ABORDAGEM BAYESIANA EM DADOS BINÁRIOS CORRELACIONADOS: UM ESTUDO LONGITUDINAL DA OCORRÊNCIA DE MENSTRUAÇÃO EM PACIENTES COM SÍNDROME DE OVÁRIOS POLICÍSTICOS APÓS TRATAMENTO

Jorge Alberto ACHCAR[1]
Edson Zangiacomi MARTINEZ[2]
Eliza OMAI[3]
Adriana de Fátima LOURENÇON[2]
Gleici Castro PERDONÁ[3]

▪ RESUMO: Neste artigo desenvolvemos uma análise de regressão logística com efeitos aleatórios considerando dados binários correlacionados longitudinais. A proposta é ilustrada com um exemplo real envolvendo a avaliação da ocorrência de menstruação em pacientes com síndrome de ovários policísticos após um tratamento (Penna et al., 2005, Human Reproduction v.20, n.9, p.2396-401). O conjunto de dados foi obtido de um estudo conduzido na Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo. A ocorrência ou não de menstruação foi observada em cada indivíduo em três instantes dentro de um intervalo de tempo. Conseqüentemente, espera-se uma correlação não nula entre as observações de um mesmo indivíduo, sendo este efeito capturado pela adição de um efeito aleatório no modelo. Analisamos os dados pelo modelo Bayesiano assumindo que os efeitos aleatórios têm distribuição normal. Em um segundo modelo, assumimos que os efeitos aleatórios são distribuídos segundo uma mistura de duas distribuições normais. Inferências para os parâmetros de interesse são baseadas em métodos Monte Carlo via cadeias de Markov (MCMC). Para comparações entre modelos utilizamos o *Deviance Information Criteria* (DIC).

▪ PALAVRAS-CHAVE: Dados binários correlacionados; regressão logística; análise Bayesiana; ensaios clínicos.

[1]Departamento de Estatística, Universidade Federal de São Carlos - UFSCar, CEP 13565-905, São Carlos, SP, Brasil. E-mail: *jachcar@power.ufscar.br*
[2]Centro de Métodos Quantitativos, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo - CEMEQ/FMRP/USP, CEP 14049-900, Ribeirão Preto, SP, Brasil. E-mail: *elizaomai@hotmail.com* / *aflourencon@yahoo.com.br*
[3]Departamento de Medicina Social, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo - FAEPA/FMRP/USP, CEP 14049-900, Ribeirão Preto, SP, Brasil. E-mail: *edson@fmrp.usp.br* / *pgleici@fmrp.usp.br*

1 Introdução

O presente estudo foi motivado pela análise de um banco de dados real, introduzido no Centro de Métodos Quantitativos (CEMEQ) da Faculdade de Medicina de Ribeirão Preto (FMRP/USP). Estes dados pertencem a um ensaio clínico controlado conduzido nesta instituição, que objetivou investigar o efeito da acarbose em dose moderada no padrão menstrual de pacientes com síndrome de ovários policísticos (SOP) e resistência à insulina (Penna et al., 2005). Foram selecionadas 30 pacientes elegíveis ao estudo segundo alguns critérios de inclusão, distribuídas aleatoriamente em dois grupos. As pacientes alocadas no primeiro grupo receberam tratamento com acarbose, e às alocadas no outro grupo foi administrado um placebo (um agente farmacologicamente inativo dado à paciente como um substituto a um agente ativo, para efeitos de comparação com o tratamento sob investigação). O ensaio foi conduzido de forma duplo-cega (Pocock, 1983), tal que o pesquisador e as pacientes desconheciam o tratamento administrado ao longo de seis meses de tratamento. Entretanto, três pacientes não aderiram ao estudo, o que resultou em um tamanho amostral efetivo de 27 mulheres.

A Tabela 1 apresenta os dados obtidos em cada período do estudo, em que foi observada a ocorrência ou não de menstruação em três períodos bimestrais (T_1, após dois meses de tratamento, T_2, realizada no quarto mês de tratamento, e T_3, realizada no sexto mês de tratamento).

Os dados apresentados na Tabela 1 sugerem uma maior frequência de menstruação para pacientes recebendo o tratamento com acarbose no sexto mês do estudo (85% das mulheres alocadas no grupo tratado menstruaram, contra 50% das mulheres do grupo que recebeu placebo). Para a modelagem destes dados, assumimos um modelo de regressão logística com efeitos aleatórios, dado por

$$P(Y_{ji} = y_{ji}) = p_{ji}^{y_{ji}}(1 - p_{ji})^{1-y_{ji}}, \tag{1}$$

onde y_{ji} é uma observação de Y_{ji}, uma variável binária tal que $y_{ji} = 1$ se observada ocorrência de menstruação para a i-ésima paciente no j-ésimo período, e $y_{ji} = 0$ caso contrário, dado que $j = 1, 2, 3$ corresponde às três avaliações (períodos T_1, T_2 e T_3, respectivamente), para $i = 1, 2, ..., 27$ (número de pacientes envolvidas no estudo), e

$$p_{ji} = \frac{\exp\{\alpha_i + \beta_{0j} + \beta_{1j}x_i\}}{1 + \exp\{\alpha_i + \beta_{0j} + \beta_{1j}x_i\}}, \tag{2}$$

sendo x_i uma observação de X_{ji}, uma variável indicadora associada ao tratamento realizado, tal que $x_i = 0$ se a i-ésima paciente foi alocada no grupo placebo e $x_i = 1$ se a i-ésima paciente recebeu tratamento com acarbose.

