ESTIMAÇÃO BAYESIANA OBJETIVA DO MODELO DE REGRESSÃO DE FEIGL E ZELEN

Teresa Cristina Martins DIAS[1]
Vera Lúcia Damasceno TOMAZELLA[1]
Eder Angelo MILANI[1]

▪ RESUMO: Em inferência bayesiana, a especificação da distribuição *a priori* para os parâmetros de interesse pode ser complexa, vaga ou muito subjetiva. Esta distribuição expressa o conhecimento ou ignorância a respeito dos parâmetros. Porém, nem sempre é fácil caracterizar ou formular tal distribuição, mas é importante identificar a forma matemática de uma função inicial que tenha efeito mínimo na inferência *a posteriori*, o que nos leva ao uso da metodologia bayesiana objetiva. Neste artigo consideramos a metodologia de análise de referência bayesiana objetiva, introduzida por Bernardo (1979), para a construção da distribuição *a priori* e *a posteriori* com o objetivo de estimar a função de sobrevivência do modelo de regressão exponencial proposto por Feigl e Zelen (1965). Utilizamos o método de simulação Monte Carlo via cadeia de Markov (MCMC) para obtenção dos resultados *a posteriori* dos parâmetros de interesse. Um estudo de simulação e uma aplicação são apresentadas para ilustrar a metodologia proposta.

▪ PALAVRAS-CHAVE: Análise de referência bayesiana; análise de sobrevivência; modelo de regressão, *priori* de referência.

1 Introdução

A teoria de análise de sobrevivência tem sido muito difundida com aplicações nas mais diversas áreas, tais como, biologia, medicina, ciências sociais e engenharia.

O termo análise de sobrevivência refere-se à análise de tempos de vida de pacientes ou ainda a tempos de ocorrência de um evento sendo que os tempos inicial e final de observação são bem definidos. Uma introdução às técnicas usadas

[1]Universidade Federal de São Carlos - UFScar, Centro de Ciências Exatas e Tecnológicas, Departamento de Estatística, Caixa Postal 676, CEP: 13.565-905, São Carlos, São Paulo, Brasil. E-mail: *dtmd@ufscar.br* / *vera@ufscar.br* / *edinhomilani@hotmail.com*

em análise de dados de sobrevivência pode ser encontrada em Armitage e Berry (1987) e Altman (1991). Muitos textos apareceram nas décadas de 70 e 80, como por exemplo, Mann Schaffer e Singpurwalla em 1974, Barlow e Proschan e, Gross e Clark em 1975, Kalbfleisch e Prentice em 1980, Nelson em 1982 e Lawless em 1982 (mais detalhes em Collett, 1994).

Com o objetivo de estudar a relação entre o tempo de ocorrência de um evento e as covariáveis explicativas, modelos de regressão são utilizados para explicar a dependência entre o tempo e estas covariáveis. Esta metodologia permite determinar quais variáveis afetam a forma da função de risco e obter estimativas desta função para cada indivíduo.

Várias técnicas apropriadas para a análise de dados desta natureza estão disponíveis na literatura, tais como: técnicas de regressão não-paramétrica (Miller, 1976; Buckley e James, 1979 e Lawless, 1982) e paramétrica (Feigl e Zelen, 1965 e Lawless, 1982).

Feigl e Zelen (1965) introduziram um modelo de sobrevivência no qual a variável resposta (tempo de sobrevivência) tem distribuição exponencial. Para o i-ésimo paciente de uma amostra de tamanho n, os autores trabalharam com a função de densidade de probabilidade exponencial na parametrização $f(t_i;\lambda_i) = \lambda_i \exp\{-\lambda_i t_i\}, t > 0$, considerando que a relação linear entre o tempo médio de sobrevivência e uma covariável x é dada por,

$$E[T_i] = \frac{1}{\lambda_i} = \alpha + \beta x_i.$$

Para dados não censurados, os autores introduziram o modelo log-linear no qual

$$\frac{1}{\lambda_i} = \alpha \exp\{\beta x_i\} \tag{1}$$

e, Zippin e Armitage (1966) estenderam a análise para dados censurados.

Assumimos que os tempos de vida de pacientes sob estudo seguem uma distribuição exponencial. O objetivo está em estimar a função de sobrevivência $S(t)$ especificamente em um tempo, t_0, dado x_0, considerando a relação (1) entre o tempo médio de vida (λ) e a covariavel. Na inferência clássica estimamos os parâmetros em (1) pelo método de máxima verossimilhança, obtendo assim uma estimativa pontual para S(t). Estimativas intervalares são obtidas via métodos aproximados, como por exemplo, o método delta (Miller, 1981). Tais estimativas podem não ser apropriadas (precisas) principalmente quando a estimativa da função de sobrevivência está próxima ou de 0 ou de 1, e quando a amostra é pequena.

No enfoque bayesiano construímos uma densidade *a posteriori* para a função de sobrevivência, obtendo então estimativas intervalares precisas, via métodos de MCMC (Gamerman e Lopes, 2006). Nesta metodologia consideramos a distribuição *a priori* introduzida por Bernardo (1979) e mais adiante desenvolvida por Berger e Bernardo (1989, 1992a, 1992c), sendo considerada um método para encontrar uma distribuição *a posteriori* objetiva. Para especificar a distribuição *a priori* de referência, usamos a teoria de informação estatística, como uma função matemática

que descreve a situação em que os dados dominam o conhecimento *a priori* sobre a quantidade de interesse.

