DISTRIBUIÇÃO EXPONENCIAL GENERALIZADA: USO DE MÉTODOS BAYESIANOS

Juliana BOLETA
Jorge Alberto ACHCAR

- RESUMO: Neste trabalho introduzimos a distribuição exponencial generalizada como uma alternativa para algumas distribuições usadas em análise de sobrevivência. Inferimos sobre os parâmetros do modelo considerando dados completos, dados censurados e a presença de covariáveis, sob o enfoque clássico e Bayesiano, e no que se refere à análise Bayesiana adotamos diferentes distribuições a priori para os parâmetros. Como ilustração foram feitas algumas aplicações utilizando métodos de simulação MCMC (Monte Carlo em Cadeias de Markov) e o software Winbugs.

- PALAVRAS-CHAVE: distribuição exponencial generalizada, dados censurados, análise Bayesiana, métodos de simulação MCMC.

1 Introdução

Uma distribuição exponencial generalizada (ver Gupta e Kundu, 1999) pode ser uma boa alternativa ao uso das populares distribuições Gama e Weibull usadas na análise de dados de sobrevivência (ver também, Raqab, 2002; Raqab e Ahsanullah, 2001; Zheng, 2002; Sarhan,2007; Gupta e Kundu, 2007 a, b).

A distribuição exponencial generalizada com dois parâmetros tem densidade dada por

$$f(t;\alpha,\lambda) = \alpha\lambda\left[1-\exp(-\lambda t)\right]^{\alpha-1}\exp(-\lambda t)\,, \qquad (1)$$

onde $t > 0$; $\alpha > 0$ e $\lambda > 0$ são, respectivamente, parâmetros de forma e escala.

A densidade (1) tem grande flexibilidade de formas dependendo do parâmetro α: se $\alpha < 1$, temos uma função decrescente e se $\alpha > 1$ temos uma função unimodal com moda dada por $\lambda^{-1}\log\alpha$. Observar que se $\alpha = 1$, temos uma distribuição exponencial com parâmetro λ.

As funções de sobrevivência e de risco associadas à densidade (1) são dadas, respectivamente, por

$$S(t;\alpha,\lambda) = P(T > t) = 1-\left[1-\exp(-\lambda t)\right]^{\alpha}\,, \qquad (2)$$

e

$$h(t;\alpha,\lambda) = \frac{f(t;\alpha,\lambda)}{S(t;\alpha,\lambda)} = \frac{\alpha\lambda\left[1-\exp(-\lambda t)\right]^{\alpha-1}\exp(-\lambda t)}{1-\left[1-\exp(-\lambda t)\right]^{\alpha}}\,. \qquad (3)$$

Observar que a função de risco $h(t;\ \alpha,\ \lambda)$ é crescente de 0 a λ se $\alpha > 1$; decrescente se $\alpha < 1$ e constante se $\alpha = 1$. Esse comportamento da função de risco é similar ao comportamento da função de risco da distribuição gama.

Também observar que o tempo de sobrevivência mediano obtido de $S(t;\ \alpha,\ \lambda)=1/2$ é dado por

$$t_{med} = -\frac{1}{\lambda}\log\left[1-\left(\frac{1}{2}\right)^{\frac{1}{\alpha}}\right]. \qquad (4)$$

2 Função de Verossimilhança para dados completos

Seja $T_1,\ T_2,\ldots,\ T_n$ denotando uma amostra aleatória de tamanho n da distribuição exponencial generalizada com densidade (1). A função de verossimilhança para α e λ é dada por

$$L(\alpha,\lambda)=\prod_{i=1}^{n} f(t_i;\alpha,\lambda)=\alpha^n\lambda^n\left[\prod_{i=1}^{n}(1-e^{-\lambda t_i})^{\alpha-1}\right]\exp\left(-\lambda\sum_{i=1}^{n}t_i\right) \tag{5}$$

O logaritmo da função de verossimilhança (5) é dado por

$$l(\alpha,\lambda)=\log L(\alpha,\lambda)=n\log\alpha+n\log\lambda-\lambda\sum_{i=1}^{n}t_i+(\alpha-1)\sum_{i=1}^{n}\log(1-e^{-\lambda t_i})\cdot \tag{6}$$

Os estimadores de máxima verossimilhança (EMV) para α e λ são obtidos igualando as primeiras derivadas de $l(\alpha,\lambda)$ em relação à α e λ, à zero, isto é,

$$\frac{\partial l}{\partial\alpha}=\frac{n}{\alpha}+\sum_{i=1}^{n}\log\left(1-e^{-\lambda t_i}\right)=0\,,$$

$$\frac{\partial l}{\partial\lambda}=\frac{n}{\lambda}+\sum_{i=1}^{n}t_i+(\alpha-1)\sum_{i=1}^{n}\frac{t_i e^{-\lambda t_i}}{(1-e^{-\lambda t_i})}=0\,. \tag{7}$$

