ESTIMAÇÃO PONTUAL E INTERVALAR NO MODELO LOGÍSTICO HIERÁRQUICO DE DOIS NÍVEIS

Artur José LEMONTE[1]
Tatiane Ferreira Nascimento Melo da SILVA[2]

■ RESUMO: Apresentamos neste artigo um estudo dos estimadores dos parâmetros no modelo logístico hierárquico de dois níveis. Consideramos estimadores baseados na quasi-verossimilhança marginal de primeira ordem, cujo algoritmo de estimação pode ser encontrado em Goldstein (1991). Avaliamos o desempenho destes estimadores pela estimação pontual e intervalar. Adicionalmente, apresentamos taxas de cobertura dos intervalos de confiança e poder dos testes de significância. O estudo foi feito por meio de simulação de Monte Carlo. Finalmente, uma aplicação a dados reais é apresentada e discutida.

■ PALAVRAS-CHAVE: Efeitos aleatórios; estimação intervalar; estimação pontual; modelos hierárquicos; modelo logístico; simulação de Monte Carlo.

1 Introdução

Pesquisadores em diversas áreas de aplicação analisam conjuntos de dados coletados especificamente para responder as perguntas que motivaram sua investigação. Comumente, tais dados apresentam uma estrutura hierárquica, que é caracterizada pela presença de unidades experimentais agrupadas em outras unidades experimentais ainda maiores, que por sua vez também podem ou não formar novos grupos. Dados com esta estrutura são encontrados, por exemplo, em investigações sociais cujo objetivo principal é estudar o efeito do contexto social sob o comportamento individual, assim como explicar as diferenças entre estes efeitos para diferentes contextos sociais.

[1]Departamento de Estatística, Universidade Federal de Pernambuco - UFPE, CEP 50740-540, Recife, PE, Brasil. E-mail: *artur@cox.de.ufpe.br*
[2]Instituto de Matemática e Estatística, Universidade de São Paulo - USP, CEP 05508-090, São Paulo, SP, Brasil. E-mail: *tafename@ime.usp.br*

Freqüentemente, a estrutura hierárquica é uma propriedade intrínseca da população de interesse. Porém, muitas vezes a abordagem estatística aplicada para a análise dos dados não leva em conta tal estrutura, obtendo como conseqüência conclusões erradas, isto é, concluir a favor da significância estatística de um parâmetro quando na verdade este não é significativo. Isto pode acontecer porque indivíduos pertencentes a um mesmo grupo tendem a ser mais "parecidos" que indivíduos de diferentes grupos, assim, suposições clássicas de independência não são válidas. Em conseqüência, métodos estatísticos de modelagem que ignoram a estrutura hierárquica dos dados vão, possivelmente, superestimar a variância dos estimadores dos parâmetros.

Entre os estudos mais comuns em que os dados da análise estatística apresentam uma estrutura hierárquica estão os de comparação institucional, os quais têm como objetivo principal estudar uma população de instituições em relação a uma medida de desempenho previamente especificada, para, posteriormente, explicar as diferenças entre os desempenhos das instituições por seus atributos e características. De modo geral, a medida de desempenho pode ser considerada como a medida resumo mais apropriada para os dados em estudo, sendo a proporção e a média, as medidas mais apropriadas nos casos dicotômico e contínuo, respectivamente. Por outro lado, estudos desta classe são encontrados, por exemplo, na área médica, em que o interesse é avaliar o desempenho de Unidades de Tratamento Intensivo (UTI's) em relação as taxas de mortalidade e explicar as diferenças no desempenho pelos atributos das UTI's (isto é, estrutura física, técnica, médica, fisioterapêutica etc.). Adicionalmente, na área financeira pode-se ter interesse em comparar o desempenho das agências bancárias em relação as taxas de clientes inadimplentes e, em seguida, explicar as diferenças entre as taxas pelas características e atributos das agências (isto é, estrutura técnica, atendimento etc.). Evidentemente, nestes estudos os dados não podem ser analisados usando métodos padrão de regressão como, por exemplo, Modelos Lineares Generalizados (McCullagh e Nelder, 1989), pois estes ignoram a natureza hierárquica dos dados.

Uma outra estratégia de análise consiste em resumir a informação da instituição numa única medida (a medida de desempenho selecionada) e tentar explicar as diferenças entre as instituições em relação a esta medida usando suas características e seus atributos. Por exemplo, se a medida de desempenho é uma proporção, isto pode ser feito usando regressão beta (Ferrari e Cribari Neto, 2004). Entretanto, este método de análise assume que os indivíduos pertencentes a cada uma das instituições são homogêneos e diretamente comparáveis, isto é, que as diferenças entre as medidas de desempenho são atribuídas somente a diferenças nos atributos e características das instituições, o qual não é válido nestes estudos, pois, por exemplo, no caso do estudo das UTI's, diferentes indivíduos com diferentes doenças chegam a diferentes UTI's, impossibilitando assim uma comparação direta das medidas de desempenho.

Os modelos hierárquicos proporcionam um marco teórico e conceitual para o tratamento e a análise estatística de dados com estrutura hierárquica, sendo uma nova formulação para os modelos de efeitos aleatórios. Os modelos hierárquicos além de levar em conta a estrutura hierárquica dos dados, permitem incorporar aspectos importantes do procedimento amostral pelo qual foram obtidos os dados, ou seja, permitem incluir pesos diferentes às observações na etapa do ajuste do modelo, os quais dependem do planejamento amostral por meio das probabilidades de inclusão das observações na amostra. Modelos hierárquicos, em geral, são estimados pelos métodos de máxima verossimilhança. Entretanto, quando a variável resposta é dicotômica, estudos de simulação apresentados em Breslow e Clayton (1993) e em Rodríguez e Goldman (1995), mostraram que alguns algoritmos de estimação utilizados na prática não são adequados para estimar os parâmetros do modelo. Detalhes sobre algoritmos de estimação em modelos hierárquicos com resposta dicotômica podem ser encontrados em Rodríguez e Goldman (1995); veja também Goldstein e Rasbash (1996).