Observe que α_i em (2) denota um efeito aleatório para a i-ésima observação, capturando a possível correlação entre y_{1i}, y_{2i} e y_{3i} (ver, por exemplo, Dey e Chen, 1998, Prentice, 1988, ou Ochi e Prentice, 1984). Podemos assumir diferentes distribuições para o efeito aleatório α_i. Dey e Chen (1998) assumem efeitos

aleatórios α_i independentes, com distribuição normal com média zero e variância σ_α^2, isto é,

$$\alpha_i \sim N(0; \sigma_\alpha^2). \tag{3}$$

Tabela 1 – Dados referentes à ocorrência ou não de menstruação nos três períodos de avaliação, segundo o tratamento (acarbose ou placebo). A não ocorrência de menstruação é denotada por 0, e a ocorrência de menstruação é denotada por 1

Acarbose				Placebo			
Paciente	$T_1^{(a)}$	$T_2^{(b)}$	$T_3^{(c)}$	Paciente	$T_1^{(a)}$	$T_2^{(b)}$	$T_3^{(c)}$
1	0	0	1	14	1	1	0
2	1	0	0	15	1	1	1
3	0	0	0	16	0	1	1
4	1	0	1	17	1	1	1
5	0	1	1	18	0	0	0
6	0	0	1	19	1	1	0
7	1	1	1	20	1	1	1
8	0	0	1	21	0	0	1
9	1	1	1	22	1	1	1
10	1	1	1	23	0	1	1
11	1	1	1	24	0	0	0
12	0	1	1	25	1	0	0
13	0	0	1	26	0	0	0
				27	0	0	0
Total	6	6	11	Total	7	8	7
	(46%)	(46%)	(85%)		(50%)	(57%)	(50%)

(a)Avaliação realizada após dois meses de tratamento.
(b)Avaliação realizada no quarto mês de tratamento.
(c)Avaliação realizada no sexto mês de tratamento.

Achcar; Janeiro e Mazucheli (2003) assumem que os efeitos aleatórios α_i são independentes e apresentam uma mistura de distribuições normais dada por

$$\pi(\alpha_i) = \sum_{k=1}^{K} \lambda_k \phi_k(\alpha_i | \mu_k, \sigma_k^2), \tag{4}$$

onde $\sum_{k=1}^{K} \lambda_k = 1$, K denota o número de componentes da mistura e ϕ_k denota uma densidade normal $N(\mu_k, \sigma_k^2)$. Na análise de dados reais, o uso de uma mistura de distribuições normais para o efeito aleatório como alternativa ao pressuposto de normalidade apresentado em (3) pode melhorar o ajuste do modelo (Achcar; Janeiro; Mazucheli, 2003). Na análise de dados binários longitudinais, muitos autores buscam tratar o tempo decorrido desde a intervenção como uma variável contínua. Por exemplo, Heagerty (1999) propõe um modelo linear generalizado

de efeitos mistos com uma função de ligação que descreve o comportamento da probabilidade de sucesso de dados binários ao longo do tempo. Preisser et al. (2000) demonstraram a utilidade de modelos com estrutura *generalized estimating equations* (GEE) em descrever a redução das taxas de tabagismo em alguns grupos específicos de indivíduos longitudinalmente. Uma vantagem destes modelos que tratam o tempo de forma contínua é a possibilidade de inferências em instantes de tempos não observados no estudo. Entretanto, a determinação de uma função de ligação que descreva com satisfatória precisão as variações de p_{ij} ao longo do tempo, este tratado como uma variável contínua, é bastante limitada quando observamos a ocorrência ou não de uma variável binária em apenas três instantes fixos, tal como os dados de Penna et al. (2005) apresentados na Tabela 1. Por este motivo, optamos por tratar o fator tempo como uma variável discreta no modelo aqui apresentado. Outros modelos também úteis na análise de dados binários correlacionados foram propostos por Chib e Greenberg, 1998, e Albert e Jais, 1998.

Na análise Bayesiana de modelos de regressão para os dados binários correlacionados da Tabela 1, consideramos o uso de métodos Monte Carlo em Cadeias de Markov (MCMC) para simular amostras da distribuição *a posteriori* conjunta para os parâmetros de interesse (ver, por exemplo, Gelfand e Smith, 1990). Atualmente, observa-se um crescente uso de métodos Bayesianos na análise de dados de ensaios clínicos (são exemplos Thall; Simon e Shen (2000), e Shaffer e Chinchilli (2004)). Na seção 2, apresentamos um modelo assumindo uma distribuição normal para os efeitos aleatórios do modelo. Na seção 3, assumimos uma mistura de duas distribuições normais para os efeitos aleatórios. Resultados da aplicação destes modelos aos dados da Tabela 1 são apresentados na seção 4. Na seção 5 apresentamos algumas considerações finais.

