USO DO FATOR DE BAYES E CRITÉRIOS DE INFORMAÇÃO PARA COMPARAR MODELOS PARA DADOS AGRUPADOS E CENSURADOS.

Sophia Lanza de ANDRADE[1]
Liciana Vaz de Arruda SILVEIRA[1]
in memoriam Francisco Javier Torres AVILÉS[2]

■ RESUMO: Dados agrupados é um caso particular de dados de sobrevivência com censura intervalar, que ocorre quando as observações são avaliadas nos mesmos intervalos de tempo. Geralmente tal caso está associado a dados com grande número de empates, podendo assim serem analisados de forma a considerar o tempo como discreto e ajustando-se modelos à probabilidade do indivíduo falhar em um certo intervalo, dado que sobreviveu ao intervalo anterior (LAWLESS, 2002). Dentre os possíveis modelos adaptados a este tipo de dados, tem-se o Modelo Logístico e o Modelo de Cox. O objetivo deste artigo é comparar o ajuste dos modelos citados anteriormente utilizando critérios clássicos e bayesianos de seleção de modelos. Como ilustração, foi usado um conjunto de dados referente a uma manifestação clínica da doença de Chagas, conhecida como megacolo chagásico (ALMEIDA, 1996).

■ PALAVRAS-CHAVE: Análise de Sobrevivência; fator de Bayes; modelo logístico; modelo de Cox; dados agrupados e censurados.

1 Introdução

Em pesquisas da área biológica é comum encontrar dados em que a variável de interesse é o tempo até a ocorrência de um determinado evento, denominado como tempo de falha. "Por razões históricas, o evento de interesse pode ser referenciado como 'morte' e o período de observação de cada indivíduo como 'tempo

[1]Universidade Estadual Paulista "Júlio de Mesquita Filho"-UNESP, Instituto de Biociências, Departamento de Bioestatística, Caixa Postal 510, CEP: 18618-000, Botucatu, SP, Brasil. E-mail: *sofs_la@hotmail.com*; *liciana@ibb.unesp.br*
[2]Universidad de Santiago de Chile, Departamento de Matemática y Ciencia de Computación, CEP: 9170-022, Santiago, Chile. E-mail: *francisco.torres@usach.cl*

de sobrevivência'. " (CARVALHO *et al.*, 2005). O conjunto de técnicas para a análise desses dados é denominado Análise de Sobrevivência.

Este tipo de dados podem se apresentar com censura, que é a observação parcial da resposta, o que significa que toda informação que se possui resume-se ao fato de que o tempo de falha é superior àquele observado, e empates, isto é, a ocorrência do evento de interesse é observada em mais de um indivíduo no mesmo instante. Tal situação ocorre quando os experimentos são conduzidos de forma a não permitir a observação do tempo exato da ocorrência do evento, restando apenas a informação do intervalo em que ele sucedeu.

Durante a coleta dos dados, pode ocorrer de os mesmos serem registrados em intervalos de tempo, obtendo assim a censura intervalar, já que o tempo exato da ocorrência da falha passa a ser desconhecido. Conforme Colosimo e Giolo (2006, p.246-247), um caso particular de censura intervalar são os dados agrupados, que ocorre quando todas as unidades amostrais são avaliadas nos mesmos instantes. Também é possível ocorrer um grande número de empates, ou seja, proporção maior que 25% (CHALITA *et al.*, 2002), ocasionando também o caso anteriormente discutido. Dessa forma deve-se considerar os tempos de vida como discretos (KALBFLEISCH e PRENTICE, 1980) e ajustar modelos à probabilidade do indivíduo morrer no intervalo, dado que ele sobreviveu ao anterior. O Modelo de Riscos Proporcionais de Cox (1972) e o Logístico (LAWLESS, 2002) podem ser ajustados a este tipo de dados utilizando as transformações complemento log-log e logística, respectivamente.

2 Material e métodos

2.1 Descrição dos dados

O megacólon chagásico é uma das manifestações digestivas da doença de Chagas, podendo vir acompanhada de megaesôfago ou não. Ele altera o trânsito intestinal do cólon, resultado do comprometimento de órgãos de tal sistema, uma vez que houve degeneração do sistema nervosos entérico. Como seu primeiro sintoma é a constipação, seu diagnóstico clínico e anatômico ocorrem tardiamente, sendo que muitos pacientes só buscam tratamento quando surgem complicações como fecaloma e volvo, como pode ser verificado em Rezende (1997) e Silveira (2007).

Muitas técnicas cirúrgicas têm sido realizadas para corrigir as mudanças do trânsito intestinal. Em particular, a técnica descrita por Duhamel (1956) e aperfeiçoada por Haddad (1968) é uma delas.

Os dados deste estudo, obtidos em Almeida (1996), foram divididos em dois grupos de pacientes: o grupo controle, formado por 19 indivíduos com ritmo intestinal normal, e sem doença de Chagas e o grupo tratado, formado por 11 pacientes que sofriam do megacólon chagásico e foram submetidos à cirurgia Duhamel-Haddad.

Todos os indivíduos receberam um marcador radiológico no início do estudo.

O evento de interesse foi o tempo até a eliminação do marcador radiológico. O pesquisador verificou a ocorrência do evento a cada 24h, usando um exame de raio-X. Como cada evento era conhecido como tendo ocorrido num período entre 24h, mas o tempo exato não era conhecido, utilizou-se para estes dados a censura intervalar.

2.2 Modelos utilizados no estudo

Como a presença de censura invalida a obtenção de medidas de tendência central e variabilidade diretamente dos dados de sobrevivência, inicialmente deve-se encontrar uma estimativa para a função de sobrevivência, $S(t)$, e a partir dela estimar as estatísticas de interesse.

Dentre as técnicas utilizadas para se estimar a função de sobrevivência, podem-se destacar as não paramétricas, tais como o estimador de Kaplan-Meier, sugerido por Kaplan e Meier (1958), utilizado quando se conhece o tempo exato de falha, e o estimador de tabela de vida, que pode ser utilizado quando se dispõe apenas da informação sobre os intervalos nos quais ocorreram falhas ou censuras. Estas técnicas não levam em consideração as covariáveis relacionadas com o tempo de vida. Para considerar tais covariáveis, devem-se utilizar os modelos paramétricos, em que se supõe uma distribuição de probabilidade conhecida para os tempos, ou o Modelo de Riscos Proporcionais de Cox (1972), para o qual não é necessário supor distribuição para a variável tempo.

2.2.1 Modelos para dados agrupados e censurados

Por sugestão de Chalita (1997), considere os tempos de vida agrupados em k intervalos denotados por $I_i = [a_{i-1}, a_i)$, $i = 1, \ldots, k$, em que $0 = a_0 < a_1 < \ldots < a_k = \infty$, e assuma que as censuras ocorrem no final do intervalo. Seja D_i o conjunto de indivíduos que falharam no intervalo I_i, R_i o conjunto dos indivíduos sob risco no início de I_i e δ_{li} a variável indicadora de falha do l-ésimo indivíduo em I_i ($\delta_{li} = 1$ se a observação for uma falha e 0 caso contrário).