Este artigo tem como objetivo avaliar, por meio de simulações de Monte Carlo, o desempenho dos estimadores do modelo hierárquico com dois níveis para resposta binária, especificamente o modelo com função de ligação logit. Para estimar os parâmetros do modelo, consideramos o algoritmo de estimação baseado na quasi-verossimilhança marginal de primeira ordem, proposto por Goldstein (1991), pois este algoritmo vem sendo amplamente explorado em aplicações práticas.

O artigo está organizado da seguinte forma: na Seção 2, apresentamos alguns conceitos sobre os modelos hierárquicos de resposta binária e expomos as principais características da estimação pontual e intervalar dos parâmetros. Na Seção 3, explicamos em mais detalhes a metodologia usada para conduzir os experimentos de simulação e os resultados numéricos. Na Seção 4 apresentamos uma aplicação a dados reais. Finalmente, na Seção 5 descrevemos as conclusões.

2 Modelos hierárquicos de resposta binária

Considere uma população de interesse com estrutura hierárquica de dois níveis, que pode ser descrita pela presença de indivíduos ou unidades experimentais de primeiro nível, agrupadas em outras unidades ainda maiores, chamadas grupos ou unidades de segundo nível. Geralmente, o estudo de populações com tais características é feito selecionando uma amostra de J unidades de segundo nível para, posteriormente, selecionar uma amostra de n_j indivíduos em cada grupo selecionado inicialmente. Para o i-ésimo indivíduo no j-ésimo grupo, observamos a resposta dicotômica Y_{ij}, que pode tomar somente um de dois valores possíveis, denotados por conveniência 1 e 0 {1: sucesso; 0: fracasso}. Assume-se que, condicionalmente a um vetor de efeitos aleatórios $\boldsymbol{\tau}$, Y_{ij} sejam variáveis aleatórias independentes seguindo a distribuição *Bernoulli* (π_{ij}), $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, em que $\pi_{ij} = P(Y_{ij} = 1)$ é a probabilidade de sucesso para o i-ésimo indivíduo no j-ésimo grupo. Entretanto, para explicar e/ou predizer o comportamento de Y_{ij}, são observadas um conjunto de K variáveis sobre todas as unidades do primeiro nível, as quais estão relacionadas à probabilidade de sucesso pela relação

$$f(\pi_{ij}) = \mathbf{X}_{ij}^{\top}\boldsymbol{\beta}_j, \tag{1}$$

em que $\mathbf{X}_{ij}^{\top}$ representa uma matriz de dimensão $n_j \times K$ contendo as K medições para as variáveis regressoras nos n_j indivíduos do j-ésimo grupo, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})^{\top}$ é um vetor de parâmetros representando os efeitos fixos e aleatórios do modelo no j-ésimo grupo e $f(\cdot)$ é uma função denomida *função de ligação*, cuja escolha depende do fenômeno em estudo e da relação entre as probabilidades de sucesso e a informação na matriz $\mathbf{X}_{ij}^{\top}$.

No segundo nível, a variabilidade dos J coeficientes $\beta_{1k}, \ldots, \beta_{Jk}$ pertencendo à k-ésima variável ($k = 1, \ldots, K$) é explicada pela variabilidade num conjunto de variáveis medidas no nível dos grupos, também conhecidas como variáveis macro. Esta relação pode ser descrita como

$$\beta_{jk} = \mathbf{G}_{jk}\boldsymbol{\theta}_k + \tau_{jk}, \quad \tau_{jk} \sim \mathcal{N}(0, \sigma_{kk}), \tag{2}$$

em que $\mathbf{G}_{jk}$ representa as observações sobre l_k variáveis macro e $\boldsymbol{\theta}_k = (\theta_{k1}, \ldots, \theta_{k\,l_k})^{\top}$ é o vetor de coeficientes. Tem-se que para $k \neq k'$,

$$\operatorname{cov}(\tau_{jk}, \tau_{j'k'}) = \begin{cases} \sigma_{kk'}, & \text{se} \quad j = j', \\ 0, & \text{caso contrário.} \end{cases}$$

Pode-se escrever (2) da forma

$$\boldsymbol{\beta}_j = \mathbf{G}_j\boldsymbol{\theta} + \boldsymbol{\tau}_j, \quad \boldsymbol{\tau} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{3}$$

em que $\mathbf{G}_j = \operatorname{diag}\left\{\mathbf{G}_{j1}^{\top}, \ldots, \mathbf{G}_{jK}^{\top}\right\}$ é uma matriz bloco diagonal de dimensão $K \times L$, $L = l_1 + \cdots + l_K$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)^{\top}$ é um vetor $L \times 1$ de coeficientes macro e $\boldsymbol{\Sigma}$ é a matriz de variâncias e covariâncias de $\boldsymbol{\tau}$ com elementos $\sigma_{kk'}$.

Estimação pontual e intervalar dos parâmetros

Estimação em modelos hierárquicos de resposta binária, em geral, é feita por métodos de máxima verossimilhança, isto é, o processo de estimação consiste em maximizar a função de verossimilhança em relação aos parâmetros do modelo. A função de verossimilhança condicional pode ser expressa como

$$L(\boldsymbol{\theta}|\boldsymbol{\tau}) = \prod_{j=1}^{J}\prod_{i=1}^{n_j}(\pi_{ij})^{y_{ij}}(1 - \pi_{ij})^{1-y_{ij}}. \tag{4}$$

Para obter a verossimilhança não-condicional, devemos multiplicar a equação (4) por $\phi(\boldsymbol{\tau})$ e integrar a função resultante em relação a $\boldsymbol{\tau}$, isto é,

$$L(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \int \cdots \int L(\boldsymbol{\theta}|\boldsymbol{\tau})\phi(\boldsymbol{\tau})\mathrm{d}\boldsymbol{\tau}, \tag{5}$$

em que $\phi(\cdot)$ representa a função densidade de $\boldsymbol{\tau}$. Porém, a expressão (5) é muito complicada e, em geral, não apresenta forma "fechada". Assim, há necessidade da aplicação de algoritmos de integração numérica (por exemplo, quadratura gaussiana), os quais proporcionam resultados satisfatórios somente para modelos com menos de três níveis (Breslow e Clayton, 1993). Dessa forma, muitos pesquisadores preferem trabalhar com soluções aproximadas (isto é, quasi-verossimilhança), que é o resultado da aproximação da função de verossimilhança usando os primeiros termos de uma expansão em série de Taylor.