De (9), encontramos o EMV para α dado por

$$\hat{\alpha}=-\frac{n}{\sum_{i=1}^{n}\log\left[1-\exp(-\hat{\lambda}t_i)\right]}. \tag{8}$$

O EMV para λ é obtido resolvendo-se a equação não linear,

$$\frac{n}{\hat{\lambda}}+n\bar{t}+(\hat{\alpha}-1)\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{t_i e^{-\hat{\lambda}t_i}}{(1-e^{-\hat{\lambda}t_i})}=0\,, \tag{9}$$

onde $n\bar{t}=\sum_{i=1}^{n}t_i$.

3 Uma Análise Bayesiana

Para uma análise Bayesiana da distribuição exponencial generalizada, consideramos diferentes distribuições a priori para os parâmetros α e λ. A priori não-informativa de Jeffreys (ver por exemplo, Box e Tiao, 1973) para α e λ é dada por

$$\prod{}_1(\alpha,\lambda)\propto\{\det I(\alpha,\lambda)\}^{1/2}\,, \tag{10}$$

onde, *I(α ,λ)* é a matriz de informação de Fisher esperada para α e λ dada por

$$I(\alpha,\lambda)=\begin{pmatrix} E\left(-\frac{\partial^2 l}{\partial\alpha^2}\right) & E\left(-\frac{\partial^2 l}{\partial\alpha\partial\lambda}\right)\\ E\left(-\frac{\partial^2 l}{\partial\alpha\partial\lambda}\right) & E\left(-\frac{\partial^2 l}{\partial\lambda^2}\right)\end{pmatrix}, \tag{11}$$

onde

$$E\left(-\frac{\partial^2 l}{\partial\alpha^2}\right)=\frac{n}{\alpha^2}\,, \tag{12}$$

$$E\left(-\frac{\partial^2 l}{\partial\alpha\partial\lambda}\right)=\frac{n}{\lambda}\left\{\left[\psi(\alpha+1)-\psi(1)\right]-\frac{\alpha}{\alpha-1}\left[\psi(\alpha)-\psi(1)\right]\right\},$$

$$E\left(-\frac{\partial^2 l}{\partial\lambda^2}\right)=\frac{n}{\lambda^2}\{1+\frac{\alpha(\alpha-1)}{\alpha-2}\left[\psi'(1)-\psi'(\alpha-1)+[\psi(\alpha-1)-\psi(1)]^2\right]- \quad .$$
$$-\alpha\left[\psi'(1)-\psi'(\alpha)+[\psi(\alpha)-\psi(1)]^2\right]\}$$

(ver Gupta e Kundu, 1999).

Uma possível simplificação é considerar uma priori não-informativa obtida a partir de $\prod(\alpha,\lambda)=\prod(\lambda|\alpha)\prod_0(\alpha)$. Usando a regra de Jeffreys, temos,

$$\prod_2(\alpha,\lambda)\propto\left\{E\left(-\frac{\partial^2 l}{\partial\lambda^2}\right)\right\}^{1/2}\prod_0(\alpha), \qquad (13)$$

onde $E\left(-\frac{\partial^2 l}{\partial\lambda^2}\right)$ é dada em (12) e $\prod_0(\alpha)$ é uma priori não-informativa dada por $\prod_0(\alpha)\propto 1/\alpha$, α >0 (uma priori imprópria). Assim,

$$\prod_2(\alpha,\lambda)\propto\frac{1}{\alpha\lambda}A^{1/2}(\alpha), \qquad (14)$$

onde, $A(\alpha)=1+\frac{\alpha(\alpha-1)}{\alpha-2}\left[\psi'(1)-\psi'(\alpha-1)+[\psi(\alpha-1)-\psi(1)]^2\right]-\alpha\left[\psi'(1)-\psi'(\alpha)+[\psi(\alpha)-\psi(1)]^2\right]$.

Uma terceira priori não-informativa é considerada assumindo independência a priori entre os parâmetros α e λ. Observar que se $E\left(-\frac{\partial^2 l}{\partial\alpha\partial\lambda}\right)\approx 0$, isto é, $[\psi(\alpha+1)-\psi(1)]/[\psi(\alpha)-\psi(1)]\approx\alpha/(\alpha-1)$, os parâmetros α e λ são ortogonais (ver Cox e Reid, 1987)..

