Modelos de Sobrevivência Bivariados derivados da Cópula Arquimediana de Clayton: Uma Abordagem Bayesiana

Thiago Ramos Biondo
Universidade de São Paulo-USP
thiagob@usp.br

Adriano Kamimura Suzuki
Universidade de São Paulo-USP
suzuki@icmc.usp.br

RESUMO

Neste trabalho consideramos modelos baseados na cópula Arquimediana de Clayton com marginais Weibull ou Exponencial Generalizada para modelar a dependência de dados de sobrevivência bivariados na presença de covariáveis e observações censuradas. Para fins inferenciais, realizamos uma abordagem Bayesiana usando métodos Monte Carlo em Cadeias de Markov (MCMC). Além disso, algumas discussões sobre os critérios de seleção de modelos são apresentadas. Com o objetivo de detectar observações influentes nos dados analisados foi utilizado o método Bayesiano de análise de influência de deleção de casos baseados na divergência ψ. Mostramos a aplicabilidade dos modelos propostos a conjuntos de dados simulados e reais.

ABSTRACT

In this work we consider models based on Clayton Archimedian copulas with marginal Weibull or Generalized Exponential to model the dependence of bivariate survival data in the presence of covariates and censored data. For inferential purposes, a Bayesian approach via Markov Chain Monte Carlo (MCMC) was considered. Further, some discussions on the model selection criteria are given. In order to examine outlying and influential observations, we present a Bayesian case deletion influence diagnostics based on the divergence ψ. The applicability of the proposed models are illustrated on artificial and real data.

Palavras-chave: Cópulas Arquimedianas; Análise de sobrevivência; Inferência Bayesiana.

1 Introdução

Um conjunto de dados de sobrevivência multivariados geralmente apresenta associação entre os tempos de sobrevivência, em que essa possível associação é frequentemente modelada por meio de modelos de fragilidade. Nestes modelos, os quais foram proposto por [34], um ou mais efeitos aleatórios, denominado fragilidade, são introduzidos na função de risco para descrever essa possível heterogeneidade entre as unidades em estudo. Neste caso, os tempos marginais são condicionalmente independentes dada a variável fragilidade.

Recentemente, uma alternativa para modelar a dependência entre dados multivariados, especialmente em áreas biológicas, ciências atuariais e finanças é a modelagem por meio de cópulas (ver, por exemplo, [26], [19]). Por exemplo, em ciências atuariais cópulas são

usadas na modelagem de mortalidade e perdas ([13]). Em finanças, na classificação de crédito e modelagem de risco ([12]). Já em estudos biomédicos, na modelagem de eventos correlacionados e riscos competitivos ([1], [32]).

Uma cópula é uma função que conecta as distribuições marginais univariadas com sua distribuição multivariada conjunta. Diferentes cópulas representam diferentes estruturas de dependência entre as variáveis ([26]). Uma cópula é uma ferramenta conveniente para estudar a estrutura de dependência e é flexível em aplicações, uma vez que quando a dispersão dos dados não se ajusta a nenhuma família conhecida, pode ser difícil especificar a distribuição conjunta.

Dados bivariados são encontrados em situações em que se observa dois tempos de vida para um mesmo equipamento ou paciente. Por exemplo, na área médica pode ocorrer o interesse em estudar os tempos de vida de orgãos humanos emparelhados como rins e olhos, o tempo entre a primeira e a segunda internação, dentre outros. Já em aplicações industriais, este tipo de dados é observado, por exemplo, em sistema cujo o tempo de duração depende da durabilidade de dois componentes, como o tempo de vida de motores de um avião bimotor.

Neste trabalho, sob uma abordagem Bayesiana, modelamos a dependência de dados de sobrevivência bivariados na presença de covariáveis e observações censuradas por meio de cópulas Arquimedianas, em particular para a cópula de Clayton e, para ambas as distribuições marginais, assumimos a distribuição Exponencial Generalizada ou a distribuição Weibull. Para fins inferenciais, serão utilizados métodos Monte Carlo em Cadeias de Markov (MCMC). Com o objetivo de detectar observações influentes nos dados foi utilizado o método Bayesiano de análise de influência caso a caso baseado na divergência ψ. Mostramos a aplicabilidade dos modelos propostos a conjuntos de dados simulados e reais.

2 UMA BREVE INTRODUÇÃO ÀS FUNÇÕES CÓPULAS

As funções cópulas permitem construir modelos multivariados que separam o comportamento das variáveis aleatórias marginais da estrutura de dependência existente entre elas.

Cópulas são funções que ligam (conectam) a função distribuição conjunta com suas funções distribuição marginais univariadas. Alternativamente, cópulas são funções distribuição multivariadas cujas marginais unidimensionais são Uniformes em $[0,1]$ (para maiores detalhes, ver os livros [25], [20] e [19]).

Uma distribuição bivariada pertence à família de cópulas Arquimedianas se tem a seguinte representação:

$$C_\phi(u,v) = \varphi(\varphi(u)^{-1} + \varphi(v)^{-1}), \quad 0 \leq u, v \leq 1 \tag{1}$$

em que ϕ é o parâmetro de dependência da cópula e $\varphi : [0,+\infty) \to [0,1]$ com $\varphi(0) = 1$ e $\lim_{x\to+\infty} \varphi(x) = 0$.

A representação da cópula Arquimediana permite reduzir o estudo de cópula multivariada ao estudo de uma função univariada φ, comumente chamada de gerador de uma cópula Arquimediana.

Considere $C(u,v)$ uma cópula Arquimediana bivariada. Temos então as seguintes propriedades:

1. $C(u,v) = C(v,u)$, ou seja, C é simétrica (permutável);
2. $C(C(u,v),w) = C(u,C(v,w))$ para todo u,v,w em $[0,1]$, ou seja, C é associativa;
3. Seja ϕ a geradora de C. Então para qualquer constante $k > 0$ tem-se que $k\phi$ também é uma geradora de C.

Neste trabalho utilizamos a cópula Arquimediana bivariada de Clayton ([7]), que tem a forma:

$$C_\phi(u,v) = (u^{-\phi} + v^{-\phi} - 1)^{-\frac{1}{\phi}}, \tag{2}$$

em que $\phi > 0$ e a função geradora é dada por $\varphi(t) = \frac{1}{\phi}(t^{-\phi} - 1)$.

Considerando as funções de sobrevivência $u = S(t_1)$ e $v = S(t_2)$, temos que a função de sobrevivência é dada por:

$$S(t_1,t_2) = C_\phi(S(t_1),S(t_2)) = (S(t_1)^{-\phi} + S(t_2)^{-\phi} - 1)^{\frac{-1}{\phi}}, \tag{3}$$

em que $S(t_1)$ e $S(t_2)$ são as funções de sobrevivência associadas a cada uma das distribuições marginais.

2.1 FUNÇÕES DENSIDADES DE PROBABILIDADE

Nesse trabalho para ambas as distribuições marginais, assumimos a distribuição Exponencial Generalizada ou a distribuição Weibull, que são apresentadas brevemente a seguir.