Este trabalho está organizado da seguinte forma. Na Seção 2 apresentamos o modelo de regressão exponencial proposto por Feigl e Zelen (1965) e a obtenção de estimativas clássicas para os parâmetros de interesse; na Seção 3 descrevemos a construção da *priori* de referência, para os casos de dois ou mais parâmetros, e mostramos as construções das distribuições *a priori* de referência para os modelos estudados. Na Seção 4 ilustramos a teoria apresentada com um estudo de simulação e duas aplicações em dados reais. As conclusões são apresentadas na Seção 5.

2 Modelo de regressão de Feigl e Zelen

Seja T uma variável aleatória não negativa denotando tempos de sobrevivência com distribuição exponencial, cuja função de densidade de probabilidade é dada por,

$$f(t \mid \lambda) = \frac{1}{\lambda} \exp\left\{-\left(\frac{t}{\lambda}\right)\right\}, \tag{2}$$

sendo $t \geq 0$ e $\lambda > 0$ o parâmetro desconhecido representando a taxa de falha constante. A partir de (2), o tempo médio de sobrevivência é dado por $E(T) = \lambda$.

Utilizando a relação dada em (1), proposta por Feigl e Zelen (1965), e a reparametrização $\lambda = \theta_1 e^{\theta_2 x}$, $\theta_1 > 0$ e $-\infty < \theta_2 < \infty$, a função de densidade de probabilidade (2) é reescrita e dada por,

$$f(t|\theta_1, \theta_2, x) = \frac{1}{\theta_1 e^{\theta_2 x}} \exp\left\{-\frac{t}{\theta_1 e^{\theta_2 x}}\right\}. \tag{3}$$

A partir de (3) obtemos as funções de risco, $h(t)$, e de sobrevivência, $S(t)$, dadas por,

$$\begin{aligned} h(t|\theta_1, \theta_2,) &= \frac{1}{\theta_1 e^{\theta_2 x}}, \\ S(t|\theta_1, \theta_2, x) &= \exp\left\{-\frac{t}{\theta_1 e^{\theta_2 x}}\right\}. \end{aligned} \tag{4}$$

A função de verossimilhança para os parâmetros do modelo (3), para uma amostra aleatória de tamanho n na qual observamos $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)$ e $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ é dada por,

$$L(\theta_1, \theta_2 \mid \boldsymbol{t}, \boldsymbol{x}) = L(\theta_1, \theta_2) = \theta_1^{-n} \exp\left\{-\theta_2 n\bar{x} - \sum_{i=1}^{n} \frac{t_i}{\theta_1 e^{\theta_2 x_i}}\right\}. \tag{5}$$

Denotando $l(\theta_1, \theta_2) = \log(L(\theta_1, \theta_2))$ encontramos,

$$l(\theta_1, \theta_2) = -n \log(\theta_1) - \theta_2 n\bar{x} - \sum_{i=1}^{n} \frac{t_i}{\theta_1 e^{\theta_2 x_i}}. \tag{6}$$

As primeiras e segundas derivadas para os parâmetros θ_1 e θ_2, obtidas de (6) são,

$$\left\{\begin{array}{lcl} \frac{\partial l(\theta_1,\theta_2)}{\partial \theta_1} & = & -\frac{n}{\theta_1}+\frac{1}{\theta_1^2}\sum_{i=1}^n t_i e^{-\theta_2 x_i}; \\ \frac{\partial l(\theta_1,\theta_2)}{\partial \theta_2} & = & -\sum_{i=1}^n x_i+\frac{1}{\theta_1}\sum_{i=1}^n t_i x_i e^{-\theta_2 x_i}; \\ \frac{\partial l(\theta_1,\theta_2)}{\partial \theta_1^2} & = & \frac{n}{\theta_1^2}-\frac{2}{\theta_1^3}\sum_{i=1}^n t_i e^{-\theta_2 x_i}; \\ \frac{\partial l(\theta_1,\theta_2)}{\partial \theta_2^2} & = & -\frac{1}{\theta_1}\sum_{i=1}^n t_i x_i^2 e^{-\theta_2 x_i}; \\ \frac{\partial l(\theta_1,\theta_2)}{\partial \theta_1 \partial \theta_2} & = & -\frac{1}{\theta_1^2}\sum_{i=1}^n t_i x_i e^{-\theta_2 x_i}. \end{array}\right. \tag{7}$$

A matriz de informação de Fisher e sua inversa, construídas a partir das equações em (7), são dadas, respectivamente, por

$$H(\theta_1,\theta_2) = \left[\begin{array}{cc} \frac{n}{\theta_1^2} & \frac{n\bar{x}}{\theta_1} \\ \frac{n\bar{x}}{\theta_1} & \sum_{i=1}^n x_i^2 \end{array}\right] \tag{8}$$

e

$$H^*(\theta_1,\theta_2) = H^{-1}(\theta_1,\theta_2) = \left[\begin{array}{cc} \frac{\theta_1^2\sum_{i=1}^n x_i^2}{n^2 s^2} & -\frac{\theta_1\bar{x}}{ns^2} \\ -\frac{\theta_1\bar{x}}{ns^2} & \frac{1}{ns^2} \end{array}\right] \tag{9}$$

para $\bar{x}=\frac{\sum_{i=1}^n x_i}{n}$ e $s^2=\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n}$.

Os parâmetros do modelo de Feigl e Zelen (1965), assim como a função de sobrevivência, podem ser estimados utilizando as matrizes H e H^* dadas em (8) e (9), respectivamente.

