UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Uso de Métodos MCMC para Análise Bayesiana de Dados de Sobrevivência na Presença de Covariáveis

Cillene Nunes de Souza
Jorge Alberto Achcar
Josmar Mazuchelli

Nº 64

NOTAS

São Carlos - SP

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2577

Uso de Métodos MCMC para Análise Bayesiana de Dados de Sobrevivência na Presença de Covariáveis

Cillene Nunes de Souza
Jorge Alberto Achcar
Josmar Mazuchelli

Nº 64

NOTAS

Série Estatística

São Carlos – SP
Mai./2001

USO DE MÉTODOS MCMC PARA ANÁLISE BAYESIANA DE DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE COVARIÁVEIS

CILLENE NUNES DE SOUZA

JORGE ALBERTO ACHCAR

SCE ,ICMC /USP, C. P. 668, 13560-970, São Carlos- S. P.

JOSMAR MAZUCHELLI

DES,Universidade Estadual de Maringá, Maringá, PR.,Brasil

Resumo

Neste artigo, apresentamos uma análise Bayesiana para modelos log-lineares com diferentes distribuições para a variável erro e a presença de observações censuradas. Sumários a posteriori de interesse como médias, desvios padrões e intervalos de credibilidade são obtidos a partir de estimadores de Monte Carlo usando amostras da distribuição a posteriori conjunta geradas pelo amostrador de Gibbs. Também introduzimos técnicas Bayesianas para discriminar diferentes modelos. Um exemplo com dados médicos ilustra a metodologia proposta.

Palavras-chave: modelo log-linear, dados censurados, análise Bayesiana, amostrador de Gibbs.

1 Introdução

Em várias aplicações, podemos ter dados representados pelos tempos até a ocorrência de um evento como falha de um componente ou morte de um paciente num ensaio clínico. Dados desse tipo são comuns em medicina, biologia, engenharia, indústria entre várias outras áreas.

Nesta situação é usual a existência de observações censuradas ou truncadas (censuras à direita, à esquerda ou por intervalo). Também é comum a presença de covariáveis ou fatores de risco associados a cada unidade.

Supor que T seja uma variável aleatória contínua, não-negativa, representando o tempo de sobrevivência de uma unidade com função densidade de probabilidade $f(t)$, e função de sobrevivência $S(t) = P(T > t)$.

A função de risco ou taxa instantânea de falha é dada por $h(t) = f(t)/S(t)$ (ver por exemplo, Lawless, 1982).

Em geral, nas aplicações médicas ou de engenharia, os pesquisadores modelam $h(t)$ por uma função paramétrica levando em consideração a forma da curva (risco constante, crescente, decrescente, em forma de U, etc) o que leva a diferentes distribuições paramétricas como a distribuição exponencial, a distribuição de Weibull, a distribuição Gama-Generalizada (ver por exemplo, Kalbfleish e Prentice 1980; ou Lawless 1982).

Na presença de covariáveis, a literatura apresenta alguns modelos de regressão como o modelo log-linear na forma

$$Y = \ln(T) = \mu(\mathbf{x}) + \sigma Z \qquad (1)$$

onde **x** é um vetor **1 x p** de covariáveis e Z é uma variável erro com uma dada distribuição de probabilidade. Uma distribuição muito usada para Z é dada pela densidade de valor extremo padrão $f(z) = \exp(z - \exp(z))$, $-\infty < z < \infty$ que implica numa distribuição Weibull para o tempo de sobrevivência. Observar que quando $\sigma = 1$, temos uma distribuição exponencial para T. Uma escolha usual para $\mu(\mathbf{x})$ é dada por

$$\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \qquad (2)$$

Outros modelos de regressão são introduzidos na literatura. Um caso especial é dado pelo modelo de riscos proporcionais (Cox 1972) dado por

$$h(t \mid \mathbf{x}) = h_0 \exp(\mathbf{x}'\boldsymbol{\beta}) \qquad (3)$$

onde $h_0(t)$ é desconhecida. Neste caso, temos um modelo semi-paramétrico, pois não assumimos uma distribuição paramétrica para o tempo de sobrevivência T. Em geral, inferências clássicas para os modelos (1) e (3) são obtidas a partir de resultados assintóticos considerando dados censurados. Esses resultados podem não ser muito precisos quando o tamanho amostral é pequeno ou quando a proporção de dados censurados é muito grande.

Neste caso, podemos usar métodos Bayesianos baseados em técnicas de simulação MCMC (Monte Carlo em Cadeias de Markov) para obter inferências precisas para os parâmetros do modelo considerado (ver por exemplo, Gelfand e Smith 1990). O uso dessas técnicas também permite a utilização de modelos paramétricos mais sofisticados envolvendo um grande número de parâmetros, como por exemplo, uma mistura de distribuições paramétricas. Além disso, o enfoque Bayesiano fornece vários critérios para verificação do ajuste do modelo aos dados.