DISTRIBUIÇÃO EXPONENCIAL GENERALIZADA

Uma distribuição Exponencial Generalizada ([16]) é uma boa alternativa ao uso das populares distribuições Gama e Weibull utilizadas na análise de dados de sobrevivência ([2]).

A distribuição Exponencial Generalizada de dois parâmetros tem função densidade de probabilidade dada por:

$$f(t) = \alpha\lambda\,(1 - exp(-\lambda t))^{\alpha-1}\,exp(-\lambda t), \tag{4}$$

em que $t > 0$, $\alpha > 0$ e $\lambda > 0$ são os parâmetros de forma e escala, respectivamente.

As funções de sobrevivência e de risco associadas à essa densidade são dadas, respectivamente, por:

$$S(t) = P(T > t) = 1 - (1 - exp(-\lambda t))^{\alpha} \tag{5}$$

e

$$h(t) = \frac{f(t)}{S(t)} = \frac{\alpha\lambda\,(1 - exp(-\lambda t))^{\alpha-1}\,exp(-\lambda t)}{1 - (1 - exp(-\lambda t))^{\alpha}}. \tag{6}$$

A Figura 1 apresenta o comportamento, para diferentes valores de α, da função densidade de probabilidade, da função de sobrevivência e da função de risco dadas em (4), (5) e (6), respectivamente.

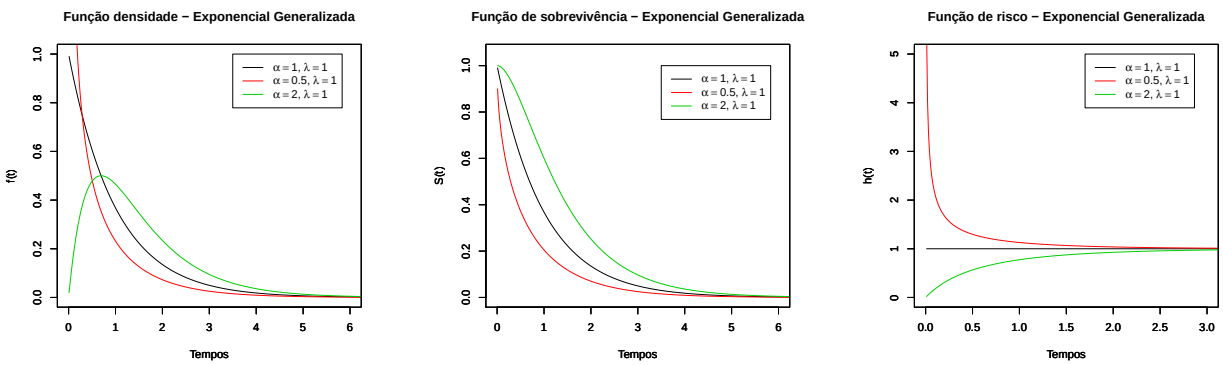

FIGURA 1: Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores de α (com $\lambda = 1$).

DISTRIBUIÇÃO WEIBULL

A distribuição Weibull ([37]) é amplamente conhecida em virtude de sua simplicidade e flexibilidade em acomodar diferentes formas de função de risco, sendo um dos modelos mais utilizados em análise de sobrevivência.

Para uma variável aleatória T com distribuição Weibull, a função densidade de probabilidade é dada por:

$$f(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1} exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right) \tag{7}$$

em que $t \geq 0$, $\alpha > 0$ e $\beta > 0$ são os parâmetros de forma e escala, respectivamente.

A função de sobrevivência do modelo Weibull é dada por:

$$S(t) = exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right) \tag{8}$$

e a função de risco por:

$$h(t) = \frac{\alpha}{\beta^{\alpha}} t^{\alpha-1}. \tag{9}$$

Esta distribuição possui riscos crescentes para $\alpha > 1$, decrescentes para $\alpha < 1$ e constantes para $\alpha = 1$, em que o modelo se reduz a distribuição Exponencial.

A Figura 2 apresenta o comportamento, para diferentes valores de α e β, da função densidade de probabilidade, da função de sobrevivência e da função de risco dadas em (7), (8) e (9), respectivamente.

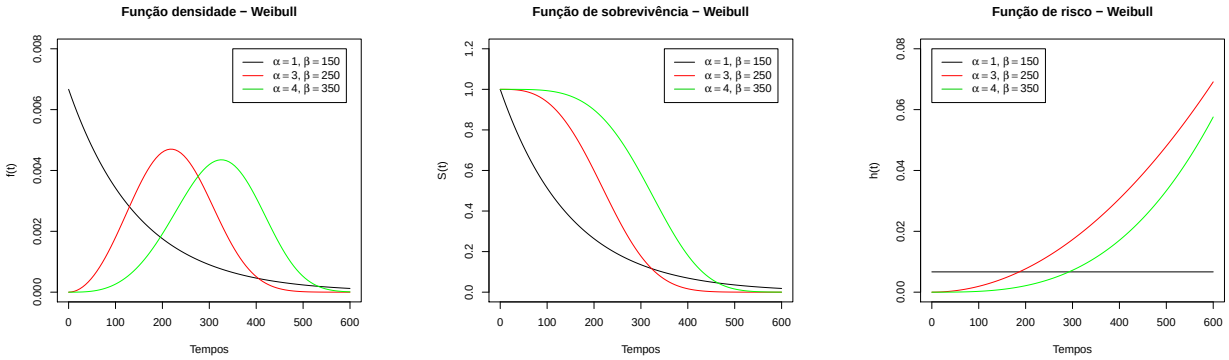

FIGURA 2: Gráfico da função densidade de probabilidade (esquerda), da função de sobrevivência (centro) e da função de risco (direita) para diferentes valores de α e β.

2.2 INFERÊNCIA

Considere (T_{i1}, T_{i2}) e (C_{i1}, C_{i2}) os i-ésimos tempos de vida e de censura bivariados, para $i = 1, \ldots, n$. Suponha que (T_{i1}, T_{i2}) e (C_{i1}, C_{i2}) são independentes. Para cada indivíduo i, as quantidades individuais são representadas pelas variáveis aleatórias $t_{ij} = \min(T_{ij}, C_{ij})$, para $j = 1, 2$. Para especificação paramétrica das marginais é possível ligar covariáveis em cada um dos componentes, isto é, $\boldsymbol{\gamma}_j$=$\varphi(x_j)$. Por simplicidade e para evitar restrições no espaço paramétrico, consideramos $\varphi(x_j)$=exp($\boldsymbol{x}_j^{\top}\boldsymbol{\beta}_j$), em que $\boldsymbol{\beta}_j$ corresponde ao vetor de coeficientes desconhecidos de ordem $q_j \times 1$, associado às covariáveis x_j, $j = 1, 2$.