A função de verossimilhança é escrita em termos da probabilidade de falha do l-ésimo indivíduo em I_i, dado que ele sobreviveu em a_{i-1} e os valores das covariáveis $\mathbf{x}_l$, sendo tal probabilidade representada por $p_i(\mathbf{x}_l)$. Após considerar as contribuições de uma observação censurada e uma não censurada, a função de verossimilhança é dada por:

$$\prod_{i=1}^{k} \prod_{l \in R_i} \{p_i(\mathbf{x}_l)\}^{\delta_{li}} \{1 - p_i(\mathbf{x}_l)\}^{1-\delta_{li}}. \tag{1}$$

A estrutura de regressão representada pela probabilidade $p_i(\mathbf{x}_l)$ pode ser modelada por meio de um modelo de riscos proporcionais ou de chances proporcionais. A seguir são apresentados os dois modelos para $p_i(\mathbf{x}_l)$.

Modelo de riscos proporcionais: Assumindo o Modelo de Cox, $p_i(\mathbf{x}_l)$ assume a seguinte forma:

$$p_i(\mathbf{x}_l) = 1 - \left[\frac{S_0(a_i)}{S_0(a_{i-1})}\right]^{\exp\left\{\mathbf{x}_l'\beta\right\}}, \tag{2}$$

em que $S_0(t)$ é a função de sobrevivência base. Desse modo, é obtida a função de verossimilhança a seguir:

$$\prod_{i=1}^{k}\prod_{l\in R_i}\left\{1 - \left[\frac{S_0(a_i)}{S_0(a_{i-1})}\right]^{\exp\left\{\mathbf{x}_l'\beta\right\}}\right\}^{\delta_{li}}\left\{\left[\frac{S_0(a_i)}{S_0(a_{i-1})}\right]^{\exp\left\{\mathbf{x}_l'\beta\right\}}\right\}^{(1-\delta_{li})}. \tag{3}$$

Assumindo a reparametrização $\gamma_i = \ln(-\ln\rho_i)$, em que $\rho_i = \frac{S_0(a_i)}{S_0(a_{i-1})}$, sugerida por Prentice e Gloeckler (1978), o logaritmo da função de verossimilhança fica

$$\begin{aligned}\ln\mathcal{L}(\beta,\gamma) = &\sum_{i=1}^{k}\sum_{l\in R_i}\Big[\delta_{li}\ln(1-\exp(-\exp(\gamma_i + \mathbf{x}_l'\beta)))- \\ &- (1-\delta_{li})\exp(\gamma_i\mathbf{x}_l'\beta)\Big].\end{aligned} \tag{4}$$

Modelo logístico: Conforme Lawless (2002, p.23), assumindo o modelo logístico, tem-se

$$p_i(\mathbf{x}_l) = 1 - \frac{1}{\left(1+\gamma_i\exp\left\{\mathbf{x}_l'\beta\right\}\right)}. \tag{5}$$

Desse modo, a função de verossimilhança pode ser escrita como

$$\mathcal{L}(\beta,\gamma) = \prod_{i=1}^{k}\prod_{l\in R_i}\left\{\frac{\gamma_i\exp\left\{\mathbf{x}_l'\beta\right\}}{1+\gamma_i\exp\left\{\mathbf{x}_l'\beta\right\}}\right\}^{\delta_{li}}\left\{\frac{1}{1+\gamma_i\exp\left\{\mathbf{x}_l'\beta\right\}}\right\}^{(1-\delta_{li})}. \tag{6}$$

Fazendo a reparametrização $\alpha_i = \ln(\gamma_i)$, é possível escrever o logaritmo da função de verossimilhança da seguinte forma:

$$\ln\mathcal{L}(\beta,\alpha) = \sum_{i=1}^{k}\sum_{l\in R_i}\left[\delta_{li}\left(\alpha_i + \mathbf{x}_l'\beta\right) - \ln\left(1+\exp\left(\alpha_i + \mathbf{x}_l'\beta\right)\right)\right]. \tag{7}$$

A estimação dos parâmetros, em ambos os modelos, é feita pelo método da máxima verossimilhança. As estimativas de máxima verossimilhança $\hat{\beta}$ e $\hat{\alpha}$ de β e α são obtidas pela maximização do logaritmo da função de verossimilhança. A função de verossimilhança é maximizada resolvendo as equações resultantes das

derivadas de primeira ordem, com relação aos parâmetros β e α. Se o sistema de equações formado não admitir soluções analíticas, então estas devem ser resolvidas numericamente por meio de um método iterativo, por exemplo, Newton Raphson.

Como já foi visto, os dados de Almeida (1996) estão divididos em dois grupos: grupo controle e grupo tratado. Sendo assim, é necessária uma expressão para a função de verossimilhança para o caso de dois grupos. Tal função foi proposta por Cox (1975) e é dada por

$$\prod_{i=1}^{k}\prod_{j=1}^{2}\left(\begin{array}{c} n_{ij} \\ f_{ij} \end{array}\right) p_{ij}^{f_{ij}}(1-p_{ij})^{(n_{ij}-f_{ij})}, \tag{8}$$

em que:
n_{ij} é o número de indivíduos sob risco em I_i e pertencentes ao grupo j;
f_{ij} é o número de falhas em I_i e pertencentes ao grupo j.

2.2.2 Análise bayesiana

Como alternativa à análise clássica, existe os métodos Bayesianos, que permitem a incorporação de informações sobre os parâmetros antes dos dados serem observados por meio da densidade de probabilidade *a priori*, $\pi(\theta)$. Quando não há nenhuma informação a respeito dos parâmetros ou tem-se por objetivo comparar resultados da análise bayesiana com a clássica, deve-se formular uma densidade *a priori* tal que toda informação venha exclusivamente dos dados, ou seja, usa-se uma densidade *a priori* não-informativa, sendo os métodos de Jeffreys e Bayes-Laplace os mais conhecidos para tal tarefa.

O Teorema de Bayes apresentado a seguir proporciona a obtenção de uma densidade *a posteriori*, $\pi(\theta|\mathbf{Y})$, a qual combina a informação vinda dos dados, através da função de verossimilhança, com a informação prévia, representada na densidade *a priori* (ver por exemplo, Box e Tiao (1973, p.10-12)).

Teorema de Bayes: Considere uma amostra aleatória $\mathbf{Y}$ onde os dados são independentes e identicamente distribuídos com uma distribuição conjunta dada pela densidade $\mathcal{L}(\mathbf{Y}|\theta)$, também definida como função de verossimilhança para θ quando os dados foram observados e uma distribuição *a priori* para θ, dada por $\pi(\theta)$. Em Gammerman e Lopes (2006, p.43-44), a distribuição *a posteriori* é descrita como

$$\pi(\theta|\mathbf{Y}) = \frac{\mathcal{L}(\mathbf{Y}|\theta)\pi(\theta)}{\int_{\theta\in\Theta}\mathcal{L}(\mathbf{Y}|\theta)\pi(\theta)d\theta} \propto \mathcal{L}(\mathbf{Y}|\theta)\pi(\theta), \tag{9}$$

em que Θ é o espaço paramétrico.

Como o termo à esquerda da igualdade é uma densidade para θ, a observação $\mathbf{Y}$ é apenas uma constante. Logo, o teorema de Bayes pode ser reescrito em sua forma mais resumida como apresentada no termo à direita da Equação 9.

Para selecionar o modelo melhor ajustado aos dados de Almeida (1996)

Tabela 1 - Calibragem do fator de Bayes

B_{01}	Evidência a favor de H_0
1 – 3,2	Não Significativa
3,2 – 10	Significativa
10 – 100	Forte
>100	Decisiva

empregando ferramentas da Análise Bayesiana, será adotado o Fator de Bayes, descrito a seguir, e outros critérios bayesianos, que serão descritos adiante.