Como $\psi(\alpha+1)=\psi(\alpha)+1/\alpha$ (ver, por exemplo, Abramowitz e Stegun, 1970, pág. 258), observamos que

$$\frac{\psi(\alpha)+1/\alpha-\psi(1)}{\psi(\alpha)-\psi(1)}\to 1, \qquad (15)$$

se $\alpha\uparrow\infty$; da mesma forma $\alpha/(\alpha-1)\to 1$ se $\alpha\uparrow\infty$.

Assim, se α for grande podemos assumir ortogonalidade aproximada entre os parâmetros α e λ. Portanto, podemos assumir uma priori não-informativa dada por

$$\prod_3(\alpha,\lambda)\propto\frac{1}{\alpha\lambda}. \qquad (16)$$

Observar que a priori (16) é uma priori imprópria.

Também assumindo independência a priori entre os parâmetros α e λ, podemos considerar a priori informativa dada por

$$\prod_4(\alpha,\lambda)=\prod_\alpha(\alpha)\prod_\lambda(\lambda), \qquad (17)$$

onde $\prod_\alpha(\alpha)$ e $\prod_\lambda(\lambda)$ são distribuições gama,

$$\prod_\alpha(\alpha)\sim Gama(a_\alpha,b_\alpha),$$
$$\prod_\lambda(\lambda)\sim Gama(a_\lambda,b_\lambda), \qquad (18)$$

onde $a_\alpha, b_\alpha, a_\lambda$ e b_λ são hiperparâmetros conhecidos e Gama(*a,b*) denota uma distribuição Gama com média *a/b* e variância a/b^2. Observar que a escolha dos hiperparâmetros das distribuições a priori (18) pode ser feita a partir da informação de especialistas sobre a média e a variância de α e λ; com isso encontramos $a_\alpha, b_\alpha, a_\lambda$ e b_λ. Também poderíamos usar métodos Bayesianos empíricos para a escolha do hiperparâmetros.

Usando a fórmula de Bayes, combinamos a função de verossimilhança (5) com uma das distribuições a priori $\prod_j(\alpha,\lambda)$, *j*=1,...,4 para determinarmos a distribuição a posteriori para α e λ. Assumindo a priori informativa (17), a distribuição a posteriori conjunta para α e λ é dada por

$$\prod(\alpha,\lambda/\mathbf{t}) \propto \alpha^{n+a_\alpha-1}\ \lambda^{n+a_\lambda-1}\times \\ \times \exp\left\{-\alpha\left[b_\alpha - \sum_{i=1}^{n}\log(1-e^{-\lambda t_i})\right]\right\}, \\ \times \exp\{-\lambda[b_\lambda + n\bar{t}]\} \qquad (19)$$

onde $\mathbf{t} = (t_1,...,t_n)$ é o vetor dos dados.

4 Presença de Dados Censurados

Supondo uma amostra aleatória com tempos de sobrevivência completos e censurados, definimos uma variável indicadora, dada por

$$\delta_i = \begin{cases} 1, \text{ se } t_i = T_i \\ 0, \text{ se } t_i = L_i \end{cases}, \qquad (20)$$

onde $t_i = min(T_i, L_i)$ é o tempo observado, T_i é o tempo de sobrevivência e L_i é o tempo de censura assumido fixo para o i-*ésimo* indivíduo (dados com censura de Tipo I).

A função de verossimilhança para α e λ é dada por

$$L(\alpha,\lambda) = \prod_{i=1}^{n} f^{\delta_i}(t_i;\alpha,\lambda)S^{1-\delta_i}(t_i;\alpha,\lambda), \qquad (21)$$

onde $f(t_i;\alpha,\lambda)$ é a densidade exponencial generalizada dada em (1) e $S(t_i;\alpha,\lambda)$ é a função de sobrevivência dada por (2). Assim, temos,

$$L(\alpha,\lambda) = \alpha^r\lambda^r \exp\{(\alpha-1)\sum_{i=1}^{n}\delta_i\log(1-e^{-\lambda t_i}) - \\ -\lambda\sum_{i=1}^{n}\delta_i t_i + \sum_{i=1}^{n}(1-\delta_i)\log\left[1-(1-e^{-\lambda t_i})^\alpha\right]\}, \qquad (22)$$

onde $r = \sum_{i=1}^{n}\delta_i$ (número de observações completas). O logaritmo da função de verossimilhança (22) é,

$$l(\alpha,\lambda) = r\log\alpha + r\log\lambda + (\alpha-1)\sum_{i=1}^{n}\delta_i\log(1-e^{-\lambda t_i}) - \\ -\lambda\sum_{i=1}^{n}\delta_i t_i + \sum_{i=1}^{n}(1-\delta_i)\log\left[1-(1-e^{-\lambda t_i})^\alpha\right]. \qquad (23)$$

Estimadores de máxima verossimilhança para α e λ são obtidos a partir das equações dadas por $dl/d\alpha = 0$ e $dl/d\lambda = 0$. Para uma análise Bayesiana do modelo, consideramos as mesmas distribuições a priori para α e λ usadas para a distribuição exponencial generalizada sem a presença de covariáveis.

