UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

**Distribuição exponencial generalizada:
uma análise bayesiana aplicada a dados de câncer**

JULIANA BOLETA

Ribeirão Preto

2012

JULIANA BOLETA

Distribuição exponencial generalizada: uma análise bayesiana aplicada a dados de câncer

Dissertação apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Mestre em Saúde na Comunidade.

VERSÃO CORRIGIDA

A versão original encontra-se na Biblioteca do Departamento de Medicina Social-FMRP-USP.

Área de Concentração: Saúde na Comunidade

Orientador: Prof. Dr. Jorge Alberto Achcar

Ribeirão Preto

2012

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Ficha Catalográfica

Boleta, Juliana

Distribuição exponencial generalizada: uma análise bayesiana aplicada a dados de câncer

93 p. :il. ; 30cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Saúde na Comunidade.

Orientador: Jorge Alberto Achcar

1. Distribuição exponencial generalizada. 2. Funções cópula.
3. Análise de sobrevivência. 4. Modelos bayesianos. 5. Leucemia
Mielóide aguda. 6. Câncer gástrico.

Folha de Aprovação

Juliana Boleta

Distribuição exponencial generalizada: uma análise bayesiana aplicada a dados de câncer

Dissertação apresentada ao Departamento de Medicina Social da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Mestre em Saúde na Comunidade.

Área de Concentração: Saúde na Comunidade

Aprovado em: ____/____/____

Banca Examinadora

Prof.(a) Dr.(a): _____

Instituição: _____ Assinatura: _____

Prof.(a) Dr.(a): _____

Instituição: _____ Assinatura: _____

Prof.(a) Dr.(a): _____

Instituição: _____ Assinatura: _____

Dedicatória

Aos meus pais, Sebastião e Janete, pelo apoio nesta caminhada e por jamais me deixar desistir.

Agradecimentos

Agradeço primeiramente à Deus cuja presença em minha vida tornou os meus sonhos possíveis, sem Ele não seria quem sou.

Agradeço aos meus pais, Sebastião e Janete, pelo imenso amor demonstrado em cada gesto, pela compreensão, por se privar de poder me ver para que eu possa cumprir com as minhas responsabilidades e mais que tudo por me ensinar a encarar os problemas de frente, sem desistir.

Agradeço às minhas irmãs, Tatiane e Viviane pelo amor e amizade que sempre me dedicaram. Vocês são os anjinhos da minha vida. Afinal, somos mais que irmãs, somos amigas.

Agradeço aos meus avós e a minha família, por todas as orações feitas, por sempre estar torcendo pelos meus estudos.

Agradeço também ao meu namorado, Julio e sua família, que em todo este tempo esteve comigo, nos momentos bons e ruins, me dando forças, me respeitando e acima de tudo me ajudando a passar por todo este período.

Agradeço ao meu orientador, Jorge, por estar me ajudando em todo este trabalho, pela dedicação e principalmente pela paciência. Obrigada pela ótima orientação e por ter acreditado no meu trabalho.

À todos meus amigos que de alguma forma contribuíram para a concretização deste trabalho. Em especial, as minhas grandes amigas Mayara, Estela e Vanessa, que não tenho

nem palavras pra descrever o quanto as estimo. Obrigada pela força e pelas ajudas que sempre me deram.

Agradeço aos meus grandes amigos do Bradesco Financiamentos, por me compreender e por toda ajuda. Agradeço, em especial, à minha grande amiga e companheira Malúbia, por estar ao meu lado e principalmente por trazer esta alegria contagiante.

Aos grandes companheiros que tive no CEMEQ e ao prof. Edson, por compartilhar de forma honrosa todo o seu conhecimento.

Agradeço a banca, pelas correções e pelas sugestões para este trabalho. Suas contribuições são sempre muito importantes.

Agradeço ao amigo, prof. Mário, que tem acompanhado a minha vida acadêmica e profissional ao longo de todos esses anos.

Agradeço a todos os colegas, funcionários e professores do Departamento de Medicina Social e do Programa de pós graduação em Saúde na Comunidade pelo companheirismo.

Resumo

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

A técnica de análise de sobrevivência tem sido muito utilizada por pesquisadores na área de saúde. Neste trabalho foi usada uma distribuição em análise de sobrevivência recentemente estudada, chamada distribuição exponencial generalizada. Esta distribuição foi estudada sob todos os aspectos: para dados completos e censurados, sob a presença de covariáveis e considerando sua extensão para um modelo multivariado derivado de uma função cópula. Para exemplificação desta nova distribuição, foram utilizados dados reais de câncer (leucemia mielóide aguda e câncer gástrico) que possuem a presença de censuras e covariáveis. Os dados referentes ao câncer gástrico tem a particularidade de apresentar dois tempos de sobrevida, um relativo ao tempo global de sobrevida e o outro relativo ao tempo de sobrevida livre do evento, que foi utilizado para a aplicação do modelo multivariado. Foi realizada uma comparação com outras distribuições já utilizadas em análise de sobrevivência, como a distribuição Weibull e a Gama. Para a análise bayesiana adotamos diferentes distribuições a priori para os parâmetros. Foi utilizado, nas aplicações, métodos de simulação de MCMC (Monte Carlo em Cadeias de Markov) e o software Winbugs.

Palavras - chave: Distribuição exponencial generalizada. Funções cópula. Análise de sobrevivência. Modelos bayesianos. Leucemia mielóide aguda. Câncer gástrico.

Abstract

BOLETA, J. **Generalized exponential distribution: a Bayesian analysis applied to cancer data.** 2012. 93 f. Dissertation (master degree) - Faculty of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, 2012.

Survival analysis methods has been extensively used by health researchers. In this work it was proposed the use a survival analysis model recently studied, denoted as generalized exponential distribution. This distribution was studied in all respects: for complete data and censored, in the presence of covariates and considering its extension to a multivariate model derived from a copula function. To exemplify the use of these models, it was considered real cancer lifetime data (acute myeloid leukemia and gastric cancer) in presence of censored data and covariates. The assumed cancer gastric lifetime data has two survival responses, one related to the total lifetime of the patient and another one related to the time free of the disease, that is, multivariate data associated to each patient. In these applications there was considered a comparative study with standard existing lifetime distributions, as Weibull and gamma distributions. For a Bayesian analysis we assumed different prior distributions for the parameters of the model. For the simulation of samples of the joint posterior distribution of interest, we used standard MCMC (Markov Chain Monte Carlo) methods and the software Winbugs.

Keywords: Generalized exponential distribution. Copula functions. Survival analysis. Bayesian models. Acute myeloid leukemia. Gastric cancer.

Lista de Figuras

1	Diagrama de dispersão para a variável GB($\times 10^9/L$) versus tempo de sobrevida (escala logarítmica)	18
2	Diagrama de dispersão para a variável Uréia (mg/dL) versus tempo de sobrevida (escala logarítmica)	18
3	Curva de sobrevida estimada para os dados de sobrevida da leucemia mielóide aguda	19
4	Curva de sobrevida estimada para a variável TSG, segundo o tratamento .	21
5	Curva de sobrevida estimada para a variável TSL, segundo o tratamento .	22
6	Gráfico da função densidade para vários valores de alfa ($\lambda=1$)	37
7	Gráfico da função de risco para vários valores de alfa ($\lambda=1$)	38
8	Gráfico da função de sobrevivência empírica e ajustadas pelas distribuições exponencial generalizada, gama e Weibull	59
9	Gráfico da função de sobrevivência empírica para o TSG e ajustadas pelas distribuições exponencial generalizada e Weibull	62
10	Gráfico da função de sobrevivência empírica para o TSL e ajustadas pelas distribuições exponencial generalizada e Weibull	62

Lista de Tabelas

1	Mortalidade proporcional não ajustada por câncer, Brasil, homens e mulheres, entre 1997 e 2007	14
2	Média dos tempos de sobrevida em relação ao tratamento	20
3	Sumários a posteriori	60
4	DIC	60
5	Sumários a posteriori para o TSG - Média (IC-95%)	63
6	Sumários a posteriori para o TSG	63
7	Sumários a posteriori para o TSL - Média (IC-95%)	63
8	DIC	64
9	Sumários a posteriori para a Distribuição Exponencial Generalizada	65
10	Sumários a posteriori para a Distribuição Weibull	65
11	DIC	65

Sumário

1	Introdução	13
1.1	Motivação: Dois conjuntos de Dados de Sobrevida associados ao Câncer	16
1.1.1	Leucemia Mielóide Aguda	17
1.1.2	Câncer gástrico	19
2	Uma breve revisão sobre Análise de Sobrevida	23
2.1	Função de Sobrevida	24
2.2	Técnicas não paramétricas	24
2.3	Técnicas paramétricas	25
2.3.1	Distribuição Exponencial	26
2.3.2	Distribuição Weibull	26
2.3.3	Distribuição Gama	27
2.4	Método de Máxima Verossimilhança	28
3	Uma breve revisão da metodologia bayesiana	31
3.1	Fórmula de Bayes	31
3.2	Distribuição a Priori	32
3.3	Um Critério para Discriminação de Modelos - DIC	33
3.4	Métodos de simulação para amostras da distribuição a posteriori	33
3.4.1	Amostrador de Gibbs	34
3.4.2	Algoritmo de Metropolis Hastings	35
4	Distribuição exponencial generalizada	36
4.1	Função Densidade	36
4.2	Função de Sobrevida e de Risco	36
4.3	Função Geradora de Momentos	37
4.4	Função de Verossimilhança para dados completos	38
4.5	Uma análise bayesiana	40
4.6	Presença de dados censurados	44
4.7	Presença de Covariáveis	45
5	Distribuição Multivariada derivadas de funções de cópula	48
5.1	Funções Cópula	48
5.2	Distribuição Exponencial Generalizada bivariada derivada da Cópula de Farlie-Gumbel-Morgenstern	50

5.3	Análise bayesiana na presença de dados censurados	53
5.4	Análise bayesiana na presença de covariáveis	56
6	Aplicações	58
6.1	Aplicação1: Leucemia Mielóide Aguda	58
6.2	Aplicação 2: Dados de Câncer Gástrico	60
6.2.1	Aplicação do Modelo Univariado	61
6.2.2	Aplicação do Modelo Multivariado usando Função Cópula	64
7	Considerações Finais	67
	Referências	69
	APÊNDICE A - Programas	74
A.1	- Modelo Univariado para os dados de Leucemia Mielóide Aguda	74
A.1.1	- Modelo 1	74
A.2	- Modelo Univariado para os dados de Câncer Gástrico	75
A.2.1	- Modelo 2: TSG	75
A.2.2	- Modelo 3: TSL	76
A.3	- Modelo Multivariado para os dados de Câncer Gástrico	77
A.3.1	- Modelo 4	77
	APÊNDICE D - Artigo Publicado: Revista Brasileira de Biometria	78
	ANEXO A - Comitê de Ética em Pesquisa	93

1 Introdução

O número de casos de câncer tem aumentado de maneira considerável em todo o mundo, tornando-se, atualmente, um dos mais importantes problemas de saúde pública (GUERRA; GALLO; MENDONÇA, 2005).

Em sua definição, o câncer é designado como “um grupo de doenças que tem como característica a desordenada multiplicação de células malignas”, ou seja, “células normais que sofreram alterações em seu DNA”. Quando essas células se desenvolvem de forma agressiva e incontrolável, formam-se os tumores malignos (neoplasias malignas), e quando invadem células vizinhas e distantes, caracteriza-se a metástase” (BORGES et al., 2007).

No ano de 1919, na cidade do Rio de Janeiro, o câncer ocupava o sexto lugar na causa de morte da população, sendo superado por algumas doenças como a tuberculose, gastroenterites, doenças cardiovasculares, pneumonias e nefrites (WUNSCH FILHO; MONCAU, 2002).

No Brasil, em 2001, considerando-se pessoas do sexo masculino, constatou-se que as principais mortes por câncer se deram pelos tumores de pulmão, próstata, estômago, esôfago, boca e faringe. Já no que concerne às pessoas do sexo feminino, destacou-se os tumores de mama, pulmão, cólon e reto, colo de útero e estômago (GUERRA; GALLO; MENDONÇA, 2005).

Um pouco adiante, em 2005, o câncer de estômago foi considerado como a segunda causa de morte, tanto na população masculina como na feminina. Ressalta-se que, contabilizando apenas pessoas do sexo masculino, o tipo acima referenciado correspondeu a 12,7% de todas as mortes decorrentes do câncer (WUNSCH FILHO; MONCAU, 2002).

Na Tabela 1, pode-se observar as proporções de mortes devido ao câncer (em geral) no Brasil, no período de 1997 a 2007. Segundo o Instituto Nacional do Câncer (BRASIL, 2003), as neoplasias vêm ganhando cada vez maior importância no perfil de mortalidade do país, ocupando o segundo lugar como causa de óbito.

É importante salientar que, entre os vários tipos de câncer, encontram-se alguns muito

mais agressivos, que encurtam o tempo de sobrevida dos pacientes.

A leucemia, por exemplo, é uma doença muito comum na infância, correspondendo a aproximadamente 30% dos casos de câncer em crianças. São registrados no Brasil, cerca de 7400 casos por ano, totalizando 5500 mortes anuais (BRASIL, 2002). Em virtude de tais fatores, esta doença se apresenta como o 8º tipo de câncer mais frequente na população brasileira (BRASIL, 2002).

Tabela 1 – Mortalidade proporcional não ajustada por câncer, Brasil, homens e mulheres, entre 1997 e 2007

Ano	Porcentagem
1997	11,6
1998	11,6
1999	11,9
2000	12,3
2001	12,6
2002	12,8
2003	13
2004	13,4
2005	14,2
2006	14,5
2007	14,8

Fonte: BRASIL (<http://mortalidade.inca.gov.br/Mortalidade/>)

Um dos tipos de leucemia considerado bastante agressivo é a Leucemia Mielóide Aguda (LMA). Isso porque, esta é uma “doença clonal do tecido hematopoético que se caracteriza pela proliferação anormal de células progenitoras da linhagem mielóide, ocasionando produção insuficiente de células sanguíneas maduras normais” (MARTINS; FALCÃO, 2000).

A faixa etária acometida pela LMA se encontra entre 15 a 20% na infância e cerca de 80% nos adultos. Principalmente na infância, este tipo de leucemia é uma das doenças que possui menor taxa de sobrevida, variando entre 30 e 40% (MENDONÇA, 2003).

Outro tipo de câncer bastante agressivo é o gastrointestinal. Entre os tumores gastrointestinais incluem-se aqueles no cólon (intestino grosso), reto, estômago, pâncreas, esôfago, intestino delgado, ânus e outros órgãos do sistema digestivo.

No mundo, no que se refere à frequência, o câncer de estômago está em segundo lugar

em relação aos outros tipos desta doença, ficando atrás apenas do câncer de pulmão (PISANI; PARKIN; FERLAY apud AZEVEDO; MENDONÇA, 1997). O câncer gástrico é responsável por aproximadamente 12,1% das mortes, e sua incidência é mais acentuada em algumas regiões da China, Ásia e América do Sul (MAGALHÃES et al., 2008). Esse tipo de câncer representa grande parte das causas de óbito no Brasil, estando posicionado entre as cinco localizações mais comuns de óbitos. A taxa de sobrevida geral em cinco anos é de cerca de 10%, no entanto, quando detectado o tumor no início, a sobrevida passa a estar entre 25 e 40% (TONETO, 2006).

Alguns autores associam a ocorrência do câncer gástrico a fatores intrínsecos (genética), e extrínsecos, que são aqueles relacionados ao tipo de alimentação (ver, por exemplo, TEIXEIRA; NOGUEIRA, 2003; TONETO, 2006).

Diante do acima exposto, consigna-se que os coeficientes mundiais de mortalidade relativos ao câncer ainda são altos, apesar de apresentarem tendência de declínio. Neste sentido, cada nova possibilidade terapêutica traz a esperança de maior tempo de sobrevida e de melhor qualidade de vida aos pacientes. Atualmente, o aumento do número de pessoas com câncer tem se constituído como um desafio para os serviços de saúde e para a sociedade.