O fator de Bayes nada mais é do que uma razão de chances das distribuições *a posteriori* para as distribuições *a priori* que, por meio de manipulações algébricas, resume-se numa razão de chances das funções de verossimilhança marginais. Considerando duas hipóteses H_0 e H_1 correspondentes aos modelos M_0 e M_1, respectivamente, o Fator de Bayes a favor de H_0 é dado por

$$B_{01} = \frac{P(\mathbf{Y}|M_0)}{P(\mathbf{Y}|M_1)} \tag{10}$$

em que,

$$P(\mathbf{Y}|M_{k'}) = \int_{\Theta_k} \mathcal{L}(\theta_{M_{k'}}, M_{k'})\pi(\theta_{M_{k'}}|M_{k'})d\theta_{M_{k'}} \tag{11}$$

é a verossimilhança marginal do modelo $M_{k'}$, $\mathcal{L}(\theta_{M_{k'}}, M_{k'})$ é a função de verossimilhança para o modelo $M_{k'}$, $\pi(\theta_{M_{k'}}|M_{k'})$ é a distribuição *a prori* conjunta para os parâmetros do modelo $M_{k'}$ e $\theta_{M_{k'}}$ é o vetor de parâmetros do modelo $M_{k'}$.

Quando $B_{01} > 1$, o modelo M_0 apresenta um melhor ajuste aos dados.

Na maioria das vezes $P(\mathbf{Y}|M_{k'})$ é muito difícil de ser calculada (Paulino et al., 2003), sendo necessário adotar métodos numéricos para sua resolução, como por exemplo, Métodos de Monte Carlo.

Na Tabela 1 é reproduzida a tabela de calibragem introduzida por Jeffreys (1961, app.B) do Fator de Bayes, útil para sua interpretação.

Densidades *a priori* e *a posteriori*: Uma vez definidos o modelo para os dados e a distribuição *a priori*, devemos combinar a informação prévia sobre os parâmetros com a informação contida nos dados por meio da função de verossimilhança, obtendo assim uma distribuição *a posteriori* para os parâmetros do modelo. Para o modelo de Cox, consideramos as densidades *a priori* de ρ e β dadas por $\rho \sim Weibull(a_1, b_1)$ ou $\rho \sim Gama(a_2, b_2)$ e $\beta \sim Normal(\mu, \sigma^2)$. Já para o modelo Logístico, consideramos as densidades *a priori* de γ e β dadas por $\gamma \sim Weibull(c_1, d_1)$ ou $\gamma \sim Gama(c_2, d_2)$ e $\beta \sim Normal(\mu, \sigma^2)$, supondo que ρ, β e γ são densidades *a priori* independentes.

Dessa forma, temos como distribuições *a posteriori*:

Modelo de Cox, com $\rho \sim Weibull(a_1, b_1)$ e $\beta \sim Normal(\mu, \sigma^2)$

$$\begin{aligned}\pi(\rho,\beta|D) &\propto \prod_{i=1}^{k}\prod_{j=1}^{2}\left(1-\rho_i^{\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{f_{ij}}\left(\rho_i^{\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{n_{ij}-f_{ij}}\times \\ &\times \frac{b_1}{a_1\sigma\sqrt{2\pi}}\left(\frac{\rho}{a_1}\right)^{b_1-1}\exp\left[-\left(\frac{\rho}{a_1}\right)^{b_1}-\frac{(\beta-\mu)^2}{2\sigma^2}\right] \qquad (12)\end{aligned}$$

Agora com $\rho \sim Gama(a_2, b_2)$ e $\beta \sim Normal(\mu, \sigma^2)$

$$\begin{aligned}\pi(\rho,\beta|D) &\propto \prod_{i=1}^{k}\prod_{j=1}^{2}\left(1-\rho_i^{\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{f_{ij}}\left(\rho_i^{\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{n_{ij}-f_{ij}}\times \\ &\times \frac{{b_2}^{a_2}\rho^{a_2-1}}{\Gamma(a_2)\sigma\sqrt{2\pi}}\exp\left[-b_2\rho-\frac{(\beta-\mu)^2}{2\sigma^2}\right] \qquad (13)\end{aligned}$$

Modelo logístico, com $\gamma \sim Weibull(c_1, d_1)$ e $\beta \sim Normal(\mu, \sigma^2)$

$$\begin{aligned}\pi(\gamma,\beta|D) &\propto \prod_{i=1}^{k}\prod_{j=1}^{2}\left(\frac{\gamma_i\exp\left\{\mathbf{x}_j'\beta\right\}}{1+\gamma_i\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{f_{ij}}\left(\frac{1}{1+\gamma_i\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{n_{ij}-f_{ij}}\times \\ &\times \frac{d_1}{c_1\sigma\sqrt{2\pi}}\left(\frac{\gamma}{c_1}\right)^{d_1-1}\exp\left[-\left(\frac{\gamma}{c_1}\right)^{d_1}-\frac{(\beta-\mu)^2}{2\sigma^2}\right] \qquad (14)\end{aligned}$$

Já com $\gamma \sim Gama(c_2, d_2)$ e $\beta \sim Normal(\mu, \sigma^2)$

$$\begin{aligned}\pi(\gamma,\beta|D) &\propto \prod_{i=1}^{k}\prod_{j=1}^{2}\left(\frac{\gamma_i\exp\left\{\mathbf{x}_j'\beta\right\}}{1+\gamma_i\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{f_{ij}}\left(\frac{1}{1+\gamma_i\exp\left\{\mathbf{x}_j'\beta\right\}}\right)^{n_{ij}-f_{ij}}\times \\ &\times \frac{{d_2}^{c_2}\gamma^{c_2-1}}{\Gamma(c_2)\sigma\sqrt{2\pi}}\exp\left[-d_2\gamma-\frac{(\beta-\mu)^2}{2\sigma^2}\right] \qquad (15)\end{aligned}$$

Como é possível observar, na primeira parte das Equações 12, 13, 14 e 15 está a verossimilhança dos modelos de Cox e Logístico, respectivamente, para o caso de dois grupos, apresentado por Cox (1975). Já na segunda parte estão as densidades *a priori* conjuntas dos parâmetros dos modelos.

Os métodos de Monte Carlo são uma alternativa apropriada aos métodos numéricos para a resolução de integrais, conforme afirma Paulino et al., (2003). Neste trabalho, foi utilizado métodos *Markov Chain Monte Carlo* (MCMC), cuja ideia principal é obter uma amostra da distribuição *a posteriori* e calcular estimativas amostrais de características desta distribuição. Para tanto, foi utilizado um estimador da verossimilhança marginal, conhecido com estimador Monte Carlo,

apresentado abaixo

$$\hat{P}(\mathbf{Y}|M_{k'}) = \frac{1}{S}\sum_{s=1}^{S}\mathcal{L}\left(\theta_{M_{k'}}^{(s)}, M_{k'}\right), \tag{16}$$

em que S é o tamanho da amostra obtida por simulação (Gamerman e Lopes, 2006, p.239).

