{3}{2}\times\binom{3}{0}\times\binom{3}{0}=3$
2	0	1	3	2	$c(3,2)=\binom{3}{2}\times\binom{3}{0}\times\binom{3}{1}=9$
2	0	2	4	4	$c(4,4)=\binom{3}{2}\times\binom{3}{0}\times\binom{3}{2}=9$
2	0	3	5	6	$c(5,6)=\binom{3}{2}\times\binom{3}{0}\times\binom{3}{3}=3$
2	1	0	3	1	$c(3,1)=\binom{3}{2}\times\binom{3}{1}\times\binom{3}{0}=9$

2	1	1	4	3	$c(4,3) = \binom{3}{2} \times \binom{3}{1} \times \binom{3}{1} = 27$
2	1	2	5	5	$c(5,5) = \binom{3}{2} \times \binom{3}{1} \times \binom{3}{2} = 27$
2	1	3	6	7	$c(6,7) = \binom{3}{2} \times \binom{3}{1} \times \binom{3}{3} = 9$
2	2	0	4	2	$c(4,2) = \binom{3}{2} \times \binom{3}{2} \times \binom{3}{0} = 9$
2	2	1	5	4	$c(5,4) = \binom{3}{2} \times \binom{3}{2} \times \binom{3}{1} = 27$
2	2	2	6	6	$c(6,6) = \binom{3}{2} \times \binom{3}{2} \times \binom{3}{2} = 27$
2	2	3	7	8	$c(7,8) = \binom{3}{2} \times \binom{3}{2} \times \binom{3}{3} = 9$
2	3	0	5	3	$c(5,3) = \binom{3}{2} \times \binom{3}{3} \times \binom{3}{0} = 3$
2	3	1	6	5	$c(6,5) = \binom{3}{2} \times \binom{3}{3} \times \binom{3}{1} = 9$
2	3	2	7	7	$c(7,7) = \binom{3}{2} \times \binom{3}{3} \times \binom{3}{2} = 9$
2	3	3	8	9	$c(8,9) = \binom{3}{2} \times \binom{3}{3} \times \binom{3}{3} = 3$
3	0	0	3	0	$c(3,0) = \binom{3}{3} \times \binom{3}{0} \times \binom{3}{0} = 1$
3	0	1	4	2	$c(4,2) = \binom{3}{3} \times \binom{3}{0} \times \binom{3}{1} = 3$
3	0	2	5	4	$c(5,4) = \binom{3}{3} \times \binom{3}{0} \times \binom{3}{2} = 3$
3	0	3	6	6	$c(6,6) = \binom{3}{3} \times \binom{3}{0} \times \binom{3}{3} = 1$
3	1	0	4	1	$c(4,1) = \binom{3}{3} \times \binom{3}{1} \times \binom{3}{0} = 3$
3	1	1	5	3	$c(5,3) = \binom{3}{3} \times \binom{3}{1} \times \binom{3}{1} = 9$
3	1	2	6	5	$c(6,5) = \binom{3}{3} \times \binom{3}{1} \times \binom{3}{2} = 9$
3	1	3	7	7	$c(7,7) = \binom{3}{3} \times \binom{3}{1} \times \binom{3}{3} = 3$
3	2	0	5	2	$c(5,2) = \binom{3}{3} \times \binom{3}{2} \times \binom{3}{0} = 3$

3	2	1	6	4	$c(6,4) = \binom{3}{3} \times \binom{3}{2} \times \binom{3}{1} = 9$
3	2	2	7	6	$c(7,6) = \binom{3}{3} \times \binom{3}{2} \times \binom{3}{2} = 9$
3	2	3	8	8	$c(8,8) = \binom{3}{3} \times \binom{3}{2} \times \binom{3}{3} = 3$
3	3	0	6	3	$c(6,3) = \binom{3}{3} \times \binom{3}{3} \times \binom{3}{0} = 1$
3	3	1	7	5	$c(7,5) = \binom{3}{3} \times \binom{3}{3} \times \binom{3}{1} = 3$
3	3	2	8	7	$c(8,7) = \binom{3}{3} \times \binom{3}{3} \times \binom{3}{2} = 3$
3	3	3	9	9	$c(9,9) = \binom{3}{3} \times \binom{3}{3} \times \binom{3}{3} = 1$

Tabela 3.3- Cálculo de $\sum \prod_{i=1}^{3} \binom{3}{y_i}$

M	T	$C(M,T)$	M	T	$C(M,T)$	M	T	$C(M,T)$
0	0	1	4	1	3	6	5	18
1	0	3	4	2	12	6	6	28
1	1	3	4	3	30	6	7	18
1	2	3	4	4	36	6	8	9
2	0	3	4	5	30	6	9	1
2	1	9	4	6	12	7	5	3
2	2	12	4	7	3	7	6	9
2	3	9	5	2	3	7	7	12
2	4	3	5	3	12	7	8	9
3	0	1	5	4	30	7	9	3
3	1	9	5	5	36	8	7	3
3	2	18	5	6	30	8	8	3
3	3	28	5	7	12	8	9	3
3	4	18	5	8	3	9	9	1
3	5	9	6	3	1			
3	6	1	6	4	9			

3.1.4- Exemplos Revisados

Exemplo 1 revisado: Considerando os dados do **Exemplo 1**, a Tabela 3.4 mostra os resultados do ajuste do modelo logístico através do *software* LogXact.

Tabela 3.4- Resultados do ajuste do modelo logístico usando o método exato.

Covariável	$\hat{\beta}$	**IC para** β **(95%)**	$valor - p$
X_1 **: gravidade do caso (0: baixa e 1: alta)**	1,3302	[-1,2657 ; ∞)	0,3125
X_2 **: duração da cirurgia (em horas)**	0,3776	[-0,1778 ; 0,9665]	0,1711

Exemplo 3 revisado: Considerando os dados do **Exemplo 3**, a Tabela 3.5 mostra os resultados do ajuste do modelo logístico através do *software* LogXact.