Com a procura de novas terapêuticas para aumentar a sobrevida dos pacientes, as técnicas de análise de sobrevivência têm sido muito utilizadas na área médica, principalmente na área de oncologia, para se avaliar a sobrevida dos pacientes após uma cirurgia, ou após uma quimioterapia (ver, por exemplo, BUSTAMANTE-TEIXEIRA; FAERSTEIN; LATORRE, 2002; MENDONÇA; SILVA; CAULA, 2004; TONETO, 2006).

Alguns modelos estatísticos são muito usados nesta área para estudar os tempos de sobrevivência, tais como, a distribuição Weibull, a distribuição exponencial e a distribuição gama, na presença ou não de covariáveis e dados censurados. Atualmente, vários modelos novos estão sendo propostos na literatura para analisar dados de sobrevida, sendo que, com a complexidade dos dados na área médica, é sempre importante ter uma variedade de opções a se utilizar, para se obter melhores inferências e previsões.

Recentemente foi introduzida a distribuição exponencial generalizada (GUPTA; KUNDU, 1999; RAQAB, 2002; RAQAB; AHSABULLAH, 2001; ZHENG, 2002; SARHAN, 2007). Essa distribuição apresenta algumas similaridades com a distribuição gama em termos de comportamento da função de risco, mas com grandes vantagens na obtenção de inferências de interesse quando os dados de sobrevivência apresentam censuras, um fato comum em dados médicos e industriais. Tendo em vista que a função de sobrevivência para a distribuição gama é dada por uma função gama incompleta, observa-se que, em regra, isso dificulta a obtenção de inferências de interesse, o que não ocorre com a distribuição exponencial generalizada.

Tendo em vista a motivação acima explanada, o presente trabalho visa explorar o uso da distribuição exponencial generalizada, utilizando dados reais de câncer e, ao mesmo tempo, verificar a possibilidade de existência de uma melhoria deste modelo ajustado quando comparado a outros modelos tradicionais (Weibull e gama). Para isso, serão consideradas algumas metodologias estatísticas existentes na literatura a fim de analisar dados de sobrevivência sob o enfoque paramétrico e não-paramétrico (ver, por exemplo, KLEIN; MOESCHBERGER, 1997). Além disso, alguns modelos usuais de regressão para a análise de dados de sobrevivência na presença de censuras também serão considerados.

As inferências para esses modelos de regressão com dados de sobrevivência na presença de censuras serão obtidas usando-se métodos de inferência bayesiana (ver, por exemplo, BOX E TIAO, 1973).

1.1 Motivação: Dois conjuntos de Dados de Sobrevivência associados ao Câncer

Neste trabalho serão utilizados dois conjuntos de dados de câncer. O primeiro deles é relativo à Leucemia Mielóide Aguda e o outro ao Câncer Gástrico (câncer de estômago). A seguir, serão descritas as principais características de cada banco de dados. Este trabalho teve aprovação no comitê de Ética em Pesquisa do HCRP e da FMRP-USP ("Anexo B").

1.1.1 Leucemia Mielóide Aguda

O banco de dados relativo à Leucemia Mielóide Aguda foi coletado no Hospital das Clínicas de Ribeirão Preto - FMRP/USP. Os dados foram cedidos pela pesquisadora Luciana Correa Oliveira de Oliveira para que fossem utilizados nesta dissertação de mestrado. A pesquisadora possui um artigo publicado no Medical Oncology (OLIVEIRA et al., 2009).

Neste banco de dados foram revistos 189 prontuários de pacientes com Leucemia Mielóide Aguda entre Janeiro de 1998 e Junho de 2008. Esses dados apresentam algumas observações censuradas (censura à direita) e neste exemplo será considerada a presença de duas covariáveis, GBx1000 ($\times 10^9/L$) e Uréia (mg/dL).

A covariável GBx1000 (glóbulos brancos) são células que possuem a função de defesa do organismo. São compostas de cinco diferentes células: neutrófilos, eosinófitos, linfócitos, monócitos e basófilos (PEREIRA, 2011). E a covariável Uréia é um resíduo de excreção, proveniente do metabolismo de compostos nitrogenados tipo proteínas e aminoácidos.

Do total de 189 pacientes, 166 tiveram a sobrevida completa observada e 23 foram censurados. A média de sobrevida geral foi de aproximadamente 432 dias. O menor tempo de sobrevida foi de 1 dia e o maior foi de 2627 dias. A covariável GB possui uma média de $35,36 \times 10^9/L$ (desvio padrão igual a 62,91) e a covariável Uréia apresentou uma média de 40,17 mg/dL (desvio padrão igual a 31,88).

Na Figura 1, pode-se observar que para os maiores tempos de sobrevida a covariável GB apresenta menores valores. Isto também ocorre com relação à variável Uréia (Figura 2).

Figura 1 – Diagrama de dispersão para a variável GB($\times 10^9/L$) versus tempo de sobrevivência (escala logarítmica)

Figura 2 – Diagrama de dispersão para a variável Uréia (mg/dL) versus tempo de sobrevivência (escala logarítmica)

Na Figura 3 é representado o gráfico da curva de sobrevivência dos pacientes, estimado pelo estimador não-paramétrico de Kaplan e Meier (1958). Observa-se que uma pequena proporção de pacientes (em torno de 20%) sobrevive mais de 500 dias.

Figura 3 – Curva de sobrevida estimada para os dados de sobrevida da leucemia mielóide aguda

No capítulo 6, este banco de dados será utilizado para exemplificar o uso da distribuição exponencial generalizada, na análise de dados de sobrevivência na presença de censuras e covariáveis.

1.1.2 Câncer gástrico

O banco de dados relativo a pacientes com câncer gástrico foi coletado no Hospital de Câncer de Barretos pelo pesquisador Alexandre Andrade dos Anjos Jácome (JÁCOME et al., 2012). O pesquisador cedeu os dados para que fossem utilizados nesta dissertação de mestrado.

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) 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.

O pesquisador definiu algumas covariáveis a serem consideradas no modelo, tais como sexo (masculino e feminino), idade, tipos de tratamentos (controle e quimioterapia), finalidade (R0-Cirurgia curativa, R1-Cirurgia com doença residual microscópica e R2-Cirurgia com doença residual macroscópica), linfadenectomia (D0, D1, D2, D3 e D1 plus), localização (antro, corpo, fundo, dois ou mais segmentos, transição esofagogástrica e coto gástrico) e TNM (estádio 0, estágio IA, estágio IB, estágio II, estágio IIIA, estágio IIIB, estágio IV M0 e estágio IV M1).

Esse banco de dados é composto por 84 pacientes do sexo feminino e 146 pacientes do sexo masculino, possuindo no geral, uma média de idade de aproximadamente 60 anos.

O TSG apresenta 73 observações completas e 157 observações censuradas e sua média de sobrevida é de aproximadamente 26,15 meses. O TSL possui 93 observações completas e 137 censuradas tendo uma média de sobrevida de 26,29 meses.

A idade média do grupo que recebeu a quimioterapia foi de 56 anos (desvio padrão igual a 11,26) e o grupo controle tem uma idade média de 62 anos (desvio padrão igual a 13,35).

A Tabela 2 apresenta a média de sobrevida para o TSL e o TSG com relação ao tratamento que os pacientes receberam.

Tabela 2 – Média dos tempos de sobrevida em relação ao tratamento

Variáveis	Quimioterapia	Controle
TSL	27,02	25,68
TSG	27,80	26,50

Observa-se (Tabela 2) que os tempos de sobrevida do grupo que recebeu a quimioterapia é maior daquela referente ao grupo que não recebeu tratamento.

A fim de visualizar o comportamento dos tempos de sobrevida considerando os dois tempos (TSG e TSL), tem-se nas Figuras 4 e 5 os gráficos dos estimadores produto-limite de Kaplan e Meier para as duas variáveis e, tendo sido também considerados os dois tratamentos (controle e quimioterapia). Pode-se observar, a partir desses gráficos, que não há uma diferença evidente entre os tratamentos considerando as duas variáveis TSG e TSL.

Figura 4 – Curva de sobrevida estimada para a variável TSG, segundo o tratamento

Figura 5 – Curva de sobrevida estimada para a variável TSL, segundo o tratamento

Este segundo banco de dados também será analisado considerando a distribuição exponencial generalizada para uma análise multivariada derivadas de funções de cópula.

2 Uma breve revisão sobre Análise de Sobrevivência

A análise de sobrevivência é uma técnica estatística aplicada a determinado evento relacionado a um período de tempo, isto é, ao tempo transcorrido entre o evento inicial, no qual o indivíduo entra em um estado particular e o evento final, onde este estado é modificado. Define-se sobrevida como o intervalo de tempo que compreende desde a entrada do indivíduo no estudo (por exemplo, data do diagnóstico) até a ocorrência do evento de interesse (falha, recuperação ou óbito), ou até a censura, que pode ser a saída do paciente de um estudo, ou, ainda, perda de contato. Portanto, nessa área, o tempo de falha se refere ao lapso temporal existente até a ocorrência de um evento de interesse, podendo este ser a morte, o aparecimento de um tumor, o desenvolvimento de uma doença, entre outros.

Neste contexto, considera-se que esta análise estatística tem como diferencial a capacidade de tratar dados censurados, isto é, indivíduos que apresentam apenas observação parcial da resposta. Tal fato se refere às situações em que por alguma razão houve a perda de seguimento durante o estudo, isto é, o acompanhamento do paciente foi interrompido, seja porque o paciente mudou de cidade ou porque o paciente morreu por uma causa divergente da ora estudada (ver, por exemplo, COLOSIMO; GIOLO, 2006). Sem a presença de censuras, as técnicas estatísticas clássicas, tais como a análise de regressão, poderiam ser utilizadas na análise desses dados.

Os dados censurados podem ser classificados de diferentes formas, quais sejam, censuras do Tipo I, censuras do Tipo II e a censuras aleatórias. A censura do Tipo I é aquela em que o estudo é terminado em um período pré-estabelecido de tempo; a censura do Tipo II é quando o estudo é terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos; e a censura aleatória, sendo a mais frequente na prática, é quando um indivíduo sai do estudo por causas aleatórias. Além destas, podem ser mencionadas as censuras à direita, à esquerda e intervalar. A censura à direita se dá quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado, a censura à esquerda acontece quando o tempo registrado é maior do que o tempo de falha

e, por fim, a censura intervalar é quando o evento de interesse acontece em um intervalo de tempo, no entanto não se sabe o momento exato em que aconteceu.

2.1 Função de Sobrevivência

Seja o tempo de falha denotado por T , uma variável aleatória não negativa, caracterizada pela função de sobrevivência dada por,

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{\infty} f(t)dt \quad (1)$$

que é a probabilidade do indivíduo sobreviver até o tempo t e $F(t)$ é a função de distribuição acumulada da variável aleatória T . A função $f(t)$ na equação (1) é a função densidade de probabilidade da variável aleatória T .

A função de taxa de falha de T é definida por,

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (2)$$

Essa distribuição é útil para descrever a distribuição do tempo de vida das observações em estudo.

Observar que a partir de (1) e (2) temos a relação $\lambda(t) = f(t)/S(t)$.

2.2 Técnicas não paramétricas

As técnicas não paramétricas são muito utilizadas em artigos científicos publicados na área médica. O estimador produto limite de Kaplan e Meier (1958) é um estimador não paramétrico da função de sobrevivência de dados de vida e permite a inclusão dos dados censurados.

O referenciado estimador trata-se de uma adaptação da função de sobrevivência empírica, que na ausência de censuras é definido pela equação (3)

$$\hat{S}(t) = \frac{(\# \text{observações} > t)}{n} \quad (3)$$

onde $\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $\frac{1}{n}$, em que n é o tamanho da amostra, caso não haja empates.

Sejam $t_{(1)} \leq t_{(2)} \leq \dots \leq t_{(r)}$ os r tempos de falhas observadas ordenadas ($r \leq n$). O estimador Produto Limite de $S(t)$ é definido como,

$$\hat{S}(t) = \prod_{j:t_j < t} \left[\frac{n_j - d_j}{n_j} \right] = \prod_{j:t_j < t} \left(1 - \frac{d_j}{n_j} \right) \quad (4)$$

A variância assintótica para o estimador de Kaplan-Meier é dados por

$$\hat{Var}(\hat{S}(t)) = \sum_{j:t_j < t} \frac{d_j}{n_j(n_j - d_j)} \quad (5)$$

onde: d_j = número de falhas no tempo j ; n_j = número de unidades em risco no tempo j (Kaplan e Meier, 1958).

Em várias áreas de aplicação, nas quais o objetivo principal é o estudo de tempos até a ocorrência de um evento, existe a possibilidade de realizar a análise com base em curvas de sobrevivência. No caso não paramétrico, uma forma de comparar estas curvas é através do teste não paramétrico de log-rank. Nesta forma de teste, a informação obtida é simplesmente da existência ou não de diferenças entre as curvas, não nos fornecendo o tamanho da diferença e nem mesmo um intervalo de confiança para o mesmo. Este teste é apropriado quando os dados assumem as propriedades de riscos proporcionais. Se este pressuposto não for aceito utiliza-se o teste de Wilcoxon (ver, por exemplo, LEE, 1992).

2.3 Técnicas paramétricas

O uso de técnicas paramétricas assume uma distribuição conhecida de probabilidade para modelar os dados. A seguir serão descritas algumas distribuições paramétricas importantes, como a distribuição Exponencial, a Weibull e a Gama.

2.3.1 Distribuição Exponencial

A distribuição exponencial é a forma mais simples para descrever tempos de falha. A distribuição é caracterizada por uma função risco constante dada por,

$$\lambda(t) = \frac{1}{\alpha}, t \geq 0, \alpha \geq 0 \quad (6)$$

em que α é o tempo médio de vida.

A função risco é conhecida também como a taxa de falha instantânea, se caracterizando, nesta distribuição pelo fato de ser independente de t , ou seja, a chance de falha em um intervalo de tempo não depende do tempo que o objeto permanece no estudo (falta de memória). A função de sobrevivência $S(t)$ é dada por,

$$S(t) = \exp \left\{ - \left(\frac{t}{\alpha} \right) \right\}. \quad (7)$$

A função densidade de probabilidade para o tempo de falha T com distribuição exponencial é dada por,

$$f(t) = \frac{1}{\alpha} \exp \left\{ - \frac{t}{\alpha} \right\}, t \geq 0. \quad (8)$$

Denota-se a distribuição exponencial por $T \sim Exp(\alpha)$.

2.3.2 Distribuição Weibull

A distribuição Weibull, amplamente conhecida em virtude de sua simplicidade e flexibilidade em acomodar diferentes formas de função de risco, é talvez o modelo de distribuição mais utilizado em análise de tempos de vida. Para uma variável aleatória T com distribuição Weibull, a função de densidade probabilidade é dada por,

$$f(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \exp \left\{ - \left(\frac{t}{\beta} \right)^\alpha \right\}, t \geq 0, \quad (9)$$

em que α é o parâmetro de forma e β o de escala, ambos positivos.

Para esta distribuição, a função de sobrevivência é,

$$S(t) = \exp \left\{ - \left(\frac{t}{\beta} \right)^\alpha \right\} \quad (10)$$

e tem taxa de falha igual a,

$$\lambda(t) = \frac{\alpha}{\beta^\alpha} t^{\alpha-1} \quad (11)$$

em que $t > 0$, $\alpha > 0$ e $\beta > 0$.

Neste caso, podem-se ter riscos (taxas de falha) crescentes para $\alpha > 1$; decrescentes para $\alpha < 1$ e constantes para $\alpha = 1$. Para $\alpha = 1$, o modelo se resume a distribuição exponencial.

Denota-se a distribuição Weibull por $T \sim Weibull(\alpha, \beta)$.