Para o modelo de Cox, o estimador de Monte Carlo da verossimilhança marginal é dado por:

$$\begin{aligned}\hat{P}(\mathbf{Y}|M) =& \frac{1}{S}\sum_{s=1}^{S}\left\{\prod_{i=1}^{k}\prod_{j=1}^{2}\left(1-\rho_i^{(s)^{\exp\left\{\mathbf{x}_j'\beta^{(s)}\right\}}}\right)^{f_{ij}}\times\right.\\ &\left.\times\left(\rho_i^{(s)^{\exp\left\{\mathbf{x}_j'\beta^{(s)}\right\}}}\right)^{n_{ij}-f_{ij}}\right\}.\end{aligned} \tag{17}$$

Para o modelo logístico, o estimador de Monte Carlo da verossimilhança marginal é dado por:

$$\begin{aligned}\hat{P}(\mathbf{Y}|M) =& \frac{1}{S}\sum_{s=1}^{S}\left\{\prod_{i=1}^{k}\prod_{j=1}^{2}\left(\frac{\gamma_i^{(s)}\exp\left\{\mathbf{x}_j'\beta^{(s)}\right\}}{1+\gamma_i^{(s)}\exp\left\{\mathbf{x}_j'\beta^{(s)}\right\}}\right)^{f_{ij}}\times\right.\\ &\left.\times\left(\frac{1}{1+\gamma_i^{(s)}\exp\left\{\mathbf{x}_j'\beta^{(s)}\right\}}\right)^{n_{ij}-f_{ij}}\right\}.\end{aligned} \tag{18}$$

Dentre os métodos MCMC os mais utilizados são os algoritmos *Metropolis-Hastings*, desenvolvido por Metropolis *et al.* (1953) e generalizado por Hastings (1970), e *Gibbs Sampling*, introduzido por Geman e Geman (1984) e aperfeiçoado por Gelfand e Smith (1990). Neste trabalho foi empregado o algoritmo de *Metropolis-Hastings*. Isto ocorreu pois não foi identificada uma distribuição conhecida para as densidades condicionais *a posteriori* de cada parâmetro, o que inviabiliza o uso do amostrador de *Gibbs Sampling*, já que seu núcleo de transição é composto por tais distribuições.

Em relação a análise da convergência, existem propostas de verificações visuais baseadas no comportamento gráfico das iterações na literatura. Nesse estudo foi utilizado o método formal proposto por Gelman e Rubin (1992) para monitorar a convergência do algoritmo de Metropolis-Hastings. Esse método é baseado em técnicas de análise de variância e sugere a convergência da cadeia apenas quando a variância entre as cadeias for bem menor que a variância dentro de cada cadeia ou, equivalentemente, quando histogramas das cadeias misturadas são similares aos de cada uma delas isoladas. Quando o fator de redução $\hat{R} \cong 1$, significa que o período de aquecimento é suficiente e as inferências podem ser realizadas baseadas nos valores das cadeias na segunda metade das iterações.

2.2.3 Critérios de seleção de modelos

Critérios clássicos: Dentre os principais critérios de seleção de modelos na análise clássica estão o critério de Akaike - AIC (AKAIKE, 1974), o critério de informação de Akaike corrigido - AICc (BOZDOGAN, 1987) e o critério de informação de Schwarz - BIC (SCHWARZ, 1978), que são baseados no valor do logaritmo da função de verossimilhança do modelo e dependem do número de observações, n, e do número de parâmetros estimados do modelo, p.

Tais critérios de seleção de modelos podem ser calculados como mostra a Tabela 2.

Tabela 2 - Cálculos dos critérios de seleção de modelos

Critério de seleção de modelos	Cálculo
AIC	$-2\ln\mathcal{L}(\theta)+2p$
BIC	$-2\ln\mathcal{L}(\theta)+p\ln(n)$
AICc	AIC $+\frac{2p(p+1)}{n-p-1}$

Observa-se que dentre todos os possíveis modelos considerados, aquele que possuir o menor valor de AIC, AICc e\ou BIC será considerado o modelo mais adequado, podendo haver divergências entre os mesmos. Uma característica do critério BIC é penalizar os modelos com um maior número de parâmetros, caracterizando a seleção de modelos com número menor dos mesmos. A correção para o AIC proposta por Bozdogan (1987) baseia-se no fato de o AIC poder ter um desempenho ruim se existirem muitos parâmetros em comparação com o tamanho da amostra. Segundo Burnham e Anderson (2004), o AICc deve ser utilizado quando a razão $\frac{n}{p}$ é pequena ($\frac{n}{p}<40$).

Critérios bayesianos: Além do Fator de Bayes, que já foi apresentado, também será usado o DIC (Deviance Information Criterion) que foi introduzido em Spiegelhalter *et al.* (2002) e adaptações do AIC, AICc e BIC utilizando as amostras *a posteriori* geradas pelo método de simulação (no caso deste estudo, Metropolis-Hastings), encontradas em Ntzoufras (2009).

A função *deviance*, proposta por Nelder e Wedderburn (1972), nada mais é do que uma medida de discrepância a fim de discriminar os modelos e medir seus ajustes. Tal função é calculada como:

$$D(\theta)=2\times(\ln\mathcal{L}(\theta)_{max}-\ln\mathcal{L}(\theta)_{est}), \tag{19}$$

em que $\ln\mathcal{L}(\theta)_{max}$ e $\ln\mathcal{L}(\theta)_{est}$ são o logaritmo da função de verossimilhança do modelo completo (saturado) e do modelo em estudo, respectivamente.

No nosso estudo, a função de verossimilhança gera a seguinte função *deviance*

sob o modelo $M_{k'}$:

$$D(\theta_{M_{k'}}, M_{k'}) = 2\sum_{i=1}^{k}\sum_{j=1}^{2}\left\{f_{ij}\ln\left[\frac{f_{ij}}{p_{ij}(\mathbf{x})}\right] + (n_{ij}-f_{ij})\ln\left[\frac{n_{ij}-f_{ij}}{n_{ij}-p_{ij}(\mathbf{x})}\right]\right\}, \quad (20)$$

em que f_{ij} e n_{ij} são o número de falhas e de indivíduos sob risco no intervalo I_i e no grupo j, respectivamente, conforme descritos anteriormente.

Utilizando as amostras *a posteriori*, haverá dois estimadores para $p_{ij}(\mathbf{x})$, descrito anteriormente como a probabilidade de falha no intervalo I_i e no grupo j: $p_{ij}(\mathbf{x})^{(s)}$, calculado na s-ésima iteração da amostra gerada por simulação de cada parâmetro, e $\bar{p}_{ij}(\mathbf{x})$, calculado sobre a média das iterações da amostra gerada por simulação de cada parâmetro. Dessa forma, para o Modelo de Cox:

$$p_{ij}(\mathbf{x})^{(s)} = 1 - \left[\rho_i^{(s)}\right]^{\exp\left\{\mathbf{x}'\beta^{(s)}\right\}} \text{ e } \bar{p}_{ij}(\mathbf{x}) = 1 - [\bar{\rho}_i]^{\exp\left\{\mathbf{x}'\bar{\beta}\right\}}. \quad (21)$$

Já para o modelo logístico:

$$p_{ij}(\mathbf{x})^{(s)} = 1 - \frac{1}{\left(1+\gamma_i^{(s)}\exp\left\{\mathbf{x}'\beta^{(s)}\right\}\right)} \text{ e } \bar{p}_{ij}(\mathbf{x}) = 1 - \frac{1}{\left(1+\bar{\gamma}_i\exp\left\{\mathbf{x}'\bar{\beta}\right\}\right)}. \quad (22)$$

Dessa forma, haverá duas funções *deviances*:

