

MODELAGEM MULTINÍVEL

Carlos A. de S. Teles Santos *

Leila D. A. Ferreira **

Nelson F. Oliveira ***

Maria Inês C. Dourado ****

Maurício L. Barreto *****

RESUMO — *Recentemente os modelos lineares hierárquicos ou multiníveis começaram a ser mais conhecidos em diversos campos do conhecimento científico, principalmente na área da Educação e das Ciências Sociais. Importantes propriedades dos modelos lineares hierárquicos permitem que a variabilidade da variável resposta: (a) seja explicada através de variáveis preditoras incluídas em diferentes níveis hierárquicos e (b) possa quantificar em cada nível, tal que a proporção da variabilidade explicada possa ser comparada diretamente. Assim, as análises multiníveis apresentam estimativas mais fidedignas, uma vez que não assumem erroneamente o pressuposto de independência entre as observações das unidades pertencentes a um agregado, como ocorre nas análises contextuais e naquelas com observações repetidas.*

PALAVRAS-CHAVE: *Modelos Lineares Hierárquicos; Modelos Lineares Multiníveis; Modelos Lineares Generalizados.*

ABSTRACT — *Recently hierarchical linear models or multilevel models have become better known in several scientific fields, particularly in the areas of education and social science. Hierarchical linear models possess important properties, the first of which allows the variability of the response variable to be explained through explanatory variables included in the model at different levels. The second enables researchers to quantify how much of the variability of the response variable is due to each level,*

*Prof. Assistente do Dep. de Ciências Exatas (UEFS). Mestre em Saúde Comunitária - Epidemiologia e Pesquisador (Instituto de Saúde Coletiva - ISC/UFBA). E-mail: carlosts@uefs.br

** Prof. do Dep. de Estatística (UFBA). Mestre em Saúde Comunitária - Epidemiologia e Pesquisadora (ISC/UFBA). E-mail: leiladen@ufba.br

***Prof. do Dep. de Estatística (UFBA). Phd em Bioestatística-Shapel Hill/California-USA. E-mail: genaro@ufba.br

**** Pesquisadora (ISC/UFBA). Phd em Epidemiologia. E-mail: maines@ufba.br

***** Prof. Titular e Pesquisador (ISC/UFBA). Phd em Epidemiologia (London School-Inglaterra). E-mail: mauricio@ufba.br

such that the proportion of the explained variability can be compared directly. Thus, multilevel analyses present more reliable estimates because they do not erroneously presume independence among the observations of the units belonging to an aggregate, as occurs in contextual analysis; also they do not waste information by taking averages of the repeated observations.

KEY WORDS: *Hierarchical Linear Models; Multinivel Linear Models; Generalized Linear Models.*

INTRODUÇÃO

Na década de 70, com o desenvolvimento computacional, alguns modelos que exigiam a utilização de processos iterativos para a estimação dos parâmetros começaram a ser utilizados. McCULLAGH & NELDER (1989) propuseram os modelos lineares generalizados (GLMs), cuja idéia básica consiste na possibilidade de ampliar a distribuição da variável resposta, permitindo que a mesma pertença à família exponencial de distribuições, bem como de dar maior flexibilidade para a relação funcional entre a média da variável resposta (μ) e o preditor linear (η). Assim, para dados de contagem, em vez de aplicarmos algum tipo de transformação na variável resposta Y com o intuito de normalizá-la, podemos, por exemplo, usar a distribuição de Poisson e a relação funcional entre a média de Y e o preditor linear $\eta = \log \mu$. Essa relação é conveniente, uma vez que garante um valor positivo para quaisquer valores dos parâmetros do preditor linear.

A partir de 1972, inúmeros artigos sobre os modelos lineares generalizados foram publicados, e o programa computacional GLIM (*Generalized Linear Interactive Models*) foi desenvolvido para ajustar tais modelos. Hoje, outros programas, tais como, o SPSS, SAS, e SPLUS, apresentam procedimentos para o ajuste dos GLMs.