5 Presença de Covariáveis

Na presença de um vetor de covariáveis $\mathbf{x} = (1, x_1, x_2,..., x_k)'$, assumimos a densidade exponencial generalizada (1) com o parâmetro de escala λ dependendo das covariáveis na forma

$$\lambda(\mathbf{x}_i) = e^{\beta\mathbf{x}_i'}, \qquad (24)$$

onde $\mathbf{x}_i = (1, x_{1i}, x_{2i},..., x_{ki})$, $\boldsymbol{\beta} = (\beta_0, \beta_1,..., \beta_k)$ e $\boldsymbol{\beta}x_i' = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$, para *i*=1,2,...*n*.

Assumindo dados de sobrevivência na presença de covariáveis e dados censurados, a função de verossimilhança para α e $\boldsymbol{\beta}$ é dada por

$$L(\alpha,\boldsymbol{\beta}) = \prod_{i=1}^{n} f^{\delta_i}(t_i/\alpha,\boldsymbol{\beta},\mathbf{x}_i) S^{1-\delta_i}(t_i/\alpha,\boldsymbol{\beta},\mathbf{x}_i)\,, \tag{25}$$

onde δ_i é uma função indicadora de censuras dada em (20), $f(t_i/\alpha,\boldsymbol{\beta},\mathbf{x}_i)$ e $S(t_i/\alpha,\boldsymbol{\beta},\mathbf{x}_i)$ são dados, respectivamente em (1) e (2) com $\lambda(\mathbf{x}_i)$ definido por (24) em lugar de λ. Dessa forma, a função de verossimilhança (25) é dada por

$$\begin{aligned} L(\alpha,\boldsymbol{\beta}) = \alpha^r \exp\{&\sum_{i=1}^{n}\delta_i\boldsymbol{\beta}'\mathbf{x}_i - \sum_{i=1}^{n}\delta_i t_i e^{\boldsymbol{\beta}'\mathbf{x}_i} + \\ &+(\alpha-1)\sum_{i=1}^{n}\delta_i \log(1-e^{-t_i e^{\boldsymbol{\beta}'\mathbf{x}_i}}) + \\ &+\sum_{i=1}^{n}(1-\delta_i)\log\left[1-(1-e^{-t_i e^{\boldsymbol{\beta}'\mathbf{x}_i}})^{\alpha}\right]\}\quad, \end{aligned} \tag{26}$$

onde *r* é o número de observações completas definido em (22).

Para uma análise Bayesiana do modelo, assumimos uma distribuição a priori gama para α (ver (18)) com hiperparâmetros conhecidos a_α e b_α e distribuições a priori normais para β_l, $l = 0,1,...,k$, isto é,

$$\beta_l \sim N(0;b_l^2)\,, \tag{27}$$

onde b_l são hiperparâmetros conhecidos, $l = 0,1,...,k$. Também assumimos independência a priori entre os parâmetros.

A distribuição a posteriori conjunta para α e $\boldsymbol{\beta}$ é dada por

$$\begin{aligned} \prod(\alpha,\boldsymbol{\beta}\,|\,\mathbf{t},\mathbf{x}) \;\propto\; &\alpha^{r+a_\alpha-1}e^{-b_\alpha\alpha}\times\left(\prod_{i=1}^{n}\exp\left[-\frac{\beta_l^2}{2b_l^2}\right]\right)\times \\ &\times\exp\{\sum_{i=1}^{n}\delta_i\boldsymbol{\beta}'\mathbf{x}_i - \sum_{i=1}^{n}\delta_i t_i e^{\boldsymbol{\beta}'\mathbf{x}_i} + (\alpha-1)\sum_{i=1}^{n}\delta_i\log(1-e^{-t_i e^{\boldsymbol{\beta}'\mathbf{x}_i}}) + \\ &+\sum_{i=1}^{n}(1-\delta_i)\log\left[1-(1-e^{-t_i e^{\boldsymbol{\beta}'\mathbf{x}_i}})^{\alpha}\right]\}\quad. \end{aligned} \tag{28}$$

Amostras da distribuição a posteriori conjunta para α e $\boldsymbol{\beta}$ são geradas usando métodos MCMC.