A primeira calculada sobre $p_{ij}(\mathbf{x})^{(s)}$, definida como sendo a média das *deviances a posteriore* e dada por:

$$\begin{aligned}\bar{D}(\theta_{M_{k'}}, M_{k'}) =& \frac{1}{S}\sum_{s=1}^{S} 2\sum_{i=1}^{k}\sum_{j=1}^{2}\left\{f_{ij}\ln\left[\frac{f_{ij}}{p_{ij}(\mathbf{x})^{(s)}}\right] + \right.\\ &\left. + (n_{ij}-f_{ij})\ln\left[\frac{n_{ij}-f_{ij}}{n_{ij}-p_{ij}(\mathbf{x})^{(s)}}\right]\right\}, \end{aligned} \quad (23)$$

e a segunda, calculada sobre $\bar{p}_{ij}(\mathbf{x})$ e definida como sendo a *deviance* das médias *a posteriore*, dada por:

$$D(\bar{\theta}_{M_{k'}}, M_{k'}) = 2\sum_{i=1}^{k}\sum_{j=1}^{2}\left\{f_{ij}\ln\left[\frac{f_{ij}}{\bar{p}_{ij}(\mathbf{x})}\right] + (n_{ij}-f_{ij})\ln\left[\frac{n_{ij}-f_{ij}}{n_{ij}-\bar{p}_{ij}(\mathbf{x})}\right]\right\}, \quad (24)$$

cada uma calculada sob o modelo $M_{k'}$ e sendo S o tamanho da amostra obtida por simulação.

Dessa forma é possível definir $p_D = \bar{D}(\theta_{M_{k'}}, M_{k'}) - D(\bar{\theta}_{M_{k'}}, M_{k'})$, que pode ser interpretado, segundo Ntzoufras (2009, p.219), como o número de parâmetros "efetivos" do modelo $M_{k'}$.

Finalmente, são introduzidos os demais critérios de seleção de modelos apresentados na Tabela 3, em que $\bar{D}(\theta) = \bar{D}(\theta_{M_{k'}}, M_{k'})$ e $D(\bar{\theta}) = D(\bar{\theta}_{M_{k'}}, M_{k'})$ e n é o tamanho da amostra.

Tabela 3 - Cálculos dos critérios de seleção de modelos bayesianos

Critério de seleção de modelos	Cálculo
AIC_B	$\bar{D}(\theta) + 2p_D$
BIC_B	$\bar{D}(\theta) + p_D \ln(n)$
$AICc_B$	AIC $+ \frac{2p_D(p_D+1)}{n-p_D-1}$
DIC	$D(\bar{\theta}) + 2p_D$

Para selecionar o modelo melhor adequado aos dados, opta-se por aquele que tiver o menor valor de AIC_B, BIC_B, $AICc_B$ e\ou DIC, conforme também é sugerido na análise clássica.

3 Resultados e discussão

Os dados referentes aos casos de megacólon chagásico, obtidos em Almeida (1996), foram organizados em relação ao intervalo de falha e/ou censura, resultando na Tabela 4, para que prosseguisse a análise dos mesmos.

Tabela 4 - Dados de megacólon chagásico

I_i	Intervalo (h)	grupo	f_{ij}	censuras	n_{ij}
1	$0 \vdash 24$	0	4	0	19
2	$24 \vdash 48$	0	3	0	15
3	$48 \vdash 72$	0	8	0	12
4	$72 \vdash 96$	0	1	3	4
5	$96 \vdash 168$	0	0	0	0
1	$0 \vdash 24$	1	0	0	11
2	$24 \vdash 48$	1	0	0	11
3	$48 \vdash 72$	1	1	0	11
4	$72 \vdash 96$	1	3	0	10
5	$96 \vdash 168$	1	3	4	7

3.1 Análise de sobrevivência

Utilizando o programa SAS (versão 9.3, 2011), foram estimados os parâmetros de cada modelo, assim como seus respectivos AIC's, AICc's e BIC's. Porém, os parâmetros calculados pelo programa estão sob o efeito das reparametrizações de cada modelo, sendo assim, é necessário aplicar suas respectivas inversas. Tais resultados estão reproduzidos nas Tabelas 5 e 6.

Neste caso, é indicado analisar o AICc ao invés do AIC. Note que, pela análise dos critérios, o modelo melhor ajustado aos dados seria o Modelo Logístico.

Tabela 5 - Critérios de discriminação de modelos

Critério	Mod. Logístico	Mod. Cox
AIC	34,6511	35,0071
AICc	76,6511	77,0071
BIC	35,8345	36,1904

Tabela 6 - Análise da estimativa dos parâmetros pelo método da máxima verossimilhança

Parâmetro	Est.	E.P.	L.I.	L.S.	p-*valor*	E.S.R.
Modelo de Cox						
ρ_1	-3,2297	0,7143	-4,6297	-1,8297	<,0001	0,9612
ρ_2	-3,3152	0,7614	-4,8074	-1,8229	<,0001	0,9643
ρ_3	-1,7085	0,5612	-2,8085	-0,6086	0,0023	0,8343
ρ_4	-1,8998	0,6363	-3,1469	-0,6527	0,0028	0,8610
ρ_5	-0,5805	0,5849	-1,7269	0,5659	0,3210	0,5714
β	1,6717	0,5717	0,5513	2,7922	0,0035	-
Modelo logístico						
γ_1	-3,4904	0,8348	-5,1266	-1,8542	<,0001	0,0305
γ_2	-3,5782	0,8799	-5,3027	-1,8536	<,0001	0,0279
γ_3	-1,6482	0,6615	-2,9446	-0,3517	0,0127	0,1924
γ_4	-1,6577	0,7192	-3,0672	-0,2481	0,0212	0,1906
γ_5	-0,2877	0,7638	-1,7846	1,2093	0,7064	0,7500
β	2,0623	0,6928	0,7044	3,4202	0,0029	-

em que:
Est.: estimativas dos parâmetros pelo método da máxima verossimilhança calculadas pelo programa SAS e, portanto, sob o efeito da reparametrização;
E.P.: Erro Padrão;
L.I. e L.S.: são, respectivamente, os limites inferiores e superiores dos Intervalos de Confiança 95%;
E.S.R.: Estimativa dos parâmetros sem a reparametrização utilizada pelo SAS.

É possível observar que a covariável relacionada ao intervalo 5 possui um p-valor $> 0,05$, significando que o efeito da sobrevivência neste intervalo foi muito baixo.

Será analisado agora o ajuste dos modelos aos dados graficamente.

Em relação ao ajuste dos modelos aos dados do grupo controle, mostrado na Figura 1, ambos os modelos ficam próximos aos dados até $t = 168$h, onde há um distanciamento maior.

Quanto ao grupo tratado, mostrado na Figura 2, é possível notar que até $t = 96$h, o modelo logístico está melhor ajustado aos dados, apesar do evidente distanciamento de ambos os modelos dos dados reais, o que não ocorre após esse instante, quando ambos os modelos estão bem ajustados aos mesmos.