Dentre os artigos publicados, merece destaque o trabalho de WEDDERBURN (1974) sobre modelos de quasi-verossimilhança. Trata-se de uma extensão natural dos GLMs para modelos mais gerais que podem incluir dados correlacionados, o que não era possível com os demais GLMs que assumem respostas independentes. Os modelos de quasi-verossimilhança assumem apenas a existência

dos dois primeiros momentos da distribuição da variável resposta Y , sem ser necessário conhecer a forma da distribuição, e impõem uma relação funcional conveniente para a média de Y e o preditor η , que aqui pode ser não-linear nos parâmetros. Com essa proposta, é mostrada a consistência e a normalidade assintótica dos estimadores de quasi-verossimilhança, os quais saem como solução de uma equação de estimação particular, que se dá de forma iterativa. Apesar dessa proposta, somente na segunda metade da década de 80, os modelos de quasi-verossimilhança começaram a ganhar mais atenção, principalmente com a publicação de LIANG & ZEGER (1986), que propuseram uma forma simples de tratar dados longitudinais através dos modelos lineares generalizados (PAULA, 1997).

CARACTERÍSTICAS DOS MODELOS LINEARES GENERALIZADOS

Os modelos lineares generalizados (*GLMs*) incluem os modelos lineares tradicionais (erros com distribuição normal), bem como os modelos logísticos, para respostas com distribuição binomial, os modelos log-lineares, para respostas com distribuição multinomial, e os modelos de regressão de Poisson, para respostas com distribuição de Poisson.

Muitos outros modelos estatísticos podem ser gerados pela seleção apropriada de uma distribuição de probabilidade para a variável resposta e de uma função de ligação. Os modelos lineares generalizados são constituídos pelos seguintes componentes:

⇒ A variável resposta $Y = \{Y_1, \dots, Y_n\}$, que segue alguma distribuição de probabilidade da família exponencial ;

⇒ A função de ligação, $g(\cdot)$, que é uma função monótona e diferenciável, definida nos *GLMs* por $g(\mu_i) = \eta$, onde $\eta = X^T \beta$ é o preditor linear; $\beta = (\beta_1, \dots, \beta_p)^T$, $p < n$, é um vetor de parâmetros desconhecidos a serem estimados e $X^T = (x_{i1}, \dots, x_{ip})^T$ representa os valores de p variáveis explicativas.

Nos modelos lineares tradicionais, a distribuição de probabilidade dos erros é a normal e a função de ligação é a identidade:

$g(\mu)=\mu$. Na regressão logística, a distribuição é a binomial e a função de ligação é o *logit*: $g(\mu)=\log(\mu/1-\mu)$; e na regressão de Poisson, a distribuição é Poisson e a função de ligação é $g(\mu)=\log(\mu)$.

Um tratamento parecido é dado aos modelos lineares multiníveis quando a análise apropriada envolve variáveis de contagem ou binárias ou mesmo quando as transformações não normalizam os dados. Esses modelos multiníveis são uma extensão direta dos modelos lineares generalizados de McCULAGH & NELDER (1989). Portanto, ter-se-ão os modelos lineares multiníveis generalizados, nos quais a natureza da variável vai determinar o tipo de distribuição de probabilidade que será utilizada, sendo as mais comuns a binomial, para respostas dicotômicas, e a Poisson, para respostas resultantes de processos de contagem.

DADOS COM ESTRUTURAS MULTINÍVEIS OU HIERÁRQUICAS

Umas das estruturas de dados que podem ser analisadas pelos modelos multiníveis ou hierárquicos consistem naquelas estruturas cujas unidades são agrupadas (aninhadas) em diferentes níveis hierárquicos. Isso ocorre em estudos de organizações (instituições), por exemplo, nos quais o pesquisador pode estar interessado em investigar como as características do local de trabalho influenciam na produtividade do trabalhador. Tais dados consistem em uma estrutura hierárquica, com as unidades de trabalhadores aninhados dentro das firmas. Assim, variáveis poderão ser mensuradas em ambos os níveis: trabalhador e firma.

O mesmo tipo de situação poderia ocorrer nos estudos de corte-transversal. Suponha, como exemplo, que demógrafos desejassem examinar como diferenças no desenvolvimento da economia nacional pudessem interferir na relação entre o grau educacional dos adultos e a taxa de fertilidade. Essas pesquisas combinam indicadores econômicos coletados em nível nacional com informações coletadas a nível domiciliar sobre a educação

e a fertilidade. Dessa forma, domicílios e países são unidades na pesquisa, estando os domicílios aninhados dentro dos países.