Tabela 3.5- Resultados do ajuste do modelo logístico usando o método exato.

Covariável	$\hat{\beta}$	**IC para** β **(95%)**	$valor - p$
X_1 **: uso de corticóide (0: não e 1: sim)**	-2,6641	[-6,9216;-0,1995]	0,0266
X_2 **: tipo de parto (0:cesariano e 1: vaginal)**	-0,2811	[-2,8769;1,7770]	1,0000
X_3 **: peso (gramas) ao nascer (0:** $\geq$ **1000 e 1: <1000)**	1,1308	[-1,0547;3,6539]	0,4308
X_4 **: idade gestacional calculada (semanas)**	0,2195	[-0,1494;0,5949]	0,2412
X_{51} **: indicadora de infecção (1:provável)**	-1,1499	[-3,8715;+∞)	1,0000
X_{52} **: indicadora de infecção (2:confirmada)**	-2,6728	[-6,3364; +∞)	1,0000
X_6 **: VM**	0,0887	[0,0319;0,1608]	0,0010
X_7 **: VPP (0:não e 1:sim)**	1,8947	[-0,1780;4,6421]	0,0845

Pode-se observar que o método exato encontra estimativas para os parâmetros do modelo. No entanto, a covariável que gera a condição de separação quase-completa

apresenta uma grande incerteza na sua estimativa caracterizada pelo valor infinito do seu limite superior do intervalo de confiança. Este fato ocorre nos dois conjuntos de dados. Isso significa que as estimativas geradas para a variável que gera a situação de quase-separação não são confiáveis.

3.2 - Adicionar uma Constante aos Dados

O procedimento de adicionar uma pequena constante K, próxima de zero, ao conjunto de dados categórico tem sido muito comum, principalmente em tabelas de contingência esparsas. Uma solução que consiste em deixar o conjunto de dados indicar qual é a melhor constante a ser usada no intervalo (0, 0,5] aparece em Colosimo, Franco e Couto (1995). Este procedimento é simples e consiste em adicionar uma constante selecionada para cada N_{1i} e N_{0i}. Colosimo, Franco e Couto (1995) utilizaram de repetições de Monte Carlo para encontrar a melhor constante. Todas as repetições usadas nas simulações geraram conjuntos de dados na categoria de separação quase-completa. O Erro Quadrático Médio (EQM) foi a medida estatística usada como critério para escolher, dentre uma série de constantes, a mais adequada. Serão usadas neste trabalho as constantes 0,02 e 0,03.

3.2.1- Exemplos Revisados

Exemplo 1 revisado: Considerando os dados do **Exemplo 1**, as Tabelas 3.6 e 3.7 mostram os resultados do ajuste do modelo logístico após adicionar as constantes $K = 0,02$ e $K = 0,03$ aos dados.

Tabela 3.6- Ajuste do modelo de regressão logística, adicionando $k = 0,02$.

Covariável	**Estimativa**	**IC para β (95%)**	*Valor – p*
X_1 : gravidade do caso ($0 : baixa \quad e \; 1 : alta$)	2,0968	[- 1,0380 ; 5,2316]	0,1899
X_2 : duração da cirurgia ($em \quad horas$)	0,3434	[- 0,1407 ; 0,8275]	0,1644

Tabela 3.7- Ajuste do modelo de regressão logística, adicionando $k = 0,03$.

Covariável	**Estimativa**	**IC para β (95%)**	*Valor – p*
X_1 : gravidade do caso $(0 : baixa \quad e\ 1 : alta)$	1,7687	[- 0,8989 ; 4,4363]	0,1937
X_2 : duração da cirurgia $(em \quad horas)$	0,3229	[- 0,1373 ; 0,7831]	0,1690

Exemplo 3 revisado: Considerando os dados do **Exemplo 3**, as Tabelas 3.8 e 3.9 mostram os resultados do ajuste do modelo logístico após adicionar as constantes $K = 0,02$ e $K = 0,03$ aos dados.

Tabela 3.8- Ajuste do modelo de regressão logística, adicionando $k = 0,02$.

Covariável	**Estimativa**	**IC para β (95%)**	*Valor – p*
X_1	-1,6317	[-3,3745 ; 0,1111]	0,0665
X_2	-0,3095	[-1,8953 ; 1,2763]	0,7021
X_3	0,1187	[-0,1682 ; 0,4056]	0,4176
X_4	0,0406	[0,0028 ; 0,0784]	0,0352
X_{51}	1,1095	[-0,3905 ; 2,6095]	0,1471
X_{52}	0,6434	[-0,8476 ; 2,1344]	0,3977
X_6	1,4393	[-3,1877 ; 6,0663]	0,5421
X_7	0,5538	[-4,0812 ; 5,1888]	0,8148

Tabela 3.9- Ajuste do modelo de regressão logística, adicionando $k = 0,03$.

Covariável	**Estimativa**	**IC para β (95%)**	*Valor – p*
X_1	-1,4333	[-3,0166 ; 0,1499]	0,0760
X_2	-0,2609	[-1,7842 ; 1,2624]	0,7371
X_3	0,1050	[-0,1692 ; 0,3792]	0,4529
X_4	0,0381	[0,0018 ; 0,0744]	0,0399
X_{51}	1,0100	[-0,3971; 2,4171]	0,1595
X_{52}	0,5754	[-0,8458 ; 1,9966]	0,4274
X_6	1,1673	[-2,6827 ; 5,0173]	0,5524
X_7	0,3640	[-3,4819 ; 4,2099]	0,8528

Pode-se notar que os resultados obtidos para $K = 0,02$ e $K = 0,03$ são similares, tanto no **Exemplo 1** quanto no **Exemplo 3**.