Seja $\mathcal{D} = \{(t_{i1}, t_{i2}, x_{ij}, \delta_{ij}); j = 1, 2; i = 1, ..., n\}$ os dados observados, em que x_{ij} é um vetor de covariáveis de dimensão q, e $\delta_{ij} = I(t_{ij} = T_{ij})$ denota o indicador de censura. Considerando a função de sobrevivência bivariada $S(t_1, t_2)$ dada em (3), a função de verossimilhança ([21]) é dada por:

$$L(\boldsymbol{\vartheta}|\mathcal{D}) = \prod_{i=1}^{n} (f(t_{i1}, t_{i2}))^{\delta_{i1}\delta_{i2}} \left(-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i1}}\right)^{\delta_{i1}(1-\delta_{i2})} \left(-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_{i2}}\right)^{\delta_{i2}(1-\delta_{i1})} (S(t_{i1}, t_{i2}))^{(1-\delta_{i1})(1-\delta_{i2})}, \tag{10}$$

em que $\boldsymbol{\vartheta}$=$(\phi, \alpha_1, \alpha_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ e $f(t_{i1}, t_{i2}) = \frac{\partial^2 S(t_{i1}, t_{i2})}{\partial t_{i1} \partial t_{i2}}$ é a função de densidade bivariada. Para obter a função de verossimilhança é preciso obter as derivadas de $S(t_1, t_2)$, que são dadas abaixo:

$$\frac{\partial S(t_1, t_2)}{\partial t_1} = -f(t_1) S(t_1)^{-\phi-1} (S(t_1)^{-\phi} + S(t_2)^{-\phi} - 1)^{-\frac{1}{\phi}-1},$$

$$\frac{\partial S(t_1, t_2)}{\partial t_2} = -f(t_2) S(t_2)^{-\phi-1} (S(t_1)^{-\phi} + S(t_2)^{-\phi} - 1)^{-\frac{1}{\phi}-1}$$

e

$$\frac{\partial S(t_1, t_2)}{\partial t_1 \partial t_2} = \frac{\partial S(t_1, t_2)}{\partial t_2 \partial t_1} = (S(t_1)^{-\phi} + S(t_2)^{-\phi} - 1)^{\frac{-1}{\phi}-2} \prod_{i=1}^{2} f(t_i) S(t_i)^{-\phi-1} (1+\phi).$$

Consideramos uma distribuição *a priori* conjunta própria para os parâmetros do modelo para garantir que a distribuição *a posteriori* conjunta seja própria (17). Assumindo as distribuições *a priori* independentes, a densidade *a priori* conjunta de $\boldsymbol{\vartheta}$=$(\phi, \alpha_1, \alpha_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ é dada por:

$$\pi(\boldsymbol{\vartheta}) = \pi(\phi) \prod_{j=1}^{2} \pi(\alpha_j) \prod_{k=1}^{2} \pi(\boldsymbol{\beta}_k), \tag{11}$$

em que

$$\pi(\boldsymbol{\beta_j}) = \pi(\beta_{j1}, ..., \beta_{jq}) = \prod_{i=1}^{q} \pi(\beta_{ji}).$$

Assumimos as seguintes distribuições *a priori*:

$$\phi \sim Gama(0.1, 0.01)$$

$$\alpha_k \sim Gama(0.1, 0.01)$$

$$\beta_{ji} \sim Normal(0, 10^3), \tag{12}$$

para $k = 1, 2$, $j = 1, 2$ e $i = 1, ..., q$.

Combinando a função de verossimilhança (10) e a distribuição *a priori* (11), a distribuição conjunta *a posteriori* de $\boldsymbol{\vartheta}$ é dada por:

$$\pi(\boldsymbol{\vartheta}|\mathcal{D}) \propto L(\boldsymbol{\vartheta}|\mathcal{D}) \pi(\boldsymbol{\vartheta}). \tag{13}$$

Neste artigo assumimos que as marginais T_j têm distribuição Exponencial Generalizada ou Weibull com parâmetros α_j e $\lambda_{ij} = \exp(\beta_{0j} + \beta_{1j} x_i)$, $i = 1, ..., n$ e $j = 1, 2$. Esta densidade *a posteriori* conjunta é analiticamente intratável. Assim, para fins inferenciais utilizamos métodos MCMC. Todas as implementações computacionais foram realizadas utilizando os sistemas JAGS -Just Another Gibbs Sampler ([28]) e R ([30]) por meio do pacote *rjags* ([10]). As estimativas dos parâmetros são dadas pelas médias da distribuição *a posteriori*.

2.3 CRITÉRIOS DE COMPARAÇÃO DE MODELOS E DIAGNÓSTICO DE OBSERVAÇÕES INFLUENTES

Na literatura encontramos diversas metodologias que se propõem a analisar a adequabilidade de um modelo a um certo conjunto de dados, além de, dentre uma coleção de modelos, selecionar o melhor.

Analogamente à [23], neste trabalho utilizamos quatro critérios de seleção de modelos: o DIC (*Deviance Information Criterion*) proposto por [31], o EAIC (*Expected Akaike Information Criterion*) por [3], o EBIC (*Expected Bayesian* (ou *Schwarz*) *Information Criterion*) por [5] e o B (*Mean of Logarithm of the Pseudo Marginal Likelihood*) que é derivado das ordenadas da densidade preditiva condicional (CPO) ([6]).

Uma forma utilizada de avaliação da influência de uma observação no ajuste de um modelo é por meio da exclusão de casos ([8]). Atualmente, técnicas de influência local têm sido amplamente utilizadas, por exemplo em [4], [35] e [22].

Neste trabalho vamos considerar a análise de influência de deleção de casos baseado na divergência ψ. Seja $D_\psi(P; P_{(-i)})$ a divergência ψ entre P e $P_{(-i)}$, em que P indica a distribuição *a posteriori* de $\boldsymbol{\vartheta}$ para os dados completos e, $P_{(-i)}$ a distribuição *a posteriori* sem o i-ésimo caso. Especificamente,

$$D_\psi(P; P_{(-i)}) = \int \psi\left(\frac{\pi(\boldsymbol{\vartheta}|\mathcal{D}^{(-i)})}{\pi(\boldsymbol{\vartheta}|\mathcal{D})}\right)\pi(\boldsymbol{\vartheta}|\mathcal{D})\,d\boldsymbol{\vartheta}, \tag{14}$$

em que ψ é uma função convexa com $\psi(1) = 0$. Várias escolhas de ψ são dadas em [11]. Por exemplo, $\psi(z) = -\log(z)$ define a divergência de Kullback-Leibler (K-L), $\psi(z) = (z-1)\log(z)$ a distância J (ou a versão simétrica da divergência de K-L), $\psi(z) = 0.5|z-1|$ a distância variacional ou norma L_1 e $\psi(z) = (z-1)^2$ define a divergência χ^2.

Podemos calcular $D_\psi(P; P_{(-i)})$ considerando uma amostra da distribuição *a posteriori* de $\boldsymbol{\vartheta}$ via métodos MCMC. Considere $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(V)}$ uma amostra de tamanho V de $\pi(\boldsymbol{\vartheta}|\mathcal{D})$. Então, uma estimativa Monte Carlo é dada por:

$$\widehat{D_\psi}(P; P_{(-i)}) = \frac{1}{V}\sum_{q=1}^{V}\psi\left(\frac{\pi(\boldsymbol{\vartheta}^{(q)}|\mathcal{D}^{(-i)})}{\pi(\boldsymbol{\vartheta}^{(q)}|\mathcal{D})}\right). \tag{15}$$

A medida $D_\psi(P; P_{(-i)})$ pode ser interpretada como a divergência ψ do efeito da exclusão do i-ésimo caso dos dados completos na distribuição *a posteriori* de $\boldsymbol{\vartheta}$.