2.3.3 Distribuição Gama

Essa distribuição também inclui a distribuição exponencial como um caso especial quando $k = 1$. A função de densidade da distribuição gama, que é caracterizada por dois parâmetros, k e α , em que $k > 0$ é chamado parâmetro de forma e $\alpha > 0$ de escala, e é expressa por,

$$f(t) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp \left\{ - \left(\frac{t}{\alpha} \right) \right\}, t \geq 0, \quad (12)$$

com função gama $\Gamma(k)$ definida por $\Gamma(k) = \int_0^\infty x^{k-1} \exp \{-x\} dx$.

A função de sobrevivência é dada por,

$$S(t) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du \quad (13)$$

A função de taxa de falha, obtida da relação $\lambda(t) = \frac{f(t)}{S(t)}$, apresenta um padrão crescente ou decrescente, no entanto, convergindo para um valor constante quando t tende a infinito.

Denota-se a distribuição Gama por $T \sim Gama(k, \alpha)$. Vários outros modelos paramétricos são propostos na literatura para tempos de sobrevida, como a distribuição log-normal e a distribuição gama generalizada.

Um modelo de regressão bastante utilizado na análise de dados de sobrevivência na presença de covariáveis é o Modelo de Regressão de Cox (COX, 1972). O modelo de regressão de Cox, é um modelo semiparamétrico, pois não assume uma distribuição paramétrica para os tempos de sobrevida, mas assume um modelo de riscos proporcionais para a função de risco. A razão de o modelo ser muito utilizado, principalmente na área médica, é a disponibilidade em softwares estatísticos mais usuais, o que os torna mais acessíveis ao profissional da área da saúde. Vale mencionar que este é também denominado como modelo de riscos proporcionais, pois a razão de taxas de falha de dois indivíduos é constante no tempo, sendo esta situação um pressuposto do modelo.

2.4 Método de Máxima Verossimilhança

Uma amostra de observações não censuradas $t_1, \dots, t_n$ de uma população, onde os tempos de sobrevida tenham uma densidade $f(t; \theta)$, onde θ é um parâmetro desconhecido. A função de verossimilhança para o parâmetro θ é dada por

$$L(\theta) = \prod_{i=1}^n f(t_i; \theta). \quad (14)$$

Na expressão (14), θ pode estar representando um único parâmetro ou um vetor de parâmetros (ver, por exemplo, COLOSIMO; GIOLO, 2006).

Para definir a verossimilhança para dados censurados, considere T uma variável aleatória representando o tempo de falha de um paciente e C uma variável aleatória, independente de T , representando o tempo de censura. Para o paciente temos que, $t = \min(T, C)$ e defina

$$\delta_i = \begin{cases} 0, & \text{se } T > C \\ 1, & \text{se } T \leq C. \end{cases}$$

Suponha que os pares (Ti, Ci) , para $i = 1, \dots, n$ formam uma amostra aleatória de n pacientes.

As observações podem ser divididas em dois conjuntos, as r primeiras observações ordenadas são as observações não censuradas $(1, 2, \dots, r)$ e as $n-r$ seguintes são observações censuradas $(r+1, r+2, \dots, n)$.

Para a censura do Tipo I, considere r falhas e $n-r$ censuras observadas ao término do experimento; portanto $L(\theta)$ é dada por

$$L(\theta) = \prod_{i=1}^r f(t_i; \theta) \prod_{i=r+1}^n S(t_i; \theta) \quad (15)$$

onde o segundo termo tem a forma $\prod_{i=r+1}^n S(c; \theta) = [S(c; \theta)]^{n-r}$, quando todos tempos censurados forem iguais a c . Observar que r é aleatório.

Para dados com censura do Tipo II, somente os r menores tempos são observados e r é fixo. Assim $L(\theta)$ é dada por

$$L(\theta) = \frac{n!}{(n-r)!} \prod_{i=1}^r f(t_i; \theta) \prod_{i=r+1}^n S(t_i; \theta) \quad (16)$$

em que $\prod_{i=r+1}^n S(t_i; \theta) = [S(t_r; \theta)]^{n-r}$ com t_r o maior tempo observado. Como o termo $\frac{n!}{(n-r)!}$ é uma constante e não depende de nenhum parâmetro, pode ser desprezado. Então se tem

$$L(\theta) \propto \prod_{i=1}^r f(t_i; \theta) \prod_{i=r+1}^n S(t_i; \theta) \quad (17)$$

Os estimadores de máxima verossimilhança são obtidos maximizando-se a função de verossimilhança. Na prática, é mais comum maximizar o logaritmo da função de verossimilhança.

Testes de hipóteses e intervalos de confiança, em geral, são obtidos a partir da distribuição normal assintótica dos estimadores de máxima verossimilhança, onde a variância é obtida do inverso da matriz de informação de Fisher. Isso pode ser um problema nas aplicações médicas, pois muitas vezes podemos ter amostras pequenas ou grande proporção de dados censurados.

3 Uma breve revisão da metodologia bayesiana

Os métodos bayesianos têm sido considerados alternativas muito eficazes e poderosas na análise de dados. Esse método possui uma filosofia muito diferente do método frequentista. No método clássico, os parâmetros do modelo são considerados constantes desconhecidas, e no método bayesiano todos os parâmetros são considerados quantidades aleatórias. Uma característica muito importante é o fato desta análise permitir a incorporação da informação de um especialista junto à informação dos dados. A inferência bayesiana tem como fundamentação a Fórmula de Bayes, a qual combina os dados com a informação a priori e, então, se obtém a posteriori (priori já complementada pela informação dos dados), onde é realizado todo o processo inferencial (BOX; TIAO, 1973; PAULINO; TURKMAN; MURTEIRA, 2003).

3.1 Fórmula de Bayes

Considere uma partição do espaço amostral Ω , onde os eventos $A_1, A_2, \dots, A_k$ formam uma sequência de eventos mutuamente exclusivos e exaustivos, isto é, $\bigcup_{j=1}^k A_j = \Omega$ e $A_i \cap A_j = \phi$ (conjunto vazio) para $i \neq j$ tal que $P\left(\bigcup_{j=1}^k A_j\right) = \sum_{j=1}^k P(A_j) = 1$.

Sendo assim, para qualquer outro evento B ($B \subset \Omega$), temos,

$$P(A_i | B) = \frac{P(B | A_i) P(A_i)}{\sum_{j=1}^k P(B | A_j) P(A_j)} \quad (18)$$

para todo i variando de 1 até k .

Assumindo agora um vetor de dados $\mathbf{y} = (y_1, \dots, y_n)'$ e θ (quantidades desconhecidas) os parâmetros de uma distribuição de probabilidade associada com a variável aleatória Y_i com valores observados $y_i, i = 1, \dots, n$.

Considere $\mathbf{y} = (y_1, \dots, y_n)$ uma amostra aleatória, onde os dados são independentes e identicamente distribuídos. Suponha que $\mathbf{y}$ é um vetor de observações de uma distribuição conjunta dada por $f(\mathbf{y} | \theta)$ e seja $\pi(\theta)$ uma distribuição a priori para θ . Então assumindo

os valores discretos $\theta_1, \dots, \theta_k$, temos de (18), a distribuição a posteriori para θ_i dado $\mathbf{y}$ é dado por

$$\pi(\theta_i | \mathbf{y}) = \frac{f(\mathbf{y} | \theta_i) \pi(\theta_i)}{\sum_{j=1}^k f(\mathbf{y} | \theta_j) \pi(\theta_j)} \quad (19)$$

onde o parâmetro θ também é considerado como uma quantidade aleatória, no enfoque bayesiano.

Para θ assumindo valores contínuos num dado intervalo, podemos escrever (19) por

$$\pi(\theta | \mathbf{y}) = \frac{f(\mathbf{y} | \theta) \pi(\theta)}{\int f(\mathbf{y} | \theta) \pi(\theta) d\theta} \quad (20)$$

em que a integral no denominador de (20) é definida no intervalo de variação de θ .

3.2 Distribuição a Priori

Na análise bayesiana, a distribuição a priori é utilizada a fim de representar o que já é conhecido sobre parâmetros desconhecidos, antes de se avaliar os dados. Deve-se ter muita cautela ao definir uma distribuição a priori, afinal se esta informação não for bem definida pode-se chegar a interpretações errôneas. Uma distribuição a priori para um parâmetro pode se dar de várias formas, sendo possível ocorrerem a partir de procedimentos subjetivos ou objetivos.

A distribuição a priori conjugada é uma priori informativa, onde a distribuição a priori e a posteriori pertencem à mesma classe de distribuições. A passagem da priori para a posteriori envolve apenas uma simples mudança nos parâmetros, sem a necessidade de cálculos adicionais (PAULINO; TURKMAN; MURTEIRA, 2003).

A distribuição a priori também pode ser a incorporação do conhecimento de um pesquisador, ou seja, o pesquisador se baseia na sua prática e no seu “feeling” para definir a priori. Em muitas situações práticas esse conhecimento do especialista não existe ou, se existe, não é fidedigno. Nesse caso, caracteriza-se uma “ignorância a priori” (PAULINO; TURKMAN; MURTEIRA, 2003).

Nestas situações são utilizadas as prioris não informativas, ou seja, prioris de referência “neutras”(BOX; TIAO, 1973). A utilização deste tipo de distribuição a priori permite a comparação com os resultados obtidos pela inferência clássica, haja vista que através de uma priori não informativa, o modelo é baseado apenas na informação dos dados amostrais. Existem vários métodos para se definir a priori não informativa, como por exemplo: Método de Bayes- Laplace, Método de Jeffreys, entre outros (BOX; TIAO, 1973; PAULINO; TURKMAN; MURTEIRA, 2003).

3.3 Um Critério para Discriminação de Modelos - DIC

O critério DIC, proposto por Spiegelhalter et al. (2002) é baseado na média a posteriori da deviance, definida por

$$D(\boldsymbol{\theta}) = -2\ln L(\boldsymbol{\theta}) + C, \quad (21)$$

em que $\boldsymbol{\theta}$ é um vetor de parâmetros desconhecidos do modelo; $L(\boldsymbol{\theta})$ é a verossimilhança e C é uma constante (nem sempre conhecida) na comparação de dois modelos.

O critério DIC é dado então por

$$DIC = D(\hat{\boldsymbol{\theta}}) + 2p_D, \quad (22)$$

em que $D(\hat{\boldsymbol{\theta}})$ é o desvio calculado na média a posteriori $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{y})$ e p_D é o número de parâmetros do modelo, dado por $p_D = \bar{D} - D(\hat{\boldsymbol{\theta}})$, em que $\bar{D} = E[D(\boldsymbol{\theta})|y]$ é a média a posteriori do desvio que mede a qualidade do ajuste dos dados para cada modelo. Para a conclusão, os menores valores de DIC indicam os melhores modelos.

3.4 Métodos de simulação para amostras da distribuição a posteriori

Para a obtenção de sumários a posteriori de interesse, geralmente é necessário resolver integrais bayesianas que não apresentam solução analítica. Na prática, observa-se que os

modelos utilizados nem sempre são simples para se obter os resumos a posteriori. Mesmo que se tenha uma priori e uma verossimilhança simples, a junção delas pode produzir uma distribuição a posteriori muito complicada (PAULINO; TURKMAN; MURTEIRA, 2003).

Os métodos com base em amostragem, como, por exemplo, o método de Monte Carlo com cadeias de Markov (MCMC), passaram a ser utilizados com o avanço das técnicas computacionais. Esse método consiste na simulação de uma variável aleatória através de uma cadeia de Markov, no qual a sua distribuição assintoticamente se aproxima da distribuição a posteriori (BERNARDO; SMITH, 1994).

A cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia depende somente do estado atual e dos dados. No entanto, existe uma relação com o estado inicial, que é descartado após um período de aquecimento, o chamado "Burn-in".

As formas mais usuais dos métodos MCMC são os amostradores de Gibbs e o algoritmo de Metropolis-Hastings. Estas duas formas simulam amostras da distribuição a posteriori conjunta a partir das distribuições condicionais (GELFAND; SMITH, 1990; CHIB; GREENBERG, 1995).

O amostrador de Gibbs nos permite gerar amostras da distribuição a posteriori conjunta desde que as distribuições condicionais completas possuam formas fechadas ou conhecidas. Por outro lado, o algoritmo de Metropolis-Hastings permite gerar amostras da distribuição a posteriori conjunta com distribuições condicionais completas possuindo ou não uma forma conhecida ou fechada.

3.4.1 Amostrador de Gibbs

Suponha $\pi(\boldsymbol{\theta}|y)$ uma distribuição a posteriori conjunta, sendo $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$, no qual desejamos obter inferências. Para isso, simulam-se quantidades aleatórias de distribuições condicionais completas $\pi(\theta_i|\mathbf{y}, \boldsymbol{\theta}_{(i)})$.

Considere os valores iniciais (arbitrários) para $\boldsymbol{\theta}$: $\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_k^{(0)}$. Desta forma, segue o seguinte algoritmo:

- Gerar $\theta_1^{(1)}$ de $\pi(\theta_1|y, \theta_2^{(0)}, \dots, \theta_k^{(0)})$;
- Gerar $\theta_2^{(1)}$ de $\pi(\theta_2|y, \theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)})$;
- (...)
- Gerar $\theta_k^{(1)}$ de $\pi(\theta_k|y, \theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_{k-1}^{(1)})$.

Substitua os valores iniciais por $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \dots, \theta_k^{(1)})$, para uma nova realização. Os valores $\theta_1^{(z)}, \theta_2^{(z)}, \dots, \theta_k^{(z)}$, para z suficientemente grande, convergem para um valor da quantidade aleatória com distribuição $\pi(\boldsymbol{\theta}|y)$ (BERNARDO; SMITH, 1994, pág 353; CASELA e GEORGE, 1992).

3.4.2 Algoritmo de Metropolis Hastings

Suponha uma amostra de densidade não regular $\pi(\theta_i|\boldsymbol{\theta}_{(i)})$, em que $\boldsymbol{\theta}_{(i)} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_k)$. Seja $q(\theta, \beta)$ o núcleo de transição da distribuição $p(\theta)$ que representa $\pi(\theta_i|\boldsymbol{\theta}_{(i)})$ e que transforma θ em β .

Desta forma o algoritmo é dado por:

- Inicie com $\theta^{(0)}$ e indicador de estágio $j = 0$;
- Gerar um ponto β do núcleo de transição $q(\theta^{(j)}, \beta)$;
- Atualizar $\theta^{(j)}$ por $\theta^{(j+1)} = \beta$, com probabilidade, $p = \min \left\{ 1, \frac{p(\beta)q[\theta^{(j)}, \beta]}{p[\theta^{(j)}]q[\beta, \theta^{(j)}]} \right\}$, ficar com $\theta^{(j)}$ com probabilidade $1 - p$;
- Repetir os dois últimos passos até conseguir uma distribuição estacionária.

Importante observar: se um valor candidato é rejeitado, então o valor atual é considerado na próxima etapa; o terceiro passo o valor de p não depende da constante normalizadora; o algoritmo de Metropolis Hastings é especificado pela densidade candidata para geração $q(x, y)$ (BERNARDO; SMITH, 1994; CHIB; GREENBERG, 1995).

4 Distribuição exponencial generalizada

Uma distribuição exponencial generalizada (GUPTA; KUNDU, 1999) pode ser uma boa alternativa ao uso das populares distribuições Gama e Weibull usadas na análise de dados de sobrevivência (RAQAB, 2002; RAQAB e AHSANULLAH, 2001; ZHENG, 2002; SARHAN, 2007; GUPTA; KUNDU, 2007).

4.1 Função Densidade

A distribuição exponencial generalizada com dois parâmetros tem função densidade de probabilidade dada por,

$$f(t; \alpha, \lambda) = \alpha\lambda[1 - \exp(-\lambda t)]^{\alpha-1}\exp(-\lambda t), \quad (23)$$

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

A densidade (23) 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 $1/\lambda$.

Como um caso especial, assumindo $\lambda = 1$, temos na figura 6, gráficos da função de densidade (23) considerando diferentes valores de α .