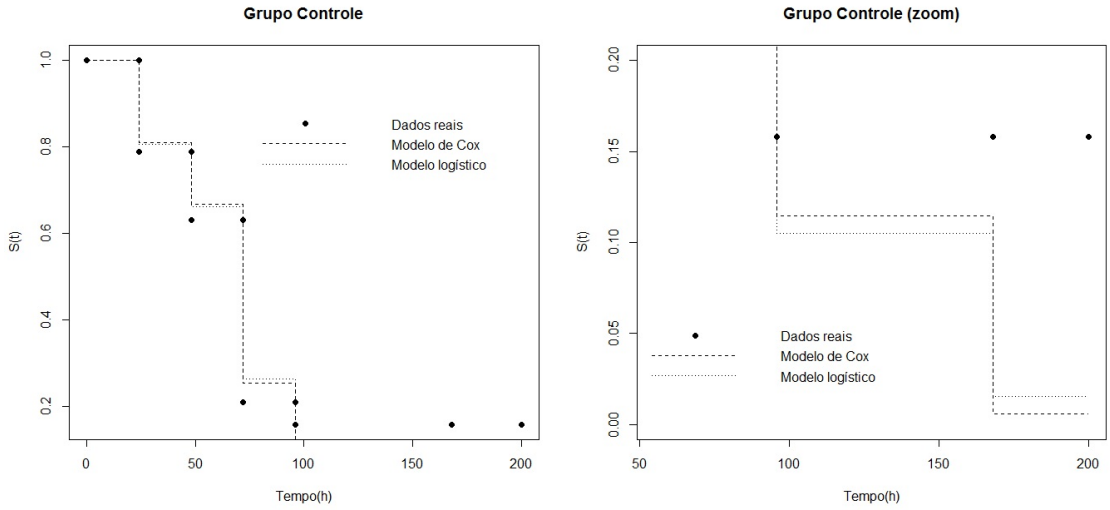

Figura 1 - Ajuste dos modelos aos dados do grupo controle.

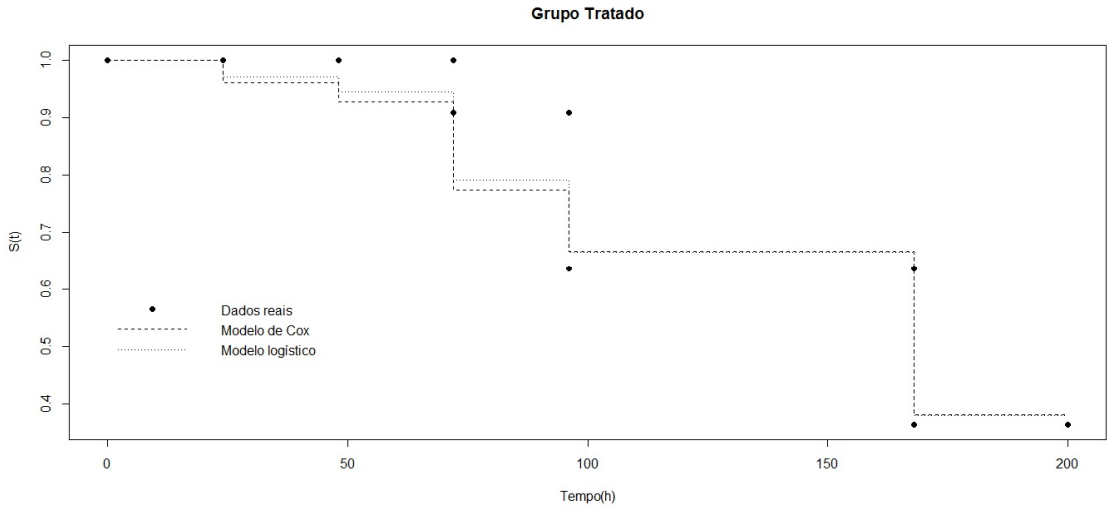

Figura 2 - Ajuste dos modelos aos dados do grupo tratado.

3.2 Análise bayesiana

Após organizar os dados conforme é mostrado na Tabela 4, foram estimados os parâmetros de cada modelo (Modelo Logístico e Modelo de Cox) pelo método

da máxima verossimilhança e contando com o auxílio do programa SAS. Porém, tais estimativas estavam sob o efeito das reparametrizações de cada modelo ($\gamma_i = \ln(-\ln \rho_i)$ para o Modelo de Cox e $\alpha_i = \ln \gamma_i$ para o Modelo logístico), logo, foi necessário aplicar suas respectivas inversas em cada estimativa.

Utilizando agora o programa MatLab (versão 7.14, 2012)e as estimativas dos parâmetros calculadas pelo programa SAS sem a reparametrização como suporte para os chutes iniciais das cadeias, iniciou-se o algoritmo de Metropolis-Hastings, uma vez que não foi identificada uma distribuição conhecida para as densidades condicionais *a posteriori* de cada parâmetro, o que inviabiliza o uso do amostrador de Gibbs Sampling, já que seu núcleo de transição é composto por tais distribuições.

A amostra gerada para cada parâmetro dos modelos foi dividida em 5 cadeias com 2000 iterações cada. Os chutes iniciais de cada cadeia levaram em consideração a estimativa do respectivo parâmetro sem a reparametrização obtida na análise clássica da seguinte forma: os valores do vetor de chutes iniciais respectivo a um parâmetro deveriam englobar o valor do mesmo, de forma que a estimativa do parâmetro se encontrasse entre o menor e o maior valor do vetor.

Por exemplo: suponha que a estimativa do parâmetro sem a reparametrização resultou em 0,85. Uma opção para os chutes iniciais das 5 cadeias de tal parâmetro seria: [0,5 0,6 0,7 0,8 0,9]. Observe que o valor da estimativa do parâmetro (0,85) se encontra entre o menor (0,5) e o maior (0,9) valor do vetor de chutes iniciais. Lembrando que os chutes iniciais de um determinado parâmetro devem ser igualmente espaçados.

Para que a amostra gerada fosse aceita, foi verificado o valor do critério de Gelman e Rubin (R), o histograma dos últimos 1000 valores da amostra *a posteriori* de cada parâmetro dos modelos gerados pelo método de Metropolis-Hastings e a média dos mesmos.

O valor de R deveria ser próximo de 1 para garantir a convergência do algoritmo de Metropolis-Hastings e poder usar a segunda metade das iterações geradas pelo mesmo para a realização das inferências. O histograma deveria ter uma aparência "bem comportada", sendo que em algumas situações isto poderia não ocorrer. Já a média da segunda metade da amostra gerada *a posteriori* para um determinado parâmetro deveria coincidir com a estimativa do mesmo obtida na análise clássica.

Caso algum desses critérios não fosse verificado, os chutes dos hiperparâmetros das densidades *a priori* deveriam ser modificados e o algoritmo de Metropolis-Hastings reiniciado, uma vez que o conhecimento em relação a tais densidades se resumia apenas na distribuição que seguiriam, sendo os valores de seus hiperparâmetros desconhecidos.

Aceita a amostra, utilizou-se a segunda metade da amostra *a posteriori* dos parâmetros para calcular o estimador de Monte Carlo da verossimilhança marginal de cada modelo. Na Tabela 7 é apresentado um resumo dos resultados *a posteriori*, bem como as distribuições *a priori* de cada parâmetro.