2 Análise Bayesiana para o Modelo de Regressão Log-Linear

Assumindo o modelo log-linear (1) com $\mu(\mathbf{x})$ dado em (2) e dados com censuras à direita, a função de verossimilhança é dada por

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta} \qquad (4)$$

onde $\delta_i=1$ para uma observação completa e $\delta_i=0$ para uma observação censurada.

Assumindo independência a priori, considerar distribuições a priori dadas por

$$(i)\sigma \sim IG(a,b) \qquad ; \text{ a, b conhecidos}$$

$$(ii)\beta_j \sim N(c_j, d_j^{\,2}) \qquad , \text{c}_\text{j}, \text{d}_\text{j} \text{ conhecidos, j} = 0,1,...,\text{p} \tag{5}$$

onde $IG(a,b)$ denota uma distribuição gama inversa com média $b/(a-1)$ e variância $b^2/(a-1)^2\,(a-2)$ e $N(c_j, d_j^2)$ denota uma distribuição normal com média c_j e variância d_j^2.

A partir de valores escolhidos para os hiperparâmetros das distribuições a priori podemos obter inferências a posteriori de interesse a partir de amostras geradas da distribuição a posteriori conjunta (Gelfand e Smith, 1990) e quando necessário usamos o método de ARMS (adaptive rejection Metropolis sampling)(ver Gilks et al, 1995) . Isso pode ser obtido considerando diferentes escolhas para a distribuição da variável aleatória Z no modelo (1). Casos especiais são dados pela distribuição de valor extremo padrão $f(z) = \exp(z - \exp(z))$, $-\infty < z < \infty$ (distribuição de Weibull para T); pela distribuição normal padronizada $f(z) \propto \exp(-z^2/2)$, $-\infty < z < \infty$ (distribuição log-normal para T); pela distribuição logística padrão $f(z) \propto \exp(z)/(1+\exp(z))^2$, $-\infty < z < \infty$ (distribuição log-logística para T); ou uma distribuição log-gama $f(z) \propto \exp\{\sqrt{k}z - k\exp(z/\sqrt{k})\}$, $-\infty < z < \infty$ (distribuição gama generalizada para T). Observar que a distribuição gama generalizada é um supermodelo que engloba várias distribuições usuais como casos especiais. Em especial, se $k = 1$ temos a distribuição de Weibull para T e para $k \uparrow \infty$, temos a distribuição log-normal .

2.1 Distribuição de Valor Extremo para Z

Como um caso especial, considerar a distribuição valor extremo padrão para Z em (1) na presença de observações censurados. A função de verossimilhança para $\boldsymbol{\beta}$ e σ é dada por

$$L(\boldsymbol{\beta},\sigma)=\prod_{i\in D}\frac{1}{\sigma}\exp\left\{\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}-\exp\left(\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\}\prod_{i\in C}\exp\left\{-\exp\left(\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\} \quad (6)$$

onde D denota o conjunto de observações completas e C denota o conjunto de observações incompletas ou censuradas e $\mathbf{x}_i'\boldsymbol{\beta}=\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi}$, i=1,...,n. Isto é,

$$L(\boldsymbol{\beta},\sigma)=\frac{1}{\sigma^r}\exp\left\{\frac{1}{\sigma}\left[\sum_{i\in D}y_i-r\beta_0-\sum_{j=1}^{p}\sum_{i\in D}\beta_j x_{ji}\right]\right\}\exp\left\{-\sum_{i=1}^{n}\exp\left(\frac{y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}}{\sigma}\right)\right\} \quad (7)$$

onde $r=\sum_{i=1}^{n}\delta_i$ é o número de observações não censuradas.

Aplicando o logaritmo na equação (7), temos

$$l(\boldsymbol{\beta},\sigma)=-r\log(\sigma)+\frac{1}{\sigma}\left[\sum_{i\in D}y_i-r\beta_0-\sum_{j=1}^{p}\sum_{i\in D}\beta_j x_{ji}\right]-\sum_{i=1}^{n}\exp\left(\frac{y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}}{\sigma}\right) \quad (8)$$

Para a utilização do algoritmo ARMS, utilizamos as distribuições a priori dadas em (5) e a distribuição condicional para cada parâmetro.