4.2 Função de Sobrevivência e de Risco

As funções de sobrevivência e de risco associadas à densidade (23) são dadas, respectivamente por

$$S(t; \alpha, \lambda) = P(T > t) = 1 - [1 - \exp(-\lambda t)]^\alpha, \quad (24)$$

Figura 6 – Gráfico da função densidade para vários valores de alfa (lambda=1)

$$h(t; \alpha, \lambda) = \frac{f(t; \alpha, \lambda)}{S(t; \alpha, \lambda)} = \frac{\alpha \lambda [1 - \exp(-\lambda t)]^{\alpha-1} \exp(-\lambda t)}{1 - [1 - \exp(-\lambda t)]^\alpha}. \quad (25)$$

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.

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

$$t_{med} = -\frac{1}{\lambda} \log \left[1 - \left(\frac{1}{2} \right)^{\frac{1}{\alpha}} \right]. \quad (26)$$

Como um caso especial, assumindo $\lambda = 1$, temos na figura 7, gráficos da função de risco (25) considerando diferentes valores de α .

4.3 Função Geradora de Momentos

A função geradora de momentos da variável aleatória T com distribuição exponencial generalizada com densidade (23) é dada (ver Gupta e Kundu, 2008) por,

Figura 7 – Gráfico da função de risco para vários valores de alfa (lambda=1)

$$M(s) = E[e^{sT}] = \frac{\Gamma(\alpha + 1)\Gamma(1 - \frac{s}{\lambda})}{\Gamma(\alpha - \frac{s}{\lambda} + 1)}, \quad (27)$$

para $s < \lambda$.

De (27), encontramos os momentos de interesse. A média e variância de T são dadas respectivamente por

$$\begin{aligned} E(T) &= \frac{1}{\lambda}[\psi(\alpha + 1) - \psi(1)], \\ Var(T) &= \frac{1}{\lambda^2}[\psi'(1) - \psi'(\alpha + 1)], \end{aligned} \quad (28)$$

onde $\psi(\cdot)$ é a função digamma dada por $\psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ e $\Gamma(x)$ é a função gama.

4.4 Função de Verossimilhança para dados completos

Seja $T_1, T_2, \dots, T_n$ uma amostra aleatória de tamanho n da distribuição exponencial generalizada com densidade (23). A função de verossimilhança para α e λ é dada por

$$\begin{aligned}
L(\alpha, \lambda) &= \prod_{i=1}^n f(t_i; \alpha, \lambda) \\
&= \alpha^n \lambda^n \left[\prod_{i=1}^n (1 - e^{-\lambda t_i})^{\alpha-1} \right] \exp\left(-\lambda \sum_{i=1}^n t_i\right).
\end{aligned} \tag{29}$$

O logaritmo da função de verossimilhança (29) é dado por

$$\begin{aligned}
l(\alpha, \lambda) &= \ln L(\alpha, \lambda) \\
&= n \ln \alpha + n \ln \lambda - \lambda \sum_{i=1}^n t_i + (\alpha - 1) \sum_{i=1}^n \ln(1 - e^{-\lambda t_i}).
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \frac{n}{\alpha} + \sum_{i=1}^n \ln(1 - e^{-\lambda t_i}) = 0, \\
\frac{\partial l}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n t_i + (\alpha - 1) \sum_{i=1}^n \frac{t_i e^{-\lambda t_i}}{(1 - e^{-\lambda t_i})} = 0.
\end{aligned} \tag{31}$$

De (31), encontramos o EMV para α dado por,

$$\hat{\alpha} = - \frac{n}{\sum_{i=1}^n \log[1 - \exp(-\hat{\lambda} T_i)]}. \tag{32}$$

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

$$\frac{n}{\hat{\lambda}} - n\bar{t} + (\hat{\alpha} - 1) \sum_{i=1}^n \frac{t_i e^{-\hat{\lambda} t_i}}{(1 - e^{-\hat{\lambda} t_i})} = 0, \tag{33}$$

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

Observar que precisamos usar um método iterativo para achar o EMV $\hat{\lambda}$ de (33).

As segundas derivadas de $l(\alpha, \lambda)$ são dadas por,

$$\begin{aligned}\frac{\partial^2 l}{\partial \alpha^2} &= -\frac{n}{\alpha^2} \\ \frac{\partial^2 l}{\partial \lambda^2} &= -\frac{n}{\lambda^2} - (\alpha - 1) \sum_{i=1}^n \frac{t_i^2 e^{-\lambda t_i}}{(1 - e^{-\lambda t_i})^2} \\ \frac{\partial^2 l}{\partial \alpha \partial \lambda} &= \sum_{i=1}^n \frac{t_i e^{-\lambda t_i}}{1 - e^{-\lambda t_i}}\end{aligned}\tag{34}$$

Testes de hipóteses e intervalos de confiança para α e λ podem ser obtidos usando a distribuição normal assintótica para $\hat{\alpha}$ e $\hat{\lambda}$, isto é,

$$(\hat{\alpha}, \hat{\lambda}) \stackrel{a}{\sim} N\{(\alpha, \lambda), I_0^{-1}\}\tag{35}$$

em que I_0 é a matriz de informação observada de Fisher dada por

$$I_0 = \left(\begin{array}{cc} E\left(-\frac{\partial^2 l}{\partial \alpha^2}\right) & E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) \\ E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) & E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) \end{array} \right) \bigg|_{(\hat{\alpha}, \hat{\lambda})}\tag{36}$$

4.5 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 (BOX; TIAO, 1973) para α e λ é dada por

$$\Pi_1(\alpha, \lambda) \propto \{det I(\alpha, \lambda)\}^{1/2}\tag{37}$$

em que, $I(\alpha, \lambda)$ é a matriz de informação de Fisher esperada para α e λ dada por,

$$I(\alpha, \lambda) = \begin{pmatrix} E(-\frac{\partial^2 l}{\partial \alpha^2}) & E(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}) \\ E(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}) & E(-\frac{\partial^2 l}{\partial \lambda^2}) \end{pmatrix} \quad (38)$$

em que

$$\begin{aligned} E\left(-\frac{\partial^2 l}{\partial \alpha^2}\right) &= \frac{n}{\alpha^2} \\ E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) &= \frac{n}{\lambda} \left\{ [\psi(\alpha + 1) - \psi(1)] - \frac{\alpha}{\alpha - 1} [\psi(\alpha) - \psi(1)] \right\} \\ E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) &= \frac{n}{\lambda^2} \left\{ 1 + \frac{\alpha(\alpha - 1)}{\alpha - 2} \times \right. \\ &\quad \times [\psi'(1) - \psi'(\alpha - 1) + [\psi(\alpha - 1) - \psi(1)]^2] - \\ &\quad \left. - \alpha [\psi'(1) - \psi'(\alpha) + [\psi(\alpha) - \psi(1)]^2] \right\} \end{aligned} \quad (39)$$

(ver Gupta e Kundu, 1999).

Uma possível simplificação é considerar uma distribuição a priori não-informativa obtida a partir de $\Pi(\alpha, \lambda) = \Pi(\lambda|\alpha)\Pi_0(\alpha)$. Usando a regra de Jeffreys, temos,

$$\Pi_2(\alpha, \lambda) \propto \left\{ E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) \right\}^{1/2} \Pi_0(\alpha) \quad (40)$$

em que $E(-\frac{\partial^2 l}{\partial \lambda^2})$ é dada em (39) e $\Pi_0(\alpha)$ é uma distribuição a priori não-informativa dada por $\Pi_0(\alpha) \propto 1/\alpha$, $\alpha > 0$.

Assim,

$$\Pi_2(\alpha, \lambda) \propto \frac{1}{\alpha \lambda} A^{1/2}(\alpha) \quad (41)$$

em que

$$\begin{aligned} A(\alpha) &= 1 + \frac{\alpha(\alpha - 1)}{\alpha - 2} [\psi'(1) - \psi'(\alpha - 1) + [\psi(\alpha - 1) - \psi(1)]^2] - \\ &\quad - \alpha [\psi'(1) - \psi'(\alpha) + [\psi(\alpha) - \psi(1)]^2]. \end{aligned}$$

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

Como $\psi(\alpha+1) = \psi(\alpha) + 1/\alpha$ (ABRAMOWITZ; STEGUN, 1970, pág. 258), observamos que

$$\frac{\psi(\alpha) + 1/\alpha - \psi(1)}{\psi(\alpha) - \psi(1)} \rightarrow 1$$

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

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

$$\Pi_3(\alpha, \lambda) \propto \frac{1}{\alpha\lambda} \quad (42)$$

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

$$\Pi_4(\alpha, \lambda) = \Pi_\alpha(\alpha)\Pi_\lambda(\lambda) \quad (43)$$

em que $\Pi_\alpha(\alpha)$ e $\Pi_\lambda(\lambda)$ são distribuições gama,

$$\begin{aligned} \Pi_\alpha(\alpha) &\sim Gama(a_\alpha, b_\alpha) \\ \Pi_\lambda(\lambda) &\sim Gama(a_\lambda, b_\lambda) \end{aligned} \quad (44)$$

em que a_α , b_α , a_λ 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 (44) pode ser feita a partir da informação de especialistas sobre

a média e a variância de α e λ ; com isso encontramos a_α , b_α , a_λ e b_λ . Ressalta-se que também seria possível a utilização de métodos bayesianos empíricos para a escolha do hiperparâmetros.

Usando a fórmula de Bayes, combina-se a função de verossimilhança (29) com uma das distribuições a priori $\Pi_j(\alpha, \lambda)$, $j = 1, \dots, 4$, para encontrarmos a distribuição a posteriori para α e λ . Assumindo a priori informativa (43), a distribuição a posteriori conjunta para α e λ é dada por

$$\begin{aligned} \Pi(\alpha, \lambda | \mathbf{t}) &\propto \alpha^{n+a_\alpha-1} \lambda^{n+a_\lambda-1} \times \\ &\times \exp \left\{ -\alpha \left[b_\alpha - \sum_{i=1}^n \log(1 - e^{-\lambda t_i}) \right] \right\} \times \\ &\times \exp \{ -\lambda [b_\lambda + n\bar{t}] \} \end{aligned} \quad (45)$$

em que $\mathbf{t} = (t_1, \dots, t_n)$ é o vetor dos dados e $\bar{t}$ é a média dos dados.

Amostras de distribuição a posteriori (45) podem ser geradas usando métodos de simulação de Monte Carlo em Cadeias de Markov (MCMC). As distribuições condicionais necessárias para o amostrador de Gibbs são dadas por

$$\begin{aligned} (i) \alpha | \lambda, \mathbf{t} &\sim \text{Gama}(n + a_\alpha, b_\alpha - \sum_{i=1}^n \log(1 - e^{-\lambda t_i})); \\ (ii) \lambda | \alpha, \mathbf{t} &\sim \text{Gama}(n + a_\lambda, b_\lambda + n\bar{t}) \end{aligned} \quad (46)$$

Uma grande simplificação na geração de amostras a posteriori (45) é obtida usando o software Winbugs (SPIEGELHALTER, 2001) o qual requer somente a especificação da distribuição conjunta para os dados e as distribuições a priori para os parâmetros do modelo.

É importante salientar que a convergência do amostrador de Gibbs pode melhorar considerando transformações dos parâmetros α e λ . Uma possível reparametrização é

dada por $\rho_1 = \log \alpha$ e $\rho_2 = \log \lambda$. Neste caso, assumindo independência a priori entre ρ_1 e ρ_2 , uma priori não informativa para ρ_1 e ρ_2 é dada por

$$\Pi(\rho_1, \rho_2) \propto \text{constante}, \quad (47)$$

em que $-\infty < \rho_1, \rho_2 < \infty$.

Outra possibilidade é considerar uma transformação nos dados. Uma possível transformação é dada por $Y = \log T$, onde o Jacobiano da transformação é dado por $|dt/dy| = e_y$.

Assim, a densidade para $Y = \log(T)$ é dada por

$$f(y; \alpha, \lambda) = \alpha \lambda [1 - \exp(-\lambda e^y)]^{\alpha-1} \exp(y - \lambda e^y), \quad (48)$$

em que $-\infty < y < \infty$.

4.6 Presença de dados censurados

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

$$\delta_i = \begin{cases} 0, & \text{se } t_i = L_i \\ 1, & \text{se } t_i = T_i. \end{cases}$$

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

A função de verossimilhança para α e λ é dada por

$$L(\alpha, \lambda) = \prod_{i=1}^n f^{\delta_i}(t_i; \alpha, \lambda) S^{1-\delta_i}(t_i; \alpha, \lambda), \quad (49)$$

em que $f(t_i; \alpha, \lambda)$ é a densidade exponencial generalizada dada em (23) e $S(t_i; \alpha, \lambda)$ é a

função de sobrevivência dada por (24).

Assim, temos,

$$L(\alpha, \lambda) = \alpha^r \lambda^r \exp \left\{ (\alpha - 1) \sum_{i=1}^n \delta_i \log(1 - e^{-\lambda t_i}) - \right. \\ \left. - \lambda \sum_{i=1}^n \delta_i t_i + \sum_{i=1}^n (1 - \delta_i) \log [1 - (1 - e^{-\lambda t_i})^\alpha] \right\}, \quad (50)$$

em que $r = \sum_{i=1}^n \delta_i$ (número de observações completas).

O logaritmo da função de verossimilhança (50) é dada por

$$l(\alpha, \lambda) = r \log \alpha + r \log \lambda + (\alpha - 1) \sum_{i=1}^n \delta_i \log(1 - e^{-\lambda t_i}) - \\ - \lambda \sum_{i=1}^n \delta_i t_i + \sum_{i=1}^n (1 - \delta_i) \log [1 - (1 - e^{-\lambda t_i})^\alpha]. \quad (51)$$

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.

4.7 Presença de Covariáveis

Na presença de um vetor de covariáveis $\mathbf{x} = (1, x_1, x_2, \dots, x_k)'$, assumimos a densidade exponencial generalizada (23) com o parâmetro de escala λ dependendo das covariáveis na forma

$$\lambda(\mathbf{x}_i) = e^{\beta \mathbf{x}_i'}, \quad (52)$$

em que $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \dots, x_{ki})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ e $\beta \mathbf{x}_i' = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$, para $i = 1, 2, \dots, n$.

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

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^n f^{\delta_i}(t_i|\alpha, \boldsymbol{\beta}, \mathbf{x}_i) S^{1-\delta_i}(t_i|\alpha, \boldsymbol{\beta}, \mathbf{x}_i), \quad (53)$$

em que δ_i é uma função indicadora de censuras dada em (49), $f(t_i|\alpha, \boldsymbol{\beta}, \mathbf{x}_i)$ e $S(t_i|\alpha, \boldsymbol{\beta}, \mathbf{x}_i)$ são dados, respectivamente em (23) e (24) com $\lambda(\mathbf{x}_i)$ definido por (52) em lugar de λ .

Desta forma, a função de verossimilhança (53) é dada por

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}) = & \alpha^r \exp \left\{ \sum_{i=1}^n \delta_i \boldsymbol{\beta}' \mathbf{x}_i - \sum_{i=1}^n \delta_i t_i e^{\boldsymbol{\beta}' \mathbf{x}_i} + \right. \\ & + (\alpha - 1) \sum_{i=1}^n \delta_i \log \left(1 - e^{-t_i e^{\boldsymbol{\beta}' \mathbf{x}_i}} \right) + \\ & \left. + \sum_{i=1}^n (1 - \delta_i) \log \left[1 - \left(1 - e^{-t_i e^{\boldsymbol{\beta}' \mathbf{x}_i}} \right)^\alpha \right] \right\}, \end{aligned} \quad (54)$$

em que r é o número de observações completas definido em (50).

