Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

MODELOS PARA DADOS DE SOBREVIVÊNCIA NA PRESENÇA DE DIFERENTES ESQUEMAS DE ATIVAÇÃO BASEADOS NA DISTRIBUIÇÃO GEOMÉTRICA

Mari Roman

São Carlos - SP
Maio/2013

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

Modelos para dados de sobrevivência na presença de diferentes esquemas de ativação baseados na distribuição Geométrica

Mari Roman

Orientador: Prof. Dr. Francisco Louzada
Co-orientador: Prof. Dr. Vicente Garibay Cancho

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de São Carlos PPGEs/UFSCar, como parte dos requisitos necessários para obtenção do título de Doutor em Estatística.

São Carlos - SP
Maio/2013

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

R758md

Roman, Mari.
Modelos para dados de sobrevivência na presença de diferentes esquemas de ativação baseados na distribuição geométrica / Mari Roman. -- São Carlos : UFSCar, 2013.
118 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2013.

1. Análise de sobrevivência. 2. Variáveis latentes. 3. Modelo de mistura. 4. Distribuição geométrica exponencial. 5. Distribuição gama generalizada. I. Título.

CDD: 519.9 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 3351-8292 – Email: ppgest@ufscar.br
13565-905 - SÃO CARLOS-SP - BRASIL

FOLHA DE APROVAÇÃO

Aluno(a) : Mari Roman

TESE DE DOUTORADO DEFENDIDA E APROVADA EM 08/04/2013 PELA COMISSÃO JULGADORA:

Presidente _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP/Orientador)

1º Examinador _____
Prof. Dr. Heleno Bolfarine (IME-USP)

2º Examinador _____
Prof. Dr. Josmar Mazucheli (UEM)

3º Examinador _____
Profa. Dra. Maria Aparecida de Paiva Franco (DEs-UFSCar)

4º Examinador _____
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP)

Resumo

Nesta tese, novas famílias de distribuições são propostas para modelar dados de tempo de vida. Essas distribuições são obtidas assumindo que a ocorrência do evento de interesse é explicada por uma estrutura latente de ativação. Em geral, as causas competitivas podem ter diferentes mecanismos de ativação, consideramos os casos: primeiro, último e aleatório. A presença de fração de curados é considerada nestes contextos. Os modelos assumem que o número de causas de risco tem distribuição de probabilidade Geométrica; e o tempo de ativação desses fatores segue distribuição Exponencial ou Gama Generalizada. Propriedades das distribuições propostas são discutidas, incluindo obtenção da função densidade de probabilidade e fórmulas explícitas da função de risco, momentos, estatística de ordem e valor modal. Outro objetivo deste trabalho é o desenvolvimento de processos inferenciais nas perspectivas clássica e bayesiana. Além disso, as medidas bayesianas de diagnóstico baseadas na divergência ψ, que incluem a divergência de Kulback–Leibler como caso particular, são consideradas para detectar observações influentes. Estudos de simulação são realizados e resultados experimentais são obtidos para conjuntos de dados reais.

Abstract

In this thesis new families of survival distributions are proposed. Those distributions are derived by assuming a latent activation structure to explain the occurrence of the event of interest. In general, the competitive causes may have different activation mechanisms. Here we assume three different ones, namely, fisrt, random and last actvation mechanisms. The presence of cure fraction are also addressed in two contexts. The models assumed that the number of causes follows a Geometric distribution and the lifetime for these causes follows an Exponential distribution, and a Gamma Generalized distribution. The properties of the proposed distributions are discussed, including a formal proof of its probability density function and explicit algebraic formulas for its reliability and failure rate functions, moments, order statistics and modal value. Inferetial procedure is based on frequentist and Bayesian perspectives. Moreover, Bayesian case influence diagnostics based in ψ-divergence, with include Kulback–Leibler divergence measure as a particular case, are developed. Simulation studies are performed and experimental results are illustrated based in real datasets.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A Análise de Sobrevivência (AS) consiste em uma importante área da ciência Estatística, sua aplicabilidade é ampla e diversificada, tendo como principal elemento o tempo que decorre entre um instante inicial e a ocorrência de determinado evento de interesse.

Ela está fortemente ligada às ciências biomédicas, que exercem, em princípio, grande influência na terminologia utilizada. Resultados dessas ciências podem, por exemplo, embasar campanhas governamentais tais como de prevenção ao câncer de mama, que incentiva e convoca com solicitude à mamografia anual mulheres acima de 40 anos, visando o diagnóstico precoce do câncer, e auxilia na escolha de tratamentos, comparando a eficácia destes em diferentes períodos da doença, considerando fatores que possam influenciar no processo.

No entanto, essa não é a única área que se beneficia do conjunto de ferramentas estatísticas compreendido na AS. Como afirma Papoila (2011), sua aplicação vai desde a demografia, em que o interesse pode recair na análise da duração de casamentos, passando pela Indústria, onde é importante estudar o tempo até a falha de determinados equipamentos ou componentes eletrônicos (teoria da confiabilidade), adentrando a Economia e as Ciências Atuariais. Além disso, em muitas situações ela é a única abordagem possível em decorrência do tipo de variáveis envolvidas e as consequentes limitações no seu uso.

Buscando contribuir com essa grande quantidade de possíveis aplicações e resultados que movem e norteiam tantas áreas de conhecimento, esta tese de doutoramento tem

por objetivo principal a formulação de novas distribuições de probabilidade para dados de sobrevivência, com a utilização de estruturas que traduzam situações reais.

1.1 Ativação por múltiplas causas

Muitas distribuições se baseiam no pressuposto de que apenas uma causa é responsável pela ocorrência do evento de interesse. Todavia, cada vez mais a etiologia - ciência das causas - está ganhando espaço e agregando áreas como a biologia, a criminologia, a psicologia, a medicina e várias outras ciências, que buscam, em suas pesquisas, as causas que deram origem ao seu objeto de estudos.

Minayo (1988) afirma que a etiologia de doenças indica que não existe apenas um único causador para as enfermidades e ainda observa que é consideravelmente difícil definir qual o causador mais importante e quais são irrelevantes. Este fato conduz à polêmicas no meio científico e são grandes as interrogações, por exemplo sobre os agentes etiológicos implicados no aparecimento da maioria das neoplasias malignas.

Enquanto as ciências médicas rumam no sentido de desvendar as possíveis causas das doenças, a modelagem estatística precisa encontrar formas de introduzir essas informações nos modelos estatísticos propostos, acomodando essas novas situações. No contexto de análise de sobrevivência, esses dados serão tempos até a ocorrência do evento de interesse.

Neste contexto, para introduzir a ideia que será abordada utilizamos uma situação da área das ciências médicas. Consideremos que o evento de interesse é a ocorrência de certa neoplasia que está associada a quantidade M de fatores de risco. Porém, pode-se desconhecer o valor de M e conhecer tampouco quantos ou qual fator de risco acionou a neoplasia. Como suposição razoável, considera-se que M seja uma variável aleatória (v.a.) latente e $M > 0$.

Quanto à ocorrência do evento de interesse, uma possibilidade é supor que sua ativação é concomitante com a do primeiro fator de risco e assim o tempo observado, $W = w$, refere-se ao mínimo dos tempos de ocorrência de cada um dos fatores de risco envolvidos. Vários autores consideraram esta suposição, dentre os quais: Goetghebeur &

Ryan (1995), Reiser *et al.* (1995), Adamidis & Loukas (1998), Louzada-Neto (1999), Lu & Tsiatis (2001) e Lu & Tsiatis (2005).

Outra possibilidade é considerarmos que a ocorrência da neoplasia é ocasionada pela ativação do último fator de risco envolvido, sendo necessário que todos os fatores de risco sejam ativados para que a neoplasia seja detectada. Sendo T_i a v.a. que denota o tempo de ativação de cada um dos fatores de risco, o índice i estando associado ao valor que a v.a. M assume, o tempo observado refere-se ao máximo dos tempos dentre os tempos T_i. Neste enfoque Cooner *et al.* (2006), Kus (2007) e Cancho *et al.* (2010a) são referências.

No entanto, ainda persiste a possibilidade do evento de interesse ter sido desencadeado por uma situação intermediaria à estas apresentadas, como abordam Cooner *et al.* (2006). Neste caso, o tempo observado refere-se ao tempo entre o mínimo e máximo, que é denominada de ativação aleatória.

Considerando a suposição de que a v.a. latente M que denota a quantidade de causas de risco (CR) assume somente valores positivos, impomos a condição de que toda a população está em risco. Entretanto há a possibilidade de que a neoplasia não ocorra em alguns indivíduos, fato este que inclui a informação de censura do tempo à ser observado, um dos diferenciais da análise de sobrevivência.

Uma particularidade ocorre quando a maioria das observações censuradas corresponde aos valores mais elevados, sendo esse um dos indícios que há proporção de curados na população. Este fato traduz-se graficamente na estabilização da curva da estimativa de Kaplan-Meier (KM) da função de sobrevivência, como apresentado na Figura 1.1. Neste contexto, modelos de longa duração são imprescindíveis.

Nesta abordagem, suspeita-se da existência de indivíduos que nunca vão experimentar o evento de interesse, ainda que o tempo de observação pudesse ser prolongado indefinidamente. Tais indivíduos são considerados imunes ao acontecimento de interesse. Neste âmbito, a população em estudo é constituída por uma mistura de indivíduos imunes (curados ou não suscetíveis) e de indivíduos não imunes (doentes ou suscetíveis). Esta divisão na população foi considerada em estudos como os apresentados por: Maller & Zhou (1996), Yakovlev & Tsodikov (1996), Chen & Ibrahim (2001), Ibrahim *et al.* (2001) e Rodrigues *et al.* (2009c).

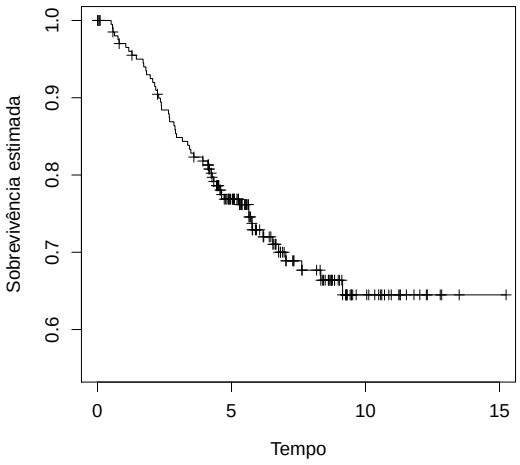

FIGURA 1.1: Gráfico das estimativas da curva de Kaplan-Meier para o valor da função de sobrevivência, dando indicativo de longa duração.

Um dos trabalhos precursores nessa área é o modelo de mistura de Berkson e Gage (1958). Cooner *et al.* (2006) e Cooner *et al.* (2007) abordam que a quantidade M de fatores de risco pode assumir o valor nulo, e neste caso estamos admitindo que na população poderá haver indivíduos que jamais experimentarão o evento de interesse. Nessa nova situação para M pode-se considerar novamente as três formas de ativação das CRs, apresentadas anteriormente, o que resultará em três modelos de longa duração.

1.2 Apresentação dos capítulos

Esta tese está assim organizada. No Capítulo 2 apresentamos a formulação das distribuições assumindo que a v.a. latente M tem distribuição de probabilidade Geométrica e considerando $M = 1, 2, \cdots, m$ (situação em que o evento de interesse sempre acontece) e $M = 0, 1, \cdots, m$ (situação em que o evento de interesse pode não ocorrer). Três formas de ativação são consideradas, quando somente o tempo de sobrevivência mínimo é observado, quando somente o tempo de sobrevivência máximo é observado e quando o tempo de sobrevivência é um valor aleatório entre os tempos mínimo e máximo. A v.a. que denota o tempo de ativação de cada CR, não tem sua distribuição especificada neste capítulo, pois o intuito principal do capítulo é apresentar como são obtidas as funções densidade, sobrevivência e risco de cada uma das distribuições decorrentes da composição,

considerando a distribuição de probabilidade Geométrica para a v.a. M.

Utilizando a formulação do Capítulo 2 e atribuindo uma distribuição para a v.a. T_i temos distribuições específicas que permitem avaliações das formas da funções de densidade de probabilidade, sobrevivência e risco bem como propriedades e aplicações em conjuntos de dados reais ou artificiais, tais considerações estão apresentadas nos Capítulos 3, 4, 5 e 6. Nestes quatro capítulos não expomos as provas de como as distribuições foram obtidas (já apresentadas no Capítulo 2) mas abordamos aspectos mais práticos e direcionados a cada uma das distribuições. Ao leitor que visa o caráter aplicado das distribuições a leitura do Capítulo 2 pode ser desnecessária sendo suficiente a leitura dos demais.

No Capítulo 3 apresentamos a distribuição Exponencial Geométrica (EGcr) para os esquemas de ativação mínimo e máximo. Atenção especial é atribuída a distribuição Exponencial Geométrica para o máximo dos tempos (EG_{mx}). Propriedades, estudo de simulação e aplicação são abordados amplamente neste capítulo. Tais resultados estão sumarizados no artigo Louzada *et al.* (2011). O Capítulo 4 é uma extensão do Capítulo 3, no qual contempla-se o grupo de indivíduos imunes da distribuição EGcr considerando a modelagem de longa duração proposta por Berkson e Gage (1952). Propriedades e aplicações à conjuntos de dados são apresentados bem como a comparação com outros modelos de longa duração. Estes resultados estão condensados nos artigos Louzada *et al.* (2012) e Roman *et al.* (2012). O Capítulo 5 é independente dos Capítulos 3 e 4 e apresenta a distribuição para o grupo de suscetíveis e imunes, quando os tempos T_i possuem distribuição Gama Generalizada. São abordados os três esquemas de ativação. Propriedades, casos particulares e aplicações em dados reais também são apresentadas. Como continuação, utilizamos a inferência bayesiana para o diagnóstico de ponto influente, com simulação e análise em dados reais, no Capítulo 6.

No Capítulo 7 sumarizamos as contribuições da tese, apresentamos uma generalização dos conceito utilizados, permitindo visualizar novas distribuições que podem ser obtidas. Pontuamos outras questões que podem ser contempladas em futuros trabalhos. No Apêndice apresentamos gráficos e resultados omitidos do texto principal.

Na Seção 1.3 deste capítulo introdutório apresentamos os conjuntos de dados utilizados na tese.

1.3 Conjuntos de dados

A seguir apresentamos a descrição dos conjuntos de dados utilizados ao longo do trabalho, com seus respectivos gráficos Kaplan-Meier (Kaplan & Meier, 1958) e TTT para avaliar a forma de função de risco.

Para obter o gráfico TTT consideramos $r = 1, \ldots, n$ e $Y_{r:n}$ as estatísticas de ordem da amostra. Para os conjuntos de dados não censurados utilizamos a proposta de Aarset (1985) em que $G(r/n) = \dfrac{(\sum_{i=1}^{r} Y_{i:n}) + (n-r)Y_{r:n}}{\sum_{i=1}^{r} Y_{i:n}}$. Para os conjunto de dados com observações censuradas utilizamos a versão apresentada em Rinne (2009), em que $G(r/n) = \sum_{i=1}^{r} (n-i+1)(Y_{i:n} - Y_{(i-1):n})$ e define-se r^* como o indicador de ordem das observações não censuradas, $r^* = 1, 2, \cdots, n^*$. Selecionando os n^* valores de $G(r/n)$ que correspondem as observações não censuradas, que são denotadas por $G(r^*/n^*)$, o gráfico TTT corresponde a relação gráfica de r^*/n^* $versus$ TTT_{r^*}, em que $TTT_{r^*} = \dfrac{G(r^*/n^*)}{G(n^*/n^*)}$.

A forma da função de risco dos dados é classificada com base no comportamento da curva do gráfico TTT (Aarset, 1987) e a evidência de longa duração é baseada no comportamento da curva das estimativas da função sobrevivência pelo estimador de Kaplan-Meier. Parametricamente, dispondo do modelo de sobrevivência, a evidência de longa duração pode ser testada via teste de hipótese, comparando os valores da verossimilhança dos modelos utilizando os parâmetros ajustados, como apresentam Maller & Zhou (1996).

$T1$: **Resistência de esferas**

O conjunto $T1$ foi extraído de Lawless (2003), os dados correspondem ao resultado de um teste de resistência à rotação de 23 esferas e o valor observado é o número de milhões de revoluções antes de ocorrer a falha na rotação. O número de revoluções é considerado o tempo de vida e varia de 18,88 a 173,40 milhões, com média igual a 72,23 milhões e desvio padrão $s = 37,47$ milhões. Não há presença de censura nessa amostra e os dados apresentam função de risco crescente (Figura 1.2).

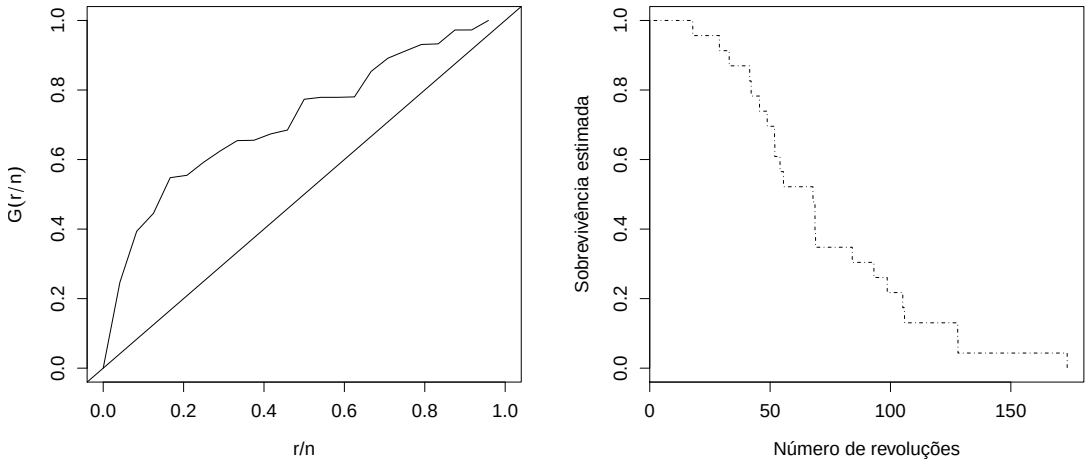

FIGURA 1.2: Gráficos TTT e curva de Kaplan-Meier para o conjunto de dados $T1$.

$T2$: **Dados de reversão sorológica de HIV em crianças**

O segundo conjunto de dados ($T2$) refere-se ao tempo até a reversão sorológica de 143 crianças que tiveram exposição ao HIV via vertical, o que ocorre ainda na gestação. Esses dados foram colhidos no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto entre os anos de 1986 e 2001 e utilizados por Perdoná (2006).

O tempo de reversão sorológica (ou sororeversão) varia de 2 a 1.021 dias, sendo 50% deles menores ou iguais a 446 dias. Das 143 crianças acompanhadas, 24 foram censuradas por algum motivo. Sendo que o paciente ainda não apresentou o evento de interesse (sororeversão) à medida que o tempo passa, o risco dele apresentar o evento de interesse aumenta, conforme podemos observar na gráfico TTT da Figura 1.3.

$T3$: **Falha no sistema de ar condicionado**

O conjunto de dados ($T3$) consiste no número de falhas sucessivas no sistema de ar condicionado para cada elemento da frota de 13 *boeing* e 720 aviões a jato. Em aviões que voam em grandes altitudes o sistema de ar-condicionado é imprescindível e permanece ligado durante todo o voo a fim de manter a temperatura e a pressurização do avião.

Os dados consistem em 214 observações não censuradas e foram utilizados por Adamidis & Loukas (1998). O conjunto original tem 217 dados, sendo 3 censurados,

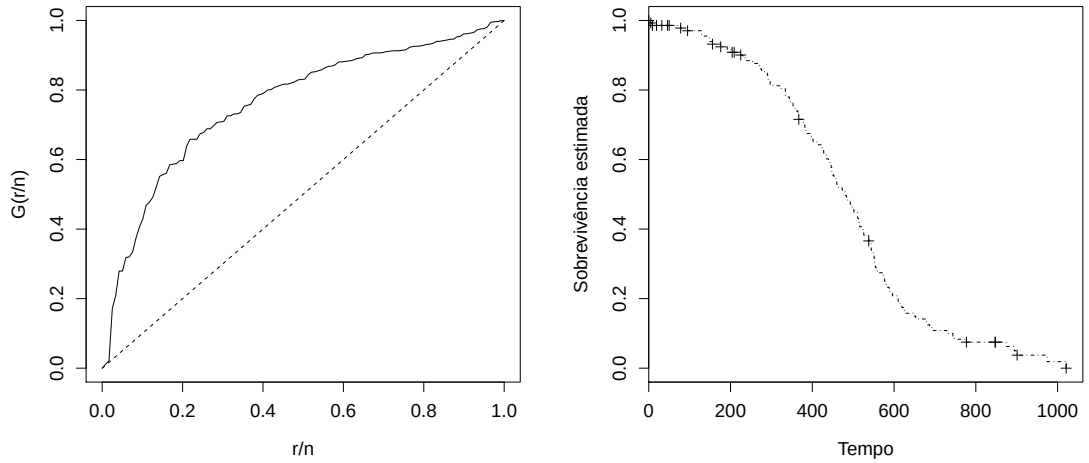

FIGURA 1.3: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T2$.

e foram analisados primeiramente por Proschan (1963) e discutidos posteriormente por Dahiya & Gurland (1972), Gleser (1989), Kus (2007) e Barreto-Souza *et al.* (2010) entre outros. A amplitude dos dados é de 602 falhas, sendo no mínimo 1 falha e metade dos sistemas não apresenta mais de 57 falhas.

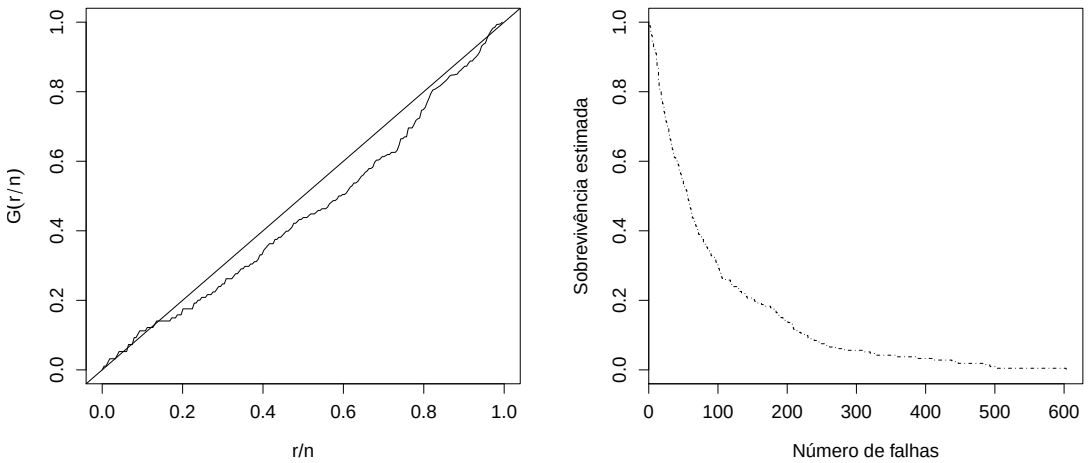

FIGURA 1.4: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T3$.

Como pode-se observar na Figura 1.4, o referido conjunto de dados não apresenta longa duração e a forma da função de risco é decrescente, porém quase constante.

$T4$ e $T5$: - Câncer de pulmão

Um estudo com 137 pacientes diagnosticados com câncer de pulmão - considerando dois tratamentos quimioterápicos - é realizado e observa-se o tempo de sobrevivência, em anos, de cada um desses pacientes. Destes, nove foram censurados e foram observadas também várias covariáveis, a saber: z_{i1}: Classificação de Karnofsky (100=bom); z_{i2}: idade (em anos); z_{i3}: meses entre o diagnóstico e o início do tratamento; z_{i4}: tipo de célula doente; z_{i5}: tratamento e z_{i6}: ter ou não recebido tratamento anterior. Os dados estão disponíveis no pacote *survival* do programa `R` (R Core Team, 2011) e são estudado por Kalbfleisch & Prentice (2002).

Considerando a covariável z_{i6}, tem-se o subconjunto dos 40 pacientes que receberam tratamento anterior ao do estudo (Lawless, 2003). Este subconjunto é denominado de $T5$ e contém três observações censuradas, referentes à pacientes que ainda permaneciam no estudo.

O tempo de vida dos pacientes do conjunto $T4$ é menos disperso em torno da média que no conjunto $T5$. Temos média de 144,6 dias e 121,62 dias com desvio padrão $s = 157,81$ e $s = 222,45$ dias para $T4$ e $T5$, respectivamente.

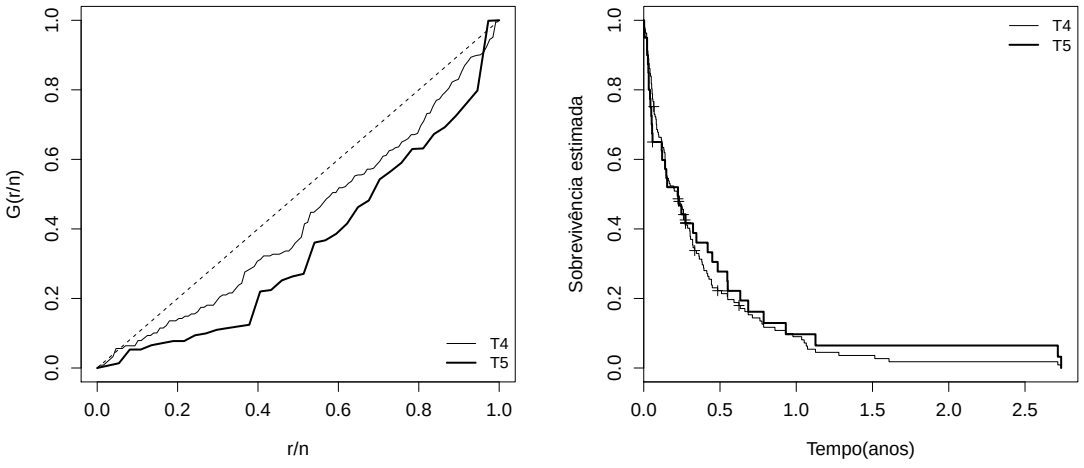

FIGURA 1.5: Gráfico TTT e curva de Kaplan-Meier para os conjuntos de dados $T4$ e $T5$.

Os dados de tempos que compõem ambos os conjuntos apresentam forma de risco decrescente e não há presença de longa duração, conforme sugere a Figura 1.5.

$T6$: **Tempo de permanência em clínica de repouso**

O conjunto de dados, aqui denominado $T6$, é estudado por Ibrahim *et al.* (2001) e corresponde ao estudo de 36 casas de repouso privadas na avaliação dos efeitos do incentivo financeiro na duração da estadia. A clínica recebe ajuda de custo para os pacientes que são medicados e bônus pela recuperação e alta médica dos mesmos.

No total foram considerados 1.601 indivíduos - dos quais 1.279 foram censurados -, com pouco mais de 25% de homens e 50% dos pacientes com idade igual ou superior a 83 anos. As covariáveis consideradas no estudo são: z_{i1}: idade do residente; z_{i2}: trabalho da clínica; z_{i3}: sexo; z_{i4}: estado civil e z_{i5}: estado de saúde.

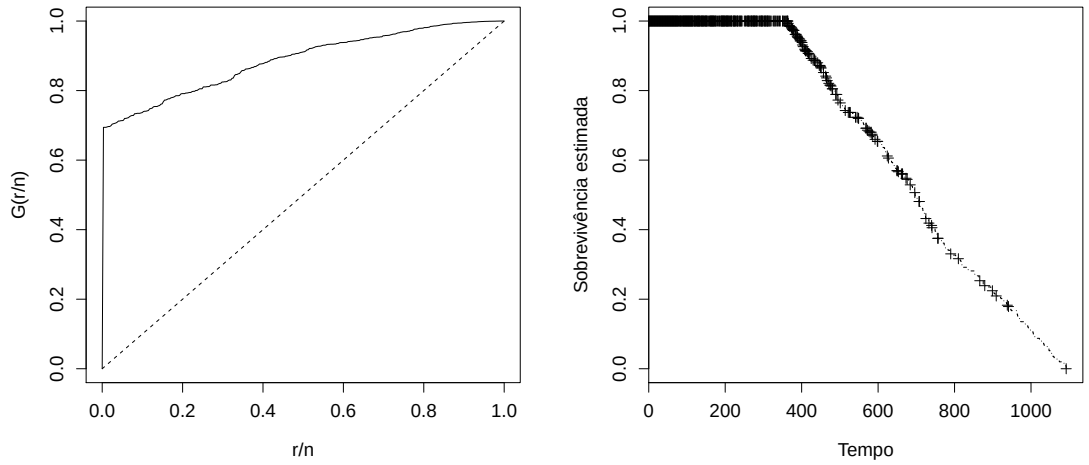

FIGURA 1.6: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T6$.

Sem considerarmos as covariáveis para formar níveis da sobrevivência, os dados não apresentam longa duração e apresentam-se com função de risco crescente (Figura 1.6).

$T7$: **Mielomatose**

O conjunto $T7$ foi extraído de Allison (1995) e consiste em dados de Mielomatose, uma neoplasia maligna de células plasmáticas, em que as células do plasma se proliferam e invadem a medula óssea, o que provoca a destruição do osso, resultando em fraturas e dores ósseas[1]. O conjunto de dados consiste de 25 pacientes com diagnóstico de mielomatose,

[1]http://medical-dictionary.thefreedictionary.com/myelomatosis

cujo tempo foi registrado, em dias, até a morte ou censura do paciente, sendo que a censura foi caracterizada pela desistência do acompanhamento ou pelo encerramento do mesmo. Há 32% de censura nos dados, e eles variam de 8 a 2.240 dias, com média de 613,36 dias. Podemos observar na Figura 1.7 que esse conjunto de dados apresenta longa duração (fração de cura) e risco decrescente.

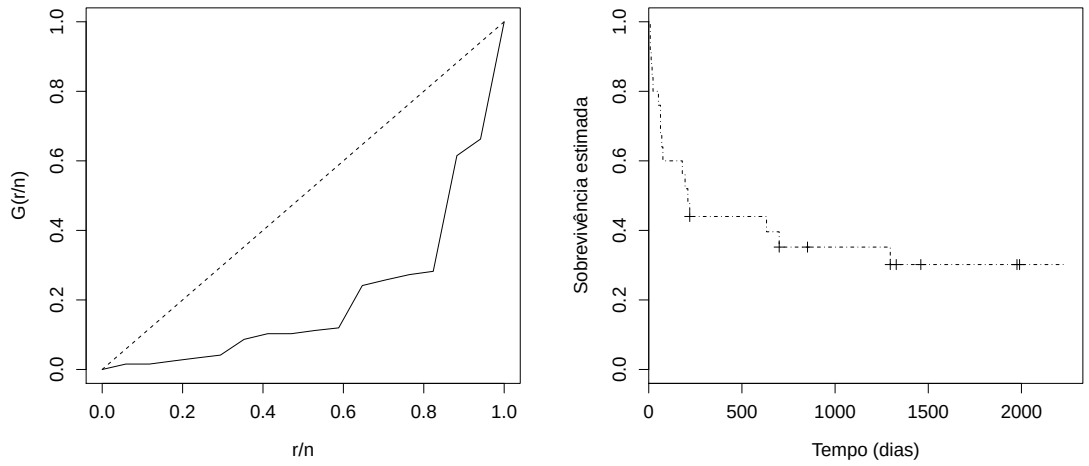

FIGURA 1.7: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T7$.