Os logaritmos das distribuições condicionais necessárias para o algoritmo ARMS (Gibbs sampling) são dados por,

$$(i)\, l(\sigma) = -r\log(\sigma) + \frac{1}{\sigma}\left[\sum_{i\in D} y_i - r\beta_0 - \sum_{j=1}^{p}\sum_{i\in D}\beta_j x_{ji}\right] - \sum_{i=1}^{n}\exp\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}}{\sigma}\right) \quad (9)$$

$$(ii)\, l(\beta_0) = -\frac{r\beta_0}{\sigma} - \sum_{i=1}^{n}\exp\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}}{\sigma}\right) \quad (10)$$

$$(iii) l(\beta_j) = \frac{1}{\sigma}\sum_{j=1}^{p}\sum_{i\in D}\beta_j x_{ij} - \sum_{i=1}^{n}\exp\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}}{\sigma}\right), \text{j} = 1,\ldots,\text{p}. \quad (11)$$

De maneira análoga, considerando outras distribuições para Z, geramos amostras da distribuição a posteriori conjunta a partir das distribuições a posteriori condicionais para cada parâmetro.

3 Discriminação de Modelos

Para um conjunto de dados podemos ter interesse em discriminar dois ou mais modelos. Em especial, podemos considerar diferentes distribuições para a variável erro Z no modelo log-linear (1) ou a inclusão de algumas covariáveis para a análise estatística. Para discriminar modelos, podemos usar técnicas tradicionais como os gráficos de resíduos ou testes de ajuste. Uma forma alternativa é dada pelo uso de densidades preditivas para selecionar o melhor modelo. A densidade preditiva ordenada para uma observação t_i é dada por,

$$c_i = f(t_i \mid \mathbf{t}_{(i)}, \mathbf{x}_i) = \int f(t_i \mid \boldsymbol{\theta}, \mathbf{x}_i)\pi(\boldsymbol{\theta} \mid \mathbf{t}_{(i)}, \mathbf{x}_{(i)})d\boldsymbol{\theta} \quad (12)$$

onde $\pi(\boldsymbol{\theta} \mid t_{(i)}, x_i)$ é a densidade a posteriori de $\boldsymbol{\theta}$ dado $\mathbf{t}_{(i)} = (t_1,\ldots,t_{i-1},t_{i+1},\ldots,t_n)$ e $\mathbf{x}_i$ é o vetor de covariáveis associado à $t_{(i)}$.

Usando as amostras geradas pelo amostrador de Gibbs, podemos aproximar $f(t_i \mid t_{(i)}, x_i)$ pela estimativa de Monte Carlo

$$\hat{f}\left(t_i \mid t_{(i)}, \hat{x}_i\right) = \frac{1}{S}\sum_{r=1}^{S} f(t_i \mid x_i, \theta^{(s)}) \tag{13}$$

onde S é o número de amostras geradas pelo algoritmo.

Podemos usar c_i para selecionar modelos. Deste modo, consideramos gráficos de c_i versus i (i=1,2,...,n) para diferentes modelos. Considerando os valores observados, o melhor modelo tem, em geral, maiores valores de c_i.

Além disso, podemos escolher o modelo cujo produto de c_i seja máximo, ou seja, $c(l) = \prod c_i(l)$,onde l indexa modelos.

Outros critérios também poderiam ser utilizados na discriminação dos modelos propostos. Um desses critérios é dado pelo critério da mínima informação de Akaike (AIC) (ver por exemplo, Klein & Moeschberger , 1997) dado por

$$AIC = -2\log(L) + 2p \tag{14}$$

onde p é o número de parâmetros.

4 Uma Aplicação : Dados de Câncer na Laringe

Considerar os dados da Tabela 1(anexo) de 90 pacientes com câncer na laringe (dados introduzidos por Kardaum, 1983). Para esse conjunto de dados, T representa o intervalo em anos, entre o primeiro tratamento e a ocorrência de morte ou término do estudo em primeiro de janeiro de 1983. Alguns fatores de risco associados a esses pacientes são dados pela idade do paciente na data do diagnóstico da doença e o estágio da doença.

Como existem quatro estágios da doença avaliados em termos do tipo de tumor, envolvimento nodal e graduação de metastase (ver Klein e Moeschberger, 1997),

introduzimos variáveis dummy para indicar os estágios da doença nos pacientes: x_1=1 (estágio II), 0 (e.o.p); x_2=1 (estágio III), 0 (e.o.p); x_3=1 (estágio IV), 0 (e.o.p).

Para analisar esses dados, assumimos o modelo de regressão log-linear (1) com diferentes densidades para Z e $\mu(\mathbf{x})$ dado por

$$\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad (15)$$

onde x_4 é a idade do paciente.

A partir de escolhas especiais dos valores dos hiperparâmetros a, b, c_j e d_j, j= 1, 2, 3, 4 nas distribuições a priori (5) envolvendo a opinião de especialistas e uma análise preliminar dos dados, temos na Tabela 2 os estimadores de Monte Carlo para as quantidades a posteriori de interesse para cada modelo, baseados em 2 cadeias com 11000 amostras geradas pelo amostrador de Gibbs. Em cada cadeia, descartamos as 1000 primeiras amostras ("burn-in sampling") e tomamos observações de 10 em 10 o que totaliza uma amostra final de S= 2000 amostras geradas.