A função de verossimilhança aproximada ou quasi-verossimilhança, como é chamada comumente, pode ser usada num processo iterativo para encontrar as estimativas de máxima verossimilhança, começando com valores aproximados e incrementando sua exatidão a cada iteração.

A inferência nestes modelos (intervalos de confiança e testes de hipóteses) está baseada em propriedades assintóticas dos estimadores de máxima verossimilhança, ou seja, supondo que o número de grupos e o número de indivíduos por grupo sejam suficientemente grandes. A estatística de teste Wald para o p-ésimo elemento do vetor de parâmetros $\boldsymbol{\theta}$ avaliada no ponto θ^* é definida como

$$W_p(\theta^*) = \frac{(\widehat{\theta}_p - \theta^*)^2}{\widehat{\text{var}}(\widehat{\theta}_p)},$$

em que $\widehat{\theta}_p$ e $\widehat{\text{var}}(\widehat{\theta}_p)$ são os valores estimados de θ_p e da variância do estimador de θ_p, respectivamente. Assim, o intervalo de confiança para θ_p utilizando a estatística W_p é dado por

$$\text{IC}[\theta_p; 1-\alpha] = \{\theta \mid W_p(\theta) < q_{(1-\alpha)}\}, \tag{6}$$

em que $q_{(1-\alpha)}$ representa o percentil $100(1-\alpha)\%$ da distribuição qui-quadrado com um grau de liberdade. Entretanto, é possível testar as hipóteses $\mathcal{H}_0\colon \theta_p = \theta_p^{(0)}$ *versus* $\mathcal{H}_1\colon \theta_p \neq \theta_p^{(0)}$, utilizando o intervalo de confiança (6), isto é, verificando se $\text{IC}[\theta_p; 1-\alpha]$ contém $\theta_p^{(0)}$. Então, um teste de hipóteses para θ_p com nível de significância α pode ser descrito da seguinte forma: rejeitar $\mathcal{H}_0$ se $\theta_p^{(0)} \notin \text{IC}[\theta_p, 1-\alpha]$ ou, equivalentemente, rejeitar $\mathcal{H}_0$ se $W_p\big(\theta_p^{(0)}\big) > q_{(1-\alpha)}$. Particularmente, para testar a significância de θ_p no modelo, a estatística de teste Wald é representada por $W_p(0)$.

3 Experimentos de Monte Carlo

Com o objetivo de estudar o desempenho dos estimadores dos parâmetros do modelo logístico hierárquico de dois níveis, apresentado na seção anterior, um número de experimentos de Monte Carlo foi conduzido. Os resultados numéricos apresentados neste artigo correspondem ao modelo

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_1 + \beta_{2j}x_{ij},$$

$i = 1, \ldots, n_j$, $j = 1, \ldots, J$, sendo

$$\beta_1 = \theta_{11}, \quad \beta_{2j} = \theta_{21} + \theta_{22}w_j + \tau_j, \quad \tau_j \sim \mathcal{N}(0, \sigma^2), \tag{7}$$

em que x_{ij} é uma variável medida no nível dos indivíduos e w_j é uma variável medida no nível dos grupos. Este modelo foi considerado em diferentes cenários de simulação os quais são definidos pelas seguintes características: (i) Número de grupos ou unidades de segundo nível; (ii) Número de unidades de primeiro nível amostrados em cada unidade de segundo nível; (iii) Valores para $\theta_{11}, \theta_{21}, \theta_{22}$ e σ^2. Utilizamos nos estudos de simulação $r = 5.000$ (número de réplicas de Monte Carlo).

Consideramos $J = 20, 25$ e 30, $\theta_{11} = 0.3$, $\theta_{21} = -0.5$ e -0.8, $\theta_{22} = 0.5$ e 0.8, e $\sigma^2 = 1$. Nas Tabelas 1 a 6 apresentamos a média das 5.000 estimativas, o viés relativo (o viés relativo de um estimador $\widehat{\theta}$ de um parâmetro escalar θ é definido como $\{\mathrm{E}(\widehat{\theta}) - \theta\}/\theta$, e o viés relativo estimado é obtido estimando $\mathrm{E}(\widehat{\theta})$ via Monte Carlo) e o desvio padrão das 5.000 estimativas. Nas Tabelas 7 a 12 apresentamos as taxas de cobertura (TC) dos intervalos de confiança para os parâmetros do modelo, considerando os níveis de cobertura 90%, 95% e 99%. Além disso, apresentamos o poder empírico do teste de significância (PTS) considerando níveis de significância 1%, 5% e 10% . Estas medidas foram estimadas da seguinte forma

$$\mathrm{TC} = \frac{\#\{\theta_p \in \mathrm{IC}[\theta_p; 1-\alpha]\}}{r},$$

$$\mathrm{PTS} = 1 - \frac{\#\{0 \in \mathrm{IC}[\theta_p; 1-\alpha]\}}{r},$$

em que r é o número de réplicas de Monte Carlo. O esquema de geração dos dados foi assim conduzido:

1. gerar x_{ij}, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, seguindo distribuição $\mathcal{N}(0, 1)$. Estes valores permanecem fixos durante todo o processo de simulação;