$T8$: **Leucemia**

O conjunto $T8$ representa dados característicos de longa duração com forma decrescente da função risco (Figura 1.8), é originário de Kersey *et al.* (1987) e disponibilizado em Maller & Zhou (1996). Correspondem à recorrência de leucemia, em anos, para um grupo de 46 pacientes que receberam o tratamento *autologous*. Os autores relatam que o percentual de pacientes curados foi estimado em 20%. No conjunto de dados tem-se 28% de censura, com o tempo de vida variando de 11 dias a 5 anos (tempo máximo do estudo).

$T9$: **Câncer de ovário**

O conjunto T9 apresenta dados de câncer de ovário e foi extraído de Maller & Zhou (1996). Entre as 26 mulheres acompanhadas no estudo, tivemos 46% de censura.

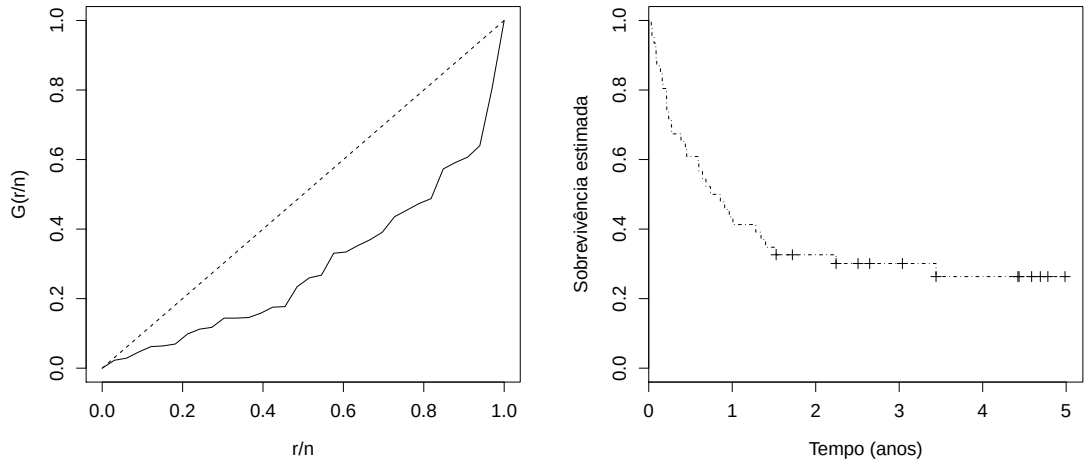

FIGURA 1.8: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T8$.

Os tempos são fornecidos em anos e variam de 0,16 a 3,36 anos, com média de 1,64 anos e $s = 0,93$ anos.

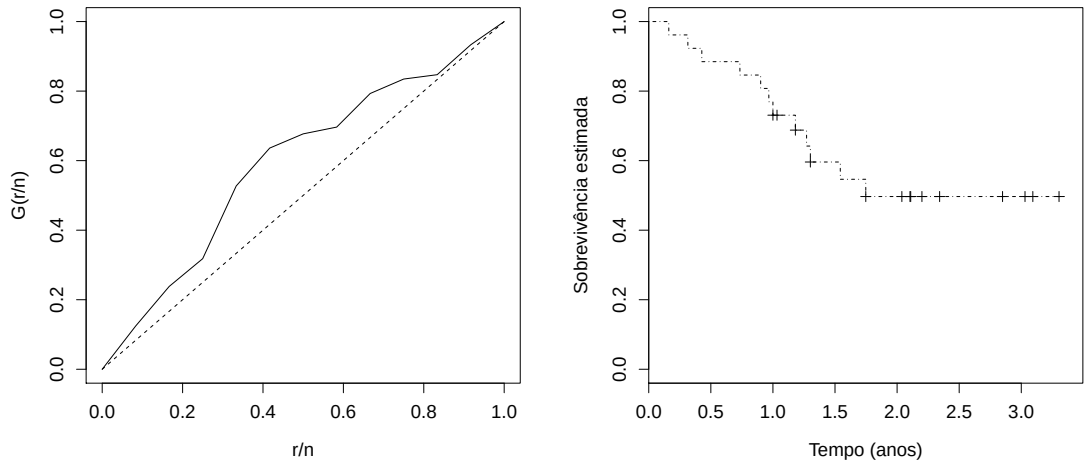

FIGURA 1.9: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T9$.

Tristemente, no início o risco de morte para uma mulher com câncer de ovário aumenta com o tempo, mas felizmente há estabilização da função de risco da mesma, fornecendo indicativo da existência de proporção de curadas, hipótese confirmada no gráfico KM que acena para uma proporção de curados em torno de 50%.

$T10$: **Glioma**

O $T10$[2] contém dados de tempo de vida de 411 pacientes com glioma, tumor do sistema nervoso central, com origem nas células glias que dão suporte e nutrição aos neurônios. Esse conjunto de dados possui $33,3\%$ de censura. Os tempos observados têm média de 324,08 dias, desvio padrão de 278,74 dias e valor máximo de 1617 dias.

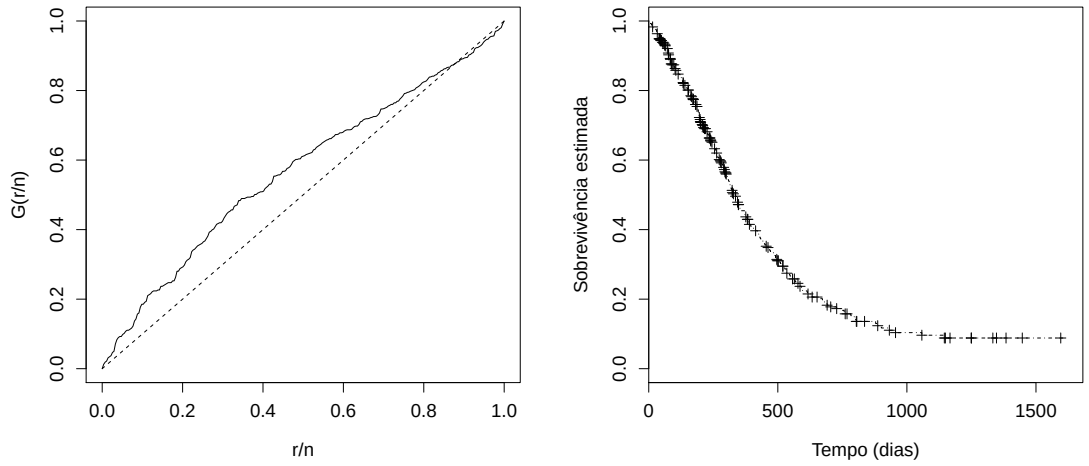

FIGURA 1.10: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T10$.

Esse tumor é um dos mais severos quando se trata de doenças do sistema nervoso e, pelo apresentado na Figura 1.10, a proporção de curados, embora não ausente, é pequena e o risco de morte é crescente inicialmente e com o passar do tempo torna-se praticamente constante.

$T11$: **Escore de crédito**

Esse conjunto de dados consiste no tempo até a inadimplência de 507 cliente de um banco brasileiro. Esses clientes emprestaram dinheiro do banco e o saldo devedor foi dividido em 24 parcelas de igual valor. A base foi balanceada para uma proporção de um cliente ruim para 2 clientes bons, resultando em quase 67% de censura. Os tempos variam de 1,23 a 24 meses, com mediana de 14,10.

Os dados apresentam forma de risco crescente e há estabilização da sobrevivência estimada indicando longa duração (Figura 1.11).

[2]http://portal.uni-freiburg.de/imbi/Royston-Sauerbrei-book

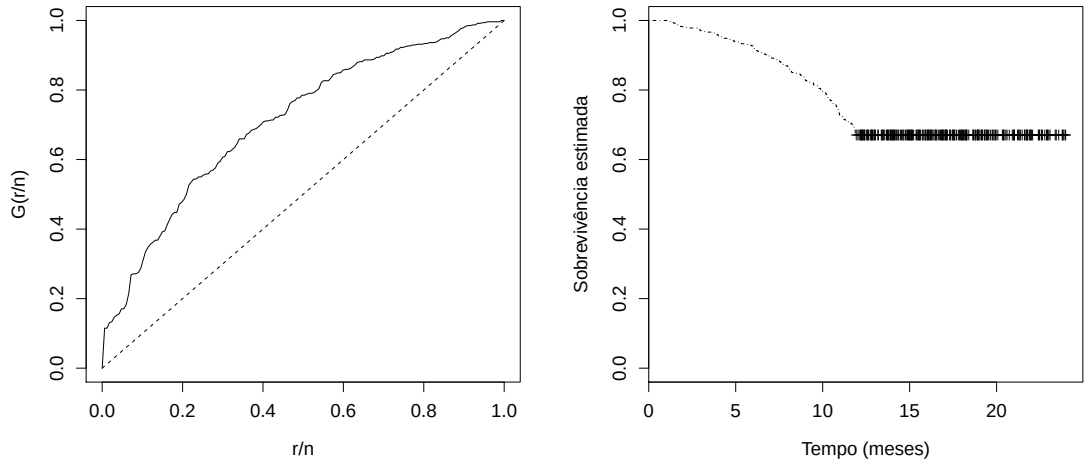

FIGURA 1.11: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T11$.

$T12$: **Melanoma**

O conjunto $T12$ foi extraído de Scheike (2009), com dados de sobrevivência de 205 pacientes após o procedimento cirúrgico para remoção de melanoma maligno. O melanoma cutâneo é um tipo de câncer de pele que tem origem nos melanócitos (células produtoras de melanina) e tem predominância em adultos brancos. Embora o câncer de pele seja o mais frequente no Brasil e corresponda a 25% de todos os tumores malignos registrados, o melanoma representa apenas 4% das neoplasias malignas do órgão, apesar de ser o mais grave devido à sua alta possibilidade de metástase. O prognóstico desse tipo de câncer pode ser considerado bom, se detectado nos estádios iniciais. Nos últimos anos houve uma grande melhora na sobrevida dos pacientes com melanoma, principalmente devido à detecção precoce do tumor.

Os tempos observados correspondem ao intervalo entre a remoção do melanoma e a morte ou censura do paciente e variam de 0,0274 até 15,25 anos (10 à 5.565 dias) com média de 5,9 anos e desvio padrão de $s = 3,1$ anos. Considerou-se censurado o paciente que morreu de outra causa, ou continuava vivo no término do estudo. O estudo apontou 72% de censura. Como covariáveis, consideramos: estado de ulceração (ausente: $n = 115$; presente: $n = 90$) e espessura do tumor (em milímetros, média de 2,92 e desvio padrão $s = 2,96$). A estimativa de Kaplan-Meier (Figura 1.12) indica proporção de curados acima de 0,6.

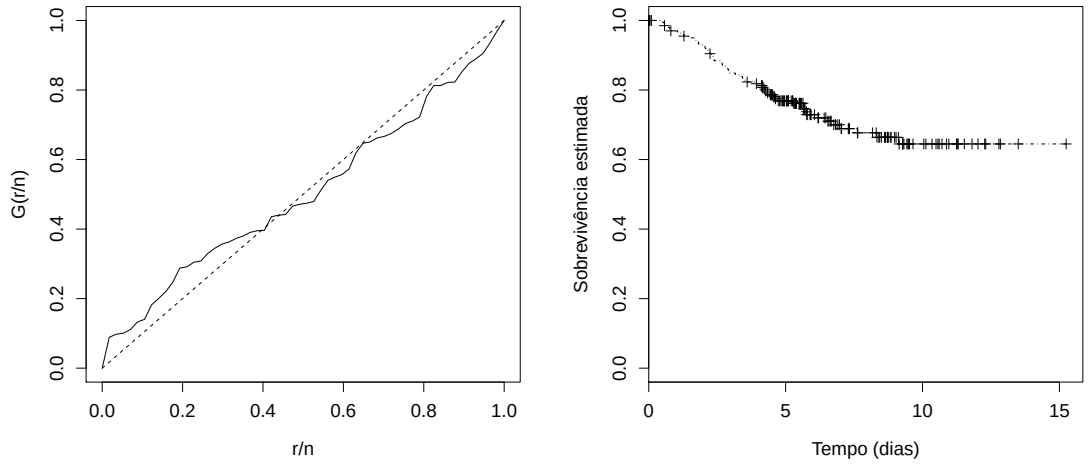

FIGURA 1.12: Gráfico TTT e curva de Kaplan-Meier para o conjunto de dados $T12$.

1.3.1 Presença de múltiplos riscos nos conjuntos de dados

Entre os 12 conjuntos de dados apresentados, temos que sete deles referem-se a dados de câncer. No estudo de câncer da população feminina Mendonça (1993) afirma que para determinados tipos de câncer é possível identificar fatores que estão associados ao seu aparecimento e que muitos estudos têm sido conduzidos nesse sentido. Para algumas neoplasias os processos etiológicos já são largamente aceitos, no entanto, há necessidade de estudos para descrever como, os principais fatores que se relacionam no desenvolvimento do câncer, se distribuem e se comportam na população. Esse conceito pode ser estendido para toda a população, e a comprovação da existência dos fatores de risco já justifica a utilização de modelos de múltiplos riscos na modelagem de dados de câncer.

Para o conjunto $T2$ que trata do tempo até a reversão da sorologia positiva para HIV em crianças, Perdoná (2006) afirma que mesmo a criança não tendo propriamente o vírus o exame pode dar positivo em decorrência dos anticorpos transmitidos pela mãe soropositivo. O índice de carga viral é um dos componente medidos para confirmação do HIV positivo, e Carvalho *et al.* (2003) afirmam que vacinas como a influenza ativam esse mecanismo. Outros trabalhos, não especificamente com crianças indicam que quadros infecciosos também podem influenciar nos exames, fatos estes que norteiam a conduta de exames de diagnóstico e posterior acompanhamento da evolução da doença. Desta forma, podemos supor que fatores do organismo da criança - sejam eles provocados ou

não - podem produzir resultado positivo no teste. Quais, quantos e de que forma foram ativados esses fatores não é conhecido ainda. Com base nessa suposição, consideramos modelos de múltiplos riscos para modelar esses dados.

Para os dados de área da engenharia: $T1$ - resistência das esferas - e $T3$ - falha de ar condicionado - a justificativa da utilização de modelos de múltiplos riscos é o fato de que na prática uma máquina geralmente não quebra totalmente de uma só vez, mas para de trabalhar quando alguma parte vital de seu conjunto se danifica.

Para o dados de tempo de inadimplência ($T11$) julgamos que é razoável considerarmos que não há apenas um fator envolvido, sejam pelas condições sociais ou morais. E para o tempo de permanência do paciente na clínica ($T6$) fatores como demanda, lucratividade, trabalho despendido com o paciente, além do estado de saúde do mesmo, podem ser fatores que influenciam na ocorrência do evento de interesse.

Capítulo 2

Distribuições para múltiplos fatores de risco: Formulação geral

Este capítulo apresenta a formulação de distribuições de probabilidade em que a ocorrência do evento de interesse está associada a uma quantidade latente de causas de falha (CR). O tempo observado corresponde ao tempo até a ocorrência de uma dessas CR: a primeira, a última, ou uma intermediária. Consideramos que a quantidade M de CR possui distribuição de probabilidade Geométrica e que o tempo até a ocorrência de cada uma das causas segue uma distribuição base, cuja f.d.p. é representada por $f_o(t)$.

2.1 Introdução

Em estudos clínicos, cujo evento de interesse é a morte do paciente ou a reincidência de um tumor, o evento de interesse pode ocorrer devido a diferentes causas, aqui tratadas de causas de risco, as quais são latentes no sentido de que não existe qualquer informação sobre quantas são, quantas foram ativadas e/ou qual delas é responsável pela manifestação do evento. Por exemplo, a recorrência de um tumor pode ser atribuída às células tumorais que ficaram com o componente de metástase ativo após o tratamento. Uma célula tumoral com componente de metástase ativo é uma célula cancerosa com potencial para metástase (Yakovlev & Tsodikov, 1996), mas não se tem certeza de quantas são essas células e sequer se foi uma delas a responsável pela recorrência do tumor. Assim, temos um cenário de múltiplos riscos.

A literatura de distribuições que acomodam múltiplos riscos com diferentes formas de ativação é expressiva e cresce rapidamente. O livro de Ibrahim *et al.* (2001), o artigo de revisão de Tsodikov *et al.* (2003) e os trabalhos de Louzada-Neto (1999), Cooner *et al.* (2007), Ortega *et al.* (2008), Rodrigues *et al.* (2009b), Ortega *et al.* (2009), Cancho *et al.* (2009) e Kim *et al.* (2011) que estendem o modelo proposto por Cooner *et al.* (2007) e fazem uma abordagem bayesiana para dados de câncer de próstata, podem ser mencionados como referências.

Cooner *et al.* (2007) consideram que os fatores de risco, mesmo existindo podem não ativarem o evento de interesse, incutindo nos modelos a proporção de curados existente na população. Consideração, esta, muito importante para a análise de sobrevivência e para a realidade do estudo, e ainda consideram que ocorrendo o evento de interesse, este pode estar vinculado ao primeiro ou ao último fator de risco.

E do ponto de vista prático, devido aos ainda insuficientes, mas já significantes avanços nos procedimentos de tratamentos de doenças como o câncer, é louvavelmente comum encontrarmos dados de sobrevivência em que parte da população é não suscetível ao evento de interesse. Por exemplo, em estudos clínicos, o indivíduo pode responder favoravelmente ao tratamento, sendo considerado curado, justificando a inclusão dessa classe de modelos nos estudos.

Seguimos Cooner *et al.* (2006) e Cooner *et al.* (2007) e formulamos misturas de distribuição que consideram primeiramente apenas a população dos suscetíveis e posteriormente a população de curados nos esquemas de ativação.

2.2 Fundamentação

Consideremos um mecanismo e supomos que a ele estão associados alguns elementos (fatores ou causas de risco). A ativação do mecanismo corresponde à ocorrência do evento de interesse e está condicionada à ocorrência de ao menos uma das CRs. A não ocorrência do evento de interesse indica que nenhum elemento foi ativado ou os elementos ativados não foram suficientes para ocasioná-lo. Todavia, ainda há a possibilidade de que ao mecanismo não estejam associados elementos que provoquem sua ativação.

A variável aleatória latente M representa a quantidade de CR associada ao mecanismo, assumimos M com distribuição de probabilidade Geométrica. Se considerarmos que sempre haverá elementos associados ao mecanismo e, portanto, $M = 1, 2, 3, \cdots$, então sua função massa de probabilidade é

$$p(m) = P(M = m) = \theta(1 - \theta)^{m-1}, \tag{2.1}$$

no entanto, se M pode assumir o valor nulo, indicando que pode não haver CR associado ao mecanismo, a função massa de probabilidade da v.a. M é dada por,

$$p(m) = P(M = m) = \theta(1 - \theta)^{m}. \tag{2.2}$$

Sejam $T_i,\ i = 1, 2, \cdots, M$ v.a. i.i.d., não nulas e independentes de M, representando o tempo até a ativação da *i-ésima* causa de risco associadas ao evento de interesse, cuja função densidade de probabilidade é $f_o(t_i)$, de forma que a f.d.a. de T_i é dada por $F_o(t_i)$, a função de sobrevivência é $S_o(t_i)$ e $h_o(t_i)$ é a função de risco de T_i. Para não carregar a notação, em geral, utilizamos $T = t$ ao invés de $T_i = t_i$.

Consideramos $T_{(1)}, T_{(2)}, \cdots, T_{(M)}$, em que $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(M)}$ são as M primeiras estatísticas de ordem de T_i e apenas o tempo $T_{(K)},\ K = 1, 2, \cdots, M$, que ativa o mecanismo, será observado.

Em muitos processos biológicos, K pode ser interpretado como um fator de resistência do sistema imunológico do indivíduo. Se o evento de interesse ocorre (i.e., há reincidência do câncer), então a variável aleatória que denota o tempo até a ocorrência do evento de interesse assume o valor da *k-ésima* estatística de ordem $T_{(K)}$. Em Cooner *et al.* (2006) e Cooner *et al.* (2007), são necessárias K de M causas para que ocorra o evento de interesse, utilizando-nos da mesma terminologia, consideramos três especificações para K:

- $K = 1$, o tempo observado é $X = T_{(1)} = \min(T_1, T_2, \cdots, T_M)$ - correspondente a ativação da primeira CR, denotamos esse cenário por CR_{mn}.
- $K = q$, neste caso observa-se $W = T_{(q)},\ 1 < q < M$ - correspondente a ativação da *q-ésima* CR. Denotamos esse cenário por CR_{al}.
- $K = M$, em que observa-se $Y = T_{(M)} = \max(T_1, T_2, \cdots, T_M)$ - correspondente a ativação de todas as CRs (ou de forma simples com a ativação da última de todas

as CR envolvida). Denotamos esse cenário por CR_{mx}.

Relacionando as duas possibilidades para a v.a. M com os três casos especificados para K, temos as seis ramificações da mistura com a distribuição de probabilidade Geométrica apresentadas nesta Tese.

Na Seção 2.3 apresentamos os conceitos relacionados aos três cenários de ativação e em seguida as distribuições para cada uma das possibilidades adotadas para a v.a M.

2.3 Diferentes esquemas de ativação

Uma dificuldade da AS no contexto de múltiplos riscos consiste em discernir qual cenário (CR_{mn}, CR_{al} ou CR_{mx} ou outros possíveis) originou os eventos de interesse e cujas observações correspondem aos dados de sobrevivência. Esta dificuldade pode ser pelo fato da quantidade de CRs ser geralmente uma v.a. latente e aos tempos relativos às CR's envolvidas dificilmente serem observáveis.

Considerando a quantidade de fatores de risco envolvidos como uma v.a. latente, assumimos que a ocorrência do evento de interesse é ocasionado pela ativação de uma, todas ou algumas das CRs. Porém, desconhecendo a quantidade de CRs ativadas para desencadear o evento de interesse, induz-se que houve a ocorrência de k, $0 < k \leq M$ fatores de risco.

Se o evento de interesse não ocorre, temos a presença de censura, mas isso não exclui a possibilidade de algum fator de risco ter sido ativado, nos dando uma informação incerta: ou não temos causas de risco associadas ao evento de interesse ou não ocorreu a quantidade mínima necessária de CRs para a ativação do mecanismo.

A seguir, apresentamos detalhes referentes aos três cenários considerados.

2.3.1 Ativação pelo mínimo dos tempos

No contexto do mínimo dos tempos, a ocorrência do evento de interesse coincide com o tempo até a ocorrência da primeira CR associada a ele. A variável aleatória que denota esse tempo é $X = T_{(1)} = \min(T_1, T_2, \cdots, T_M)$. Este contexto se compara a um

sistema ligado em série, no qual a falha do sistema é motivada pela falha de qualquer um dos componentes do mesmo. É comum a utilização do termo *riscos competitivos* para designar esse cenário, que é denotado por CR_{mn}.

Neste enfoque, Goetghebeur & Ryan (1995) abordam o problema de avaliação dos efeitos de covariáveis com base na estrutura de risco proporcional semiparamétrico para cada tipo de falha, quando o tipo de falha é desconhecido para alguns indivíduos. Reiser *et al.* (1995) consideram procedimentos estatísticos para analisar dados mascarados, não aplicados quando todas as observações têm uma causa desconhecida de fracasso.

Adamidis & Loukas (1998) propõem a distribuição Exponencial Geométrica (EG), que acomoda dados sobrevivência, Louzada-Neto (1999) propõe o modelo de múltiplos riscos, que trata com versatilidade desses tempos. Lu & Tsiatis (2001) apresentam um método para estimar os coeficientes de regressão para dados dessa classe e uma comparação de duas verossimilhanças parciais para esses dados é apresentada em Lu & Tsiatis (2005).

Procedimentos estatísticos e discussões a respeito desse assunto podem ser encontrados também em Lawless (2003), Crowder *et al.* (1991) e Cox & Oakes (1984).

2.3.2 Ativação pelo máximo dos tempos

Nessa classe de eventos, intuitivamente a ocorrência do evento de interesse é tardia, porém, não menos problemática. Neste cenário o tempo observado está associado ao tempo até a ocorrência do último fator de risco, $Y = \max(T_1, T_2 \ldots, T_M)$. Esse processo de ativação é relacionado com um sistema ligado em paralelo, em que o sistema só apresentará falha quando todos os seus componentes falharem. Pelo fato de uma causa de risco necessitar das demais para que o evento de interesse ocorra, essa classe de múltiplos riscos pode ser denominada de riscos complementares e a denotamos por CR_{mx}.

Cooner *et al.* (2006) expõem sobre a situação de riscos latentes complementares, Kus (2007) propõe uma distribuição para acomodar dados de sobrevivência neste cenário e Cancho *et al.* (2010a) apresentam a distribuição Exponencial Poisson.

2.3.3 Ativação aleatória

Neste cenário o tempo observado $W = T_{(K)}$ é relativo à ocorrência da K-*ésima* CR, $(1 < K < M)$ e será representado por CR_{al}. Uma de suas características é que, dependendo das suposições feitas, as distribuições decorrentes das misturas de distribuições, podem coincidir com distribuições tradicionais já existentes.

2.4 Modelos de Mistura

Esta formulação não considera a presença de longa duração, a qual pode ser introduzida no modelo considerando a metodologia principiada por Boag (1949) e Berkson e Gage (1952). Mas nem sempre a proporção de curados é o objetivo de estudo e muitos conjuntos de dados da literatura não apresentam essa particularidade.

Consideramos a v.a. M que denota o número de CRs cuja f.d. é dada em (2.1) e $T_i,\ i = 1, 2, ...M$ o tempo até a ocorrência da *i-ésima* CR com f.d.p. dada por $f_o(t)$.

2.4.1 Modelos de mistura no cenário CR_{mn}

Proposição 2.1 *Seja X a variável aleatória que denota o tempo até a ocorrência do evento de interesse, em que $X = \min(T_1, T_2 \ldots, T_M)$, $M \sim G(\theta)$ cuja f.p. é dada em (2.1) e T_i com f.d.p. $f_o(t)$. Então a f.d.p. de X é dada por*

$$f(x) = \frac{\theta f_o(x)}{[1-(1-\theta)S_o(x)]^2}. \tag{2.3}$$

Prova 2.1 A função densidade de probabilidade condicional de X dado $M = m$ é $f(X|M = m) = m\,[S_o(x)]^{m-1} f_o(x)$. Então, a função densidade marginal de X é dada por

$$\begin{aligned} f(x) =& \sum_{m=1}^{\infty} f(X|M=m)P(M) = \sum_{m=1}^{\infty} m\,[S_o(x)]^{m-1} f_o(x)\theta(1-\theta)^{m-1} \\ =& \frac{\theta f_o(x)}{(1-\theta)S_o(x)} \sum_{m=1}^{\infty} m\,[(1-\theta)S_o(x)]^m. \end{aligned}$$

Como $(1-\theta)S_o(x) \in (0,1)$, então como mostrado em (A.1) segue que

$$f(x) = \frac{\theta f_o(x)}{(1-\theta)S_o(x)} \frac{(1-\theta)S_o(x)}{[1-(1-\theta)S_o(x)]^2} = \frac{\theta f_o(x)}{[1-(1-\theta)S_o(x)]^2}. \quad \blacksquare$$

Proposição 2.2 *A função distribuição e a função de sobrevivência atreladas à f.d.p. (2.3) são dadas, respectivamente, por*

$$F(x) = \frac{F_o(x)}{1-(1-\theta)S_o(x)} \qquad e \qquad S(x) = \frac{\theta S_o(x)}{1-(1-\theta)S_o(x)}.$$

Prova 2.2 A função distribuição é obtida como segue:

$$\begin{aligned} F(x) =& \int_0^x \frac{\theta f_o(t)}{[1-(1-\theta)S_o(t)]^2} dt = \frac{\theta}{1-\theta}\int_\theta^{1-(1-\theta)S_o(x)} \frac{1}{u^2} du \\ =& \frac{\theta}{1-\theta}\left[-\frac{1}{1-(1-\theta)S_o(x)} + \frac{1}{\theta}\right] = \frac{F_o(x)}{1-(1-\theta)S_o(x)}. \end{aligned}$$

Usando a relação $S(x) = 1 - F(x)$ obtêm-se a função de sobrevivência. ■

Consequentemente, a função de risco é

$$h(x) = \frac{f_o(x)}{S_o(x)\left(1-(1-\theta)S_o(x)\right)} = \frac{f_o(x)}{S_o(x)}\frac{1}{1-(1-\theta)S_o(x)} = \frac{h_o(x)}{1-(1-\theta)S_o(x)}. \quad (2.4)$$

Trabalhos como Adamidis & Loukas (1998) utilizam esses resultados particularmente para $f_o(t)$ sendo a f.d.p. da distribuição Exponencial, mas não apresentam o caso geral para qualquer f.d.p. como o exposto, embora Marshall & Olkin (1997) apresentem tais formulações porém, com menor riqueza de detalhes.

2.4.2 Modelos de mistura no cenário CR_{mx}

A variável aleatória que denota o tempo até a ocorrência do evento de interesse é $Y = \max(T_1, T_2 \ldots, T_M)$. Consideremos que $M \sim G(\theta)$ (2.1) e T_i com f.d.p. $f_o(t)$.

