DISTRIBUIÇÃO EXPONENCIAL GENERALIZADA BIVARIADA DERIVADA DE FUNÇÕES CÓPULAS: UMA APLICAÇÃO A DADOS DE CÂNCER GÁSTRICO

Juliana BOLETA[1]
Jorge Alberto ACHCAR[1]

▪ RESUMO: Neste artigo utilizamos a distribuição exponencial generalizada em uma análise de dados de sobrevivência multivariados utilizando funções cópula de Farlie-Gumbel-Morgenstern. Inferências sobre os parâmetros do modelo com dados censurados e na presença de covariáveis, são obtidas sob o enfoque Bayesiano. Adotamos diferentes distribuições a priori para os parâmetros. Como ilustração analisamos um conjunto de dados de sobrevivência reais bivariados relacionados ao câncer gástrico, utilizando métodos de simulação MCMC (Monte Carlo em Cadeias de Markov) e o software Winbugs, para obter os sumários a posteriori de interesse.

▪ PALAVRAS-CHAVE: Distribuição exponencial generalizada; funções cópula; modelos Bayesianos; câncer gástrico.

1 Introdução

Uma distribuição exponencial generalizada (Gupta e Kundu, 1999) pode ser uma boa alternativa ao uso das populares distribuições Gama e Weibull usadas na análise de dados de sobrevivência (Raqab e Ahsanullah, 2001; Raqab, 2002; Zheng, 2002; Gupta e Kundu, 2007; Sarhan, 2007). Essa distribuição tem algumas vantagens que podem ser úteis na análise de dados de sobrevivência: flexibilidade de ajuste e forma simples para a função de sobrevivência.

[1]Universidade de São Paulo – USP, Faculdade de Medicina de Ribeirão Preto – FMRP, Departamento de Medicina Social, CEP: 14048-900, Ribeirão Preto, SP, Brasil. E-mail: *jujuboleta@yahoo.com.br / achcar@fmrp.usp.br*

1.1 Função densidade da distribuição exponencial generalizada

A distribuição exponencial generalizada com dois parâmetros para uma variável aleatória T, representando o tempo de sobrevivência de um indivíduo, tem densidade dada por

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

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

1.2 Função de sobrevivência e de risco

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

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

e

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

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 (mais detalhes dessa distribuição são dados por Achcar e Boleta, 2009).

2 Análise multivariada derivadas de funções de cópula

Algumas áreas, que utilizam análises de tempos de sobrevida, costumam se deparar com mais de um tempo de sobrevida associado ao mesmo indivíduo. Nestas situações pode-se considerar distribuições de sobrevida bivariadas, no caso de dois tempos de sobrevivência ou mesmo distribuições multivariadas no caso de mais de dois tempos de sobrevivência. Neste capítulo será introduzido o conceito de funções cópula, e também será descrito a derivação da distribuição exponencial generalizada bivariada de uma função cópula.

2.1 Funções cópula

As funções cópula podem ser usadas para relacionar as distribuições marginais independentes com as distribuições conjuntas. Para isso relacionamos as funções distribuições marginais univariadas $F_1(t_1), F_2(t_2)$, ..., $F_m(t_m)$, a partir da função,

$$C(F_1(t_1), F_2(t_2), ..., F_m(t_m)) = F(t_1, t_2, ..., t_m); \quad (4)$$

definida como uma função cópula C, resultando em uma função distribuição multivariada com funções distribuições marginais univariadas dadas por $F_1(t_1), F_2(t_2), ..., F_m(t_m)$.

É importante salientar que a função distribuição multivariada F pode ser escrita na forma de uma função cópula (Sklar, 1959); isto é, se $F(t_1, t_2, ..., t_m)$ é a junção entre a função distribuição multivariada com as funções distribuições marginais univariadas $F_1(t_1), F_2(t_2), ..., F_m(t_m)$, então existe uma função cópula $C(U_1, U_2, ..., U_m)$ dada por

$$F(t_1, t_2, ..., t_m) = C(F_1(t_1), F_2(t_2), ..., F_m(t_m)). \tag{5}$$

Se todas as funções F_i são contínuas, então C é único.

Para o caso especial de uma distribuição bivariada, temos $m = 2$.

A abordagem para a formulação de uma distribuição multivariada é baseada na ideia que uma simples transformação pode ser feita em cada variável marginal, onde cada variável marginal possui uma distribuição uniforme. Feito isso, a estrutura de dependência pode ser expressa como uma distribuição multivariada das uniformes obtidas, e a função cópula é precisamente a distribuição multivariada das variáveis aleatórias marginais uniformes.

Desta forma, existem várias famílias de funções cópulas que se diferem na dependência que eles representam.