2. gerar w_j, $j = 1, \ldots, J$, seguindo distribuição $\mathcal{N}(0, 1)$. Estes valores permanecem fixos durante todo o processo de simulação;

3. gerar τ_j, $j = 1, \ldots, J$, seguindo distribuição $\mathcal{N}(0, \sigma^2)$;

4. calcular a probabilidade de sucesso $\pi_{ij} = f^{-1}(\beta_1 + \beta_{2j}x_{ij})$, com β_1 e β_{2j} descritos em (7) e f a função de ligação logit;

5. gerar u_{ij}, $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, seguindo distribuição $\mathcal{U}(0, 1)$;

6. calcular os valores de Y_{ij} da forma

$$Y_{ij} = \begin{cases} 1, & \text{se } u_{ij} \leq \pi_{ij}, \\ 0, & \text{se } u_{ij} > \pi_{ij}. \end{cases}$$

No processo de simulação de Monte Carlo, os passos 3 a 6 foram repetidos r vezes. Todas as simulações foram feitas utilizando a linguagem matricial de programação `Ox` (Doornik, 2001) em um computador Athlon XP 1800. Mais informações sobre esta linguagem de programação podem ser encontradas em *http://www.doornik.com.*

Resultados numéricos

Note nas Tabelas 1 a 6 que em todos os casos, os vieses relativos estimados são negativos, indicando que a força de associação entre a variável resposta e as variáveis explicativas está sendo subestimada pelo algoritmo de estimação considerado. Observe que a variância σ^2 também é subestimada pelo algoritmo. Note também que o desvio padrão do estimador $\widehat{\theta}_{11}$ de θ_{11} é consideralmente menor que os desvios padrão de todos os outros estimadores. Note ainda que os devios padrão dos estimadores $\widehat{\theta}_{21}$ e $\widehat{\theta}_{22}$ de θ_{21} e θ_{22}, respectivamente, são similares. Observe que, na maioria dos casos, à medida que o número de grupos aumenta, os vieses relativos diminuem (em valor absoluto). Da mesma forma, quando a diferença entre o número de indivíduos amostrados por grupo diminuem, os vieses relativos decrescem (em módulo).

Observe nas Tabelas 7 a 12 que para os parâmetros θ_{21} e θ_{22} do modelo, as taxas de cobertura dos intervalos de confiança estão bem próximas dos níveis de cobertura, ou seja, a estimação intervalar destes parâmetros apresenta bom desempenho. Já para o parâmetro θ_{11}, nota-se que as taxas de cobertura foram consideravelmente maiores que os níveis de cobertura considerados. É interessante observar que analogamente a estimação pontual, o desempenho da estimação intervalar melhora (na maioria dos casos) com o aumento do número de grupos, ou seja, as taxas de cobertura aproximam-se dos verdadeiros níveis de cobertura considerados. Isto também foi observado quando a diferença entre o número de indivíduos amostrados por grupo diminuem. Note ainda que o poder do teste de significância para os parâmetros do modelo aumenta com o número de grupos. Além disso, o poder também aumenta com a força da associação entre a variável resposta e explicativa, isto é, o poder cresce quando os valores (em módulo) de θ_{21} e θ_{22} aumentam.

De acordo com estes resultados, percebe-se que o desempenho dos estimadores dos parâmetros estão diretamente relacionados com o número de grupos. Isto acontece pois, assintoticamente (isto é, número de grupos e número de indivíduos por grupo suficientemente grande) estes estimadores são não viesados.

Em situações práticas, o que se observa é grupos muito pequenos na amostra. Uma possível solução para este problema seria fazer algum tipo de refinamento assintótico (por exemplo, correção de viés; veja Cox e Snell, 1968) sobre estes estimadores, o que melhoraria seu desempenho, quando os grupos fossem pequenos.

Tabela 1 - Estimativas dos parâmetros. Número total de grupos: 20; $\theta_{11} = 0,3$; $\theta_{21} = -0,5$; $\theta_{22} = 0,8$; $\sigma^2 = 1$

	Parâmetro	Média	Viés relativo	Desvio padrão
	θ_{11}	$0,2456$	$-0,1813$	0,0424
Mín.$(n_j) = 30$	θ_{21}	$-0,4534$	$-0,0932$	0,2383
Máx.$(n_j) = 70$	θ_{22}	$0,7724$	$-0,0345$	0,2594
	σ^2	$0,8171$	$-0,1829$	0,4239
	θ_{11}	$0,2392$	$-0,2027$	0,0412
Mín.$(n_j) = 30$	θ_{21}	$-0,4563$	$-0,0874$	0,2375
Máx.$(n_j) = 60$	θ_{22}	$0,7763$	$-0,0296$	0,2577
	σ^2	$0,7981$	$-0,2019$	0,6690
	θ_{11}	$0,2243$	$-0,2523$	0,0412
Mín.$(n_j) = 50$	θ_{21}	$-0,4588$	$-0,0824$	0,2404
Máx.$(n_j) = 50$	θ_{22}	$0,7897$	$-0,0129$	0,2627
	σ^2	$0,8562$	$-0,1438$	0,5635

Tabela 2 - Estimativas dos parâmetros. Número total de grupos: 25; $\theta_{11} = 0,3$; $\theta_{21} = -0,5$; $\theta_{22} = 0,8$; $\sigma^2 = 1$