Tabela 7 - Resultados *a posteriori*

Parâmetro	dist. *a priori*	Média	L.I.	L.S.	R
Modelo de Cox com distribuição *a priori* Weibull					
ρ_1	Weibull(0.1,70)	0,9609	0,9269	0,9816	1,0060
ρ_2	Weibull(0.1,70)	0,9619	0,9321	0,9832	0,9988
ρ_3	Weibull(0.1,15)	0,8307	0,7224	0,9134	1,0020
ρ_4	Weibull(0.1,20)	0,8656	0,7732	0,9337	1,0012
ρ_5	Weibull(0.1,5)	0,5754	0,3674	0,7765	1,0010
β	Normal(2.1,1)	1,6713	1,0557	2,2698	1,0058
Modelo de Cox com distribuição *a priori* Gama					
ρ_1	Gama(3,2.5)	0,9505	0,8951	0,9859	1,0194
ρ_2	Gama(3,2)	0,9514	0,8818	0,9869	1,0138
ρ_3	Gama(3,2.5)	0,8183	0,6814	0,9137	1,0089
ρ_4	Gama(3,2.5)	0,8400	0,6871	0,9465	1,0046
ρ_5	Gama(3,2.5)	0,5900	0,3127	0,8443	0,9984
β	Normal(1.9,0.09)	1,6688	1,1374	2,1858	1,0122
Modelo logístico com distribuição *a priori* Weibull					
γ_1	Weibull(0.1,0.7)	0,0346	0,0074	0,0894	1,0068
γ_2	Weibull(0.1,0.6)	0,0299	0,0050	0,0826	1,0163
γ_3	Weibull(0.1,1.4)	0,2033	0,0719	0,4134	1,0070
γ_4	Weibull(0.1,1.2)	0,1928	0,0573	0,4014	1,0075
γ_5	Weibull(0.2,7)	0,7413	0,4607	0,9497	1,0023
β	Normal(1.9,0.81)	2,0828	1,2206	3,0817	1,0122
Modelo logístico com distribuição *a priori* Gama					
γ_1	Gama(0.5,20)	0,0303	0,0061	0,0821	1,0136
γ_2	Gama(0.5,25)	0,0259	0,0048	0,0687	1,0077
γ_3	Gama(0.5,50)	0,1830	0,0562	0,4108	1,0039
γ_4	Gama(0.5,5.26)	0,1965	0,0502	0,4507	1,0111
γ_5	Gama(2,2.86)	0,7475	0,2033	1,5901	1,0040
β	Normal(1.6,0.81)	2,1480	1,2462	3,1394	1,0144

em que:
Média: média *a posteriori* dos valores dos parâmetros gerados pelo método iterativo de Metropolis-Hastings;
L.I. e L.S.: são, respectivamente, os limites inferiores e superiores dos Intervalos de Credibilidade 95%;
R: Critério de convergência de métodos iterativos de Gelman e Rubin ($R \sim 1$ indica convergência).

O algoritmo de Metropolis-Hastings gerou, para cada modelo e após o período de aquecimento, uma amostra de tamanho 6000, em que 5000 desses valores estão associados aos 5 parâmetros referentes aos intervalos e os outros 1000 valores à β, parâmetro associado ao grupo a que a observação pertence. Após obter os valores dos parâmetros *a posteriori*, foi calculado o valor do estimador de Monte Carlo da verossimilhança marginal de cada modelo, apresentados na Tabela 8.

Tabela 8 - Valor do estimador de Monte Carlo

Mod. Logístico		Mod. de Cox	
Priori Gama – M_0	*Priori* Weibull – M_1	*Priori* Gama – M_2	*Priori* Weibull – M_3
$1,52739 \times 10^{-20}$	$2,02845 \times 10^{-20}$	$1,10427 \times 10^{-20}$	$2,02587 \times 10^{-20}$

O Fator de Bayes será calculado inicialmente sobre os modelos que possuem as mesmas *prioris*:

$$\hat{B}_{02} = \frac{\hat{P}(\mathbf{Y}|M_0)}{\hat{P}(\mathbf{Y}|M_2)} = \frac{1,52739 \times 10^{-20}}{1,10427 \times 10^{-20}} \simeq 1,383164 \tag{25}$$

$$\hat{B}_{13} = \frac{\hat{P}(\mathbf{Y}|M_1)}{\hat{P}(\mathbf{Y}|M_3)} = \frac{2,02845 \times 10^{-20}}{2,02587 \times 10^{-20}} \simeq 1,001272 \tag{26}$$

Observe que em ambos os casos o modelo logístico foi selecionado como o melhor ajustado aos dados, já que $\hat{B}_{02}, \hat{B}_{13} > 1$, porém de forma não significativa, segundo a Tabela 1, referente a calibragem para o Fator de Bayes de Jeffreys (1961).

Agora, comparando os modelos selecionados anteriormente:

$$\hat{B}_{10} = \frac{\hat{P}(\mathbf{Y}|M_1)}{\hat{P}(\mathbf{Y}|M_0)} = \frac{2,02845 \times 10^{-20}}{1,52739 \times 10^{-20}} \simeq 1,328049 \tag{27}$$

Logo, como $\hat{B}_{10} > 1$, podemos concluir que o Modelo Logístico com a distribuição a *priori* Weibull é o melhor ajustado aos dados, segundo o fator de Bayes, e de forma não significativa.

Assim o Fator de Bayes não especifica qual dos modelos é o mais adequado para os dados do megacolon chagásico, uma vez que o mesmo selecionou modelos de forma insignificativa. Sendo assim, é necessário utilizar os outros critérios de seleção de modelos vistos anteriormente, cujas estimativas estão apresentadas na Tabela 9.

Tabela 9 - Critérios de discriminação de modelos bayesianos

Critério	Modelo Logístico		Modelo de Cox	
	Priori Gama	*Priori* Weibull	*Priori* Gama	*Priori* Weibull
p_D	5,702429	4,807391	3,818094	2,181465
AIC_B	101,4668	97,78799	92,51038	92,14434
$AICc_B$	104,7479	100,096	93,97142	92,66191
BIC_B	109,457	104,5241	97,86028	95,201
DIC	95,76439	92,9806	88,69229	89,96287

Desse modo os demais critérios de seleção de modelos bayesianos apresentados na Tabela 9 apontam o Modelo de Cox como o melhor ajustado aos dados (com *priori* Weibull segundo AIC_B, BIC_B e $AICc_B$ e *priori* Gama segundo DIC), chegando assim a um impasse. Dessa forma, considera-se ambos os modelos possíveis de serem ajustados aos dados, dependendo do tipo de análise a ser feita, ou seja, opta-se pelo Modelo Logístico se for utilizada a Análise Clássica e, utilizando a Análise Bayesiana, pelo Modelo de Cox. Pela sua facilidade na interpretação e aplicação, foi escolhido o Modelo Logístico no ajuste dos dados de Almeida (1996), e a interpretação de seus parâmetros é feita por meio da razão de chances, como é sugerido em Colosimo e Giolo (2006, p.162-164), e aplicado como segue.

Tabela 10 - Razão de chances para o modelo logístico

Razão	valor da razão de chances	LI	LS
Fixando o grupo tratado (grupo 1)			
$\frac{\exp(\hat{\alpha}_2)}{\exp(\hat{\alpha}_1)}$	$\frac{0,0279}{0,0305} = 0,915944$	$0,838534$	$1,0006$
$\frac{\exp(\hat{\alpha}_3)}{\exp(\hat{\alpha}_1)}$	$\frac{0,1924}{0,0305} = 6,310406$	$4,492907$	$8,864017$
$\frac{\exp(\hat{\alpha}_4)}{\exp(\hat{\alpha}_1)}$	$\frac{0,1906}{0,0305} = 6,250741$	$4,983338$	$7,841264$
$\frac{\exp(\hat{\alpha}_5)}{\exp(\hat{\alpha}_1)}$	$\frac{0,7500}{0,0305} = 24,59886$	$21,40233$	$28,27562$
Fixando intervalo			
$\frac{\exp\left(\hat{\alpha}_i+\hat{\beta}\right)}{\exp(\hat{\alpha}_i)}$, $\forall i = 1,\ldots,5$	$7,864036$	$2,022477$	$30,57528$

Conforme pode ser observado na Tabela 10, mantendo o intervalo fixo, as chances de um indivíduo do grupo controle eliminar o marcador radiológico até o fim do intervalo, dado que ele não o fez até o início do mesmo e também não o fez no intervalo anterior ao analisado, é $7,8640$ vezes a chance daquele pertencente ao grupo tratado e sob as mesmas condições, e tal razão se mantém em qualquer intervalo.