Nos casos bivariados, considere T_1 e T_2 variáveis aleatórias independentes com funções distribuição contínuas dadas por F_1 e F_2.

Essa transformação pode ser aplicada separadamente para as duas variáveis aleatórias definindo $U = F_1(T_1)$ e $V = F_2(T_2)$, onde U e V possuem distribuição uniforme $(0, 1)$, mas que podem ser dependentes de T_1 e T_2 (T_1 e T_2 independentes implica que U e V são independentes). Especificar uma relação de dependência entre T_1 e T_2 equivale a assumir dependência entre U e V.

Com U e V variáveis aleatórias uniformes, o problema se reduz em especificar uma distribuição bivariada entre duas uniformes, isto é, a função cópula. Para todas as cópulas têm-se os limites Frechet-Hoeffding (Nelsen, 1999), dados por

$$max(0, u + v - 1) \leq C(u, v) \leq min(u, v). \tag{6}$$

Diferentes famílias de funções cópula são introduzidas na literatura. Na área de finanças a função cópula Gaussiana é muito utilizada (Nelsen, 1999). A cópula Gaussiana é construída baseada na distribuição normal bivariada, dada por

$$C_p(u, v) = \Phi_\rho \left[\Phi^{-1}(u), \Phi^{-1}(v)\right]; \tag{7}$$

em que Φ_ρ é a função distribuição bivariada de uma distribuição normal bivariada padrão com coeficiente de correlação ρ.Assim, considerando X e Y variáveis aleatórias, a função cópula é dada por

$$\begin{aligned} \Phi_\rho(x, y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho^2}} \times \\ &\times \quad exp\left\{-\frac{1}{2(1-\rho^2)}\left(z^2 - 2\rho zw + w^2\right)\right\} dzdw. \end{aligned} \tag{8}$$

2.2 Distribuição exponencial generalizada bivariada derivada da cópula de Farlie-Gumbel-Morgenstern

Diferentes funções cópulas têm sido introduzidas na literatura; a partir dessas cópulas e assumindo distribuições exponenciais generalizadas marginais para cada tempo de sobrevivência podemos obter uma distribuição de sobrevivência exponencial generalizada bivariada. Um caso especial é dado pela Cópula de Farlie-Gumbel-Morgensten (Morgenstern, 1956) dada por

$$\begin{aligned} C(u,v) &= \left[1-e^{ln(1-u)}\right]\left[1-e^{ln(1-v)}\right]\times \\ &\times \left[1+\theta exp\left\{(ln(1-u)+ln(1-v))\right\}\right]; \end{aligned} \tag{9}$$

em que $u = F_1(t_1)$ (distribuição marginal para a variável aleatória T_1) e $v = F_2(t_2)$ (distribuição marginal para a variável aleatória T_2). Isso é

$$C(u,v) = uv\left[1+\theta(1-u)(1-v)\right]; \tag{10}$$

onde $-1 \leq \theta \leq 1$.

Observar que θ é o parâmetro associado à dependência entre as variáveis aleatórias T_1 e T_2. Algumas relações úteis entre funções de cópula e coeficientes de correlação de Spearman e Kendall são introduzidos na literatura. O coeficiente de correlação de Spearman $\rho_S(T_1,T_2)$ e o coeficiente de correlação de Kendall $\rho_T(T_1,T_2)$ são dadas, respectivamente, por

$$\begin{aligned} \rho_S(T_1,T_2) &= 12\int_0^1\int_0^1 [C(u,v)-uv]\,dudv \\ \rho_T(T_1,T_2) &= 4\int_0^1\int_0^1 C(u,v)dC(u,v)-1. \end{aligned} \tag{11}$$

De (10), obtêm-se $\rho_S(T_1,T_2) = \theta/3$ e $\rho_T(T_1,T_2) = 2\theta/9$ (Nelsen, 1999).

Para a especificação das distribuições a priori para o parâmetro de dependência definido no intervalo (-1,1), pode-se explorar as relações dadas em (11) usando métodos bayesianos empíricos ou usar a informação de especialistas.