Como apontado por [27] e [38], pode ser difícil para um profissional (por exemplo, um médico) avaliar o ponto de corte da medida de divergência, de modo a determinar se uma observação ou um pequeno subconjunto de observações é influente ou não. Neste contexto, usamos a proposta dada por [27] e [38]. Considere uma moeda viesada com probabilidade de sucesso p. Então, a divergência ψ entre a moeda viesada e a não viesada é:

$$D_\psi(f_0; f_1) = \int \psi\left(\frac{f_0(x)}{f_1(x)}\right) f_1(x)dx, \tag{16}$$

em que $f_0(x) = p^x(1-p)^{1-x}$ e $f_1(x) = 0.5$, $x = 0, 1$. Se $D_\psi(f_0, f_1) = d_\psi(p)$, então pode ser facilmente verificado que d_ψ satisfaz a seguinte equação:

$$d_\psi(p) = \frac{\psi(2p) + \psi(2(1-p))}{2}. \tag{17}$$

Não é difícil notar que, para as medidas de divergência consideradas, d_ψ aumenta à medida que p afasta-se de 0.5. Além disso, $d_\psi(p)$ é simétrica em torno de $p = 0.5$ e d_ψ atinge seu mínimo em $p = 0.5$. Neste ponto, $d_\psi(0.5) = 0$ e $f_0 = f_1$. Portanto, se considerarmos $p > 0.80$ (ou $p \leq 0.20$) como uma moeda muito viciada, então $d_{L_1}(0.80) = 0.30$. Esta relação implica que o i-ésimo caso é considerado influente quando $d_{L_1}(0.80) > 0.30$.

Logo, se utilizarmos a divergência de Kullback-Leibler, podemos considerar que uma observação é influente quando $d_{\text{K-L}} > 0.223$. Da mesma forma, usando a distância J ou a divergência χ^2, uma observação na qual $d_J > 0.416$ ou $d_{\chi^2}(0.80) > 0.360$ pode ser considerada influente.

3 ESTUDO DE SIMULAÇÃO

Inicialmente empregamos dados simulados para estudar as propriedades frequentistas dos estimadores Bayesianos quando os parâmetros do modelo são conhecidos. O objetivo deste estudo de simulação é mostrar o bom comportamento das estimativas Bayesianas, com base na média frequentista e nas medidas utilizadas para comparação de modelos: EAIC, EBIC, DIC e B.

Para simular n observações (t_{i1}, t_{i2}) do modelo baseado na cópula de Clayton, assumindo que as marginais T_j têm distribuição Exponencial Generalizada ou Weibull, com paramêtros α_j e $\lambda_{ij} = \exp(\beta_{0j} + \beta_{1j}x_i)$, $j = 1, 2$, realizamos o seguinte algoritmo:

Algoritmo implementado no *software* R

Passo 1: Gerar as covariáveis x_i de uma distribuição Bernoulli com parâmetro 0.5.

Passo 2: Gerar os tempos de censura C_{ij} a partir de uma distribuição Uniforme $U(0, \tau_j)$, com τ_j controlando o percentual de observações censuradas, $j = 1, 2$.

Passo 3: Gerar $u_{i1} \sim U(0, 1)$ para obter o T_{i1} e calcular t_{i1} da seguinte forma:
- Se for para a distribuição Weibull: Gerar $T_{i1} = (-\log(1 - u_{i1})/\lambda_{i1})^{1/\alpha_1}$ em que $u_{i1} \sim U(0, 1)$. Comparar T_{i1} com o valor de censura C_{i1} a fim de determinar o indicador de censura δ_{i1} e o valor observado dado por $t_{i1} = \min(T_{i1}, C_{i1})$.
- Se for para a distribuição Exponencial Generalizada: Gerar $T_{i1} = ((-log(u_{i1})/\lambda_{i1}))^{1/\alpha_1}$ em que $u_{i1} \sim U(0, 1)$. Comparar T_{i1} com o valor de censura C_{i1} a fim de determinar o indicador de censura δ_{i1} e o valor observado dado por $t_{i1} = \min(T_{i1}, C_{i1})$.

Passo 4: Gerar $u_{i2} \sim U(0, 1)$ para obter o T_{i2} e calcular t_{i2} da seguinte forma:
- Se for para a distribuição Weibull: calcular $w_i = [u_{i1}^{-\phi}(u_{i2}^{-\phi/(\phi+1)} - 1) + 1]^{(-1/\phi)}$.
Obter $T_{i2} = (-\log(1 - w_i)/\lambda_{i2})^{1/\alpha_2}$ e então comparar T_{i2} com o valor de censura C_{i2} a fim de determinar o indicador de censura δ_{i2} e o valor observado dado por $t_{i2} = \min(T_{i2}, C_{i2})$.
- Se for para a distribuição Exponencial Generalizada: calcular $w_i = [u_{i1}^{-\phi}(u_{i2}^{-\phi/(\phi+1)} - 1) + 1]^{(-1/\phi)}$.
Obter $T_{i2} = (-\log(w_i)/\lambda_{i2})^{1/\alpha_2}$ e então comparar T_{i2} com o valor de censura C_{i2} a fim de determinar o indicador de censura δ_{i2} e o valor observado dado por $t_{i2} = \min(T_{i2}, C_{i2})$.

Neste trabalho, simulamos os conjuntos de dados assumindo $(0\%, 0\%)$ e $(30\%, 20\%)$ de censuras para dois diferentes tamanhos de amostras $N = 50$ e $N = 200$. Para cada caso, geramos 200 conjuntos de dados. Para o modelo com marginais Weibull, considerando os seguintes parâmetros: $\alpha_1 = 2$, $\beta_{01} = -1$, $\beta_{11} = 0.5$, $\alpha_2 = 3$, $\beta_{02} = 1$, $\beta_{12} = -0.5$ e $\phi = 4$. Já para o modelo com marginais Exponencial Generalizada, os seguintes parâmetros: $\alpha_1 = 0.5$, $\beta_{01} = 1.5$, $\beta_{11} = -0.5$, $\alpha_2 = 1$, $\beta_{02} = 3$, $\beta_{12} = -1$ e $\phi = 2$.

Para cada conjunto de dados gerados simulamos duas cadeias de tamanho 65.000 para cada parâmetro. Foram desconsideradas as primeiras 10.000 iterações para eliminar o efeito dos valores iniciais. Para evitar problemas de autocorrelação, consideramos um espaçamento de tamanho 10, obtendo uma amostra final de tamanho 11.000 sobre a qual a inferência *a posteriori* é baseada. Para cada amostra, a média e o desvio padrão *a posteriori* dos parâmetros e os valores dos critérios de seleção de modelo são gravados.

A convergência das cadeias foi monitorada de acordo com os métodos recomendados por [9] (pacote CODA [28]). Em todos os casos, a convergência foi verificada por meio do diagnóstico de Gelman-Rubin [14] sendo muito próximo a 1 (≤ 1.01).