Como o interesse está em estimar a função de sobrevivência (4), especificamente em t_0 dado x_0, a função de sobrevivência, neste cenário, é escrita da seguinte forma,

$$S(t_0|x_0) = P\left(T > t_0|x_0\right) = \exp\left\{-\frac{t_0}{\theta_1 e^{\theta_2 x_0}}\right\}.$$

Chamando $S(t_0|x_0)=S$ e considerando a reparametrização,

$$\left\{\begin{array}{ll} S= & \exp\left\{-\frac{t_0}{\theta_1 e^{\theta_2 x_0}}\right\} \\ \phi= & \theta_2, \end{array}\right. \tag{10}$$

obtemos,

$$\left\{\begin{array}{ll} \theta_1= & \frac{t_0}{-\log(S)e^{\phi x_0}} \\ \theta_2= & \phi. \end{array}\right.$$

Logo,

$$f(t|S,\phi,x) = \frac{-\log(S)e^{\phi x_0}}{t_0 e^{\phi x}}\exp\left(\frac{t\log(S)e^{\phi x_0}}{t_0 e^{\phi x}}\right), \tag{11}$$

sendo $S \in [0,1]$ e $\phi \in \mathbf{R}$.

Para o modelo (11) a função de verossimilhança, considerando uma amostra aleatória de tamanho n na qual observamos $\boldsymbol{t} = (t_1, t_2, \cdots, t_n)$ e $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, é

$$\begin{aligned} L(S, \phi \mid \boldsymbol{t}, \boldsymbol{x}) &= t_0^{-n}[-log(S)]^n \\ &\times \exp\left[n\phi(x_0 - \bar{x}) + \frac{log(S)}{t_0}\sum_{i=1}^{n} t_i \exp(\phi(x_0 - x_i))\right]. \end{aligned} \quad (12)$$

A matriz de informação de Fisher em termos dos parâmetros S e ϕ pode ser obtida da seguinte forma,

$$I(S, \phi) = J^T H(\theta_1, \theta_2) J$$

sendo

$$J = \frac{\partial(\theta_1, \theta_2)}{\partial(S, \phi)} = \begin{bmatrix} \frac{t_0 \exp(-\phi x_0)}{S(\log(S))^2} & \frac{t_0 x_0 \exp(-\phi x_0)}{\log(S)} \\ 0 & 1 \end{bmatrix}$$

o jacobiano da transformação inversa. Assim, é possível obter a matriz de informação de Fisher e sua inversa, dadas respectivamente por,

$$I(S, \phi) = \begin{bmatrix} \frac{n}{S^2(\log(S))^2} & \frac{n(x_0 - \bar{x})}{S\log(S)} \\ \frac{n(x_0 - \bar{x})}{S\log(S)} & nx_0^2 - 2nx_0\overline{x} + \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

e

$$I^*(S, \phi) = I^{-1}(S, \phi) = \begin{bmatrix} \frac{-S^2(\log(S))^2(nx_0(x_0 - 2\bar{x}) + \sum_{i=1}^{n} x_i^2)}{n(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2)} & \frac{S\log(S)(\bar{x} - x_0)}{-n\bar{x}^2 + \sum_{i=1}^{n} x_i^2} \\ \frac{S\log(S)(\bar{x} - x_0)}{-n\bar{x}^2 + \sum_{i=1}^{n} x_i^2} & \frac{1}{-n\bar{x}^2 + \sum_{i=1}^{n} x_i^2} \end{bmatrix}.$$

3 Análise de referência

A análise de referência fornece uma função matemática que descreve a situação na qual os dados dominam melhor o conhecimento *a priori* sobre a quantidade de interesse. Esta metodologia foi introduzida por Bernardo (1979) e mais adiante desenvolvida por Berger e Bernardo (1989, 1992a, 1992b, 1992c).

Uma característica importante na abordagem de Berger - Bernardo para construir uma *priori* não-informativa é o tratamento diferenciado para os parâmetros de interesse e *nuisance*. Na presença de parâmetros *nuisance*, devemos estabelecer uma parametrização ordenada com os parâmetros de interesse apontados.

Para obtenção da *priori* de referência conjunta de ψ e γ, parâmetros de interesse e *nuisance*, respectivamente, considere $H(\psi, \gamma)$ a matriz de informação

de Fisher em termos de ψ e γ e $H^*(\psi,\gamma) = H^{-1}(\psi,\gamma)$. Quando as funções $h_{11}^*(\psi,\gamma)^{-1/2}$ e $h_{22}(\psi,\gamma)^{1/2}$ podem ser fatoradas na forma

$$\{h_{11}^*(\psi,\gamma)\}^{-1/2} = f_1(\psi)g_1(\gamma) \text{ e } \{h_{22}(\psi,\gamma)\}^{1/2} = f_2(\psi)g_2(\gamma),$$

sendo h_{ij}^* o elemento da linha i e coluna j da matriz H^* e, h_{ij} o elemento da linha i e coluna j da matriz H. Então,

$$\pi(\psi) \propto f_1(\psi) \text{ e } \pi(\gamma|\psi) \propto g_2(\gamma)$$

e *a priori* de referência relativa ao parâmetro ordenado (ψ,γ) é dada por

$$\pi(\psi,\gamma) = f_1(\psi)\, g_2(\gamma).$$

Para o caso multiparamétrico, a obtenção da *priori* de referência conjunta de ψ e $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_m)$, parâmetros de interesse e *nuisances*, respectivamente, quando as funções $h_{11}^*(\psi,\boldsymbol{\gamma})^{-1/2}$ e $h_{22}(\psi,\boldsymbol{\gamma})^{1/2}$ podem ser fatoradas na forma

$$\{h_{11}^*(\psi,\boldsymbol{\gamma})\}^{-1/2} = f_1(\psi)g_1(\boldsymbol{\gamma}) \text{ e}$$

$$\{h_{i+1,i+1}(\psi,\boldsymbol{\gamma})\}^{1/2} = f_{i+1}(\psi,\gamma_1,\ldots,\gamma_{i-1},\gamma_{i+1},\ldots,\gamma_m)g_{i+1}(\gamma_i),$$

sendo $i = 1,\ldots,m$. Então,

$$\pi(\psi) \propto f_1(\psi) \text{ e } \pi(\gamma_i|\psi,\gamma_1,\ldots,\gamma_{i-1},\gamma_{i+1},\ldots,\gamma_m) \propto g_{i+1}(\gamma_i)$$

e *a priori* de referência relativa ao parâmetro ordenado $(\psi,\boldsymbol{\gamma})$ é dada por

$$\pi(\psi,\boldsymbol{\gamma}) = f_1(\psi)\prod_{i=1}^{m} g_{i+1}(\gamma_i).$$

Este resultado é apresentado em Berger e Bernardo (2005).