Proposição 2.3 *A f.d.p. da variável aleatória Y é*

$$f(y) = \frac{\theta f_o(y)}{\left[1-(1-\theta)F_o(y)\right]^2}. \quad (2.5)$$

Prova 2.3 Sendo a função densidade de probabilidade condicional de Y dado $M = m$ dada por $f(Y|M=m) = m\left[F_o(y)\right]^{m-1} f_o(y)$, então, a f.d.p. marginal de Y é

$$\begin{aligned} f(y) =& \sum_{m=1}^{\infty} f(Y|M=m)P(M) = \sum_{m=1}^{\infty} m\left[F_o(y)\right]^{m-1} f_o(y)\theta(1-\theta)^{m-1} \\ =& \frac{\theta f_o(y)}{(1-\theta)F_o(y)} \sum_{m=1}^{\infty} m\left[(1-\theta)F_o(y)\right]^m. \end{aligned}$$

Como $(1-\theta)F_o(y) \in (0,1)$, então como mostrado em (A.1) segue que

$$f(y) = \frac{\theta f_o(y)}{(1-\theta)F_o(y)} \frac{(1-\theta)F_o(y)}{[1-(1-\theta)F_o(y)]^2} = \frac{\theta f_o(y)}{[1-(1-\theta)F_o(y)]^2}. \quad \blacksquare$$

Proposição 2.4 *A função distribuição e a função de sobrevivência da variável Y com f.d.p. (2.5) são dadas, respectivamente, por*

$$F(y) = \frac{\theta F_o(y)}{1-(1-\theta)F_o(y)} \qquad e \qquad S(y) = \frac{S_o(y)}{1-(1-\theta)F_o(y)}. \tag{2.6}$$

Prova 2.4 A obtenção da f.d.a. é dada como segue:

$$\begin{aligned} F(y) = &\int_0^y \frac{\theta f_o(t)}{[1-(1-\theta)F_o(t)]^2} dt = -\frac{\theta}{1-\theta} \int_1^{1-(1-\theta)F_o(y)} \frac{1}{u^2} du \\ = &\frac{\theta}{1-\theta} \left[\frac{1}{1-(1-\theta)F_o(y)} - 1 \right] = \frac{\theta F_o(y)}{1-(1-\theta)F_o(y)}, \end{aligned}$$

e função de sobrevivência é obtida fazendo $S(y) = 1 - F(y)$. $\blacksquare$

No cenário CR_{mx}, a função de risco para o tempo amostrado é

$$h(y) = \frac{\theta f_o(y)}{S_o(y)\,[1-(1-\theta)F_o(y)]} = \frac{f_o(y)}{S_o(y)} \frac{\theta}{1-(1-\theta)F_o(y)} = \frac{\theta h_o(y)}{1-(1-\theta)F_o(y)}. \tag{2.7}$$

2.4.3 Modelos de mistura no cenário CR_{al}

Assumimos que dado $M \geq 1$, a distribuição condicional de K é discreta em $\{1, 2, \ldots, M\}$, com probabilidade $1/M$. Assim a função distribuição para $W = T_{(K)}$, dado $M = m$ e $K = k$, é dada por

$$F_{W|m,k}(w) = P[W \leq w | M = m, K = k] = \sum_{j=k}^{m} \binom{m}{j} [F_o(w)]^j [S_o(w)]^{m-j}. \tag{2.8}$$

Proposição 2.5 *Considerando a distribuição condicional (2.8), a f.d.a. marginal de W é a própria $F_o(w)$.*

Prova 2.5

$$\begin{aligned} F(w) &= \sum_{m=1}^{\infty} \sum_{k=1}^{m} P[W \leq w | M = m, K = k)] P[K = k | M = m] P[M = m] \\ &= 1 - \sum_{m=1}^{\infty} \left\{ \sum_{k=0}^{m} (m-k) B(k, m, F_o(w)) \right\} \frac{1}{m} P[M = m] \\ &= 1 - [1 - F_o(w)] \sum_{m=1}^{\infty} p_m = F_o(w), \end{aligned} \tag{2.9}$$

em que $B(k, m, F_o(w))$ é a função massa de probabilidade da distribuição Binomial com parâmetros m, $F_o(w)$ e $P[M = m]$. ■

Portanto a distribuição da v.a. W é a mesma da variável aleatória latente T_i. Desta forma, as demais funções associadas ao CR_{al} também se resumem às funções associadas à v.a. T_i, ou seja

$$S(w) = S_o(w) \qquad e \qquad f(w) = f_o(w). \tag{2.10}$$

2.5 Modelo de longa duração com fatores de risco latentes

Supomos que a v.a. M, que denota a quantidade de CR envolvidas na ocorrência do evento de interesse em um indivíduo, tem uma distribuição de probabilidade com massa no zero, isto é, $P(M = 0) > 0$, havendo portanto a possibilidade de que as supostas CR não provoquem a ocorrência do evento de interesse. Nessa situação há uma condição de censura e o tempo pode ser tomado como infinito, nos remetendo ao cenários dos conhecidos e já mencionados modelos de sobrevivência de longa duração.

Consideramos a v.a. M que denota o número de CRs cuja f.d. é dada em (2.2) e $T_i,\ i = 1, 2, ...M$ o tempo até a ocorrência da *i-ésima* CR com f.d.p. dada por $f_o(t_i)$.

Assumimos que dado $M \geq 1$, a distribuição condicional de K é Uniforme em $\{1, 2, \ldots, M\}$. Nesta configuração a função de sobrevivência para a população é dada por

$$\begin{aligned} S_{\text{pop}}(w) &= P(W \geq w) = P(M = 0) + P(T_{(K)} > w, 1 \leq K \leq M) \\ &= P(M = 0) + \sum_{m=1}^{\infty} \sum_{K=1}^{m} P(T_{(K)} > w | K, M = m) P(K | M = m) P(M = m) \\ &= P(M = 0) + \sum_{m=1}^{\infty} \sum_{K=1}^{m} \sum_{i=0}^{K-1} \binom{m}{i} [F_o(w)]^i [S_o(w)]^{m-i} P(K | M = m) P(M = m). \end{aligned} \tag{2.11}$$

Note que $P(T_{(K)} > w | K, M = m) = \sum_{i=0}^{K-1} \binom{m}{i} [F_o(w)]^i [S_o(w)]^{m-i}$ fornece a f.d.a. da distribuição Binomial para m ensaios e probabilidade de sucesso $F_o(w)$.

Em (2.11) temos que $\sum_{K=1}^{m} \sum_{i=0}^{K-1} \binom{m}{i} [F_o(w)]^i [S_o(w)]^{m-i}$ é específica para cada

valor de K considerado, sendo

$$\sum_{i=0}^{K-1}\binom{m}{i}[F_o(w)]^i[S_o(w)]^{m-i} = \begin{cases} [S_o(w)]^m, & \text{se } K=1; \\ S_o(w), & \text{se } K=k,\ 0<k<m; \\ 1-[F_o(w)]^m, & \text{se } K=m. \end{cases} \tag{2.12}$$

2.5.1 Modelos de mistura com longa duração no cenário CR_{mn}

A variável aleatória que denota o tempo até a ocorrência do evento de interesse é $X = \min(T_1, T_2 \ldots, T_M)$.

Proposição 2.6 *A função de sobrevivência de longa duração da v.a. X, sendo $M \sim G(\theta)$ com f.p. dada em (2.2) e T_i com função de sobrevivência $S_o(t)$, é dada por*

$$S_{pop}(x) = \frac{\theta}{1-(1-\theta)S_o(x)}. \tag{2.13}$$

Prova 2.6 Substituindo (2.12) para $K=1$ e $P(K|M=m) = \frac{1}{m}$ em (2.11), segue que

$$\begin{aligned} S_{pop}(x) =& P(M=0) + \sum_{m=1}^{\infty}\sum_{K=1}^{m}[S_o(x)]^m P(K|M=m)P(M=m) \\ =& \theta + \sum_{m=1}^{\infty}[S_o(x)]^m\theta(1-\theta)^m = \theta + \theta\sum_{m=1}^{\infty}[S_o(x)(1-\theta)]^m \\ =& \theta + \frac{\theta(1-\theta)S_o(x)}{1-(1-\theta)S_o(x)} = \frac{\theta}{1-(1-\theta)S_o(x)}. \quad \blacksquare \end{aligned}$$

Proposição 2.7 *A função densidade de probabilidade correspondente à função de sobrevivência dada em (2.13) é*

$$f_{pop}(x) = \frac{\theta(1-\theta)f_o(x)}{[1-(1-\theta)S_o(x)]^2}. \tag{2.14}$$

Prova 2.7 A função densidade de probabilidade pode ser obtida por

$$\begin{aligned} f_{pop}(x) =& -\frac{\partial}{\partial x}\left[\frac{\theta}{1-(1-\theta)S_o(x)}\right] = \frac{\theta(1-\theta)}{[1-(1-\theta)S_o(x)]^2}\left[\frac{-\partial}{\partial x}S_o(x)\right] \\ =& \frac{\theta(1-\theta)f_o(x)}{[1-(1-\theta)S_o(x)]^2}. \quad \blacksquare \end{aligned}$$

Consequentemente, a função de risco é dada por

$$h_{pop}(x) = \frac{(1-\theta)f_o(x)}{1-(1-\theta)S_o(x)}. \tag{2.15}$$

2.5.2 Modelo de mistura com longa duração no cenário CR_{mx}

A variável aleatória que denota o tempo até a ocorrência do evento de interesse é $Y = \max(T_1, T_2 \ldots, T_M)$.

Proposição 2.8 *A função de sobrevivência de longa duração da v.a. Y, sendo $M \sim G(\theta)$ (2.2) e T_i com f.d.a. $F_o(t)$, é dada por*

$$S_{pop}(y) = 1 - \frac{\theta(1-\theta)F_o(y)}{1-(1-\theta)F_o(y)}. \tag{2.16}$$

Prova 2.8 Considerando (2.12) para $K = m$ em (2.11), segue que

$$\begin{aligned} S_{pop}(y) =& P(M=0) + \sum_{m=1}^{\infty}\sum_{K=1}^{m}\left\{1-[F_o(y)]^m\right\}P(K|M=m)P(M=m) \\ =& \theta + \sum_{m=1}^{\infty}\left\{1-[F_o(y)]^m\right\}\theta(1-\theta)^m = \theta + \theta\left[\sum_{m=1}^{\infty}(1-\theta)^m - \sum_{m=1}^{\infty}\left[(1-\theta)F_o(y)\right]^m\right] \\ =& \theta + \theta\left[\frac{1-\theta}{\theta} - \frac{(1-\theta)F_o(y)}{1-(1-\theta)F_o(y)}\right] = 1 - \frac{\theta(1-\theta)F_o(y)}{1-(1-\theta)F_o(y)}. \end{aligned}$$

■

A fração de curados é $p_0 = \theta$ e as correspondentes f.d.p. e função de risco são

$$f_{pop}(y) = \frac{\theta(1-\theta)f_o(y)}{[1-(1-\theta)F_o(y)]^2} \quad \text{e} \quad h_{pop}(y) = \frac{(1-\theta)f_o(y)}{[1-(1-\theta)F_o(y)]\,[S_o(y)+\theta^2F_o(y)]}.$$

2.5.3 Modelos de mistura com longa duração no cenário CR_{al}

Proposição 2.9 *Seja $W = T_{(K)}$, em que $1 < K < M$, a v.a. que denota o tempo até a ocorrência do evento de interesse, $M \sim G(\theta)$ com f.d. dada em (2.2) e T_i com f.d.p. $f_o(t_i)$. A função de sobrevivência de longa duração da v.a. W é dada por*

$$S_{pop}(w) = \theta + (1-\theta)S_o(w), \tag{2.17}$$

que é o modelo de longa duração de Berkson e Gage (BG).

Prova 2.9 Considerando (2.12) para $K = k$ em (2.11), temos

$$\begin{aligned} S_{pop}(w) =& P(M=0) + \sum_{m=1}^{\infty} mS_o(w)\frac{1}{m}P(M=m) = \theta + \sum_{m=1}^{\infty} S_o(w)\theta(1-\theta)^m \\ =& \theta + \theta S_o(w)\sum_{m=1}^{\infty}(1-\theta)^m = \theta + \theta S_o(w)\left(\frac{1-\theta}{\theta}\right) = \theta + (1-\theta)S_o(w). \end{aligned}$$

■

A função densidade de probabilidade e a função de risco correspondentes à função de sobrevivência dada em (2.17) são

$$f_{pop}(w) = (1-\theta)f_o(w), \qquad \text{e} \qquad h_{pop}(w) = \frac{(1-\theta)f_o(w)}{\theta + (1-\theta)S_o(w)}.$$

2.6 Comentários

Uma relação matemática envolvendo os modelos apresentados nas seções (2.4) e (2.5) e os modelos de mistura com fração de cura (Boag, 1949; Berkson & Gage, 1952), seria

$$S_{pop}(t) = p + (1-p)S(t), \tag{2.18}$$

em que $S(t)$ é definido em (2.2) ou (2.4).

Deste modo, $S_{\text{pop}}(t)$ é uma mistura de modelos com fração de cura (em que a fração de cura é $p = \theta$) e função de sobrevivência $S(t)$ para a população de não curados. Considerando ainda os diferentes esquemas de ativação, temos que $S(t)$ são dados nas Proposições 2.2 e 2.4 para o mínimo e o máximo dos tempos T_i, respectivamente, e para o aleatório na equação (2.10).

Este resultado implica que cada modelo de mistura com fração de cura corresponde a algum modelo de mistura sem fração de cura. Em cada umas das situações consideradas - com e sem longa duração -, as funções de sobrevivências associadas às três ativações consideradas estão relacionadas.

Proposição 2.10 *Sendo $S_{mn}(t)$, $S_{al}(t)$ e $S_{mx}(t)$ as funções de sobrevivência das três ativações, dos modelos de mistura sem longa duração para o cenário de múltiplos riscos CR_{mn}, CR_{al} e CR_{mx} respectivamente, então*

$$S_{mn}(t) \leq S_{al}(t) \leq S_{mx}(t), \ \forall t > 0.$$

Prova 2.10 Seja $V_1 = \min\{T_1, T_2, \cdots, T_M\}$, $V_M = \max\{T_1, T_2, \cdots, T_M\}$. Sendo $K \in \{1, 2, \cdots, M\}$ então para $t > 0$ temos que $T_K > t \Rightarrow V_M > t \Rightarrow [T_K > t] \subset [V_M > t]$ desta forma temos que $P(T_K > t) \leq P(V_M > t)$ ou seja $S_{mx}(t) \geq S_{al}(t), \ \forall t > 0$. Com relação similar prova-se que $S_{mn}(t) \leq S_{al}(t), \ \forall t > 0$. ■

Desta forma, independentemente da distribuição base ou de ter longa duração, a relação é mantida.

As misturas aqui apresentadas permitem que facilmente se obtenha as funções que descrevem uma distribuição de probabilidade para qualquer distribuição densidade de probabilidade adotada para os tempos.

O objetivo de apresentar o aspecto das misturas, como elas são geridas e como se dá a relação entre o parâmetro da distribuição Geométrica e a distribuição de probabilidade atribuída aos tempos, muitas vezes camuflada quando apresentamos a função final, foi alcançado. Isto facilita a comparação entre as funções que descrevem a distribuição para contexto do mínimo e do máximo dos tempos. Por exemplo, observamos que a função densidade de probabilidade no contexto do mínimo utiliza a informação da função de sobrevivência dos tempos T_i, enquanto que no contexto do máximo dos tempos é utilizada a função distribuição acumulada e como essa diferença impacta nas funções de risco e sobrevivência.

Capítulo 3

Distribuições da família Exponencial Geométrica

Neste capítulo apresentamos três distribuições da família Exponencial Geométrica. Estas distribuições são derivadas assumindo estrutura latente de ativação para explicar a ocorrência do evento de interesse, tal qual apresentado na Seção 2.4 do capítulo anterior. Consideremos que o tempo até a ocorrência de cada uma das causas possíveis de serem ativadas tem distribuição Exponencial. Apresentamos propriedade e propomos um modelo de regressão. O método de estimação de máxima verossimilhança é utilizado para se obter as estimativas para conjuntos de dados reais e simulados.

3.1 Introdução

A distribuição Exponencial (ED) apresenta-se como solução eficaz, simples, elegante e fechada em inúmeros problemas associados ao tempo de vida e/ou estudos de confiabilidade (Adamidis & Loukas (1998), Kus (2007), Tahmasbi & Rezaei (2008)). Uma característica da distribuição Exponencial é ter função de risco constante, como consequência, em muitos casos, ela não fornece ajuste paramétrico razoável para aplicações sem essa característica. A fim de superar tal problema, novas classes de distribuições são introduzidas com base na modificação da ED. Temos por exemplo Adamidis & Loukas (1998), que introduzem a distribuição Exponencial Geométrica. Esta apresenta risco decrescente e é obtida considerando CR latentes. Gupta & Kundu (1999) propõem a ED

Generalizada, que acomoda dados com risco crescente e decrescente.

Neste mesmo enfoque Kus (2007) propõe a distribuição Exponencial-Poisson com taxa de falha decrescente, mesma característica da distribuição apresentada por Tahmasbi & Rezaei (2008) e denominada de distribuição Exponencial Logarítmica. Essas distribuições consideram a composição da distribuição Exponencial com as distribuições Poisson e Logarítmica, respectivamente, em um cenário de riscos competitivos (CR_{mn}). Chahkandi & Ganjali (2009) unificaram as distribuições propostas por Adamidis & Loukas (1998), Kus (2007) e Tahmasbi & Rezaei (2008), utilizando a composição da distribuição Exponencial com a distribuição Série de Potência.

Em trabalhos recentes, Cancho *et al.* (2010a) introduzem a distribuição Exponencial Poisson com taxa de falha crescente, esta distribuição tem como base o cenário de riscos complementares (CR_{mx}).

A formulação aqui apresentada inclui as distribuições propostas por Adamidis & Loukas (1998) e Louzada *et al.* (2011).

3.2 A distribuição Exponencial Geométrica com causas de risco latentes

No contexto apresentado na Seção 2.4 - M, $M = 1, 2, 3, \ldots$, v.a. que denota a quantidade de CR com f.d. dada por (2.1) - consideramos as v.a.s T_i - *i.i.d.*, $i = 1, \cdots, M$ - que denotam o tempo de ativação da *i-ésima* CR com distribuição Exponencial de parâmetro λ, assim $f_o(t) = f(t_i)$ dada por

$$f(t_i) = \lambda e^{-\lambda t_i} I_{[0,\infty)}(t_i), \tag{3.1}$$

resultando na distribuição composta Exponencial Geométrica com causas de risco latentes (EGcr). Abordamos, conforme modelagem formulada no Capítulo 2, as distribuições para as ativações de mínimo e máximo, que são as duas situações em que a distribuição originada se diferencia da distribuição base - Exponencial -.

A função densidade de probabilidade para a v.a. W que denota o tempo observado

tem suporte em $[0, \infty)$ e é função da v.a. T_i, cuja expressão é dada por

$$f_{\text{EG}}(W = w) = \begin{cases} \frac{\lambda\theta e^{-\lambda w}}{\left[1-(1-\theta)e^{-\lambda w}\right]^2}, & \text{para } W = \min(T_1, T_2, \cdots, T_M); \\ \\ \frac{\lambda\theta e^{-\lambda w}}{\left[\theta+(1-\theta)e^{-\lambda w}\right]^2}, & \text{para } W = \max(T_1, T_2, \cdots, T_M). \end{cases} \tag{3.2}$$

Em (3.2) a primeira equação refere-se à distribuição EG apresentada por Adamidis & Loukas (1998) e a denotamos por EG_{mn}. Para a última ativação temos a distribuição EG_{mx}, que será discutido na Seção 3.4 (Louzada *et al.*, 2011). Ressaltamos que Adamidis & Loukas (1998) consideram $1 - \theta$ como a probabilidade de sucesso da distribuição Geométrica.

A Figura 3.1 dispõe os gráficos da f.d.p. das respectivas distribuições para o conjunto de valores $\theta = 0,001;\ 0,01;\ 0,2;\ 0,5;\ 0,7$ e $0,99$ com $\lambda = 1$.

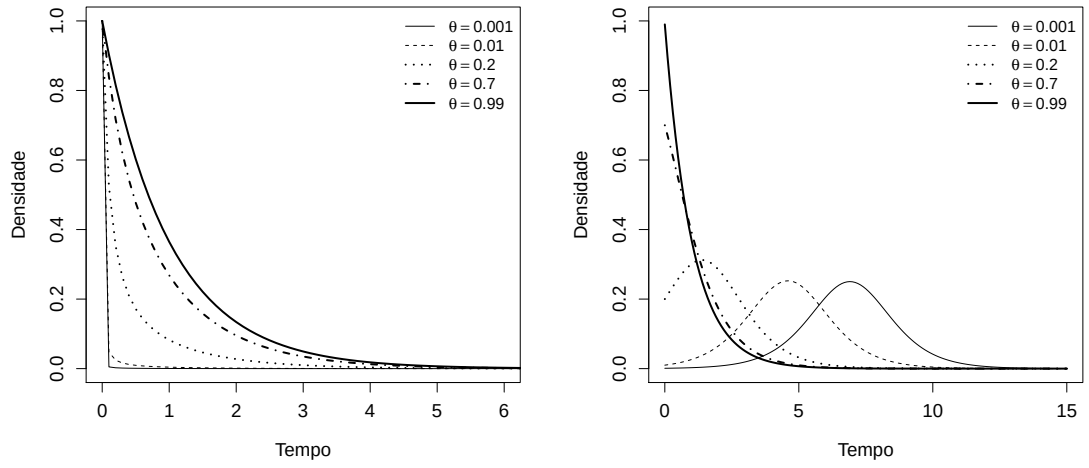

FIGURA 3.1: Função densidade de probabilidade das distribuições EG_{mn} (esquerdo) e EG_{mx} (direito), variando θ e com $\lambda = 1$.

Na distribuição EG_{mx} o parâmetro λ controla a escala e θ é o parâmetro de forma da f.d.p. tornando-a decrescente se $\theta \geq 1/2$ e unimodal para $\theta < 1/2$.

As funções de sobrevivência associadas a estas distribuições são dada, respectivamente por

$$S_{\text{EG}}(W = w) = \begin{cases} \frac{\theta e^{-\lambda w}}{1-(1-\theta)e^{-\lambda w}}, & \text{para } W = \min(T_1, T_2, \cdots, T_M); \\ \\ \frac{e^{-\lambda w}}{\theta+(1-\theta)e^{-\lambda w}}, & \text{para } W = \max(T_1, T_2, \cdots, T_M). \end{cases} \tag{3.3}$$

Na Figura 3.2 estão apresentadas algumas curvas da função de sobrevivência para diferentes valores de θ, considerando $\lambda = 1$, pois seu valor interfere apenas na escala. Observamos que a função de sobrevivência para o mínimo dos tempos atinge seu limite inferior em um intervalo de tempo menor se comparada com a mesma função para o máximo dos tempos.

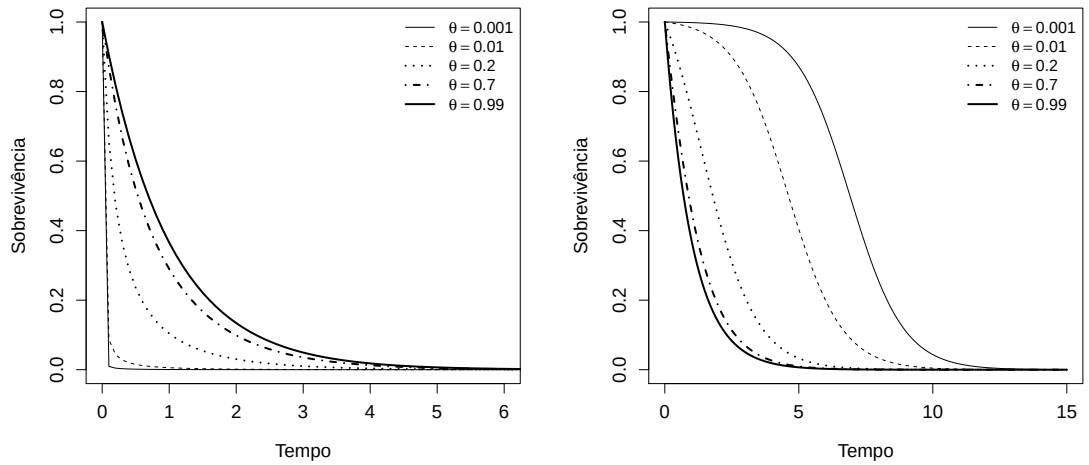

FIGURA 3.2: Gráficos das funções de sobrevivência das distribuições EG_{mn} (esquerdo) e EG_{mx} (direito), variando θ e com $\lambda = 1$.

A distribuição EG_{al}, em concordância ao afirmado na Seção 2.4, é a própria distribuição Exponencial, distribuição de probabilidade dos tempos T_i's.

3.3 Relação entre os esquemas de ativação

Nesta seção são apresentados alguns aspectos comparativos entre os esquemas de ativação CR_{mn}, CR_{mx} e CR_{al}.

Primeiramente vamos apresentar as funções de risco (taxa de falha) para os diferentes esquemas, lembrando que iremos preterir as funções relativas à ativação aleatória (CR_{al}) por ser a mesma da distribuição Exponencial. A função de risco para a distribuição

EG é dada por

$$h_{\mathrm{EG}}(W=w)=\begin{cases}\frac{\lambda}{1-(1-\theta)e^{-\lambda w}}, & \text{para } W=\min(T_1,T_2,\cdots,T_M);\\ \\ \frac{\lambda\theta}{\theta+(1-\theta)e^{-\lambda w}}, & \text{para } W=\max(T_1,T_2,\cdots,T_M).\end{cases} \tag{3.4}$$

A função taxa de falha para a última ativação (EG_{mx}) é crescente no tempo (w), enquanto que para a primeira ativação (EG_{mn}) é decrescente.

O comportamento das funções de risco das distribuições EG_{mn} e EG_{mx} dadas em (3.4) e a função de risco da distribuição EG_{al} estão relacionados conforme proposição a seguir, sendo que $h_{EG_{al}}(w)=\lambda$.

Proposição 3.1 *(a) Quando $w\to 0$, a função de risco da distribuição EG_{mx} converge para $\lambda\theta$ e a função de risco da distribuição EG_{mn} converge para $\dfrac{\lambda}{\theta}$. (b) As funções de risco das distribuições EG_{mn} e EG_{mx} convergem para a função de risco da distribuição EG_{al} quando $w\to\infty$*

Prova 3.1 (a) Considerando $w\to 0$ e a última ativação (EG_{mx}), temos

$$\lim_{w\to 0}h(w)=\lim_{w\to 0}\frac{\lambda\theta}{\theta+(1-\theta)e^{-\lambda w}}=\frac{\lambda\theta}{\theta+(1-\theta)}=\lambda\theta,$$

enquanto que para a primeira ativação (EG_{mn}) temos

$$\lim_{w\to 0}h(y)=\lim_{w\to 0}\frac{\lambda}{1-(1-\theta)e^{-\lambda w}}=\frac{\lambda}{1-(1-\theta)}=\frac{\lambda}{\theta}.$$

(b) Quando $w\to\infty$, para a distribuição EG_{mn} temos

$$\lim_{w\to\infty}h(w)=\lim_{w\to\infty}\frac{\lambda}{1-(1-\theta)e^{-\lambda w}}=\lambda,$$

e para a distribuição EG_{mx},

$$\lim_{w\to\infty}h(w)=\lim_{w\to\infty}\frac{\lambda\theta}{\theta+(1-\theta)e^{-\lambda w}}=\frac{\lambda\theta}{\theta}=\lambda. \quad \blacksquare$$

Corroborando com a Proposição 3.1, a convergência para λ das funções de risco de EG_{mn} e EG_{mx}, quando $w\to\infty$, também é evidenciada na Figura 3.3, que apresenta o comportamento da taxa de falha dos dois esquemas de ativação. Fixamos $\lambda=1$ e estudamos o comportamento da função de risco para alguns valores de θ. E também

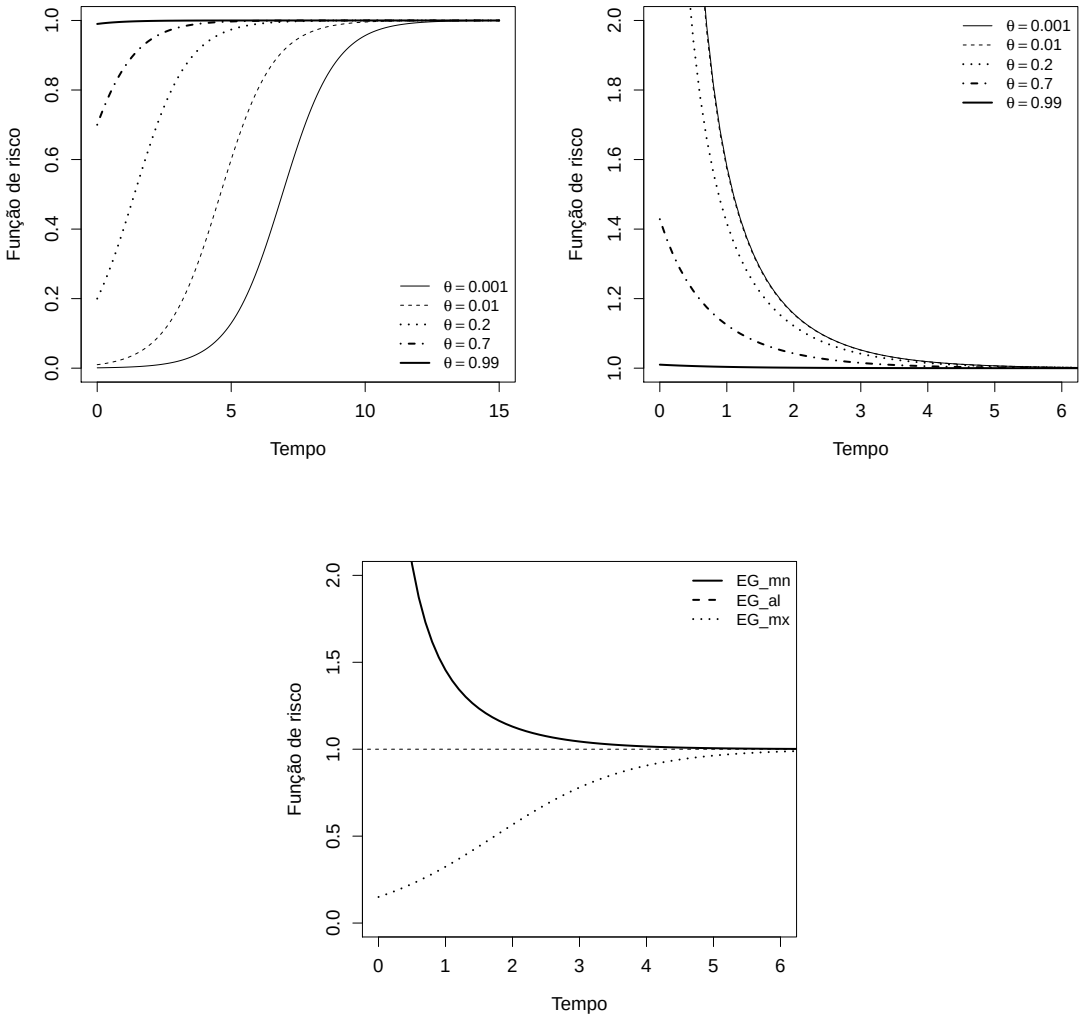

FIGURA 3.3: Superior: Função de risco das distribuições EG_{mx} (esquerda), EG_{mn} (direita) com $\lambda = 1$. Inferior: comparação da função de risco das três distribuições para $\theta = 0,15$ e $\lambda = 1$.

consideramos θ fixo para efeitos de comparação entre as duas situações - Figura 3.3 painel inferior.

As funções de sobrevivência para as distribuições $EGcr$ nos diferentes esquemas de ativação estão relacionadas, conforme apresenta a Proposição 2.10 para o caso geral, a qual está graficamente apresentada na Figura A.2 do Apêndice.

3.3.1 Estudo de simulação

No processo de simulação consideramos cinco tamanhos amostrais: $n = 10, 20, 30, 50$ e 100 e quatro níveis de censura: *percentual de censura* $= 0,1;\ 0,2;\ 0,35$ e $0,50$.