	Parâmetro	Média	Viés relativo	Desvio padrão
	θ_{11}	$0,2346$	$-0,2180$	0,0361
Mín.$(n_j) = 30$	θ_{21}	$-0,4622$	$-0,0756$	0,2166
Máx.$(n_j) = 70$	θ_{22}	$0,7825$	$-0,0219$	0,2332
	σ^2	$0,8694$	$-0,1306$	0,5612
	θ_{11}	$0,2373$	$-0,2090$	0,0346
Mín.$(n_j) = 30$	θ_{21}	$-0,4836$	$-0,0328$	0,2163
Máx.$(n_j) = 60$	θ_{22}	$0,7877$	$-0,0154$	0,2415
	σ^2	$0,8648$	$-0,1352$	0,4488
	θ_{11}	$0,2320$	$-0,2267$	0,0374
Mín.$(n_j) = 50$	θ_{21}	$-0,4824$	$-0,0352$	0,2149
Máx.$(n_j) = 50$	θ_{22}	$0,7874$	$-0,0158$	0,2296
	σ^2	$0,8735$	$-0,1265$	0,4441

Tabela 3 - Estimativas dos parâmetros. Número total de grupos: 30; $\theta_{11} = 0,3$; $\theta_{21} = -0,5$; $\theta_{22} = 0,8$; $\sigma^2 = 1$

	Parâmetro	Média	Viés relativo	Desvio padrão
	θ_{11}	$0,2564$	$-0,1453$	0,0346
Mín.$(n_j) = 30$	θ_{21}	$-0,4490$	$-0,1020$	0,1916
Máx.$(n_j) = 70$	θ_{22}	$0,7765$	$-0,0294$	0,2093
	σ^2	$0,8825$	$-0,1175$	0,4353
	θ_{11}	$0,2604$	$-0,1320$	0,0346
Mín.$(n_j) = 30$	θ_{21}	$-0,4780$	$-0,0440$	0,1934
Máx.$(n_j) = 60$	θ_{22}	$0,7824$	$-0,0220$	0,2088
	σ^2	$0,9234$	$-0,0766$	0,3245
	θ_{11}	$0,2661$	$-0,1130$	0,0332
Mín.$(n_j) = 50$	θ_{21}	$-0,4770$	$-0,0460$	0,1887
Máx.$(n_j) = 50$	θ_{22}	$0,7768$	$-0,0290$	0,2049
	σ^2	$0,8761$	$-0,1239$	0,3167

Tabela 4 - Estimativas dos parâmetros. Número total de grupos: 20; $\theta_{11} = 0,3$; $\theta_{21} = -0,8$; $\theta_{22} = 0,5$; $\sigma^2 = 1$

	Parâmetro	Média	Viés relativo	Desvio padrão
	θ_{11}	$0,2479$	$-0,1737$	0,0424
Mín.$(n_j) = 30$	θ_{21}	$-0,7402$	$-0,0748$	0,2385
Máx.$(n_j) = 70$	θ_{22}	$0,4827$	$-0,0346$	0,2553
	σ^2	$0,8144$	$-0,1856$	0,6713
	θ_{11}	$0,2408$	$-0,1973$	0,0412
Mín.$(n_j) = 30$	θ_{21}	$-0,7455$	$-0,0681$	0,2398
Máx.$(n_j) = 60$	θ_{22}	$0,4869$	$-0,5131$	0,2592
	σ^2	$0,8141$	$-0,1859$	0,6115
	θ_{11}	$0,2264$	$-0,2453$	0,0412
Mín.$(n_j) = 50$	θ_{21}	$-0,7534$	$-0,0583$	0,2425
Máx.$(n_j) = 50$	θ_{22}	$0,4954$	$-0,0092$	0,2642
	σ^2	$0,8704$	$-0,1296$	0,5162

Tabela 5 - Estimativas dos parâmetros. Número total de grupos: 25; $\theta_{11} = 0,3$; $\theta_{21} = -0,8$; $\theta_{22} = 0,5$; $\sigma^2 = 1$

	Parâmetro	Média	Viés relativo	Desvio padrão
	θ_{11}	$0,2379$	$-0,2070$	0,0346
Mín.$(n_j) = 30$	θ_{21}	$-0,7532$	$-0,0585$	0,2152
Máx.$(n_j) = 70$	θ_{22}	$0,4906$	$-0,0188$	0,2300
	σ^2	$0,8799$	$-0,1201$	0,3365
	θ_{11}	$0,2412$	$-0,1930$	0,0346
Mín.$(n_j) = 30$	θ_{21}	$-0,7763$	$-0,0296$	0,2173
Máx.$(n_j) = 60$	θ_{22}	$0,4939$	$-0,0122$	0,2287
	σ^2	$0,8748$	$-0,1252$	0,5449
	θ_{11}	$0,2350$	$-0,2167$	0,0361
Mín.$(n_j) = 50$	θ_{21}	$-0,7770$	$-0,0288$	0,2154
Máx.$(n_j) = 50$	θ_{22}	$0,4910$	$-0,0180$	0,2287
	σ^2	$0,8740$	$-0,1260$	0,5646

Tabela 6 - Estimativas dos parâmetros. Número total de grupos: 30; $\theta_{11} = 0,3$; $\theta_{21} = -0,8$; $\theta_{22} = 0,5$; $\sigma^2 = 1$

	Parâmetro	Média	Viés relativo	Desvio padrão
	θ_{11}	$0,2570$	$-0,1433$	0,0332
Mín.$(n_j) = 30$	θ_{21}	$-0,7354$	$-0,0808$	0,2069
Máx.$(n_j) = 70$	θ_{22}	$0,4876$	$-0,0248$	0,2076
	σ^2	$0,8821$	$-0,1179$	0,5307
	θ_{11}	$0,2608$	$-0,1307$	0,0346
Mín.$(n_j) = 30$	θ_{21}	$-0,7689$	$-0,0389$	0,1931
Máx.$(n_j) = 60$	θ_{22}	$0,4885$	$-0,0230$	0,2064
	σ^2	$0,9259$	$-0,0741$	0,4498
	θ_{11}	$0,2666$	$-0,1113$	0,0332
Mín.$(n_j) = 50$	θ_{21}	$-0,7652$	$-0,0435$	0,1947
Máx.$(n_j) = 50$	θ_{22}	$0,4853$	$-0,0294$	0,2030
	σ^2	$0,8687$	$-0,1313$	0,5104