3.3- Proposta Nova

Este procedimento consiste em retirar aleatoriamente uma observação de uma das caselas não-nulas (com mesmo valor da covariável ou mesmo valor da resposta) e adicioná-la à casela nula. Ou seja, deve-se trocar o resultado de uma observação. Com isso, torna-se possível usar o modelo logístico na sua forma usual através dos *softwares* estatísticos disponíveis, pois o conjunto de dados não mais estará na condição de separação quase-completa.

Para maior clareza desta proposta, considere a representação da Tabela 3.10 resultante do cruzamento da covariável binária X_1 com a resposta Y. Note que a casela c=0.

Tabela 3.10- Cruzamento da covariável binária X_1 com a resposta Y, com c=0

Y	X_1		***Total***
	0	***1***	
0	30	15	45
1	0	2	2
Total	30	17	47

De acordo com o **Teorema 1**, o conjunto de dados está na categoria de separação quase-completa. Esta proposta sugere a retirada aleatória de uma observação da casela não-nula $(Y = 0 \; e \; X_1 = 0, \; ou \; Y = 1 \; e \; X_1 = 1)$ que represente o maior total (neste caso, 30) e adicioná-la à casela nula. Assim, a tabela 3.10 fica alterada para

Tabela 3.11- Cruzamento da covariável binária X_1 com a resposta Y, com c=1

Y	X_1		***Total***
	0	***1***	
0	29	15	44
1	1	2	3
Total	30	17	47

Os dados dispostos agora, como na Tabela 3.11, permitem que o modelo logístico possa ser ajustado normalmente.

Note que, retirando a observação da casela que representa o maior valor, não está sendo questionado o quão sério seria mudar de $(Y = 0 \; e \; X_1 = 0)$ para $(Y = 1 \; e \; X_1 = 0)$.

Para exemplificar melhor esta proposta de solução, serão usados os dados dos **Exemplos 1 e 3**.

3.3.1- Exemplos Revisados

Exemplo 1 revisado : Considere a representação dada na **Tabela 2.1** resultante do cruzamento de Y com X_1 binária. De acordo com os resultados apresentados na **Tabela 2.2**, temos que a casela $c = 0$. Será apresentada a solução proposta em que a observação retirada aleatoriamente pertence às caselas a (mesmo valor de X_1) e d (mesmo valor da resposta N_1).

- Retirada aleatória de uma observação da casela $a = 68$ resultante do cruzamento de $X_1 = 0$ e $N_1 = 0$.

  O número 22 foi sorteado de forma aleatória, que corresponde ao paciente 26, conforme a Tabela 3.10.

Tabela 3.10- Seleção aleatória da observação da casela $a = 68$ a ser adicionada à casela nula.

PACIENTE	X1	X2	N1	ORDEM	Nº ALEATORIO
26	0	3,5	0	22	22

Ao adicionar esta observação à casela nula, a **Tabela 2.2** fica mudada conforme a Tabela 3.11 e os resultados para o ajuste do modelo logístico encontram-se na Tabela 3.12.

Tabela 3.11- Cruzamento da covariável binária X_1 com a resposta N_1, após adicionar à casela nula a observação da casela $a = 68$.

	X_1: gravidade do caso		
N_1: ocorrência de meningite	***0: baixa***	***1: alta***	***Total***
0: não	67	32	99
1: sim	1	2	3
Total	68	34	102

Tabela 3.12- Ajuste do modelo de regressão logística, após adicionar à casela nula a observação da casela $a = 68$.

Parâmetro	$\hat{\beta}$	**IC para β (95%)**	*Valor – p*
X_1 : gravidade do caso $(0 : baixa \quad e \; 1 : alta)$	1,246	[-1,2569 ; 3,7489]	0,329
X_2 : duração da cirurgia $(em \quad horas)$	0,3425	[-0,1385 ; 0,8235]	0,163

- Retirada aleatória de uma observação da casela $d = 2$ resultante do cruzamento de $X_1 = 1$ e $N_1 = 1$.

O número 1 foi sorteado de forma aleatória, que corresponde ao paciente 100, conforme a Tabela 3.13.

Tabela 3.13- Seleção aleatória da observação da casela $d = 2$ a ser adicionada à casela nula.

PACIENTE	X1	X2	N1	ORDEM	Nº ALEATORIO
100	1	1,5	1	1	1

Ao adicionar esta observação à casela nula, tem-se a Tabela 3.14 e os resultados para o ajuste do modelo logístico encontram-se na Tabela 3.15.

Tabela 3.14- Cruzamento da covariável binária X_1 com a resposta N_1, após adicionar à casela nula a observação da casela $d = 2$.

	X_1: gravidade do caso		
N_1: ocorrência de meningite	***0: baixa***	***1: alta***	***Total***
0: não	68	32	100
1: sim	1	1	2
Total	69	33	102

Tabela 3.15- Ajuste do modelo de regressão logística, após adicionar à casela nula a observação da casela $d = 2$.

Parâmetro	$\hat{\beta}$	**IC para β (95%)**	*Valor – p*
X_1 : gravidade do caso $(0 : baixa \quad e \; 1 : alta)$	0,286	[-2,7598 ; 3,3318]	0,854
X_2 : duração da cirurgia $(em \quad horas)$	0,4998	[-0,1111 ; 1,1107]	0,109

Exemplo 3 revisado: Considere a representação dada na **Tabela 2.3** resultante do cruzamento de Y com X_1 categórica. De acordo com os resultados da **Tabela 2.6**, tem-se que a casela $d = 0$. Será apresentada a solução proposta em que a observação retirada aleatoriamente pertence às caselas a e $(e \; ou \quad f)$.

- Retirada aleatória de uma observação da casela $a = 11$ resultante do cruzamento de $X_5 = 0$ e $Y = 0$.