Fixados o tamanho amostral e o percentual de censura, obtivemos 1.000 amostras da distribuição EG_{mx} utilizando a função quantil

$$Q_{mx}(u) = F_{mx}^{-1}(u) = -\lambda^{-1} \log\left(\frac{\theta(1-u)}{\theta(1-u)+u}\right),$$

sendo $\lambda = 0,5$, $\theta = 0,25$ e $u \sim U(0,1)$. Para cada amostra ajustamos a distribuição que originou os dados (EG_{mx}) e sua complementar (EG_{mn}) e julgamos a distribuição que melhor se ajustou utilizando o critério AIC. Ao final das 1.000 amostras obteve-se o percentual de vezes que a distribuição utilizada para gerar os dados foi o que melhor se ajustou. Repetimos o procedimento para cada uma das 20 combinações entre tamanho amostral e percentual de censura, considerando a distribuição EG_{mx} para gerar os dados. Procedimento similar foi adotado para a distribuição EG_{mn}, sendo as amostras obtidas pela função quantil

$$Q_{mn}(u) = F_{mn}^{-1}(u) = -\lambda^{-1} \log\left(\frac{u-1}{u(1-\theta)-1}\right).$$

A indicação de censura correspondeu a uma amostra de tamanho n da distribuição Bernoulli cuja probabilidade de fracasso corresponde ao percentual de censura fixado.

A estimação dos parâmetros dos modelos foi obtida no programa `R` (R Core Team, 2011) utilizando a rotina *optim* que encontrar a solução que minimiza o negativo do logaritmo da função de verossimilhança. O método optimização Nelder-Mead foi utilizado com vetor nulo como valores iniciais do processo.

A Tabela 3.1 dispõe do percentual de vezes que a distribuição da qual os dados foram gerados foi a que melhor se ajustou aos mesmos. Observamos que mesmo mediante grandes percentuais de censura é possível diferenciar as distribuições em todos os tamanhos de amostras considerados; e mesmo adotando valores complementares para o parâmetro θ para não comprometer a comparação entre os resultados, a distribuição EG_{mx} mostrou-se mais eficiente na identificação das amostras geradas de seu modelo,

principalmente ao considerarmos tamanhos amostrais pequenos e elevados níveis de censura.

TABELA 3.1: Percentual de vezes que a distribuição que gera os dados (EG_{mn}/EG_{mx}) proporcionou um melhor ajuste aos mesmos pelo critério AIC.

Percentual	Tamanho amostral				
de censura	10	20	30	50	100
$0,10$	$0,726/0,998$	$0,875/0,999$	$0,936/0,999$	$0,988/0,999$	$0,999/0,999$
$0,20$	$0,711/0,995$	$0,846/0,999$	$0,923/0,999$	$0,980/0,999$	$0,999/0,999$
$0,35$	$0,672/0,984$	$0,837/0,999$	$0,909/0,999$	$0,971/0,999$	$0,999/0,999$
$0,50$	$0,648/0,978$	$0,791/0,992$	$0,880/0,999$	$0,953/0,993$	$0,995/0,999$

3.4 Distribuição Exponencial Geométrica para o máximo dos tempos

Algumas das mais importantes características de uma distribuição podem ser estudadas através de seus momentos. A expressão geral do *r-ésimo* momento $\mu_r^{'} = E(Y^r)$ da distribuição EG_{mx} é obtido analiticamente como segue:

Proposição 3.2 *Seja Y uma v.a. com distribuição EG_{mx}. Sua função característica é dada por*

$$\Phi_Y(t) = \frac{\lambda}{\theta(\lambda - it)}\Psi(2, 1 - \frac{it}{\lambda}, 2 - \frac{it}{\lambda}, -\beta), \tag{3.5}$$

em que, $\beta = \frac{1-\theta}{\theta}$ e $i^2 = -1$.

Prova 3.2 Temos que

$$\begin{aligned}\Phi_Y(t) = \int_0^\infty e^{ity} f(y)dy &= \int_0^\infty e^{ity}\frac{\theta\lambda e^{-\lambda y}}{\left[e^{-\lambda y}(1-\theta)+\theta\right]^2}dy \\ &= \frac{1}{\theta}\int_0^\infty e^{ity}\frac{\lambda e^{-\lambda y}}{\left[1+e^{-\lambda y}\frac{(1-\theta)}{\theta}\right]^2}dy = \frac{1}{\theta}\int_0^1 \frac{u^{-\frac{it}{\lambda}}}{(1+\frac{1-\theta}{\theta}u)^2}du,\end{aligned}$$

em que $u = e^{-\lambda y}$.

Comparando a integral acima com o resultado

$\Psi(a,b,c,z) = \dfrac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \displaystyle\int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a} dt$ - Abramowitz & Stegun (1972), *p.*558 -,

temos $b = 1 - \frac{it}{\lambda}$, $c = 2 - \frac{it}{\lambda}$, $-z = \beta = \frac{1-\theta}{\theta}$ e $a = 2$. Fazendo as substituições apropriadas a prova está concluída. ■

Proposição 3.3 *Se a variável aleatória Y tem distribuição EG_{mx} e $r \in N$, então*

$$E(Y^r) = -\frac{r!}{\lambda^r(1-\theta)} L\left(-\beta, r\right), \tag{3.6}$$

em que $\beta = \dfrac{1-\theta}{\theta}$, *e* $L(-\beta, r) = \displaystyle\sum_{k=1}^{\infty} \frac{(-\beta)^k}{k^r}$.

Prova 3.3 Sendo t um número real e $\lambda > 0$, considerando a relação

$$\frac{1}{(\lambda - it)} \Psi(2, 1 - \frac{it}{\lambda}, 2 - \frac{it}{\lambda}, -\beta) = \sum_{k=0}^{\infty} \frac{(k+1)(-\beta)^k}{\lambda(k+1) - it} \tag{3.7}$$

$-\infty < t < \infty$, $i = \sqrt{-1}$ e $1- < \beta < 1$, (Jodrá, 2008), a função característica (3.5) pode ser escrita como

$$\Phi_Y(t) = \frac{\lambda}{\theta} \sum_{k=0}^{\infty} \frac{(k+1)(-\beta)^k}{\lambda(k+1) - it}, \tag{3.8}$$

de forma que sua *r-ésima* derivada é

$$\Phi_Y^r(t) = \frac{\lambda \Gamma(r+1) i^r}{\theta} \sum_{k=0}^{\infty} \frac{(k+1)(-\beta)^k}{(\lambda(k+1) - it)^{r+1}}, \qquad r = 1, 2, \ldots, \tag{3.9}$$

fazendo $t = 0$ e sendo $E(Y^r) = \Phi_Y^{(r)}(0)/i^r$, segue que

$$E[Y^r] = -\frac{\Gamma(r+1)}{(1-\theta)\lambda^r} \sum_{k=1}^{\infty} \frac{(-\beta)^r}{k^r} = -\frac{r!}{(1-\theta)\lambda^r} L(-\beta, r),$$

para $\beta \in (0, \infty)$ e $r = 1, 2, \cdots$. ■

Proposição 3.4 *A variável aleatória Y com distribuição EG_{mx} tem média e variância dadas respectivamente por*

$$E(Y) = \frac{-\log(\theta)}{\lambda(1-\theta)} \qquad e \qquad Var(Y) = -\frac{1}{\lambda^2(1-\theta)} \left(2L(-\beta, 2) + \frac{[\log(\theta)]^2}{1-\theta} \right). \tag{3.10}$$

Prova 3.4 Basta utilizar o resultado $L(-\beta, 1) = -\log(1+\beta)$ (Adamidis & Loukas, 1998) aplicado na Proposição 3.3. ■

Sendo que a variância é função da média, nas situações em que esta última assume valores grandes a variância pode não ser representativa, então o coeficiente de variação CV apresenta-se como alternativa. Logo, temos que

$$CV = -\frac{1-\theta}{\log(\theta)}\sqrt{-\frac{2L(-\beta,2)}{1-\theta} - \left(\frac{\log(\theta)}{1-\theta}\right)^2}. \tag{3.11}$$

A moda para a v.a. Y com distribuição EG_{mx} é $\widetilde{Y} = \frac{1}{\lambda}\log\frac{1-\theta}{\theta}$.

A distribuição Exponencial Geométrica para o mínimo dos tempos EG_{mn} tem suas particularidades e propriedades abordadas em Adamidis & Loukas (1998).

3.4.1 Inferência

Assumimos que os tempos de vida são v.a. *i.i.d.* oriundos de um mecanismo aleatório de censura à direita e δ_i é o indicador de censura (0 se censura, 1 se observado). O logaritmo da função verossimilhança para a distribuição EG_{mx} é dado por

$$\ell(\theta,\lambda) = \log(\lambda\theta)\sum_{i=1}^{n}\delta_i - \lambda\sum_{i=1}^{n}y_i - \sum_{i=1}^{n}(\delta_i+1)\log\left(e^{-\lambda y_i}(1-\theta)+\theta\right). \tag{3.12}$$

A Figura A.1 (Apêndice) dispõe o gráfico de curvas de nível da função $\ell(\theta,\lambda)$, indicando que ela apresenta um único ponto de máximo.

Os componentes do vetor escore $U(\theta,\lambda) = U(\boldsymbol{\vartheta}) = (\partial\ell(\boldsymbol{\vartheta})/\partial\theta, \partial\ell(\boldsymbol{\vartheta})/\partial\lambda)$ são dados por

$$\frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\theta} = \frac{\sum_{i=1}^{n}\delta_i}{\theta} + \sum_{i=1}^{n}(\delta+1)\frac{e^{-\lambda y_i}+1}{e^{\lambda y_i}(1-\theta)+\theta},$$

$$\frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\lambda} = \frac{\sum_{i=1}^{n}\delta_i}{\lambda} - \sum_{i=1}^{n}y_i + \sum_{i=1}^{n}(\delta+1)\frac{y_i e^{-\lambda y_i}(1-\theta)}{e^{\lambda y_i}(1-\theta)+\theta}.$$

Os EMVs - $\hat{\theta}$ e $\hat{\lambda}$ - para os parâmetros θ e λ não podem ser obtidos explicitamente pela solução do sistema de equações $U(\theta,\lambda) = \mathbf{0}$, sendo obtidos pela maximização direta da função log-verossimilhança pela rotina *optim* do `R` (R Core Team, 2011), considerando as mesmas particularidades descritas na Seção 3.3.1.

Sob certas condições de regularidade para a função de verossimilhança e com o tamanho da amostra grande, a distribuição assintótica para os EMVs é a distribuição Normal com média ϑ e matriz de covariâncias $I^{-1}(\vartheta)$, em que $I(\vartheta)$ é a matriz de informação de Fisher, permitindo que o intervalo de confiança e o teste de hipóteses possam ser obtidos. Em muitos casos, $I(\vartheta)$ não tem forma fechada e dificilmente é calculada. Uma solução é usar a negativa da matriz hessiana da função log-verossimilhança avaliada no ponto $\vartheta = \hat{\vartheta}$, que é um estimador consistente para a matriz de covariâncias assintótica (Mudholkar *et al.*, 1996).

Para a comparação de modelos encaixados - distribuição EG_{mx} com a ED - consideramos as hipóteses $H_0 : \theta = 1$ contra $H_1 : \theta \neq 1$, e o teste da razão de verossimilhanças, cuja estatística é

$$w_n = 2\{\ell_{EG_{mx}} - \ell_{ED}\}, \tag{3.13}$$

em que ℓ_{ED} e $\ell_{EG_{mx}}$ são os logaritmos das funções de verossimilhança para os modelos sob a hipótese restrita H_0 e sob a hipótese irrestrita H_1, respectivamente, para uma amostra de tamanho n.

Como o teste é realizado no limite do espaço paramétrico, seguindo Maller & Zhou (1995), presume-se que w_n assintoticamente assume uma mistura de distribuição Qui-Quadrado com 1 grau de liberdade e ponto de massa no zero. Então, $\lim_{n\to\infty} P(w_n \leq c) = 1/2 + 1/2\ P(\chi^2_1 \leq c)$, em que $P(\chi^2_1 \leq c)$ denota uma variável aleatória com uma distribuição Qui-Quadrado com 1 grau de liberdade. Valores positivos elevados de w_n dão evidência favorável ao modelo completo.

3.4.2 Simulação

Visamos estudar o comportamento assintótico dos estimadores de máxima verossimilhança pela quantificação da probabilidade de cobertura dos intervalos assintóticos, em situações nas quais observam-se dados censurados; avaliar o decaimento da variância dos estimadores com o aumento do número de amostras e se este decaimento interfere no probabilidade de cobertura dos intervalos de confiança assintóticos.

Fixamos o vetor paramétrico $\vartheta = (\theta; \lambda) = (0,25; 1)$ e geramos 1000 repetições para cada uma das combinações entre tamanho amostral e percentual de censura con-

siderados. Obtivemos as estimativas $\hat{\vartheta}$ de ϑ pela maximização direta da função $\ell(\theta;\lambda)$ (rotina *optim* do `R`). Calculamos os intervalos normais assintóticos de 95% de confiança para ϑ e a variância dos estimadores para cada uma das repetições. A probabilidade de cobertura empírica foi obtida calculando-se a proporção de intervalos para os quais o valor fixado para ϑ pertence ao intervalo, e esta foi superior ao valor nominal para todas as combinações consideradas. A variância dos estimadores diminuiu com o aumento do tamanho amostral (ver Tabela 3.2), mas devido ao alto percentual de probabilidade de cobertura para todas as combinações de tamanho amostral e censuras, não foi possível verificar sua influência no mesmo.

TABELA 3.2: Valor médio das variância das EMVs para 1000 repetições amostrais em cada nível.

Parâmetro	Percentual	Tamanho amostral				
	de censura	10	20	30	50	100
	0,10	0,1965	0,1052	0,0726	0,0435	0,0213
λ^*	0,20	0,2386	0,1292	0,0900	0,0526	0,0257
	0,35	0,3874	0,1901	0,1290	0,0758	0,0359
	0,10	1693955	197285	65832	3822	0,2350
θ^*	0,20	4098119	628442	246994	6185	0,2717
	0,35	5489072	931602	86555	28036	0,3420

$\lambda^* = e^{\lambda}$, $\theta^* = \log \frac{\theta}{1-\theta}$.

3.4.3 Aplicação

Comparamos o ajuste das distribuições EG_{mx} com EG_{mn} para quatro conjuntos de dados da literatura. Os conjuntos de dados considerados são: T1, T2, T3 e T5, que estão apresentados na Seção 1.3, sendo dois deles com censura.

A forma da função de risco foi identificada mediante o gráfico TTT, que indica forma crescente para os conjuntos $T1$ e $T2$. Portanto, presumimos que eles podem ser convenientemente acomodados pela distribuição EG_{mx}. Para $T3$ e $T5$, a forma da função de risco é decrescente e possivelmente acomodada pela distribuição EG_{mn}.

Nos ajustes das distribuições, a convergência da rotina empregada foi rápida, sem necessidade de interferência na escolha valor inicial do processo. A Tabela 3.3 fornece os valores dos critérios $-\ell(\widehat{\boldsymbol{\vartheta}}_g)$, AIC e BIC para ambas distribuições, as quais apontam como a melhor escolha para ajuste a distribuição EG_{mx} para os conjuntos $T1$ e $T2$ e EG_{mn} para os conjuntos $T3$ e $T5$, concordando com a indicativa dada pela forma da função de risco dos dados com a distribuição proposta. A supremacia dos ajustes é visível no gráfico da Figura 3.4 em que se encontram as estimavas de sobrevivência por KM, sobrepostas pelas curvas das funções de sobrevivência estimadas das das distribuições ajustadas.

TABELA 3.3: Critérios $-\ell(\widehat{\boldsymbol{\vartheta}}_g)$, AIC e BIC para os ajustes das distribuições EG_{mn} e EG_{mx}.

Dados	EG_{mn}			EG_{mx}		
	$-\ell(\boldsymbol{\vartheta})$	AIC	BIC	$-\ell(\boldsymbol{\vartheta})$	AIC	BIC
T1	121,43	246,87	249,15	114,35	232,70	234,97
T2	860,87	1725,74	1731,66	805,47	1614,95	1620,87
T3	1175,92	2355,85	2362,57	1178,76	2361,52	2368,25
T5	219,31	442,62	446,60	223,92	451,84	455,22

As estimativas de máxima verossimilhança (EMVs) e seus respectivos desvios padrões (entre parênteses) dos parâmetros da distribuição EG_{mx} são $\hat{\lambda} = 0,0435(0,0096)$ e $\hat{\theta} = 0,0554(0,0449)$ para $T1$; $\hat{\lambda} = 0,0083(0,0007)$ e $\hat{\theta} = 0,0194(0,0076)$ para $T2$.

Para a distribuição EG_{mn}, temos $\hat{\lambda} = 0,0080(0,0013)$ e $\hat{\theta} = 0,4736(0,1432)$ para $T3$; $\hat{\lambda} = 0,0021(0,0016)$ e $\hat{\theta} = 0,1574(0,1267)$ para o conjunto $T5$.

A comparação entre as distribuições EG_{mx} e ED é realizada via teste de hipótese para os dados $T1$ e $T2$. Os valores da estatística do teste, w_n, são iguais a $14,178$ e $110,788$, respectivamente para $T1$ e $T2$, cujos valores são expressivamente maiores que o valor de referência $\frac{1}{2} + \frac{1}{2}P(\chi^2_1 \leq c) = 2,42$, fornecendo forte evidência a favor da distribuição EG_{mx} para ambos os conjuntos de dados.

Outra questão que pode surgir é a utilização de distribuições tradicionais para dados de sobrevivência, como a distribuição Weibull. Nos conjuntos de dados $T1$ e $T2$, a comparação das distribuições Weibull e EG_{mx} não pode ser feita via teste de hipótese por pertencerem a famílias distintas de hipóteses (Cox, 1961), mas pode ser avaliada pelos

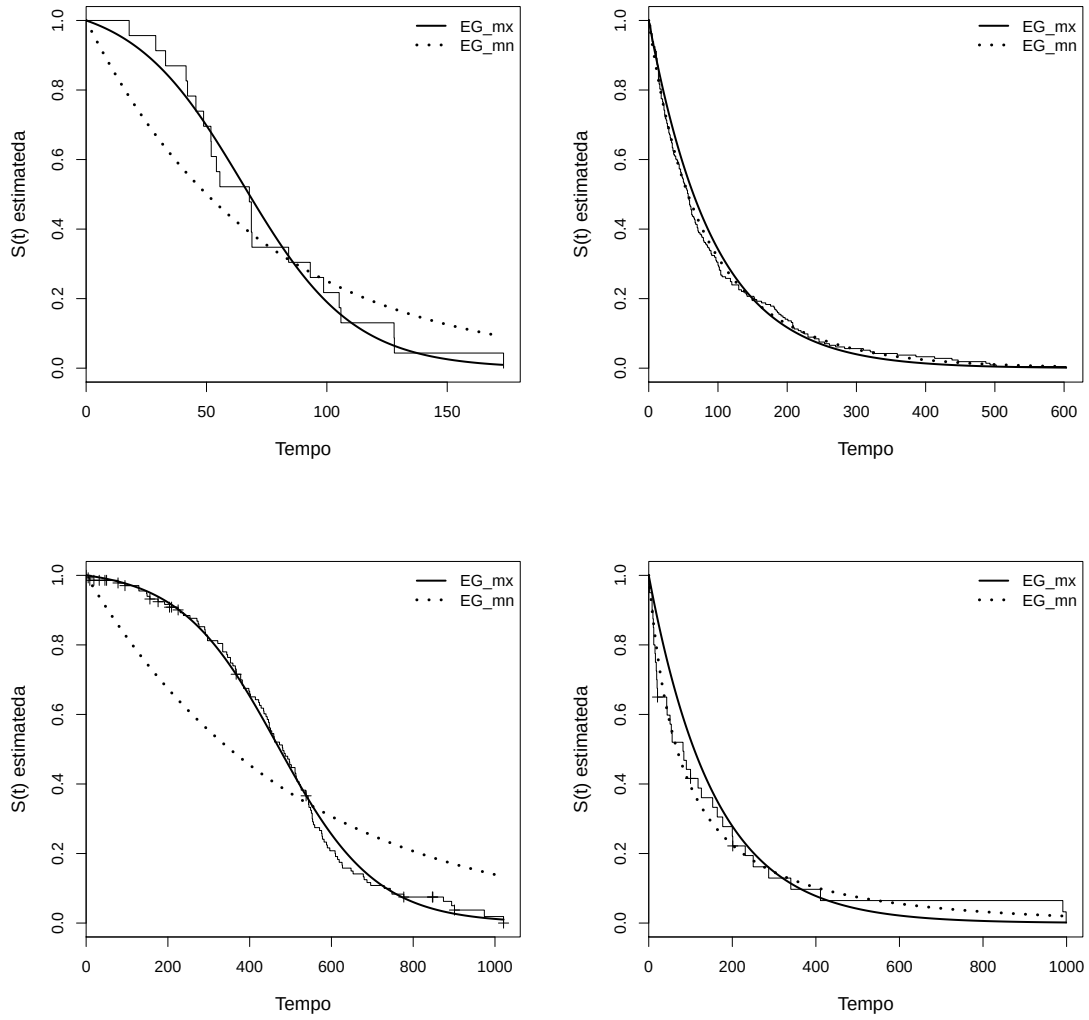

FIGURA 3.4: Curvas de Kaplan-Meier com função de sobrevivência estimada considerando os ajustes das distribuições EG_{mn} e EG_{mx} nos conjuntos de dados. Painéis: superior esquerdo ($T1$), superior direito ($T3$), inferior esquerdo ($T2$) e inferior direito ($T5$).

critérios AIC e BIC. Para a distribuição Weibull no conjunto $T1$, obtivemos $231,36$ e $233,63$, respectivamente para AIC e BIC e para $T2$ $1.630,52$ e $1.636,46$. Esses resultados dão evidência favorável à distribuição Weibull para $T1$ e à distribuição EG_{mx} para $T2$, mostrando a importância da distribuição EG_{mx}, que pode ser considerada como uma distribuição concorrente às distribuições usuais.

3.4.4 Especificação do modelo de regressão

Como a inclusão de covariáveis no modelo podem impactar de forma diferente dependendo do parâmetro escolhido para sua inclusão, consideramos três diferentes formas de parametrização que consideram tanto a taxa de falha da distribuição dos tempos quanto a média da distribuição que representa o número de CRs.

Parametrização I

Neste caso as covariáveis são inseridas no parâmetro λ que representa a taxa de falha da distribuição Exponencial, distribuição dos tempos de ocorrência de cada uma das CRs.

Considerando $\lambda = \lambda(z_i) = e^{(z_i^{'}\beta)}$, a função log-verossimilhança para os modelos EG_{mn} e EG_{mx} será dada respectivamente pelas equações (3.14) e (3.15).

$$\ell_{1mn}(\boldsymbol{\vartheta};\mathbf{D}) \propto \sum_{i=1}^{n}\left[\delta_i z_i^{'}\beta + \log\theta - x_i e^{z_i^{'}\beta} - (\delta_i+1)\log\left(1-(1-\theta)e^{-x_i e^{(z_i^{'}\beta)}}\right)\right]. \tag{3.14}$$

$$\ell_{1mx}(\boldsymbol{\vartheta};\mathbf{D}) \propto \sum_{i=1}^{n}\left[\delta_i z_i^{'}\beta + \delta_i\log\theta - y_i e^{z_i^{'}\beta} - (\delta_i+1)\log\left(\theta+(1-\theta)e^{-y_i e^{(z_i^{'}\beta)}}\right)\right]. \tag{3.15}$$

Parametrização II

Nesta parametrização as covariáveis são inseridas na média μ da distribuição Geométrica, em que $\mu = \frac{1}{\theta}$, fazendo $\mu = e^{(z_i^{'}\beta)}$. Neste caso, a função log-verossimilhança para os dois casos considerados é dada por (3.16) e (3.17).

$$\ell_{2mn}(\boldsymbol{\vartheta};\mathbf{D}) \propto \sum_{i=1}^{n}\left[-z_i^{'}\beta + \delta_i\log\lambda - \lambda x_i - (\delta_i+1)\log\left(1-(1-e^{z_i^{'}\beta})e^{-\lambda x_i}\right)\right]. \tag{3.16}$$

$$\ell_{2mx}(\boldsymbol{\vartheta};\mathbf{D}) \propto \sum_{i=1}^{n}\left[\delta_i\log\lambda - \delta_i z_i^{'}\beta - \lambda y_i - (\delta_i+1)\log\left(e^{z_i^{'}\beta}+\left(1-e^{z_i^{'}\beta}\right)e^{-\lambda y_i}\right)\right]. \tag{3.17}$$

Parametrização III

Na terceira situação, consideramos a combinação do modelo I e II, em que algumas covariáveis são indexadas nos modelos através do parâmetro taxa de falha da distribuição Exponencial ou na média da distribuição Geométrica.

3.4.5 Aplicação do modelo de regressão

Aos conjuntos $T6$ e $T4$ ajustamos os modelos EG_{mx} e EG_{mn}, respectivamente, considerando

$$\log[\lambda(\mathbf{z}_i)] = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} + \beta_5 z_{i5}, \text{ e } \theta = \frac{e^{\phi}}{1+e^{\phi}},$$

para a parametrização I, e

$$\log[\mu(\mathbf{z}_i)] = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} + \beta_5 z_{i5}, \text{ e } \lambda = e^{\gamma},$$

para a parametrização II.

Para a parametrização III, temos que

$$\log[\mu(\mathbf{z}_i)] = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} \text{ e } \log[\lambda(\mathbf{z}_i)] = \beta_0^* + \beta_4 z_{i4} + \beta_5 z_{i5}.$$

A Tabela 3.4 dispõe das EMVs para as reparametrizações $\lambda = e^{\gamma}$ e $\theta = \frac{e^{\phi}}{1+e^{\phi}}$, com seus erros padrões para os modelos ajustados em cada conjunto de dados, juntamente com o valor-p indicando significância para os três parâmetros nos três modelos considerados no conjunto $T6$, sendo que a única covariável significativa foi a relacionada ao tempo de estada do paciente com o trabalho que a clínica teve para mantê-lo, representado por β_2.

Em termos de qualidade de ajuste, a parametrização III mostrou-se mais adequada que a I e a II e seus respectivos valores para $-\ell(\boldsymbol{\vartheta})$ em relação ao ajuste do conjunto de dados $T6$ foram 2016,147; 2029,724 e 2016,182. Para o conjunto de dados $T4$ tivemos também a predominância da parametrização III com $-\ell(\boldsymbol{\vartheta}) = 24,284$, enquanto as parametrizações I e II obtiveram os valores 31,421 e 26,122, respectivamente.

Os gráficos Quantil-quantil para os modelos ajustados nas diferentes parametrizações estão na Figura A.3 (Apêndice) e indicam bons ajustes.

3.5 Comentários

A partir do estudo de simulação, observamos que os modelos EG_{mn} e EG_{mx} conseguem discriminar as amostras por eles geradas, mesmo que estas sejam pequenas e com níveis de censuras elevados. Na avaliação dos gráficos TTT, observamos que os dados

TABELA 3.4: Estimativas de máxima verossimilhança para os modelos EG_{mx} para $T6$ e EG_{mn} para $T4$.

Parame-		T6			T4		
trização	Parâmetro	Estimativa	EP	p-valor	Estimativa	EP	p-valor
	ϕ	9,4941	0,3152	0,0000	1,1719	1,3529	0,3863
	β_0	-4,0957	0,2090	0,0000	2,7314	0,7604	0,0003
	β_1	-0,0001	0,0021	0,9469	-0,0364	0,0051	0,0000
I	β_2	-0,4951	0,0204	0,0000	-0,0015	0,0096	0,8738
	β_3	-0,0215	0,0291	0,4597	0,0010	0,0089	0,9028
	β_4	0,0173	0,0356	0,6275	0,1286	0,0814	0,1141
	β_5	0,0163	0,0133	0,2215	0,1627	0,1924	0,3976
	γ	0,0149	0,0000	0,0000	0,5705	0,2146	0,0078
	β_0	8,7028	1,0999	0,0000	-0,7044	0,3100	0,0230
	β_1	-0,0001	0,0116	0,9935	0,0144	0,0058	0,0143
II	β_2	5,3818	0,3026	0,0000	0,0124	0,0051	0,0164
	β_3	0,2504	0,2798	0,3709	0,0042	0,0031	0,1705
	β_4	-0,1886	0,3358	0,5742	0,1099	0,0438	0,0121
	β_5	-0,1786	0,1141	0,1176	-0,2918	0,1183	0,0136
	β_0	8,7103	0.5351	0,0000	-0,3404	0,1658	0,0400
	β_1	0,0003	0,0012	0,7902	0,0118	0,0047	0,0137
	β_2	5,3830	0,3177	0,0000	0,0033	0,0018	0,0785
III	β_3	0,2466	0,2772	0,3736	-0,0006	0,0022	0,7781
	β_0^*	-4,2316	0,1113	0,0000	0,3173	0,4913	0,5184
	β_4	-0,1981	0,3325	0,5512	0,0746	0,0932	0,4235
	β_5	-0,1824	0,1171	0,1192	-0,0108	0,2219	0,9608

EP: Erro padrão

de risco crescentes (decrescentes), mesmo com elevados percentuais de censura, tendem a serem melhores ajustados por modelos que também apresentam a forma de risco crescente (decrescente). A distribuição EG_{mx} pode ser considerada como uma opção no ajuste de dados de sobrevivência, pois concorre com modelos já tradicionais nessa área, como o Weibull. Além disso, dispõe apenas de dois parâmetros e não há dificuldades de estimação

no contexto clássico e com boas propriedades.

Ao considerarmos as diferentes parametrizações para indexar as covariáveis, observamos que na parametrização III os ajustes obtidos não foram bons quando agrupamos covariáveis verdadeiramente quantitativas com as de cunho qualitativo no mesmo parâmetro, por isso optamos pela seriação, a saber: em λ indexamos as covariáveis z_{i4}, z_{i5} para $T4$ e z_{i1} para $T6$, as demais foram indexadas em μ.

Capítulo 4

Distribuições Exponenciais Geométricas de longa duração

Este capítulo apresenta o modelo de mistura de longa duração BG para as distribuições EG_{mn} e EG_{mx}, representadas, respectivamente, pelas siglas LEG_{mn} e LEG_{mx}. Tecemos comparações entre as distribuições obtidas e relatamos algumas propriedades. Em termos de ajustes, aplicamos os modelos em dados da área médica e finanças e comparamo-os com distribuições de longa duração BG provenientes das distribuições Exponencial, Weibull e Log-Logística.

4.1 Introdução

Uma dificuldade na modelagem tradicional surge quando parte da população não é suscetível ao evento de interesse. Na área médica esse subconjunto corresponde ao grupo dos indivíduos imunes ao problema considerado ou, aos que uma vez acometidos pela doença, desenvolvem meios de resistir ou até mesmo vencer o problema, sendo considerados curados.

Modelos que consideram essa divisão na população são chamados de modelos de longa duração ou modelos de cura. O trabalho precursor é apresentado por Boag (1949), que utiliza o método de máxima verossimilhança para estimar a proporção de cura em um experimento referente ao câncer de mama. Embasados nesse conceito, Berkson &

Gage (1952) propõem um modelo de mistura para estimar essa proporção considerando um tratamento de câncer de estômago.

Muitos outros trabalhos têm acrescido à teoria dos modelos de mistura de longa duração, em que assume-se que a proporção de indivíduos imunes é p. Igualmente a Farewell & Prentice (1977), Farewell (1982), Greenhouse & Wolf (1984), Ghitany & Maller (1992), Ghitany *et al.* (1994), Maller & Zhou (1995), MacKenzie (1996), Chen & Ibrahim (2001), Cancho & Bolfarine (2001), Pons & Lemdani (2003) e Perperoglou *et al.* (2007) consideram os modelos de proporção de curados em seus estudos.