Para uma análise bayesiana do modelo, assumimos uma distribuição a priori gama para α (ver (44)) com hiperparâmetros conhecidos a_α e b_α e distribuições a priori normais para β_l , $l = 0, 1, \dots, k$, isto é,

$$\beta_l \sim N(0; b_l^2), \quad (55)$$

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

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

$$\begin{aligned}
\Pi(\alpha, \boldsymbol{\beta} | \mathbf{t}, \mathbf{x}) &\propto \alpha^{r+a_\alpha-1} e^{-b_\alpha \alpha} \times \left(\prod_{i=1}^n \exp \left[-\frac{\beta_l^2}{2b_l^2} \right] \right) \times \\
&\times \exp \left\{ \sum_{i=1}^n \delta_i \boldsymbol{\beta}' \mathbf{x}_i - \sum_{i=1}^n \delta_i t_i e^{\boldsymbol{\beta}' \mathbf{x}_i} + \right. \\
&+ (\alpha - 1) \sum_{i=1}^n \delta_i \log(1 - e^{-t_i e^{\boldsymbol{\beta}' \mathbf{x}_i}}) + \\
&\left. + \sum_{i=1}^n (1 - \delta_i) \log \left[1 - (1 - e^{-t_i e^{\boldsymbol{\beta}' \mathbf{x}_i}})^\alpha \right] \right\}.
\end{aligned} \tag{56}$$

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

5 Distribuição 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.

5.1 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 $F_1(t_1), F_2(t_2), \dots, F_m(t_m)$, a partir da função,

$$C(F_1(t_1), F_2(t_2), \dots, F_m(t_m)) = F(t_1, t_2, \dots, t_m) \quad (57)$$

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

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

$$F(t_1, t_2, \dots, t_m) = C(F_1(t_1), F_2(t_2), \dots, F_m(t_m)) \quad (58)$$

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, 2006), dados por

$$\max(0, u + v - 1) \leq C(u, v) \leq \min(u, v) \quad (59)$$

Diferentes famílias de funções cópula são introduzidas na literatura. Na área de finanças a função cópula Gaussiana é muito utilizada (SANTOS E PEREIRA, 2010). A cópula Gaussiana é construída baseada na distribuição normal bivariada, dada por

$$C_p(u, v) = \Phi_\rho [\Phi^{-1}(u), \Phi^{-1}(v)] , \quad (60)$$

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

$$\begin{aligned}\Phi_\rho(x, y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^x \int_{-\infty}^y \frac{1}{2\pi\sqrt{1-\rho^2}} \times \\ &\times \exp\left\{-\frac{1}{2(1-\rho^2)}(z^2 - 2\rho zw + w^2)\right\} dz dw\end{aligned}\tag{61}$$

5.2 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-Morgenstern (MORGENSTERN, 1956) dada por

$$\begin{aligned}C(u, v) &= [1 - e^{\ln(1-u)}] [1 - e^{\ln(1-v)}] \times \\ &\times [1 + \theta \exp\{(\ln(1-u) + \ln(1-v))\}],\end{aligned}\tag{62}$$

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

$$C(u, v) = uv [1 + \theta(1-u)(1-v)],\tag{63}$$

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,

$$\begin{aligned}\rho_S(T_1, T_2) &= 12 \int_0^1 \int_0^1 [C(u, v) - uv] dudv \\ \rho_T(T_1, T_2) &= 4 \int_0^1 \int_0^1 C(u, v) dC(u, v) - 1\end{aligned}\tag{64}$$

De (63), 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 (64) usando métodos bayesianos empíricos ou usar a informação de especialistas.

Seja distribuições exponenciais generalizadas marginais dadas por,

$$\begin{aligned}u &= F_1(t_1) = P\{T_1 \leq t_1\} = (1 - \exp(-\lambda_1 t_1))^{\alpha_1} \\ v &= F_2(t_2) = P\{T_2 \leq t_2\} = (1 - \exp(-\lambda_2 t_2))^{\alpha_2}\end{aligned}\tag{65}$$

A função distribuição conjunta para T_1 e T_2 dada por,

$$\begin{aligned}F(t_1, t_2 | \lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) &= C(F_1(t_1), F_2(t_2)) \\ &= F_1(t_1)F_2(t_2) \times [1 + \theta(1 - F_1(t_1))(1 - F_2(t_2))]\end{aligned}\tag{66}$$

Então,

$$\begin{aligned}
F(t_1, t_2 | \lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) &= (1 - \exp(-\lambda_1 t_1))^{\alpha_1} (1 - \exp(-\lambda_2 t_2))^{\alpha_2} \times \\
&\times [1 + \theta(1 - (1 - \exp(-\lambda_1 t_1))^{\alpha_1}) \times \\
&\times (1 - (1 - \exp(-\lambda_2 t_2))^{\alpha_2})]
\end{aligned} \tag{67}$$

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

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

$$f(t_1, t_2 | \lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) = \frac{\partial^2 F(t_1, t_2)}{\partial t_1 \partial t_2} \tag{68}$$

De (66) temos

$$\begin{aligned}
f(t_1, t_2 | \lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) &= f_1(t_1) f_2(t_2) + \theta f_1(t_1) f_2(t_2) \times \\
&\times [(1 - 2F_1(t_1))(1 - 2F_2(t_2))]
\end{aligned} \tag{69}$$

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

$$\begin{aligned}
f_1(t_1) &= \alpha_1 \lambda_1 [1 - \exp(-\lambda_1 t_1)]^{\alpha_1 - 1} \exp(-\lambda_1 t_1) \\
f_2(t_2) &= \alpha_2 \lambda_2 [1 - \exp(-\lambda_2 t_2)]^{\alpha_2 - 1} \exp(-\lambda_2 t_2)
\end{aligned} \tag{70}$$

onde $F_1(t_1)$ e $F_2(t_2)$ são dados por (65).

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

$$S(t_1, t_2) = P\{T_1 > t_1, T_2 > t_2\} = 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2) \quad (71)$$

em que $F_1(t_1)$ e $F_2(t_2)$ são dadas por (65) e $F(t_1, t_2)$ é dado em (67). Então, tem-se

$$\begin{aligned} S(t_1, t_2) = & 1 - (1 - \exp(-\lambda_1 t_1))^{\alpha_1} - (1 - \exp(-\lambda_2 t_2))^{\alpha_2} + \\ & + (1 - \exp(-\lambda_1 t_1))^{\alpha_1} (1 - \exp(-\lambda_2 t_2))^{\alpha_2} \times \\ & \times [1 + \theta(1 - (1 - \exp(-\lambda_1 t_1))^{\alpha_1})(1 - (1 - \exp(-\lambda_2 t_2))^{\alpha_2})] \end{aligned} \quad (72)$$

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

5.3 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, $i = 1, \dots, n$. 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 sobrevivência observados;
- C_2 : t_{1i} é o tempo de sobrevivência 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 sobrevivência;
- 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

$$\begin{aligned}
L &= \prod_{i \in C_1} f(t_{1i}, t_{2i}) \prod_{i \in C_2} \left(-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right) \times \\
&\times \prod_{i \in C_3} \left(-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right) \prod_{i \in C_4} S(t_{1i}, t_{2i})
\end{aligned} \tag{73}$$

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; $\left(-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right)$ 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 $i = 1, \dots, n$, onde n é o número de observações.

Desta forma, reescrevendo a função de verossimilhança tem-se

$$\begin{aligned}
L &= \prod_{i=1}^n [f(t_{1i}, t_{2i})]^{\delta_{1i}\delta_{2i}} \prod_{i=1}^n \left(-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} \right)^{\delta_{1i}(1-\delta_{2i})} \times \\
&\times \prod_{i=1}^n \left(-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} \right)^{\delta_{2i}(1-\delta_{1i})} \prod_{i=1}^n [S(t_{1i}, t_{2i})]^{(1-\delta_{1i})(1-\delta_{2i})}
\end{aligned} \tag{74}$$

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

$$L = \prod_{i=1}^n f(t_{1i}, t_{2i}) \tag{75}$$

Em (74), substituindo $S(t_{1i}, t_{2i})$ por (72), $f(t_{1i}, t_{2i})$ por (69), onde $f_1(t_1)$ e $f_2(t_2)$ são dados em (70) e $F_1(t_1)$ e $F_2(t_2)$ são dados por (65), assim

$$\begin{aligned}
f(t_{1i}, t_{2i} | \lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta) &= \alpha_1 \alpha_2 \lambda_1 \lambda_2 (1 - \exp(-\lambda_1 t_{1i}))^{\alpha_1 - 1} \times \\
&\times (1 - \exp(-\lambda_2 t_{2i}))^{\alpha_2 - 1} \exp(-\lambda_1 t_{1i} - \lambda_2 t_{2i}) \times \\
&\times [1 + \theta(1 - 2(1 - \exp(-\lambda_1 t_{1i}))^{\alpha_1}) \times \\
&\times (1 - 2(1 - \exp(-\lambda_2 t_{2i}))^{\alpha_2})]
\end{aligned} \tag{76}$$

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

$$\begin{aligned}
-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} &= f_1(t_{1i}) \{1 - F_2(t_{2i}) [1 + \theta(1 - F_2(t_{2i}))(1 - 2F_1(t_{1i}))]\} \\
-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{2i}} &= f_2(t_{2i}) \{1 - F_1(t_{1i}) [1 + \theta(1 - F_1(t_{1i}))(1 - 2F_2(t_{2i}))]\}
\end{aligned} \tag{77}$$

Isto é,

$$\begin{aligned}
-\frac{\partial S(t_{1i}, t_{2i})}{\partial t_{1i}} &= \alpha_1 \lambda_1 (1 - \exp(-\lambda_1 t_{1i}))^{\alpha_1 - 1} \exp(-\lambda_1 t_{1i}) \times \\
&\times \{1 - (1 - \exp(-\lambda_2 t_{2i}))^{\alpha_2} \times \\
&\times [1 + \theta(1 - (1 - \exp(-\lambda_2 t_{2i}))^{\alpha_2}) \times \\
&\times (1 - 2(1 - \exp(-\lambda_1 t_{1i}))^{\alpha_1})]\}
\end{aligned} \tag{78}$$

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

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

$$\begin{aligned}\lambda_j &\sim U(a_j, b_j) \\ \alpha_j &\sim U(c_j, d_j) \\ \theta &\sim U(-1, 1)\end{aligned}\tag{79}$$

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 $\mathbf{v} = (\lambda_1, \lambda_2, \alpha_1, \alpha_2, \theta)'$ é dada por

$$\pi(\mathbf{v}|\mathbf{t}) \propto \pi(\mathbf{v})L(\mathbf{v}|\mathbf{t})\tag{80}$$

em que $\pi(\mathbf{v})$ é a distribuição priori conjunta para $\mathbf{v}$; $L(\mathbf{v}|\mathbf{t})$ é a função de verossimilhança (74), e $\mathbf{t} = (\mathbf{t}_1, \dots, \mathbf{t}_n)$ e $\mathbf{t}_i = (t_{1i}, t_{2i})$, $i = 1, \dots, n$ é o vetor dos dados observados.

Para se obter os sumários a posteriori de interesse, simula-se amostras da distribuição a posteriori conjunta (80) usando métodos MCMC(Monte Carlo em Cadeias de Markov).

5.4 Análise bayesiana na presença de covariáveis

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

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

$$\lambda_{1i} = \gamma_1 \exp(\boldsymbol{\beta}'_1 \mathbf{x}_i) \quad (81)$$

$$\lambda_{2i} = \gamma_2 \exp(\boldsymbol{\beta}'_2 \mathbf{x}_i)$$

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

Considera-se também a presença de dados censurados.

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

$$\begin{aligned} \alpha_j &\sim U(a_j, b_j) \\ \gamma_j &\sim U(c_j, d_j) \\ \theta &\sim U(-1, 1) \\ \beta_{jk} &\sim N(0, g^2) \end{aligned} \quad (82)$$

para $j = 1, 2$; em que $k = 1, \dots, p$ 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.

6 Aplicações

Neste capítulo são apresentados alguns modelos de regressões aplicados a dados reais de câncer, descritos na sessão 1.1. Serão considerados três modelos de regressão, o modelo de regressão Weibull, modelo de regressão gama e o modelo de regressão exponencial generalizado. Todos serão utilizados levando em consideração a presença de covariáveis e censuras.

6.1 Aplicação1: Leucemia Mielóide Aguda

Como descrito anteriormente (ver sessão 1.1.1), nesta aplicação considerou-se um estudo retrospectivo com 189 pacientes com leucemia mielóide aguda. Foram considerados duas covariáveis, GBx1000 e Uréia. Nesta aplicação, assumimos todos os modelos de regressão da seguinte forma

$$\alpha(\mathbf{x}_i) = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}, \quad (83)$$

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

Na Figura 8, observa-se o gráfico de Kaplan Meier juntamente com as distribuições Weibull, gama e exponencial generalizada. Observa-se um bom ajuste da distribuição exponencial generalizada.

Para os modelos foi adotado uma distribuição a priori gama para α (44) com hiperparâmetros dados por $a_\alpha = 2$, $b_\alpha = 1$ e distribuições a priori normais (83) para β_0 , β_1 e β_2 . Para a distribuição Weibull e a distribuição gama, os hiperparâmetros foram $b_0 = 10$, $b_1 = 0,5$ e $b_2 = 0,5$ e para a distribuição exponencial generalizada foi utilizado $b_0 = 10$, $b_1 = 0,01$ e $b_2 = 0,01$.

Os modelos Weibull e exponencial generalizada foram baseados em 20.000 amostras simuladas de Gibbs, com saltos de 20 em 20, para a distribuição a posteriori conjunta para

Figura 8 – Gráfico da função de sobrevivência empírica e ajustadas pelas distribuições exponencial generalizada, gama e Weibull

β_0 , β_1 e β_2 e o modelo gama foi baseado em 20.000 amostras simuladas com saltos de 10 em 10. Foram descartadas as 5.000 primeiras amostras ("burn-in-samples"). A convergência do algoritmo foi observada usando métodos gráficos. Usando o software Winbugs, temos na Tabela 3, os sumários da distribuição a posteriori de interesse.

Da Tabela 3, observa-se para a distribuição exponencial generalizada que as covariáveis GBx1000 e uréia apresentam um efeito significativo nos tempos de sobrevivência, pois zero não está incluído nos respectivos intervalos de credibilidade 95% para β_1 e β_2 . Esta conclusão é equivalente para os modelos utilizando as distribuições Weibull e gama.

Em resumo, as três distribuições levam a resultados parecidos apesar de algumas diferenças nos erros-padrões dos estimadores e consequentemente dos intervalos de credibilidade para os parâmetros.

Para comparar os modelos de regressão utilizados, foi utilizado o critério de discriminação - DIC (21). Na Tabela 4, tem-se os valores do DIC para cada um dos modelos de regressões aplicadas aos dados.

Tabela 3 – Sumários a posteriori

Distribuições	Parâmetro	Média	D. P.	IC 95%
Exp.Generalizada	α	0,570	0,050	(0,480; 0,670)
	β_0	-7,200	0,180	(-7,570; -6,860)
	β_1	0,011	0,001	(0,008; 0,013)
	β_2	0,016	0,002	(0,010; 0,022)
Gama	α	0,667	0,058	(0,559; 0,786)
	β_0	-6,928	0,167	(-7,260; -6,60)
	β_1	0,010	0,001	(0,008; 0,012)
	β_2	0,015	0,002	(0,009; 0,020)
Weibull	α	0,677	0,041	(0,598; 0,759)
	β_0	-4,484	0,308	(-5,09; -3,89)
	β_1	0,007	0,001	(0,005; 0,009)
	β_2	0,012	0,002	(0,007; 0,016)

D. P. é referente ao desvio padrão do modelo

IC 95% é referente ao intervalo de credibilidade com 95%

Tabela 4 – DIC

Distribuições	DIC
Exponencial Generalizada	2.161,95
Gama	2.451,29
Weibull	2.151,02

A partir dos resultados da Tabela 4, observa-se que a distribuição exponencial generalizada e a distribuição de Weibull são melhores modelos para analisar os dados quando comparados com a distribuição Gama (menores valores do DIC). Esses resultados estão de acordo com as curvas exibidas na Figura 8.

É importante salientar que nas aplicações deveríamos usar outras técnicas de discriminação de modelos e também comparar os valores preditos com os valores observados para decidirmos pelo melhor modelo. Na prática é interessante usarmos os gráficos de Kaplan-Meier como dado na Figura 8 juntamente com técnicas existentes de discriminação de modelos.