2 Análise Bayesiana assumindo distribuição normal para os efeitos aleatórios α_i

A função de verossimilhança para $\alpha = (\alpha_1, ..., \alpha_n)'$ e $\beta = (\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \beta_{03}, \beta_{13})'$ é dada por

$$L(\alpha, \beta) = \prod_{i=1}^{n}\prod_{j=1}^{3} \frac{\exp\{(\alpha_i + \beta_{0j} + \beta_{1j}x_i)y_{ji}\}}{1 + \exp\{\alpha_i + \beta_{0j} + \beta_{1j}x_i\}}, \tag{5}$$

onde $n = 27$ é o tamanho amostral. Assumindo independência *a priori* entre os parâmetros, considere distribuições *a priori* normais para β_{lj}, dadas por

$$\beta_{lj} \sim N(a_{lj}; b_{lj}^2), \tag{6}$$

onde a_{lj} e b_{lj} são conhecidos, $l = 0, 1$ e $j = 1, 2, 3$.

Assumindo efeitos aleatórios com uma distribuição normal (3), consideramos uma distribuição *a priori* para σ_α^2 dada por

$$\sigma_\alpha^2 \sim IG(c; d), \tag{7}$$

onde c e d são constantes conhecidas e $IG(c;d)$ denota uma distribuição gama inversa com média $d/(c-1)$ e variância $d^2/[(c-1)^2(c-2)]$.

Combinando (5) com (3), (6) e (7), obtemos a distribuição *a posteriori* conjunta para α, β e σ_α^2 dada por

$$\begin{aligned}\pi(\alpha,\beta,\sigma_\alpha^2|\mathbf{y},\mathbf{x}) \quad \propto \quad & (\sigma_\alpha^2)^{-(c+1)}\exp\left\{-\frac{d}{\sigma_\alpha^2}\right\}\times \\ & \times\left\{\prod_{l=0}^{1}\prod_{j=1}^{3}\exp\left(-\frac{1}{2b_{lj}^2}\left(\beta_{lj}-a_{lj}\right)^2\right)\right\}\times \\ & \times\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left(-\frac{\alpha_i^2}{2\sigma_\alpha^2}\right)\right\}\times \\ & \times\left\{\prod_{i=1}^{n}\prod_{j=1}^{3}\frac{\exp\left[(\alpha_i+\beta_{0j}+\beta_{1j}x_i)y_{ji}\right]}{\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]}\right\}. \qquad (8)\end{aligned}$$

De (8), as distribuições condicionais *a posteriori* necessárias para o algoritmo de amostradores de Gibbs são dadas por

(i) $\pi(\alpha_i|\theta_{(\alpha_i)},\mathbf{y},\mathbf{x}) \propto \exp\left\{-\frac{\alpha_i^2}{2\sigma_\alpha^2}\right\}\psi_{1i}(\alpha,\beta)$,

onde $\psi_{1i}(\alpha,\beta)=\exp\left\{\alpha_i y_{\bullet i}-\sum_{j=1}^{3}\log\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]\right\}$,

$y_{\bullet i}=\sum_{j=1}^{3}y_{ji}$, $i=1,...,n$ e $\theta_{(\alpha_i)}$ denota o vetor de todos os parâmetros, exceto α_i;

(ii) $\sigma_\alpha^2|\theta_{(\sigma_\alpha^2)},\mathbf{y},\mathbf{x} \sim IG\left(\frac{n}{2}+c;d+\frac{1}{2}\sum_{i=1}^{n}\alpha_i^2\right)$,

$\theta_{(\sigma_\alpha^2)}$ denota o vetor de todos os parâmetros, exceto σ_α^2;

(iii) $\pi(\beta_{0j}|\theta_{(\beta_{0j})},\mathbf{y},\mathbf{x}) \propto \exp\left\{-\frac{1}{2b_{0j}^2}\left(\beta_{0j}-a_{0j}\right)^2\right\}\psi_{2i}(\alpha,\beta)$,

onde $\psi_{2i}(\alpha,\beta)=\exp\left\{\beta_{0j}y_{j\bullet}-\sum_{i=1}^{n}\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]\right\}$,

$y_{j\bullet}=\sum_{i=1}^{n}y_{ji}$, $j=1,2,3$ e $\theta_{(\beta_{0j})}$ denota o vetor de todos os parâmetros, exceto β_{0j};

(iv) $\pi(\beta_{1j}|\theta_{(\beta_{1j})},\mathbf{y},\mathbf{x}) \propto \exp\left\{-\frac{1}{2b_{1j}^2}\left(\beta_{1j}-a_{1j}\right)^2\right\}\psi_{3i}(\alpha,\beta)$,

onde $\psi_{3i}(\alpha,\beta)=\exp\left\{\beta_{1j}\sum_{i=1}^{n}x_i y_{ji}-\sum_{i=1}^{n}\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]\right\}$,

$j=1,2,3$ e $\theta_{(\beta_{1j})}$ denota o vetor de todos os parâmetros, exceto β_{1j}.