As análises hierárquicas também podem ocorrer nos estudos da área da educação, como, por exemplo, na mensuração do rendimento escolar das crianças que compõem as turmas de alunos. Essas, por sua vez, compõem as escolas. E as escolas são dirigidas por órgãos administradores. Assim, poderíamos ter quatro níveis hierárquicos associando-se as crianças ao 1º nível, as turmas de alunos, ao 2º, as escolas, ao 3º e os órgãos administradores, ao 4º nível.

Devido à larga abrangência de problemas com essas formas de estruturas dos dados, as aplicações dos modelos lineares multiníveis vêm crescendo. Nos últimos 15 anos, muitas extensões dos modelos lineares multiníveis foram desenvolvidas, possibilitando a inclusão de variáveis em ambos os níveis: individual e agregado. Dessa forma, na modelagem multinível, as estimativas dos efeitos fixos se aproximam daquelas obtidas pelas análises tradicionais, que são realizadas em nível individual. Os erros-padrão aproximam-se daqueles obtidos nas análises contextuais, ou seja, daquelas realizadas em nível do agregado. Além disso, a modelagem multinível fornece informações sobre a composição da variabilidade total. Segundo BRYK & RAUDENBUSH (1992), esses modelos não são uma solução para todos os problemas, contudo, eles representam um grande passo para auxiliar as análises por serem estatisticamente corretos e não desperdiçarem informação.

TERMINOLOGIA DOS MODELOS LINEARES MULTINÍVEIS

Os modelos lineares multiníveis são também conhecidos como modelos lineares hierárquicos (*HLM*), modelos com coeficientes aleatórios, “*empirical bayes models* ou *seemingly unrelated regressions*” (KREFT *et al.*, 1994). Apesar dessa diversidade de nomenclaturas, pode-se dizer que o termo genérico, *HLM*, refere-se à estimação Bayesiana de modelos lineares com dados que têm estruturas de erros complexas. A principal diferença desses modelos reside nas estimativas dos erros-padrão, que serão mais precisas do que os das análises tradicionais, uma

vez que a correlação intra-classe é considerada ao ajustarmos os modelos. Este fato faz com que os modelos de coeficientes aleatórios sejam mais conservadores (KREFT & LEEU, 1998). Por isso, se o pesquisador estiver interessado em componentes de covariância (variabilidade entre e intra unidades), os modelos com coeficientes aleatórios e os programas computacionais para análise de dados com estruturas hierárquicas serão uma boa escolha.

MÉTODOS DE ESTIMAÇÃO DOS MODELOS LINEARES MULTINÍVEIS

O ajuste de modelos multiníveis é baseado na estimação dos mínimos quadrados generalizados (GLS). Dessa forma, escreve-se um modelo contendo parâmetros fixos e aleatórios, como se segue: $Y = X\beta + Z\theta$, onde X é a matriz associada com o vetor de parâmetros fixos, Z é a matriz associada com o vetor de parâmetros aleatórios e Y é o vetor de respostas. A estimação dos parâmetros é feita iterativamente através dos mínimos quadrados generalizados iterativos (IGLS), ajustando-se modelos de regressão para as partes fixa e aleatória (GOLDSTEIN, 1995).

Suponha um simples modelo de componente de variância com dois níveis: $Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_{0j} + r_{ij}$. Através do GLS, estimam-se os coeficientes fixos, $\beta = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$, onde V é a matriz de variância e covariância de Y , com resíduos normais. Esse procedimento iterativo é iniciado a partir de uma estimativa razoável de mínimos quadrados ordinários (OLS) dos coeficientes fixos. Tendo estimado os parâmetros fixos, β , extrai-se um vetor residual, $Y' = Y - X\beta$, que é utilizado para estimar os parâmetros aleatórios do modelo. Calcula-se, em seguida, $E(Y^*) = Y' Y'^T = V$, construindo o vetor $Y^{**} = \text{Vec}(Y^*)$ para ser utilizado como variável resposta na equação de regressão para estimar os parâmetros aleatórios, $\theta = (Z^T V^{*-1} Z)^{-1} Z^T V^{*-1} Y^{**}$. Nesta equação, V^* é o produto de Kronecker de V ($V^* = V \otimes V$). Assim, alterna-se entre estimar o vetor de parâmetros fixos e aleatórios até a convergência do modelo. Para os modelos não-lineares é utilizada a quasi-verossimilhança (GOLDSTEIN, 1995).