6.2 Aplicação 2: Dados de Câncer Gástrico

Como descrito anteriormente (ver sessão 1.1.1), neste trabalho foram reunidos dados de 230 prontuários. Foram obtidos dois tempos de sobrevida: o TSL e o TSG. Este con-

junto de dados pode ter interferência de algumas covariáveis como: sexo, idade, tipos de tratamento, finalidade, linfadenectomia, localização e TNM. Nesta aplicação, com o objetivo de fazer uma avaliação do comportamento do modelo multivariado, foi considerado no modelo apenas a covariável que contempla os tipos de tratamentos.

$$\alpha_j(\mathbf{x}_i) = \exp\{\beta_0 + \beta_1 x_{ji}\}$$

para $i = 1, 2, \dots, 230$ (número de indivíduos) e $j = 1, 2$ (TSL e TSG), onde x_{ji} denota a covariável tratamento.

Foi utilizado dois tipos de modelos para a análise destes dados. Em um primeiro momento é utilizado um modelo univariado para cada um dos dois tempos de sobrevida e no outro momento é utilizado o modelo multivariado. Nesta aplicação foram consideradas duas distribuições: exponencial generalizada e Weibull. Com o intuito de aplicar o modelo desenvolvido para a exponencial generalizada, foi utilizada apenas a covariável que denota os tipos de tratamentos.

6.2.1 Aplicação do Modelo Univariado

Na Figura 9 e 10, observa-se o gráfico de Kaplan Meier juntamente com as distribuições Weibull e exponencial generalizada. Nota-se que as distribuições Weibull e exponencial generalizada se ajustam bem aos dados, mas visto o comportamento da sobrevida, seria necessário um modelo mais apurado e que acompanhasse a tendência do Kaplan Meier.

Figura 9 – Gráfico da função de sobrevivência empírica para o TSG e ajustadas pelas distribuições exponencial generalizada e Weibull

Figura 10 – Gráfico da função de sobrevivência empírica para o TSL e ajustadas pelas distribuições exponencial generalizada e Weibull

Nos modelos univariados para ambos os tempos de sobrevida, TSG e TSL, foi adotado uma distribuição a priori gama para α (44) com hiperparâmetros dados por $a_\alpha = 10$, $b_\alpha = 10$ e distribuição a priori normal (55) para o β_0 e β_1 , com hiperparâmetros dados

por $b_0 = 0$ e $b_1 = 0,5$. Para a distribuição a posteriori conjunta os modelos utilizando a distribuição exponencial generalizada foram baseadas em 20.000 amostras simuladas de Gibbs, com saltos de 10 em 10. E para os modelos utilizando a distribuição Weibull foram baseadas em 20.000 amostras simuladas de Gibbs, com saltos de 20 em 20.

Foram descartadas as 5.000 primeiras amostras ("burn-in-samples"). A convergência do algoritmo foi observada usando métodos gráficos.

Usando o software Winbugs, temos na Tabela 5, os sumários a posteriori de interesse para o TSG e na Tabela 7 os sumários a posteriori de interesse para o TSL.

Tabela 5 – Sumários a posteriori para o TSG - Média (IC-95%)

Parâmetros	Exp. Generalizada	Weibull
α	1,40 (1,09; 1,77)	1,05 (0,86; 1,25)
β_0	-3,68 (-4,05; -3,34)	-4,22 (-4,92; -3,56)
β_1 (Químio x Cont)	-0,22 (-0,60; 0,15)	-0,26 (-0,71; 0,17)

Pela Tabela 5, o modelo de regressão utilizando a distribuição exponencial generalizada chegou às mesmas conclusões do modelo de regressão utilizando a distribuição Weibull, ou seja, que não existe nenhuma influência do tratamento no tempo de sobrevida geral dos pacientes (o "zero" está incluso nos intervalos de credibilidade).

Na Tabela 6, é apresentado o valor do DIC para o TSG, a fim de se escolher o melhor modelo de regressão aplicado a este tempo de sobrevida. Através dela, observa-se que o melhor modelo, utilizado para analisar o TSG, é o que utiliza a distribuição exponencial generalizada.

Tabela 6 – Sumários a posteriori para o TSG

Distribuições	DIC
Exponencial Generalizada	761,2
Weibull	765,4

Tabela 7 – Sumários a posteriori para o TSL - Média (IC-95%)

Parâmetros	Exp. Generalizada	Weibull
α	1,40 (1,11; 1,74)	1,08(0,91; 1,27)
β_0	-3,46(-3,80;-3,17)	-4,09(-4,73;-3,48)
β_1 (Químio x Cont)	-0,13(-0,45;0,20)	-0,14(-0,54;0,23)

Pela Tabela 7, obtêm-se as mesmas conclusões tomadas para o modelo que considerou o TSG, ou seja, que não existe nenhum efeito do tratamento no tempo de sobrevida livre dos pacientes (o "zero" está incluso no intervalo de credibilidade).

Tabela 8 – DIC	
Distribuições	DIC
Exponencial Generalizada	911,9
Weibull	916,1

Pela Tabela 8, nota-se que, para o TSL, a regressão exponencial generalizada também foi o melhor modelo aplicado aos dados.

6.2.2 Aplicação do Modelo Multivariado usando Função Cópula

Para o modelo multivariado foram utilizados duas distribuições, a exponencial generalizada e a Weibull.

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 2 para γ_1 e γ_2 , parâmetros da distribuição. Para α_1 e α_2 foi utilizado uma distribuição a priori 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-samples"). A convergência do algoritmo foi observada usando métodos gráficos.

Tabela 9 – 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)

O modelo utilizando a distribuição exponencial generalizada, através da Tabela 9, não identificou efeito do tratamento no tempo de sobrevida.

Tabela 10 – 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)

Observa-se pela Tabela 10, que o modelo utilizando a distribuição Weibull identificou efeito do tratamento no tempo de sobrevida livre de evento.

Tabela 11 – DIC

Distribuições	DIC
Exponencial Generalizada	1.590,8
Weibull	1.596,9

Através da Tabela 11, 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 dependência 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.

7 Considerações Finais

O uso da distribuição exponencial generalizada pode ser uma alternativa para a análise de dados de sobrevivência, dada a grande flexibilidade de ajuste e forma analítica simples para a função de sobrevivência.

Como descrito, essa distribuição apresenta similaridades com a distribuição gama com relação ao comportamento da função de risco, mas na presença de censuras ela se mostra muito eficiente para a obtenção de inferências de interesse. Na função de sobrevivência para a distribuição gama observa-se uma função gama incompleta, o que, em geral, dificulta a obtenção de inferências de interesse e isso não ocorre com a distribuição exponencial generalizada.

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 explorado estes tipos de estudo, como por exemplo, 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 boa 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 por exemplo, Achcar e Santos (2010) e 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).

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 é uma boa 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; BOLETA, 2009).

Pela aplicação considerada com dados de sobrevivência bivariados, conseguimos observar o quanto é fundamental utilizar um parâmetro de dependência entre os tempos de sobrevida. Este modelo, associando os dois tempos de sobrevida, permite aprimorar os resultados para a tomada de decisão.

O uso de métodos bayesianos considerando técnicas de simulação MCMC (Monte Carlo em Cadeias de Markov) é uma boa opção na obtenção de inferências 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.

Este trabalho deixa outras oportunidades de estudo, como por exemplo, explorar a forma da função risco da distribuição exponencial generalizada para ser usada em diferentes conjuntos de dados reais, na tentativa de se conhecer melhor esta função e também achar possíveis ganhos com relação ao ajuste em comparação a outras distribuições comumente usadas na análise de dados de sobrevivência, como a distribuição exponencial, Weibull, log-normal entre várias outras. Mesmo não sendo o objetivo deste trabalho, para os dados multivariados, seria interessante investir em um estudo de especificação de distribuições a priori e considerar outros tipos de função de cópula, a fim de se buscar um modelo apropriado para cada conjunto de dados.

Referências

ABRAMOWITZ, M.; STEGUN, I. A. Handbook of mathematical functions with formulas, graphs, and mathematical tables. New York: Dover Publications, 1970.

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

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

AZEVEDO, G.; MENDONÇA, S. Evolução da mortalidade por câncer de estômago no Estado do Rio de Janeiro: uma comparação entre a região metropolitana e o interior no período de 1979 a 1986. *Cad. Saúde Pública*, Rio de Janeiro, v. 13, p. 79-84, 1997.

BERNARDO, J. M.; SMITH, A. F. M. Bayesian theory. New York: Wiley, 1994.

BORGES, E. C.; CAMARGO, G. C.; SOUZA, M. O.; PONTUAL, N. A.; NOVATO, T. S. Qualidade de vida em pacientes ostomizados: uma comparação entre portadores de câncer colorretal e outras patologias. *Rev Inst Ciênc Saúde*, v.25,n. 4, p.357-363, 2007.

BOX, G. E. P.; TIAO, G. C. Bayesian inference in statistical analysis. Reading: Addison-Wesley, 1973.

BRASIL. Ministério da Saúde. Instituto nacional de câncer. Atlas de mortalidade por câncer no Brasil: 1979-1999. Rio de Janeiro: INCA, 2002.

BRASIL. Ministério da Saúde. Instituto nacional de câncer. Câncer no Brasil: dados dos registros de base populacional. Rio de Janeiro: INCA, 2003.

BUSTAMANTE-TEIXEIRA, M. T.; FAERSTEIN, E.; LATORRE, M. R. Técnicas de análise de sobrevivência. *Cad. Saúde Pública*, Rio de Janeiro, v.18, n. 3, p.579-594, mai-jun, 2002.

CHIB, S.; GREENBERG, E. Understanding The Metropolis-Hasting Algorithm. *Am.*

Stat., Vol. 49, n. 4, 1995.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blücher, 2006.

COX, D.R. Regression Models and Life Tables. *J. R. Stat. Soc.*, Ser. B, v.34, p.187-220, 1972.

COX, D.R.; REID, N. Parameter Orthogonality and Approximate Conditional Inference (with discussion). *J. R. Stat. Soc.*, Ser. B, v.49, p.1-39, 1987.

GELFAND, A. E.; SMITH, F. M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, n. 85, p. 398-409, 1990.

GUERRA, M. R.; GALLO, C. V. M.; MENDONÇA, G. A. S. Risco de câncer no Brasil: tendências e estudos epidemiológicos mais recentes. *Rev. Bras. de Cancerol.*, v.51, n. 3, p.227-234, 2005.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions. *Aust. N. Z. J. Stat.*, v. 41, p. 173-188, 1999.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions: different methods of estimation. *J. Stat. Comput. Simul.*, v. 69, p. 315-338, 2001.

GUPTA, R. D.; KUNDU, D. Generalized exponential distribution: existing results and some recent developments. *J. Stat. Plan Inference*, 2007.

GUPTA, R. D.; KUNDU, D. Generalized exponential distribution; Bayesian inference. *Comput. Stat. Data Anal.*, vol 52, no. 4, p. 1873-1883, 2008.

JÁCOME, A. A. A.; WOHNATH, D. R.; NETO, C. S.; 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*. 2012.

KAPLAN, E.L.; MEIER, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Associ.*, v. 53, p. 457-81, 1958.

KLEIN, J. P.; MOESCHBERGER, M. L. Survival analysis: techniques for censored and truncated data. New York: Springer, 1997.

LAWLESS, J. F. Statistical models and methods for lifetime data. John Wiley e Sons, 1982.

LEE, E. T. Statistical Methods for survival data analysis. 2nd ed. Oklahoma: John Wiley e Sons, 1992.

MAGALHÃES, L. P.; OSHIMA, C. T. F.; SOUZA, L. G.; LIMA, J. M.; CARVALHO, L.; FORONES, N. M. Variação de peso, grau de escolaridade, saneamento básico, etilismo, tabagismo e hábito alimentar pregresso em pacientes com câncer de estômago. *Arq Gastroenterol.* v. 45, n.2, p. 111-116, 2008.

MARTINS, S. L. R.; FALCÃO, R. P. A importância da imunofenotipagem na Leucemia Mielóide Aguda. *Rev. Assoc. Med. Bras.*, São Paulo, v. 46, n. 1, Jan./Mar. 2000.

MENDONCA, N. Leucemia mielóide aguda na criança: como andamos no Brasil?. *J. Pediatria, Rio de Janeiro*, v. 79, n. 6, p. 476-477, 2003.

MENDONCA, G. A. S.; SILVA, A. M.; CAULA, W. M. Características tumorais e sobrevida de cinco anos em pacientes com câncer de mama admitidas no Instituto Nacional de Câncer. Rio de Janeiro, Brasil. *Cad. Saúde Pública*, v. 20, n. 5, p. 1232-1239, 2004.

MORGENSTERN, D.; Einfache Beispiele zweidimensionaler Verteilungen. *Mitteilungsblatt für Mathematische Statistik.* v. 8 p. 234-235, 1956.

NELSEN, R. B. An introduction to copulas. New York. Springer: 1999.

NELSEN, R. An introduction to copulas. New York. 2nd. Springer: 2006.

OLIVEIRA, L. C. O.; ROMANO, L. G. M.; PRADO-JUNIOR, B. P. A.; COVAS, D. T.; REGO, E. M.; SANTIS, G. C. Outcome of acute myeloid leukemia patients with hyperleukocytosis in Brazil. *Med Oncol*, v. 27, p. 1254-1259, 2010.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. Estatística Bayesiana. Lisboa. Fundação Calouste Gulbenkian, 2003.

RAQAB, M. Z. Inferences for generalized exponential distribution based on record statistics. *J. Stat Plan Inference*, v. 104, 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.*, v. 69, p. 109-124, 2001.

SANTOS, R. P. S.; PEREIRA, P. L. V. Modelando contágio financeiro através de cópulas. *Rev. Bras. Finanças*, Rio de Janeiro, Vol. 9, No. 3, 2011.

SARHAN, A. M. Analysis of incomplete, censored data in competing risks models with generalized exponential distributions. *IEEE Trans. Reliab*, 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.; THOMAS, A.; BEST, N. G.; LUND, D. Winbugs user manual. Cambridge: MRC Biostatistics Unit., 2001.

SPIEGELHALTER, D. J., Best, N. J., CARLIN, B. P., LINDE, A. V. D. Bayesian measures of model complexity and fit. *J. Royal Stat. Soc.*, v. 64, 2002.

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

TEIXEIRA, J. B. A.; NOGUEIRA, M. S. Câncer gástrico: fatores de risco em clientes atendidos nos serviços de atenção terciária em um município do interior paulista. *Rev. Latino-Am. Enfermagem*, v. 11, n. 1, p. 43-48, 2003.

TONETO, M. G. Análise imunoistoquímica da expressão do fator tecidual e da densidade microvascular no câncer gástrico: correlação com fatores prognósticos e sobrevida. 100f. Tese (Doutorado em Medicina) - Faculdade de Medicina, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2006.

VIOLA, M. L. L. Tipos de Dependência entre Variáveis Aleatórias e Teoria de Cópulas. 97f. Instituto de Matemática, Estatística e Computação Científica (IMECC-UNICAMP), Campinas, 2009.

WUNSCH FILHO, V.; MONCAU, J. E. Mortalidade por câncer no Brasil 1980-1995: padrões regionais e tendências temporais. *Rev. Assoc. Med. Bras.*, v. 48, n. 3, p. 250-257, 2002.

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

APÊNDICE A - Programas

Nesse apêndice são apresentados os programas computacionais utilizados nesta dissertação de mestrado para a distribuição exponencial generalizada.