6 Aplicação: Um exemplo com dados censurados e presença de covariáveis

Neste exemplo consideramos um estudo retrospectivo com 190 pacientes com leucemia mielóide aguda (um tipo de câncer). Esse conjunto de dados de sobrevivência (em dias) foi analisado no CEMEQ (Centro de Métodos Quantitativos) da Faculdade de Medicina da Universidade de São Paulo, campus de Ribeirão Preto. Esses dados apresentam algumas observações censuradas (censura à direita) e presença de covariáveis (GBx1000 e Uréia). Para analisar esses dados, assumimos um modelo de regressão exponencial generalizado na presença de dados censurados e covariáveis.

Assim, assumimos a densidade exponencial generalizada (1) para os tempos de sobrevivência com parâmetro de escala (ver (24)) dado por

$$\lambda(\mathbf{x}_i) = e^{\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}}\,, \tag{29}$$

para i=1,...,189 (número de indivíduos), onde x_{1i} denota a covariável GBx1000 e x_{2i} denota a covariável uréia.

Assumindo uma distribuição a priori gama para α, com hiperparâmetros dados por $a_\alpha = 2$, $b_\alpha = 1$ e distribuições a priori normais (29) para β_0, β_1 e β_2 com hiperparâmetros b_0=10, b_1=0,5 e b_2=0,5, e usando o software Winbugs, temos na Tabela 4, os sumários a posteriori de interesse baseados em 1000 amostras simuladas de Gibbs para a distribuição a posteriori conjunta para α, β_0, β_1 e β_2 após descartar as 5000 primeiras amostras ("burn-in-samples"). A convergência do algoritmo foi observada usando métodos gráficos.

Tabela 4 - Sumários a posteriori (dados com censuras e covariáveis).

Distribuição	Parâmetro	Média	DP	Intervalo de credibilidade 95%
Exp.	α	3,086	0,7686	(1,870; 5,044)
Generalizada	β_0	-6,603	0,3233	(-7,256; -5,925)
	β_1	-0,001	0,0065	(-0,015; 0,008)
	β_2	-0,003	0,0087	(-0,022; 0,011)

Dos resultados da Tabela 4, observamos que as covariáveis GBx1000 e uréia não apresentam um efeito significativo nos tempos de sobrevivência, pois zero esta incluído no intervalo de credibilidade 95% para β_1 e β_2, respectivamente, assumindo a distribuição exponencial generalizada.

Algumas Notas Conclusivas

O uso da distribuição exponencial generalizada pode ser uma boa alternativa para a análise de dados de sobrevivência, dado a grande flexibilidade de ajuste e forma analítica simples para a função de sobrevivência (ver (2)). Essa distribuição apresenta algumas similaridades com a distribuição gama em termos de comportamento da função de risco, mas grandes vantagens na obtenção de inferências de interesse quando os dados de sobrevivência apresentam censuras, um fato comum com dados médicos e industriais. Observe que a função de sobrevivência para a distribuição gama é dada por uma função gama incompleta, o que, em geral, dificulta a obtenção de inferências de interesse.

Referências

ABRAMOWITZ, M.; STEGUN, I. A. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, New York: Dover Publications, 1970.

BOX, G.E.P.; TIAO, G.C. Bayesian Inference in Statistical Analysis. Reading: Addison-Wesley, 1973.

COX, D.R.; REID, N. Parameter Orthogonality and Approximate Conditional Inference (with discussion). Journal of the Royal Statistical Society, Ser. B, 49, 1-39, 1987.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions. Australian and New Zealand Journal of Statistics, v.41, p.173-188, 1999.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions: different methods of estimation. Journal of Statistical Computation and Simulation, 69, 315-338, 2001.

GUPTA, R. D.; KUNDU, D. Generalized exponential distribution:existing results and some recent developments. Journal of Statistical Planning and Inference, doi: 10.1016/j.jspi.2007.03.030, 2007.

RAQAB,M.Z. Inferences for generalized exponential distribution based on record statistics. Journal of Statistical Planning and Inference, 104, 339-350, 2002.

RAQAB,M.Z.;AHSANULLAH,M. Estimation of the location and scale parameters of generalized exponential distribution based on order statistics. Journal of Statistical Computation and Simulation,69,109-124, 2001.

SARHAN, A. M. Analysis of incomplete, censored data in competing risks models with generalized exponential distributions. IEEE Transactions on Reliability, 56, 132-138, 2007.

ZHENG, G. (2002) Fisher information matrix in type-II censored data from exponentiated exponential family. Biometrical Journal, 44, 353-357, 2002.