Observe que precisamos usar o algoritmo de Metropolis-Hastings (ver, por exemplo, Smith e Roberts, 1993) para gerar α_i e β_{lj}, $i = 1, ..., n$, $l = 0, 1$, $j = 1, 2, 3$, dado que suas distribuições condicionais *a posteriori* não assumem uma forma conhecida.

3 Análise Bayesiana assumindo uma mistura de duas distribuições normais para os efeitos aleatórios α_i

Assumindo efeitos aleatórios α_i com uma mistura de duas distribuições normais (4) e mesmas distribuições *a priori* para β_{lj} dadas em (6), considere as seguintes distribuições *a priori* para μ_k, σ_k^2 e λ_1, $k = 1, 2$:

$$\begin{array}{rcl} \mu_k & \sim & N(c_k; d_k^2),\ c_k \text{ e } d_k \text{ conhecidos;} \\ \sigma_k^2 & \sim & IG(e_k; f_k),\ e_k \text{ e } f_k \text{ conhecidos;} \\ \lambda_1 & \sim & B(g; h),\ g \text{ e } h \text{ conhecidos,} \end{array} \tag{9}$$

onde $B(g;h)$ denota uma distribuição Beta com média $g/(g+h)$ e variância $gh/[(g+h)^2(g+h+1)]$, e $\lambda_2 = 1 - \lambda_1$. Sendo $\mu = (\mu_1, \mu_2)'$, $\sigma^2 = (\sigma_1^2, \sigma_2^2)'$ e $\lambda = (\lambda_1, \lambda_2)'$, e assumindo independência *a priori* entre os parâmetros, a distribuição *a posteriori* conjunta para $\theta = (\alpha, \mu, \sigma^2, \lambda, \beta)$ é dada por

$$\pi(\theta|\mathbf{y}, \mathbf{x}) \propto \psi(\theta) \left\{ \prod_{i=1}^{n} \sum_{k=1}^{2} \lambda_k \phi_k \left(\alpha_i | \mu_k, \sigma_k^2\right) \right\}, \tag{10}$$

onde

$$\begin{aligned} \psi(\theta) \quad = \quad & \left\{ \prod_{k=1}^{2} \left(\sigma_k^2\right)^{-(e_k+1)} \exp\left(-\frac{f_k}{\sigma_k^2}\right) \right\} \times \left\{ \prod_{k=1}^{2} \exp\left(-\frac{1}{2d_k^2} \left(\mu_k - c_k\right)^2\right) \right\} \times \\ & \times \lambda_1^{g-1} \left(1 - \lambda_1\right)^{h-1} \left\{ \prod_{l=0}^{1} \prod_{j=1}^{3} \exp\left[-\frac{1}{2b_{lj}^2} \left(\beta_{lj} - a_{lj}\right)^2\right] \right\} \times \\ & \times \left\{ \prod_{i=1}^{n} \prod_{j=1}^{3} \frac{\exp\left[\left(\alpha_i + \beta_{0j} + \beta_{1j} x_i\right) y_{ji}\right]}{1 + \exp\left(\alpha_i + \beta_{0j} + \beta_{1j} x_i\right)} \right\}. \end{aligned} \tag{11}$$

Para simplificar as expressões das distribuições condicionais necessárias para o algoritmo de amostradores de Gibbs, podemos introduzir variáveis latentes (ver, por exemplo, Tanner e Wang, 1987), dadas por $\mathbf{Z}_i = (Z_{i1}, Z_{i2})$, $Z_{i1} + Z_{i2} = 1$, $i = 1, .., n$, onde Z_{i1} tem uma distribuição de Bernoulli com probabilidade de sucesso h_{i1} dada por

$$h_{i1} = \frac{\lambda_1 \phi_1 \left(\alpha_i | \mu_1, \sigma_1^2\right)}{\sum_{k=1}^{2} \lambda_k \phi_k \left(\alpha_i | \mu_k, \sigma_k^2\right)}. \tag{12}$$

Combinando (10) e (11), a distribuição *a posteriori* conjunta para θ é dada por

$$\pi(\theta|\mathbf{y},\mathbf{x},\mathbf{Z}) \propto \psi(\theta)\left\{\prod_{i=1}^{n}\prod_{k=1}^{2}\left[\lambda_k\phi_k\left(\alpha_i|\mu_k,\sigma_k^2\right)\right]^{Z_{ik}}\right\}, \tag{13}$$

onde $\psi(\theta)$ é dada em (11). A partir de (13), determinamos as distribuições condicionais necessárias para o algoritmo de amostradores de Gibbs, dadas por:

(i) $\pi(\alpha_i|\beta,\mu,\sigma^2,\lambda_1,\mathbf{y},\mathbf{x},\mathbf{Z}) \propto N\left(\frac{Z_{i1}\mu_1\sigma_2^2+Z_{i2}\mu_2\sigma_1^2}{Z_{i1}\sigma_2^2+Z_{i2}\sigma_1^2}, \frac{\sigma_1^2\sigma_2^2}{Z_{i1}\sigma_2^2+Z_{i2}\sigma_1^2}\right)\psi_{1i}(\alpha,\beta)$,

onde $\psi_{1i}(\alpha,\beta) = \exp\left\{\alpha_i y_{\bullet i} - \sum_{j=1}^{3}\log\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]\right\}$

e $y_{\bullet i} = \sum_{j=1}^{3} y_{ji}$, $i = 1,...,n$;