O número 9 foi sorteado de forma aleatória, que corresponde à paciente 60.
Ao adicionar essa observação à casela nula, a Tabela 2.6 fica mudada conforme a Tabela 3.16 e, os resultados para o ajuste do modelo logístico encontram-se na Tabela 3.17.

Tabela 3.16- Cruzamento da resposta Y com a variável categórica X_5, após adicionar à casela nula a observação da casela $a = 11$.

	Infecção			
Hemorragia	***0: não***	***1: provável***	***2: confirmada***	***Total***
0: grau 1	10	22	58	90
1: graus 2 a 4	1	5	8	14
Total	11	27	66	104

Tabela 3.17- Ajuste do modelo de regressão logística, após adicionar à casela nula a observação da casela $a = 11$.

Parâmetro	$\hat{\beta}$	**IC para β (95%)**	*Valor – p*
X_1	-2,1062	[-4,0145 ; -0,19794]	0,031
X_2	-0,0095	[-1,8682 ; 1,84917]	0,992
X_3	1,0578	[-0,7129 ; 2,82846]	0,242
X_4	0,2884	[-0,0568 ; 0,63356]	0,102
X_{51}	-1,086	[-3,8868 ; 1,71484]	0,447
X_{52}	-2,5	[-5,4322 ; 0,43216]	0,095
X_6	0,10076	[0,0363 ; 0,16519]	0,002
X_7	1,7344	[-0,0131 ; 3,48194]	0,052

- Retirada aleatória de uma observação da casela $(e\ ou\ f)$ resultante do cruzamento de $X_5 = 1$ e $Y = 1$ ou $X_5 = 2$ e $Y = 1$.

O número 9 foi sorteado de forma aleatória que corresponde à paciente 53.
Ao adicionar esta observação à casela nula, a Tabela 2.6 fica mudada conforme a Tabela 3.18 e os resultados para o ajuste do modelo logístico encontram-se na Tabela 3.19.

Tabela 3.18- Cruzamento da resposta Y com a variável categórica X_5, após adicionar à casela nula a observação da casela $f = 8$.

	Infecção			
Hemorragia	***0: não***	***1: provável***	***2: confirmada***	***Total***
0: grau 1	11	22	58	91
1: graus 2 a 4	1	5	7	13
Total	12	27	65	104

Tabela 3.19- Ajuste do modelo de regressão logística, após adicionar à casela nula a observação da casela $f = 8$.

Parâmetro	$\hat{\beta}$	**IC para β (95%)**	$Valor - p$
X_1	-3,047	[-5,6871 ; -0,40688]	0,024
X_2	-0,196	[-2,2756 ; 1,88356]	0,853
X_3	1,1675	[-0,7919 ; 3,12691]	0,243
X_4	0,3020	[-0,0904 ; 0,69439]	0,131
X_{51}	0,016	[-3,3415 ; 3,37348]	0,993
X_{52}	-1,762	[-5,1469 ; 1,62292]	0,308
X_6	0,11402	[0,0413 ; 0,18672]	0,002
X_7	2,307	[0,1608 ; 4,45320]	0,035

Esta proposta que consiste simplesmente em retirar aleatoriamente uma observação de uma das caselas não-nulas (com mesmo valor da covariável ou mesmo valor da resposta) e adicioná-la à casela nula, gera estimativas parecidas para os parâmetros do

modelo independente de onde é tirada a observação a ser acrescentada à casela nula, exceto para a covariável que gera a condição de separação quase-completa.

Capítulo 4

Simulações de Monte Carlo

Com os resultados da análise dos dados, surgem perguntas como: Qual é a melhor proposta de solução? Que medida estatística usar para ajudar a identificar qual é a melhor proposta?

Após gerar em torno de 1.000 simulações, o critério que será usado para ajudar a identificar a melhor proposta de solução (Regressão Logística Exata; Soma de uma pequena constante K aos dados; Alocação de uma observação da casela vizinha à casela nula) será baseado no Erro Quadrático Médio (EQM).

Para obter o EQM foi usado parte de um programa em *Fortran* desenvolvido para gerar os resultados apresentados em Colosimo, Franco e Couto (1995). Este programa foi atualizado (*Fortran PowerStation 4.0*) na proposta apresentada que *soma uma pequena constante* K *aos dados* e, ainda, acrescentado a ele, a proposta que *aloca uma observação da casela vizinha à casela nula*. Foram fixados os valores das covariáveis X_1 *e* X_2 do **Exemplo 1** e os betas foram escolhidos, de modo que, a chance dos dados cairem na situação de quase-separação fosse mais freqüente. Dados na situação *overlap* foram totalmente descartados. Os mesmos $Y's$ simulados que foram gerados neste programa em *Fortran* foram armazenados em um arquivo para serem processados no *LogXact* para a execução da proposta que usa a *regressão logística exata*. Gerado o arquivo de saída do *LogXact*, foram guardadas as estimativas dos betas e usados os seus resultados para calcular o EQM num programa em *Delphi*.

Na seção seguinte apresenta-se o cálculo do EQM.

4.1- Cálculo do Erro Quadrático Médio

As três propostas de solução apresentadas nesta dissertação têm como objetivo gerar as estimativas dos parâmetros e, conseqüentemente, o EQM, dado por

$$EQM = \left(\overline{\hat{\beta}} - \beta \right)^2 + \hat{Var}\left(\hat{\beta}\right)$$

em que:

$$\overline{\hat{\beta}} = \frac{\sum_{i=1}^{N} \hat{\beta}_i}{N} \qquad \hat{Var}\left(\hat{\beta}\right) = \frac{\sum_{i=1}^{N} \left(\hat{\beta}_i - \overline{\hat{\beta}} \right)^2}{(N-1)}$$

em que:

N = Número de Repetições Válidas,

β = valores verdadeiros.

As Tabelas 4.1, 4.2 e 4.3 apresentam os resultados do Erro Quadrático Médio para as três propostas de solução, respectivamente.