Posteriormente temos também Ortega *et al.* (2009) que propõem uma distribuição Gama Generalizada modificada, considerando a presença de curados, Rodrigues *et al.* (2009a) trazem uma unificação dos modelos de sobrevivência de longa duração, Rodrigues *et al.* (2009d) utilizam a distribuição COM-Poisson com fração de cura em dados de melanoma, Perdoná & Louzada-Neto (2011) apresentam uma generalização do modelo de mistura e Cancho *et al.* (2012) introduzem o modelo Geométrico Birbaun Saunders com fração de cura e Mazucheli *et al.* (2012).

A divisão da população em dois grupos é facilmente aceitável, então parece-nos adequado proceder conforme Maller & Zhou (1996) e considerar dois componentes em um modelo de mistura. Um dos componentes se incumbe das informações referentes ao grupo de indivíduos suscetíveis ao evento, enquanto o outro relata as informações referentes ao grupo de indivíduos imunes ao evento.

Na medicina costuma-se observar o tempo até a ocorrência ou recorrência de uma doença ou até a morte do paciente, observando a proporção de curados ou imunes. Podemos citar, por exemplo, o caso da leucemia, geralmente de origem desconhecida, que em 2012 estima-se que acometerá 8.510 indivíduos só no Brasil (Brasil, 2011). Mas graças às melhorias no tratamento ao longo das décadas, a taxa de cura pode atingir proporções bem elevadas (Kersey *et al.*, 1987). O câncer de ovário, tumor ginecológico mais difícil de ser diagnosticado e o de menor chance de cura, a previsão para 2012 é que o Brasil tenha 6.190 novos casos. De acordo com a *American Cancer Society*, a taxa de sobrevivência de mulheres com câncer de ovário epitelial invasivo, em 5 anos, deverá atingir a marca de 94%, dependendo do estágio da doença[1].

[1]http://www.cancer.org/cancer/ovariancancer/overviewguide/ovarian-cancer-overview-survival-rates

4.2 Formulação do modelo

Considerando uma população em que existe possibilidade de cura para o evento de interesse e $W = w$ a v.a. que denota o tempo até a ocorrência do evento de interesse, a função de sobrevivência imprópria para $W = w$ é

$$S(w) = pS_{CI}(w) + (1-p)S_{CS}(w),$$

em que $S_{CI}(w)$ e $S_{CS}(w)$ são as funções de sobrevivência dos indivíduos nos conjuntos dos imunes (CI) e conjunto dos suscetíveis (CS), respectivamente (Maller & Zhou, 1996). Esses autores também ressaltam que o indivíduo do conjunto CI pode não apresentar o evento de interesse, tornando seu tempo de falha infinito, assim $S_{CI}(w) = P(W > w|CI) = 1$, $\forall w \geq 0$. Portanto, $S(w)$ pode ser escrita como

$$S(w) = p + (1-p)S_{CS}(w), \tag{4.1}$$

sendo que todo indivíduo que pertence ao conjunto CS apresenta o evento de interesse, assim $\lim_{w\to\infty} S_{CS}(w) = 0$ e, consequentemente, $\lim_{w\to\infty} S(w) = p$. Portanto, a função de sobrevivência (não condicionada) é imprópria e seu limite corresponde à proporção de indivíduos que pertencem ao conjunto dos imunes.

O modelo (4.1) corresponde ao modelo de mistura padrão proposto por Berkson & Gage (1952) e geralmente chamado simplesmente de modelo Berkson e Gage (modelo BG). Além disso é um dos modelos de longa duração mais conhecidos na análise de sobrevivência.

Considerando $S_{CS}(w)$ como em (3.3), temos que a distribuição Exponencial Geométrica de longa duração (LEG), para CR_{mn} e CR_{mx}, têm funções de sobrevivência dadas por

$$S_{\text{LEG}}(W = w) = \begin{cases} \dfrac{p + (\theta - p)e^{-\lambda w}}{1 - (1-\theta)e^{-\lambda w}}, & \text{para } W = \min(T_1, T_2, \cdots, T_M); \\ \dfrac{\theta p + (1-\theta p)e^{-\lambda w}}{\theta + (1-\theta)e^{-\lambda w}}, & \text{para } W = \max(T_1, T_2, \cdots, T_M), \end{cases} \tag{4.2}$$

cuja forma dessas funções, para alguns valores de θ com $p = 0,25$, são apresentadas na Figura 4.1. (Para $p = 0$ temos as distribuições apresentadas no Capítulo 3).

Nesta metodologia para a ativação aleatória, a função de sobrevivência é $S(t) = p + (1-p)e^{-\lambda t}$ que é o próprio modelo de Berkson e Gage da distribuição Exponencial,

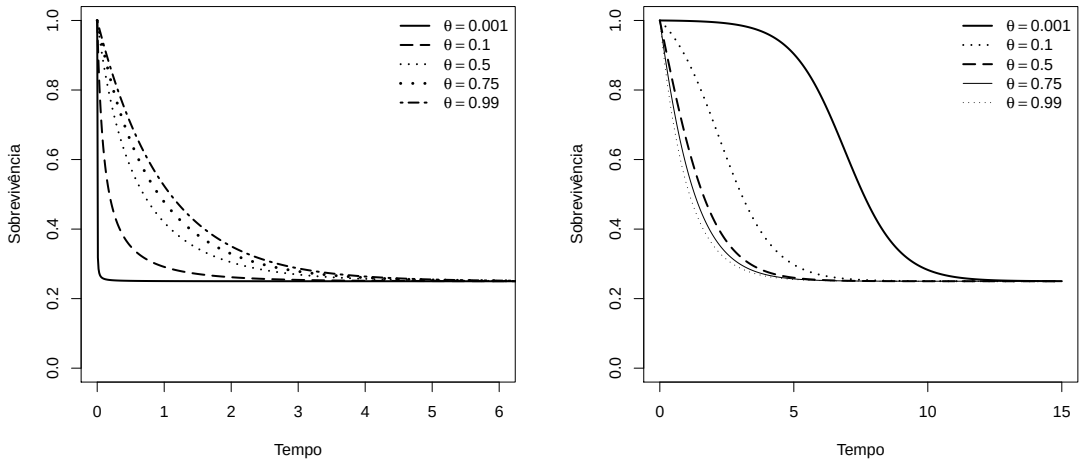

FIGURA 4.1: Função de sobrevivência para a distribuição LEG, considerando a primeira ativação (superior) e a última ativação (inferior), para $\lambda = 1$ e $p = 0,25$.

o qual não será trabalhado neste capítulo e será denominado simplesmente por LE, cuja f.d.p. é $f(w) = (1-p)\lambda e^{-\lambda w}$.

Considerando $S_{CS}(w)$ como em (3.2), temos que a distribuição Exponencial Geométrica de longa duração (LEG) tem f.d.p. dada por $(1-p)f_{CS}(w)$. Para CR_{mn} e CR_{mx} as f.d.p são dada por

$$f_{\text{LEG}}(W = w) = \begin{cases} \dfrac{(1-p)\theta\lambda e^{-\lambda w}}{(1-(1-\theta)e^{-\lambda w})^2}, & \text{para } W = \min(T_1, T_2, \cdots, T_M); \\ \dfrac{(1-p)\theta\lambda e^{-\lambda w}}{(\theta+(1-\theta)e^{-\lambda w})^2}, & \text{para } W = \max(T_1, T_2, \cdots, T_M). \end{cases} \tag{4.3}$$

As distribuições apresentadas em (4.3) são representadas por LEG_{mn} para a primeira ativação e LEG_{mx} para a última ativação. Sendo discutidas nas Seções 4.3 e 4.4, respectivamente. Na Figura 4.2 estão os gráficos da f.d.p. das distribuições LEG_{mn} e LEG_{mx} mediante variação do parâmetro θ. Para $p = 0$ os gráficos correspondem aos apresentados na Figura 3.1.

A função que descreve o risco do evento ocorrer no tempo $W = w$ é dada por

$$h_{\text{LEG}}(w) = \begin{cases} \dfrac{(1-p)\lambda\theta e^{-\lambda w}}{[1-(1-\theta)e^{-\lambda w}][p+(\theta-p)e^{-\lambda w}]}, & \text{para o tempo mínimo;} \\ \dfrac{(1-p)\lambda\theta e^{-\lambda w}}{[\theta+(1-\theta)e^{-\lambda w}][\theta p+(1-\theta p)e^{-\lambda w}]}, & \text{para o tempo máximo.} \end{cases} \tag{4.4}$$

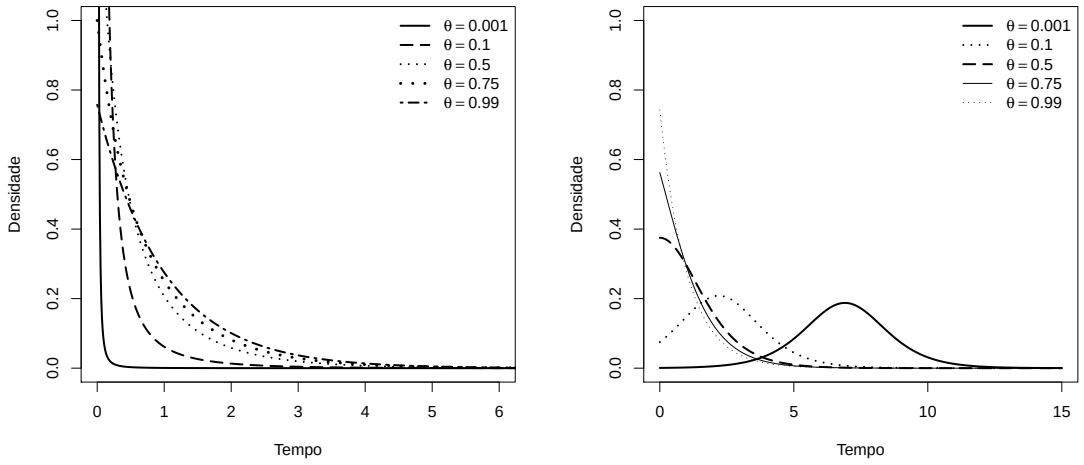

FIGURA 4.2: Função densidade de probabilidade para a LEG. Painel esquerdo considerando CR_{mn} e no painel direito a CR_{mx}, com $\lambda = 1$ e $p = 0,25$.

Os valores iniciais das função de risco são finitos para os dois contextos apresentados. Para CR_{mn} o valor inicial é $(1-p)\dfrac{\lambda}{\theta}$ e para CR_{mx} é $(1-p)\lambda\theta$.

Enquanto a função de risco da distribuição LEG_{mn} é sempre decrescente (conforme resultados apresentado por Glaser (1980)), a função de risco da distribuição LEG_{mx} apresenta as formas decrescente, crescente e unimodal. Nas Figuras 4.3 e 4.4 dispõem-se as representações gráficas da função $h_{\mathrm{LEG}}(w)$ com $\lambda = 1$.

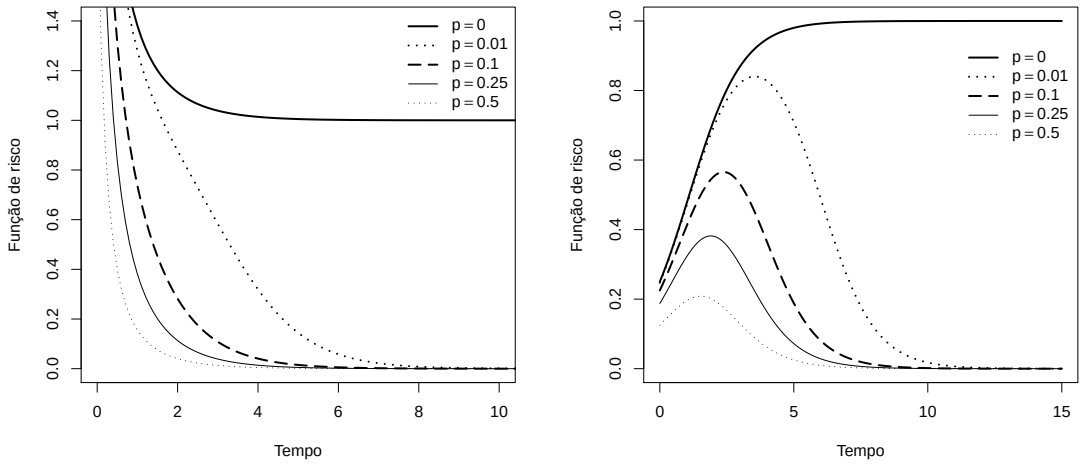

FIGURA 4.3: Função de risco das distribuições LEG_{mn} (esquerda) e LEG_{mx} (direita), para $\theta = 0,25$.

O parâmetro p que informa sobre a proporção de curados, também denominado de parâmetro de longa duração produz mudanças de forma na função de risco, sendo elas mais evidentes no modelo de máximo (Figura 4.4).

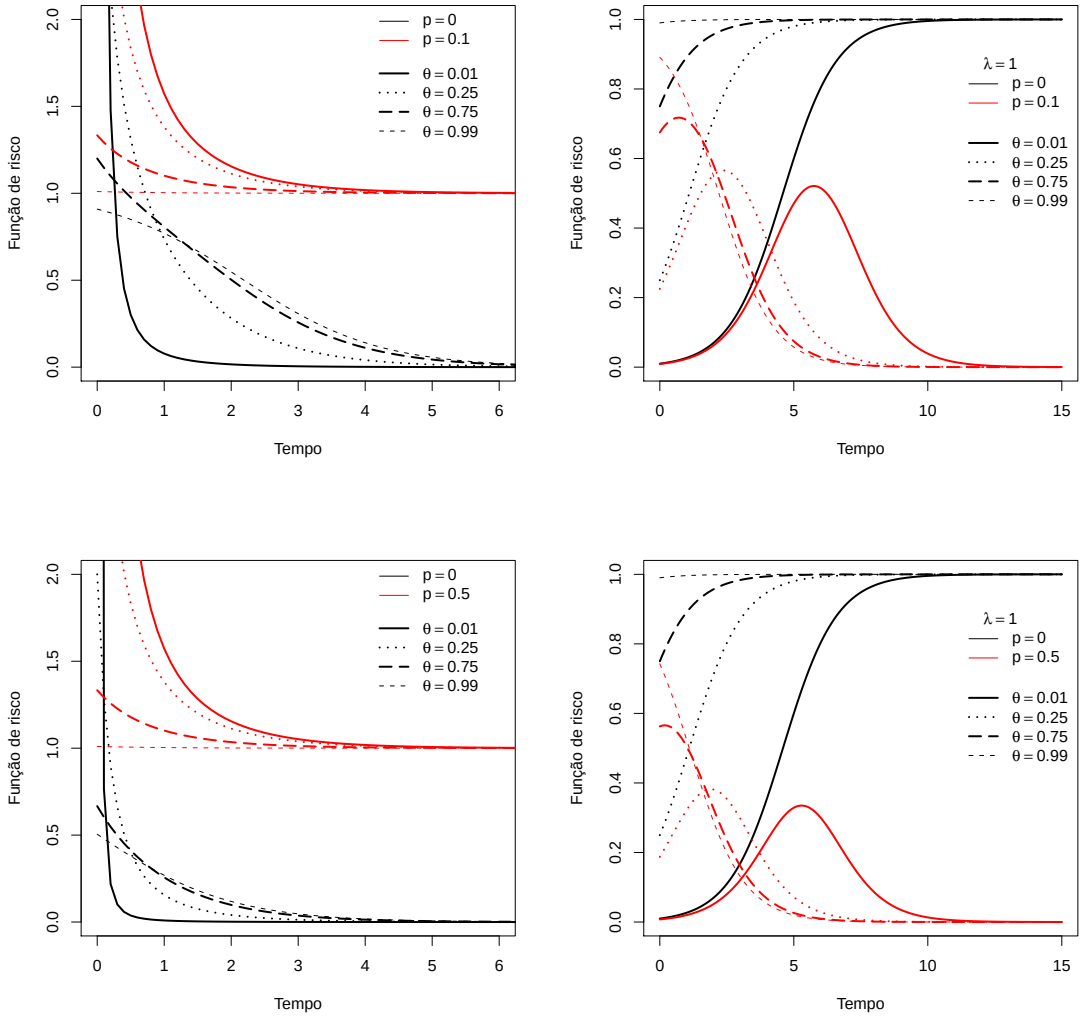

FIGURA 4.4: Função de risco da distribuição LEG_{mn} (esquerda) e LEG_{mx} (direita), com $\lambda = 1$ e dois valores de p.

4.3 Distribuição Exponencial Geométrica de longa duração para o mínimo dos tempos

Nesta seção abordamos propriedades e aplicações da distribuição LEG_{mn}. A v.a. que denota o tempo até a ocorrência do evento de interesse é $X = \min(T_1, T_2, \cdots, T_M)$.

4.3.1 Propriedades

Note que $f_{LEG_{mn}}(x)$ é uma f.d.p. imprópria, isto é, a função distribuição da v.a. X, $F_X(x)$, é deficiente no sentido de que $F_X(+\infty) = lim_{x\to+\infty} < 1$ e, consequentemente, $P(X = +\infty) = 1 - F_X(+\infty) > 0$ (Feller, 1965), o que implica diretamente que o r-*ésimo* momento ordinário $\mu_r^{'} = E(X^r) = +\infty$. No entanto, temos a função quantil, que é dada por

$$Q(u) = \frac{-1}{\lambda} \log\left(\frac{u-(1-p)}{u(1-\theta)-(1-p)}\right), \tag{4.5}$$

em que u tem distribuição uniforme $U(0, 1-p)$, em que $1-p$ é o limite superior da f.d.a. que é imprópria. Temos que $Q(u)$ nos permite gerar amostras de variáveis aleatórias com distribuição LEG_{mn}, cuja mediana teórica é dada por

$$\widetilde{W} = Q\left(\frac{1-p}{2}\right) = \frac{1}{\lambda}\log(1+\theta). \tag{4.6}$$

Em testes de controle de qualidade e confiabilidade, ocorrem situações em que o profissional precisa prever a falha futura de itens com base nos tempos de falhas ocorridos anteriormente. Os indicadores utilizados nessas situações são muitas vezes baseados em momentos de estatística de ordem.

Proposição 4.1 *Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias independentes e identicamente distribuídas, tais que que $X_k \sim LEG_{mn}(p, \lambda, \theta)$ para $k = 1, 2, \ldots, n$. A f.d.p. da k-ésima estatística de ordem, $X_{k:n}$, é*

$$f_{k:n}(x) = g_{k;n}(x)(1-p)^k \left(\frac{p+(\theta-p)e^{-\lambda x}}{\theta e^{-\lambda x}}\right)^{n-k}$$

em que $x > 0$ e $g_{k;n}(x)$ é a f.d.p. da estatística de ordem da distribuição $EG_{mn}(\theta, \lambda)$ dada em dada em (A.3).

Prova 4.1 Derivamos uma expressão explícita para a função densidade da k-*ésima* estatística de ordem, $X_{k:n}$, digamos $f_{k:n}(x)$, em uma amostra aleatória de tamanho n da distribuição LEG_{mn}, partindo da definição e fazendo as devidas substituições, temos

$$f_{k:n}(x) = \frac{n!}{(n-k)!(k-1)!} f_{LEG}(x)[F_{LEG}(x)]^{k-1}[S_{LEG}(x)]^{n-k}$$

substituindo temos

$$\begin{aligned} f_{k:n}(x) =& \frac{n!}{(n-k)!(k-1)!}(1-p)f_{EG}(x)[(1-p)F_{EG}(x)]^{k-1}[p+(1-p)S_{EG}(x)]^{n-k} \\ =& \frac{n!}{(n-k)!(k-1)!}f_{EG}(x)[F_{EG}(x)]^{k-1}[S_{EG}]^{n-k}(1-p)^k\left[\frac{p+(1-p)S_{EG}(x)}{S_{EG}(x)}\right]^{n-k} \\ =& g_{k:n}(x)(1-p)^k\left[\frac{p+(1-p)S_{EG}(x)}{S_{EG}(x)}\right]^{n-k}, \end{aligned} \tag{4.7}$$

em que $g_{k:n}(x)$ e S_{EG} são f.d.p. da estatística de ordem e função de sobrevivência da distribuição EG_{mn}, assim

$$f_{k:n}(x) = g_{k;n}(x)(1-p)^k\left[\frac{p+(\theta-p)e^{-\lambda x}}{\theta e^{-\lambda x}}\right]^{n-k}. \qquad \blacksquare$$

4.3.2 Inferência

Considerando as funções densidade e sobrevivência para a primeira ativação, dadas respectivamente em (4.3) e (4.2), a função log-verossimilhança é dada por

$$\begin{aligned} \ell(\theta,\lambda,p) =& \log[(1-p)\lambda\theta]\sum_{i=1}^{n}\delta_i - \lambda\sum_{i=1}^{n}\delta_i x_i - \sum_{i=1}^{n}(\delta_i+1)\log\left[1-(1-\theta)e^{-\lambda x_i}\right] \\ &+\sum_{i=1}^{n}(1-\delta_i)\log\left[p+(\theta-p)e^{-\lambda x_i}\right]. \end{aligned} \tag{4.8}$$

em que δ_i é o indicador de censura e x_i é o tempo observado.

A Figura A.1 (Apêndice) dispõe o gráfico de curva de níveis da função $\ell(\theta,\lambda,p)$, com p fixo, indicando que ela apresenta um único ponto de máximo.

O vetor escore é dado por

$$U(\theta,\lambda,p) = U(\boldsymbol{\vartheta}) = \left(\frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\theta}, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\lambda}, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial p}\right),$$

em que

$$\begin{aligned} \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\theta} &= \frac{\sum_{i=1}^{n}\delta_i}{\theta} + \sum_{i=1}^{n}(\delta_i+1)\frac{e^{-\lambda x_i}}{1-(1-\theta)e^{-\lambda x_i}} + \sum_{i=1}^{n}(1-\delta_i)\frac{e^{-\lambda x_i}}{p+(\theta-p)e^{-\lambda x_i}}, \\ \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\lambda} &= \frac{\sum_{i=1}^{n}\delta_i}{\lambda} - \sum_{i=1}^{n}\delta_i x_i - \sum_{i=1}^{n}\frac{(\delta_i+1)(1-\theta)x_i e^{-\lambda x_i}}{1-(1-\theta)e^{-\lambda x_i}} - \sum_{i=1}^{n}\frac{(1-\delta_i)(\theta-p)x_i e^{-\lambda x_i}}{p+(\theta-p)e^{-\lambda x_i}}, \\ \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial p} &= \frac{-\sum_{i=1}^{n}\delta_i}{1-p} + \sum_{i=1}^{n}(1-\delta_i)\frac{1-e^{-\lambda x_i}}{p+(\theta-p)e^{-\lambda x_i}}. \end{aligned}$$

O sistema de equações $U(\boldsymbol{\vartheta}) = \mathbf{0}$ não tem solução explícita, desta forma utilizamos o método de otimização proposto por Nelder & Mead (1965) vinculado à rotina

optim do R (R Core Team, 2011) para obter $\hat{\boldsymbol{\vartheta}}$ estimativas de máxima verossimilhança de $\boldsymbol{\vartheta}$. Os critérios AIC, BIC, $-\ell(\boldsymbol{\vartheta})$ foram utilizados para comparar os ajustes.

4.3.3 Aplicação a dados de câncer

Nossa intenção é mostrar a aplicabilidade da distribuição LEG_{mn} e sua competitividade em termos de ajustes com distribuições difundidas na análise de sobrevivência, como é o caso das distribuições Exponencial e Weibull. Comparamos o ajuste da distribuição LEG_{mn} com a distribuição Exponencial de longa duração (LE) e com a distribuição Weibull de longa duração (LW). Para tanto consideramos os conjuntos de dados de câncer $T7$ (mielomatose) e $T8$ (leucemia), descritos na Seção 1.3.

A Tabela 4.1 apresenta as EMVs e desvio padrão para os parâmetros das três distribuições consideradas. Observamos que as estimativas da proporção de curados ficaram bem próximas entre si e também do valor em que o gráfico KM estabilizou, como mostra a Figura 4.5, que traz a estimativa de KM para a função de sobrevivência e as estimativas da função de sobrevivência das distribuições ajustadas, utilizando as EMVs.

TABELA 4.1: EMVs e desvio padrão (nos parênteses) para os dados de *Mielomatose* ($T7$) e *Leucemia* $T8$.

Dados	Distribuição	λ	θ	ϕ	p
$T7$	LEG_{mn}	0,0002 (0,0084)	0,9857 (0,0659)	-	0,2658 (0,1229)
	LW	0,0049 (0,0022)	-	0,6749 (0,1431)	0,2901 (0,1019)
	LE	0,0042 (0,0011)	-	-	0,3035 (0,0966)
$T8$	LEG_{mn}	0,9980 (0,6723)	0,5567 (0,4564)	-	0,2636 (0,0712)
	LW	1,4517 (0,3078)	-	0,9452 (0,1377)	0,2688 (0,0690)
	LE	1,4331 (0,2795)	-	-	0,2710 (0,0684)

Na Tabela 4.2 encontram-se os valores dos critérios de comparação de modelos, os quais fornecem evidências em favor da distribuição LEG_{mn} para os dois conjuntos de dados. Os critérios de comparação de modelos fornecem evidências em favor da distribuição LEG_{mn} para os dois conjuntos de dados, Tabela 4.2. Comparamos o ajuste da distribuição LEG_{mn} com a LE consideramos o procedimento de teste de hipótese para modelos encaixados com estatística de teste $\omega_n = 2\{\ell_{LEG_{mn}} - \ell_{LE}\}$. O valor obtido para a

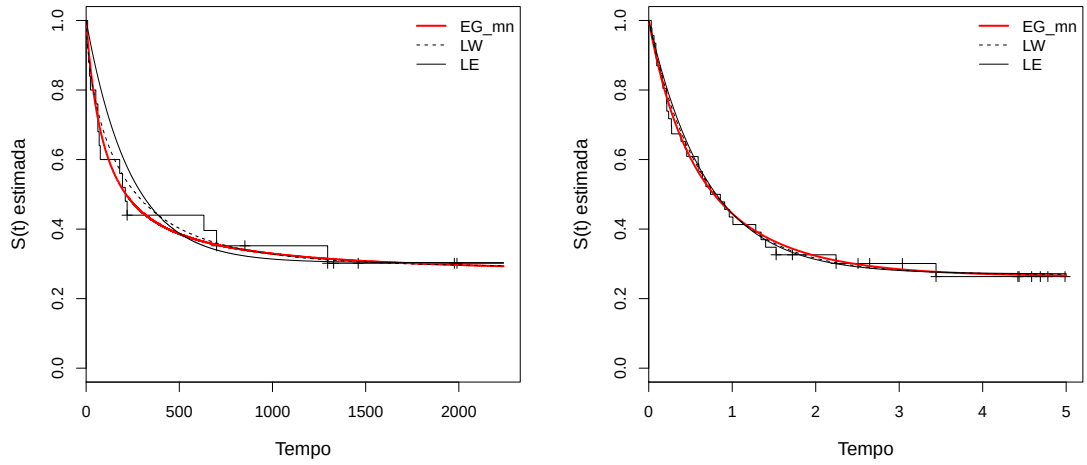

FIGURA 4.5: Curva de Kaplan-Meier com função de sobrevivência estimada considerando os ajustes das distribuições LEG_{mn}, LW e LE para os conjuntos de dados $T7$ e $T8$, respectivamente.

estatística é $\omega_n = 6,526$, maior que o valor de comparação $1/2 + 1/2\ P(\chi^2_1 \leq c) = 2,421$, evidenciando a distribuição LEG_{mn} como a melhor opção de ajuste para $T7$. Entretanto, não existe evidências que a distribuição LEG_{mn} ajusta-se melhor que a distribuição LE para $T8$ a 5% de significância.

TABELA 4.2: Critérios de comparação de ajustes - $-\ell(\widehat{\boldsymbol{\vartheta}}_g)$, AIC e BIC - para as distribuições Exponencial Geométrica com longa duração, Weibull com longa duração e Exponencial com longa duração.

Distribuição	$T7$			$T8$		
	$-\ell(\boldsymbol{\vartheta})$	AIC	BIC	$-\ell(\boldsymbol{\vartheta})$	AIC	BIC
LEG_{mn}	121,0429	248,0853	251,7424	45,9017	97,8035	103,2895
LW	121,9174	249,8348	253,4915	46,1484	98,2969	103,7828
LE	124,3070	252,6141	255,0518	46,2279	96,4559	100,1132

4.3.4 Comentários

A distribuição LEG_{mn} é uma extensão da distribuição EG proposta por Adamidis & Loukas (1998) e está fundamentada no cenário de riscos latentes competitivos com uma

população mista que considera grupos de indivíduos suscetíveis e imunes ao evento de interesse.

Essa distribuição apresenta funções de sobrevivência, risco e densidade com formas fechadas, sendo de fácil estimação quando utilizados procedimentos computacionais de otimização. Sua importância prática foi demonstrada em duas aplicações em que a distribuição LEG_{mn}, única fundamentada em risco competitivo, mostrou competitividade com as demais distribuições já tradicionais na literatura. Embora a distribuição LEG_{mn} tenha sido aqui aplicada apenas a dados de câncer, ela é flexível o suficiente para ser considerada para dados de sobrevivência de áreas como engenharia, confiabilidade, demografia, finanças, além de outras.

4.4 Distribuição Exponencial Geométrica de longa duração para o máximo dos tempos

Nesta seção detalhamos a distribuição relativa ao tempo máximo LEG_{mx} apresentando propriedades e aplicações. A variável aleatória que denota a ocorrência do evento de interesse é $Y = \max(T_1, T_2, \cdots, T_M)$.

4.4.1 Propriedades

Como no caso anterior, temos que $f_{LEG_{mx}}(y)$ é imprópria. Assim, segundo Feller (1965), a função distribuição da variável aleatória Y é deficiente no sentido de que $F_Y(+\infty) = lim_{y\to+\infty} < 1$ e, consequentemente, $P(Y = +\infty) = 1 - F_Y(+\infty) > 0$, implicando diretamente que o r-*ésimo* momento ordinário $\mu_r^{'} = E(Y^r) = +\infty$. No entanto a função quantil é possível de ser obtida, sendo que por ela podemos obter a mediana da distribuição. A função quantil é dada por

$$Q(u) = \frac{1}{\lambda}\log\left(\frac{\theta(1-p-u)+u}{\theta(1-p-u)}\right), \tag{4.9}$$

em que $u \sim U(0, 1-p)$, cujo limite superior da distribuição uniforme corresponde ao limite superior da f.d.a. e a mediana da distribuição é

$$\widetilde{Y} = Q\left(\frac{1-p}{2}\right) = \frac{1}{\lambda}\log\left(\frac{\theta+1}{\theta}\right). \tag{4.10}$$

Outra propriedade da distribuição LEG_{mx} é a f.d.p. da estatística de ordem, como apresentada a seguir.

Proposição 4.2 *Sejam $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes e identicamente distribuídas, tais que $Y_k \sim LEG_{mx}(\theta, \lambda, p)$, $k = 1, 2, \ldots, n$. A função densidade de probabilidade da k-ésima estatística de ordem é*

$$f_{k:n}(y) = g_{k;n}(y)(1-p)^k \left(\frac{\theta p + (1-\theta p)e^{-\lambda y}}{e^{-\lambda y}}\right)^{n-k},$$

em que $y > 0$ e $g_{k;n}$ é a função densidade de probabilidade da estatística de ordem da distribuição $EG_{mx}(\theta, \lambda)$.

