CONSIDERAÇÕES SOBRE A INCLUSÃO DE EFEITOS ALEATÓRIOS EM MODELOS DE TRANSIÇÃO DE MARKOV

Idemauro Antonio Rodrigues de LARA [1]
Clarice Garcia Borges DEMÉTRIO[2]
Sílvia Emiko SHIMAKURA [3]

▪ RESUMO:
Nesse trabalho é considerada a estrutura de medidas repetidas no tempo com variáveis binárias e apresenta-se um modelo de transição misto. O método de ajuste é baseado na teoria da máxima verossimilhança e foi implementado no *software* R. Verificou-se, por meio de simulação, que há uma tendência de que esses modelos "escondam" a existência de efeitos aleatórios quando o número de ocasiões é limitado, levando a estimativas viciadas. O procedimento pode ser útil em situações em que se deseja adicionar efeitos aleatórios ao modelo de transição, unificando duas classes de modelos e permitindo a interpretação das matrizes de probabilidades de transição em termos individuais.

▪ PALAVRAS-CHAVE:
dados longitudinais; modelos de transição; efeitos aleatórios; máxima verossimilhança

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.

[1]Departamento de Estatística, UFPR E-mail: *idemauro@ufpr.br*
[2]Departamento de Ciências Exatas, ESALQ/USP E-mail: *clarice@esalq.usp.br*
[3]Laboratório de Estatística e Geoinformação, DEST-UFPR E-mail: *silvia.shimakura@ufpr.br*

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*).

Quando se trabalha com variáveis categorizadas um dos objetivos é especificar as probabilidades de transição entre as categorias de resposta. Em particular, para dados binários assumindo alcance 1 para a cadeia e ligação canônica, tem-se:

$$\pi_{ab}(t) = \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})} \qquad \text{com} \;\; a, b \in \{0, 1\},$$

avaliando-se η_{it} primeiramente com relação a $Y_{i(t-1)} = 0$ e, posteriormente a $Y_{i(t-1)} = 1$, tem-se:

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

em que $\pi_{00}(t) = 1 - \pi_{01}(t)$ e $\pi_{10}(t) = 1 - \pi_{11}(t)$. Maiores detalhes sobre matrizes de probabilidades de transição bem como referências sobre o assunto podem ser encontradas em Lara et. al (2007).

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). Adicionalmente, há de se considerar o problema do uso de variáveis dicotômicas. Esses dados são usais em estudos longitudinais sobretudo pela facilidade de interpretação dos parâmetros e, obtenção da razão de chances associadas a fatores desejáveis. Apesar dessas vantagens, dados binários, muitas vezes, podem gerar problemas, como por exemplo, o excesso de zeros ou até omitir informações importantes em face da sua natureza. Nesse trabalho considera-se a estrutura de medidas repetidas com variáveis binárias e propõe-se um modelo de transição com a inclusão de efeitos aleatórios.

2 Material e Métodos

2.1 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. Nessa estrutura, a correlação entre as respostas para um indivíduo é resultante da heterogeneidade natural nos coeficientes de regressão e da história dos indivíduos.

Lara (2007) aplica essa estrutura a um conjunto de dados reais, considerando alcances 1 e 2 em processos estacionários. No entanto, os modelos ajustados não acusaram a existência de efeitos aleatórios, orientando uma discussão sobre o assunto. Por hipótese, acredita-se que a estrutura do dado binário combinada ao número de ocasiões do estudo corroboraria para camuflar a detecção do efeito aleatório, sinalizando a necessidade de se desenvolverem estudos por simulação.

2.2 Métodos

2.2.1 Definição do modelo de transição misto e estimação

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:

$$\eta_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(\boldsymbol{h}_{it}; \boldsymbol{\alpha}) + \boldsymbol{z}'_{it}\boldsymbol{d}_i, \tag{3}$$

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, $\boldsymbol{d}_i \sim N_k(\boldsymbol{0}, \boldsymbol{G})$.