A convergência do algoritmo Gibbs sampling pode ser verificada usando o critério de Gelman e Rubin (1992) e também verificando o comportamento das simulações dos parâmetros utilizando o software CODA (ver Best, Cowles e Vines, 1995). Para todos os casos foi verificada convergência.

Também temos na Tabela 2, os estimadores de máxima verossimilhança para os parâmetros e os intervalos de confiança baseados na distribuição normal assintótica dos estimadores de máxima verossimilhança e intervalos de credibilidade baseados nas amostras geradas pelo algoritmo Gibbs sampling.

Os estimadores de máxima verossimilhança foram obtidos utilizando o software SAS e as amostras geradas pelo amostrador de Gibbs foram geradas a partir do software BUGS(Gilks et al. 1994) ou Ox (ver Doornik, 1999).

Tabela 2: Estimativas clássicas e Bayesianas para os parâmetros dos modelos.

Distribuição Para T	Parâmetros	EMV	Intervalo de Confiança(95%)	Média a posteriori	Intervalo de Credibilidade(95%)
Exponencial	β_0	3.755	(1.815;5.695)	3.88	(1.950;5.923)
	β_1	-0.145	(-1.046;0.756)	-0.125	(-1.014;0.838)
	β_2	-0.648	(-1.344;0.047)	-0.66	(-1.379;0.032)
	β_3	-1.635	(-2.451;-0.855)	-1.621	(-2.400;-0.833)
	β_4	-0.019	(-0.046;0.008)	-0.02	(-0.050;0.007)
	σ	-			
Weibull	β_0	3.528	(1.756;5.299)	3.878	(1.727;6.028)
	β_1	-0.148	(-0.954;0.649)	-0.115	(-1.042;0.812)
	β_2	-0.586	(-1.211;0.392)	-0.654	(-1.373;0.065)
	β_3	-1.544	(-2.255;-0.832)	-1.609	(-2.420;-0.797)
	β_4	-0.017	(-0.04;0.006)	-0.021	(-0.048;0.006)
	σ	0.885	(0.673;1.096)	0.98	(0.721;1.238)
Logística	β_0	3.102	(1.236;4.968)	2.893	(1.518;4.272)
	β_1	-0.125	(-0.938;0.688)	-0.067	(-0.825;0.682)
	β_2	-0.805	(-1.496;-0.113)	-0.647	(-1.359;0.055)
	β_3	-1.766	(-2.599;-0.933)	-1.646	(-2.768;-0.649)
	β_4	-0.015	(-0.042;0.012)	-0.034	(-0.056;-0.012)
	σ	0.715	(0.547;0.883)	0.898	(0.805;1.007)
Lognormal	β_0	3.383	(1.550;5.215)	3.234	(1.906;4.563)
	β_1	-0.198	(-1.064;0.668)	-0.154	(-0.989;0.679)
	β_2	-0.899	(-1.611;-0.186)	-0.874	(-1.577;-0.185)
	β_3	-1.857	(-2.723;-0.990)	-1.840	(-2.661;-1.025)
	β_4	-0.018	(-0.043;0.007)	-0.016	(0.036;0.004)
	σ	1.263	(1.000;1.525)	1.343	(1.087;1.685)
Gama Generalizada	β_0	2.593	(1.817;3.368)	2.225	(2.201;2.291)
	β_1	-0.109	(-0.410;0.192)	-0.110	(-0.177;-0.012)
	β_2	-0.052	(-0.307;0.203)	-0.068	(-0.108;-0.015)
	β_3	-0.813	(-1.134;-0.431)	-0.765	(-0.894;-1.533)
	β_4	-0.004	(-0.015;0.007)	-0.004	(-0.004;-0.002)
	α	6.859	(1.116;12.603)	6.681	(5.921;6.991)
	κ	0.085	(0.004;0.167)	0.076	(0.049;0.099)

Na Figura 1, temos os gráficos das estimativas de Monte Carlo para as densidades preditivas ordenadas c_i versos i, i= 1,..., n, em cada observação t_i . A partir da Figura 1, verificamos que a distribuição Gama Generalizada para T leva, em média a maiores valores de c_i calculados nos valores observados. Isso também é comprovado a partir dos valores de $c(l) = \prod c_i(l)$ (valor máximo para a distribuição Gama Generalizada) e dos valores de AIC (ver Tabela 3).