3.1 Inferência para os parâmetros do modelo

No modelo (3), consideramos θ_1 como parâmetro de interesse e θ_2 como parâmetro *nuisance*. Usando a metodologia descrita na Seção 3, a distribuição *a priori* de referência condicional para θ_2 dado θ_1 é $\pi(\theta_2 \mid \theta_1) = g_2(\theta_2) \propto 1$ e a distribuição *a priori* marginal para θ_1 é $\pi(\theta_1) = f_1(\theta_1) \propto \frac{1}{\theta_1}$. Assim a distribuição *a priori* de referência conjunta para θ_1 e θ_2 é dada por,

$$\pi(\theta_1,\theta_2) \propto \frac{1}{\theta_1}. \tag{13}$$

A distribuição *a posteriori* de referência conjunta, construída a partir da função de verossimilhança (5) e da distribuição *a priori* de referência conjunta para θ_1 e θ_2 (13), é dada por,

$$\pi(\theta_1,\theta_2|\boldsymbol{t},\boldsymbol{x}) \propto \frac{1}{\theta_1^{n+1}} exp\left\{-n\bar{x}\theta_2 - \frac{1}{\theta_1}\sum_{i=1}^{n} t_i e^{-\theta_2 x_i}\right\}. \tag{14}$$

A partir de (14) obtemos as distribuições *a posteriori* condicionais completas para θ_1 e θ_2, dadas respectivamente por,

$$\pi(\theta_1|\theta_2, \boldsymbol{t}, \boldsymbol{x}) \propto \frac{1}{\theta_1^{n+1}} \exp\left\{-\frac{1}{\theta_1}\sum_{i=1}^{n} t_i e^{-\theta_2 x_i}\right\}$$

e

$$\pi(\theta_2|\theta_1, \boldsymbol{t}, \boldsymbol{x}) \propto \exp\left\{-n\bar{x}\theta_2 - \frac{1}{\theta_1}\sum_{i=1}^{n} t_i e^{-\theta_2 x_i}\right\}.$$

No caso multiparametrico, sejam $f(t|\theta_1, \boldsymbol{\theta}_{-1}, x)$ o modelo probabilístico com m parâmetros e θ_1 a quantidade de interesse, para $\boldsymbol{\theta}_{-1} = (\theta_2, \ldots, \theta_m)$. A função de densidade (3) é dada por,

$$f(t|\theta_1, \boldsymbol{\theta}_{-1}, x) = \frac{1}{\theta_1 e^{\mathbf{x}'\boldsymbol{\theta}_{-1}}} \exp\left\{-\frac{t}{\theta_1 e^{\mathbf{x}'\boldsymbol{\theta}_{-1}}}\right\},$$

a função a priori de referência conjunta relativa a ordenação θ_1 e $\boldsymbol{\theta}_{-1}$ é dada por,

$$\pi(\theta_1, \boldsymbol{\theta}_{-1}) \propto \frac{1}{\theta_1},$$

e a distribuição a posteriori de referência conjunta é dada por

$$\pi(\theta_1, \boldsymbol{\theta}_{-1}|\mathbf{t}, \mathbf{x}) \propto \frac{1}{\theta_1^{n+1}} \exp\left\{-\sum_{i}^{n} \mathbf{x}_i'\boldsymbol{\theta}_{-1} - \frac{1}{\theta_1}\sum_{i=1}^{n} t_i e^{-\mathbf{x}_i'\boldsymbol{\theta}_{-1}}\right\}.$$

3.2 Inferência para a função de sobrevivência do modelo

Considerando a reparametrização dada em (10), na qual S e ϕ são parâmetros de interesse e *nuisance*, respectivamente, temos que $\pi(S) = f_1(S) \propto -\frac{1}{S\log(S)}$ e $\pi(\phi|S) = g_2(\phi) \propto 1$. Portanto, a distribuição *a priori* conjunta para S e ϕ é

$$\pi(S, \phi) \propto -\frac{1}{S\log(S)}. \tag{15}$$

A Figura 1 mostra o comportamento da distribuição *a priori* de referência para o parâmetro S.

A distribuição *a posteriori* de referência conjunta para S e ϕ, construída a partir da função de verossimilhança dada em (12) e da distribuição *a priori* de referência conjunta dada em (15) é dada por,

$$\pi(S, \phi|\boldsymbol{t}, \boldsymbol{x}) \propto S^{A(\phi)-1}(-\log(S))^{n-1} \exp\left[\left(nx_0 - \sum_{i=1}^{n} x_i\right)\phi\right], \tag{16}$$

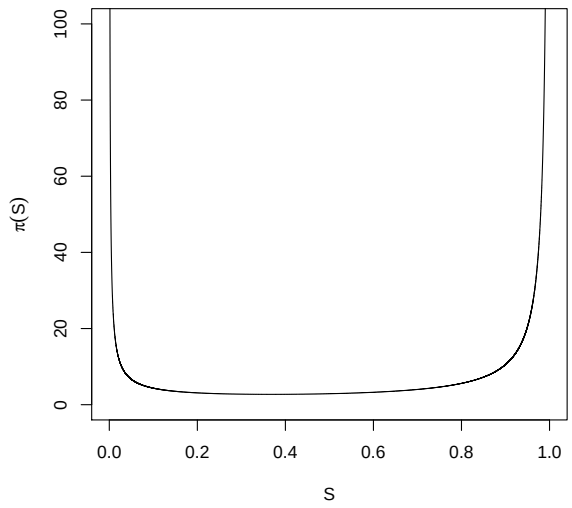

Figura 1 - Distribuição *a priori* para S.

sendo $A(\phi) = \frac{1}{t_0}\sum_{i=1}^{n} t_i e^{-(x_i - x_0)\phi}$, $0 \leq S \leq 1$ e $-\infty \leq \phi \leq \infty$.