(ii) $\pi\left(\sigma_k^2|\alpha,\beta,\mu,\sigma_{(k)}^2,\lambda_1,\mathbf{y},\mathbf{x},\mathbf{Z}\right) \sim IG\left(e_k+\frac{v_k}{2}; f_k+\frac{1}{2}\sum_{i=1}^{n}Z_{ik}\left(\alpha_i-\mu_k\right)^2\right)$,

onde $v_k = \sum_{i=1}^{n} Z_{ik}$, $k = 1,2$ e $\sigma_{(k)}^2 = \sigma_j^2$, $j \neq k$, $j,k = 1,2$;

(iii) $\pi(\beta_{0j}|\alpha,\theta_{(\beta_{0j})},\mu,\sigma^2,\lambda_1,\mathbf{y},\mathbf{x},\mathbf{Z}) \propto \exp\left\{-\frac{1}{2b_{0j}^2}\left(\beta_{0j}-a_{0j}\right)^2\right\}\psi_{2i}(\alpha,\beta)$,

onde $\psi_{2i}(\alpha,\beta) = \exp\left\{\beta_{0j}y_{j\bullet} - \sum_{i=1}^{n}\log\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]\right\}$,

$y_{j\bullet} = \sum_{i=1}^{n} y_{ji}$, $j = 1,2,3$ e $\theta_{(\beta_{0j})}$ denota o vetor dos parâmetros, exceto β_{0j};

(iv) $\pi(\beta_{1j}|\alpha,\theta_{(\beta_{1j})},\mu,\sigma^2,\lambda_1,\mathbf{y},\mathbf{x},\mathbf{Z}) \propto \exp\left\{-\frac{1}{2b_{1j}^2}\left(\beta_{1j}-a_{1j}\right)^2\right\}\psi_{3i}(\alpha,\beta)$,

onde $\psi_{3i}(\alpha,\beta) = \exp\left\{\beta_{1j}\sum_{i=1}^{n}x_i y_{ji} - \sum_{i=1}^{n}\log\left[1+\exp\left(\alpha_i+\beta_{0j}+\beta_{1j}x_i\right)\right]\right\}$,

$j = 1,2,3$ e $\theta_{(\beta_{1j})}$ denota o vetor dos parâmetros, exceto β_{1j};

(v) $\pi\left(\lambda_1|\alpha,\beta,\mu,\sigma^2,\mathbf{y},\mathbf{x},\mathbf{Z}\right) \sim B\left(g+v_1;\, h+v_2\right)$;

(vi) $\pi\left(\mu_k|\alpha,\beta,\mu_{(k)},\sigma^2,\lambda_1,\mathbf{y},\mathbf{x},\mathbf{Z}\right) \sim N\left(\frac{c_k\sigma_k^2+d_k^2\sum_{i=1}^{n}\alpha_i Z_{ik}}{\sigma_k^2+d_k^2\sum_{i=1}^{n}Z_{ik}}; \frac{c_k^2\sigma_k^2}{\sigma_k^2+d_k^2\sum_{i=1}^{n}Z_{ik}}\right)$,

onde $\mu_{(k)} = \mu_j$, $j \neq k$, $j,k = 1,2$.

Observe que precisamos usar o algoritmo de Metropolis-Hastings para gerar α_i e β_{lj}, $i = 1,...,n$, $l = 0,1$, $j = 1,2,3$. Uma grande simplificação é dada usando o programa *WinBugs* (Spiegelhalter et al., 1999), que não requer a especificação das distribuições condicionais usadas para gerar amostras.

4 Resultados e análises

Para analisar os dados da Tabela 1, inicialmente assumimos um modelo de regressão logística (1) e (2) com efeitos aleatórios apresentando uma distribuição normal (3) com as distribuições *a priori* (6) e (7) para β_{lj} e σ_α^2, respectivamente, com $a_{01} = 0$, $a_{02} = 0,288$, $a_{03} = 0$, $a_{1j} = 0$ para $j = 1,2,3$ e $b_{lj}^2 = 10^2$ para $l = 0,1$, $j = 1,2,3$, $c = d = 0,01$ (denominaremos este modelo de M_1). No ajuste do modelo com efeitos aleatórios apresentando uma mistura de duas distribuições normais (modelo M_2) foram assumidas distribuições *a priori* (6) e (9) para β_{lj} e $\mu_k, \sigma_k^2, \lambda_k$, respectivamente, com $a_{01} = 0$, $a_{02} = 0,288$, $a_{03} = 0$, $a_{1j} = 0$ para $j = 1,2,3$ e $b_{lj}^2 = 10^2$ para $l = 0,1$, $j = 1,2,3$, e $c_1 = -2.5$, $c_2 = 2.0$, $d_1^2 = d_2^2 = 1$, $e_1 = e_2 = f_1 = f_2 = 5$, $g = h = 1$.

