Uma aplicação dos Modelos Lineares Generalizados Hierárquicos duplos em dados longitudinais de contagem com excesso de zeros

Nívea B. da Silva - IMECC/UNICAMP [1][2]
Rosemeire L. Fiaccone - IM/UFBA[2]
Leila Denise A. F. Amorim - IM/UFBA[2]

1 Introdução

Modelar dados de contagem é bastante comum nas diversas áreas do conhecimento. Em aplicações biomédicas, no entanto, a estrutura destes dados apresenta muitas vezes excesso de zeros e/ou super-dispersão, também chamada de extra-variabilidade, que são dois problemas que tipicamente ocorrem em dados de contagem.

O excesso de zeros em dados de contagem pode ocorrer tanto em estudos transversais, onde a unidade de investigação é medida uma única vez, como em estudos longitudinais, onde o grande desafio é a estrutura de dependência entre observações repetidas realizadas em uma mesma unidade de investigação (Fiaccone, 2006).

O modelo clássico de Regressão de Poisson é frequentemente útil em descrever a média μ_i, contudo, ele subestima a variância dos dados quando há extra-variabilidade nos mesmos. Não levar em consideração o excesso de zeros e/ou a extra-variabilidade nos dados, pode levar à estimação incorreta dos erros-padrão, e consequentemente a uma avaliação incorreta da significância dos parâmetros da regressão individual (Lee et. al, 2006).

Tal problema pode ser resolvido estendendo-se o modelo clássico de regressão de Poisson, ou seja, é possível utilizar, por exemplo, um estimador sanduíche ou ainda estimar adicionalmente um parâmetro de dispersão. Outra forma seria usar o modelo de regressão binomial negativo. Embora tais modelos sejam capazes de capturar a super-dispersão nos dados, eles muitas vezes não são suficientes para modelar o excesso de zeros (Zeileis, et. al, 2008).

Existem na literatura diversas metodologias que lidam com o problema de excesso de zeros nos dados. Dentre elas estão o modelo de Poisson inflacionado de zero e o modelo "Hurdle". Neste trabalho o referido problema será abordado sob a ótica dos modelos lineares generalizados hierárquicos duplos (DHGLM), que permitem a adição de efeitos aleatórios em seus vários componentes. A aplicação do método será ilustrada com a análise de dados de um ensaio clínico duplo-cego, placebo controlado, realizado pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia, no período de dezembro de 1990 a dezembro de 1991, cujo objetivo foi avaliar o efeito da suplementação periódica de vitamina A sobre a morbidade e mortalidade em crianças menores de 5 anos (Barreto et al,1994).

---

[1] Mestranda do Programa de Pós-Graduação em Estatística-UNICAMP (bolsista CNPq)
[2] Contatos: nivea.bispo@gmail.com, fiaccone@ufba.br, leiladen@ufba.br

2 Metodologia

2.1 Dados de contagem com excesso de zero

Dados expresssos como contagem contabilizam o número de vezes que certo evento ocorre em um determinado período de tempo. Tal evento é uma realização de uma variável aleatória que assume distribuição de Poisson.

A distribuição de Poisson pressupõe igualdade entre sua média e variância, contudo, na prática a variância dos dados tende a ser muito maior que a média dos mesmos, fenômeno conhecido como super-dispersão. Outro fenômeno bastante comum em dados de contagem é o excesso de zeros, muitas vezes ocasionado por uma combinação dos chamados zeros estruturais e zeros amostrais, favorecendo, assim, a presença de um número de zeros muito maior do que se esperaria em uma distribuição de Poisson (Costa, 2003).

O método padrão para modelar respostas do tipo contagem é a regressão de Poisson, porém ele subestima a variância dos dados na presença dos fenômenos anteriormente citados. Existem na literatura diversos métodos que lidam com o problema de excesso de zeros. Lambert (1992) propôs o modelo de Poisson inflacionado de zero (ZIP), e mais tarde, Ridout et al. (1998) apresentaram uma revisão dos modelos que se ajustam a dados de contagem com inflação de zeros. Contudo, tais modelos possuem limitações de uso para dados de contagem com estrutura longitudinal e/ou de *cluster*. Hur & Hall (2000) propuseram o modelo RE-ZIP que dá conta do efeito do *cluster*.