Seja distribuições exponenciais generalizadas marginais dadas por,

$$\begin{aligned} u &= F_1(t_1) = P\left\{T_1 \leq t_1\right\} = (1-exp(-\lambda_1 t_1))^{\alpha_1} \\ v &= F_2(t_2) = P\left\{T_2 \leq t_2\right\} = (1-exp(-\lambda_2 t_2))^{\alpha_2}. \end{aligned} \tag{12}$$

A função distribuição conjunta para T_1 e T_2 dada por,

$$\begin{aligned} F(t_1,t_2|\lambda_1,\lambda_2,\alpha_1,\alpha_2,\theta) &= C(F_1(t_1),F_2(t_2)) \\ &= F_1(t_1)F_2(t_2)\times\left[1+\theta\left(1-F_1(t_1)\right)\left(1-F_2(t_2)\right)\right]. \end{aligned} \tag{13}$$

Então,

$$\begin{aligned} F(t_1,t_2|\lambda_1,\lambda_2,\alpha_1,\alpha_2,\theta) &= (1-exp(-\lambda_1 t_1))^{\alpha_1}(1-exp(-\lambda_2 t_2))^{\alpha_2}\times \\ &\times [1+\theta(1-(1-exp(-\lambda_1 t_1))^{\alpha_1})\times \\ &\times (1-(1-exp(-\lambda_2 t_2))^{\alpha_2})]; \end{aligned} \tag{14}$$

onde $t_1 > 0$ e $t_2 > 0$.

A função densidade de probabilidade conjunta para T_1 e T_2 é dada por

$$f(t_1, t_2|\lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) = \frac{\partial^2 F(t_1, t_2)}{\partial t_1 \partial t_2}. \tag{15}$$

De (13) temos

$$\begin{aligned} f(t_1, t_2|\lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) &= f_1(t_1) f_2(t_2) + \theta f_1(t_1) f_2(t_2) \times \\ &\times \left[(1 - 2F_1(t_1))(1 - 2F_2(t_2))\right]; \end{aligned} \tag{16}$$

em que $f_1(t_1)$ e $f_2(t_2)$ são as funções densidades marginais para T_1 e T_2 dadas, respectivamente, por

$$\begin{aligned} f_1(t_1) &= \alpha_1 \lambda_1 \left[1 - exp(-\lambda_1 t_1)\right]^{\alpha_1 - 1} exp(-\lambda_1 t_1) \\ f_2(t_2) &= \alpha_2 \lambda_2 \left[1 - exp(-\lambda_2 t_2)\right]^{\alpha_2 - 1} exp(-\lambda_2 t_2); \end{aligned} \tag{17}$$

onde $F_1(t_1)$ e $F_2(t_2)$ são dados por (12).

Observe que se $\theta = 0$, tem-se variáveis aleatórias não correlacionadas.

Observar que a função de sobrevivência bivariada para os tempos de sobrevidas T_1 e T_2 obtém-se

$$S(t_1, t_2) = P\left\{T_1 > t_1, T_2 > t_2\right\} = 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2); \tag{18}$$

em que $F_1(t_1)$ e $F_2(t_2)$ são dadas por (12) e $F(t_1, t_2)$ é dado em (14). Então, tem-se

$$\begin{aligned} S(t_1, t_2) &= 1 - (1 - exp(-\lambda_1 t_1))^{\alpha_1} - (1 - exp(-\lambda_2 t_2))^{\alpha_2} + \\ &+ (1 - exp(-\lambda_1 t_1))^{\alpha_1} (1 - exp(-\lambda_2 t_2))^{\alpha_2} \times \\ &\times \left[1 + \theta(1 - (1 - exp(-\lambda_1 t_1))^{\alpha_1})(1 - (1 - exp(-\lambda_2 t_2))^{\alpha_2})\right]. \end{aligned} \tag{19}$$

Observar que a função cópula de Farlie-Gumbel-Morgestern só é apropriada para situações onde existe dependência fraca entre os tempos de sobrevida. Em outras situações são mais adequadas outras funções cópulas (por exemplo, funções cópula de Gumbel, Clayton, entre outras; Nelsen, 1999).

2.3 Análise bayesiana na presença de dados censurados

Suponha que T_1 e T_2 são duas variáveis aleatórias relativas a tempos até a ocorrência de eventos de interesse, sujeitos a censura independentemente dos tempos observados. Sejam t_{1i} e t_{2i}observações amostrais de T_1 e T_2, respectivamente, para o i-ésimo indivíduo, $i = 1, ..., n$. Ao classificar os n pares de observações (t_{1i},t_{2i}) em quatro classes, tem-se:

- C_1: t_{1i} e t_{2i} são tempos de sobrevida observados;
- C_2: t_{1i} é o tempo de sobrevida e t_{2i} é o tempo de censura (sabe-se apenas que $T_{2i} \geq t_{2i}$);
- C_3: t_{1i} é o tempo de censura e t_{2i} é o tempo de sobrevida;

- C_4: t_{1i} e t_{2i} são os tempos de censura;