Para simplificar a notação (3), 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:

$$L(\boldsymbol{\delta}, \boldsymbol{G}; \boldsymbol{y}) \propto \prod_{i=1}^{N} \int \prod_{t=2}^{n_i} [\mu_{it}(\boldsymbol{\delta}, \boldsymbol{G})]^{y_{it}} [1 - \mu_{it}(\boldsymbol{\delta}, \boldsymbol{G})]^{1-y_{it}} f(\boldsymbol{d}_i; \boldsymbol{G}) d(\boldsymbol{d}_i), \tag{4}$$

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 (4) 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 (4), reduz-se a:

$$\prod_{i=1}^{N} \int \exp\left[\boldsymbol{\delta}' \sum_{t=2}^{n_i} \boldsymbol{x}^*_{it} y_{it} + \boldsymbol{d}'_i \sum_{t=2}^{n_i} \boldsymbol{z}_{it} y_{it} - \sum_{t=2}^{n_i} \ln\{1 + \exp(\boldsymbol{x}^{*'}_{it}\boldsymbol{\delta} + \boldsymbol{z}'_{it}\boldsymbol{d}_i)\}\right]$$
$$(2\pi)^{-1} |\boldsymbol{G}|^{-1/2} \exp\left(\frac{-\boldsymbol{d}'_i \boldsymbol{G}^{-1} \boldsymbol{d}_i}{2}\right) d(\boldsymbol{d}_i), \tag{5}$$

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

O problema para maximizar a função (5) é a presença das N integrais sob os efeitos aleatórios, que em geral, não podem ser resolvidas analiticamente. Faz-se necessário usar procedimentos de aproximações numéricas. Nesse trabalho, optou-se pelo uso da técnica baseada na aproximação dos integrandos, ou simplesmente aproximação de Laplace, devido à otimização do custo operacional das simulações. A lógica desse método é aproximar integrais na forma:

$$I = \int e^{Q(\boldsymbol{\gamma})} d\boldsymbol{\gamma}, \tag{6}$$

quando $Q(\boldsymbol{\gamma})$ é uma função conhecida e unimodal de uma variável q-dimensional, $\boldsymbol{\gamma}$, trocando na expressão (6), $Q(\boldsymbol{\gamma})$, por uma forma aproximada, obtida por intermédio de uma expansão de segunda ordem em série de Taylor dessa função ao redor de sua moda, $\hat{\boldsymbol{\gamma}}$. Como cada integral na função (5) é proporcional a uma integral da forma (6) o método de Laplace pode ser aplicado. Uma discussão mais detalhada dessa técnica bem como de outros procedimentos para maximização da função de verossimilhança podem ser encontrados em Molenberghs e Verbeke (2005).

2.2.2 Aspectos computacionais e estudo por simulação

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. Nesse trabalho a implementação computacional foi feita no *software* R, adotando-se o modelo:

$$\eta_{it} = (\beta_0 + d_{i0}) + \beta x_{it} + \alpha y_{i(t-1)}, \tag{7}$$

assumindo estacionariedade, alance 1 para a cadeia e assumindo vetor de parâmetros: $\boldsymbol{\delta} = (-2,5; 1,0; 2,0)$ e $d_i \sim N(0; 0,5)$.

Inicialmente desenvolveu-se uma função para organização do conjunto de dados na estrutura que permita incorporar ao predito linear funções das respostas prévias. A partir dessa estrutura, foram feitas 10.000 simulações considerando $t = 4, 5, 8$ e 12 ocasiões e $n = 500$ indivíduos. A seguir, ajustaram-se modelos de transição com interceptos aleatórios (estrutura real, equação 7) pela função *glmmML* e ignorando-se essa estrutura pela função *geeglm* (modelos somente com efeitos fixos). O teste da razão de verossimilhanças e o critério da informação de Akaike (AIC) foram usados para discriminar entre esses dois modelos. A hipótese nula nesse caso é de que o modelo tem somente efeitos fixos.