A Tabela 1 apresenta a média Monte Carlo (MC) das estimativas dos parâmetros ajustando as distribuições Weibull e Exponencial Generalizada na cópula de Clayton para o caso sem censura $(0\%, 0\%)$ e com censura $(30\%, 20\%)$ e dois tamanhos de amostras simuladas. Podemos observar que nos casos em que se gera e se obtém o ajuste do mesmo modelo (com e sem a presença de censura) as estimativas obtidas estão próximas, em média, do verdadeiro valor e, para os modelos cruzados, as estimativas diferem bastante, especialmente no caso em que o verdadeiro modelo é Weibull e ajustamos o modelo Exponencial Generalizada.

A Tabela 2 apresenta a média Monte Carlo (MC) dos quatro critérios de comparação de modelos para comparação dos modelos de sobrevivência bivariado baseado na cópula de Clayton com marginais Weibull ou Exponencial Generalizada. Podemos observar que, para os casos com e sem censura, o verdadeiro modelo gerado obteve melhor ajuste de acordo com todos os critérios considerados.

4 DIAGNÓSTICO DE OBSERVAÇÕES INFLUENTES

Para examinar o desempenho da medida de diagnóstico, geramos uma amostra de tamanho 400 para o modelo Clayton bivariado com marginais Weibull, considerando os seguintes parâmetros: $\alpha_1 = 2$, $\beta_{01} = -1$, $\beta_{11} = 0.5$, $\alpha_2 = 3$, $\beta_{02} = 1$, $\beta_{12} = -0.5$ e $\phi = 4$. Consideramos que 30% do tempo 1 e 25% do tempo 2 foram censurados.

Selecionamos os casos 100 e 200 (ambos os tempos observados), 300 (tempo 1 observado e tempo 2 censurado) e 400 (ambos os tempos censurados) para perturbação. Para criar observações artificialmente influentes no conjunto de dados, escolhemos um, dois ou três desses casos selecionados. Para cada caso, pertubamos um ou ambos os tempos da seguinte forma: $\widetilde{t_i} = t_i + 5D_t, i = 1, 2$, em que D_t é o desvio padrão dos t_i's. Para os casos 100, 200 e 400 ambos os tempos de vidas foram perturbados e para o caso 300 apenas o tempo de vida t_1 foi perturbado.

Para a implementação do algoritmo MCMC, assim como a verificação da convergência das cadeias, realizamos os mesmos procedimentos descritos anteriormente.

A Tabela 3 mostra que as inferências *a posteriori* são sensíveis à perturbação do(s) caso(s) selecionado(s).

A Tabela 4 apresenta os critérios Bayesianos do ajuste de diferentes casos de conjuntos de dados perturbados. Podemos observar que o conjunto de dados (a) (conjunto dos dados originais simulados) teve o melhor ajuste.

Vamos considerar as amostras da distribuição *a posteriori* dos parâmetros do modelo Clayton bivariado para obter uma estimativa das quatro medidas de divergência, cujos resultados estão apresentados na Tabela 5. A Tabela 5 mostra, antes da perturbação (conjuntos de dados (a)), que todos os casos selecionados não são influentes, com pequenas medidas de divergência. Entretanto, após perturbações (conjunto de dados (b) a (h)) as quatro medidas aumentam, indicando que os casos são influentes.

As Figuras 7 e 8 mostram os gráficos de índices das quatro medidas de divergência para o conjunto de dados (b) e (h). Claramente, podemos ver que as quatros medidas de divergência detectam os pontos influentes.

5 APLICAÇÃO À DADOS REAIS

Como aplicação à dados reais, utilizamos os dados apresentados em [24] que se refere a 38 pacientes com insuficiência renal. Os tempos (em dias) bivariados medidos são a respeito de recorrência de infecção no local onde foi inserido o catéter nos pacientes que utilizaram um aparelho portátil de diálise, sendo dado para cada paciente dois tempos de recorrência. Vamos considerar como covariável o sexo do paciente (0: masculino, 1: feminino).

Tabela 1: Média MC das estimativas dos parâmetros ajustando o modelo Clayton bivariado com marginais Weibull e Exponencial Generalizada para as duas configurações de censuras e tamanhos de amostras simuladas.

Verdadeiro modelo		Modelo ajustado							
		$N = 50$				$N = 200$			
	Parâmetro	Weibull		Exponencial Generalizada		Weibull		Exponencial Generalizada	
		(0%, 0%)	(30%, 20%)	(0%, 0%)	(30%, 20%)	(0%, 0%)	(30%, 30%)	(0%, 0%)	(30%, 20%)
Weibull	r_1	2.076	2.108	3.100	3.287	2.029	2.019	2.950	2.978
	β_{01}	-1.043	-1.054	0.166	0.184	-1.012	-1.009	0.164	0.171
	β_{11}	0.515	0.493	0.250	0.246	0.498	0.503	0.246	0.249
	r_2	3.124	3.128	10.188	9.563	3.036	3.039	9.244	8.708
	β_{02}	1.024	1.038	1.452	1.431	1.017	1.016	1.456	1.438
	β_{12}	-0.511	-0.555	-0.163	-0.169	-0.516	-0.514	-0.169	-0.169
	ϕ	4.092	4.232	5.020	4.858	3.978	4.020	4.821	4.517
Exponencial Generalizada	r_1	0.680	0.657	0.514	0.521	0.667	0.612	0.504	0.506
	β_{01}	1.551	1.714	1.513	1.387	1.507	1.350	1.499	1.465
	β_{11}	-0.354	-0.347	-0.521	-0.554	-0.338	0.340	-0.510	-0.524
	r_2	1.021	1.036	1.052	1.049	0.972	1.004	1.007	1.016
	β_{02}	3.073	3.394	3.007	2.962	2.923	3.004	2.996	2.992
	β_{12}	-1.031	-1.044	-1.006	-1.033	-0.978	-1.012	-1.006	-1.009
	ϕ	1.950	1.985	2.111	2.294	1.997	2.029	2.014	2.061

Tabela 2: Média Monte Carlo dos quatro critérios Bayesianos baseados sobre as 200 amostras geradas para as duas configurações de censura e tamanhos de amostras simuladas.

		Modelo	Modelo Ajustado							
			Exponencial Generalizada				Weibull			
			DIC	EAIC	EBIC	B	DIC	EAIC	EBIC	B
(0%, 0%)	$N = 50$	Weibull	31.814	38.412	51.796	-0.244	20.770	27.110	40.494	-0.131
		Exponencial Generalizada	-288.729	-282.334	-268.950	2.963	-286.347	-280.069	-266.685	2.491
	$N = 200$	Weibull	124.009	130.915	154.003	-0.292	78.519	85.350	108.438	-0.178
		Exponencial Generalizada	-1145.380	-1138.533	-1115.445	2.881	-1136.841	-1129.991	-1106.903	2.860
(30%, 20%)	$N = 50$	Weibull	48.332	54.787	68.172	-0.408	41.555	47.591	60.975	-0.336
		Exponencial Generalizada	-227.281	-220.928	-207.544	2.349	2290.283	-303.333	-289.949	3.173
	$N = 200$	Weibull	188.661	195.522	218.611	-0.454	158.333	165.072	188.160	-0.378
		Exponencial Generalizada	-924.268	-917.420	-894.332	2.329	-922.648	-915.788	-892.699	2.324

TABELA 3: Média e desvio padrão das estimativas *a posteriori* dos parâmetros do modelo.