Tabela 7 - Taxas de cobertura e poder empírico. Número total de grupos: 20; $\theta_{11} = 0,3$; $\theta_{21} = -0,5$; $\theta_{22} = 0,8$; $\sigma^2 = 1$

		TC			PTS		
		90%	95%	99%	10%	5%	1%
Mín.$(n_j) = 30$	θ_{11}	95,63	99,02	99,98	99,88	99,24	83,21
Máx.$(n_j) = 70$	θ_{21}	87,77	93,23	98,56	60,23	46,64	22,57
	θ_{22}	88,00	93,59	98,24	91,10	84,85	64,93
Mín.$(n_j) = 40$	θ_{11}	93,47	98,66	99,94	99,78	98,86	81,02
Máx.$(n_j) = 50$	θ_{21}	87,47	93,32	98,62	61,12	47,11	23,60
	θ_{22}	88,34	94,06	98,72	91,61	85,89	65,80
Mín.$(n_j) = 50$	θ_{11}	86,51	95,77	99,92	99,28	96,97	70,64
Máx.$(n_j) = 50$	θ_{21}	87,88	93,96	98,76	60,45	45,83	22,43
	θ_{22}	88,52	94,00	98,50	91,15	85,31	65,11

Tabela 8 - Taxas de cobertura e poder empírico. Número total de grupos: 25; $\theta_{11} = 0,3$; $\theta_{21} = -0,5$; $\theta_{22} = 0,8$; $\sigma^2 = 1$

		TC			PTS		
		90%	95%	99%	10%	5%	1%
Mín.$(n_j) = 26$	θ_{11}	88,42	96,86	99,88	99,96	99,76	94,66
Máx.$(n_j) = 74$	θ_{21}	88,27	93,91	98,64	68,75	56,30	30,81
	θ_{22}	88,99	94,29	98,96	95,52	91,51	76,40
Mín.$(n_j) = 38$	θ_{11}	90,22	97,64	99,88	99,98	99,76	95,92
Máx.$(n_j) = 62$	θ_{21}	88,45	94,06	98,78	73,02	61,30	33,66
	θ_{22}	88,82	94,54	98,78	95,94	92,38	77,95
Mín.$(n_j) = 50$	θ_{11}	86,83	96,08	99,82	99,96	99,74	93,54
Máx.$(n_j) = 50$	θ_{21}	89,09	94,29	99,04	71,95	59,88	33,77
	θ_{22}	89,21	94,93	99,06	95,72	92,21	77,20

Tabela 9 - Taxas de cobertura e poder empírico. Número total de grupos: 30; $\theta_{11} = 0,3$; $\theta_{21} = -0,5$; $\theta_{22} = 0,8$; $\sigma^2 = 1$

		TC			PTS		
		90%	95%	99%	10%	5%	1%
Mín.$(n_j) = 20$	θ_{11}	94,88	98,68	99,94	99,96	99,94	99,72
Máx.$(n_j) = 58$	θ_{21}	88,51	93,86	98,66	74,50	62,42	35,02
	θ_{22}	89,13	94,70	98,86	97,64	95,50	85,51
Mín.$(n_j) = 35$	θ_{11}	95,62	98,86	100,0	100,0	100,0	99,82
Máx.$(n_j) = 64$	θ_{21}	89,68	95,06	99,14	77,84	66,69	39,19
	θ_{22}	89,74	94,66	98,86	97,62	95,42	85,34
Mín.$(n_j) = 50$	θ_{11}	96,84	99,42	99,96	100,0	99,98	99,90
Máx.$(n_j) = 50$	θ_{21}	89,17	94,65	99,28	79,26	69,16	41,37
	θ_{22}	89,13	94,05	98,52	97,92	96,08	86,06

Tabela 10 - Taxas de cobertura e poder empírico. Número total de grupos: 20; $\theta_{11} = 0,3$; $\theta_{21} = -0,8$; $\theta_{22} = 0,5$; $\sigma^2 = 1$

		TC			PTS		
		90%	95%	99%	10%	5%	1%
Mín.$(n_j) = 30$	θ_{11}	95,70	99,16	99,96	99,88	99,20	84,75
Máx.$(n_j) = 70$	θ_{21}	87,23	93,04	98,48	92,58	86,65	64,55
	θ_{22}	88,59	93,52	98,52	63,83	49,82	24,32
Mín.$(n_j) = 26$	θ_{11}	94,31	98,84	99,94	99,76	99,00	81,94
Máx.$(n_j) = 74$	θ_{21}	87,35	92,92	98,28	93,14	87,05	65,18
	θ_{22}	88,03	93,22	98,70	63,89	50,98	25,70
Mín.$(n_j) = 38$	θ_{11}	88,50	96,41	99,90	99,40	97,02	71,88
Máx.$(n_j) = 62$	θ_{21}	87,81	93,64	98,62	92,54	85,88	63,94
	θ_{22}	88,03	93,48	98,50	63,52	49,98	24,53

Tabela 11 - Taxas de cobertura e poder empírico. Número total de grupos: 25; $\theta_{11} = 0,3$; $\theta_{21} = -0,8$; $\theta_{22} = 0,5$; $\sigma^2 = 1$

		TC			PTS		
		90%	95%	99%	10%	5%	1%
Mín.$(n_j) = 26$	θ_{11}	91,27	97,48	99,98	100,00	99,94	95,94
Máx.$(n_j) = 74$	θ_{21}	88,23	94,16	98,66	96,56	93,03	77,96
	θ_{22}	88,83	94,06	98,96	71,22	59,05	33,85
Mín.$(n_j) = 38$	θ_{11}	92,56	98,06	100,0	99,98	99,94	96,66
Máx.$(n_j) = 62$	θ_{21}	88,58	93,94	98,78	97,44	94,60	81,31
	θ_{22}	88,72	94,56	98,88	72,35	60,46	33,77
Mín.$(n_j) = 20$	θ_{11}	88,43	96,60	99,92	99,98	99,84	94,24
Máx.$(n_j) = 78$	θ_{21}	88,95	94,11	98,86	97,48	94,27	80,88
	θ_{22}	88,69	94,31	98,96	71,51	59,62	33,77

