Comparação de métodos de estimação em modelos logísticos multiníveis para dados longitudinais: Um estudo de simulação.

Renata de M. Esquivel [(1)] , Leila D.A.F. Amorim [(2,3)]; Rosemeire L. Fiaccone[(2)].

1. Mestranda do Centro Integrado de Manufatura e Tecnologia (CIMATEC)/ SENAI.
2. Departamento de Estatística da UFBA.
3. Projeto com financiamento FAPESB, Termo de Outorga nº.0082/2006.

I. Introdução

O desenvolvimento de métodos estatísticos para análise de dados obtidos em situações em que as observações são dependentes tem apresentado crescimento relevante nas últimas décadas e, em especial, na análise de dados provenientes de estudos longitudinais. Em estudos longitudinais pode-se avaliar o desenvolvimento individual ou geral da variável resposta ao longo do tempo. Os estudos longitudinais têm como característica principal a mensuração repetida de uma variável resposta em um mesmo indivíduo ou unidade de investigação ao longo do tempo. Isto implica que as observações obtidas de um mesmo indivíduo em várias ocasiões são naturalmente correlacionadas. Para considerar a dependência entre as observações repetidas de cada indivíduo em estudos longitudinais e responder às questões de investigação de maneira satisfatória, a utilização de técnicas estatísticas sofisticadas faz-se necessária (Twisk, 2003).

Como técnica de análise para dados longitudinais pode-se utilizar os modelos multiníveis que contêm variáveis mensuradas em níveis diferentes de hierarquia, possibilitando avaliar a variabilidade desses níveis através da introdução de efeitos aleatórios. Considerando o enfoque longitudinal, as observações repetidas de cada indivíduo constituem o nível mais baixo e estas observações são agrupadas dentro dos respectivos indivíduos (Twisk, 2005; Goldstein, 1995).

A estimação em modelos multiníveis lineares generalizados (GLMM) pode ser realizada sob dois enfoques: métodos de aproximação da verossimilhança como, por exemplo, a quasi- verossimilhança penalizada (PQL), ou por integração numérica (Hox, 2002). Estudos de simulação sugerem que as aproximações por integrais numéricas têm performance consideravelmente melhor do que os demais métodos de aproximação (Raudenbush e Yang, 1998; Raudenbush *et. al*., 2000). Segundo Rodriguez e Goldman (1995), quando se tem resposta binária, os procedimentos que utilizam a verossimilhança penalizada podem gerar estimativas muito viesadas para os efeitos fixos e aleatórios, sobretudo em amostras com poucos conglomerados. Vários métodos de estimação são discutidos na literatura para estimação dos parâmetros quando se considera os modelos logísticos multiníveis. Esse

trabalho objetiva comparar métodos de estimação disponíveis no *software* R para modelos logísticos multiníveis na análise de dados longitudinais.

II. Metodologia

Os modelos multiníveis logísticos fazem parte da família dos modelos multiníveis não lineares. No contexto dos modelos multiníveis logísticos para dados longitudinais utiliza-se as observações repetidas binárias de um mesmo indivíduo ao longo do tempo. Os dados longitudinais têm uma estrutura multinível e, portanto, as observações repetidas de cada indivíduo constituem o nível mais baixo da hierarquia, sendo estas agrupadas dentro dos respectivos indivíduos. Deste modo, pode-se quantificar a variação entre as observações 'dentro' de um mesmo individuo e a variância entre os diferentes indivíduos (Twisk, 2005; Goldstein, 1995; Collett, 1952).