Tabela 4.1- Cálculo do EQM para a Proposta de Solução que usa a Regressão Logística Exata.

BETAS	***Erro Quadrático Médio***			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	**N**
BETA0=-9; BETA1=3.0; BETA2=0.1	41,02618	6,397636	0,1550247	786
BETA0=-8; BETA1=3.0; BETA2=0.1	27,76298	5,922798	0,1477597	802
BETA0=-7; BETA1=3.0; BETA2=0.1	16,26940	4,615071	0,1645181	858

Dentre as 1.000 simulações de Y devem-se subtrair os casos em que a distribuição condicional da estatística suficiente para o parâmetro de interesse é degenerada. Isso

justifica o fato de que o número efetivo de simulações na Tabela 4.1, representado por N, não totaliza 1.000.

Tabela 4.2- Cálculo do EQM para a Proposta de Solução que usa a Soma da Constante $K = 0,02$ aos dados.

BETAS	***Erro Quadrático Médio***			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	**N**
BETA0=-9; BETA1=3.0; BETA2=0.1	15,81359	1,93785	0,09110	1.000
BETA0=-8; BETA1=3.0; BETA2=0.1	8,83495	1,81893	0,08955	1.000
BETA0=-7; BETA1=3.0; BETA2=0.1	4,19854	1,41465	0,08910	1.000

Tabela 4.3- Cálculo do EQM para a Proposta de Solução que Aloca aleatoriamente uma Observação da Casela Vizinha à Casela Nula.

BETAS	***Erro Quadrático Médio***			
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	**N**
BETA0=-9; BETA1=3.0; BETA2=0.1	28,01917	9,70956	0,03975	910
BETA0=-8; BETA1=3.0; BETA2=0.1	25,54767	9,66912	0,06623	902
BETA0=-7; BETA1=3.0; BETA2=0.1	12,55776	9,65856	0,09846	923

Nota-se que, em geral, a proposta que usa como solução a Regressão Logística Exata parece ser a pior, uma vez que apresentou os maiores valores para o EQM. Já a proposta que gerou os melhores valores para o EQM foi aquela que soma a constante $K = 0,02$ aos dados.

Capítulo 5

Resultados e Conclusões

5.1- Resultados

As tabelas-resumo (Tabelas 5.1 e 5.2) para os exemplos 1 e 3, citados no texto, mostram os resultados para as três propostas de solução.

Tabela 5.1- Resultados dos ajustes para o **Exemplo 1**, usando as três propostas.

Covariável	Método Exato			Adicionar k				Proposta Nova			
	$\hat{\beta}$	IC para β	valor − p	Valor de K	$\hat{\beta}$	IC para β	valor − p	Casela	$\hat{\beta}$	IC para β	valor − p
X_1	1,3302	[-1,2657 ; ∞]	0,3125	K=0,02	2,0968	[-1,0380 ; 5,2316]	0,1899	a = 68	1,246	[-1,2569 ; 3,7489]	0,329
				K=0,03	1,7687	[-0,8989 ; 4,4363]	0,1937	d = 2	0,286	[-2,7598 ; 3,3318]	0,854
X_2	0,3776	[-0,1778 ; 0,9665]	0,1711	K=0,02	0,3434	[-0,1407 ; 0,8275]	0,1644	a = 68	0,3425	[-0,1385 ; 0,8235]	0,163
				K=0,03	0,3229	[-0,1373 ; 0,7831]	0,1690	d = 2	0,4998	[-0,1111 ; 1,1107]	0,109

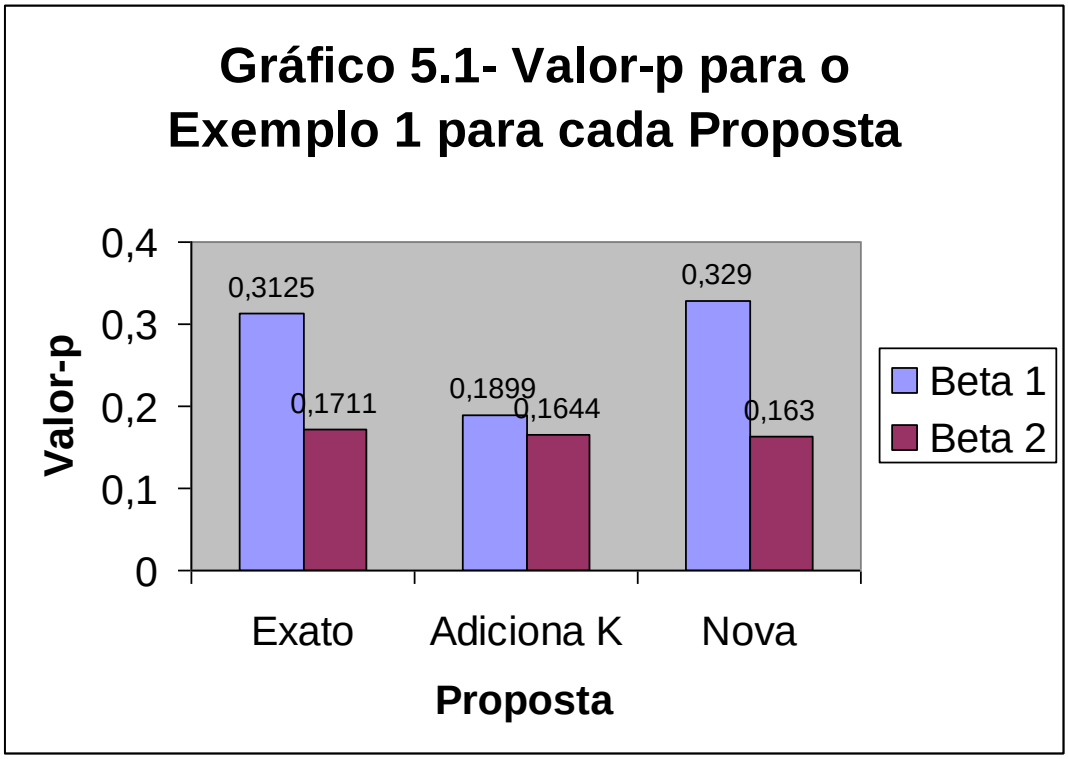

Figura 5.1- Resultados do Valor-p considerando cada proposta de solução para os dados do Exemplo 1.