Em ambos casos, os valores dos hiperparâmetros foram escolhidos após uma análise preliminar dos dados (análise Bayesiana empírica, ver, por exemplo, Carlin e Louis, 2000). Na Tabela 1, observamos que, quando um indivíduo é alocado no grupo placebo ($x_i = 0$), temos da amostra uma probabilidade de sucesso de 50% no tempo T_1, 57% (ou seja, $\frac{8}{14}$) no tempo T_2 e 50% no tempo T_3. Assim, ao desprezarmos o efeito aleatório na equação (2), especificamos ao parâmetro β_{01} uma média *a priori* igual a $a_{01} = \ln\left[\frac{1}{2}\left(1-\frac{1}{2}\right)\right] = 0$, ao parâmetro β_{02} uma média *a priori* igual a $a_{02} = \ln\left[\frac{8}{14}\left(1-\frac{8}{14}\right)\right] = 0,288$, e ao parâmetro β_{03} uma média *a priori* igual a $a_{03} = \ln\left[\frac{1}{2}\left(1-\frac{1}{2}\right)\right] = 0$.

Utilizamos o programa *WinBugs* para gerar 55.000 amostras de Gibbs da densidade *a posteriori* conjunta (8) e (13), e destas foram descartadas as 5.000 primeiras amostras ("*burn-in samples*") com a finalidade de eli-minar o efeito dos valores iniciais usados no algoritmo de simulação. Além disso, foram consideradas as iterações 5ª, 10ª, 15ª, ..., resultando uma amostra final de 10.000 amostras para cada parâmetro. A convergência do algoritmo foi verificada a partir de gráficos temporais das amostras geradas e utilizando técnicas usuais existentes na literatura (Gelman e Rubin, 1992).

Na Tabela 2 estão as estimativas de Monte Carlo para as médias *a posteriori* e intervalos de credibilidade 95% para os parâmetros de interesse de ambos modelos. Os limites inferior e superior destes intervalos de credibilidade são dados, respectivamente, pelos percentis 2,5% e 97,5% das distribuições *a posteriori* geradas pelo algoritmo. São apresentadas também na Tabela 2 estimativas das diferenças $\beta_{02} - \beta_{01}$, $\beta_{03} - \beta_{01}$, $\beta_{03} - \beta_{02}$, destinadas a inferências sobre o efeito do placebo ao longo do tempo, e $\beta_{12} - \beta_{11}$, $\beta_{13} - \beta_{11}$, $\beta_{13} - \beta_{12}$, referindo-se a inferências sobre o efeito do tratamento com acarbose ao longo do tempo. A partir dos resultados obtidos, temos evidências de que o parâmetro β_{13} e as quantidades $\beta_{13} - \beta_{11}$ e $\beta_{13} - \beta_{12}$ são, para o modelo M_1, diferentes de zero (intervalo com 95% de credibilidade no qual o valor zero não está contido).

Tabela 2 – Sumários *a posteriori* para os modelos M_1 (assumindo uma distribuição normal para os efeitos aleatórios) e M_2 (assumindo uma mistura de duas distribuições normais para os efeitos aleató rios)

Parâmetro	Média	Desvio-padrão	Mediana	Intervalo de credibilidade 95%	
Modelo M_1					
β_{01}	-0,017	1,100	-0,013	-2,236	2,213
β_{02}	0,566	1,107	0,550	-1,573	2,837
β_{03}	0,004	1,085	0,007	-2,159	2,197
β_{11}	-0,229	1,593	-0,230	-3,342	3,006
β_{12}	-0,798	1,602	-0,771	-4,008	2,412
β_{13}	3,247	1,842	3,114	0,010	7,274
$\beta_{02} - \beta_{01}$	0,583	1,094	0,562	-1,520	2,777
$\beta_{03} - \beta_{01}$	0,021	1,072	0,032	-2,070	2,157
$\beta_{03} - \beta_{02}$	-0,562	1,083	-0,543	-2,746	1,539
$\beta_{12} - \beta_{11}$	-0,569	1,543	-0,555	-3,658	2,437
$\beta_{13} - \beta_{11}$	3,475	1,852	3,376	0,082	7,421
$\beta_{13} - \beta_{12}$	4,045	1,898	3,950	0,564	8,084
Modelo M_2					
β_{01}	0,788	1,488	0,820	-2,169	3,560
β_{02}	1,352	1,441	1,389	-1,542	4,118
β_{03}	0,811	1,469	0,824	-2,134	3,619
β_{11}	-1,084	1,389	-1,067	-3,913	1,626
β_{12}	-1,652	1,397	-1,601	-4,502	0,986
β_{13}	2,223	1,533	2,178	-0,683	5,366
$\beta_{02} - \beta_{01}$	0,564	1,067	0,558	-1,474	2,715
$\beta_{03} - \beta_{01}$	0,024	1,045	0,013	-2,042	2,053
$\beta_{03} - \beta_{02}$	-0,541	1,070	-0,532	-2,680	1,521
$\beta_{12} - \beta_{11}$	-0,567	1,488	-0,554	-3,522	2,261
$\beta_{13} - \beta_{11}$	3,307	1,686	3,260	0,171	6,805
$\beta_{13} - \beta_{12}$	3,875	1,738	3,826	0,642	7,445