As distribuições *a posteriori* condicionais completas para S e ϕ são dadas respectivamente por,

$$\pi(S|\phi, \boldsymbol{t}, \boldsymbol{x}) \propto S^{A(\phi)-1}(-\log(S))^{n-1}$$

e

$$\pi(\phi|S, \boldsymbol{t}, \boldsymbol{x}) \propto S^{A(\phi)-1} \exp\left[\left(nx_0 - \sum_{i=1}^{n} x_i\right)\phi\right].$$

A distribuição *a posteriori* condicional completa, por exemplo para S, é encontrada integrando a densidade *a posteriori* conjunta dada em (16) com relação aos demais parâmetros. O mesmo procedimento é aplicado para θ_1 e θ_2. A solução analítica para estas equações é obtida integrando as expressões. Como tais integrações são complexas, o método MCMC (Gamerman e Lopes, 2006) é utilizado para, a partir da distribuição *a posteriori* conjunta, obtermos as estimativas para os parâmetros de interesse.

Considerando o caso multiparamétrico em que temos os parâmetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ associados ao modelo dado em (3), a reparametrização $S = \exp\left\{-\frac{t_0}{\theta_1 e^{\mathbf{x}'\boldsymbol{\theta}_{-1}}}\right\}$ e $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{m-1}) = \boldsymbol{\theta}_{-1}$, pelo mesmo procedimento

apresentado acima, obtemos a função a *priori* conjunta para S e $\boldsymbol{\phi}$, dada por,

$$\pi(S, \boldsymbol{\phi}) \propto -\frac{1}{S \log(S)}.$$

A distribuição a posteriori de referência conjunta de S e $\boldsymbol{\phi}$ é dada por

$$\pi(S, \boldsymbol{\phi}|\boldsymbol{t}, t_0, \boldsymbol{x}, \boldsymbol{x_0}) \propto S^{A(\boldsymbol{\phi})-1}[-\log(S)]^{n-1} \exp\left[\sum_{i=1}^{n} (\boldsymbol{x_0} - \boldsymbol{x_i})\, \boldsymbol{\phi}\right],$$

sendo $A(\boldsymbol{\phi}) = \frac{1}{t_0} \sum_{i=1}^{n} t_i e^{-(\boldsymbol{x_i} - \boldsymbol{x_0})\boldsymbol{\phi}}$.

4 Um estudo de simulação e aplicações

Nas subseções seguintes, apresentamos um estudo com dados simulados e duas aplicações em dados reais, sendo uma delas com dados censurados. Os resultados foram obtidos no *software* R *Development Core Team* (2013).

4.1 Um estudo de simulação

O estudo de simulação foi realizado com amostras de diferentes tamanhos: $n = 30$, 50, 100 e 200. No modelo (3), fixamos $\theta_1 = 80$ e $\theta_2 = -0,5$ e geramos tempos de vida usando o teorema fundamental da transformação de probabilidades. Para os diferentes n, geramos uma cadeia com 100.000 iterações das distribuições *a posteriori* marginais de θ_1 e θ_2 utilizando o algoritmo de Metropolis-Hastings (Gamerman e Lopes, 2006). Para cada cadeia, descartamos as 5.000 primeiras iterações para evitar influência dos valores iniciais e também, consideramos um salto de 10 iterações para evitar a correlação entre os valores gerados, resultando em um total de 9.500 valores gerados. As estimativas obtidas são mostradas na Tabela 1. Pelo método de Geweke (Geweke, 1992) existe indicação de convergência das cadeias geradas.

Tabela 1 - Resumos *a posteriori* para θ_1 e θ_2

	θ_1			θ_2		
n	Média	DP	IC (95%)	Média	DP	IC (95%)
30	88,94	13,03	[67,27 ; 117,78]	-0,420	0,143	[-0,699 ; -0,142]
50	87,86	10,01	[70,52 ; 109,57]	-0,497	0,112	[-0,717 ; -0,275]
100	77,18	7,85	[63,42 ; 94,28]	-0,480	0,098	[-0,674 ; -0,289]
200	81,81	6,07	[71,08 ; 93,96]	-0,507	0,067	[-0,636 ; -0,376]

Como estamos interessados na função de sobrevivência, para os mesmos conjuntos de valores gerados, obtivemos uma amostra das distribuições *a posteriori* marginais de S e ϕ, para cada n, no tempo $t_0 = 100$ e no valor $x_0 = 0,5$. As estimativas para os parâmetros S e ϕ são mostradas na Tabela 2.