Na próxima sessão será apresentada uma classe de modelos que se ajusta bem ao problema de excesso de zero em dados de contagem com estrutura longitudinal.

2.2 DHGLM

Classe de modelos proposta por Lee e Nelder (2006) como uma extensão dos Modelos Lineares Generalizados Hierárquicos-MLGH (Lee e Nelder, 1996), onde efeitos aleatórios podem ser especificados para ambos os componentes do modelo (média e dispersão).

Nesta nova classe de modelos a heteroscedasticidade entre *clusters* pode ser modelada através da introdução de efeitos aleatórios no modelo de dispersão, como ocorre com a heterogeneidade no modelo para a média. Além disso, ela permite que inferências mais robustas sejam feitas sobre os valores discrepantes, permitindo assim distribuições com caudas mais pesadas.

Desta forma, os DHGLM podem ser decompostos dentro de um conjunto de MLG interligados, o que permite que uma grande variedade de modelos possa gerada, ajustada e comparada a partir de procedimentos iterativos de mínimos quadrados ponderados. A seguir a referida classe de modelos é sucintamente descrita.

Suponha que, condicional ao par de efeitos aleatórios (a, ν), a resposta y satisfaça as condições:

$$E(y|a,u) = \mu \quad \text{e} \quad Var(y|a,u) = \phi V(\mu),$$

onde ϕ é o parâmetro de dispersão e $V(\cdot)$ a função de variância.

A ideia chave desta nova classe de modelos é introduzir efeitos aleatórios no componente ϕ. Assim, temos:

(i) Dado 'u', o preditor linear para μ será um MLGH da forma:

$$\eta = g(\mu) = X\beta + Z\upsilon, \tag{1}$$

onde $g(\cdot)$ é uma função de ligação, X e Y são matrizes do modelo, υ são os efeitos aleatórios, com $\upsilon = g_m(\nu)$ para alguma função monótona $g_m(\cdot)$, e β são os efeitos fixos.

Além disso, o parâmetro de dispersão λ, para o efeito u, segue um MLGH da forma:

$$\xi_m = h_m(\lambda) = G_m\gamma_m, \tag{2}$$

onde $h_m(\cdot)$ é uma função de ligação, G_m é a matriz do modelo e γ_m são os efeitos fixos.

(ii) Dado 'a', o preditor linear para ϕ será um MLGH da forma:

$$\xi = h(\phi) = G_\gamma + Fb, \tag{3}$$

onde $h(\cdot)$ é uma função de ligação, G e F são matrizes do modelo, b são os efeitos aleatórios, com $b = g_d(a)$ para alguma função monótona $g_d(\cdot)$, e γ são os efeitos fixos.

O parâmetro de dispersão α, para o efeito $'a'$, segue um MLG da forma:

$$\xi_d = h_d(\alpha) = G_d\gamma_d, \tag{4}$$

onde $h_d(\cdot)$ é uma função de ligação, G_m é uma matriz do modelo e γ_d são os efeitos fixos.

É válido ressaltar que o número de componentes do MLG em (2) e (4) é igual ao número de componentes aleatórios em (1) e (3), respectivamente. Além disso, se $b = 0$ na equação (3), o DHGLM torna-se um MLGH com estrutura de dispersão.

2.2.1 Método de estimação

Para realizar inferências sobre os componentes do DHGLM, Lee e Nelder (2006) propuseram o uso da Verossimilhança H, que é definida como um caso especial da verossimilhança estendida. Este método provê um algoritmo simples, computacionalmente rápido e estatisticamente eficiente para ajuste dos MLGH e DHGLM.

► Verossimilhança H

Para estimação dos parâmetros no DHGLM, Lee e Nelder descreveram a verossimilhança H da seguinte forma:

$$h = l_0(y|\upsilon, b; \beta, \phi) + l_1(\upsilon; \lambda) + l_2(b; \alpha), \tag{5}$$

onde $l_0(y|\upsilon, b; \beta, \phi)$, $l_1 f(\upsilon; \lambda)$ e $l_2(b; \alpha)$ denotam, respectivamente, o log da verossimilhança para $y|(\upsilon, b)$, υ e b, sendo (λ, α) os parâmetros de dispersão.