A função de verossimilhança para o modelo contínuo (Lawless, 1982, p. 479) é dada por

$$\begin{aligned} L &= \prod_{i\in C_1} f(t_{1i},t_{2i}) \prod_{i\in C_2} \left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}}\right) \times \\ &\times \prod_{i\in C_3} \left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}}\right) \prod_{i\in C_4} S(t_{1i},t_{2i}); \end{aligned} \tag{20}$$

onde $f(t_{1i},t_{2i})$ é a função densidade de probabilidade conjunta para T_{1i} e T_{2i}; $S(t_{1i},t_{2i})$ é função de sobrevivência conjunta; $\left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}}\right)$ e $\left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}}\right)$ são as derivadas parciais de $S(t_{1i},t_{2i})$. As variáveis indicadoras δ_{1i} e δ_{2i} são definidas por

$$\delta_{ji} = \begin{cases} 0, & \text{se } t_{ji} \text{ é uma observação censurada} \\ 1, & \text{se } t_{ji} \text{ é o tempo de sobrevida observado;} \end{cases}$$

para $j = 1, 2$ e $i = 1, ..., n$, onde n é o número de observações.

Desta forma, reescrevendo a função de verossimilhança tem-se

$$\begin{aligned} L &= \prod_{i=1}^{n} [f(t_{1i},t_{2i})]^{\varsigma_{1i}\varsigma_{2i}} \prod_{i=1}^{n} \left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}}\right)^{\varsigma_{1i}(1-\varsigma_{2i})} \times \\ &\times \prod_{i=1}^{n} \left(-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}}\right)^{\varsigma_{2i}(1-\varsigma_{1i})} \prod_{i=1}^{n} [S(t_{1i},t_{2i})]^{(1-\varsigma_{1i})(1-\varsigma_{2i})}. \end{aligned} \tag{21}$$

Observe que se não houver dados censurados, a função de verossimilhança se reduz a

$$L = \prod_{i=1}^{n} f(t_{1i},t_{2i}). \tag{22}$$

Em (21), substituindo $S(t_{1i},t_{2i})$ por (19), $f(t_{1i},t_{2i})$ por (16), onde $f_1(t_1)$ e $f_2(t_2)$ são dados em (17) e $F_1(t_1)$ e $F_2(t_2)$ são dados por (12), assim

$$\begin{aligned} f(t_{1i},t_{2i}|\lambda_1,\lambda_2,\alpha_1,\alpha_2,\theta) &= \alpha_1\alpha_2\lambda_1\lambda_2(1-exp(-\lambda_1 t_{1i}))^{\alpha_1-1} \times \\ &\times (1-exp(-\lambda_2 t_{2i}))^{\alpha_2-1} exp(-\lambda_1 t_{1i} - \lambda_2 t_{2i}) \times \\ &\times [1+\theta(1-2(1-exp(-\lambda_1 t_{1i}))^{\alpha_1}) \times \\ &\times (1-2(1-exp(-\lambda_2 t_{2i}))^{\alpha_2})]. \end{aligned} \tag{23}$$

As primeiras derivadas de $S(t_{1i},t_{2i})$ em relação a t_{1i} e t_{2i} são dados por

$$\begin{aligned} -\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}} &= f_1(t_{1i})\{1-F_2(t_{2i})[1+\theta(1-F_2(t_{2i}))(1-2F_1(t_{1i}))]\} \\ -\frac{\partial S(t_{1i},t_{2i})}{\partial t_{2i}} &= f_2(t_{2i})\{1-F_1(t_{1i})[1+\theta(1-F_1(t_{1i}))(1-2F_2(t_{2i}))]\} \end{aligned} \tag{24}$$

Isto é,

$$
\begin{aligned}
-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}} &= \alpha_1\lambda_1(1-exp(-\lambda_1 t_{1i}))^{\alpha_1-1}exp(-\lambda_1 t_{1i})\times \\
&\times \{1-(1-exp(-\lambda_2 t_{2i}))^{\alpha_2}\times \\
&\times [1+\theta(1-(1-exp(-\lambda_2 t_{2i}))^{\alpha_2})\times \\
&\times (1-2(1-exp(-\lambda_1 t_{1i}))^{\alpha_1})]\}
\end{aligned} \tag{25}
$$

$$
\begin{aligned}
-\frac{\partial S(t_{1i},t_{2i})}{\partial t_{1i}} &= \alpha_2\lambda_2(1-exp(-\lambda_2 t_{2i}))^{\alpha_2-1}exp(-\lambda_2 t_{2i})\times \\
&\times \{1-(1-exp(-\lambda_1 t_{1i}))^{\alpha_1}\times \\
&\times [1+\theta(1-(1-exp(-\lambda_1 t_{1i}))^{\alpha_1})\times \\
&\times (1-2(1-exp(-\lambda_2 t_{2i}))^{\alpha_2})]\}
\end{aligned}
$$

Para uma análise bayesiana, são assumidas as seguintes distribuições a priori para λ_1, λ_2, α_1, α_2 e θ,

$$
\begin{aligned}
\lambda_j &\sim U(a_j,b_j) \\
\alpha_j &\sim U(c_j,d_j) \\
\theta &\sim U(-1,1);
\end{aligned} \tag{26}
$$

para $j=1,2$; $U(a,b)$ denota uma distribuição uniforme no intervalo (a,b); a_j, b_j, c_j e d_j são hiperparâmetros. É assumida independência a priori entre os parâmetros.