No presente trabalho considerou-se o modelo multinível logístico para dados longitudinais com apenas o intercepto aleatório, dois níveis de hierarquia e sem covariável variante no tempo, que pode ser descrito por:

$$\log it(\pi_{ij}) = \gamma_{00} + \left(\sum_{k=1}^{p} \beta_k x_{ki} \right) + u_i$$

em que ***k*** , ***i*** e ***j*** denotam os índices dos parâmetros, dos indivíduos e das observações repetidas ao longo do tempo, respectivamente. Nesse caso γ_{00} é o intercepto que representa a resposta esperada para a população, β_k (k= 1, ... , p) representa o k-ésimo coeficiente de regressão e π_{ij} denota a probabilidade de ocorrência do evento de interesse para o indivíduo i (i= *1,..., n*) no j-ésimo tempo (j=*1, .., T)*. O termo u_i denota o efeito aleatório do i-ésimo indivíduo, que pode ser usado para investigar as mudanças na probabilidade de resposta ao longo do tempo. Assume-se que $u_i \sim N(0, \tau_{00})$, onde τ_{00} denota a variância entre as medidas repetidas dos indivíduos, considerando que a probabilidade de ocorrência da resposta em cada tempo compartilha o mesmo efeito do indivíduo (Collett , 1952; Hox, 2000). Sendo assim, as respostas y_{ij} podem ser escritas como: $y_{ij} = \pi_{ij} + e_{ij}$, em que $e_{ij} \sim Bin[0, \pi_{ij}(1-\pi_{ij})]$ são os erros residuais e no contexto dos dados longitudinais estes erros são correlacionados (Goldstein, 1995). Esta característica é considerada na estrutura de covariância dos dados.

Para quantificar o grau de dependência entre as medidas repetidas de um mesmo indivíduo pode-se utilizar o coeficiente de correlação intraclasse (ICC), que estima a proporção da variabilidade da resposta explicada pela variabilidade que ocorre entre as observações repetidas dos indivíduos. O ICC para um modelo logístico multinível com apenas o intercepto aleatório e dois níveis de variação pode ser definido como:

$$ICC = \frac{\tau_{00}}{\tau_{00} + \left(\frac{\pi^2}{3} \right)}$$

em que τ_{00} e $\tau_{00} + \left(\frac{\pi^2}{3}\right)$ representam, respectivamente, a variância entre as medidas repetidas de um mesmo indivíduo e a variância total (soma das variâncias entre e dentro dos indivíduos) (Goldstein,1995; Twisk, 2005; Collet, 1952). Segundo Collet (1952), o principal uso desta medida estatística está relacionado com a comparação da magnitude dos efeitos entre as medidas repetidas do individuo por diferentes estudos.

II.1 Métodos de Estimação

Na regressão multinível linear a estimação dos parâmetros pode ser realizada através do método de máxima verossimilhança (ML), assumindo-se normalidade dos erros, ou alternativamente a máxima verossimilhança restrita (REML) (Goldstein, 1995). Para os modelos logísticos multiníveis as estimativas são obtidas através da maximização da função de verossimilhança marginal. Em termos gerais, a estimação para modelos multiníveis lineares generalizados pode ser realizada por dois enfoques: métodos de aproximação da verossimilhança ou de integração numérica (Hox, 2002). Para o primeiro enfoque a linearização da função de ligação é feita pela expansão da série de Taylor, podendo ser de primeira ou segunda ordem. Caso a expressão resultante da integração dos efeitos aleatórios não seja analiticamente possível, pode-se recorrer à integração numérica, como quadratura Gaussiana. Alguns procedimentos, no entanto, podem evitar a alta demanda computacional, como, por exemplo, a quasi-verossimilhança preditiva ou penalizada (PQL), a aproximação de Laplace e a quasi-verossimilhança marginal (MQL). Estudos de simulação sugerem que quando ambos os enfoques são possíveis, o método da integração numérica alcança estimativas mais precisas do que as obtidas pelos MQL ou PQL. Porém, o método de integração numérica pode encontrar problemas para obtenção de convergência (Goldstein, 1995 ; Venables, 2002).