A LÓGICA DOS MODELOS LINEARES MULTINÍVEIS COM 2 NÍVEIS

O modelo para a análise no nível 1 pode ser denotado por:

$$(I) \quad Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij},$$

onde r_{ij} tem distribuição normal, $N(0, \sigma^2)$, com n_i unidades no nível 1 dentro de cada J -ésima unidade no nível 2;

Y_{ij} é uma função do conjunto de características individuais das unidades no nível 1 (estudantes, ocasiões, indivíduos, estados etc.).

Utilizando o modelo (I) no nível 1, podemos estudar a(s) associação(es) da(s) variável(is) X_{ij} com a resposta Y_{ij} dentro de uma população de unidades no nível 2 (escolas, crianças, organizações, regiões, domicílios etc.).

Cada unidade no nível 2 é descrita pelo par de valores (β_{0j}, β_{1j}) , que podem ser preditos através das características das unidades no nível 2. Por exemplo, considere a variável indicadora W_j . Assim, o modelo no nível 2 será:

$$(IIa) \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \quad \text{e} \quad (IIb) \beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}.$$

Substituindo as equações (IIa) e (IIb) em (I), produz-se a equação de predição para a resposta:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \bar{X}_{.j}) + \gamma_{11} W_j(X_{ij} - \bar{X}_{.j}) + u_{0j} + u_{1j}(X_{ij} - \bar{X}_{.j}) + r_{ij}.$$

Este é um exemplo simples de um modelo linear hierárquico.

Note que a equação acima não é um típico modelo linear padrão assumindo mínimos quadrados ordinários (OLS), pois os erros têm uma estrutura mais complexa. Tais erros, u_{0j} e u_{1j} , são dependentes dentro de cada j -ésima unidade no nível 2.

E esses erros têm variâncias desiguais porque u_{0j} e u_{1j} variam ao redor das unidades do nível superior (2), e u_{1j} também varia de acordo com os valores de $(X_{ij} - \bar{X}_{.j})$ das unidades no nível 1.

Assim, os modelos lineares multiníveis ou hierárquicos fornecem uma maior flexibilidade na formulação de modelos explicativos nas pesquisas em diferentes áreas do conhecimento: educacional, geográfica, sociológica, epidemiológica, demográfica e econômica.

PROGRAMAS PARA ANÁLISE DE DADOS MULTINÍVEIS

Tradicionalmente, pacotes estatísticos, seja para as análises lineares ou para uso de modelos lineares generalizados, consideram apenas um nível hierárquico e uma variável aleatória no ajuste dos modelos. Porém, em meados dos anos 80, foram desenvolvidos vários programas para o ajuste de modelos multiníveis. Entre eles, destacam-se o BMDP-5V, GENMOD, HLM3, MLN3 e VARCL.

KREFT *et al.* (1994) descreve esses programas, ressaltando o tipo de análise realizada, o ambiente computacional onde os pacotes podem ser executados, o algoritmo utilizado e, ao final, faz uma comparação de desempenho entre eles. Segundo esses autores, o BMDP-5V foi desenvolvido para analisar dados de medidas repetidas com especial ênfase para dados não-balanceados, sendo eficiente para pequenos conjuntos de dados no contexto biológico. O BMDP-5V possui versões em ambientes computacionais de grande porte: MVS, VM, VAX/VMS e, mais recentemente, no ambiente MS-DOS para PC. As estimativas de todos os parâmetros através do método de máxima verossimilhança são realizadas por um dos três diferentes algoritmos: *Newton-Raphson*, *Fisher scoring* e algoritmo *EM* generalizado. Esses algoritmos produzem estimativas idênticas, mas com diferente número de iterações.

Depois disso, o programa VARCL foi desenvolvido, usando o algoritmo *Fisher scoring* para implementar as análises de coeficientes aleatórios sem permitir termos de interação entre variáveis de níveis diferentes. O VARCL pode ser utilizado para analisar dados multivariados. Em 1988, criaram o GENMOD com o algoritmo *EM* para ser aplicado em análises comparativas e contextuais, principalmente na área de demografia. O GENMOD foi escrito em FORTRAN77, podendo ser executado em MS-DOS. O procedimento de estimação é a máxima verossimilhança restrita. Naquele mesmo ano, escreveram a versão 2.1 do HLM, utilizando o algoritmo *EM* com acelerador de *Aitkin*.