Pode-se notar, através da Tabela 5.1 e da Figura 5.1, que nenhuma das covariáveis são importantes para explicar a ocorrência de meningite nas três propostas de solução. As estimativas de β_2 são bem parecidas nas três propostas de solução, conforme Figura 5.1(b).

Tabela 5.2- Resultados dos ajustes para o **Exemplo 3**, usando as três propostas.

Covariável	Método Exato			Adicionar k				Proposta Nova			
	$\hat{\beta}$	IC para β	valor − p	K	$\hat{\beta}$	IC para β	valor − p	Casela	$\hat{\beta}$	IC para β	valor − p
X_1	-2,664	[-6,9216 ; -0,1995]	0,0266	K=0,02	-1,6317	[-3,3745 ; 0,1111]	0,0665	a=11	-2,1062	[-4,0145 ; -0,1979]	0,031
				K=0,03	-1,4333	[-3,0166 ; 0,1499]	0,0760	f = 8	-3,047	[-5,6871 ; -0,4069]	0,024
X_2	-0,281	[-2,8769 ; 1,7770]	1,0000	K=0,02	-0,3095	[-1,8953 ; 1,2763]	0,7021	a=11	-0,0095	[-1,8682 ; 1,8492]	0,992
				K=0,03	-0,2609	[-1,7842 ; 1,2624]	0,7371	f = 8	-0,196	[-2,2756 ; 1,8836]	0,853
X_3	1,131	[-1,0547 ; 3,6539]	0,4308	K=0,02	0,1187	[-0,1682 ; 0,4056]	0,4176	a=11	1,0578	[-0,7129 ; 2,8285]	0,242
				K=0,03	0,1050	[-0,1692 ; 0,3792]	0,4529	f = 8	1,1675	[-0,7919 ; 3,1269]	0,243
X_4	0,219	[-0,1494 ; 0,5949]	0,2412	K=0,02	0,0406	[0,0028 ; 0,0784]	0,0352	a=11	0,2884	[-0,0568 ; 0,6336]	0,102
				K=0,03	0,0381	[0,0018 ; 0,0744]	0,0399	f = 8	0,3020	[-0,0904 ; 0,6944]	0,131
X_{51}	-1,149	[-3,8715 ; + ∞]	1,0000	K=0,02	1,1095	[-0,3905 ; 2,6095]	0,1471	a=11	-1,086	[-3,8868 ; 1,7148]	0,447
				K=0,03	1,0100	[-0,3971 ; 2,4171]	0,1595	f = 8	0,016	[-3,3415 ; 3,3735]	0,993
X_{52}	-2,673	[-6,3364 ; + ∞]	1,0000	K=0,02	0,6434	[-0,8476 ; 2,1344]	0,3977	a=11	-2,5	[-5,4322 ; 0,4322]	0,095
				K=0,03	0,5754	[-0,8458 ; 1,9966]	0,4274	f = 8	-1,762	[-5,1469 ; 1,6229]	0,308
X_6	0,089	[0,0319 ; 0,1608]	0,0010	K=0,02	1,4393	[-3,1877 ; 6,0663]	0,5421	a=11	0,10076	[0,0363 ; 0,1652]	0,002
				K=0,03	1,1673	[-2,6827 ; 5,0173]	0,5524	f = 8	0,1140	[0,0413 ; 0,1867]	0,002
X_7	1,895	[-0,1780 ; 4,6421]	0,0845	K=0,02	0,5538	[-4,0812 ; 5,1888]	0,8148	a=11	1,7344	[-0,0131 ; 3,4819]	0,052
				K=0,03	0,3640	[-3,4819 ; 4,2099]	0,8528	f = 8	2,307	[0,1608 ; 4,4532]	0,035