Considere o seguinte modelo hierárquico para estimação dos parâmetros para os efeitos fixos, aleatórios e de dispersão, respectivamente:

$$y|\upsilon, b \sim f_1(y|\upsilon, b; \beta, \phi) \quad , \quad \upsilon \sim f_2(\upsilon; \lambda) \quad \text{e} \quad b \sim f_3(b; \alpha) \tag{6}$$

onde f_1, f_2 e f_3 são, respectivamente, densidades arbitrárias de $y|\upsilon$, υ e b.

A expressão (5) tem os componentes: $l_0(y|\upsilon, b; \beta, \phi) = \log f_1(y|\upsilon, b; \beta, \phi)$, $l_1(\upsilon; \lambda) = \log f_2(\upsilon; \lambda)$ e $l_2(b; \alpha) = \log f_3(b; \alpha)$, podendo, portanto, ser reescrita da seguinte forma:

$$h = l_0(y|\upsilon, b; \beta, \phi) + L_{\upsilon,b}$$

onde $L_{\upsilon,b}$ é a verossimilhança marginal, e pode ser obtida a partir da integral:

$$L_{\upsilon,b} = \log \int \exp(h) d\upsilon db = \log \int \exp(L_\upsilon) db = \log \int \exp(L_b) d\upsilon$$

sendo que $L_\upsilon = \log \int \exp(h) d\upsilon$ e $L_b = \log \int \exp(h) db$.

Um critério proposto por Lee e Nelder (2006) para estimação dos parâmetros no DHGLM é a utilização da verossimilhança H para inferências sobre υ, da verossimilhança marginal $L_{\upsilon,b}$ para β, e da verossimilhança restrita $p_\beta(L_\upsilon)$ para os parâmetros de dispersão. Porém, quando L_υ é numericamente díficil de ser obtida, os autores propõem o uso de verossimilhanças H perfiladas ajustadas, $p_\upsilon(h)$ e $p_{\beta,\upsilon}(h)$, como aproximações para L_υ e $p_\beta(L_\upsilon)$, respectivamente.

3 Aplicação e Resultados

Para efeito de ilustração, utilizaremos os dados de um ensaio clínico duplo-cego, placebo controlado, realizado pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia, no período de dezembro de 1990 a dezembro de 1991, cujo objetivo foi avaliar o efeito da suplementação periódica de vitamina A sobre a morbidade e mortalidade em crianças menores de 5 anos. As crianças foram designadas aleatoriamente aos grupos de vitamina A e placebo. Uma cápsula de placebo ou vitamina A foi oferecida às crianças a cada quatro meses durante um ano. Definiu-se como diarréia o registro de três ou mais dejeções líquidas e/ou amolecidas em um período de 24 horas, e delimitou-se como um novo episódio de diarréia o intervalo de três ou mais dias sem diarréia (Barreto et al,1994).

Neste trabalho a resposta de interesse representa o número de episódios severos de diarréia ocorridos na criança i, no período t, $t = 1, 2, 3$. Esse período representa o espaçamento entre duas suplementações ocorridas de 4 em 4 meses, o que caracteriza um evento recorrente.

Nesta aplicação o modelo DHGLM será utilizado como estratégia na análise de dados de contagem com excesso de zero. O referido modelo permite que sejam adicionados efeitos aleatórios no componente de dispersão, dando conta da super-dispersão ocasionada pelo excesso de zeros nos dados. Portanto, o modelo (6), descrito inicialmente terá a forma:

$$y|\upsilon, b \sim Poisson(\exp(\upsilon + b + \beta x)) \quad , \quad \upsilon \sim Normal(0, \lambda) \quad \text{e} \quad b \sim Normal(0, \alpha)$$

Foram consideradas cinco covariáveis para ajuste do modelo: sexo, idade, grupo de tratamento (Vitamina A ou Placebo), ausência de banheiro na residência e ausência de tratamento de água. O *software* estatístico GENSTAT (versão 9.0) foi utilizado para ajuste do modelo proposto.