A.1 - Modelo Univariado para os dados de Leucemia Mielóide Aguda

A.1.1 - Modelo 1

Listagem 1 – Programa Desenvolvido no software *Winbugs* (Modelo 1).

```

1 model
2 {
3   for ( i in 1:N)
4     {
5       zeros[i] <- 0
6       phi[i] <- -log(L[i])
7       zeros[i] ~ dpois(phi[i])
8
9       L[i]<-exp(( delta[i])*( log(alpha)+log(lambda[i])-lambda[i]*t[i]+(alpha-1)*
10        *log(1-exp(-lambda[i]*(t[i])))))+(1-delta[i])*
11        *log(1-pow(1-exp(-lambda[i]*(t[i])), alpha)))
12
13       lambda[i] <- exp(beta0+beta1*GB.1000[i] +beta2* ureia[i] )
14     }
15   alpha ~ dgamma(2,1)
16   beta0 ~ dnorm(0,0.1)
17   beta1 ~ dnorm(0,10)
18   beta2 ~ dnorm(0,10)
19 }
```

A.2 - Modelo Univariado para os dados de Câncer Gástrico

A.2.1 - Modelo 2 - TSG

Listagem 2 – Programa Desenvolvido no software *Winbugs* (Modelo 2).

```

1  model
2  {
3  for (i in 1:N)
4  {
5    zeros[i] <- 0
6    phi[i] <- -log(L[i])
7    zeros[i] ~ dpois(phi[i])
8
9    L[i] <- exp((delta[i] * (log(alpha) + log(lambda[i]) - lambda[i] * t[i] + (alpha - 1) *
10      * log(1 - exp(-lambda[i] * (t[i]))))) + (1 - delta[i]) *
11      * log(1 - pow(1 - exp(-lambda[i] * (t[i])), alpha))))
12
13    lambda[i] <- exp(beta0 + beta1 * QTRT[i])
14  }
15
16  alpha ~ dgamma(10, 10)
17  beta0 ~ dnorm(0, 1)
18  beta1 ~ dnorm(0, 1)
19  }

```

A.2.2 - Modelo 3 - TSL

Listagem 3 – Programa Desenvolvido no software *Winbugs* (Modelo 3).

```

1  model
2  {
3    for (i in 1:N)
4      {
5        zeros[i] <- 0
6        phi[i] <- -log(L[i])
7        zeros[i] ~ dpois(phi[i])
8
9        L[i] <- exp((delta[i] * (log(alpha) + log(lambda[i]) - lambda[i] * t[i] + (alpha - 1) *
10          * log(1 - exp(-lambda[i] * (t[i]))))) + (1 - delta[i]) *
11          * log(1 - pow(1 - exp(-lambda[i] * (t[i])), alpha))))
12
13        lambda[i] <- exp(beta0 + beta1 * QIRT[i])
14      }
15
16  alpha ~ dgamma(10, 10)
17  beta0 ~ dnorm(0, 1)
18  beta1 ~ dnorm(0, 1)
19 }

```

A.3 - Modelo Multivariado para os dados de Câncer Gástrico

A.3.1 - Modelo 4

Listagem 4 – Programa Desenvolvido no software *Winbugs* (Modelo 4).

```

1 model {
2   for (i in 1:N) {
3     zeros[i] <- 0
4     phi[i] <- -log(L[i])
5     zeros[i] ~ dpois(phi[i])
6
7     lambda1[i] <- gamma1*exp(beta13*QIRT[i])
8
9     lambda2[i] <- gamma2*exp(beta23*QIRT[i])
10
11    a1[i] <- 1-exp(-lambda1[i]*t1[i])
12    a2[i] <- 1-exp(-lambda2[i]*t2[i])
13    b1[i] <- pow(a1[i], alpha1)
14    b2[i] <- pow(a2[i], alpha2)
15    c1[i] <- pow(a1[i], alpha1-1)
16    c2[i] <- pow(a2[i], alpha2-1)
17
18    f1[i] <- alpha1*alpha2*lambda1[i]*lambda2[i]*c1[i]*c2[i]*exp(-lambda1[i]*t1[i]-
19      -lambda2[i]*t2[i])*(1+theta*(1-2*b1[i])*(1-2*b2[i]))
20
21    S11[i] <- alpha1*lambda1[i]*c1[i]*exp(-lambda1[i]*t1[i])*
22      *(1-(b2[i])*(1+theta*(1-b1[i])*(1-2*b2[i])))
23
24    S12[i] <- alpha2*lambda2[i]*c2[i]*exp(-lambda2[i]*t2[i])*
25      *(1-(b1[i])*(1+theta*(1-b1[i])*(1-2*b2[i])))
26
27    S1[i] <- 1-b1[i]-b2[i]+b1[i]*b2[i]*(1+theta*(1-b1[i])*(1-b2[i]))
28
29    L[i] <- exp(delta1[i]*delta2[i]*log(f1[i])+ delta1[i]*(1-delta2[i])*log(S11[i]) +
30      +(1-delta1[i])*delta2[i]*log(S12[i]) + (1- delta1[i])*(1-delta2[i])*log(S1[i]))
31  }
32  k1 <- 59.94
33  gamma1 ~ dunif(0,0.2)
34  gamma2 ~ dunif(0,0.2)
35  theta ~ dunif(-1,1)
36  alpha1 ~ dgamma(15,10)
37  alpha2 ~ dgamma(16,10)
38  beta13 ~ dnorm(0,1)
39  beta23 ~ dnorm(0,1)
40 }

```

APÊNDICE D - Artigo Publicado

Revista Brasileira de Biometria

DISTRIBUIÇÃO EXPONENCIAL GENERALIZADA: USO DE MÉTODOS BAYESIANOS

Jorge Alberto ACHCAR¹
Juliana BOLETA¹

- **RESUMO:** Neste artigo 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

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

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

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:

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

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

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.

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

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

Como um caso especial, assumindo $\lambda = 1$, temos na Figura 1, gráficos da função de risco (3) considerando diferentes valores de α .

Figura 1 - Gráfico da função de risco para vários valores de alfa (lambda=1).

A função geradora de momentos da variável aleatória T com distribuição exponencial generalizada com densidade (1) é dada (ver Gupta e Kundu, 2008) por

$$M(s) = E[e^{sT}] = \frac{\Gamma(\alpha+1) \Gamma\left(1 - \frac{s}{\lambda}\right)}{\Gamma\left(\alpha - \frac{s}{\lambda} + 1\right)}, \quad (5)$$

para $s < \lambda$.

De (5), encontramos os momentos de interesse. A média e variância de T são dadas respectivamente por

$$\begin{aligned} E(T) &= \frac{1}{\lambda} [\psi(\alpha+1) - \psi(1)], \\ Var(T) &= \frac{1}{\lambda^2} [\psi'(1) - \psi'(\alpha+1)], \end{aligned} \quad (6)$$

onde $\psi(\cdot)$ é a função digamma dada por $\psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ e $\Gamma(x)$ é a função gama.

Neste artigo, vamos introduzir uma análise Bayesiana para a distribuição exponencial generalizada usando métodos de simulação de Monte Carlo em Cadeias de Markov (MCMC), como o popular amostrador de Gibbs (ver por exemplo, Casela e George, 1992 ou Gelfand e Smith, 1990) e o algoritmo de Metropolis- Hastings (ver por exemplo, Chib e Greenberg, 1995) para obter os sumários a posteriori de interesse.

O artigo é organizado na forma: na seção 2 apresentamos a função de verossimilhança para dados completos; na seção 3 introduzimos uma análise Bayesiana considerando diferentes distribuições a priori para os parâmetros; na seção 4 consideramos a presença de dados censurados; na seção 5 consideramos a presença de dados censurados e covariáveis; finalmente na seção 6, introduzimos algumas aplicações.

2 Função de verossimilhança para dados completos

Seja $T_1, T_2, \dots, T_n$ 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

$$\begin{aligned} L(\alpha, \lambda) &= \prod_{i=1}^n f(t_i; \alpha, \lambda) \\ &= \alpha^n \lambda^n \left[\prod_{i=1}^n (1 - e^{-\lambda t_i})^{\alpha-1} \right] \exp \left(-\lambda \sum_{i=1}^n t_i \right). \end{aligned} \quad (7)$$

O logaritmo da função de verossimilhança (7) é dado por

$$l(\alpha, \lambda) = \log L(\alpha, \lambda) = n \log \alpha + n \log \lambda - \lambda \sum_{i=1}^n t_i + (\alpha-1) \sum_{i=1}^n \log(1 - e^{-\lambda t_i}). \quad (8)$$

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

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha} + \sum_{i=1}^n \log(1 - e^{-\lambda t_i}) = 0, \quad (9)$$

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

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

$$\hat{\alpha} = -\frac{n}{\sum_{i=1}^n \log[1 - \exp(-\hat{\lambda} t_i)]}. \quad (10)$$

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

$$\frac{n}{\hat{\lambda}} + n\bar{t} + (\hat{\alpha} - 1) \sum_{i=1}^n \frac{t_i e^{-\hat{\lambda} t_i}}{(1 - e^{-\hat{\lambda} t_i})} = 0, \quad (11)$$

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

Observar que precisamos usar em método iterativo para achar o EMV $\hat{\lambda}$ de (11).

As segundas derivadas de $l(\alpha, \lambda)$ são dadas por

$$\begin{aligned} \frac{\partial^2 l}{\partial \alpha^2} &= -\frac{n}{\alpha^2}, \\ \frac{\partial^2 l}{\partial \alpha \partial \lambda} &= \sum_{i=1}^n \frac{t_i e^{-\lambda t_i}}{(1 - e^{-\lambda t_i})}, \\ \frac{\partial^2 l}{\partial \lambda^2} &= -\frac{n}{\lambda^2} - (\alpha - 1) \sum_{i=1}^n \frac{t_i^2 e^{-\lambda t_i}}{(1 - e^{-\lambda t_i})^2}. \end{aligned} \quad (12)$$

Testes de hipóteses e intervalos de confiança para α e λ podem ser obtidos usando a distribuição normal assintótica para $\hat{\alpha}$ e $\hat{\lambda}$, isto é,

$$(\hat{\alpha}, \hat{\lambda}) \overset{a}{\sim} N\{(\alpha, \lambda), I^{-1}\}. \quad (13)$$

onde I é a matriz de informação observada de Fisher dada por

$$I = \begin{pmatrix} E\left(-\frac{\partial^2 l}{\partial \alpha^2}\right) & E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) \\ E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) & E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) \end{pmatrix}_{|(\hat{\alpha}, \hat{\lambda})}. \quad (14)$$

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

$$\Pi_1(\alpha, \lambda) \propto \{\det I(\alpha, \lambda)\}^{1/2}, \quad (15)$$

onde, $I(\alpha, \lambda)$ é a matriz de informação de Fisher esperada para α e λ dada por

$$I(\alpha, \lambda) = \begin{pmatrix} E\left(-\frac{\partial^2 l}{\partial \alpha^2}\right) & E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) \\ E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) & E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) \end{pmatrix}, \quad (16)$$

onde

$$\begin{aligned} E\left(-\frac{\partial^2 l}{\partial \alpha^2}\right) &= \frac{n}{\alpha^2}, \\ E\left(-\frac{\partial^2 l}{\partial \alpha \partial \lambda}\right) &= \frac{n}{\lambda} \left\{ [\psi(\alpha+1) - \psi(1)] - \frac{\alpha}{\alpha-1} [\psi(\alpha) - \psi(1)] \right\}, \\ E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) &= \frac{n}{\lambda^2} \left\{ 1 + \frac{\alpha(\alpha-1)}{\alpha-2} [\psi'(1) - \psi'(\alpha-1) + [\psi(\alpha-1) - \psi(1)]^2] - \right. \\ &\quad \left. - \alpha [\psi'(1) - \psi'(\alpha) + [\psi(\alpha) - \psi(1)]^2] \right\} \end{aligned} \quad (17)$$

(ver Gupta e Kundu, 1999).

Uma possível simplificação é considerar uma priori não-informativa obtida a partir de $\Pi(\alpha, \lambda) = \Pi(\lambda|\alpha)\Pi_0(\alpha)$. Usando a regra de Jeffreys, temos,

$$\Pi_2(\alpha, \lambda) \propto \left\{ E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right) \right\}^{1/2} \Pi_0(\alpha), \quad (18)$$

onde $E\left(-\frac{\partial^2 l}{\partial \lambda^2}\right)$ é dada em (17) e $\Pi_0(\alpha)$ é uma priori não-informativa dada por

$\Pi_0(\alpha) \propto 1/\alpha$, $\alpha > 0$ (uma priori imprópria).

Assim,

$$\Pi_2(\alpha, \lambda) \propto \frac{1}{\alpha \lambda} A^{1/2}(\alpha), \quad (19)$$

onde,

$$A(\alpha) = 1 + \frac{\alpha(\alpha-1)}{\alpha-2} [\psi'(1) - \psi'(\alpha-1) + [\psi(\alpha-1) - \psi(1)]^2] - \alpha [\psi'(1) - \psi'(\alpha) + [\psi(\alpha) - \psi(1)]^2]$$

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

$$\frac{\psi(\alpha) + 1/\alpha - \psi(1)}{\psi(\alpha) - \psi(1)} \rightarrow 1, \quad (20)$$

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

$$\Pi_3(\alpha, \lambda) \propto \frac{1}{\alpha\lambda}. \quad (21)$$

Observar que a priori (21) é uma priori imprópria.

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

$$\Pi_4(\alpha, \lambda) = \Pi_\alpha(\alpha) \Pi_\lambda(\lambda), \quad (22)$$

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

$$\begin{aligned} \Pi_\alpha(\alpha) &\sim \text{Gama}(a_\alpha, b_\alpha), \\ \Pi_\lambda(\lambda) &\sim \text{Gama}(a_\lambda, b_\lambda), \end{aligned} \quad (23)$$

onde $a_\alpha, b_\alpha, a_\lambda$ e b_λ são hiperparâmetros conhecidos e $\text{Gama}(a, b)$ denota uma distribuição Gama com média a/b e variância a/b^2 . Observar que a escolha dos hiperparâmetros das distribuições a priori (23) 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 (7) com uma das distribuições a priori $\Pi_j(\alpha, \lambda)$, $j=1, \dots, 4$ para determinarmos a distribuição a posteriori para α e λ . Assumindo a priori informativa (22), a distribuição a posteriori conjunta para α e λ é dada por

$$\begin{aligned}\Pi(\alpha, \lambda / \mathbf{t}) &\propto \alpha^{n+a_\alpha-1} \lambda^{n+a_\lambda-1} \times \\ &\times \exp \left\{ -\alpha \left[b_\alpha - \sum_{i=1}^n \log(1 - e^{-\lambda t_i}) \right] \right\}, \\ &\times \exp \left\{ -\lambda [b_\lambda + n\bar{t}] \right\}\end{aligned}\quad (24)$$

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

Amostras de distribuição a posteriori (24) podem ser geradas usando métodos de simulação de Monte Carlo em Cadeias de Markov (MCMC). As distribuições condicionais necessárias para o amostrador de Gibbs são dadas por

$$(i) \quad \alpha | \lambda, \mathbf{t} \sim \text{Gama}(n + a_\alpha, b_\alpha - \sum_{i=1}^n \log(1 - e^{-\lambda t_i})); \quad (25)$$

$$(ii) \quad \lambda | \alpha, \mathbf{t} \sim \text{Gama}(n + a_\lambda, b_\lambda + n\bar{t}).$$

Uma grande simplificação na geração de amostras a posteriori (24) é obtida usando o software Winbugs (Spiegelhalter et al, 2001) que só requer a especificação da distribuição conjunta para os dados e as distribuições a priori para os parâmetros do modelo.

É importante salientar que a convergência do amostrador de Gibbs pode melhorar considerando transformações dos parâmetros α e λ . Uma possível reparametrização é dada por $\rho_1 = \log \alpha$ e $\rho_2 = \log \lambda$. Neste caso, assumindo independência a priori entre ρ_1 e ρ_2 , uma priori não informativa para ρ_1 e ρ_2 é dada por

$$\Pi(\rho_1, \rho_2) \propto \text{constante}, \quad (26)$$

onde $-\infty < \rho_1, \rho_2 < \infty$.