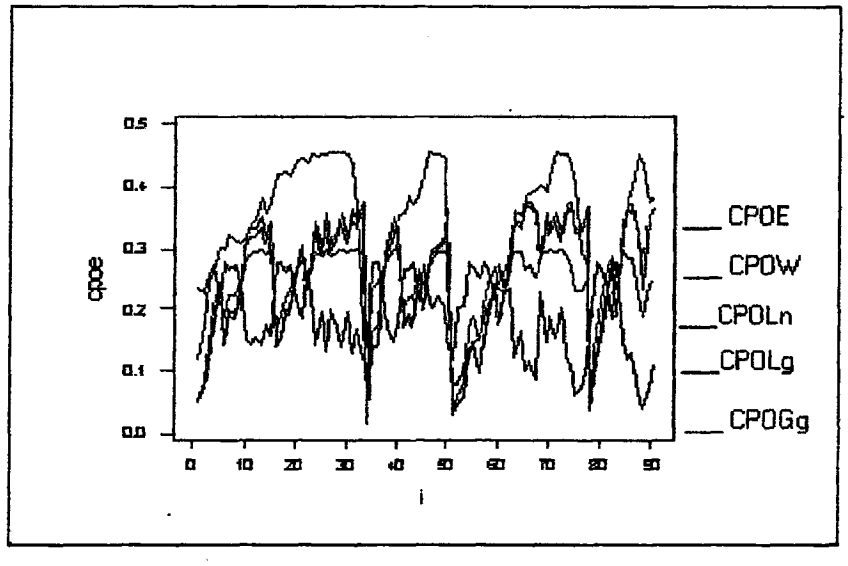

Figura 1: Gráficos de c_i versus i.

Acima, temos os gráficos de CPO para os vários modelos ajustados. Na legenda sempre teremos CPO acrescido de alguma letra que indica o modelo, por exemplo, CPOE indica a CPO do modelo Exponencial (E), CPOW (modelo Weibull) e assim sucessivamente.

Tabela 3: Valores de AIC e c(l)

Distribuição	AIC	c(l)
Exponencial	220.05	c(1) = 5.904 e-60
Weibull	219.00	c(2) = 1.144 e-58
Logística	220.99	c(3) = 7.254 e-60
Lognormal	219.99	c(4) = 1.057 e-67
Gama Generalizada	206.00	c(5) = 8.552 e -54

A partir dos resultados da Tabela 3, verificamos que a distribuição Gama Generalizada $f(z) \propto \exp\{\sqrt{k}z - k\exp(z/\sqrt{k})\}$,-∞ <z< ∞ no modelo log-linear (1) é melhor ajustada ao conjunto de dados da Tabela 1.

Assumindo a distribuição Gama Generalizada para o erro, concluímos que a covariável x_3 apresenta um efeito significativo nos tempos de sobrevivência T(ver Tabela 2). As demais covariáveis não apresentam efeito significativo como observamos nos intervalos de credibilidade para os parâmetros de regressão.

5 Mistura de Distribuições Paramétricas

Em muitas aplicações com dados de sobrevivência podemos ter funções de risco em forma de U ou com multimodalidade. Quando isto acontece, os modelos paramétricos usuais podem não ser apropriados e precisamos de modelos mais sofisticados. Uma possibilidade é considerar uma mistura de distribuições para a variável Z no modelo log-linear (1) dada por,

$$f(z) = \sum_{j=1}^{k} p_j f(z \backslash \theta_j) \tag{16}$$

onde $\sum_{j=1}^{k} p_j = 1$ e θ_j é um vetor de parâmetros associados com a j-ésima distribuição de probabilidade componente na mistura.

Casos especiais são dados com uma mistura de distribuições normais com densidades $f(z \mid \mu_j, \beta_j) \propto \exp\left\{-\frac{1}{2\sigma_j}(z-\mu_j)^2\right\}$, -∞<z<∞ (mistura de distribuições log-normais para T) ou uma mistura de distribuições de valor extremo com densidades,

$f_j(z \mid u_j, b_j) = \frac{1}{b_j}\exp\left\{\frac{z-u_j}{b_j} - \exp\left(\frac{z-u_j}{b_j}\right)\right\}$,-∞<z<∞ (mistura de distribuições Weibull para T).

5.1 Mistura de Distribuições de Weibull para T

Como um caso especial, considere o modelo com uma mistura de distribuições Weibull para T, com densidade,

$$f(t) = \sum_{j=1}^{k} p_j f_j(t \setminus \alpha_j, \beta_j) \tag{17}$$

$$\text{onde } f_j(t \mid \alpha_j, \beta_j) = \left(\frac{\beta_j}{\alpha_j}\right)\left(\frac{t}{\alpha_j}\right)^{\beta_j - 1} \exp\left\{-\left(\frac{t}{\alpha_j}\right)^{\beta_j}\right\}, \tag{18}$$

$\alpha_j = \exp(\mu(\mathbf{x}) + \sigma u_j b_j)$ e $\beta_j = 1/\sigma b_j$, j = 1,...,k, onde **x** é um vetor de covariáveis.