Outras distribuições a priori podem ser consideradas, como uma distribuição gama para α_j e λ_j, $j=1,2$.

A distribuição a posteriori conjunta de interesse para $\boldsymbol{\upsilon}=(\lambda_1,\lambda_2,\alpha_1,\alpha_2,\theta)'$ é dada por

$$
\pi(\boldsymbol{\upsilon}|\mathbf{t}) \propto \pi(\boldsymbol{\upsilon})L(\boldsymbol{\upsilon}|\mathbf{t}); \tag{27}
$$

em que $\pi(\boldsymbol{\upsilon})$ é a distribuição priori conjunta para $\boldsymbol{\upsilon}$; $L(\boldsymbol{\upsilon}|\mathbf{t})$ é a função de verossimilhança (21), e $\mathbf{t}=(\mathbf{t}_1,...,\mathbf{t}_n)$ e $\mathbf{t}_i=(t_{1i},t_{2i})$, $i=1,...,n$ é o vetor dos dados observados.

Para se obter os sumários a posteriori de interesse, simula-se amostras da distribuição a posteriori conjunta (27) usando métodos MCMC(Monte Carlo em Cadeias de Markov).

2.4 Análise bayesiana na presença de covariáveis

Agora vamos assumir um vetor de covariáveis $\mathbf{X}=(X_1,...,X_p)'$ associados aos pares de tempos de sobrevida $(T_1,\ T_2)$.

Na presença do vetor de covariáveis $\mathbf{X}$, considere o seguinte modelo

$$\begin{aligned} \lambda_{1i} &= \gamma_1 exp(\boldsymbol{\beta}_1' \mathbf{x}_i) \\ \lambda_{2i} &= \gamma_2 exp(\boldsymbol{\beta}_2' \mathbf{x}_i); \end{aligned} \tag{28}$$

em que $\boldsymbol{\beta}_j = (\beta_{j1}, ..., \beta_{jp})'$ é o vetor de parâmetros da regressão, $j = 1, 2$, associados com o vetor de covariáveis $\mathbf{x} = (x_{1i}, ..., x_{ip})'$, $i = 1, ..., n$.

Considera-se também a presença de dados censurados.

Para uma análise bayesiana, assumem-se a seguintes distribuições a priori para γ_j, α_j, β_{jk} e θ:

$$\begin{array}{rcl} \alpha_j & \sim & U(a_j, b_j) \\ \gamma_j & \sim & U(c_j, d_j) \\ \theta & \sim & U(-1, 1) \\ \beta_{jk} & \sim & N(0, g^2); \end{array} \tag{29}$$

para $j = 1, 2$; em que $k = 1, ..., p$ e a_j, b_j, c_j, d_j e g são hiperparâmetros conhecidos e $N[0, g^2]$ denota a distribuição normal com média zero e variância g^2.

Observe que outras distribuições de sobrevivência marginais poderiam ser usadas na construção da distribuição bivariada usando a função cópula de Farlie-Gumbel-Morgenstern. Em especial, poderíamos usar distribuições marginais de Weibull.

2.4.1 Uma aplicação com dados de sobrevivência para pacientes com câncer gástrico

Considerar um conjunto de dados de pacientes com câncer gástrico coletados no Hospital de Câncer de Barretos (Jácome *et al.*, 2012),estado de São Paulo. As coletas foram feitas através de prontuários e da seleção dos espécimes tumorais(adenocarcinoma gástrico) conservados em parafina e armazenados no arquivo do Departamento de Anatomia Patológica do Hospital de Câncer de Barretos. Foram analisados prontuários de 2006 a 2008, totalizando 230 pacientes. Foram coletados dois tempos de sobrevida: o tempo de sobrevida livre de evento (TSL) e o tempo de sobrevida global (TSG). Foi definido pelo pesquisador que a sobrevida livre de evento (TSL) é o tempo decorrente (em meses) entre a gastrectomia e a ocorrência de recidiva ou progressão documentada por exames de imagem ou novo procedimento cirúrgico ou óbito por qualquer causa; e sobrevida global (TSG)é definida como o tempo decorrente (em meses) entre a gastrectomia e o óbito por qualquer causa. Pacientes que perderam seguimento ao longo de sua evolução foram censurados na data do último contato. Foram relatadas pelo pesquisador, algumas covariáveis que poderiam ter interferência no tempo de sobrevida, mas para este modelo, em particular, foi utilizado apenas a covariável de define os tratamentos (quimioterapia e controle). Nesta aplicação, assumimos os modelos de regressão na forma,

