Considerações sobre a inclusão de efeitos aleatórios em modelos de transição de Markov

Idemauro Antonio Rodrigues de Lara
Departamento de Estatística, Universidade Federal do Paraná
Clarice Garcia Borges Demétrio
Departamento de Ciências Exatas, ESALQ-Universidade de São Paulo
Sílvia Emiko Shimakura
Departamento de Estatística, Universidade Federal do Paraná

1 Introdução

Como se sabe, a escolha do modelo para a análise de dados longitudinais deve levar em conta a natureza da variável resposta e os objetivos do estudo. Nesse contexto, a literatura freqüentemente apresenta os modelos marginais, de transição e de efeitos aleatórios como três propostas independentes e de objetivos específicos. Uma discussão comparativa desses modelos pode ser vista em Zeger e Liang (1992). Embora a interpretação dos parâmetros e inferências decorrentes difiram de um modelo para outro, na prática, essa segmentação estrutural dos três modelos não é tão restritiva quanto parece.

Considere um estudo longitudinal e seja, $\boldsymbol{y}_i$ o vetor de variáveis respostas do i-ésimo indivíduo, de dimensões $n_i \times 1$, isto é, $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$, de tal forma que numa ocasião t, a observação y_{it} esteja associada um vetor de variáveis explicativas $(p \times 1)$. De acordo com Diggle et al. (2002), um modelo de transição de Markov especifica um modelo linear generalizado para a distribuição condicional de Y_{it} dadas as respostas passadas e um conjunto de covariáveis. Seja $\boldsymbol{h}_{it} = (y_{i1}, y_{i2}, \ldots, y_{i,(t-1)})$ o vetor de dimensão q das respostas prévias e $\boldsymbol{x}_{it} = (x_{it1}, \ldots, x_{itp})'$ o vetor de variáveis explicativas, um modelo de transição caracteriza-se por:

$$g(\mu_{it}^C) = \eta_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(\boldsymbol{h}_{it}; \boldsymbol{\alpha}) \quad \text{e} \quad v_{it}^C = v(\mu_{it}^C)\phi, \tag{1}$$

em que $g(\mu_{it}^C)$ e $v(\mu_{it}^C)$ denotam a função de ligação e a função de variância, respectivamente. As respostas prévias ou funções delas mesmas são tratadas como variáveis explicativas adicionais. Os parâmetros de interesse, a priori, podem ser representados pelo vetor $\boldsymbol{\delta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$, em que $\boldsymbol{\beta}$ está associado às covariáveis e $\boldsymbol{\alpha}$ às respostas prévias. Logicamente, assim como no caso dos modelos marginais, esses parâmetros referem-se a efeitos fixos e em decorrência disso têm interpretações para a média populacional, isto é, são modelos da classe PA (*population-averaged*).

No entanto, não há razões para se restringir apenas aos efeitos fixos, nem ao menos para se escolher entre o uso do modelo de tranisção ou de efeitos aleatórios, quando se sabe que o modelo ideal para a interpretação dos resultados é aquele que considera na estrutura do preditor linear (1) a propriedade markoviana com a inclusão de fatores aleatórios convenientes. Apesar da idéia de modelos de transição com efeitos aleatórios não ser usual (em geral são modelos definidos para efeitos fixos), essa técnica foi considerada sob o enfoque de séries temporais, nos trabalhos de Korn e Whittemore (1979) e Stiratelli, Laird e Ware (1984). Nesse trabalho considera-se a estrutura de medidas repetidas com variáveis categorizadas e propõe-se um modelo de transição com a inclusão de efeitos aleatórios.

2 Métodos

2.1 Definição do modelo de transição misto

Sob um modelo de transição de Markov misto, as variáveis aleatórias Y_i condicionadas à história do processo ($\boldsymbol{h}_i$) e aos efeitos aleatórios ($\boldsymbol{d}_i$) seguem um modelo linear generalizado. A parte sistemática do modelo inclui as respostas prévias como efeitos fixos, variáveis explicativas associadas exclusivamente aos efeitos aleatórios ou a efeitos fixos e aleatórios. O intercepto do modelo também pode ser um parâmetro fixo ou aleatório, dependendo das pressuposições estabelecidas. Assim, de um modelo geral, para uma cadeia de ordem q, tem-se como preditor linear:

$$\eta_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(\boldsymbol{h}_{it}; \boldsymbol{\alpha}) + \boldsymbol{z}'_{it}\boldsymbol{d}_i,$$

em que $\boldsymbol{\beta}, \boldsymbol{\alpha}$ são os parâmetros associados aos efeitos fixos, $\boldsymbol{z}_{it}$ é a i-ésima linha da matriz do modelo associada ao vetor de efeitos aleatórios, $\boldsymbol{d}_i(k \times 1)$. Sendo usual a escolha da distribuição normal multivariada para esses efeitos aleatórios, ou seja, $\boldsymbol{d}_i \sim N_k(\boldsymbol{0}, \boldsymbol{G})$.