Outra possibilidade é considerar uma transformação nos dados. Uma possível transformação é dada por $Y = \log T$, onde o Jacobiano da transformação é dado por

$$\left| \frac{dt}{dy} \right| = e^y.$$

Assim, a densidade para $Y = \log(T)$ é dada por

$$f(y; \alpha, \lambda) = \alpha \lambda \left[1 - \exp(-\lambda e^y) \right]^{\alpha-1} \exp(y - \lambda e^y), \quad (27)$$

onde $-\infty < y < \infty$.

4 Presença de dados censurados

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

$$\delta_i = \begin{cases} 1, & \text{se } t_i = T_i \\ 0, & \text{se } t_i = L_i \end{cases}, \quad (28)$$

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

$$L(\alpha, \lambda) = \prod_{i=1}^n f^{\delta_i}(t_i; \alpha, \lambda) S^{1-\delta_i}(t_i; \alpha, \lambda), \quad (29)$$

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,

$$L(\alpha, \lambda) = \alpha^r \lambda^r \exp\left\{(\alpha-1) \sum_{i=1}^n \delta_i \log(1 - e^{-\lambda t_i}) - \lambda \sum_{i=1}^n \delta_i t_i + \sum_{i=1}^n (1-\delta_i) \log\left[1 - (1 - e^{-\lambda t_i})^\alpha\right]\right\}, \quad (30)$$

onde $r = \sum_{i=1}^n \delta_i$ (número de observações completas).

O logaritmo da função de verossimilhança (30) é dado por

$$l(\alpha, \lambda) = r \log \alpha + r \log \lambda + (\alpha-1) \sum_{i=1}^n \delta_i \log(1 - e^{-\lambda t_i}) - \lambda \sum_{i=1}^n \delta_i t_i + \sum_{i=1}^n (1-\delta_i) \log\left[1 - (1 - e^{-\lambda t_i})^\alpha\right]. \quad (31)$$

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

$$\lambda(\mathbf{x}_i) = e^{\beta \mathbf{x}_i'}, \quad (32)$$

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

Assumindo dados de sobrevivência na presença de covariáveis e dados censurados, a função de verossimilhança para α e β é dada por

$$L(\alpha, \beta) = \prod_{i=1}^n f^{\delta_i}(t_i / \alpha, \beta, \mathbf{x}_i) S^{1-\delta_i}(t_i / \alpha, \beta, \mathbf{x}_i), \quad (33)$$

onde δ_i é uma função indicadora de censuras dada em (28), $f(t_i / \alpha, \beta, \mathbf{x}_i)$ e $S(t_i / \alpha, \beta, \mathbf{x}_i)$ são dados, respectivamente em (1) e (2) com $\lambda(\mathbf{x}_i)$ definido por (32) em lugar de λ .

Dessa forma, a função de verossimilhança (33) é dada por

$$\begin{aligned} L(\alpha, \beta) = & \alpha^r \exp \left\{ \sum_{i=1}^n \delta_i \beta' \mathbf{x}_i - \sum_{i=1}^n \delta_i t_i e^{\beta' \mathbf{x}_i} + \right. \\ & + (\alpha - 1) \sum_{i=1}^n \delta_i \log(1 - e^{-t_i e^{\beta' \mathbf{x}_i}}) + \\ & \left. + \sum_{i=1}^n (1 - \delta_i) \log \left[1 - (1 - e^{-t_i e^{\beta' \mathbf{x}_i}})^\alpha \right] \right\}, \end{aligned} \quad (34)$$

onde r é o número de observações completas definido em (30).

Para uma análise Bayesiana do modelo, assumimos uma distribuição a priori gama para α (ver (23)) com hiperparâmetros conhecidos a_α e b_α e distribuições a priori normais para $\beta_l, l = 0, 1, \dots, k$, isto é,

$$\beta_l \sim N(0; b_l^2), \quad (35)$$

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

A distribuição a posteriori conjunta para α e β é dada por

$$\begin{aligned} \Pi(\alpha, \beta | \mathbf{t}, \mathbf{x}) \propto & \alpha^{r+a_\alpha-1} e^{-b_\alpha \alpha} \times \left(\prod_{i=1}^n \exp \left[-\frac{\beta_i^2}{2b_i^2} \right] \right) \times \\ & \times \exp \left\{ \sum_{i=1}^n \delta_i \beta' \mathbf{x}_i - \sum_{i=1}^n \delta_i t_i e^{\beta' \mathbf{x}_i} + (\alpha - 1) \sum_{i=1}^n \delta_i \log(1 - e^{-t_i e^{\beta' \mathbf{x}_i}}) + \right. \\ & \left. + \sum_{i=1}^n (1 - \delta_i) \log \left[1 - (1 - e^{-t_i e^{\beta' \mathbf{x}_i}})^\alpha \right] \right\}. \end{aligned} \quad (36)$$

Amostras da distribuição a posteriori conjunta para α e β são geradas usando métodos MCMC.

6 Aplicações

6.1 Um exemplo com dados completos

Consideremos um conjunto de dados relativos à durabilidade de um componente (dados introduzidos em Lawless, 1982, pág. 228): 17,88; 28,92; 33,0; 41,52; 42,12; 45,60; 48,80; 51,84; 51,96; 54,12; 55,56; 67,80; 68,64; 68,64; 68,88; 84,12; 93,12; 98,64; 105,12; 105,84; 127,92; 128,04 e 173,40. Os dados representam os números (milhões) de ciclos até a falha do componente.

Assumir uma distribuição exponencial generalizada com densidade (1) para analisar os dados. Os estimadores de máxima verossimilhança para α e λ (ver (10) e (11)) são dados por $\hat{\alpha} = 5,28$ e $\hat{\lambda} = 0,03$.

Para uma análise Bayesiana do modelo, assumimos as distribuições a priori gama (23) para α e λ com $a_\alpha = b_\alpha = 0,01$ e $a_\lambda = b_\lambda = 0,1$.

Usando o software Winbugs, geramos inicialmente 5000 amostras (“burn-in-samples”) da distribuição a posteriori conjunta para α e λ (24) que foram descartadas para eliminar o efeito dos valores iniciais para os parâmetros α e λ usados no processo iterativo e a seguir geramos outras 2000 amostras escolhidas de 10 em 10 para se ter amostras aproximadamente não correlacionadas.

A convergência do algoritmo foi verificada a partir de gráficos de séries temporais das amostras simuladas.

Os sumários de interesse considerando as 1000 amostras simuladas de Gibbs são dados na Tabela 1.

Tabela 1 - Sumários a posteriori

Distribuições	Parâmetro	Média	DP	Intervalo de Credibilidade 95%
Exp. Generalizada	α	5,230	2,054	(2,279; 10,000)
	λ	0,031	0,006	(0,019; 0,044)
Weibull	α	2,035	0,3157	(1,472; 2,678)
	λ	3,095E-4	5,147E-4	(6,734E-6; 0,001)
Gama	α	3,953	1,131	(2,099; 6,39)
	λ	0,054	0,016	(0,026; 0,090)

Na Figura 2, temos os gráficos das funções de sobrevivência empírica e ajustada pela distribuição exponencial generalizada (2) com os estimadores Bayesianos $\tilde{\alpha} = 5,23$ e $\tilde{\lambda} = 0,031$. Da Figura 2, observamos um bom ajuste da distribuição exponencial generalizada para os dados.

Figura 2 - Gráfico da função de sobrevivência empírica e ajustada pelas distribuições Exponencial Generalizada, Gama e Weibull.

Na Figura 2, também temos os gráficos das funções de sobrevivência ajustadas pelas distribuições de Weibull e Gama assumindo distribuições a priori aproximadamente não-informativas (os sumários a posteriori para as três distribuições são apresentados na Tabela 1).

6.2 Um exemplo com dados censurados

No estudo feito neste exemplo, são apresentados na Tabela 2, os tempos em dias até a ocorrência dos primeiros sinais de alterações indesejadas no estado geral de 45 pacientes de ambos os sexos que receberam tratamento quimioterápico após terem sido submetidos à cirurgia do intestino (dados introduzidos em Colosimo e Giolo, 2006, pág. 76). Foi registrado um total de 250 dias desde a entrada do primeiro paciente até o término do estudo.

Tabela 2 - Dados de cirurgia de intestino

7, 8, 10, 12, 13, 14+, 19, 23, 25+, 26, 27, 31, 31+, 49, 59+, 64+,
 87, 89, 107, 117, 119, 130, 148, 153, 156, 159, 191, 222, 200+,
 203+, 210+, 220+, 220+, 228+, 230+, 233+, 235+, 240+, 240+,
 240+, 241+, 245+, 247+, 248+, 250+

(+ indica censura)

Assumindo a distribuição exponencial generalizada, consideramos as distribuições a priori gama (23) para α e λ com $a_\alpha = b_\alpha = a_\lambda = b_\lambda = 0,1$. Usando o software Winbugs, temos na Tabela 3, os sumários a posteriori de interesse baseados em 1000 amostras simuladas de Gibbs da distribuição a posteriori para α e λ após um descarte (“burn-in-samples”) de 5000 amostras.

Tabela 3 - Sumários a posteriori (dados com censuras)

Distribuição	Parâmetro	Média	D.P	Intervalo de credibilidade 95%
Exp.	α	0,8028	0,1913	(0,4808; 1,2380)
	λ	0,0029	0,0010	(0,0010; 0,0053)
Weibull	α	0,8268	0,1500	(0,5614; 1,135)
	λ	0,0122	0,0098	(0,0016; 0,0393)
Gama	α	0,7955	0,1807	(0,4817; 1,1880)
	λ	0,0027	0,0011	(9,019E-4; 0,0053)

Na Figura 3, temos os gráficos das funções de sobrevivência estimada pelo estimador não-paramétrico de Kaplan e Meier e ajustada pela distribuição exponencial generalizada com os estimadores Bayesianos $\tilde{\alpha} = 0,8028$ e $\tilde{\lambda} = 0,00295$. É interessante observar que os estimadores de máxima verossimilhança para α e λ são dados respectivamente por, $\hat{\alpha} = 0,81$ e $\hat{\lambda} = 0,003$.

A partir dos gráficos da Figura 3, observamos bom ajuste da distribuição exponencial generalizada para os dados. Também temos na Figura 3, os gráficos das funções de sobrevivência ajustadas pelas distribuições Weibull e Gama usando distribuições a priori aproximadamente não-informativas (os sumários a posteriori são apresentados na Tabela 3).

Figura 2 - Gráfico da função de sobrevivência empírica e ajustadas pelas distribuições exponencial generalizada, gama e Weibull (dados com censuras).

6.3 Um exemplo com dados censurados e presença de covariáveis

Neste exemplo consideramos um estudo retrospectivo com 190 pacientes com leucemia mieloide 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 (32)) dado por

$$\lambda(\mathbf{x}_i) = e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}}, \quad (37)$$

para $i=1, \dots, 189$ (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 (ver (23)) com hiperparâmetros dados por $a_\alpha = 2$, $b_\alpha = 1$ e distribuições a priori normais (37) 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. Generalizada	α	3,086	0,7686	(1,870; 5,044)
	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 GBx1000e uréia não apresentam um efeito significativo nos tempos de sobrevivência, pois zero está incluído no intervalo de credibilidade 95% para β_1 e β_2 , respectivamente, assumindo a distribuição exponencial generalizada.

Notas conclusivas

O uso da distribuição exponencial generalizada pode ser uma boa alternativa para a análise de dados de sobrevivência, dada 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.

O uso de métodos Bayesianos combinados com métodos de simulação MCMC (Monte Carlo em cadeias de Markov) para gerar amostras da distribuição a posteriori de interesse é uma boa alternativa para a obtenção dos sumários a posteriori de interesse, especialmente usando o software Winbugs.

Agradecimentos

Os autores são extremamente agradecidos às sugestões propostas pelos dois revisores desse artigo que levaram a uma grande melhoria e correções de uma revisão prévia do artigo.

ACHCAR, J. A.; BOLETA, J. Generalized exponential distribution: use of Bayesian methods. *Rev. Bras. Biom.*, São Paulo, v.27, n.4, p.644-658, 2009.

- **ABSTRACT:** In this paper, we introduce the generalized exponential distribution as an alternative for some existing lifetime distributions used in lifetime data analysis. We obtain inferences on the parameters of the model considering complete lifetime data, censoring data and

the presence of covariates, under the classical and Bayesian approaches, and related to the Bayesian approach, we assume different prior distributions for the parameters. As illustration, we considered some applications using MCMC (Markov Chain Monte Carlo) simulation methods and the software Winbugs.

- **KEYWORDS:** Generalized exponential distribution; censored lifetime data; Bayesian analysis; MCMC simulation methods.

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.
- CASELA, G.; GEORGE, E. I. Explaining the Gibbs sampler. *Am. Stat.* Washington, v. 46, n. 3, p.167-74, 1992.
- CHIB, S.; GREENBERG, E. Understanding the metropolis-hasting algorithm. *Am. Stat.* Washington, v.49, n.4, 1995.
- COLOSIMO, E.A.; GIOLO, S.R. *Análise de sobrevivência aplicada*. São Paulo: Edgard Blücher, 2006.
- COX, D.R.; REID, N. Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Stat. Soc., Ser. B*, London, v.49, n.1, p.1-39, 1987.
- GELFAND, A. E.; SMITH, F. M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.*, new York, v.85, p.398-409, 1990.
- GUPTA, R. D.; KUNDU, D. Generalized exponential distributions. *Aust. N.Z. J. Stat.*, Oxford, v.41, p.173-188, 1999.
- GUPTA, R. D.; KUNDU, D. Generalized exponential distributions: different methods of estimation. *J. Stat. Comp. Simul.*, v.69, p.315-338, 2001.
- GUPTA, R. D.; KUNDU, D. Generalized exponential distribution:existing results and some recent developments. *J. Stat. Plann. Infer.*, doi: 10.1016/j.jspi.2007.03.030, 2007.
- GUPTA, R. D.; KUNDU, D. Generalized exponential distribution. Bayesian inference. *Comput. Stat. Data Anal.*, v.52, n.4, 1873-1883, 2008.
- LAWLESS, J. F. *Statistical models and methods for lifetime data*. Canadá: John Wiley & Sons, 1982.
- RAQAB, M. Z. Inferences for generalized exponential distribution based on record statistics. *J. Stat. Plann. Infer.*, v.104, 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.*, v.69, p.109-124, 2001.
- SARHAN, A. M. Analysis of incomplete, censored data in competing risks models with generalized exponential distributions. *IEEE Trans. Reliab.*, v.56, p.132-138, 2007.

SPIEGELHALTER, D. J.; THOMAS, A.; BEST, N. G.; LUND, D. *Winbugs user manual*. Cambridge: MRC Biostatistics Unit., 2001.

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

Recebido em 26.08.2009.

Aprovado após revisão 11.12.2009.

ANEXO A - Comitê de Ética em Pesquisa

HOSPITAL DAS CLÍNICAS DA FACULDADE DE MEDICINA
DE RIBEIRÃO PRETO DA UNIVERSIDADE DE SÃO PAULO

www.hcrp.usp.br

Ribeirão Preto, 25 de maio de 2011

Ofício nº 1833/2011
CEP/MGV

Prezados Senhores,

O trabalho intitulado **“DISTRIBUIÇÃO EXPONENCIAL GENERALIZADA: UMA APLICAÇÃO A DADOS DE CÂNCER”** foi analisado pelo Comitê de Ética em Pesquisa, em sua 324ª Reunião Ordinária realizada em 23/05/2011 e enquadrado na categoria: **APROVADO**, de acordo com o Processo HCRP nº 4291/2011.

Este Comitê segue integralmente a Conferência Internacional de Harmonização de Boas Práticas Clínicas (IGH-GCP), bem como a Resolução nº 196/96 CNS/MS.

Lembramos que devem ser apresentados a este CEP, o Relatório Parcial e o Relatório Final da pesquisa.

Atenciosamente.

DRª MARCIA GUIMARÃES VILLANOVA
Coordenadora do Comitê de Ética em
Pesquisa do HCRP e da FMRP-USP

Ilustríssimos Senhores
JULIANA BOLETA
JORGE ALBERTO ACHCAR
Depto. de Medicina Social