Considerando dados sem censuras e sem covariáveis, a função de verossimilhança para $\mathbf{p} = (p_1,...,p_k)$, $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_k)$ e $\boldsymbol{\beta} = (\beta_1,...,\beta_k)$ é dada por,

$$L(\mathbf{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{j=1}^{k} p_j f_j(t \mid \alpha_j, \beta_j) \tag{19}$$

Para simplificar as distribuições condicionais necessárias para o amostrador de Gibbs, introduzimos variáveis latentes (ver por exemplo Tanner e Wong, 1987) $\mathbf{Z}_i = (Z_{i1},...Z_{ik})$ com uma distribuição multinomial $(1, h_{ij})$, onde

$$h_{ij} = \frac{p_j f_j(t \setminus \alpha_j, \beta_j)}{\sum_{j=1}^{k} p_j f_j(t \setminus \alpha_j, \beta_j)} \quad \text{i=1,....,n ; j=1,...,k} \tag{20}$$

Assim,

$$\pi(\mathbf{Z}_1,..,\mathbf{Z}_n) = \frac{\prod_{i=1}^{n}\prod_{j=1}^{k}\left[p_j f_j(t \mid \alpha_j, \beta_j)\right]^{Z_{ij}}}{\prod_{i=1}^{n}\sum_{j=1}^{k} p_j f_j(t \mid \alpha_j, \beta_j)} \tag{21}$$

Assim, dada uma distribuição a priori $\pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\beta})$ para $\mathbf{p},\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ a distribuição a posteriori para $\mathbf{p}$, $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$ é dada por,

$$\pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\beta}\mid \mathbf{t},\mathbf{Z}) \propto \pi(\mathbf{p},\boldsymbol{\alpha},\boldsymbol{\beta})\prod_{i=1}^{n}\prod_{j=1}^{k} p_j^{Z_{ij}} f_j^{Z_{ij}}(t\mid \alpha_j,\beta_j) \tag{22}$$

Como um caso especial, considere uma mistura de $k=2$ distribuições de Weibull $(p_1 = p, p_2 = 1-p_1)$ com as seguintes distribuições a priori:

$$\begin{aligned} &(i) p \sim B(a,b);\quad \text{a, b conhecidos}\\ &(ii)\alpha_j \sim \Gamma(c_j,d_j);\quad \text{c}_j, d_j \text{ conhecidos}\\ &(iii)\beta_j \sim \Gamma(e_j,f_j);\quad \text{e}_j, f_j \text{ conhecidos} \end{aligned} \tag{23}$$

onde $B(a,b)$ representa uma distribuição Beta com média $a/a+b$ e variância $ab/\left[(a+b)^2(a+b+1)\right]$; $\Gamma(c,d)$ denota uma distribuição gama com média c/d e variância c/d^2 .

Os logaritmos das distribuições condicionais para o algoritmo ARMS são dados por,

$$l_{\beta_j} = \sum_{i=1}^{n} q_{ij}\log\beta_j - \beta_j\sum_{i=1}^{n} q_{ij}\log\alpha_j + (\beta_j - 1)\sum_{i=1}^{n} q_{ij}\log t_i - \sum_{i=1}^{n} q_{ij}\left(\frac{t_i}{\alpha_j}\right)^{\beta_j} + \log\pi(\beta_j) \tag{24}$$

$$l_{\alpha_j} = \beta_j\sum_{i=1}^{n} q_{ij}\log\alpha_j - \sum_{i=1}^{n} q_{ij}\left(\frac{t_i}{\alpha_j}\right)^{\beta_j} + \log\pi(\alpha_j) \tag{25}$$

$$l_p = \sum_{i=1}^{n} q_{i1}\log p + \sum_{i=1}^{n} q_{i2}\log(1-p) + \log\pi(p) \tag{26}$$

Para cada iteração temos o seguinte algoritmo:

(a)Gerar uma amostra $\mathbf{z}^{(s)} = (\mathbf{z}_1,...,\mathbf{z}_n)$ onde $\mathbf{Z}_i = (\mathbf{Z}_{i1}, \mathbf{Z}_{i2})$ e Z_{i1} é gerado de uma distribuição Bernoulli com probabilidade de sucesso h_{i1} (ver (20)).

(b)Atualizar p, α_1,α_2,β_1 e β_2 a partir das distribuições condicionais dadas em (24),(25) e (26).

5.2 Uma ilustração numérica

Na tabela 4 temos os números de ciclos até falha de um grupo de 60 aparelhos elétricos em um teste de vida (Lawless, 1982, pg 56). Os tempos de falha foram ordenados por conveniência.