A verossimilhança marginal, obtida pela integração da distribuição conjunta em relação ao efeito aleatório, é analiticamente intratável, porém várias aproximações têm sido propostas na literatura. As aproximações numéricas disponíveis na literatura para linearizar a função de ligação e que são utilizadas neste trabalho incluem aproximação de Laplace, PQL e integrações numéricas. A aproximação de Laplace lineariza a parte não linear usando expansão em série de Taylor até a segunda ordem com base na maximização de uma função conhecida e unimodal. A aproximação realizada através do PQL, por sua vez, consiste na linearização da parte não linear usando expansão em série de Taylor com base nos coeficientes de regressão estimados e nos resíduos atuais u_i , podendo ser de primeira ou segunda ordem. Já a integração aproximada através do MQL considera a linearização apenas baseado na parte fixa do preditor linear (Hox, 2002; Goldstein e Rasbash, 1996).

Considerando outro enfoque de estimação, a verossimilhança analiticamente intratável pode ser solucionada numericamente através da quadratura Gaussiana ordinária (GQ) ou a sua versão melhorada, a quadratura Gaussiana Adaptativa (AGQ) (Rabe- Hesketh *et.al*., 2005). Os métodos de integração numérica maximizam a verossimilhança correta e, neste caso, testes e índices para avaliação da bondade do ajuste baseados na *deviance*

podem ser usados (Hox, 2002). A quadratura de Gauss-Hermite ou quadratura Gaussiana Adaptativa aproxima a integral intratável através da soma ponderada do integrando em um conjunto de valores da variável que está sendo integrada (Skrondal e Rabe-Hesketh, 2004).

II.2 Estudos de simulação

Para comparar os diferentes métodos de estimação para o modelo logístico multinível foram conduzidos estudos de simulação com diferentes graus de dependência (ICC=0,2 ; 0,5; e 0,8), número de observações repetidas (3; 6; 10) e número de indivíduos (50; 150; 300). Para cada configuração foram geradas 2 000 amostras, que foram ajustadas considerando-se os diferentes métodos de estimação. Os métodos de estimação são avaliados através do vício associado a efeitos fixos e aleatórios, da estimação da variabiliade do parâmetro e da probabilidade de cobertura (CP, em inglês) no intervalo de 95% de confiança de Wald.

Nos estudos de simulação foram geradas variáveis resposta binárias correlacionadas, sendo fixados os seguintes valores para os parâmetros fixos do modelo : γ_{00} = 0,05; β_1=0,9 e β_2=0,6. A geração dos dados seguiu procedimento descrito em Santos e colaboradores (2008). Considerou-se duas variáveis independentes (X_{1i} e X_{2i}) , sendo X_1 ~ Bernoulli (π_{ij}) e X_2 ~ Normal (μ, σ^2). O efeito aleatorio foi definido como u_i ~ Normal (0, τ_{00}).

As simulações foram conduzidas no software R, versão 2.10.1. Utilizou-se a função *glmer* do *software* R considerando o seu *default*, referente à aproximação por Laplace, além da Quadratura Gaussiana Adaptativa com dois diferentes pontos de integração (8 e 25). Utilizou-se também a função *glmmPQL* , cujo método de estimação é o PQL.

III. Resultados dos Estudos de Simulação

Verificou-se, de forma geral, para todos os métodos de estimação que o vício médio, o erro padrão médio do estimador do efeito fixo, o desvio padrão das estimativas dos parâmetros do modelo e a variância média do efeito aleatório tendem a crescer com o aumento do ICC. A prevalência média da variável resposta variou entre 54.8% a 59.6%.

Ao se considerar observações repetidas para 50 indivíduos o método da Quadratura Gaussiana foi aquele que apresentou menores vícios, particularmente para correlação alta entre as medidas repetidas e uma maior quantidade de medidas no tempo. Por exemplo, quando o ICC=0,8 e com 10 medidas repetidas, verifica-se os seguintes vícios: 0,067 (Laplace), 0,046 (Quadatura Gaussiana – ambos pontos de integração) e 0,168 (PQL). Em geral, o PQL tem CP mais próximos de 95% , apresentando a melhor estimação da variabilidade associada aos parâmetros fixos, apesar do maior vício. Em relação as médias da variância do efeito aleatório observou-se que a aproximação por Laplace apresenta valores menores do que os obtidos pela PQL.