Prova 4.2 Pela definição de estatística de ordem tal qual apresentada na demonstração (4.7) temos que

$$f_{k:n}(y) = g_{k;n}(y)(1-p)^k \left[\frac{\theta p + (1-\theta p)e^{-\lambda y}}{e^{-\lambda y}}\right]^{n-k},$$

em que $g_{k:n}(y)$ é dada em (A.4). ■

4.4.2 Inferência

Considerando as funções densidade e sobrevivência do evento produzido pela ativação da última CR, dadas em (4.3) e (4.2), respectivamente, a função log-verossimilhança para $Y = \max(t_1, t_2, \cdots, t_m)$ com longa duração e δ_i indicador de censura, é dada por

$$\begin{aligned}\ell(\theta, \lambda, p) =& \log[\lambda\theta(1-p)]\sum_{i=1}^{n}\delta_i - \lambda\sum_{i=1}^{n}\delta_i y_i - \sum_{i=1}^{n}(1+\delta_i)\log\left[\theta + e^{-\lambda y_i}(1-\theta)\right] \\ &+ \sum_{i=1}^{n}(1-\delta_i)\log\left[p\theta + e^{-\lambda y_i}(1-p\theta)\right].\end{aligned} \tag{4.11}$$

A Figura A.1 (Apêndice) dispõe o gráfico de curva de níveis da função $\ell(\theta, \lambda, p = 0,30)$ e sugere que a mesma apresenta um único ponto de máximo.

O vetor escore é dado por

$$U(\theta, \lambda, p) = U(\boldsymbol{\vartheta}) = \left(\frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\theta}, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial\lambda}, \frac{\partial\ell(\boldsymbol{\vartheta})}{\partial p}\right),$$

em que

$$\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \theta} = \frac{1}{\theta}\sum_{i=1}^{n}\delta_i - \sum_{i=1}^{n}(\delta_i+1)\frac{1-e^{-\lambda y_i}}{\theta+(1-\theta)e^{-\lambda y_i}} + \sum_{i=1}^{n}\frac{(1-\delta_i)(p-pe^{-\lambda y_i})}{p\theta+(1-p\theta)e^{-\lambda y_i}},$$

$$\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial \lambda} = \frac{1}{\lambda}\sum_{i=1}^{n}\delta_i - \sum_{i=1}^{n}\delta_i y_i + \sum_{i=1}^{n}\frac{(\delta_i+1)y_i e^{-\lambda y_i}}{\theta(1-\theta)^{-1}+e^{-\lambda y_i}} - \sum_{i=1}^{n}\frac{(1-\delta_i)y_i e^{-\lambda y_i}}{p\theta(1-p\theta)^{-1}+e^{-\lambda y_i}},$$

$$\frac{\partial \ell(\boldsymbol{\vartheta})}{\partial p} = \frac{-1}{1-p}\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}(1-\delta_i)\frac{\theta-\theta e^{-\lambda y_i}}{p\theta+e^{-\lambda y_i}(1-p\theta)}.$$

Como no caso anterior, o sistema de equações $U(\boldsymbol{\vartheta}) = \mathbf{0}$ não tem solução explícita e utilizamos o método de otimização da rotina *optim* do `R` (R Core Team, 2011) utilizando o método quase-Newton para obter as estimativas de máxima verossimilhança de $\boldsymbol{\vartheta}$. Na comparação dos ajustes utilizamos os critérios AIC, BIC, $-\ell(\boldsymbol{\vartheta})$.

4.4.3 Estudo de simulação

Para avaliar a performance dos EMVs dos parâmetros da distribuição LEG_{mx}, geramos tempos de vidas desta distribuição para $\theta = 0,1$; $\lambda = 0,1$ e fração de curados de 50%, considerando a função quantil (4.9).

Consideramos os tamanhos amostrais $n = 20, 40, \ldots, 800$. Para cada tamanho amostral foram obtidas 1.000 simulações e calculou-se a média do vício dos EMVs de λ, θ e p, como também o erro quadrático médio (EQM) dos EMVs. A Figura A.5 (no Apêndice) mostra que tanto o vício quanto o EQM variam com respeito à n (linha tracejada na corresponde ao vício tendendo a zero). Observamos, como era esperado, que a média do vício e o EQM estão próximas de zero e os EQMs decrescem quando o tamanho amostral aumenta. Resultados similares são observados para outros valores do vetor paramétrico (θ, λ, p).

4.4.4 Aplicação

Comparamos a distribuição $LEG_{mx}(\theta, \lambda, p)$ com a distribuição LE, bem como as distribuições de longa duração Weibull (LW) e Log-Logística (LLL), para três conjuntos de dados, em que o terceiro é da área de finanças enquanto os dois primeiros são da área médica, com o objetivo de expor a aplicabilidade da distribuição LEG_{mx} e sua competitividade em termos de qualidade de ajuste com distribuições usuais na análise

de sobrevivência e apontar a flexibilidade da distribuição LEG_{mx} em acomodar dados de diversas áreas.

Foram utilizados os conjuntos $T9$, $T10$ e $T11$ descritos na Seção 1.3, observando a forma da função de risco dos mesmo (gráfico TTT) e a indicação de longa duração dos mesmos. Esta última é empiricamente proposta pelo percentual de censura dos dados - 50%, 10% e 65%, respectivamente - e posteriormente pela forma do gráfico de KM.

A Tabela 4.3 fornece as EMVs e desvio padrão (entre parênteses) para os parâmetros das distribuições - LEG_{mx}, LE, LW, LLL -, que se compararmos os ajustes pela proporção de imunes, p, a distribuição LE apresenta as menores proporções para $T9$ e $T11$ e, neste último, em torno de 80% a menos que os demais. No conjunto $T10$, a distribuição LLL praticamente não identificou o grupo dos imunes, enquanto a distribuição LEG_{mx} e o LW indicam que cerca de 8% da população pertence a esse grupo.

TABELA 4.3: EMVs e desvio padrão (parênteses) para as distribuições LEG_{mx}, LE, LW e LLL.

Dado	Dist.	λ	θ	ϕ	p
$T9$	LEG_{mx}	3,0213 (1,0499)	0,04533 (0,0571)		0,4882 (0,1080)
	LE	0,5286 (0,4031)			0,3271 (0,2760)
	LW	0,8672 (0,1384)		2,1078 (0,5752)	0,4928 (0,1074)
	LLL	1,0905 (0,3182)		2,1942 (0,7442)	0,4178 (0,1501)
$T10$	LEG_{mx}	0,0052 (0,0005)	0,1915 (0,0331)		0,0852 (0,0214)
	LE	0,0021 (0,0001)			0,0232 (0,0272)
	LW	0,0025 (0,0001)		1,3893 (0,0716)	0,0815 (0,0218)
	LLL	328,4439 (18,1483)		1,6634 (0,0835)	0,0001 (0,0034)
$T11$	LEG_{mx}	0,5106 (0,0405)	0,01238 (0,0046)		0,6593 (0,0217)
	LE	0,0273 (0,0142)			0,1221 (0,3723)
	LW	0,1077 (0,0031)		2,9685 (0,2162)	0,6640 (0,0213)
	LLL	8,9401 (0,4782)		3,0089 (0,2693)	0,6158 (0,0278)

A qualidade de ajuste das distribuições aos dados é dada por $-\ell(\widehat{\boldsymbol{\vartheta}})$ (Tabela 4.4) e, para os três conjuntos de dados, a distribuição LEG_{mx} supera as demais distribuições nos três critérios. Esta predileção pela distribuição LEG_{mx} pode ser visualizada no gráfico Kaplan-Meier (Figura 4.6), com as funções de sobrevivência estimadas para distribuições

ajustadas.

TABELA 4.4: Critérios de comparação de ajustes - $-\ell(\widehat{\boldsymbol{\vartheta}}_g)$, AIC e BIC - para as distribuições Exponencial Geométrica com longa duração, Exponencial com longa duração, Weibull com longa duração e Log-Logística com longa duração.

Dado	Distribuição	$-\ell(\boldsymbol{\vartheta})$	AIC	BIC
$T9$	LEG_{mx}	24,526	55,050	58,827
	LE	26,988	57,979	60,493
	LW	24,675	55,350	59,124
	LLL	25,524	57,049	60,824
$T10$	LEG_{mx}	1952,567	3911,135	3923,191
	LE	1968,706	3941,413	3949,450
	LW	1952,869	3911,738	3923,794
	LLL	1959,587	3925,174	3937,229
$T11$	LEG_{mx}	740,799	1487,599	1500,285
	LE	794,907	1593,814	1602,271
	LW	742,668	1491,336	1504,022
	LLL	758,874	1523,750	1536,435

A curva da sobrevivência estimada para $T10$ (gráfico KM) não é completamente visível em decorrência da grande quantidade de dados que os conjunto apresenta.

4.4.5 Comentários

A distribuição LEG_{mx} é uma extensão da distribuição EG_{mx} apresentada no Capítulo 3 (denominada de CEG em Louzada *et al.* (2011)) e está fundamentada no cenário de riscos latentes complementares com uma população mista que considera grupos de indivíduos suscetíveis e imunes ao evento de interesse.

Essa distribuição apresenta funções de sobrevivência, risco e densidade com formas fechadas, tendo seus EMVs obtidos facilmente com a utilização de procedimentos computacionais de otimização. Sua importância prática foi demonstrada em três aplicações em que a distribuição LEG_{mn}, única fundamentada em riscos complementares

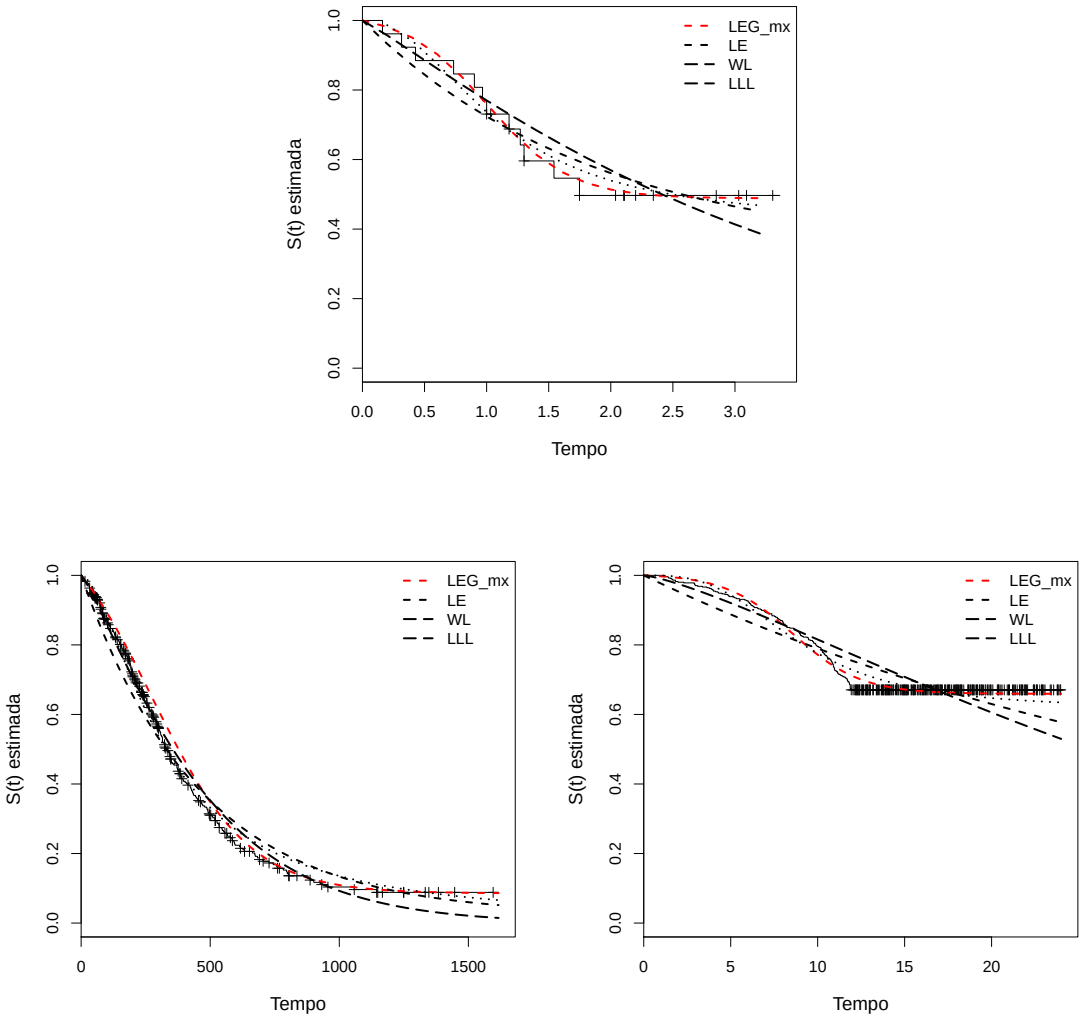

FIGURA 4.6: Curva de Kaplan-Meier com as funções de sobrevivência estimada a partir das EMVs dos parâmetros das distribuições LEG_{mx}, LE, LW e LLL para os dados $T9$, $T10$ e $T11$, respectivamente.

entre as utilizadas, mostrou competitividade com as demais distribuições já tradicionais na literatura.

Do ponto de vista prático, as hipóteses subjacentes à LEG_{mx} parecem ser adequadas para todos os conjuntos de dados considerados. Como observado no KM, para os três exemplos, é evidente que parte das unidades não pode ser afetada pelo evento de interesse (Figuras 1.9, 1.10 e 1.11). Além disso, a suposição de riscos complementares parece ser apropriada, uma vez que, para os dados de câncer de ovário ($T9$) e glioma ($T10$), é necessário um número máximo de células que compõem as metástases tumorais

para desencadear a morte do paciente.

Para os dados $T11$ pode-se supor que um cliente pode ser considerado "ruim", quando tornou-se inadimplente após ter utilizado o último recurso que dispunha para quitar sua dívida. Por outro lado, a suposição de riscos complementares pode ser questionada nos dados de escore de crédito, uma vez que é possível considerar que um mesmo conjunto de causas de não pagamento pode não afetar a cada um dos clientes e, assim, observa-se apenas a causa do tempo de vida mínimo. Embora estes pressupostos dependam do tipo de carteira em análise, vale a pena considerar os riscos complementares em uma perspectiva de modelagem de crédito. Finalmente, é importante ressaltar que toda a análise apresentada é apenas o passo preliminar para desenvolvimento de um modelo, uma vez que o objetivo principal da análise de crédito é desenvolver um modelo preditivo que dê uma probabilidade de não pagamento baseado em um conjunto de covariáveis.

4.5 Conclusões

Os modelos de longa duração LEG_{mn} e LEG_{mx} apresentam boas propriedades de estimação de máxima verossimilhança, são competitivos se comparados aos modelos de mistura de longa duração BG considerando as distribuições Weibull, Exponencial e o Log-Logístico. Os conjuntos considerados neste capítulo necessitam de modelos de longa duração, os modelos Weibull, Exponencial, Log-Logístico, EG_{mn} e EG_{mx} sem longa duração produziram ajustes muito aquém do aceitável e foram omitidos.

Capítulo 5

Distribuições da família Gama Generalizada Geométrica

Neste capítulo apresentamos as distribuições Gama Generalizada Geométrica *GGGcr* e Gama Generalizada Geométrica com longa duração *LGGGcr*, nos três esquemas de ativação conforme conceitos apresentados no Capítulo 2. Algumas comparações entre os esquemas de ativação são abordadas, bem como propriedades e particularidades das distribuições. Para finalizar fazemos uma aplicação a um conjunto de dados reais.

5.1 Introdução

As distribuições compostas Gama Generalizada Geométrica com causa de risco latente denotada por (GGGcr) e Gama Generalizada Geométrica com longa duração e causa de risco latente (LGGGcr) são concebidas no cenário latente de múltiplos riscos, com e sem fração de cura, respectivamente, em que as CR possuem distribuição de probabilidade Geométrica e os tempos de ocorrência de cada uma das CR são distribuídos conforme a distribuição Gama Generalizada de três parâmetros (GG) (Stacy, 1962; Farewell & Prentice, 1977), cuja função densidade de probabilidade é descrita por

$$f_{GG}(t) = \begin{cases} \frac{c(\lambda)}{\sigma t} \exp\left\{\frac{1}{\lambda}\left[\frac{\log(t)-\mu}{\sigma}\right] - \frac{1}{\lambda^2}\exp\left\{\lambda\left[\frac{\log(t)-\mu}{\sigma}\right]\right\}\right\}, & \text{se} \quad \lambda \neq 0 \\ \frac{1}{t\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left[\frac{\log(t)-\mu}{\sigma}\right]^2\right\}, & \text{se} \quad \lambda = 0, \end{cases} \tag{5.1}$$

em que $t > 0$, $\mu \in \mathbb{R}$, $\sigma > 0$ e $\lambda \in \mathbb{R}$ são, respectivamente, parâmetros de locação, escala e forma e $c(\lambda) = \frac{|\lambda|}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}$ com $\Gamma(\cdot)$ a função gama. Lawless (1980), faz $Y = log(t)$ e denomina de distribuição Log-Gama Generalizada.

A f.d.a. da GG para $t > 0$ é dada por

$$F_{GG}(t) = \begin{cases} \Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(t)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}, & \text{se} \quad \lambda > 0; \\ \Phi\left[\frac{\log(t)-\mu}{\sigma}\right], & \text{se} \quad \lambda = 0, \\ 1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(t)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}, & \text{se} \quad \lambda < 0, \end{cases} \tag{5.2}$$

em que $\Phi(\cdot)$ denota a f.d.a. da distribuição Normal padrão; $\Gamma_G(t,\gamma) = \int_0^t x^{\gamma-1}e^{-x}/\Gamma(\gamma)dx$ é a f.d.a. para o caso particular da distribuição Gama com média e variância igual a γ, $\gamma > 0$.

A distribuição GG assim definida tem como casos particulares a distribuição Gama de dois parâmetros - quando $\lambda = \sigma$; a distribuição Weibull - se $\lambda = 1$; a distribuição Exponencial - se $\lambda = \sigma = 1$; a distribuição LogNormal - para $\lambda = 0$ e a distribuição Weibull Invertida - se $\lambda = -1$. Ela tem sido usada em várias áreas de pesquisa, principalmente por sua capacidade de reproduzir as propriedades das distribuições mais tradicionais, como Weibull, LogNormal e Exponencial, tornando-se uma opção vantajosa. Kao (1958) comenta que a *GG*, muitas vezes, é utilizada até mesmo para determinar qual das referidas distribuições deve ser usada para modelar um conjunto de dados.

Dados de sobrevivência podem ser tratados adequadamente pela distribuição GG. Yamaguchi (1992) utiliza essa distribuição na análise de tempos de trabalho fixo no Japão e Allenby *et al.* (1999) apresentam um modelo dinâmico baseado na distribuição GG. Cox *et al.* (2007) apresentam uma análise de sobrevivência paramétrica e de classificação da função de risco da distribuição *GG* e justificam sua importância, evidenciando a extensa família de distribuição que contém as distribuições mais comumente usadas, como Exponencial, Weibull, LogNormal e Gama. Outro aspecto importante apresentado por eles é que esta família de distribuição inclui todas as quatro formas mais comuns que a função de risco assume: unimodal, decrescente, crescente e U.

Além disso, Ortega *et al.* (2009) propõem a distribuição *GG* modificada que permite o tratamento de dados com presença de longa duração. Temos também a distribuição Gama Generalizada Exponencial proposta por Cordeiro *et al.* (2011). No contexto de mistura de distribuições, Ortega *et al.* (2011) introduzem uma distribuição Gama

Generalizada Geométrica e expressam sua função densidade como combinação linear da função densidade da Gama Generalizada. Abd El-Fatah *et al.* (2011) trabalham com uma modificação da distribuição Gama Generalizada e propõem um método de estimação de máxima verossimilhança para os cinco parâmetros da mesma. Pascoa *et al.* (2011) apresentam a distribuição Gama Generalizada Kumaraswamy. Por fim, Tojeiro *et al.* (2012) trabalham com a distribuição Gama Generalizada no contexto de modelos de risco utilizando a inferência bayesiana para a estimação dos parâmetros do modelo.

Gupta & Kundu (1999), ao explanar a respeito da distribuição Gama, argumentam que uma das maiores desvantagens da mesma seria a dificuldade de computar a função distribuição ou sobrevivência para λ (parâmetro de forma) não inteiro, fazendo-se necessária a utilização de tabelas matemáticas ou programas computacionais para a obtenção dos valores. Ressaltam ainda que este fator torna a distribuição Gama pouco popular se comparada com a distribuição Weibull. No entanto, ousamos dizer que, perante os notáveis e constantes avanços computacionais ocorridos nos últimos anos, essa 'desvantagem' já foi vencida.

5.2 Distribuição Gama Generalizada Geométrica

As distribuições que fazem parte da mistura da distribuição Gama Generalizada com a distribuição Geométrica (GGGcr), considerando as 3 possíveis ativações dos fatores de riscos, possuem 4 parâmetros, sendo elas: Gama Generalizada Geométrica para a primeira ativação (GGG_{mn}), Gama Generalizada Geométrica para a ativação aleatória (GGG_{al}) e Gama Generalizada Geométrica para a última ativação (GGG_{mx}).

Nas distribuições temos que $S_{\mathrm{GGG}}(w) = P(W > w|M \geq 1) = S_{GGG}(W = w)$. E sendo $W = X$ o tempo observado é referente à ocorrência do primeiro fator de risco envolvido e se $W = Y$ o tempo observado é referente ao último fator de risco envolvido.

Para o esquema de ativação aleatório a distribuição GGG_{al} é a mesma distribuição da variável aleatória latente T_j's, de modo que $S_{\mathrm{GGG}}(x) = S_{\mathrm{GG}}(x)$, conforme apresentado na Seção 2.3.3, consequentemente as demais funções que descrevem a distribuição correspondem as funções da distribuição GG. Para coesão de notação, adotaremos a função densidade de probabilidade $f_{\mathrm{GGG}}(w) = f_{\mathrm{GG}}(w)$ e função de risco $h_{\mathrm{GGG}}(w) = h_{\mathrm{GG}}(w)$.

5.2.1 Distribuição Gama Generalizada Geométrica para a primeira ativação

A função de sobrevivência para a população de não curados para a primeira ativação, $X = \min\{T_1, T_2, \cdots, T_M\}$, considerando o método exposto na Seção 2.4.1, é dada por

$$S_{\mathrm{GGG}}(x) = \begin{cases} \frac{\theta\left(1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(x)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)}{1-(1-\theta)\left(1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(x)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)}, & \text{se} \quad \lambda > 0; \\ \frac{\theta\left\{1-\Phi\left[\frac{\log(x)-\mu}{\sigma}\right]\right\}}{1-(1-\theta)\left\{1-\Phi\left[\frac{\log(x)-\mu}{\sigma}\right]\right\}}, & \text{se} \quad \lambda = 0, \\ \frac{\theta\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(x)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}}{1-(1-\theta)\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(x)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}}, & \text{se} \quad \lambda < 0, \end{cases} \tag{5.3}$$

que é uma função de sobrevivência própria, pois $S_{\mathrm{GGG}}(0) = 1$ e $S_{\mathrm{GGG}}(\infty) = 0$.

A f.d.p. da GGG_{mn}, conforme apresentado na Seção 2.4.1, é dada por

$$f_{\mathrm{GGG}}(x) = \begin{cases} \frac{\frac{\lambda}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}\theta\exp\left\{\frac{1}{\lambda}\left[\frac{\log(x)-\mu}{\sigma}\right]-\frac{1}{\lambda^2}\exp\left\{\lambda\left[\frac{\log(x)-\mu}{\sigma}\right]\right\}\right\}}{x\sigma\left[1-(1-\theta)\left(1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(x)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)\right]^2}, & \text{se} \quad \lambda > 0; \\ \frac{\theta\exp\left\{-\frac{1}{2}\left[\frac{\log(x)-\mu}{\sigma}\right]^2\right\}}{x\sqrt{2\pi}\sigma\left\{1-(1-\theta)\left[1-\Phi\left(\frac{\log(x)-\mu}{\sigma}\right)\right]\right\}^2}, & \text{se} \quad \lambda = 0, \\ \frac{\frac{\lambda}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}\theta\exp\left\{\frac{1}{\lambda}\left[\frac{\log(x)-\mu}{\sigma}\right]-\frac{1}{\lambda^2}\exp\left\{\lambda\left[\frac{\log(x)-\mu}{\sigma}\right]\right\}\right\}}{x\sigma\left[1-(1-\theta)\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(x)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right]^2}, & \text{se} \quad \lambda < 0, \end{cases} \tag{5.4}$$

cuja forma gráfica indica que a distribuição GGG_{mn} é flexível, como mostra a Figura 5.1, cujos gráficos indicam também que os parâmetros λ e θ, que foram fixados nos gráficos, possuem um efeito substancial sobre a assimetria e curtose da distribuição.

A função de risco correspondente à distribuição GGG para a população de não curados é dada por

$$h_{\mathrm{GGG}}(x) = \frac{h_{\mathrm{GG}}(x)}{1-(1-\theta)S_{\mathrm{GG}}(x)}, \tag{5.5}$$

em que $h_{GG}(x)$ é a função de risco da distribuição GG e $S_{\mathrm{GG}}(x)$ é dada em (5.2).

Comparando as funções de risco das distribuições GGG e GG, observamos que o risco associado a um tempo $X = x$, qualquer, quando considerada a distribuição GG, é maior ou igual ao risco desse mesmo tempo quando considerada a distribuição GGG, isto é, $h(x)_{GG} \leq h_{GGG}(x)$, pois a razão $\dfrac{h_{GGG}(x)}{h_{GG}(x)}$ é crescente, e temos ainda que

- $\lim\limits_{x\to\infty} h_{\mathrm{GGG}}(x) = \lim\limits_{x\to\infty} h_{\mathrm{GG}}(x)$,

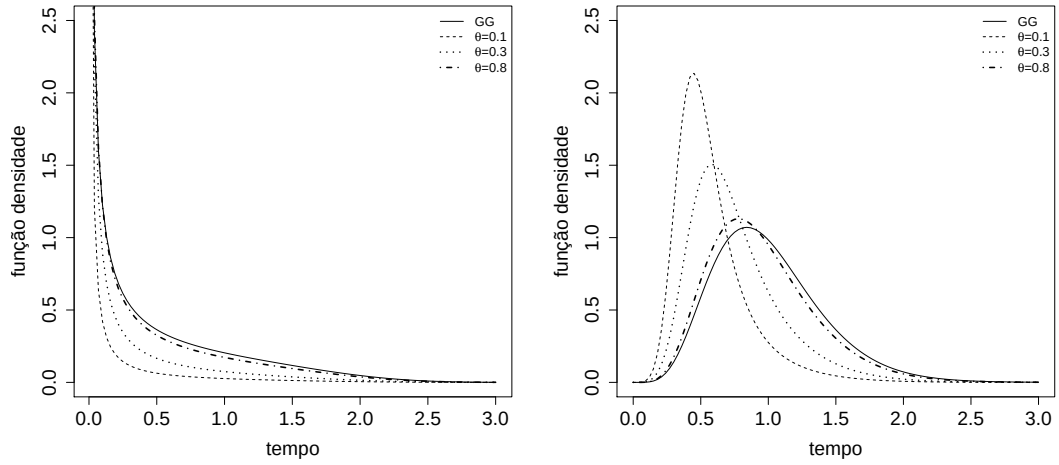

FIGURA 5.1: Função densidade de probabilidade da distribuição GGG_{mn} com parâmetro de escala $\mu = 0$, $\sigma = 1$ e $\lambda = 4$ (esquerda); $\sigma = 0,4$ e $\lambda = 0,4$ (direita).

- $\lim_{x\to 0} h_{\mathrm{GGG}}(x) = \dfrac{1}{\theta}\lim_{x\to 0} h_{\mathrm{GG}}(x)$.

Portanto, no limite, a função de risco da distribuição GGG_{mn} é igual à função de risco da distribuição GG. A Figura 5.2 mostra algumas formas da função de risco GGG_{mn}, considerando-se diferentes valores para λ e θ.

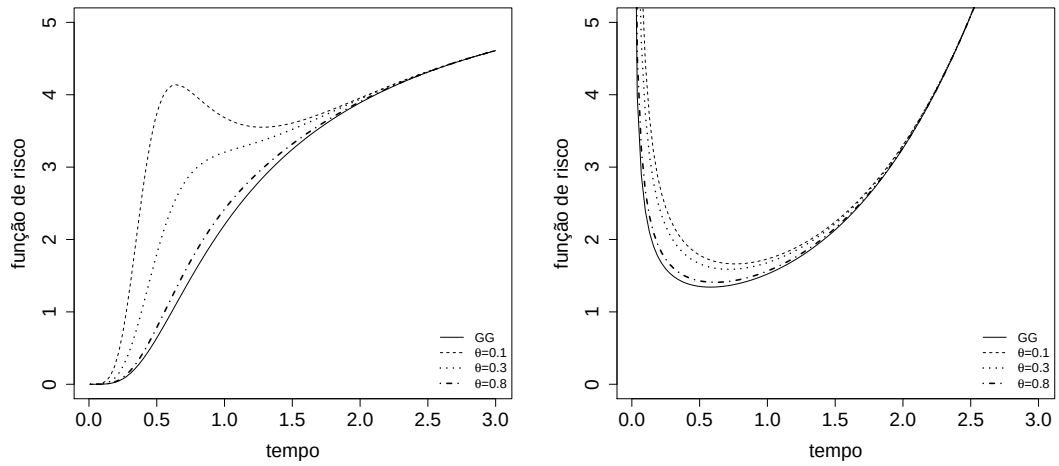

FIGURA 5.2: Função de risco da distribuição GGG_{mn} $\mu = 0$, $\sigma = 0,4$ e $\lambda = 0,4$ (esquerda) e $\sigma = 1$ e $\lambda = 4$ (direita).

5.2.2 Distribuição Gama Generalizada Geométrica para a última ativação

Considerando que o tempo Y amostrado refere-se ao tempo da última CR ativada e o exposto na Seção 2.4.2, a função de sobrevivência para a população de não curados é dada por

$$S_{\mathrm{GGG}}(y) = \begin{cases} \frac{\left(1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(y)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)}{1-(1-\theta)\left(\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(y)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)}, & \text{se} \quad \lambda > 0; \\ \frac{1-\Phi\left[\frac{\log(y)-\mu}{\sigma}\right]}{1-(1-\theta)\Phi\left[\frac{\log(y)-\mu}{\sigma}\right]}, & \text{se} \quad \lambda = 0, \\ \frac{\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(y)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}}{1-(1-\theta)\left(1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(y)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)}, & \text{se} \quad \lambda < 0, \end{cases} \tag{5.6}$$

a qual é própria, pois $S_{\mathrm{GGG}}(0) = 1$ e $S_{\mathrm{GGG}}(\infty) = 0$.