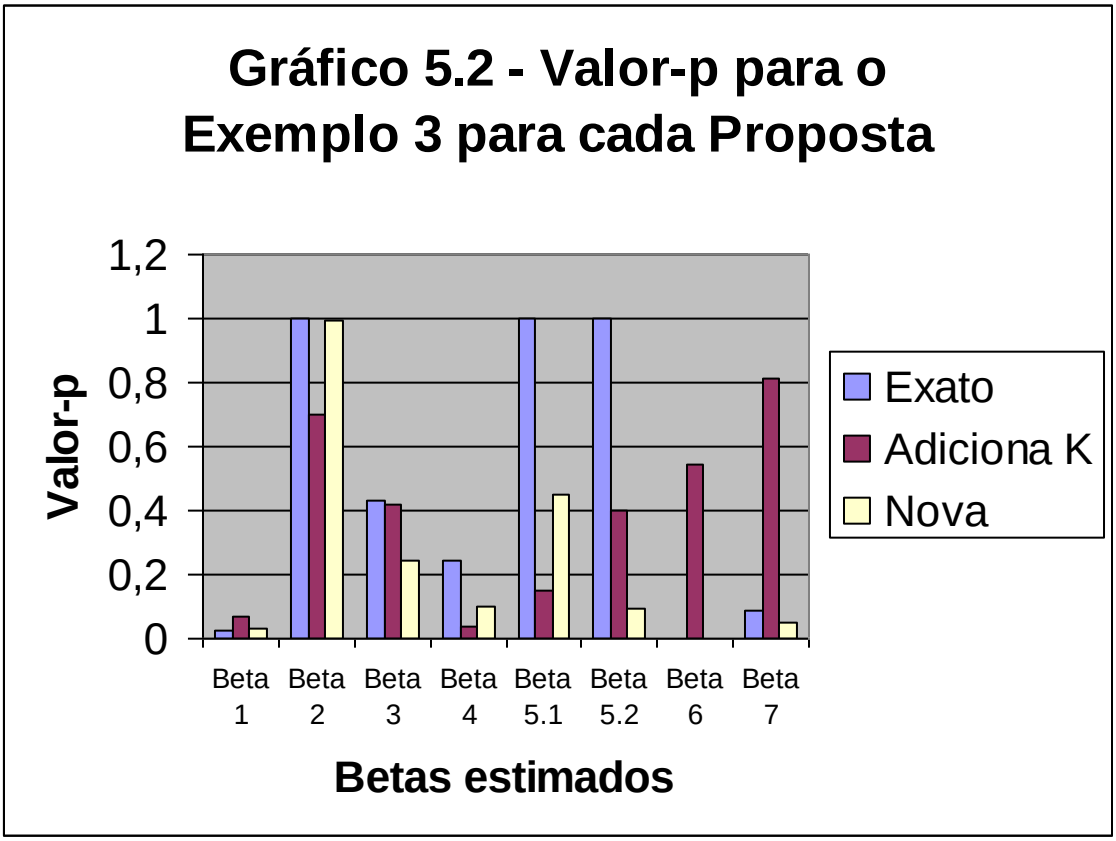

Figura 5.2- Resultados do Valor-p considerando cada proposta de solução para os dados do Exemplo 3.

Pode-se notar, através da Tabela 5.2 e da Figura 5.2, que as covariáveis importantes para explicar a *presença (ou não) de hemorragia peri-intraventricular no parto* em cada uma das três propostas de solução são, para um nível de significância de 5% :

- **Método Exato :**

  X_1 : uso de corticóide (0: Não; 1: Sim)

  X_6 : número de dias em que a criança permaneceu sob ventilação mecânica assistida (VM)

- **Nova Proposta :**

  X_1 : uso de corticóide (0: Não; 1: Sim)

  X_6 : número de dias em que a criança permaneceu sob ventilação mecânica assistida (VM)

  X_7 : ventilação por pressão positiva (VPP) – ambu (0: Não; 1: Sim)

- **Adiciona a constante K=0,02 :**

  X_4 : IGC (Idade Gestacional Calculada : em semanas)

5.2- Conclusões

Três soluções foram propostas para uma situação de regressão logística em que a análise padrão não pode ser aplicada. As propostas: *Soma de uma Pequena Constante* $K = 0,02$ *aos dados* e, *Alocação de uma Observação Vizinha à Casela Nula* tornaram possível usar o modelo logístico na sua forma usual, numa situação em que os estimadores de máxima verossimilhança não existem. Os resultados baseados no Erro Quadrático Médio foram melhores para tais propostas, se comparados com aquela que usa a *Regressão Logística Exata*.

O procedimento que consiste em adicionar uma pequena constante, próxima de zero, aos dados parece ser uma solução razoável e simples quando os dados estão classificados na categoria de Separação Quase Completa.

O uso da Regressão Logística Exata, através do *software LogXact*, como uma das propostas de solução, apresentou as seguintes limitações :

- Quando a função de verossimilhança condicional não pode ser maximizada, o *LogXact* calcula uma estimativa pontual não-viciada da mediana (MUE), que, neste caso, deve ser usada com cautela. A variável que gera a situação de quase-separação apresenta intervalo de confiança para o parâmetro correspondente com limite indeterminado;

- O *Logxact* trabalha com um número limitado de variáveis (colunas) no arquivo de dados, o que gerou dificuldades para gerar os resultados;

- Em determinadas situações esse programa não gerou alguns betas, apresentando a informação de que a distribuição condicional da estatística suficiente para o parâmetro de interesse era degenerada.

Referências Bibliográficas

ALBERT, A., ANDERSON, J.A. **On the Existence of Maximum Likelihood Estimates in Logistic Regression Models.** Biometrika, 71, pp.1-10, 1984.

COLOSIMO, E. A., FRANCO, G. C., COUTO, B. M. **The Logistic Regression Model and Rare Events.** Estadistica, 47, 148, 149, pp. 1-16, 1995.

COX, D.R. **Analysis of Binary Data.** New York: Chapman and Hall, 1970.

HOSMER, D.W., LEMESHOW, S. **Applied Logistic Regression.** Wiley, New York, 1989.

HIRJI, K. E. , MEHTA, C. R. , PATEL, N. R**. Computing Distributions for Exact Logistic Regression,** Journal of the American Statistical Association, 82, 1110-1117, 1987.

HIRJI, K. E. , MEHTA, C. R. , PATEL, N. R. **Exact Inference for Matched Case-Control Studies**, Biometrics, 44, 803-814, 1988.

HIRJI, K. E. **Exact Distributions for Polytomous Data**, Journal of the American Statistical Association, 87, 487-492, 1992.

KING, E. N., RYAN, T. P. **A Preliminary Investigation of Maximum Likelihood Logistic Regression versus Exact Logistic Regression.** The American Statistician, vol. 56, No. 3, pp. 162-171, August 2002.

LOGXACT FOR WINDOWS, **Logistic Regression Software Featuring Exact Methods**, Cytel Software Corporation, Cambridge, MA, 2000.

McCULLAGH, P., J. A. NELDER. **Generalized Linear Models.** London: Chapman and Hall, 1989.

MEHTA, C. R. , PATEL, N. R. **Exact Logistic Regression: Theory and Examples.** Statistics in Medicine, vol. 14, 2143-2160, 1995.

MEHTA, C. R. , PATEL, N. R., SENCHAUDHURI, P. **Exact Stratified Linear Rank Tests for Ordered Categorical and Binary Data**, Journal of Computational and Graphical Statistics, 1, 21-40, 1992.