A Tabela 1 apresenta resultados preliminares para o ajuste dos modelos HGLM e DHGLM. Os dois modelos foram ajustados para avaliar a importância ou não de se incluir um componente de dispersão no ajuste. É possível notar que as estimativas obtidas nos dois modelos são bem similares, contudo, para o modelo DHGLM os erros-padrão estimados tendem a ser um pouco menores. A diferença na deviance entre os modelos, para a ausência de $\alpha = Var(b_i) = 0$ é 41,49. Tal diferença aponta que o componente de dispersão b_i é necessário, e sua adição o reflete o efeito do seu impacto

sobre as estimativas dos parâmetros, indicando assim, a importância de se ajustar um modelo que incorpore o efeito ocasionado pelo excesso de zeros na resposta analisada.

Tabela 1: Estimativas dos parâmetros baseadas na verossimilhança H para o número de episódios severos de diarréia

Parâmetros	HGLM		DHGLM	
	est.	ep	est.	ep
β_0	0.544	0.103	0.5436	0.102
β_1 (sexo:M)	-0.016	0.071	-0.015	-0.070
β_2 (grupo: vit A)	-0.093	0.071	-0.093	-0.070
β_3 (idade)	-0.042	0.003	-0.042	0.003
β_4 (banheiro:N)	0.427	0.078	0.427	0.077
β_5 (água:N)	0.212	0.102	0.212	0.102
$log(\phi)$ (cte.)	-0.010	0.029	-0.013	0.028
$log(\lambda)$ (cte.)	-0.460	-0.063	-0.459	0.063
$log(\alpha)$ (cte.)			-3.69	3.61
$-2p_{\upsilon,b,\beta}(h)$	6835.934		6877.424	

4 Considerações Finais

Neste trabalho um procedimento alternativo foi apresentado para lidar com dados de contagem contendo excesso de zeros. A proposta do trabalho foi apresentar a classe de modelos lineares generalizados generalizados hierárquicos duplos, que permite que efeitos aleatórios possam ser especificados em ambos os componentes do modelo: média e dispersão. Assim, a heterogenidade entre *clusters* pode ser modelada através da introdução de efeitos aleatórios no modelo de dispersão, dando conta, assim, do problema de excesso de zeros.

Existem na literatura metodologias (modelos ZIP e "Hurdle") que lidam com o referido problema, contudo, elas possuem limitações de uso quando os dados possuem uma estrutura longitudinal e/ou de *cluster*.

Referências

[1] BARRETO, M.L., SANTOS, L.M.P., ASSIS, A.M.O., ARAÚJO, M.P.N. FARENZENA, G. G., SANTOS, P.A.G., FIACCONE, R.L. Effect of vitamin A suplementation on diarrhoea and acute lower respiration infection in young children in Brazil. *Lancet*, **344**, 228 − 231, 1994.

[2] COSTA, S. C. *Modelos Lineares Generelizados Mistos para dados longitudinais*. Tese (Doutorado), ESALQ, 2003.

[3] FIACCONE, R. L. *Modelling Multivariate Binary and Count Data, with Application to Infant Diarrhoea in Brazil*. Tese (Doutorado em Estatística), Lancaster University, 2006.

[4] LEE, Y., NELDER, J.A. & NOH M. H-likelihood: problems and solutions. *Statistics and Computing*, **17**, 49 − 55, 2007.

[5] LEE, Y. & NELDER, J.A. Double hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**, 1 − 29, 2006.

[6] LEE, Y., NELDER, J.A. & PAWITAN, Y. Generalized Linear Models with random effects - Unified Analysis via H-Likelihood. Chapman & Hall/CRC, 2006.

[7] LEE, Y., NELDER, J.A. Hierarchical generalized linear models. *Journal of the Royal Statistical Society*, **58**, 681 − 694, 1996.

[8] ZEILEIS, A., KLEIBER, C. & JACKMAN, S. Regression Models for Count Data in R. *Journal of Statistical Software)*, **17**, 1 − 21, 2008.