Tabela 2 - Resumos *a posteriori* para S e ϕ

	S			ϕ		
n	Média	DP	IC-95%	Média	DP	IC (95%)
30	0,249	0,056	[0,148 ; 0,370]	-0,418	0,137	[-0,697 ; -0,149]
50	0,231	0,043	[0,153 ; 0,320]	-0,497	0,109	[-0,710 ; -0,288]
100	0,192	0,036	[0,127 ; 0,268]	-0,481	0,096	[-0,671 ; -0,296]
200	0,206	0,025	[0,159 ; 0,259]	-0,505	0,066	[-0,635 ; -0,374]

Observamos que quando o tamanho de amostra é pequeno ou moderado as estimativas estão razoáveis; porém, quando aumentamos o tamanho da amostra, as estimativas ficam mais próximas dos verdadeiros valores. Este fato ocorre para todos os parâmetros. Notamos também que as estimativas de θ_2 e de ϕ estão muito próximas, o que já era esperado, pois estes parâmetros são iguais (pelas Tabelas 1 e 2).

Calculamos a probabilidade de cobertura e o tamanho do intervalo de credibilidade (IC), para os diversos tamanhos de amostra, para os parâmetros θ_1 e θ_2 (Tabela 3) e S e ϕ (Tabela 4). Para o cálculo destas medidas, repetimos 5.000 vezes a geração dos tempos e a estimação, como descrito no início desta seção. Note que as probabilidades de cobertura aumentam e os tamanhos dos intervalos de credibilidade diminuem com o aumento dos tamanhos das amostras. Notamos que os resultados encontrados para θ_2 e ϕ estão muito próximos, indicando que a escolha de estimar S ou θ_1 não afeta a estimação do outro parâmetro.

Tabela 3 - Probabilidade de cobertura e amplitude dos IC's para θ_1 e θ_2

	θ_1		θ_2	
n	prob. de cobertura	amplitude	prob. de cobertura	amplitude
30	0,9562	60,5873	0,9227	0,7763
50	0,9544	46,2565	0,9544	0,5838
100	0,9519	32,0341	0,9541	0,3982
200	0,9671	22,4740	0,9574	0,2815

Tabela 4 - Probabilidade de cobertura e amplitude dos IC's para S e ϕ

	S		ϕ	
n	prob. de cobertura	amplitude	prob. de cobertura	amplitude
30	0,9473	0,2516	0,9452	0,7750
50	0,9431	0,1963	0,9343	0,5798
100	0,9413	0,1401	0,9608	0,4059
200	0,9502	0,1001	0,9574	0,2815

4.2 Aplicação com os Dados de Feigl e Zelen

Os dados de Feigl e Zelen (1965) consistem de tempos de sobrevivência (em semanas) de pacientes com leucemia e uma variável concomitante WBC, representando a contagem de glóbulos brancos por 10.000 unidades na célula dos pacientes. Baseado no exame das células com leucemia, os pacientes foram classificados como AG positivo e AG negativo (Tabela 5).

Tabela 5 - Dados de pacientes com leucemia

WBC/10.000	Tempo	AG+	WBC/10.000	Tempo	AG-
0,23	65	1	0,44	56	0
0,075	156	1	0,3	65	0
0,43	100	1	0,4	17	0
0,26	134	1	0,15	7	0
0,6	16	1	0,9	16	0
1,05	108	1	0,53	22	0
1	121	1	1	3	0
1,7	4	1	1,9	4	0
0,54	39	1	2,7	2	0
0,7	143	1	2,8	3	0
0,94	56	1	3,1	8	0
3,2	26	1	2,6	4	0
3,5	22	1	2,1	3	0
10	1	1	7,9	30	0
10	1	1	10	4	0
5,2	5	1	10	43	0
10	65	1			

Para ajustar o modelo dado em (3), assumimos a covariável x_1 como sendo o valor da contagem de glóbulos brancos medidos em unidades de 10.000 e a covariável x_2 como sendo uma covariável *dummy* assumindo 1 para indicar o grupo do pacientes com a característica AG positivo e assumindo 0 para indicar o grupo dos pacientes com a característica AG negativo.

Para o ajuste do modelo dado em (16) consideramos um tempo de aproximadamente dois anos ($t_0 = 96$) e a função de sobrevivência em pacientes com a contagem de glóbulos brancos igual a 6.000 unidades ($x_0 = 0,6$) e com a característica AG positivo, sendo que S é dado por $S = \exp\left(-\frac{96}{\theta_1 \exp(0,6\theta_2+\theta_3)}\right)$.

Analogamente ao estudo de simulação, construímos as cadeias a partir da distribuição *a posteriori*. As estimativas obtidas para os parâmetros foram baseadas em cadeias geradas de tamanho 100.000, com *burn in* de 5.000 e salto igual a 10, resultando em uma amostra de tamanho 9.500. O método de Geweke (1992) indicou a convergência das cadeias geradas.

As medidas resumo para os parâmetros θ_1, θ_2 e θ_3 e para S, ϕ_1 e ϕ_2 são apresentadas, respectivamente, nas Tabelas 6 e 7. A Figura 2 mostra a densidade

de probabilidade *a posteriori* para os parâmetros dos modelos.

Tabela 6 - Resumos *a posteriori* para θ_1, θ_2 e θ_3

	Média	DP	IC (95%)	Geweke
θ_1	23,0700	5,0119	[14,5192 ; 34,1714]	1,6869
θ_2	-0,0637	0,0369	[-0,1312 ; 0,0118]	-0,8078
θ_3	1,1410	0,2689	[0,6153 ; 1,6601]	-0,4346

Tabela 7 - Resumos *a posteriori* para S e ϕ

	Média	DP	IC (95%)	Geweke
S	0,2390	0,1473	[0,0529 ; 0,5887]	0,9158
ϕ_1	-0,0642	0,0378	[-0,1396 ; 0,0091]	-0,9599
ϕ_2	1,0660	0,3141	[0,4745 ; 1,7039]	0,0882