$$\lambda_j(\mathbf{x}_i) \quad = \quad exp\left\{\beta_0 + \beta_1 x_{1i}\right\};$$

para $i = 1, 2, ..., 230$ (número de índivíduos) e $j = 1, 2$ (TSL e TSG), em que x_{1i} denota a covariável tipos de tratamento.

Para o modelo multivariado foram utilizados duas distribuições, a exponencial generalizada e a Weibull. A partir de uma análise preliminar dos dados considerando-se os tempos de sobrevivência completos, foi verificado uma dependencia fraca entre os dois tempos de sobrevivência. Isso justifica o uso da função cópula Farlie-Gumbel-Morgensten.

Para os dois modelos foi utilizado uma distribuição a priori $Uniforme(-1, 1)$ para o parâmetro associado à dependência entre as variáveis aleatórias (θ), e para o β_0 e β_1 (referente a covariável), adotou-se uma distribuição a priori normal com hiperparâmetros $a_\beta = 0$, $b_\beta = 0, 1$.

Para o modelo multivariado baseado na distribuição exponencial generalizada, foi utilizado uma distribuição a priori com distribuição uniforme variando entre 0 e $0, 2$ para α_1 e α_2, parâmetros da distribuição. Para γ_1 e γ_2 foi utilizado uma distribuição a prori Gama com hiperparâmetros $a_\alpha = 15$ e $b_\alpha = 10$.

Para o modelo multivariado baseado na distribuição Weibull, os parâmetros γ_1 e γ_2 receberam uma distribuição a priori com distribuição uniforme variando entre 0 e 2. Para o parâmetro de forma α_2 foi utilizado uma gama com hiperparâmetros $a_\alpha = 540$ e $b_\alpha = 10$, e para o α_1 foi realizado uma reparametrização utilizando um parâmetro ς com distribuição exponencial, portanto, $\alpha_1 \sim exp(\gamma)$ e $\gamma \sim Gama(40, 10)$.

Para a distribuição a posteriori dos parâmetros dos modelos com distribuição Weibull e exponencial generalizada foram baseadas em 20.000 amostras simuladas de Gibbs, com saltos de 10 em 10. Foram descartadas as 5.000 primeiras amostras ("burn-in-sample"). A convergência do algoritmo foi observada usando métodos gráficos.

Tabela 1 - Sumários a posteriori para a Distribuição Exponencial Generalizada

Parâmetros	Média	DP	IC (95%)
α_1	1,57	0,19	(1,22; 1,96)
α_2	1,60	0,18	(1,28; 1,97)
γ_1	0,03	0,01	(0,02; 0,03)
γ_2	0,03	0,01	(0,02; 0,04)
θ	0,97	0,03	(0,90; 1,00)
β_{11}(Quimio x Cont)	-0,13	0,15	(-0,16; 0,43)
β_{12}(Quimio x Cont)	0,14	0,13	(-0,13; 0,41)

(DP = desvio-padrão)

O modelo utilizando a distribuição exponencial generalizada não identificou efeito do tratamento no tempo de sobrevida.

Tabela 2 - Sumários a posteriori para a Distribuição Weibull

Parâmetros	Média	DP	IC (95%)
α_1	56,11	7,72	(44,11;73,78)
α_2	52,54	2,21	(48,32;57,02)
ς	4,01	0,13	(3,78;4,30)
γ_1	1,35	0,13	(1,10;1,62)
γ_2	1,23	0,09	(1,04;1,43)
θ	0,96	0,05	(0,89;1,00)
β_{11}(Quimio x Cont)	-0,14	0,15	(-0,43;0,15)
β_{12}(Quimio x Cont)	-0,28	0,12	(-0,52;-0,02)

(DP = desvio-padrão)

O modelo utilizando a distribuição Weibull identificou efeito do tratamento no tempo de sobrevida livre de evento.

Tabela 3 - DIC

Distribuições	DIC
Exponencial Generalizada	1.590,8
Weibull	1.596,9

Para comparação dos dois modelos considerados, usamos o critério bayesiano DIC (Apêndice). Através da Tabela 3 pode-se observar que os valores do DIC obtidos dos ajustes dos modelos estão muito próximos, mas o modelo utilizando a distribuição exponencial generalizada se mostrou mais efetivo.