Para a Quadratura Gaussiana Adaptativa verificou-se que não houve grandes diferenças nas propriedades dos estimadores de acordo com o número de pontos de integração considerados. Contudo, o vicio médio é ligeiramente maior para uma menor quantidade de pontos de integração, especialmente para valores altos do ICC. Além disso, pode ser observado uma pequena diferença entre as variâncias médias do efeito aleatório para um ICC de maior magnitude para qualquer que seja o número de medidas repetidas.

Com o aumento do número de indivíduos e de medidas repetidas, verificou-se uma melhoria das propriedades dos estimadores considerados. De modo geral, usando-se observações em 150 indivíduos, os resultados apontam para uma redução dos vícios médios, sobretudo para a Quadratura Gaussiana (ambos número de pontos de integração), seguido pelo método de aproximação de Laplace. Em relação a CP, a Quadratura Gaussiana e Laplace apresentaram resultados similares, contendo as maiores probabilidades de cobertura.

IV. Considerações finais

Os métodos de estimação avaliados neste estudo apresentam performance dependente sobretudo da combinação entre número de mensurações ao longo do tempo e correlação (indicada neste estudo pelo coeficiente de correlação intraclasse). De modo geral, os piores resultados foram obtidos com o aumento do ICC e com a diminuição no número de medidas repetidas. A maior magnitude do vício foi verificada através do uso do PQL, corroborando com estudos anteriores (Rodriguez e Goldman, 1995; Callens e Croux, 2003), apesar deste método apresentar a melhor performance em relação ao estimador da variabilidade dos parâmetros fixos do modelo. O número de pontos de integração investigados neste estudo não foi um fator importante na performance do método da Quadratura Gaussiana.

Os resultados apontam para evidências de que o aumento do número de indivíduos sob estudo, bem como do aumento do número de medidas repetidas, tornam os métodos de estimação considerados neste trabalho com performance similar. Neste caso, a escolha do método de estimação vai depender da sua disponibilidade em software estatístico.

Referências

Callens, M. and Croux, C. Performance of likelihood-based estimation methods for multilevel binary regression models. Journal of Statistical Computation and Simulation. v. 75, 12:1003–1017, 2005.

Collet, D. Modelling Binary Data. Texts in Statistical Science. 2nd Chapman & Hall/CRC, 1952.

Diggle, P.J., *et al*. Analysis of Longitudinal Data. 2nd Oxford, 2002.

Goldstein, H. Multilevel statistical models. 2nd Ed.London: Edward Arnold, 1995.

Goldstein, H. and Rasbash, J. .Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society A , 159:505-13 , 1996.

Hesketh, S.R. *et. al*. Reliable estimation of generalized linear mixed models using adaptive quadrature.The Stata Journal 2, Number 1, 1–21 , 2002.

Hesketh S.R., *et. al*. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics 128: 301–323, 2005.

Hox J.J. Modeling longitudinal and multilevel data. Multilevel Analyses of Grouped and Longitudinal Data. Lawrence Erlbaum Associates, 2000.

Hox, J.J Multilevel Analysis.Techniques and Application. Lawrence Erlbaum Associates, 2002.

Leyland, A.H. and Goldstein, H. Multilevel Modelling of Health Statistics. John Wiley & Sons, 2001.

Molenberghs, G. and Verbeke G. Models for Discrete Longitudinal Data. Springer Series in Statistics, 2005.

Rodriguez, G. and Goldman, N. An assessment of estimation procedures for multilevel models with binary response. Journal of the Royal Statistical Society Series *A*, 158: 73–89, 1995.

Santos C.A.S.T, *et. al*. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. BMC-Medical Research Methodology, 1471-2288 / 8-80, 2008).

Skrondal A. and Rabe-Hesketh S (2004). Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Chapman & Hall/CRC.

Twisk J.W.R. Applied Multilevel Analysis: a Practical Guide. Cambridge University Press, 2005.

Twisk J.W.R. Applied Longitudinal Data Analysis for Epidemiology. Cambridge University Press, 2003.

Venables, W.N. and Ripley, B.D. Modern Applied Statistics. Statistics and Computing with S. 4nd Springer, 2002.