Nome dos Dados	Identificação dos casos perturbados	α_1 Média DP	β_{01} Média DP	β_{11} Média DP	α_2 Média DP	β_{02} Média DP	β_{12} Média DP	ϕ Média DP
a	Nenhum	1.918	-1.003	0.487	2.812	0.929	-0.497	4.486
		0.075	0.085	0.087	0.105	0.071	0.088	0.430
b	100	1.932	-1.016	0.467	2.725	0.912	-0.524	4.510
		0.076	0.085	0.086	0.100	0.070	0.088	0.423
c	200	1.771	-0.992	0.580	2.375	0.704	-0.231	4.027
		0.073	0.085	0.090	0.083	0.070	0.087	0.409
d	300	1.787	-0.880	0.354	2.697	0.931	-0.562	4.103
		0.070	0.082	0.090	0.104	0.072	0.091	0.405
e	400	1.869	-1.014	0.510	2.713	0.859	-0.439	4.688
		0.073	0.086	0.089	0.100	0.070	0.088	0.439
f	{100, 200}	1.785	-1.002	0.558	2.331	0.697	-0.257	4.072
		0.073	0.087	0.089	0.082	0.071	0.087	0.415
g	{300, 400}	1.747	-0.893	0.377	2.605	0.868	-0.507	4.288
		0.068	0.080	0.087	0.098	0.069	0.089	0.418
h	{100, 200, 300}	1.693	-0.902	0.433	2.265	0.720	-0.342	3.773
		0.070	0.083	0.090	0.082	0.071	0.088	0.401

Legenda: DP denota o desvio padrão, o conjunto de dados (a) denota os dados originais simulados sem perturbação e os conjuntos de dados (b) a (h) denotam os conjuntos de dados com casos perturbados.

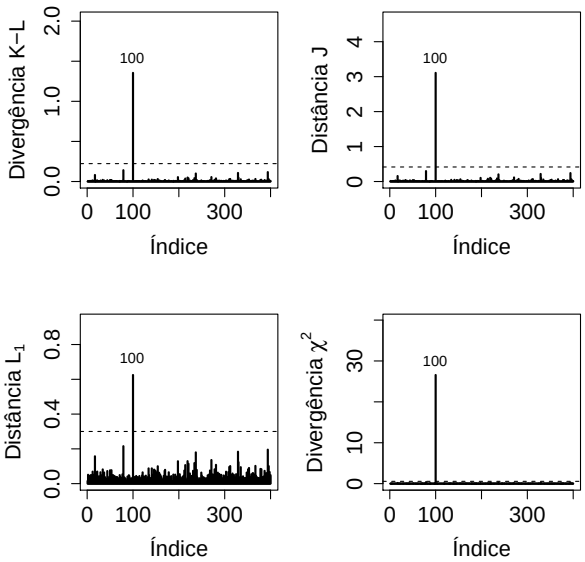

FIGURA 3: Gráficos de índices das medidas de divergência para o caso (b).

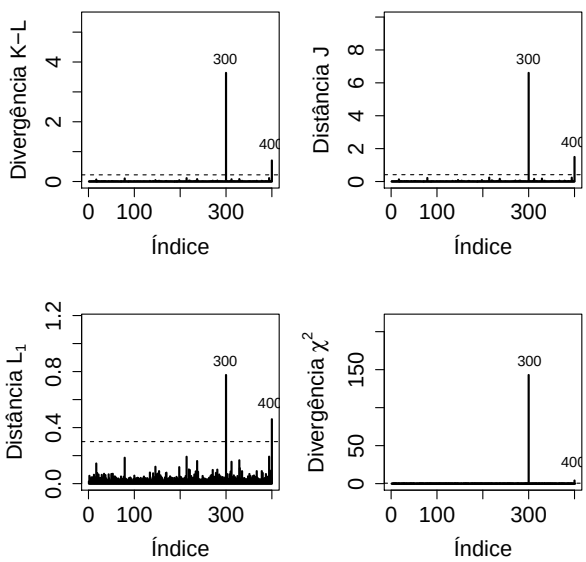

FIGURA 4: Gráficos de índices das medidas de divergência para o caso (g).

Ajustamos o modelo Clayton bivariado com ambas marginais Weibull ou Exponencial Generalizada, considerando duas cadeias de tamanho 65.000, desconsiderando as primeiras 10.000 iterações para eliminar o efeito dos valores iniciais e, para evitar problemas de autocorrelação, foi considerado um espaçamento de tamanho 10, obtendo uma amostra efetiva de tamanho 11.000 sobre a qual a inferência *a posteriori* é baseada. A convergência das cadeias foi monitorada de acordo com os métodos recomendados por [9]. Na Tabela 6 apresentamos os resumos *a posteriori* para os parâmetros do modelo Clayton bivariado

TABELA 4: Critérios Bayesianos ajustando o modelo de sobrevivência Clayton bivariado com marginais Weibull para cada conjunto de dados simulados.

Nome dos dados	Critérios Bayesianos			
	DIC	EAIC	EBIC	B
a	362.650	369.504	397.444	-0.444
b	376.624	383.720	411.660	-0.462
c	453.459	460.493	488.433	-0.558
d	422.874	429.785	457.725	-0.520
e	382.984	389.755	417.696	-0.470
f	464.639	471.396	499.336	-0.572
g	442.171	448.953	476.894	-0.544
h	520.539	527.270	555.210	-0.642

TABELA 5: Medidas de divergência para o modelo Clayton bivariado com marginais Weibull para cada conjunto de dados simulados.

Nome dos dados	Identificação dos casos	Medidas de divergência			
		K-L	J	L_1	χ^2
a	100	0.013	0.025	0.063	0.026
	200	0.010	0.017	0.052	0.018
	300	0.000	0.001	0.015	0.001
	400	0.007	0.016	0.050	0.016
b	100	1.355	3.108	0.625	26.571
c	200	11.921	27.793	1.411	14544.162
d	300	4.123	11.837	1.080	1379.935
e	400	0.821	1.762	0.490	7.134
f	100	0.337	0.794	0.333	3.297
	200	11.576	12.846	0.609	107.384
g	300	3.636	6.606	0.776	142.997
	400	0.705	1.494	0.460	4.116
h	100	0.417	0.942	0.371	3.192
	200	8.889	9.903	0.607	21.192
	300	2.786	5.992	0.804	117.299

para ambas as distribuições.

Nas Figuras 9 e 10 apresentamos os gráficos de índices considerando o modelo de Clayton com ambas marginais Weibull e Exponencial Generalizada, respectivamente. Em ambos os casos, podemos observar que todas as medidas detectam a observação 21 como possível ponto influente.

A Tabela 7 mostra que as quatro medidas de divergência que detectam a observação 21 como influente.