A f.d.p. da GGG_{mx}, conforme exposto na Seção 2.4.2, é dada por

$$f_{\mathrm{GGG}}(y) = \begin{cases} \frac{\frac{\lambda}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}\theta\exp\left\{\frac{1}{\lambda}\left[\frac{\log(y)-\mu}{\sigma}\right]-\frac{1}{\lambda^2}\exp\left\{\lambda\left[\frac{\log(y)-\mu}{\sigma}\right]\right\}\right\}}{y\sigma\left[1-(1-\theta)\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(y)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right]^2}, & \text{se} \quad \lambda > 0; \\ \frac{\theta\exp\left\{-\frac{1}{2}\left[\frac{\log(y)-\mu}{\sigma}\right]^2\right\}}{y\sqrt{2\pi}\sigma\left\{1-(1-\theta)\Phi\left[\frac{\log(y)-\mu}{\sigma}\right]\right\}^2}, & \text{se} \quad \lambda = 0, \\ \frac{\frac{\lambda}{\Gamma(\lambda^{-2})}(\lambda^{-2})^{\lambda^{-2}}\theta\exp\left\{\frac{1}{\lambda}\left[\frac{\log(y)-\mu}{\sigma}\right]-\frac{1}{\lambda^2}\exp\left\{\lambda\left[\frac{\log(y)-\mu}{\sigma}\right]\right\}\right\}}{y\sigma\left[1-(1-\theta)\left(1-\Gamma_G\left\{\lambda^{-2}\exp\left[\lambda\left(\frac{\log(y)-\mu}{\sigma}\right)\right];\lambda^{-2}\right\}\right)\right]^2}, & \text{se} \quad \lambda < 0, \end{cases} \tag{5.7}$$

Na Figura 5.3 estão algumas curvas da função densidade de probabilidade da distribuição GGG_{mx}.

Utilizando (2.7), a função de risco da distribuição GGG_{mx} para a população de não curados é

$$h_{\mathrm{GGG}}(y) = \frac{h_{\mathrm{GG}}(y)}{1-(1-\theta)F_{\mathrm{GG}}(y)}, \quad t > 0, \tag{5.8}$$

em que $h_{GG}(y)$ e $F_{\mathrm{GG}}(y)$ são, respectivamente, função de risco e distribuição da GG.

A razão entre as funções de risco das distribuições GGG e GG para a última ativação é decrescente para $0 < \theta < 1$. Temos ainda que $h(y)_{GG} \geq h_{GGG}(y)$ e

- $\lim\limits_{y\to\infty} h_{\mathrm{GGG}}(y) = \dfrac{1}{\theta}\lim\limits_{y\to\infty} h_{\mathrm{GG}}(y)$,
- $\lim\limits_{y\to 0} h_{\mathrm{GGG}}(y) = \lim\limits_{y\to 0} h_{\mathrm{GG}}(y)$.

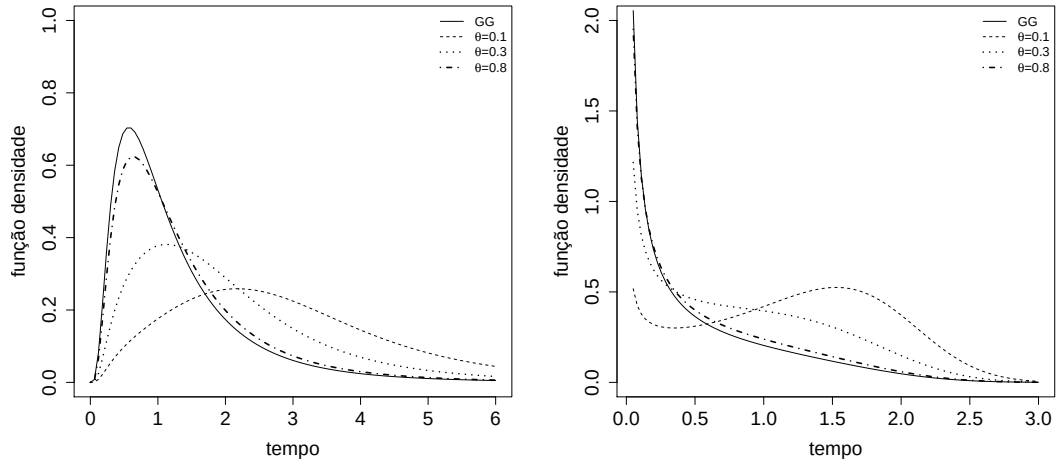

FIGURA 5.3: Função densidade de probabilidade da distribuição GGG_{mx} em comparação com a distribuição GG considerando a variação de θ para o parâmetro de escala $\mu = 0$, fixo. Esquerda: $\sigma = 0,75$ e $\lambda = 0$; Direita: $\sigma = 1$ e $\lambda = 4$.

Portanto, no limite, o comportamento da função de risco da GGG_{mx} é o mesmo da função de risco da GG. A Figura 5.4 apresenta algumas formas da função de risco GGG_{mx} para parâmetros fixados.

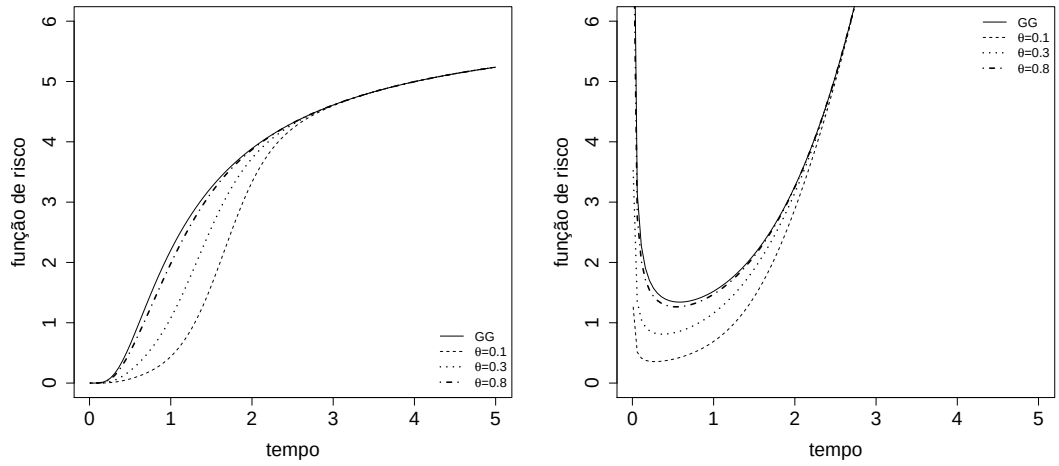

FIGURA 5.4: Gráficos da função de risco da distribuição GGG_{mx} perante a variação de θ e $\mu = 0$, comparando com a função de risco da distribuição GG. Esquerda: $\sigma = 0,4$ e $\lambda = 0,4$; Direita: $\sigma = 1$ e $\lambda = 4$.

5.2.3 Propriedades

Nesta seção apresentamos o r-*ésimo* momento da variável aleatória W. Consideramos $W = X$ para a distribuição GGG_{mn} e $W = Y$ para a distribuição GGG_{mx}. Também investigamos a variação das medidas de assimetria e curtose.

A f.d.p. considerando a ativação pelo mínimo dos tempos e $\lambda < 0$ corresponde no cenário da ativação pelo máximo dos tempos à f.d.p. para $\lambda > 0$. Representaremos ambas por f_{mn}. De forma similar, para a ativação pelo máximo dos tempos e $\lambda > 0$, a expressão da f.d.p. é a mesma do cenário de ativação pelo mínimo quando considerado $\lambda < 0$. Desta forma adotamos f_{mx} para representá-las.

Adotando $\alpha = e^{\mu}$, $\tau = \frac{\lambda}{\sigma}$, $\kappa = \lambda^{-2}$, temos que

$$f_{mn} = \frac{\tau\theta}{\alpha\Gamma(\kappa)}\left(\frac{w}{\alpha}\right)^{\tau\kappa-1}\kappa^{\kappa}e^{-\kappa\left(\frac{w}{\alpha}\right)^{\tau}}\left[1-(1-\theta)\left(1-\Gamma_G\left(\kappa\left(\frac{w}{\alpha}\right)^{\tau};\kappa\right)\right)\right]^{-2} \quad (5.9)$$

e

$$f_{mx} = \frac{\tau\theta}{\alpha\Gamma(\kappa)}\left(\frac{w}{\alpha}\right)^{\tau\kappa-1}\kappa^{\kappa}e^{-\kappa\left(\frac{w}{\alpha}\right)^{\tau}}\left[1-(1-\theta)\Gamma_G\left(\kappa\left(\frac{w}{\alpha}\right)^{\tau};\kappa\right)\right]^{-2}. \quad (5.10)$$

Esta parametrização pode apresentar problemas de convergência mesmo em conjuntos com mais de 200 observações, como afirma Lawless (2003), que acrescenta também que os métodos para obter as estimativas de máxima verossimilhança podem falhar na convergência e ainda têm como agravante que distribuições com valores muito diferentes de κ, α e τ podem parecer quase idênticas. Por tal fato não utilizaremos essa parametrização nos ajustes, apenas utilizar-se-á na obtenção das propriedades, as quais tornam-se simplificadas com a compactação das funções.

Proposição 5.1 *Se W tem função densidade de probabilidade f_{mn} ou f_{mx}, respectivamente, então o r-ésimo momento de W é dado por*

$$(i)\ u_r' = \frac{\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)}\sum_{j=0}^{\infty}(-1)^j\frac{j+1}{p^{j+2}}(1-p)^{j+1}I\left[\kappa,\frac{r}{\tau},j\right] \quad \textit{para } f_{mn}$$

ou,

$$(ii)\ u_r' = \frac{\alpha^r(1-p)}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)}\sum_{j=0}^{\infty}(j+1)p^jI\left[\kappa,\frac{r}{\tau},j\right] \quad \textit{para } f_{mx}$$

em que

$$I\left[\kappa,\frac{r}{\tau},j\right] = \frac{\Gamma\left(\frac{r}{\tau}+\kappa(j+1)\right)}{\kappa^j}F_A^{(j)}\left(\frac{r}{\tau}+\kappa(j+1);\kappa,\cdots,\kappa;\kappa+1,\cdots,\kappa+1;-1,\cdots,-1\right).$$

Prova 5.1 Por definição, temos que

$$u_r^{'}(w) = \int_0^\infty w^r f(w)dw = \alpha^r \int_0^\infty \left(\frac{w}{\alpha}\right)^r f(w)dw.$$

i) O r-ésimo da v.a. W com f.d.p. dada por f_{mn} é

$$\begin{aligned} u_r^{'} &= \frac{\theta\alpha^r}{\Gamma(\kappa)\kappa^{\frac{r}{\tau}}} \int_0^\infty \left[\kappa\left(\frac{y}{\alpha}\right)^\tau\right]^{\kappa+\frac{r}{\tau}-1} e^{-\kappa\left(\frac{y}{\alpha}\right)^\tau} \left[1-(1-\theta)\left(1-\Gamma_G\left(\kappa\left(\frac{y}{\alpha}\right)^\tau;\kappa\right)\right)\right]^{-2} \\ &\times \kappa\frac{\tau}{\alpha}\left(\frac{y}{\alpha}\right)^{\tau-1} dy. \end{aligned}$$

Assumindo $w = \kappa\left(\frac{y}{\alpha}\right)^\tau$ tem-se que $dw = \kappa\frac{\tau}{\alpha}\left(\frac{y}{\alpha}\right)^{\tau-1} dy$ então

$$u_r^{'} = \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)} \int_0^\infty w^{\kappa+\frac{r}{\tau}-1} e^{-w} \left[1-(1-\theta)\left(1-\Gamma_G(w;\kappa)\right)\right]^{-2} dw,$$

para $|x| < 1$ e $a > 0$, temos que $(1-x)^{-a} = \sum_{j=0}^{\infty} \frac{\Gamma(a+j)}{\Gamma(a)j!} x^j$, assim

$$u_r^{'} = \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)} \int_0^\infty w^{\kappa+\frac{r}{\tau}-1} e^{-w} \sum_{l=0}^{\infty} (l+1)(1-\theta)^l \left[1-\Gamma_G(w;\kappa)\right]^l dw$$

e utilizando o resultado: $(a-b)^j = \sum_{k=0}^{j} \binom{j}{k} (-1)^k a^{j-k} b^k$, segue que

$$u_r^{'} = \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)} \int_0^\infty w^{\kappa+\frac{r}{\tau}-1} e^{-w} \sum_{l=0}^{\infty}\sum_{j=0}^{l} (-1)^j (l+1)(1-\theta)^l \binom{l}{j} \left[\Gamma_G(w;\kappa)\right]^j dw.$$

Como $\sum_{j=0}^{\infty}\sum_{k=0}^{j} a_{jk} = \sum_{k=0}^{\infty}\sum_{j=k}^{\infty} a_{jk}$ (demonstração no Apêndice) e utilizando o resultado $(1-x)^{-b-1} = \sum_{a=b}^{\infty} \binom{a}{b} x^{a-b}$ (Abramowitz & Stegun (1972), p.822, equação 24.1.1) segue que

$$u_r^{'} = \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)} \int_0^\infty w^{\kappa+\frac{r}{\tau}-1} e^{-w} \sum_{j=0}^{\infty} (-1)^j \frac{(j+1)(1-\theta)^j}{\theta^{j+2}} \left[\Gamma_G(w;\kappa)\right]^j dw,$$

sendo $\Gamma(a,x) = \sum_{q=0}^{\infty} \frac{(-1)^q x^{a+q}}{q!(q+a)}$ (Johnson *et al.* (2005) p. 16), então

$$\begin{aligned} u_r^{'} &= \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)} \sum_{j=0}^{\infty} (-1)^j \frac{(j+1)(1-\theta)^j}{\theta^{j+2}} \int_0^\infty w^{\kappa+\frac{r}{\tau}-1} e^{-w} \left[\frac{w^\kappa}{\Gamma(\kappa)} \sum_{q=0}^{\infty} \frac{(-1)^q}{q!(\kappa+q)} w^q\right]^j dw \\ &= \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)} \sum_{j=0}^{\infty} (-1)^j \frac{(j+1)(1-\theta)^j}{\theta^{j+2}} I\left[\kappa, \frac{r}{\tau}, j\right]. \end{aligned}$$

ii) Considerando W com f.d.p. dada por f_{mx}, seu r-ésimo momento é obtido como segue:

$$u_r^{'} = \frac{\theta\alpha^r}{\Gamma(\kappa)\kappa^{\frac{r}{\tau}}}\int_0^\infty \left[\kappa\left(\frac{y}{\alpha}\right)^\tau\right]^{\kappa+\frac{r}{\tau}-1} e^{-\kappa\left(\frac{y}{\alpha}\right)^\tau}\left[1-(1-\theta)\Gamma_G\left(\kappa\left(\frac{y}{\alpha}\right)^\tau;\kappa\right)\right]^{-2}\kappa\frac{\tau}{\alpha}\left(\frac{y}{\alpha}\right)^{\tau-1}dy,$$

e assumindo $w = \kappa\left(\frac{y}{\alpha}\right)^\tau$ consequentemente tem-se que $dw = \kappa\frac{\tau}{\alpha}\left(\frac{y}{\alpha}\right)^{\tau-1}dy$ e então

$$u_r^{'} = \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)}\int_0^\infty w^{\kappa+\frac{r}{\tau}-1}e^{-w}\left[1-(1-\theta)\Gamma_G(w;\kappa)\right]^{-2}dw.$$

Para $|x|<1$ e $a>0$, segue que $(1-x)^{-a} = \sum_{j=0}^{\infty}\frac{\Gamma(a+j)}{\Gamma(a)j!}x^j$, assim,

$$u_r^{'} = \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)}\int_0^\infty w^{\kappa+\frac{r}{\tau}-1}e^{-w}\sum_{j=0}^{\infty}(j+1)(1-\theta)^j\Gamma_G(w;\kappa)^j dw,$$

como $\Gamma(a,x) = \sum_{q=0}^{\infty}\frac{(-1)^q x^{a+q}}{q!(q+a)}$ (Johnson *et al.* (2005) p. 16), então

$$\begin{aligned} u_r^{'} &= \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)}\sum_{j=0}^{\infty}(j+1)(1-\theta)^j\int_0^\infty w^{\kappa+\frac{r}{\tau}-1}e^{-w}\left[\frac{w^\kappa}{\Gamma(\kappa)}\sum_{q=0}^{\infty}\frac{(-1)^q}{q!(\kappa+q)}w^q\right]^j dw \\ &= \frac{\theta\alpha^r}{\kappa^{\frac{r}{\tau}}\Gamma(\kappa)}\sum_{j=0}^{\infty}(j+1)(1-\theta)^j I\left[\kappa,\frac{r}{\tau},j\right]. \quad \blacksquare \end{aligned}$$

Os momentos da variável aleatória W com distribuições GGG_{mn} ou GGG_{mx}, dados na Proposição (5.1), são os principais resultados desta seção. Investigamos graficamente a variação das medidas de assimetria e curtose, plotando seus quantis como função de λ e σ, respectivamente, e considerando um conjunto de valores para θ e $\mu = -0,7$ nos casos de $\sigma = 0,5$ ou $\lambda = 4$. Observamos ainda que o parâmetro θ não controla a forma dessas medidas (Figura A.6 do Apêndice).

5.2.4 Casos particulares

Algumas distribuições existentes na literatura são casos particulares das distribuições $GGGcr$. Apresentamos-as de forma geral para os mecanismos de ativação máximo e mínimo.

- Para $\lambda = \sigma$, temos a distribuição Gama Geométrica com f.d.p. dada por

$$f_{\text{GG}}(w) = \begin{cases} \dfrac{\frac{\theta}{\Gamma(\lambda^{-2})} w^{(\lambda^{-2}-1)} (\lambda^{-2})^{\lambda^{-2}} e^{-\mu\lambda^{-2}} \exp\left\{ w^{\frac{1}{\lambda}} \lambda^{-2} e^{\frac{-\mu}{\lambda}} \right\}}{[1-(1-\theta)(1-\Gamma\{w\lambda^{-2}e^{-\mu};\lambda^{-2}\})]^2}, & \text{para } CR_{mn}; \\ \dfrac{\frac{\theta}{\Gamma(\lambda^{-2})} w^{(\lambda^{-2}-1)} (\lambda^{-2})^{\lambda^{-2}} e^{-\mu\lambda^{-2}} \exp\left\{ w^{\frac{1}{\lambda}} \lambda^{-2} e^{\frac{-\mu}{\lambda}} \right\}}{[1-(1-\theta)\Gamma\{w\lambda^{-2}e^{-\mu};\lambda^{-2}\}]^2}, & \text{para } CR_{mx}. \end{cases} \tag{5.11}$$

- Para $\lambda = 1$, temos a distribuição Weibull Geométrica

$$f_{\text{WG}}(w) = \begin{cases} \dfrac{\theta w^{\frac{1}{\sigma}-1} \sigma^{-1} e^{-\frac{\mu}{\sigma}} \exp\left\{ -\left(we^{-\mu}\right)^{\frac{1}{\sigma}} \right\}}{\left[1-(1-\theta)\left(1-\Gamma\left\{(we^{-\mu})^{\frac{1}{\sigma}};1\right\}\right)\right]^2}, & \text{para } CR_{mn}; \\ \dfrac{\theta w^{\frac{1}{\sigma}-1} \sigma^{-1} e^{-\frac{\mu}{\sigma}} \exp\left\{ -\left(we^{-\mu}\right)^{\frac{1}{\sigma}} \right\}}{\left[1-(1-\theta)\left(\Gamma\left\{(we^{-\mu})^{\frac{1}{\sigma}};1\right\}\right)\right]^2}, & \text{para } CR_{mx}; \end{cases} \tag{5.12}$$

  sendo que para a primeira ativação, se considerarmos $\alpha = \frac{1}{\sigma}$, $p = 1-\theta$ e $\beta = e^{-\mu}$ temos a distribuição WG apresentada por Barreto-Souza *et al.* (2010), e para a última ativação temos a distribuição Weibull Geométrica Complementar (CWG) proposta por Tojeiro *et al.* (2013).

- Para $\lambda = \sigma = 1$ temos a distribuição Exponencial Geométrica

$$f_{\text{EG}}(w) = \begin{cases} \dfrac{\theta e^{-\mu} \exp\left\{ -we^{-\mu} \right\}}{[1-(1-\theta)(1-\Gamma\{we^{-\mu};1\})]^2}, & \text{para } CR_{mn}; \\ \dfrac{\theta e^{-\mu} \exp\left\{ -we^{-\mu} \right\}}{[1-(1-\theta)(\Gamma\{we^{-\mu};1\})]^2}, & \text{para } CR_{mx}; \end{cases} \tag{5.13}$$

  e para a primeira ativação, se considerarmos $p = 1-\theta$ e $\beta = e^{-\mu}$ temos a distribuição EG apresentado por Adamidis & Loukas (1998), e para a última ativação, se fizermos $\lambda = e^{-\mu}$, temos a distribuição Exponencial Geométrica Complementar (CEG) proposto por Louzada *et al.* (2011). Estas distribuições foram apresentadas no Capítulo 3.

- Para $\lambda = 0$, temos a distribuição LogNormal Geométrica

$$f_{\text{LNG}}(w) = \begin{cases} \dfrac{\theta \exp\left\{ -\frac{1}{2}\left[\frac{\log(w)-\mu}{\sigma}\right]^2 \right\}}{y\sqrt{2\pi}\sigma\left[1-(1-\theta)\Phi\left[-\frac{\log(w)-\mu}{\sigma}\right]\right]^2}, & \text{para } CR_{mn}; \\ \dfrac{\theta \exp\left\{ -\frac{1}{2}\left[\frac{\log(w)-\mu}{\sigma}\right]^2 \right\}}{w\sqrt{2\pi}\sigma\left[1-(1-\theta)\Phi\left[\frac{\log(w)-\mu}{\sigma}\right]\right]^2}. & \text{para } CR_{mx} \end{cases} \tag{5.14}$$

- Para $\lambda = -1$, temos a distribuição Weibul Inversa Geométrica

$$f_{\text{WIG}}(w) = \begin{cases} \dfrac{-\theta(\frac{1}{w})^{\frac{1}{\sigma}+1} \sigma^{-1} e^{\frac{\mu}{\sigma}} \exp\left\{ -(\frac{1}{w})^{\frac{1}{\sigma}} e^{\frac{\mu}{\sigma}} \right\}}{\left[1-(1-\theta)\left(\Gamma\left\{(\frac{1}{w})^{\frac{1}{\sigma}} e^{\frac{\mu}{\sigma}};1\right\}\right)\right]^2}, & \text{para } CR_{mn}; \\ \dfrac{-\theta(\frac{1}{w})^{\frac{1}{\sigma}+1} \sigma^{-1} e^{\frac{\mu}{\sigma}} \exp\left\{ -(\frac{1}{w})^{\frac{1}{\sigma}} e^{\frac{\mu}{\sigma}} \right\}}{\left[1-(1-\theta)\left(1-\Gamma\left\{(\frac{1}{w})^{\frac{1}{\sigma}} e^{\frac{\mu}{\sigma}};1\right\}\right)\right]^2}, & \text{para } CR_{mx}. \end{cases} \tag{5.15}$$

5.3 Distribuições Gama Generalizada Geométrica de longa duração

Fazem parte das distribuições Gama Generalizada Geométrica de longa duração e causas de risco latentes *LGGGcr* as distribuições Gama Generalizada Geométrica com longa duração para a primeira ativação ($LGGG_{mn}$), Gama Generalizada Geométrica com longa duração para a ativação aleatória ($LGGG_{al}$) e Gama Generalizada Geométrica com longa duração para a última ativação ($LGGG_{mx}$). Devido ao fato das expressões matemáticas serem demasiadamente grandes, utilizamos as funções da distribuição Gama Generalizada para expressar as funções das distribuições da mistura *LGGGcr*.

$$f_{\text{pop}}(W=w)=\begin{cases}\frac{\theta(1-\theta)f_{GG}(w)}{[1-(1-\theta)S_{GG}(w)]^2}, & \text{para } CR_{mn};\\ \theta(1-\theta)S_{GG}(w), & \text{para } CR_{al};\\ \frac{\theta(1-\theta)f_{GG}(w)}{[1-(1-\theta)F_{GG}(w)]^2}, & \text{para } CR_{mx},\end{cases} \tag{5.16}$$

em que $f_{GG}(w)$ é dada em (5.1) e $F_{GG}(w)$ especificada em (5.2).

As funções densidades de probabilidade são denotadas por $f_{LGGG_{mn}}$ o mínimo dos tempos, $f_{LGGG_{al}}$ para o cenário CR_{al} e $f_{LGGG_{mx}}$ para o máximo dos tempos.

A função de sobrevivência para cada uma das três ativações é

$$S_{\text{pop}}(w)=\begin{cases}\frac{\theta}{1-(1-\theta)S_{GG}(w)}, & \text{para a primeira ativação;}\\ (1-\theta)f_{GG}(w), & \text{para a ativação aleatória;}\\ 1-\frac{\theta(1-\theta)F_{GG}(w)}{1-(1-\theta)F_{GG}(w)}, & \text{para a última ativação.}\end{cases} \tag{5.17}$$

Conforme apresentado na Proposição 2.10 as funções de sobrevivência nos diferentes esquemas de ativação estão relacionadas. Tal hierarquia em termos de valor da função sobrevivência, indica que para a CR_{mx} a função de sobrevivência assume o valor da proporção de curados θ em momento posterior se comparada aos demais, levando-nos a crer que eventos acionados pelo CR_{mx} demoram mais para ocorrer (vide Figura 5.5).

A Figura 5.6 apresenta a função de sobrevivência para os três esquemas de ativação, comparando as curvas para diferentes valores de θ. No cenário CR_{mx} observa-se

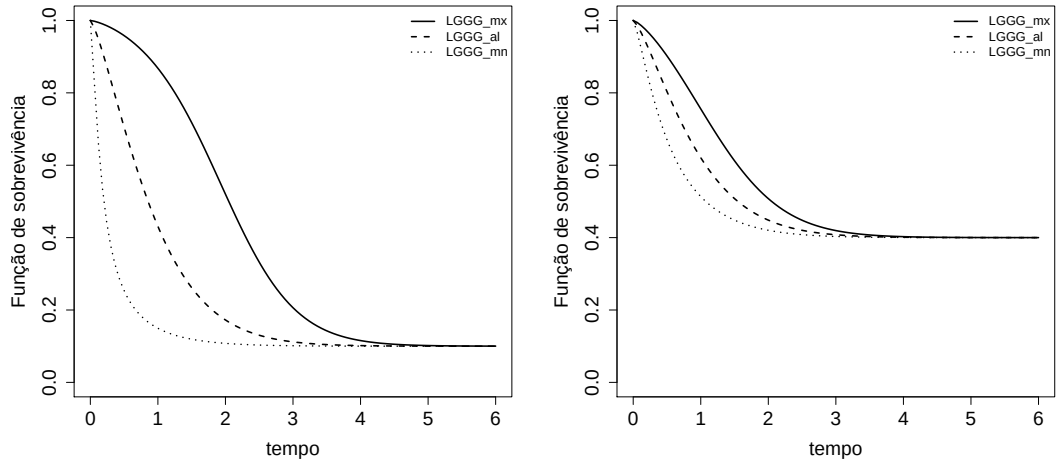

FIGURA 5.5: Função de sobrevivência das distribuições $LGGG_{mn}$ $LGGG_{al}$ e $LGGG_{mx}$ para os valores fixos $\mu = 0,\ \lambda = 1$ *e* $\sigma = 0,75$, sendo $\theta = 0,1$ e $0,4$ respectivamente.

que no quando o tempo e a proporção de imunes na população são pequenos o valor assumido pela função de sobrevivência é maior. Como interpretação, na área da saúde, podemos dizer que uma doença que atinge praticamente todas as pessoas é desencadeada de forma mais lenta que doenças que atingem apenas parte da população. Esta condição coincide com a conduta, seja de órgãos governamentais ou pessoal, de despender maior atenção para situações que podem acometer toda a população em relação aos casos menos comuns.

A existência de uma relação matemática entre os modelos de mistura com e sem longa duração dada em (2.18) com os modelos de mistura com fração de cura BG (Boag, 1949; Berkson & Gage, 1952), para este caso específico pode ser escrita como

$$S_{pop}(w) = \theta + (1 - \theta)S_{\mathrm{GGG}}(w),$$

em que $S_{\mathrm{GGG}}(y)$ é dado por (5.3) ou (5.6). Então, $S_{\mathrm{pop}}(w)$ é uma mistura de modelos com fração de cura com fração de cura $p_0 = \theta$ e função de sobrevivência $S_{\mathrm{GGG}}(y)$ para a população de não curados.

5.3.1 Função de verossimilhança

Consideremos a situação em que o tempo de falha W não é completamente observado e está sujeito a censura à direita. Seja C_i o tempo de censura. Em uma

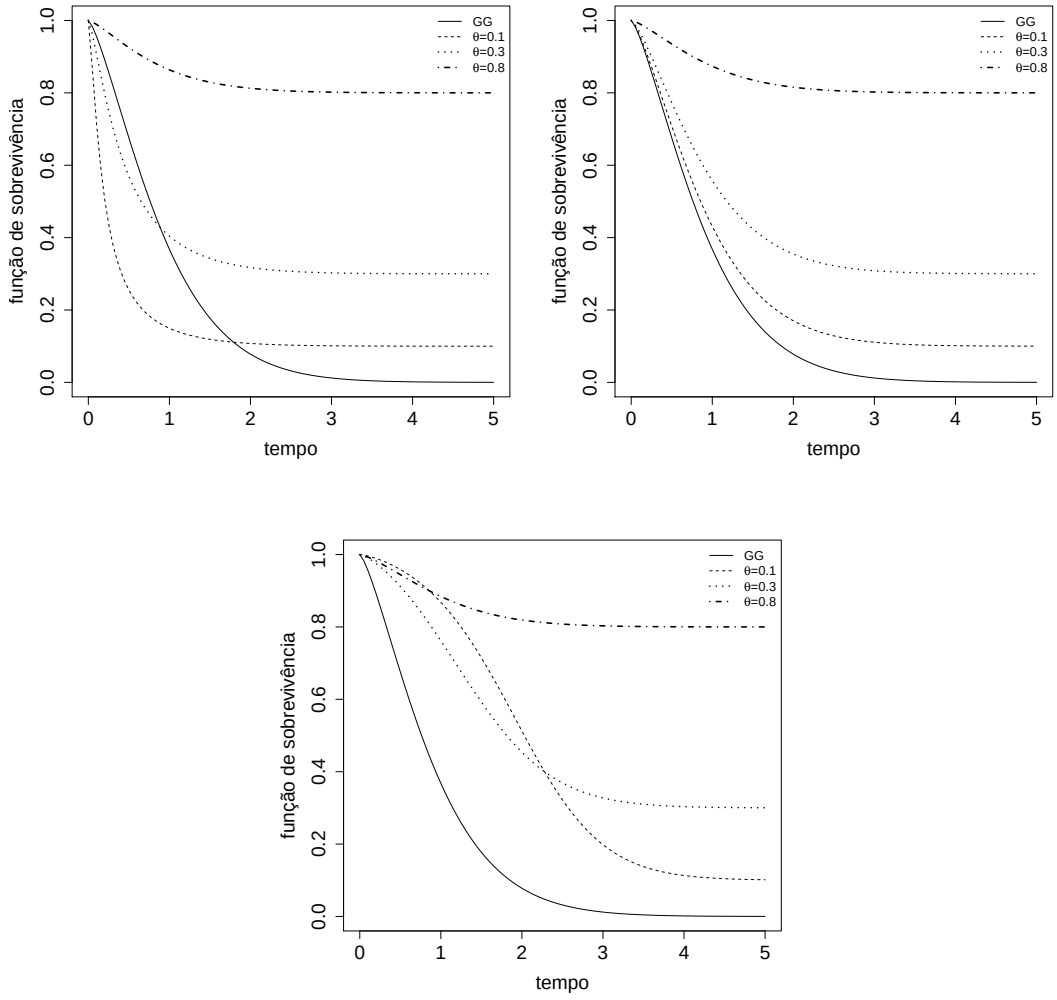

FIGURA 5.6: Função de sobrevivência para $\mu = 0,\ \sigma = 0,75,\ \lambda = 1$ e variação de θ das distribuições $LGGG_{mn}$, $LGGG_{al}$ e $LGGG_{mx}$, respectivamente.

amostra de tamanho n, observa-se $W_i = \min\{W_i, C_i\}$ e $\delta_i = \mathrm{I}(W_i \leq C_i)$, em que $\delta_i = 1$ se W_i é o tempo de falha e $\delta_i = 0$ se o dado for censurado à direita, para $i = 1, \ldots, n$.