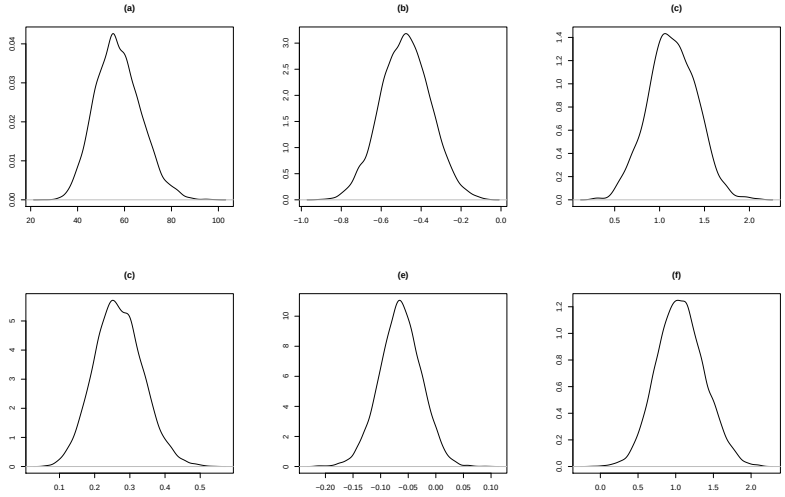

Figura 2 - Densidade de probabilidade de θ_1 em (a), θ_2 em (b), θ_3 em (c), S em (d), ϕ_1 em (e) e ϕ_2 em (f).

Os resumos *a posteriori* para os parâmetros (Tabelas 6 e 7) mostram que os resultados obtidos para θ_2 e ϕ_1 e para θ_3 e ϕ_2, são próximos, o que já era esperado, visto a reparametrização adotada em (10). Observamos para um indivíduo que apresente a quantidade de glóbulos brancos igual a 6.000 unidades e seja do grupo AG positivo a probabilidade dele sobreviver além do tempo $t_0 = 96$ é de $0,2390$

com um intervalo de credibilidade de $[0,0529\ ;0,5887]$ que esta dentro do espaço paramétrico.

Podemos notar que as densidades *a posteriori* para os parâmetros ϕ_1, ϕ_2, θ_1, θ_2 e θ_3 tem comportamento próximo à distribuição normal (Figura 2) e o parâmetro S tem um comportamento próximo à distribuição beta.

Na Figura 3 apresentamos as curvas estimadas pelo estimador de Kaplan-Meier (EKM) juntamente com as curvas obtidas através da função de sobrevivência dada em (4), da média *a posteriori* dos parâmetros e a média de x_1, igual a 2,917, como sendo a primeira covariável.

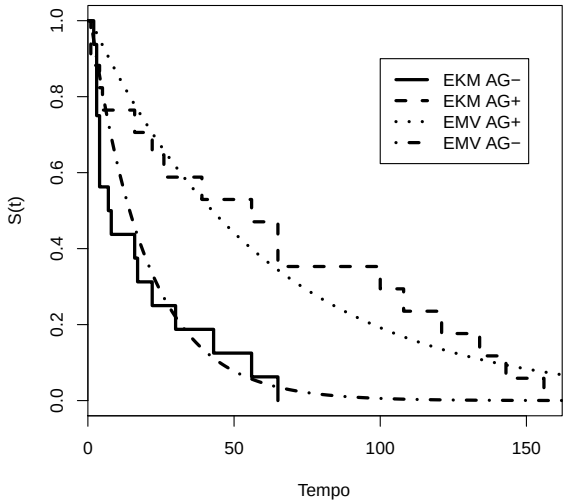

Figura 3 - Curvas de sobrevivência estimadas pelo EKM e através do estimador de máxima verossimilhança.

4.3 Aplicação com dados de câncer de pulmão

O conjunto de dados desta aplicação representa os tempos de vida (em anos) de 228 pacientes com câncer de pulmão, da Clínica Mayo, sendo que 63 tempos foram censurados. A covariável utilizada é a idade (em anos). Os dados, de nome lung (Loprinzi et al., 1994), estão disponíveis no *software* R Development Core Team (2013).

Para o ajuste do modelo dado em (16) consideramos o tempo de um ano, $(t_0 = 1)$, e a covariável com sendo igual a 63, $x = 63$, ou seja, estamos interessado em determinar a probabilidade de um paciente com 63 anos sobreviver além de um ano. A função de sobrevivência, S, do paciente é dada por $S = \exp\left(-\frac{1}{\theta_1 \exp(63\theta_2)}\right)$.

As estimativas bayesianas foram obtidas a partir da construção das cadeias geradas da distribuição *a posteriori*, analogamente ao estudo de simulação (Seção 4.1).

As medidas resumo para os parâmetros θ_1 e θ_2 e, para S e ϕ são apresentadas, respectivamente, nas Tabelas 8 e 9; a Figura 4 mostra a densidade de probabilidade *a posteriori* para os parâmetros dos modelos.

Tabela 8 - Resumos *a posteriori* para θ_1 e θ_2

	Média	DP	IC-95%	Geweke
θ_1	3,3370	0,2442	[2,8808 ; 3,8506]	0,6359
θ_2	-0,0169	0,0003	[-0,0175 ; -0,0163]	-1,1589

Tabela 9 - Resumos *a posteriori* para S e ϕ

	Média	DP	IC-95%	Geweke
S	0,4185	0,0287	[0,3633 ; 0,4745]	0,2638
ϕ	-0,0170	0,0002	[-0,0174 ; -0,0166]	-0,4036

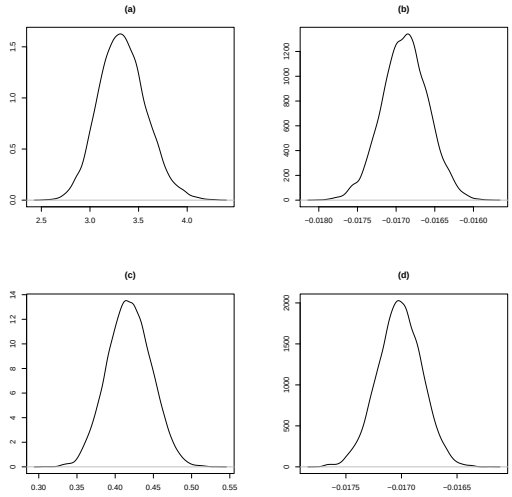

Figura 4 - Densidade de probabilidade de θ_1 em (a), θ_2 em (b), θ_3 em (c), S em (c) e ϕ em (d).