MEHTA, C. R. , PATEL, N. R., SENCHAUDHURI, P. **Efficient Monte Carlo Methods for Conditional Logistic Regression**, Journal of the American Statistical Association, 95, 99-108, 2000.

MICROSOFT FORTRAN POWERSTATION 4.0, 1994-95.

MINITAB FOR WINDOWS, Release 12.2, 1998.

NELDER , J. A., WEDDERBURN, R. W. M. **Generalized Linear Models.** J. R. Statist. Soc. A, 135, 370-84, 1972.

SANTNER, T.J., DUFFY, D.E. **A Note on A. Albert and J. A. Anderson´s Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models.** Biometrika, 73, pp. 755- 758, 1986.

SOUZA,M. C. F. M. C. **Regressão Logística Exata para Dados de Resposta Binária.** Dissertação de Mestrado. Belo Horizonte: Departamento de Estatística da UFMG; 2000.

TRITCHLER, D. **An Algorithm for Exact Logistic Regression.** Journal of the American Statistical Association,79, 709-711, 1984.

ANEXO A1

Tabela A1 - Dados referentes a 104 pacientes onde a resposta de interesse é a presença (ou não) de hemorragia peri-intraventricular no parto.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	N_1	N_0
1	0	0	30	0	0	1	0	1
0	0	1	29	0	0	0	0	1
1	0	0	30	0	1	1	0	1
1	0	0	30	0	2	1	0	1
0	0	0	29	0	0	0	0	1
1	0	0	33	0	0	0	0	2
1	0	0	29	0	2	1	0	1
0	0	0	28	0	0	1	0	1
1	0	0	32	0	0	0	0	1
0	0	0	33	0	0	0	0	1
0	0	0	34	1	2	0	0	1
0	0	0	32	1	4	1	0	1
0	0	0	31	1	5	1	1	1
1	0	0	31	1	0	0	0	1
0	1	1	24	1	37	1	1	1
1	0	0	32	1	0	0	0	2
0	0	0	32	1	1	0	0	1
0	0	0	28	1	0	0	0	1
0	0	1	27	1	10	1	0	1
0	1	0	28	1	27	1	1	1
0	1	1	24	1	23	1	1	1
0	0	1	28	1	0	0	0	1
1	0	1	32	1	0	0	0	1
0	0	0	33	1	0	0	0	1
0	1	0	28	1	2	0	0	1
0	0	1	28	1	25	0	0	1
0	1	0	32	1	12	1	0	1
1	0	0	30	1	0	0	0	1
1	0	0	30	1	2	1	0	1
1	1	1	27	1	30	0	0	1
1	0	1	28	1	14	1	0	1
1	0	0	28	1	9	0	0	1
1	0	1	28	1	3	1	0	1
1	1	0	29	1	0	0	0	1
1	0	1	34	1	16	1	1	1
1	0	0	31	1	3	0	0	1
0	0	0	28	2	25	1	0	1
0	0	1	27	2	29	1	0	1

X_1	X_2	X_3	X_4	X_5	X_6	X_7	N_1	N_0
0	0	1	27	2	49	1	1	1
0	1	0	37	2	13	0	0	1
0	0	1	30	2	13	1	0	1
0	0	1	26	2	9	1	0	1
1	1	0	29	2	6	0	0	1
0	0	1	29	2	3	1	0	1
1	0	0	29	2	5	1	0	1
0	0	1	31	2	0	1	0	1
1	0	0	32	2	0	0	0	2
1	0	0	33	2	0	0	0	1
0	1	1	27	2	40	1	0	1
0	0	0	32	2	0	0	0	2
0	1	0	26	2	6	1	0	1
0	1	1	29	2	15	1	1	1
1	0	0	31	2	0	0	0	1
0	0	0	34	2	0	0	0	1
1	0	1	28	2	14	1	0	1
1	0	0	28	2	4	0	0	1
1	0	1	27	2	0	0	0	1
1	0	0	35	2	5	1	0	1
0	1	1	25	2	84	1	1	1
1	0	0	28	2	0	1	0	1
1	0	1	29	2	4	1	0	1
1	0	0	28	2	7	1	0	1
0	0	0	34	2	2	0	0	1
0	1	0	29	2	11	0	0	1
0	1	0	29	2	7	1	0	1
0	0	0	29	2	5	1	0	1
0	0	1	28	2	18	1	1	1
0	0	0	31	2	1	0	0	1
1	0	1	29	2	4	0	0	1
0	0	0	36	2	0	0	0	1
0	0	0	30	2	7	0	0	1
0	0	0	30	2	0	0	0	1
1	0	0	31	2	40	0	0	1
1	0	0	29	2	3	0	0	2
1	0	1	30	2	34	0	0	1
1	0	0	30	2	2	1	0	1
0	0	0	29	2	57	0	1	1
1	0	0	30	2	0	0	0	2
1	1	1	27	2	7	0	0	1
1	0	0	28	2	2	1	0	1
0	0	0	29	2	3	1	0	1
0	1	0	34	2	0	0	0	1
1	0	1	28	2	7	1	0	1

X_1	X_2	X_3	X_4	X_5	X_6	X_7	N_1	N_0
0	0	1	32	2	13	1	1	1
0	0	0	31	2	2	1	0	1
1	0	1	28	2	39	1	0	1
0	0	0	36	2	0	1	0	1
1	0	1	30	2	5	0	0	1
1	0	0	29	2	11	1	0	1
0	0	0	30	2	5	1	1	1
1	0	1	27	2	18	1	0	1
0	1	0	27	2	6	1	0	1
1	0	0	32	2	23	1	0	1
0	1	1	27	2	3	1	0	1
1	0	1	25	2	26	1	0	1
0	1	1	27	2	1	0	0	1
0	0	1	27	2	39	0	0	1
0	0	1	27	2	22	0	1	1