A Tabela 8 apresenta os critérios de comparação de modelos para comparar o modelo de sobrevivência bivariado baseado na cópula de Clayton com marginais Weibull e o modelo de sobrevivência bivariado baseado na cópula de Clayton com marginais Exponencial Generalizada. Como resultado, consideramos que o modelo com marginais Weibull obteve o melhor ajuste neste conjunto de dados levando em consideração que todos os critérios utilizados dão evidências positivas a favor deste modelo.

As Figuras 10 e 11 mostram, respectivamente, as curvas de Kaplan-Meier para as variáveis T_1 e T_2 dicotomizadas pelo sexo do paciente juntamente com os ajustes do modelo de sobrevivência bivariado baseado na cópula de Clayton com marginais Weibull.

TABELA 6: Média *a posteriori*, desvio padrão (DP) e intervalo de credibilidade (IC) de 95% para os parâmetros do modelo Clayton bivariado com marginais Weibull ou Exponencial Generalizada.

	Parâmetro	Weibull			Exponencial Generalizada		
		Média	DP	IC (95%)	Média	DP	IC (95%)
Tempo 1	α_1	0.964	0.129	(0.729, 1.233)	1.000	0.218	(0.616, 1.475)
	β_{01}	-3.281	0.590	(-4.482, -2.209)	-3.427	0.335	(-4.144, -2.841)
	β_{11}	-1.877	0.418	(-2.675, -1.028)	-1.937	0.371	(-2.643, -1.188)
Tempo 2	α_2	0.800	0.107	(0.601, 1.024)	0.776	0.167	(0.492, 1.149)
	β_{02}	-3.377	0.634	(-4.690, -2.204)	-4.629	0.365	(-5.423, -3.984)
	β_{12}	-0.540	0.361	(-1.220, 0.199)	-0.513	0.408	(-1.285, 0.340)
Cópula	ϕ	0.518	0.284	(0.134, 1.224)	0.489	0.279	(0.085, 1.162)

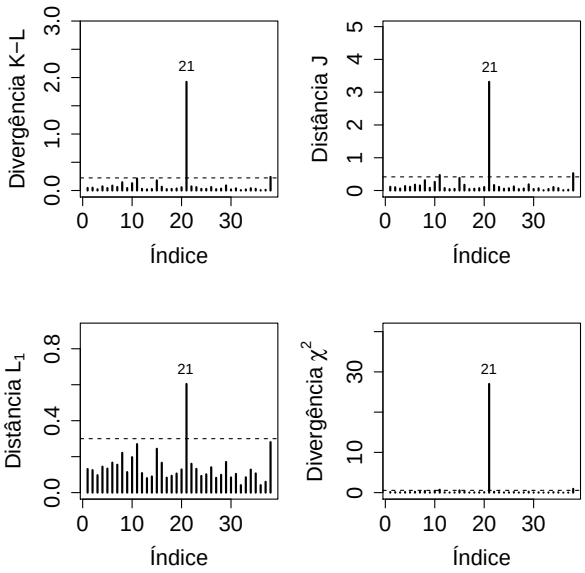

FIGURA 5: Gráficos de índices das medidas de divergência considerando a distribuição Weibull.

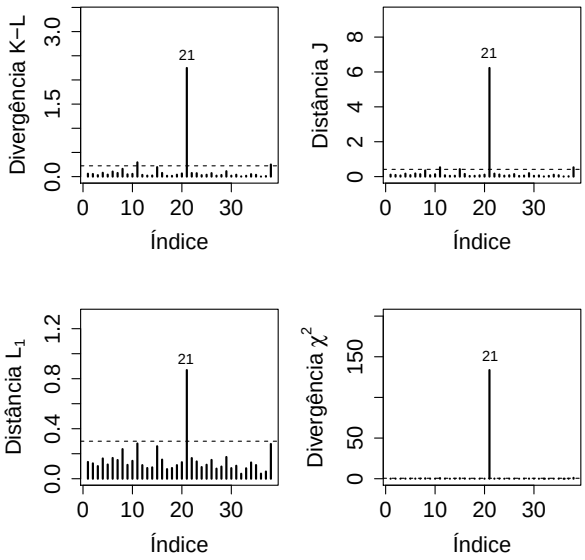

FIGURA 6: Gráficos de índices das medidas de divergência considerando a distribuição Exponencial Generalizada.

6 CONSIDERAÇÕES FINAIS

Neste trabalho apresentamos a modelagem de dados de sobrevivência por meio de cópulas Arquimedianas, em particular para a cópula de Clayton. Todo o procedimento inferencial foi realizado sob uma abordagem Bayesiana assumindo ausência de informação *a priori*. Como aplicação dos modelos estudados realizamos um amplo estudo de simulação no qual verificamos que, com diferentes tamanhos amostrais e diferentes configurações de censura, as estimativas obtidas foram próximas do verdadeiro valor.

Avaliamos a robustez do modelo relacionado às escolhas dos hiperparâmetros das distribuições a *priori*, realizando um estudo de sensibilidade no qual constatamos que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas nos resultados das aplicações aos dados artificiais e aos dados reais.

Verificamos que as quatros medidas de divergência detectam os pontos influentes do modelo com um certo padrão de perturbação que foi adotado no trabalho.

Como estudos futuros estamos interessados em utilizar outras distribuições marginais, tais como o modelo de Cox e o modelo Exponencial por partes, como também os modelos de sobrevivência com fração de cura.

TABELA 7: Medidas de convergência para as marginais Weibull e Exponencial Generalizada.

Nome do dado	Distribuição	Medidas de divergência			
		K-L	J	L_1	χ^2
Observação 21	Weibull	1.923	3.315	0.604	26.996
	Exponencial Generalizada	2.248	6.229	0.869	133.623

TABELA 8: Critérios para comparação dos modelos.

Modelo	Critérios Bayesianos			
	DIC	EAIC	EBIC	B
Weibull	725.192	737.731	749.194	-9.631
Exponencial Generalizada	725.865	739.027	750.490	-9.661

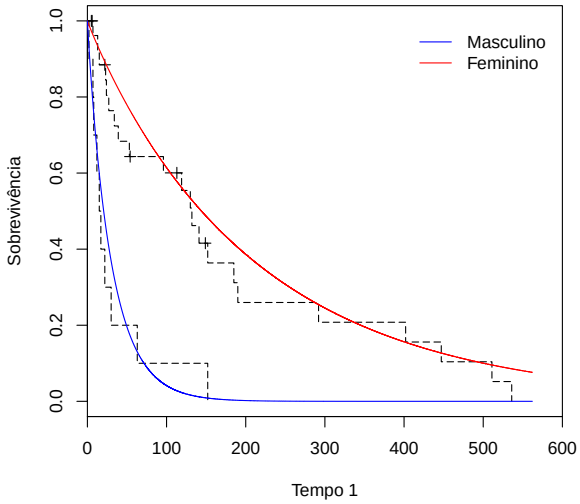

FIGURA 7: Curvas de Kaplan-Meier e curvas de sobrevivências Weibull estimadas para a variável T_1.