Utilizando a distribuição Weibull, observa-se que quando inserimos no modelo um parâmetro de depêndencia entre os tempos, consegue-se detectar uma possível intervenção do tratamento no tempo de sobrevida.

Com relação ao parâmetro de dependência, dada a dificuldade em obter informação a priori, pode-se elaborar um trabalho de especificação de distribuições a priori apenas para este parâmetro. Isso também pode ser objetivo de uma pesquisa futura.

Conclusões

Uma generalização da distribuição exponencial generalizada é considerada para dados bivariados usando funções cópula.

O uso de funções cópula pode ser um método eficaz para modelagem de dados bivariados em análise de sobrevivência considerando diferentes distribuições de sobrevivência marginais. Na literatura, observa-se que alguns autores têm investido fortemente neste tipo de estudo (Viola, 2009). Segundo estes autores, o uso de funções cópula tem sido amplamente utilizada, na área de finanças, ciências atuárias,

estudos biomédicos e engenharias e também pode ser uma alternativa para ser usada com dados de sobrevivência multivariados.

A derivação utilizando a função de cópula de Farlie- Gumbel-Morgenstern é bastante utilizada por pesquisadores dada a sua simplicidade quando comparada com outras funções cópulas. Alguns autores, como Achcar e Santos (2010, 2011) ou Suzuki (2012), consideraram este tipo de função cópula para analisar dados de sobrevivência bivariados na presença de dados censurados e com a presença de covariáveis, e mostra que este tipo de modelo pode ter um bom desempenho quando os dados apresentam uma fraca dependência. Para situações onde temos dependência moderada ou forte, devemos usar outras funções cópula existentes na literatura para construção de modelos multivariados com dados de sobrevivência (Nelsen, 1999). Outras funções cópulas também poderiam ser exploradas na análise dos dados considerados nesse artigo, como as funções de cópula de Gumbel, Clayton entre outras, mas isso será objetivo de um próximo artigo.

Na área médica, vários tipos de doenças apresentam mais de um evento de interesse, consequentemente apresentam tempo de sobrevida para cada um desses eventos. Em paralelo, esses tempos de sobrevida possuem uma dependência, mesmo se tratando de eventos distintos. Nestas situações é de suma importância a utilização de modelos multivariados com um parâmetro de dependência flexível.

Através das aplicações pode-se observar que a distribuição exponencial generalizada pode ser uma opção quando comparada à outras distribuições mais utilizadas em análise de sobrevivência. Importante observar que para alguns conjuntos de dados, a distribuição exponencial generalizada pode se destacar mais eficaz que a distribuição Weibull (Achcar e Boleta, 2009).

O uso de métodos Bayesianos considerando técnicas de simulação MCMC (Monte Carlo em Cadeias de Markov) leva à obtenção de inferências precisas para os parâmetros do modelo. Além disso, as quantidades a posteriori de interesse podem ser obtidas de forma simples usando o software WinBugs que não requer grandes custos computacionais.

Também é importante salientar que para uma análise Bayesiana dos dados, o uso de outros programas computacionais possivelmente com códigos em R podem ser alternativas melhores ao uso do software WinBugs permitindo o uso de distribuições a priori mais não-informativas, inclusive usando outros algoritmos MCMC diferentes do Gibbs e do Metropolis-Hastings existentes na literatura. Isso será o objetivo de uma pesquisa futura.

Agradecimentos

Agradecemos, em especial, ao pesquisador Alexandre Jácome, por ceder os dados para aplicação do modelo proposto.

BOLETA, J.; ACHCAR, J. A. Generalized exponential distribution of derivative bivariate copula functions: an application to gastric cancer data. *Rev. Bras. Biom.,* São Paulo, v.30, n.4, p.401-414, 2012.

▪ *ABSTRACT: In this paper, we introduce the use of generalized exponential distributions to analise multivariate lifetime data in presence of censored data and covariates derived from Farlie-Gumbel-Morgenstern copula functions.We assume different priors for the parameters of the model and we have used MCMC (Markov Chain Monte Carlo) methods to get the posterior summaries of interest.*

▪ *KEYWORDS: Generalized exponential distribution; copula functions; Bayesian models; gastric cancer.*

Referências

ACHCAR, J. A.; BOLETA, J. Distribuição exponencial generalizada: uso de métodos Bayesianos. *Rev. Bras. Biom.*, São Paulo, v.27, n.4, p.644-658, 2009.

ACHCAR, J. A. ; SANTOS, C. A. A Bayesian analysis for multivariate data in the presence of covariates. *J. Stat. Theory Appl.*, Holland, v.9, p.233-253, 2010.