Mantendo agora fixado o grupo tratado, teremos a seguinte análise para o segundo intervalo: não eliminando o marcador radiológico até o início do segundo intervalo, 24h, e dado que não o fez no intervalo anterior, as chances do paciente eliminar o marcador até o fim do intervalo em estudo, 48h, é $0,9159$ vezes a chance de um paciente eliminá-lo até o fim do primeiro intervalo, 24h, sendo que não o fez até o início do mesmo, 0h.

A análise dos parâmetros segue-se semelhante ao feito acima para os terceiro, quarto e quinto intervalos, levando em conta as razões de chances calculadas na Tabela 10.

Como o Modelo Logístico possui chances proporcionais, as razões de chances se mantém iguais para ambos os grupos, sendo assim, a interpretação acima também é válida para o grupo controle.

Conclusão

Em relação a metodologia, uma boa definição da distribuição *a priori* é fundamental para o resumo *a posteriori*, uma vez que uma pequena mudança em um único hiperparâmetro de uma das distribuições *a priori* já faz com que o Fator de Bayes seja favorável a outro modelo. Também foi uma oportunidade de estudar o fator de Bayes como critério de seleção de modelos, estudando suas dificuldades (como, por exemplo, as distribuições *a priori*), bem como suas vantagens (a tabela de calibragem).

Uma vez que o Fator de Bayes apresenta alguns problemas quanto à sua aplicação, foi necessário o emprego de outros critérios de seleção de modelos envolvendo as amostras *a posteriori* geradas por um método de simulação.

Em relação aos resultados, por haver poucas diferenças nos ajustes do Modelo Logístico e do Modelo de Cox aos dados de Almeida (1996), supõe-se que ambos poderiam ser empregados dependendo do tipo de análise empregada. Porém optou-se pelo Modelo Logístico pela sua facilidade de aplicação e interpretação.

ANDRADE, S. L.; SILVEIRA, L. V. A.; AVILÉS, F. J. T. Using Bayes factor and information criteria to compare models for grouped and censored data. *Rev. Bras. Biom.,* Lavras, v.35, n.1, p.27-47, 2017.

- *ABSTRACT: Grouped data is a particular case of survival data with interval censoring that occurs when the observations are evaluated at the same time intervals. Generally, its associated at data with a large number of draws and, therefore, it can be analyzed considering discrete-time and fitting models at the probability of an individual fails in an certain interval, given that they survived the previous one (LAWLESS, 2002). Among the possible models adapted to this type of data, we can mention the Logistic Model and Cox's Model. The purpose of this article is to compare the fit of these two models using classic and bayesian model selection criteria. As an example, was used a data set related to a clinical manifestation of Chagas disease known as chagasic megacolon (ALMEIDA, 1996).*

- *KEYWORDS: Survival analysis; Bayes factor; logistic model; Cox's model; grouped and censored data.*

Referências

AKAIKE, A.A new look at statistical model identification. *IEEE-TAC*, v.AC-19, n.6, p.716-722, 1974.

ALMEIDA, A. C. *Resultados funcionais da operação de Duhamel-Haddad no tratamento do megacolo chagásico*, 1996. 79f. Dissertação (Mestrado) - Universidade Federal de Goiás, Goiânia, 1996.

BOX, G. E.; TIAO, G. C. *Bayesian Inference in Statistical Analysis*. New York: Addison-Wesley, 1973. 588p.

BOZDOGAN, H.Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, v.52, n.3, p.345-370, 1987.

BURNHAM, K. P.; ANDERSON, D. R.Multimodel inference: understanding AIC and BIC in model selection. *Socio. Meth. Res.*, v.33, n.2, p.261-304, 2004.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; BARBOSA, M. T. S.; SHIMAKURA, S. E. *Análise de Sobrevida: Teoria e Aplicações em Saúde*. 1.ed. Rio de Janeiro: Fiocruz, 2005. 400p.

CHALITA, L. V. A. S. *Modelos para dados grupados e censurados*, 1997. 135p. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1997.

CHALITA, L. V. A. S.; COLOSIMO, E. A.; DEMÉTRIO, C. B. G.Likelihood approximations and discrete models for tied survival data. *Comun. Statist-Theor. Method.*, v.31, n.7, p.1215-1229, 2002.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. São Paulo: Edgar Blücher Ltda., 2006. 370p.

COX, D. R. Regression Models and Life-Tables (with discussion). *J. R. Stat. Soc. B*, v.34, n.2, p.187-220, 1972.

COX, D. R. Partial Likelihood. *Biometrika*, v.62, n.2, p.269-276, 1975.

DUHAMEL, B.; *Une nouvelle operation de megacolon congenita I.* Presse Méd, v.64, p.2249-2250, 1956.

GAMERMAN, D.; LOPES,H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London: Chapman & Hall, 2006. 340p.

GELFAND, A. E.; SMITH, A. F. M.Sampling-based approaches to calculating marginal densities. *JASA*, v.85, n.410, p.398-409, 1990

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.*, v.7, n.4, p.457-511, 1992.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE-TPAMI*, v.PAMI-6, n.6, p.721-741, 1984.

HADDAD, J.; *Tratamento do megacólon adquirido pelo abaixamento retro-retal do cólon com colostomia perineal (operação de Duhamel modificada)*, 1968. Tese. Faculdade de Medicina da Universidade de São Paulo, 1968

HASTINGS, W. K.Monte Carlo sampling methods using Markov chains and their aplications. *Biometrika*, v.57, n.1, p.97-109, 1970.

JEFFREYS, H. *Theory of Probability*. London: Oxford University Press, 1961.

KALBFLEISCH, J. D.; PRENTICE, R. L. *The statistical analysis of failure time data*. New York: John Wiley and Sons, 1980. 321p.

KAPLAN, L. E.; MEIER, P. Nonparametric estimation from incomplete observations. *JASA*, v.53, n.282, p.457-481, 1958.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New York: John Wiley and Sons, 2002. 664p.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of State calculations by fast computing machines. *J. Chem. Phys.*, v.21, n.6, p.1087-1092, 1953.

NELDER, J. A.; WEDDERBURN, R. W. M. *Generalized Linear Models*. J. R. Stat. Soc. A, v.135, n.3, p.370-384, 1972.

NTZOUFRAS, I. *Bayesian Modeling Using WinBUGS*. Athens: John Wiley & Sons, 2009. 520p.

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

PRENTICE, R. L.; GLOECKLER, L. A. Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics*, v.34, n.1, p.57-67, 1978

REZENDE, J. M. O aparelho digestivo na doença de Chagas: aspectos clínicos. In: DIAS, J. C. P.; COURA, J. R. *Clínica e terapêutica da doença de Chagas: uma abordagem prática para o clínico geral*. New York: Editora FIOCRUZ, 1997. p.153-176.

SCHWARZ, G.Estimating the dimension of a model. *Ann. Stat.*, v.6, n.2, p.461-464, 1978.

SILVEIRA, A. B. M. *Estudo estrutural dos componentes do sistema nervoso entérico e de células inflamatórias: uma contribuição à imunopatologia do megacólon chagásico*, 2007. 122p. Tese (Doutorado) - Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, 2007.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; VAN DER LINDE, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B*, v.64, n.4, p.583-639, 2002.

Recebido em 23.08.2015.

Aprovado após revisão em 23.09.2016.