Tabela 4: Tempos de falhas de 60 aparelhos elétricos

Tempos de falhas											
14	*34*	*59*	*61*	*69*	*80*	*123*	*142*	*165*	*210*	*381*	*464*
479	*556*	*574*	*839*	*917*	*969*	*991*	*1064*	*1088*	*1091*	*1174*	*1270*
1275	*1355*	*1397*	*1477*	*1578*	*1649*	*1702*	*1893*	*1932*	*2001*	*2161*	*2292*
2326	*2337*	*2628*	*2785*	*2811*	*2886*	*2993*	*3122*	*3248*	*3715*	*3790*	*3857*
3912	*4100*	*4106*	*4116*	*4315*	*4510*	*4584*	*5267*	*5299*	*5583*	*6065*	*9701*

Construímos, através dessa tabela, o histograma dos tempos de falha dos aparelhos.

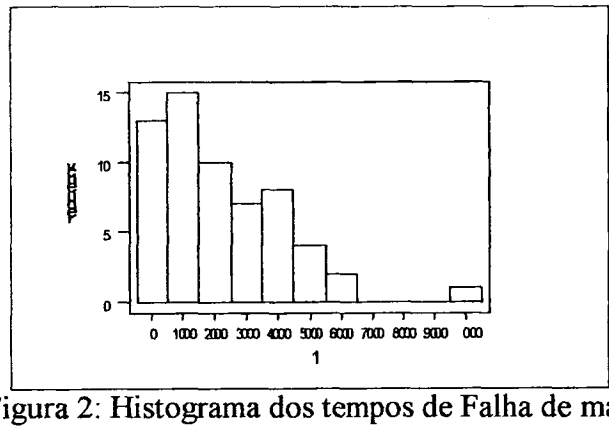

Figura 2: Histograma dos tempos de Falha de materiais elétricos.

A Figura 2 nos dá indícios que um modelo de mistura seria melhor ajustado aos dados da Tabela 4. Com isso, ajustamos um modelo de mistura de 2 distribuições de Weibull para esse conjunto de dados.

Na Tabela 5, temos as estimativas Bayesianas e clássicas para os parâmetros do modelo de mistura.

Tabela 5 Estimativas para os parâmetros.

	Parâmetro	EMV	Intervalo de Confiança (95%)	Média a Posteriori	Intervalo de Credibilidade (95%)
Misutra	$\alpha 01$	5,679	(2,327;9,030)	5,621	(5,046;6,0794)
	$\alpha 02$	19,893	(-1,474;41,254)	19,017	(18,066;19,939)
	$\beta 1$	0,785	(0,393;1,177)	0,781	(0,683;0,889)
	$\beta 2$	2,437	(-0,052;4,926)	2,332	(2,187;2,475)
	p	0,589	(0,227;0,932)	0,595	(0,348;0,867)

Podemos observar que as estimativas Bayesianas são mais precisas que as clássicas. Além disso, é importante notar que estimativas baseadas em métodos assintóticos podem nos levar a números absurdos, já que nos intervalos de Confiança os dois parâmetros α e β assumem valores negativos.

6 Algumas Conclusões

O uso de métodos MCMC permite um grande avanço na análise Bayesiana de dados de sobrevivência na presença de covariáveis e dados censurados pois usualmente um modelo apropriado para os dados envolve um grande número de parâmetros e as inferências clássicas usuais podem não ser apropriadas. Além disso, assumindo o modelo log-linear (1) podemos considerar diferentes distribuições paramétricas para os tempos de sobrevivência e usar técnicas Bayesianas de discriminação de modelos para decidir pelo melhor modelo par ajustar os dados.

A utilização de mistura de distribuições é uma alternativa bastante interessante para modelagem e inferências clássicas neste caso pode nos levar a resultados duvidosos. As técnicas Bayesianas foram melhores neste caso.

ANEXO A

***Tabela A1**: Dados de 90 pacientes com câncer na laringe.*

Tempo de estudo em meses	Indicador 1: morte 0: censura	Estágio da Doença	Idade no Diagnóstico	Ano do Diagnóstico
0.6	1	1	77	76
1.3	1	1	53	71
2.4	1	1	45	71
3.2	1	1	58	74
3.3	1	1	76	74
3.5	1	1	43	71
3.5	1	1	60	73
4.0	1	1	52	71
4.0	1	1	63	76
4.3	1	1	86	74
5.3	1	1	81	72
6.0	1	1	75	73
6.4	1	1	77	72
6.5	1	1	67	70
7.4	1	1	68	71
2.5	0	1	57	78
3.2	0	1	51	77
3.3	0	1	63	77
4.5	0	1	48	76
4.5	0	1	68	76
5.5	0	1	70	75
5.9	0	1	47	75
5.9	0	1	58	75
6.1	0	1	77	75
6.2	0	1	64	75
6.5	0	1	79	74
6.7	0	1	61	74
7.0	0	1	66	74
7.4	0	1	73	73
8.1	0	1	56	73
8.1	0	1	73	73
9.6	0	1	58	71
10.7	0	1	68	70
0.2	1	2	86	74
1.8	1	2	64	77
2.0	1	2	63	75
3.6	1	2	70	77
4.0	1	2	81	71
6.2	1	2	74	72
7.0	1	2	62	73
2.2	0	2	71	78
2.6	0	2	67	78
3.3	0	2	51	77
3.6	0	2	72	77
4.3	0	2	47	76
4.3	0	2	64	76
5.0	0	2	66	76