No início dos anos 90, desenvolveram o MLN3 como parte do projeto de modelos multiníveis do Instituto de Educação da Universidade de Londres. Subseqüentemente, pacotes estatísticos maiores, como SAS e GENSTAT, vêm incorporando os procedimentos de máxima verossimilhança (*ML*) e máxima verossimilhança

restrita (*REML*) para modelos com respostas normais. Os pacotes EGRET e SABRE, por sua vez, possuem procedimentos que ajustam os modelos *logit* com 2 níveis, enquanto o BUGS, um pacote Bayesiano, utilizando o *Gibbs Sampling*, é capaz de ajustar os mesmos modelos multiníveis do MLN3 (WOODHOUSE, 1992).

O MLwiN, que foi desenvolvido a partir do MLn , tendo como precursor o ML3, realiza diferentes análises de modelos multiníveis, tem uma interface gráfica e apresenta facilidades de diagnóstico e de manipulação dos dados. O método de estimação dos parâmetros é realizado através do algoritmo *IGLS* (mínimos quadrados generalizados iterativos) e o *RIGLS*, que é chamado de máxima verossimilhança restrita. Além disso, foram incorporados ao MLwiN diagnósticos gráficos para os modelos Bayesianos, que utilizam o método Monte Carlo com cadeia de Markov (*MCMC*), bem como foi implementado o bootstrap iterativo para a estimação dos modelos lineares generalizados multiníveis (GOLDSTEIN, 1995).

CONSIDERAÇÕES FINAIS

Recentemente, os modelos lineares hierárquicos ou multiníveis tornaram-se mais conhecidos em diversos campos do conhecimento científico, principalmente nas áreas da Educação, das Ciências Sociais e da Epidemiologia. Os pesquisadores dessas áreas passaram a sentir necessidade de considerar, em seus estudos, diferentes níveis de análise: individual e de agregado, pois os determinantes de um evento que podem afetar um indivíduo são processos dinâmicos que poderão ser melhor investigados se usarmos modelos que permitam incorporar variáveis explicativas em ambos os níveis hierárquicos.

Além disso, os modelos lineares hierárquicos possuem importantes propriedades. A primeira delas permite que a variabilidade da variável resposta nos diferentes níveis hierárquicos possa ser explicada através de variáveis preditoras incluídas no modelo nesses diferentes níveis. A segunda propriedade dos modelos hierárquicos permite quantificar quanto da variabilidade da resposta se deve a cada nível, tal que a proporção da variabilidade explicada possa ser comparada diretamente. Ao

mesmo tempo, as análises hierárquicas são mais fidedignas ao não assumirem erroneamente o pressuposto de independência entre as observações das unidades pertencentes a uma estrutura maior, como na análise contextual (KREFT & LEEU, 1998), e nem desperdiçarem informação ao tomarem as médias das medidas, como na análise com dados agregados.

Por fim, espera-se que essas informações a respeito dos modelos multiníveis tenham mostrado o potencial e a utilidade destes modelos nas análises de diferentes áreas do conhecimento e a aplicabilidade dos mesmos em situações de pesquisa que resultem em dados com estrutura multinível ou hierárquica.

REFERÊNCIAS BIBLIOGRÁFICAS

BRYK, A.S.; RAUDENBUSH, S.W. **Hierarchical linear models**. Chicago: Sage Publications, 1992.

GOLDSTEIN, H. **Multilevel statistical models**. 2. ed. London: Institute of Education, University of London, 1995.

KREFT, I.G.G.; LEEU, W.J. **Introducing multilevel modeling**. California: Sage Publications, 1998.

KREFT, I.G.G.; LEEU, W.J.; DLEEDEN, R.V.D. Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, MLN3, VARCL. **The American Statistician**, v. 48, n.4, nov., 1994.

LIANG, K.Y.; ZEGGER, S.L. Longitudinal data analysis using generalized linear models. **Biometrika**, n.73, p.13-22, 1986.

McCULLAGH, P.; NELDER, J.A. **Generalized linear models**. London: Chapman and Hall, 1989.

PAULA, G.A. **Modelos de regressão**. São Paulo: Instituto de Matemática e Estatística, Universidade de São Paulo. (Apostila de curso, jul., 1997).

WEDDERBURN, R.W.M. Quasi-likelihood functions generalized linear models and the Gauss-Newton methods. **Biometrika**, n.61, p.439-66, 1974.

WOODHOUSE, G.A. **Guide to ML3 for New Users**. 2 ed. London: Ed. G. Woodhouse, Multilevel Models Project, Institute of Education, University of London, 1992.