2.2 Aspectos computacionais

A implementação computacional para o ajuste de um modelo de transição misto requer que estrutura estocástica no preditor linear, a função de ligação e a distribuição dos efeitos aleatórios sejam corretamente especificados. Dessa forma, os programas desenvolvidos nos *softwares* estatísticos para o ajuste de modelos generalizados mistos, podem ser adaptados para o modelo de transição, a fim de se estimarem os parâmetros referentes aos efeitos fixos e aleatórios, bem como a variância desses efeitos. Como uma das opções computacionais disponíveis, para sua implementação no SAS, pode-se combinar a macro **DROPOUT** (Molenberghs e Verbeke, 2005), para o arranjo do conjunto de dados de acordo com a ordem de cadeia necessária e, a macro **GLIMMIX**, para o ajuste do modelo de transição misto. Nesse caso, a integral da função de verossimilhança é feita a partir da aproximação dos dados. Nos procedimentos NLMIXED e no glmML do R, a solução da integral se dá por quadratura gaussiana (Lara, 2007).

3 Exemplo de motivação

Considere um estudo longitudinal sobre doença respiratória, em que os indivíduos em cada uma das ocasiões são qualificados em bom (1) ou ruim (0) quanto a sua condição de saúde, tendo ainda associadas as covariáveis de tratamento, idade e sexo. O objetivo é estimar as probabilidades de transição, que são as probabilidades de os indivíduos passarem de uma categoria para a outra em ocasiões sucessivas, bem como avaliar a influência do tratamento e das demais covariáveis sob essas probabilidades. Sob um modelo de transição de Markov misto, pode-se considerar que o efeito do tratamento tem o mesmo peso nas probabilidades de transição para o estado de doença de qualquer paciente, assim como o estado do indivíduo na ocasião $(t-1)$, contribui com peso fixo para o estado do indivíduo na ocasião t. Entretanto, pode-se considerar que cada indivíduo tem uma propensão para a doença, refletindo suas predisposições genéticas e influências não mensuráveis de fatores ambientais. Assim, as probabilidades de transição são determinadas não somente por efeitos fixos, mas também por um componente aleatório (intercepto). Assumindo dependência de ordem 1, o modelo logístico de transição misto tem, então, a seguinte estrutura funcional:

$$\eta_{it} = (\beta_0 + d_{i0}) + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha y_{i(t-1)}, \tag{2}$$

em que $\boldsymbol{x}'_{it}$ engloba a covariável referente ao efeito de tratamento e $y_{i(t-1)}$ é a covariável da suposição de Markov. O termo d_{i0} representa a propensão individual para a probabilidade de doença respiratória. Da mesma forma poderíamos considerar a inclusão dos efeitos de sexo e idade no modelo. Para simplificar a notação, considere $\boldsymbol{\delta}$ o vetor de parâmetros de efeitos fixos, incluindo $\boldsymbol{\beta}, \boldsymbol{\alpha}$ e $\boldsymbol{x}^*_{it}$ a i-ésima linha da matriz de planejamento associada a esses efeitos:

$$\eta_{it} = \boldsymbol{x}^{*'}_{it}\boldsymbol{\delta} + \boldsymbol{z}'_{it}\boldsymbol{d}_i.$$

A função de verossimilhança para $\boldsymbol{\delta}$ e $\boldsymbol{G}$ no caso estacionário é proporcional a:

$$L(\boldsymbol{\delta}, \boldsymbol{G}; \boldsymbol{y}) \propto \prod_{i=1}^{N} \int \prod_{t=2}^{n_i} [\mu_{it}(\boldsymbol{\delta}, \boldsymbol{G})]^{y_{it}} [1 - \mu_{it}(\boldsymbol{\delta}, \boldsymbol{G})]^{1-y_{it}} f(\boldsymbol{d}_i; \boldsymbol{G}) d(\boldsymbol{d}_i), \tag{3}$$

em que $\mu_{it}(\boldsymbol{\delta}, \boldsymbol{d}_i) = \mathrm{E}(Y_{it} \mid \boldsymbol{d}_i; \boldsymbol{\delta})$.