Tabela A1: *Dados de 90 pacientes com câncer na laringe.*

7.5	0	2	50	73
7.6	0	2	53	73
9.3	0	2	61	71
0.3	1	3	49	72
0.3	1	3	71	76
0.5	1	3	57	74
0.7	1	3	79	77
0.8	1	3	82	74
1.0	1	3	49	76
1.3	1	3	60	76
1.6	1	3	64	72
1.8	1	3	74	71
1.9	1	3	53	74
1.9	1	3	72	74
3.2	1	3	54	75
3.5	1	3	81	74
5.0	1	3	59	73
6.3	1	3	70	72
6.4	1	3	65	72
7.8	1	3	68	72
3.7	0	3	52	77
4.5	0	3	66	76
4.8	0	3	54	76
4.8	0	3	63	76
5.0	0	3	49	76
5.1	0	3	69	76
6.5	0	3	65	74
8.0	0	3	78	73
9.3	0	3	69	71
10.1	0	3	51	71
0.1	1	4	65	72
0.3	1	4	71	76
0.4	1	4	76	77
0.8	1	4	65	76
0.8	1	4	78	77
1.0	1	4	41	77
1.5	1	4	68	73
2.0	1	4	69	76
2.3	1	4	62	71
3.6	1	4	71	75
3.8	1	4	84	74
2.9	0	4	74	78
4.3	0	4	48	76

Referência:

Best, N.G., Cowless, M.K.,Vines, S.K., 1995, *CODA: Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output,Version 0.3.* MCR Biostatistics Unit, Cambridge.

Cox, D.R.,1972, Regression Models and Life Tables (with discussion). *Journal of the Royal Statistical Society,* Series B, 24,406-423.

Doornik, J. A., 1999, *Object-Oriented Matrix Programming Using Ox*, 3rd ed. London: Timberlake Consultants Press and Oxford.

Gelfand, A. E., and Smith, A. F. M., 1990, Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A., and Rubin, D. B., 1992, Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.

Gilks, W.R., Best, N.G. e Tan, K.C. ,1995, *Adaptive rejection Metropolis sampling whithin Gibbs sampling. Journal of the Royal Statistical Society,* Series C, 44:455-472.

Kalbfleisch, J. D., Prentice, R. L. , 1980, The Statistical Analysis of Failure Time Data. New York, John Wiley & Sons.

Kardaun, O., 1983, Statistical Analysis of Male Larynx-Cancer Patients – A case study. Statistical Nederlandica, 37, 103-126.

Klein, J.P., Moeschberger, M. L. 1997, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer- Verlag, New-York.

Lawless, J. F., 1982, *Statistical Models and Methods for Lifetime Data.* New York, John Wiley & Sons.

Tanner, M. A., and Wong, W. H., 1987, The calculation of posterior distributions via data argumentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.

NOTAS DO ICMC

SÉRIE ESTATÍSTICA

063/2001 ACHCAR J.A.; JUNQUEIRA, J.J.G. – Extra-biomial variability: a bayesian approach.

062/2000 WRUCK, E.; ACHCAR J.A.; MAZUCHELI, J. – Classification and discrimination for populations with mixture of multivariate normal distributions.

061/2000 ANDRADE, M.G.; MEIRA, S.A.; FRAGOSO, M.D.; CARNEIRO, A.A.F.M. – A bayesian approach to the stochastic flood control problem.

060/2000 ACHCAR, J.A.; JANEIRO, V. – A bayesian analysis for corralated binary data in the presence of covariates.

059/2000 MAZUCHELI, J.; ACHCAR, J.A.; KASS, R.E. – Regression models for lifetime data with mixture of normal distributions.

058/99 ACHCAR, J.A.; FORTULAN, V.C. – Meta analysis: a bayesian approach.

057/99 OLIVEIRA, S.C.; ACHCAR, J.A. - Confiabilidade de redes: um enfoque bayesiano.

056/98 RODRIGUES, J.; CHAVES, J.S. – A note on bayesian exponential regression model with censored data.

055/98 RODRIGUES, J.; SILVEIRA, V.D.R. – Bayesian computation for dichotomous variables with classification errors.

054/98 ACHCAR, J.A.; FORTULAN, V.C. – Relação entre o uso de hormônio e câncer em mulheres: um aplicação de meta-análise sob um enfoque bayesiano.