Tabela 12 - Taxas de cobertura e poder empírico. Número total de grupos: 30; $\theta_{11} = 0,3$; $\theta_{21} = -0,8$; $\theta_{22} = 0,5$; $\sigma^2 = 1$

		TC			PTS		
		90%	95%	99%	10%	5%	1%
Mín.$(n_j) = 20$	θ_{11}	95,92	99,28	99,96	99,96	99,96	99,80
Máx.$(n_j) = 78$	θ_{21}	87,78	93,16	98,62	98,46	96,22	85,79
	θ_{22}	89,17	94,64	98,96	77,27	66,98	41,62
Mín.$(n_j) = 26$	θ_{11}	95,96	99,12	99,98	100,0	100,0	99,84
Máx.$(n_j) = 74$	θ_{21}	89,20	94,44	99,02	98,98	97,10	88,16
	θ_{22}	89,52	94,94	98,88	77,69	66,57	40,42
Mín.$(n_j) = 38$	θ_{11}	97,26	99,38	99,96	99,98	99,96	99,94
Máx.$(n_j) = 62$	θ_{21}	88,95	94,63	98,98	99,14	97,68	89,23
	θ_{22}	89,05	94,03	98,60	77,87	68,06	42,99

4 Aplicação a dados reais

Nesta seção aplicaremos a metodologia apresentada anteriormente a um conjunto de dados reais. Utilizamos os dados da Pequisa Nacional sobre Demografia e Saúde, realizada em 1996, pela Sociedade Civil de Bem-Estar Familiar no Brasil (BEMFAM). Esta pesquisa faz parte do Programa Internacional de Demografia e Saúde (*Demographic and Health Survey – DHS*). O objetivo da análise visa avaliar o impacto de fatores que podem estar associados à desnutrição entre crianças brasileiras no segundo ano de vida. A amostra analisada foi composta de 810 crianças, contendo 142 casos de crianças que estavam desnutridas. As variáveis utilizadas para explicar as variações na desnutrição foram: Região (Norte, Nordeste, Sudeste, Sul e Centro-Oeste); Idade materna ($\leq$ 19 anos, 20–34 e 35 ou mais); Ordem de nascimento (Primeiro filho, 2–3 filhos e 4 ou mais); Intervalo prévio ($\leq$ 18 meses e $>$ 18 meses); Educação materna (0–3 anos, 4–8 anos e 9 ou mais); Percepção do peso da criança (Grande/muito grande, médio e Pequeno/muito pequeno). Uma descrição completa e mais detalhada sobre estas variáveis e o porquê delas estarem associadas à desnutrição está dado em Cruz e Leite (2002).

Das Grupta (1990), analisando informações demográficas da área rural de Punjab (Índia), observou que em 12,6% das famílias ocorreram 62,2% das mortes na infância. Este resultado fornece forte evidência de que as mortes se concentram em determinadas famílias, formando o que tem sido denominado de "conglomerados de morte" (*death cluster*). Conglomerados com elevada prevalência de desnutrição infantil foram observados por Griffiths (1999) na Índia. Esses estudos indicam a presença de estrutura hierárquica, que implica a correlação entre observações em cada família. Neste caso, as observações não são independentes, pressuposto assumido pelos tradicionais modelos de regressão.

Dessa forma, como a variável resposta desnutrição é dicotômica {1: desnutrido; 0: não desnutrido}, e diante da possibilidade de correlação entre as observações, resultante da estrutura hierárquica, utilizaremos um modelo logístico hierárquico. Ainda, como a amostra restringe-se às crianças com um ano de idade, à exceção dos poucos casos de gemelares, a chance de se ter mais de uma criança por família é desprezível. Logo, utilizaremos um modelo logístico hierárquico de dois níves, sendo estes: 1) a criança; 2) a comunidade onde a criança reside (setor censitário).

O modelo pode ser especificado na forma

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \tau_j,$$

em que $\log\{\pi_{ij}/1-\pi_{ij}\}$ representa a chance da i-ésima criança do j-ésimo setor censitário estar desnutrida. Ainda, $\boldsymbol{\beta}$ é um vetor de parâmetros associado às variáveis explicativas. Tem-se também que τ_j, com distribuição $\mathcal{N}(0,\sigma^2)$, é o efeito aleatório que captura a variação entre as comunidades.

Na Tabela 13 apresentamos as estimativas dos parâmetros do modelo juntamente com os erros padrão. Os resultados foram obtidos utilizando o pacote estatístico `R` em sua versão 2.3.1 (www.r-project.org).

Tabela 13 - Estimativas dos parâmetros e razão de chances (usado para estimar a chance de uma criança estar desnutrida no segundo ano de vida)

Variáveis	Coeficientes	Erro padrão	Odds Ratio
Região			
Nordeste[a]	–	–	–
Norte	$0,285$	0,402	1,33
Sudeste	$-1,549^*$	0,458	0,21
Sul	$-1,354^*$	0,605	0,26
Centro-Oeste	$-0,508$	0,453	0,60
Idade materna			
20–34 anos[a]	–	–	–
$\leq$ 19 anos	$0,983^*$	0,350	2,67
35 ou mais	$-0,284$	0,455	0,75
Ordem de nascimento			
2–3 filhos[a]	–	–	–
Primeiro filho	$-0,708^*$	0,366	0,49
4 ou mais	$0,593^*$	0,322	1,81
Intervalo prévio			
$\leq$ 18 meses[a]	–	–	–
$>$ 18 meses	$-0,791^*$	0,339	0,45
Educação materna			
0–3 anos[a]	–	–	–
4–8 anos	$-0,543^*$	0,284	0,58
9 ou mais	$-0,552$	0,429	0,58
Percepção do peso da criança			
Médio[a]	–	–	–
Grande/muito grande	$-0,318$	0,284	0,73
Pequeno/muito pequeno	$1,039^*$	0,429	2,83
Efeito Aleatório	Estimativa	Erro Padrão	
σ^2	$1,465^*$	0,451	

[a] Categoria de referência,
* Significativo a 5%,

Observe que o parâmetro aleatório σ^2 associado com a área em que a criança reside foi significativo. Em relação aos riscos existentes entre crianças da região Nordeste, a proteção contra riscos de desnutrição no segundo ano de vida entre crianças residentes na região Sudeste foi de 79%. Entre as crianças da região Sul foi de 74%. Os riscos verificados entre as crianças moradoras das regiões Norte e Centro-Oeste não apresentaram significância estatística.