Conclusão

Neste artigo apresentamos uma breve revisão da metodologia de análise de referência, pois esta fornece um método para obtenção de uma distribuição *a priori* não informativa. Além disso, esta metodologia generaliza algumas alternativas, como por exemplo a distribuição *a priori* de Jeffreys (1961). A partir do modelo proposto por Feigl e Zelen (1965) e do modelo (11) construímos as distribuição *a posteriori* de referência dos modelos. No estudo de simulação obtivemos estimativas pontuais próximas dos verdadeiros valores (fixados) e observamos que para amostras pequenas, as probabilidades de cobertura ficaram próximas de 95% e as amplitudes dos intervalos de credibilidade diminuíram com o aumento do tamanho da amostra (Seção 4.1). A importância prática do modelo pode ser vistas nas aplicações com os conjuntos de tempos de vida de pacientes com leucemia, usado por Feigl e Zelen (1965), e câncer de pulmão.

DIAS, T. C. M.; TOMAZELLA, V. L. D.; MILANI, E. A. Objective Bayesian estimation for the regression model Feigl and Zelen. *Rev. Bras. Biomet.,* São Paulo, v.31, n.1, p.116-131, 2013.

- *ABSTRACT: In bayesian inference the specification of prior distribution for the parameters of interest can be complex, vague or highly subjective. This distribution expresses the knowledge or ignorance about the parameters. But it is important to identify the mathematical form of an initial function that has minimal effect on the inference a posteriori, which leads to the use of Bayesian methodology objective. In this article we consider the methodology of analysis of objective bayesian reference introduced by Bernardo (1979), for the construction of posterior distribution in order to estimate the parameters and the survival function of the exponential regression model proposed by Feigl and Zelen (1965) and used the method via simulation Monte Carlo Markov Chain (MCMC) to give the results a posteriori parameters of interest.*

- *KEYWORDS: Reference Bayesian analysis; regression model; referencepriori, survival analysis.*

Referências

BERGER, J. O.; BERNARDO, J. M. Estimating a product of means: Bayesian analysis with reference priors. *J. Am. Stat. Assoc.*, New York, v.84, p.200–207, 1989.

BERGER, J. O.; BERNARDO, J. M. Ordered group reference priors with applications to a multinomial problem. *Biometrika*, Oxford, v.79, p.25–37, 1992a.

BERGER, J. O.; BERNARDO, J. M. Reference priors in a variance components problem, *Bayesian Analysis in statistics and econometrics*. New York: Spring - Verlag, 1992b. p.323–340.

BERGER, J. O.; BERNARDO, J. M. On the development of reference priors. *Bayesian Stat.*, Oxford, v.4, p.35–60, 1992c.

BERGER, J. O.; BERNARDO, J. M.; SUN, D. *Reference priors from first principles: A general definition*, Tech. Rep., SAMSI, NC, USA, 2005.

BERNARDO, J. M. Reference posterior distributions for Bayesian inference. *J. R. Stat. Soc.* B, Chichester, v.41, p.113-147, 1979.

BERNARDO, J. M. Reference analysis. In: DEY, D. K.; RAO, C. R. (Ed.) *Handb. stat.*, Amsterdan: Elsevier, 2005, v.25, p.17-90.

BERNARDO, J. M.; SMITH, A. F. M. *Bayesian theory*. Chichester: Wiley, 1994.

BUCKLEY, J.; JAMES, I. Linear regression with censored data. *Biometrika*, Oxford, v.66, p.429–436, 1979.

COLLETT, D. *Modelling survival data in medical research*. London: Chapman & Hall, 1994. 408p.

FEIGL, P.; ZELEN, M. Estimation of exponencial survival probabilities with concomitant information. *Biometrics*, Chichester, v.21, p.826–838, 1965.

GAMERMAN, D.; LOPES, H. F. *Markov chain Monte Carlo, stochastic simulation for Bayesian inference*. Londres: Chapman & Hall/CRC, 2006.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *In Bayesian Stat.*, Oxford, v.4, p.169–193, 1992.

JEFFREYS, H. *Theory of probability*. 3rd. ed, Oxford: Claredon Press, 1961.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New York: John Wiley & Sons, 1982.

LOPRINZI, C.; LAURIE, J.; WIEAND, H. Prospective evaluation of prognostic variables from patient completed questionnaires. North central cancer treatment group, *J. Clin. Oncol.*, Alexandria, v.12, n.3, p.601–607, 1994.

MILLER, R. G. Least squares regression with censored data. *Biometrika*, Oxford, v.63, p.449-464, 1976.

MILLER, R. G. *Survival analysis*. New York: John Wiley & Sons, 1981. 238p.

R Development Core Team. *A language and environment for statistical computing*, version 3.0.0, The R Foundation for Statistical Computing, 2013.

ZIPPIN, C.; ARMITAGE, P. Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, Chichester, v.22, p.665-672, 1966.

Recebido em 15.01.2013.

Aprovado após revisão em 15.07.2013.