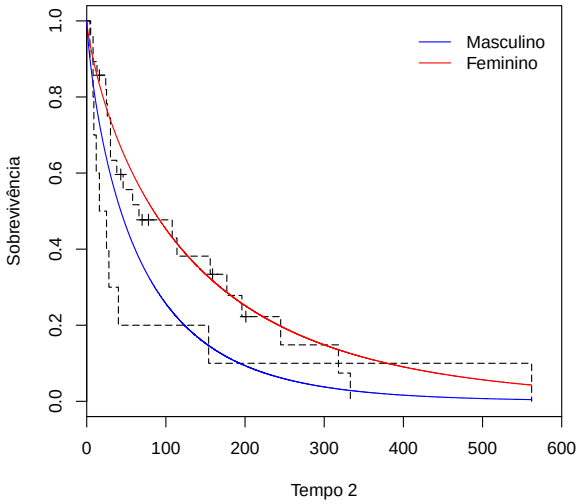

FIGURA 8: Curvas de Kaplan-Meier e curvas de sobrevivências Weibull estimadas para a variável T_2.

REFERÊNCIAS

[1] BOLETA, J.; ACHCAR, J. A. *Distribuição Exponencial Generalizada bivariada derivada de funções cópulas: Uma aplicação a dados de câncer gástrico*. Revista Brasileira de Biometria, v.30 (4), p.401-414, 2012.

[2] BOLETA, J. *Distribuição Exponencial Generalizada: uma análise bayesiana aplicada a dados de câncer*. 93 f. Dissertação (Mestrado) - Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2012.

[3] BROOKS, S. P. *Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde*, v.64, 616-618, 2002

[4] CANCHO, V.; ORTEGA, E.; PAULA, G. *On estimation and influence diagnostics for log-Birnbaum-Saunders Student-t regression models: Full Bayesian analysis*. Journal of Statistical Planning and Inference, v.140, p.2486-2496, 2010.

[5] CARLIN, B. P.; LOUIS, T. A. *Bayes and Empirical Bayes Methods for Data Analysis*. 2.ed. Boca Raton: Chapman and Hall, 2001.

[6] CHO, H.; IBRAHIM, J. G.; SINHA, D.; SHU, H. *Bayesian case influence diagnostics for survival models*. Biometrics, v.65, p.116-124, 2009.

[7] CLAYTON, D. G. *A model for association in bivariate life-tables and its application in epidemiological studies of familial tendency in chronic disease incidence*. Biometrika, v.65, p.141-151, 1978.

[8] COOK, R. D.; WEISBERG, S. *Residuals and Influence in Regression*. Boca Raton: Chapman and Hall, 1982.

[9] COWLESS, M. K.; CARLIN, B. P. *Markov chain Monte Carlo convergence diagnostics: a comparative review*. Journal of the American Statistical Association, v.91, p.883-904, 1996.

[10] DENWOOD M. J.; TUKALOV A.; PLUMMER M. *Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS*. Journal of Statistical Software, 2015.

[11] DEY, D.; BIRMIWAL, L. *Robust Bayesian analysis using divergence measures*. Statistics and Probability Letters, v.20, p.287-294, 1994.

[12] EMBRECHTS, P.; LINSKOG, F.; MCNIEL, A. *Modelling dependence with copulas and applications to risks management*. http://www.math.ethz.ch/baltes/ftp/papers.html, 2003.

[13] FREES, E., ANDWANG, P. *Credibility using copulas*. North American Actuarial Journal 9, 2005.

[14] GELMAN, A.; RUBIN, D. B. *Inference from iterative simulation using multiple sequences*. Statistical Science, v.7, p.457-511, 1992.

[15] GENEST, C. *Frank's family of bivariate distributions*. Biometrika, v.74, p.549-555, 1987.

[16] GUPTA, R. D.; KUNDU, D. *Generalized Exponential distributions*. Aust. N.Z. J. Stat., Oxford, v.41, p.173-188, 1999.

[17] IBRAHIM, J. G.; CHEN, M-H.; SINHA, D. (2001). *Bayesian Survival Analysis*. New York: Springer-Verlag, 2001.

[18] JOE, H. *Multivariate Models and Dependence Concepts*. London: Chapman and Hall, 1997.

[19] JOE, H. *Dependence Modeling with Copulas.* London: Chapman and Hall, 2014.

[20] KOLEV, N.; DOS ANJOS, U.; MENDES, B. V. M. *Copulas: A review and recent developments*. Stochastic Models, v.22 (4), p.617-660, 2006.

[21] LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. New York: Wiley and Sons, 2003.

[22] LOUZADA, F.; SUZUKI, A. K.; CANCHO, V. G.; PRINCE F. L.; PEREIRA, G. A. *The Long-Term Bivariate Survival FGM Copula Model: An Application to a Brazilian HIV Data*. Journal of Data Science, v.10, p.511-535, 2010.

[23] LOUZADA, F.; SUZUKI, A. K.; CANCHO, V. G. *The FGM Long-Term Bivariate Survival Copula Model: Model, Bayesian Estimation, and Case Influence Diagnostics*. Communications in Statistics - Theory and Methods, v.42 (4), p.673-691, 2013.

[24] MCGILCHRIST C. A.; AISBETT C. W. *Regression with frailty is survival analysis*. Biometrics, v.47, p.461-466, 1991.

[25] NELSEN, R. *Properties of a one-parametric family of bivariate distributions with specified marginals*. Communications in Statistics, v.15, p.3277-3285, 1986.

[26] NELSEN, R. *An Introduction to Copulas*. 2.ed. New York: Springer, 2006.

[27] PENG, F.; DEY, D. *Bayesian analysis of outlier problems using divergence measures*. The Canadian Journal of Statistics - La Revue Canadienne de Statistique, v.23, p.199-213, 1995.

[28] PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K.*Output analysis and diagnostics for MCMC*. http://cran.r-project.org/web/packages/coda/index.html, 2006.

[29] QUEIROS FLORES, A. *Copula functions and bivariate distributions for survival analysis: An application to political survival, 2008.*

[30] R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org, 2007.

[31] SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; VAN DER LINDE, A. *Bayesian measures of model complexity and fit*. Journal of the Royal Statistical Society Series B, v.64, p.583-639, 2002.

[32] SUZUKI, A. K.; LOUZADA-NETO, F.; CANCHO, V. G.; BARRIGA, G. D. C. *The FGM bivariate lifetime copula model: a bayesian approach*. Advances and Applications in Statistics, v.21 (1), p.55-76, 2011.

[33] SUZUKI, A. K. *Modelos de sobrevivência bivariados baseados na cópula FGM : Uma abordagem bayesiana*, 2012. Tese de Doutorado - Universidade Federal de São Carlos, São Carlos, 102 p., 2012.

[34] VAUPEL, J. W., MANTON, K. G. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

[35] VIDAL, I.; CASTRO, L. M. *Influential observations in the independent Student-t measurement error model with weak nondifferential error*. Chilean Journal of Statistics, v.1, p.17-34, 2010.

[36] VIOLA, M. L. L. *Tipos de Dependência entre Variáveis Aleatórias e Teoria de Cópulas.*, 2009.

[37] WEIBULL W. *A statistical theory of the strength of material*. Proc. Roy. Swedish Inst. Eng. Res. 151:1, 1939.

[38] WEISS, R. *An approach to Bayesian sensitivity analysis*. Journal of the Royal Statistical Society Series B, p.739-750, 1996.