3 Resultados e Discussão

Os resultados obtidos por simulação comprovaram a hipótese de que a estrutura dos modelos de transição com dados binários pode esconder a presença de efeitos aleatórios, sobretudo com um número reduzido de ocasiões, como mostra a tabela 2. Com 12 ocasiões, o efeito aleatório é detectado em aproximadamente 97% dos casos.

Tabela - 1: Erro quadrático médio das estimativas dos parâmetros com base nos dados simulados, considerando modelos somente com efeitos fixos (Modelo 1) e mistos (Modelo 2) e 4 diferentes números de possibilidades de ocasiões.

	$t = 4$	
Parâmetros	Modelo 1	Modelo 2
β_0	0,01896	0,02364
β_1	0,02749	0,02905
α	0,05632	0,04858
	$t = 5$	
Parâmetros	Modelo 1	Modelo 2
β_0	0,01515	0,01583
β_1	0,02311	0,02221
α	0,05317	0,03952
	$t = 8$	
Parâmetros	Modelo 1	Modelo 2
β_0	0,01070	0,00847
β_1	0,01783	0,01339
α	0,05040	0,02031
	$t = 12$	
Parâmetros	Modelo 1	Modelo 2
β_0	0,00825	0,00533
β_1	0,01573	0,00913
α	0,04926	0,01011

Tabela - 2: Percentual do número de vezes em que o AIC do Modelo 1 foi maior do que o do Modelo 2 de acordo com o número de ocasiões.

	Número de ocasiões			
	4	5	8	12
Percentual	$12,29\%$	$18,66\%$	$62,59\%$	$96,93\%$

Por outro lado, com relação aos vícios dos estimadores, verificou-se uma

tendência dos modelos em subestimar o efeito da covariável X e superestimar o efeito de $Y_{(t-1)}$, sendo esses vícios maiores para o Modelo 1, ou seja, ignorar a existência de efeito aleatório fatalmente acarretará em estimativas viciadas, que conseqüentemente vão distorcer as probabilidades de transição.

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. Em particular, com dados binários deve-se ficar atento para a possibilidade do efeito existir mas não ser detectado por algum critério em face de se ter um número restrito de ocasiões.

LARA, I.A.R.; DEMÉTRIO, C.G.B.; SHIMAKURA, S. E. Considerations on the inclusion of random effects in the transition models of Markov. *Rev. Bras. Biom.,* São Paulo, v.xx, n.x, p.xx-xx, 2009. *Rev. Mat. Estat.* (São Paulo), v. xx, n.x, p. 1-xx, 2009.

- *ABSTRACT: This work is considered the structure of repeated measures in time with binary variables and presents a model of mixed transition. The method of adjustment is based on the theory of maximum likelihood and was implemented in the software R. It was by means of simulation, there is a tendency of these models hide the existence of random effects when the number of occasions is limited, leading to biased estimates. The procedure can be useful in situations where you want to add random effects model of transition, merging two classes of models, allowing the interpretation of the matrices of transition probabilities in terms of individuals.*

- *KEYWORDS: longitudinal data; transition models; random effetcts; maximum likelihood*

Referências

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.

LARA, I. A. R.; DEMÉTRIO, C. G. B., ANDRADE, D. F., MOTA, J. M. A. Modelos de transição para dados binários: idéias básicas e testes para comparar tratamentos. *Revista Brasileira de Biometria*, Jaboticabal, v.25, n^0 4, p.77-100, 2007

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

R Development Core Team. *R: A language and environment for statistical computing 2.7.2*. Vienna, Austria, 2008. Disponível em <http://www.R-project.org>. Acesso em: 23 junho 2008.

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*, v. 11, Chichester, p. 1825 – 1839, 1992.

Recebido em xx.xx.200x.

Aprovado após revisão em xx.xx.200x.