Os riscos encontrados entre filhos de mães adolescentes corresponderam a 2,67 em relação aos riscos verificados entre as mães com idades entre 20 a 34 anos no momento do nascimento de seus filhos. Os primeiros filhos tiveram a metade dos riscos de desnutrição no segundo ano de vida que as crianças nascidas de ordem 2 ou 3, enquanto que crianças de ordem 4 ou mais tiveram os riscos 80% superiores aos encontrados entre o mesmo grupo de comparação. A proteção atribuída ao nascimento após um intervalo maior ou igual a 18 meses, em relação aos ocorridos após um intervalo inferior a essa idade, foi de 55%. Ter 4 a 8 anos de estudo representou riscos de desnutrição no segundo ano de vida. O impacto de ter 9 ou mais anos de escolaridade não se mostrou estatisticamente significativo. A percepção materna de que seu filho nasceu pequeno ou muito pequeno foi relevante. A chance de um filho percebido por sua mãe como pequeno ou muito pequeno ao nascer apresentar desnutrição no segundo ano de vida é quase três vezes maior do que o observado entre os considerados normais ao nascimento por suas mães.

Conclusões

Apresentamos neste artigo um estudo sobre os estimadores dos parâmetros em um modelo logístico hierárquico de dois níveis. As estimativas dos parâmetros foram obtidas utilizando o método da quasi-verossimilhança marginal de primeira ordem, proposto em Goldstein (1991). Na parte pontual, conclui-se que os estimadores apresentam vieses consideráveis. Além disso, os estimadores tendem a subestimar os verdadeiros valores dos parâmetros. Em relação a parte intervalar, observou-se que, à medida que o número de grupos aumenta, as taxas de cobertura aproximam-se dos verdadeiros níveis de cobertura. Adicionalmente, o poder do teste de significância também aumenta quando o número de grupos cresce. Este comportamento também foi observado quando a diferença entre o número de indivíduos amostrados por grupo diminuem.

Agradecimentos

Artur José Lemonte agradece ao apoio finaceiro da CAPES e Tatiane Ferreira Nascimento Melo da Silva a FAPESP. Os autores agradecem aos pareceristas pelos comentários e sugestões.

LEMONTE, A. J.; SILVA, T. F. N. M. Point and interval estimation in the hierarchic logistic model two-level. *Rev. Mat. Estat.,* São Paulo, v.24, n.2, p.147-162, 2006.

- *ABSTRACT: A study on parameter estimation in the two-level hierarchical logistic model is presented in this paper. Estimators based on first-order marginal quasi-likelihood were taken into account, and their estimation algorithm can be found in Goldstein (1991). The performance of these estimators was evaluated through exact and interval estimation. Additionally, the cover rates of the confidence intervals and power of the significance tests are also presented. Monte Carlo simulation was used, and finally, an application of real data is presented and discussed.*

- *KEYWORDS: Hierarchic models; interval estimation; logistic model; Monte Carlo simulation; punctual estimation; random effect.*

Referências

BRESLOW, N. E.; CLAYTON, D. G. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.*, New York, v.88, p.9–25, 1993.

COX, D. R.; SNELL, E. A general definition of residuals. *J. R. Stat. Soc., Ser. B*, London, v.30, p.248–275, 1968.

CRUZ, M. C. C.; LEITE, I. C. Fatores de risco para déficits estaturais no segundo ano de vida: Brasil, PNDS, 1996. *Rev. Bras. Est. Popul.*, Campinas, v.19, p.131–140, 2002.

DAS GRUPTA, M. Death clustering, mother's education and the determinants of child mortality in rural Punjab, India. *Popul. Stud.*, London, v.44, p.498–505, 1990.

DOORNIK, J. A. *Ox*: An object-oriented matrix language. 4th ed. London: Timberlake Consultants Press, 2001. 408p.

FERRARI, S. L. P.; CRIBARI–NETO, F. Beta regression for modeling rates and proportions. *J. Appl. Stat.*, Abingdon, v.31, p.799-815, 2004.

GOLDSTEIN, H. Nonlinear multilevel models with an application to discrete response data. *Biometrika*, London, v.78, p.45-51, 1991.

GOLDSTEIN, H.; RASBASH, J. Improved approximations for multilevel models with binary responses. *J. R. Stat. Soc. Ser A*, London, v.159, p.505–513, 1996.

GRIFFITHS, P. *Effects of household structure, kinship and caste on childhood nutritional status in two culturally contrasting states of India.* 1999. 123f. Dissertation (Magister in Science). Department of Social Statistics. Public Health Medicine, University of Southampton. Southampton, 1999.

McCULLAGH, P.; NELDER, J. A. *Generalized linear models*, 2nd ed. London: Chapman and Hall, 1989. 511p.

RODRÍGUES, G.; GOLDMAN, N. An assessment of estimation procedures for multilevel models with binary responses, *J. R. Stat. Soc. Ser. A*, London, v.158, p.73–89, 1995.

Recebido em 02.04.2006.

Aprovado após revisão em 23.06.2006.