Nota-se que a forma da função de verossimilhança (3) é equivalente à definida para os modelos mistos, a menos pelo fato de que os produtórios não envolvem a primeira observação, devido à suposição da cadeia de Markov. Focalizando a atenção sob função de ligação canônica e supondo distribuição normal para o efeito aleatório, $\boldsymbol{d}_i$, a expressão (3), reduz-se a:

$$\prod_{i=1}^{N} \int \exp\left[\boldsymbol{\delta}' \sum_{t=2}^{n_i} \boldsymbol{x}^*_{it} y_{it} + \boldsymbol{d}'_i \sum_{t=2}^{n_i} \boldsymbol{z}_{it} y_{it} - \sum_{t=2}^{n_i} \ln\{1 + \exp(\boldsymbol{x}^{*'}_{it}\boldsymbol{\delta} + \boldsymbol{z}'_{it}\boldsymbol{d}_i)\}\right] (2\pi)^{-1} |\boldsymbol{G}|^{-1/2} \exp\left(\frac{-\boldsymbol{d}'_i \boldsymbol{G}^{-1} \boldsymbol{d}_i}{2}\right) d(\boldsymbol{d}_i),$$

em que $\boldsymbol{G}$ é matriz de covariâncias dos efeitos aleatórios, $\boldsymbol{d}_i$.

Uma vez ajustado o modelo de transição com intercepto aleatório, estimam-se as probabilidades de transição como nos modelos com efeitos fixos, ou seja, por meio de:

$$\pi_{ab}(t) = \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})},$$

Matricialmente, pode-se escrever:

$$\boldsymbol{P}(t) = \begin{pmatrix} \pi_{00}(t) = \dfrac{1}{1 + \exp(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta})} & \pi_{01}(t) = \dfrac{\exp(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta})}{1 + \exp(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta})} \\ \pi_{10}(t) = \dfrac{1}{1 + \exp(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha)} & \pi_{11}(t) = \dfrac{\exp(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha)}{1 + \exp(\beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \alpha)} \end{pmatrix},$$

em que $\boldsymbol{x}'_{it}\boldsymbol{\beta}$ representa o efeito das variáveis explicativas referentes aos efeitos fixos, exceto os termos intercepto (efeito aleatório) e covariável da suposição de Markov.

4 Considerações finais

O modelo de transição misto pode ser útil em situações em que se deseja incluir efeitos aleatórios pertinentes no modelo. Certamente, isso muda o foco da interpretação dos resultados, uma vez que o modelo usa tanto a informação da resposta média populacional como a distribuição imposta para os efeitos aleatórios para se estimarem os parâmetros desejados. A própria matriz de probabilidades de transição, nesse sentido, poderia ser interpretada individualmente, ou seja, uma matriz para cada indivíduo. Evidentemente, se faz necessário avaliar a necessidade da inclusão de um efeito aleatório ou ainda testar a existência desse efeito no modelo. Um teste de razão de verossimilhanças pode ser feito para discriminar entre o modelo de efeitos fixos e aleatórios. A hipótese nula nesse caso será de que o modelo tem somente efeitos fixos.

5 Referências Bibliográficas

DIGGLE, P.J.; HEAGERTY, P.J.; LIANG, K.Y.; ZEGER, S.L. **Analysis of longitudinal data**. New York: Oxford University Press, 2002, 379 p.

KORN, E.L.; WHITTEMORE, A.S. Methods for analysing panel studies of acute health effects of air pllution. **Biometrics**, Washington, v. 35, p. 715 − 802, 1979.

LARA, I.A.R. **Modelos de transição para dados binários**. 2007, 111 p. Tese (Doutorado em Estatística e Experimentação Agronômica), Escola Superior de Agricultura Luiz de Queiroz-USP, Piracicaba, 2007.

MOLENBERGHS, G. VERBEKE, G. **Models for discrete longitudinal data**. New York: Springer-Verlag, 2005, 683 p.

STIRATELLI, R.; LAIRD, N.; WARE, J.H. Random effects-models for serial observations with binary response. **Biometrics**, Washington, v. 40, p. 961 − 971, 1984.

ZEGER, S.L.; LIANG, K.Y. An overview of methods for the analysis of longitudinal data. **Statistics in Medicine**, Chichester, v. 11, p. 1825 − 1839, 1992.