ACHCAR, J. A. ; SANTOS, C. A. A Bayesian analysis in the presence of covariates for multivariate survival data: an example of application. *Rev. Colomb. Estad.*, Bogotá, v.34, n.11, p.111-131, 2011.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions. *Austral. N. Zeal. J. Stat.*, Richmond, v.41, n.2, p.173-188, 1999.

GUPTA, R. D.; KUNDU, D. Generalized exponential distribution:existing results and some recent developments. *J. Stat. Plan. Inference*, Amsterdam, v.137, n.11, p.3537-3547, 2007.

JÁCOME, A. A. A.; WOHNRATH, D. R.; SCAPULATEMPO NETO, C.; FREGNANI, J. H. T. G.; QUINTO, A. L.; OLIVEIRA, A. T. T.; VAZQUEZ, V. L.; FAVA G.; MARTINEZ, E. Z.; SANTOS J. S. Effect of adjuvant chemoradiotherapy on overall survival of gastric cancer patients submitted to D2 lymphadenectomy. *Gastric Cancer*, tOKYO, N.16, N.2, P.233-238, 2013.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New Jersey: John Wiley e Sons, 1982. 330p.

MORGENSTERN, D. Einfache Beispiele zweidimensionaler Verteilungen. *Mit. Math. Statist.*, [S.L.], v.8, p.234-235, 1956.

NELSEN, R. B. *An introduction to copulas*. New York: Springer Verlag, 1999. 184p.

RAQAB, M. Z. Inferences for generalized exponential distribution based on record statistics. *J. Stat. Plan. Inference*, Amsterdam, v.104, n.2, p.339-350, 2002.

RAQAB, M. Z.; AHSANULLAH, M. Estimation of the location and scale parameters of generalized exponential distribution based on order statistics. *J. Stat. Comput. Simul.*, Abingdon v.69, n.12, p.109-124, 2001.

SARHAN, A. M. Analysis of incomplete, censored data in competing risks models with generalized exponential distributions. *IEEE Trans. Reliab.*, Piscataway, v.56, p.132-138, 2007.

SKLAR, M. Fonctions de répartition à n-dimensions leurs marges. *Publ. Inst. Statist. Univ.*, Paris, v.8, p.229-231, 1959.

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

SUZUKI, A. K. *Modelos de sovrevivência bivariados baseados na cópula FGM: uma abordagem Bayesiana*. 2012. 91f. Tese (Doutorado em Estatística) - Universidade Federal de São Carlos, São Carlos, 2012.

VIOLA, M. L. L. *Tipos de dependência entre variáveis aleatórias e teoria de cópulas*. Campinas: Instituto de Matemática, Estatística e Computação Científica (IMECC-UNICAMP), 2009. 105p.

ZHENG, G. Fisher information matrix in type-II censored data from exponentiated exponential family. *Biom. J.*, Weinheim, v.44, p.353-357, 2002.

Recebido em 30.11.2012.

Aprovado após revisão em 02.04.2013.

APÊNDICE – Critério de Informação Deviance (DIC)

Para a seleção de modelos, podemos usar diferentes critérios de seleção. Um critério Bayesiano muito popular e implementado no software WinBUGS (Lunn *et al.*, 2000) é dado pelo critério de informação deviance (DIC).

O DIC é um critério útil para a seleção de modelos sob o enfoque Bayesiano, onde amostras da distribuição a posteriori para os parâmetros do modelo são obtidas usando métodos MCMC.

O deviance (desvio) é definido por

$$D(\boldsymbol{\theta}) = -2lnL(\boldsymbol{\theta}) + C, \tag{30}$$

em que $\boldsymbol{\theta}$ é um vetor de parâmetros desconhecidos do modelo; $L(\boldsymbol{\theta})$ é a função de verossimilhança do modelo e C é uma constante que não é necessário que seja conhecida quando a comparação entre modelos é efetuada.

O critério DIC definido por SPIEGELHALTER *et al.* (2002) é dado por

$$DIC = D(\hat{\boldsymbol{\theta}}) + 2n_D, \tag{31}$$

onde $D(\hat{\boldsymbol{\theta}})$ é o desvio avaliado na média a posteriori $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|dados)$ e n_D é o número efetivo de parâmetros do modelo, dado por $n_D = \bar{D} - D(\hat{\boldsymbol{\theta}})$, onde $\bar{D} = E[D(\boldsymbol{\theta})|dados]$ é o desvio a posteriori que mede a qualidade do ajuste dos dados para cada modelo. Menores valores de DIC indicam melhores modelos. Observar que estes valores podem ser negativos.