Uma outra forma de avaliar os resultados é pelas razões de probabilidades de menstruação (Tabela 3), segundo as amostras geradas pelo algoritmo MCMC. Seja $\overline{p}_j^{(0)}$ a probabilidade de menstruação no j-ésimo tempo ($j = 1, 2, 3$) dado que a paciente foi alocada no grupo placebo e $\overline{p}_j^{(1)}$ a probabilidade de menstruação no j-ésimo tempo ($j = 1, 2, 3$) dado que a paciente foi alocada no grupo que recebeu o tratamento. Temos assim:

$$\overline{p}_j^{(0)} = \frac{\sum_{i=1}^n p_{ji}(1-x_i)}{n-\sum_{i=1}^n x_i} \qquad \text{e} \qquad \overline{p}_j^{(1)} = \frac{\sum_{i=1}^n p_{ji}x_i}{\sum_{i=1}^n x_i}, \tag{14}$$

onde a probabilidade p_{ji} é obtida da equação (2). O modelo M_1 estimou $\overline{p}_1^{(0)}$ em 0,5007, $\overline{p}_2^{(0)}$ em 0,5721, $\overline{p}_3^{(0)}$ em 0,5025, $\overline{p}_1^{(1)}$ em 0,4595, $\overline{p}_2^{(1)}$ em 0,4630 e $\overline{p}_3^{(1)}$ em 0,8464, estimativas estas bastante próximas àquelas obtidas diretamente da amostra, exibidas na Tabela 1. Temos desta maneira, por exemplo, que a probabilidade de uma mulher que recebeu o placebo menstruar no período T_2 é $\frac{\overline{p}_2^{(0)}}{\overline{p}_1^{(0)}}$ vezes a probabilidade de menstruar no período T_1. Esta razão é estimada em 1,147 pelo modelo M_1 (ver Tabela 3). Intervalos de credibilidade 95% permitem inferências sobre estas razões de probabilidade. Na Tabela 3, observa-se que os intervalos de credibilidade 95% associados às razões $\frac{\overline{p}_3^{(1)}}{\overline{p}_1^{(1)}}$ e $\frac{\overline{p}_3^{(1)}}{\overline{p}_3^{(0)}}$ não cobrem o valor 1, evidenciando que a probabilidade de menstruação de uma mulher que recebe o tratamento é maior no período T_3 do que no período T_1; e que, no período T_3, a chance de menstruação é maior para a mulher tratada que para a mulher não tratada. Estas inferências mostram, portanto, evidências de que o tratamento é eficaz.

Tabela 3 – Razões de probabilidades de menstruação, segundo os modelos M_1 (assumindo uma distribuição normal para os efeitos aleatórios) e M_2 (assumindo uma mistura de duas distribuições normais para os efeitos aleatórios)

Tratamentos e períodos comparados	Modelo M_1			Modelo M_2		
	Mediana	Intervalo de credibilidade 95%		Mediana	Intervalo de credibilidade 95%	
Placebo, T_2 *versus* T_1	1,147	0,697	1,944	1,143	0,656	1,930
Placebo, T_3 *versus* T_1	1,004	0,589	1,778	1,002	0,567	1,751
Acarbose, T_2 *versus* T_1	1,005	0,526	1,916	0,988	0,511	1,969
Acarbose, T_3 *versus* T_1	1,849	1,197	3,263*	1,820	1,167	3,263*
T_1, acarbose *vs.* placebo	0,918	0,491	1,640	0,941	0,493	1,661
T_2, acarbose *vs.* placebo	0,807	0,442	1,372	0,818	0,424	1,417
T_3, acarbose *vs.* placebo	1,685	1,158	2,744*	1,683	1,146	2,705*

∗Razão na qual o respectivo intervalo de credibilidade 95% não contém o valor 1.

Conclusões

O modelo proposto mostra evidências de que o tratamento com acarbose exerce efeito na ocorrência da menstruação, sendo este efeito também evidente ao longo do tempo. Cabe destacar que a probabilidade da paciente que recebe o tratamento menstruar no período T_3 é aproximadamente 1,9 vezes a probabilidade de menstruar no período T_1. No período T_3, a chance de menstruação para a mulher tratada é aproximadamente 1,7 vezes a chance de menstruação para a mulher não tratada.

Para comparar os dois modelos ajustados, foram obtidos os respectivos valores do *Deviance Information Criterion* (DIC) (ver, por exemplo, Paulino; Turkman e Murteira, 2003). Para o modelo M_1 obtivemos um DIC igual a $97,389$, e, para o modelo M_2, um DIC igual a $102,744$. Isto sugere que o modelo que proporciona um melhor ajuste aos dados é o M_1, ou seja, o modelo de regressão logística assumindo uma distribuição normal para os efeitos aleatórios.

Considerações Finais

O uso de aproximações Bayesianas baseadas em métodos MCMC é uma atraente alternativa para a análise de dados binários correlacionados. A correlação entre os dados pode ser capturada pelos efeitos aleatórios em que uma distribuição normal ou uma mistura de distribuições normais proporcionam maior flexibilidade de ajuste. Os métodos MCMC são facilmente implementados e não requerem sofisticação computacional, como foi observado no uso do programa *WinBugs*.

Agradecimentos

Na condução do presente estudo, a pesquisa de Eliza Omai e de Adriana de Fátima Lourençon foi fomentada pela Fundação de Apoio ao Ensino, Pesquisa e Assistência do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo (Faepa). Os autores agradecem ao Dr. Ivan Penna e aos médicos da Faculdade de Medicina de Ribeirão Preto da USP que cederam os dados utilizados neste estudo.

ACHCAR, J. A.; MARTINEZ, E. Z.; OMAI, E.; LOURENÇON, A. de F.; PERDONÁ, G. C. A Bayesian approach for correlated binary data: a longitudinal study on the occurrence of menstruation in patients with polycystic ovarian syndrome after treatment. *Rev. Mat. Estat.,* São Paulo, v.24, n.1, p.113-125, 2006.

▪ *ABSTRACT: In this paper, we develop a Bayesian logistic regression analysis with random effects considering longitudinal correlated binary data. We illustrate the approach with an example in the field of clinical medicine involving the evaluation of the occurrence of menstruation in patients with polycystic ovarian syndrome after a treatment. This data set was obtained from a real study conducted at the Medical School of the University of São Paulo, Ribeirão Preto, Brazil (Penna et al., 2005, Human Reproduction, v.20, n.9, p.2396-401). The occurrence/no occurrence of menstruation was observed on each subject at three specific times over a given period. Consequently, a no-null correlation among the observations from a same subject is expected, and this effect is captured by adding a random effect in the model. We explore Bayesian approaches to modeling the data assuming that the random effects are drawn from a normal distribution and also from a mixture of normal distributions. Inferences for the model parameters are based on MCMC methods. Model comparison is assessed via Deviance Information Criteria (DIC).*

▪ *KEYWORDS: Correlated binary data; logistic regression; Bayesian analysis; clinical trials.*

References

ACHCAR, J. A.; JANEIRO, V.; MAZUCHELI, J. Regression models for correlated binary data if random effects assuming a mixture of normal distributions. *Comput. Stat.*, Heidelberg, v.18, p.39-55, 2003.

ALBERT, J.; JAIS, J. P. Gibbs sampler for the logistic model in the analysis of longitudinal binary data. *Stat. Med.*, Chichester, v.17, n.24, p.2905-2921, 1998.

CARLIN, B. P.; LOUIS, T. A. *Bayes and empirical Bayes methods for data analysis.* 2^{nd}ed. London: Chapman & Hall, CRC, 2000. 440p.

CHIB, S.; GREENBERG, E. Analysis of mutivariate probit models. *Biometrika*, London, v.85, n.2, p.347-361, 1998.

DEY, D. K.; CHEN, M. H. Bayesian analysis of correlated binary data models. *Sankhya A*, Kolkata, v.60, p.322-343, 1998.

GELFAND, A. E.; SMITH, A. F. M. Sampling based approaches to calculating marginal densities. *J.Am.Stat.Assoc.*, New York, v.85, n.410, p.398-409, 1990.

GELMAN, A.; RUBIN, B. D. Inference from iterative simulation using multiple sequences. *Stat. Sci.*, Hayward, v.7, n.4, p.457-472, 1992.

HEAGERTY, P. J. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, Washington, v.55, n.3, p.688-698, 1999.

OCHI, Y.; PRENTICE, R. L. Likelihood inference in a correlated probit regression model. *Biometrika*, London, v.71, n.3, p.531-543, 1984.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística Bayesiana.* Lisboa: Fundação Calouste Gulbenkian, 2003. 446p.

PENNA, I. A. A. et al. Acarbose in obese patients with polycystic ovarian syndrome: a double-blind, randomized, placebo-controlled study. *Hum. Reprod.*, Oxford, v.20, n.9, p.2396-2401, 2005.

POCOCK, S. J. *Clinical trials*: a practical approach. Chichester: John Wiley, 1983. 266p.

PREISSER, J. S. et al. Analysis of smoking trends with incomplete longitudinal binary responses. *J. Am. Stat. Assoc.*, Alexandria, v.95, n.452, p.1021-1031, 2000.

PRENTICE , R. L. Correlated binary regression with covariate specific to each binary observation. *Biometrics*, Washington, v.44, n.4, p.1033-1048, 1988.

SHAFFER, M. L.; CHINCHILLI, V.M. Bayesian inference for randomized clinical trials with treatment failures. *Stat. Med.*, Chichester, v.23, n.8, p.1215-1228, 2004.

SMITH, A. F. M.; ROBERTS, G. O. Bayesian computation via the Gibbs sampler and related MCMC methods. *J. R. Stat. Soc. B*, Cambridge, v.55, n.1, p.3-23, 1993.

SPIEGELHALTER, D. J. et al. *WinBugs version 1.3: Bayesian inference using Gibbs sampling*. Cambridge: MRC Biostatistics Unit, 1999. Não paginado.

TANNER, M.; WONG, W. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, New York, v.82, n.398, p.528-540, 1987.

THALL, P. F., SIMON, R. M., SHEN, Y. Approximate Bayesian evaluation of multiple treatment effects. *Biometrics*, Washington, v.56, n.1, p.213-219, 2000.

Recebido em 28.07.2005.

Aprovado após revisão em 25.5.2006.