Seja $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^\top$ denotando o vetor de covariáveis para o i-*ésimo* indivíduo. Para completar o modelo, propomos que a função de ligação com as covariáveis seja na fração de cura e pela relação logística

$$\log\left(\frac{p_{0i}}{1-p_{0i}}\right) = \boldsymbol{z}_i^\top \boldsymbol{\beta} \quad \text{ou} \quad p_{0i} = \frac{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})}{1+\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})}, \tag{5.18}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_p)^\top$ é o vetor de coeficientes de regressão. Desta forma, para cada grupo de indivíduos representado por $\boldsymbol{z}_i$, temos uma fração de cura diferente. Com essa função de ligação os modelos se tornam identificáveis, no sentido de Li *et al.* (2001).

Com a expressão (5.18) escrevemos a função verossimilhança de $\boldsymbol{\vartheta} = (\mu, \sigma, \lambda, \boldsymbol{\beta}^\top)^\top$, considerando censura não informativa, como sendo

$$L(\boldsymbol{\vartheta}; \mathcal{D}) \propto \prod_{i=1}^{n} f_{\text{pop}}(w_i; \boldsymbol{\vartheta})^{\delta_i} \, S_{\text{pop}}(w_i; \boldsymbol{\vartheta})^{1-\delta_i}, \tag{5.19}$$

em que $\mathcal{D} = (\boldsymbol{w}, \boldsymbol{\delta}, \boldsymbol{z})$, $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$, $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$ e $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$, enquanto que $f_{\text{pop}}(\cdot; \boldsymbol{\vartheta})$ e $S_{\text{pop}}(\cdot; \boldsymbol{\vartheta})$ são funções impróprias de densidade e sobrevivência, dadas em (5.16) e (5.17), respectivamente.

5.3.2 Aplicação

Nesta seção utilizamos os conjuntos de dados $T4$ e $T5$ apresentados na Seção 1.3, cujos gráficos Kaplan-Meier indicam que $T4$ e $T5$ não apresentam fração de curados, embora em $T5$ há uma queda brusca da função de sobrevivência no final dos tempos, após um período longo de estabilidade em torno de $0,5\%$. A forma da função de risco é crescente para ambos.

Consideramos as mesmas covariáveis nos conjuntos de dados, z_{i1} e z_{i3}, que correspondem, respectivamente, à classificação de Karnofsky e ao tempo entre o diagnóstico e o início do tratamento. E lembrando que o conjunto $T5 \subset T4$, em que $T5$ corresponde aos pacientes que receberam tratamento prévio ao considerado no estudo.

Pela maximização direta do logaritmo da função verosimilhança (via rotina *optim* no programa `R`), obtivemos as EMVs para os três modelos apresentados na Seção 5.3 e do modelo de mistura de longa duração BG para a distribuição Weibull (WL). Considerando os critérios $-\ell(\boldsymbol{\vartheta})$ e AIC, cujos valores estão disponíveis na Tabela 5.1, o modelo $LGGG_{mn}$ é a melhor opção de ajuste para os dados completos $T4$, e, quando consideramos apenas os pacientes que haviam recebido algum tipo de tratamento anterior ao estudo considerado ($T5$), temos $LGGG_{mx}$ como o modelo que proporciona o melhor ajuste.

A Tabela 5.2 apresenta as EMVs e seus respectivos desvios padrão para os parâmetros dos modelos proporcionadores dos melhores ajustes segundo os critérios considerados. O gráfico Q-Q dos ajustes estão disponíveis na Figura A.4 (Apêndice).

Ressaltamos que o conjunto $T5$ corresponde a menos de 30% dos dados de $T4$ e refere-se aos pacientes que já haviam recebido um tratamento prévio ao estudo.

TABELA 5.1: Valor dos critérios $-\ell(\boldsymbol{\vartheta})$ e AIC para os modelos ajustados aos dados dos conjuntos $T4$ e $T5$.

Dados	$LGGG_{mn}$		$LGGG_{al}$		$LGGG_{mx}$		WL	
	$-\ell(\boldsymbol{\vartheta})$	AIC	$-\ell(\boldsymbol{\vartheta})$	AIC	$-\ell(\boldsymbol{\vartheta})$	AIC	$-\ell(\boldsymbol{\vartheta})$	AIC
$T4$	720,161	1452,320	746,471	1504,942	723,090	1458,180	748,091	1506,183
$T5$	3,460	18,920	0,837	13,674	-1,145	9,71	1,434	12,868

TABELA 5.2: EMVs dos conjuntos $T6$ e $T4$ nos modelos $LGGG_{mx}$ e $LGGG_{mn}$, respectivamente.

	$T4$ - $LGGG_{mn}$		$T5$ - $LGGG_{mx}$	
Parâmetro	Média	DP	Média	DP
μ	7,239	0,893	-3,661	0,765
σ	0,356	0,893	1,513	0,265
λ	1,713	1,535	0,079	0,564
β_0	-9,162	1,700	-1,292	1,074
β_1	0,065	0,008	-0,034	0,022
β_2	0,007	0,014	0,013	0,017

DP: Desvio padrão

Relacionando esta particularidade do conjunto de dados com à ideia incutida na mistura dos modelos, podemos dizer que o tratamento prévio ao estudo (recebido pelos elementos que compõem $T5$) postergou a ocorrência do evento de interesse e, desta forma, o modelo que melhor representaria o contexto dos dados seria o modelo dos máximos dos tempos, como ocorreu.

O modelo Weibull, amplamente utilizado em AS para modelar dados de tempo de sobrevida, não foi a melhor opção pelos critérios considerados entretanto não foi a pior entre as consideradas.

5.4 Comentários

As distribuições que pertencem às composições propostas - *GGGcr* e *LGGGcr* - são alternativas flexíveis e abrangentes na modelagem de dados de sobrevivência, até por terem como casos particulares vários modelos bem difundidos na área e também por abrangerem boa parte das formas de risco comumente encontradas. A possibilidade de relacionar o melhor ajuste considerado com a realidade dos dados é outro fator importante e pode ser explorado em contextos reais.

A modelagem, considerando o grupo dos imunes e os esquemas de ativação, torna o modelo mais realista, e ainda pode ser escrito como o modelo Berkson e Gage, frequentemente utilizado quando trata-se de fração de cura.

Capítulo 6

Abordagem bayesiana: Diagnóstico de ponto influente e inferência

Este capítulo traz a técnica de diagnóstico de ponto influente na amostra e técnicas de comparação de ajustes, concebidas com uma metodologia bayesiana para os modelos $LGGG_{mn}$, $LGGG_{al}$ e $LGGG_{mx}$. O ajuste bayesiano de tais modelos à dados de melanoma foi realizado considerando a análise de ponto influente e avaliando o impacto da covariável na fração de curados.

6.1 Introdução

Consideramos os modelos de mistura Gama Generalizada Geométrico cujas f.d.p. e função de sobrevivência são dadas em (5.16) e (5.17), cujos parâmetros são μ, σ, λ e θ. Sendo $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^\top$ o vetor de covariáveis e introduzido no modelo considerando a relação logística [vide equação (5.18)] através do parâmetro θ.

A inferência bayesiana é uma alternativa para a estimação dos parâmetros resultante $\boldsymbol{\vartheta} = (\mu, \sigma, \lambda, \boldsymbol{\beta}^\top)^\top$ dos modelos $LGGGcr$. Considerando distribuições *a priori* próprias asseguramos que a distribuição *a posteriori* é própria (Ibrahim *et al.*, 2001).

6.2 Distribuições *a priori* e *a posteriori*

A distribuição Normal com média μ_o e variância σ_o e a distribuição Gama com parâmetro de forma a, parâmetro de escala b e média a/b são denotadas por $N(\mu, \sigma^2)$ e $G(a, b)$. Por simplicidade, assumimos que $\boldsymbol{\beta_j}$, μ, σ e λ são independentes *a priori* e têm distribuições *a priori* de acordo com seus espaços paramétricos, de forma que obtemos

$$\pi(\boldsymbol{\vartheta}) = \prod_{i=0}^{p} \pi(\beta_i)\pi(\mu)\pi(\sigma)\pi(\lambda), \tag{6.1}$$

em que $\beta_i \sim N(0, \sigma^2_{\beta_i})$, $i = 0, 1, \ldots, p$, $\mu \sim N_1(0, \sigma^2_\mu)$, $\sigma \sim \mathrm{G}(a_0, b_0)$ e $\lambda \sim N_1(0, \sigma^2_\lambda)$.

Todos os hiperparâmetros são especificados para expressar *priori's* vagas; a distribuição *a posteriori* conjunta para $\boldsymbol{\vartheta}$ é obtida por

$$\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\vartheta}; \boldsymbol{\mathcal{D}}) \prod_{i=0}^{p} \pi(\beta_i)\pi(\mu)\pi(\sigma)\pi(\lambda). \tag{6.2}$$

Em decorrência desta função densidade *a posteriori* conjunta ser analiticamente intratável, recorremos ao método de simulação inferencial de Monte Carlo via Cadeias de Markov (MCMC), pelo algoritmo de Metropolis–Hasting, para gerar amostras da distribuição *a posteriori* dos parâmetros (vide Gamerman & Lopes (2006)). O algoritmo foi implementado na linguagem `R` (R Core Team, 2011) e a convergência foi monitorada utilizando o método proposto por Geweke (1992) e métodos gráficos.

Iniciamos o processo de implementação do algoritmo transformando as variáveis $\boldsymbol{\vartheta} = (\lambda, \mu, \log(\sigma), \boldsymbol{\beta}_i)$ para trabalharmos com todos os parâmetros na reta. Esta transformação muda o espaço paramétrico para $\mathcal{R}^{p+4}$. Como matriz de transição utilizamos uma distribuição Normal multivariada cuja vetor de médias corresponde ao ajuste clássico do logaritmo da posteriori e como matriz de covariância utilizamos a decomposição de Choleski na negativa da inversa da matriz hessiana do mesmo ajuste. Para implementar o algoritmo de Metropolis–Hastings, procedemos como segue:

(1) iniciar com o vetor $\boldsymbol{\vartheta}_{(j)}$ e o indicador de iterações em $j = 0$;

(2) gerar um ponto $\boldsymbol{\vartheta}'$ de acordo com o núcleo de transição $Q(\boldsymbol{\vartheta}', \boldsymbol{\vartheta}_j) = N_{p+4}\left(\boldsymbol{\vartheta}_j, \tilde{\Sigma}\right)$, em que $\tilde{\Sigma}$ é a matriz de covariância, que é a mesma em qualquer estágio;

(3) calcular a probabilidade de aceitação $p_j = \min\{1, \pi(\boldsymbol{\vartheta}'|\mathbf{\mathcal{D}})/\pi(\boldsymbol{\vartheta}_{(j)}|\mathbf{\mathcal{D}})\}$ e gera u de $U(0,1)$.

(4) atualizar

$$\boldsymbol{\vartheta}_{(j+1)} = \begin{cases} \boldsymbol{\vartheta}', & \text{se } u \leq p_j \\ \boldsymbol{\vartheta}_{(j)}, & \text{caso contrário} \end{cases};$$

(5) repetir os passos (2), (3) e (4) aumentando o contador até se obter a distribuição estacionária.

6.3 Critérios para comparação de modelos

Há uma variedade de metodologias para comparar modelos ajustados e determinar o melhor ajuste. Um dos critérios utilizados na literatura é o da estatística ordenada da distribuição preditiva condicional (CPO) - para maiores detalhes sobre esta estatística e aplicações em seleção de modelos, vide Gelfand *et al.* (1992) e Geisser & Eddy (1979).

Consideremos $\mathbf{\mathcal{D}}$ os dados completos e $\mathbf{\mathcal{D}}^{(-i)}$ os dados sem a i-*ésima* observação e $g(w_i|\boldsymbol{\vartheta}) = f_{\text{pop}}(w_i;\boldsymbol{\vartheta})$ se o tempo foi observado, ou $g(w_i|\boldsymbol{\vartheta}) = S_{\text{pop}}(w_i;\boldsymbol{\vartheta})$ se o tempo foi censurado. Denotamos a função densidade *a posteriori* de $\boldsymbol{\vartheta}$ dado $\mathbf{\mathcal{D}}^{(-i)}$ por $\pi(\boldsymbol{\vartheta}|\mathbf{\mathcal{D}}^{(-i)})$, $i = 1, \ldots, n$. Para a i-*ésima* observação, o CPO_i pode ser escrito como

$$CPO_i = \int_{\vartheta\in\Theta} g(w_i|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}|\mathbf{\mathcal{D}}^{(-i)})d\boldsymbol{\vartheta} = \left\{\int_{\vartheta} \frac{\pi(\boldsymbol{\vartheta}|\mathbf{\mathcal{D}})}{g(w_i|\boldsymbol{\vartheta})}d\boldsymbol{\vartheta}\right\}^{-1}. \tag{6.3}$$

O melhor modelo ajustado é o que detém o maior valor do CPO_i, pois ele pode ser interpretado como a altura da densidade marginal do tempo de vida para o evento no ponto w_i. No entanto, estamos limitados ao fato de que nem sempre há uma forma fechada para o CPO_i, o que ocorre em modelos mais complexos, como no nosso caso. Assim, uma estimativa de Monte Carlo para o CPO_i pode ser obtida usando uma amostra MCMC da distribuição *a posteriori* $\pi(\boldsymbol{\vartheta}|\mathbf{\mathcal{D}})$.

Seja $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(Q)}$ uma amostra de tamanho Q de $\pi(\boldsymbol{\vartheta}|\mathbf{\mathcal{D}})$, uma aproximação de Monte Carlo do CPO_i (Ibrahim *et al.*, 2001) é dada por

$$\widehat{CPO_i} = \left\{\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{g(w_i|\boldsymbol{\vartheta}^{(q)})}\right\}^{-1}.$$

Para a comparação de modelos, utilizamos a estatística logaritmo da pseudoverossimilhança marginal (LPML) definida por

$$LPML = \sum_{i=1}^{n} \log(\widehat{CPO}_i),$$

e quanto maior seu valor, melhor é o modelo ajustado (Cancho *et al.*, 2010b). Nos limitamos às conclusões decorrentes da LPML para a seleção de modelos.

Outros critérios, como DIC (Spiegelhalter *et al.*, 2002), $EAIC$ (Brooks, 2002) e $EBIC$ (Carlin & Louis, 2001), são baseados na média *a posteriori* do desvio, que pode ser aproximada por $\overline{d} = \sum_{q=1}^{Q} d(\boldsymbol{\vartheta}_q)/Q$, em que $d(\boldsymbol{\vartheta}) = -2\sum_{i=1}^{n} \log\left[g(w_i|\boldsymbol{\vartheta})\right]$.

O critério DIC pode ser obtido considerando amostras de MCMC, de forma que $\widehat{DIC} = \overline{d} + \widehat{\rho}_d = 2\overline{d} - \widehat{d}$, sendo ρ_d o número efetivo de parâmetro e definido como $E\{d(\boldsymbol{\vartheta})\} - d\{E(\boldsymbol{\vartheta})\}$, em que $d\{E(\boldsymbol{\vartheta})\}$ é o desvio avaliado na média da distribuição *a posteriori* e pode ser estimado por

$$\widehat{D} = d\left(\frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\beta}^{(q)}, \frac{1}{Q}\sum_{q=1}^{Q}\gamma_1^{(q)}, \frac{1}{Q}\sum_{q=1}^{Q}\gamma_2^{(q)}\right).$$

De forma similar, os critérios EAIC e EBIC podem ser estimados considerando

$$\widehat{EAIC} = \overline{d} + 2\#(\boldsymbol{\vartheta}) \quad \text{e} \quad \widehat{EBIC} = \overline{d} + \#(\boldsymbol{\vartheta})\log(n),$$

em que $\#(\boldsymbol{\vartheta})$ é o número de parâmetros do modelo.

6.4 Análise de influência bayesiana

Uma vez que os modelos de regressão são sensíveis às suposições básicas, é aconselhável realizar a análise de sensibilidade. Cook (1986) usa esta ideia para motivar sua avaliação de análise de influência, que propõe uma perturbação dos componentes do modelo, ou dos dados, utilizando a função de afastamento pela verossimilhança do modelo. Ele afirma também que o modelo estável para pequenas modificações é digno de maior confiança.

Os melhores esquemas de perturbação já apresentados são baseados na supressão de casos (Cook & Weisberg, 1982), cuja análise estuda o efeito da remoção completa do

mesmo. Este mesmo raciocínio é a base do desenvolvimento da metodologia da análise de influência global bayesiana, que possibilita determinar quais casos podem influenciar nos resultados da análise.

Seja $D_\psi(P, P_{(-i)})$ a divergência ψ entre P e $P_{(-i)}$. Temos que P é a distribuição *a posteriori* $\boldsymbol{\vartheta}$ para os dados completos e $P_{(-i)}$ a distribuição *a posteriori* de $\boldsymbol{\vartheta}$ sem o i-*ésimo* caso. Especificamente,

$$D_\psi(P, P_{(-i)}) = \int_{\vartheta\in\Theta} \psi\left(\frac{\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}^{(-i)})}{\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})}\right)\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})\,d\boldsymbol{\vartheta}, \tag{6.4}$$

em que ψ é a função convexa com $\psi(1) = 0$.

Várias opções para ψ são dadas em Dey & Birmiwal (1994). Por exemplo,

- se $\psi(z) = -\log(z)$, temos a divergência de Kullback-Leibler (K-L);
- $\psi(z) = (z-1)\log(z)$ retornamos à *J-distance* (versão simétrica de divergência K-L);
- se $\psi(z) = 0,5|z-1|$ temos a distância variacional ou norma L_1; e
- para $\psi(z) = (z-1)^2$ retorna a divergência qui-quadrado $\chi^2 - \textit{divergence}$.

Há uma relação entre o CPO e a medida de divergência ψ (Cancho *et al.*, 2010b, 2011). Para expressar matematicamente esta relação considera-se $\pi(\boldsymbol{\vartheta})$ a distribuição *a priori* e $\prod_{i=1}^n g(w_i|\boldsymbol{\vartheta})$ a função de verossimilhança para $\boldsymbol{\vartheta}$. Pelo teorema de Bayes, a distribuição *a posteriori* é

$$\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}) = \frac{\pi(\boldsymbol{\vartheta})\prod_{j\in\boldsymbol{\mathcal{D}}} g(w_j|\boldsymbol{\vartheta})}{\int_{\boldsymbol{\vartheta}\in\Theta}\pi(\boldsymbol{\vartheta})\prod_{j\in\boldsymbol{\mathcal{D}}} g(w_j|\boldsymbol{\vartheta})d\boldsymbol{\vartheta}},$$

e a razão de *posteriori's* em (6.4) pode ser escrita como

$$\begin{aligned}
\frac{\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}^{(-i)})}{\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})} &= \frac{\pi(\boldsymbol{\vartheta})\prod_{j\in D^{(i)}} g(w_i|\boldsymbol{\vartheta})}{\int_{\boldsymbol{\vartheta}\in\Theta}\pi(\boldsymbol{\vartheta})\prod_{j\in D^{(i)}} g(w_j|\boldsymbol{\vartheta})d\boldsymbol{\vartheta}} \times \frac{\int_{\boldsymbol{\vartheta}\in\Theta}\pi(\boldsymbol{\vartheta})\prod_{j\in D} g(w_j|\boldsymbol{\vartheta})d\boldsymbol{\vartheta}}{\pi(\boldsymbol{\vartheta})\prod_{j\in D} g(w_j|\boldsymbol{\vartheta})} \\
&= \frac{1}{g(w_i|\boldsymbol{\vartheta})} \times \frac{\int_{\boldsymbol{\vartheta}\in\Theta}\pi(\boldsymbol{\vartheta})\prod_{j\in D} g(w_j|\boldsymbol{\vartheta})d\boldsymbol{\vartheta}}{\int_{\boldsymbol{\vartheta}\in\Theta}\frac{1}{g(w_i|\boldsymbol{\vartheta})}\pi(\boldsymbol{\vartheta})\prod_{j\in D} g(w_j|\boldsymbol{\vartheta})d\boldsymbol{\vartheta}} \\
&= \frac{\left(\int_{\boldsymbol{\vartheta}\in\Theta}\frac{1}{g(w_i|\boldsymbol{\vartheta})}\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})d\boldsymbol{\vartheta}\right)^{-1}}{g(w_i|\boldsymbol{\vartheta})} = \frac{CPO_i}{g(w_i|\boldsymbol{\vartheta})}.
\end{aligned}$$

Que, substituindo em (6.4), tem-se

$$\begin{aligned}
D_\psi(P, P_{(-i)}) &= \int_{\vartheta\in\Theta}\psi\left(\frac{CPO_i}{g(w_i|\boldsymbol{\vartheta})}\right)\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})\,d\boldsymbol{\vartheta} \\
&= E_{\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}}\left[\psi\left(\frac{CPO_i}{g(w_i|\boldsymbol{\vartheta})}\right)\right].
\end{aligned} \tag{6.5}$$

Pela expressão (6.5) obtêm-se a divergência K-L dada por

$$\begin{aligned} D_{\text{K-L}}(P, P_{(-i)}) &= -E_{\boldsymbol{\vartheta}|\mathcal{D}}\left\{\log(CPO_i)\right\} + E_{\boldsymbol{\vartheta}|\mathcal{D}}\left\{\log\left[g(w_i|\boldsymbol{\vartheta})\right]\right\} \\ &= -\log(CPO_i) + E_{\boldsymbol{\vartheta}|\mathcal{D}}\left\{\log\left[g(w_i|\boldsymbol{\vartheta})\right]\right\}. \end{aligned} \tag{6.6}$$

E pode-se obter $D_\psi(P, P_{(-i)})$ de uma amostra da distribuição *posteriori* de $\boldsymbol{\vartheta}$ pelos métodos MCMC.

Considerando $\boldsymbol{\vartheta}^{(1)}, \ldots, \boldsymbol{\vartheta}^{(Q)}$ uma amostra de tamanho Q de $\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})$, a estimativa de Monte Carlo de $K(P, P_{(-i)})$ é dada por

$$\widehat{D_\psi}(P, P_{(-i)}) = \frac{1}{Q}\sum_{q=1}^{Q} \psi\left(\frac{\widehat{CPO}_i}{g(y_i|\boldsymbol{\vartheta}^{(q)})}\right). \tag{6.7}$$

De (6.7) uma estimativa de Monte Carlo para a divergência K-L é dada por

$$\widehat{D_{\text{K-L}}}(P, P_{(-i)}) = -\log(\widehat{CPO}_i) + \frac{1}{Q}\sum_{q=1}^{Q} \log\left[g(y_i|\boldsymbol{\vartheta}^{(q)})\right]. \tag{6.8}$$

O $D_\psi(P, P_{(-i)})$ pode ser interpretado como a divergência ψ do efeito na distribuição *a posteriori* conjunta de $\boldsymbol{\vartheta}$, ao deletar o i-*ésimo* caso do conjunto de dados.

Como apontado por Peng & Dey (1995) e Weiss (1996) (vide também Cancho *et al.*, 2010b, 2011), pode ser complicado determinar o ponto de corte da medida de divergência, de modo a determinar se um pequeno conjunto de dados é influente ou não. Consideramos como pontos influentes: $d_{L_1} > 0,3$, $d_{\chi^2} > 0,36$, $d_{\text{K-L}} > 0,22$ e $d_{\text{J}} > 0,41$; entretanto pode-se proceder de maneira sistemática conforme Peng & Dey (1995) e Weiss (1996), dado com segue:

Considere uma moeda não viciada, com probabilidade de sucesso p. A medida de divergência ψ entre uma moeda viciada e a não viciada é

$$D_\psi(f_0, f_1) = \int \psi\left(\frac{f_0(x)}{f_1(x)}\right) f_1(x)dx, \tag{6.9}$$

em que $f_0(x) = p^x(1-p)^{1-x}$ e $f_1(x) = 0,5$, $x = 0, 1$. Se $D_\psi(f_0, f_1) = d_\psi(p)$, pode ser verificado que d_ψ satisfaz a equação

$$d_\psi(p) = \frac{\psi(2p) + \psi(2(1-p))}{2}. \tag{6.10}$$

Não é difícil ver que as medidas de divergências consideradas aumentam conforme p se afasta do valor $0,5$. Além disso, $d_\psi(p)$ é simétrica em $p = 0,5$, que também é ponto de

mínimo, $d_{\psi}(0,5) = 0$ e $f_0 = f_1$. Portanto, se considerarmos $p > 0,75$ (ou $p \leq 0,25$) como um vício robusto na moeda, temos $d_{L_1}(0,75) = d_{\chi^2}(0,75) = 0,25$. Esta equação implica que o i-*ésimo* caso é considerado influente quando $d_{L_1}(0,75) > 0,25$ ou $d_{\chi^2} > 0,25$. Assim, utilizando a divergência de Kullback-Leibler, podemos considerar que uma observação é influente quando $d_{\text{K-L}} > 0,14$. De maneira similar, se usamos a *J-distance*, temos que uma observação é influente quando $d_{\text{J}} > 0,27$.

6.5 Estudo de simulação

O estudo de simulação no contexto bayesiano foi realizado com os objetivos:

a) primeiramente, avaliar as propriedades frequentistas das estimativas dos parâmetros dos modelos propostos, e

b) posteriormente, analisar o desempenho das medidas de diagnóstico propostas na Seção 6.4, considerando um conjunto de dados simulado, com um ou mais casos perturbados.

Para o caso *(a)*, um estudo de má especificação foi realizado, visando avaliar a performance dos ajustes dos modelos e a distinção entre eles - $LGGG_{mn}$, $LGGG_{mn}$ *e* $LGGG_{mx}$ -, à luz de um conjunto de dados e nos critérios descritos na Seção 6.3.

Para cumprir os propósitos *(a) e (b)*, consideramos que: para cada indivíduo i, $i = 1, 2, \cdots, n$, os tempos provêm da distribuição Gama Generalizada com parâmetros $\mu = 2$, $\sigma = 1$ *e* $\lambda = 1,2$ e o número de CR - M_i - provém da distribuição de probabilidade Geométrica com parâmetro $\theta_i = p_{0i}\Theta\left(\frac{-e^{-1/p_{0i}}}{p_{0i}}\right)$, em que $\Theta(\cdot)$ é a função de Lambert (Corless *et al.*, 1996). Assim o parâmetro da distribuição Geométrica para a geração de M_i é $p_{0i} = \frac{1}{(1+\exp(-\beta_0-\beta_1 x_i))}$. Consideramos uma covariável binária $Z = z$ com valores gerados da distribuição de $Bernoulli(0,5)$; fixamos $\beta_0 = -0,8$ e $\beta_1 = 0,5$, assim a fração de curados para os dois níveis da covariável Z é $p_0^{(0)} = 0,310$ e $p_0^{(1)} = 0,425$, respectivamente; e o tempo de censura para cada indivíduo é oriundo de uma distribuição Exponencial com parâmetro $0,03$.

6.5.1 Propriedades frequentistas

Temos por objetivo mostrar o bom comportamento das estimativas bayesianas, com base no erro quadrático médio (EQM), na média e nas medidas usadas para a comparação de modelos. Obtemos amostras de tamanho $n = 200$ para os três esquemas de ativação, resultando em três configurações para a simulação dos dados, e esse procedimento é repetido 500 vezes para cada um deles.

Em cada conjunto de dados, são ajustados os três modelos de longa duração propostos no capítulo anterior, considerando distribuições *a priori's* independentes especificadas por $\beta_j \sim N(0, 10^4)$ $j = 0, 1$, $\mu \sim N(0, 10^4)$, $\sigma \sim Gamma(1, 0.1)$ e $\lambda \sim N(0, 10^4)$, e o algoritmo Metropolis-Hasting. Após a obtenção da convergência, 25.000 valores gerados inicialmente são descartados como período de aquecimento (*burn-in*). Para reduzir a autocorrelação e obter melhores resultados na convergência, optamos por utilizar uma realização da amostra a cada 30 geradas. A convergência do algoritmo foi monitorada pelo método proposto por Geweke (1992) além de métodos gráficos.

Para cada amostra a posteriori, a média dos parâmetros e os valores do DIC, $EAIC$, $EBIC$ e $LPML$ são armazenados. Estatísticas resumo das simulações para os três modelos são dadas na Tabela 6.1 e Tabela 6.2. Nas tabelas, MC Média denota a média aritmética das 500 médias obtidas, dada por $\sum_{l=1}^{500} \hat{\vartheta}_{kl}/500$, MC EQM é o erro quadrático médio, dado por $\sum_{l=1}^{500}(\hat{\vartheta}_{kl} - \vartheta_k)^2/500$, em que ϑ_k é o verdadeiro valor do parâmetro estimado $\hat{\vartheta}_k$. Observe na Tabela 6.1 que as MCs Médias estão bem próximas dos verdadeiros valores e os MCs EQMs são pequenos quando o ajuste é realizado pelo modelo do mesmo esquema de ativação que gerou os dados.

Além disso, a Tabela 6.3 traz os percentuais de amostras em que a distribuição geradora foi indicada como a melhor opção de ajuste pelo critério LPML, cujos percentuais são similares aos dos critérios DIC, $EAIC$ e $EBIC$. Observa-se também que o modelo a partir do qual a amostra foi gerada tem percentagem mais elevada, exceto para o modelo de ativação aleatório, que é instável.

TABELA 6.1: Resultado baseado nas 500 simulações de Monte Carlo para os três esquemas de ativação. Média das médias (MC Média) e média dos EQM (MC EQM) das estimativas *a posteriori* para a fração de curados.

		Modelo ajustado					
Modelo	Fração de	Primeira ativação		Última ativação		Ativação aleatória	
verdadeiro	curados	MC Média	MC EQM	MC Média	MC EQM	MC Mean	MC EQM
$LGGG_{mn}$	$p_0^{(0)}$	0,3134	0,0022	0,3340	0,0027	0,2713	0,0096
	$p_0^{(1)}$	0,4286	0,0024	0,3841	0,0066	0,3969	0,0107
$LGGG_{mx}$	$p_0^{(0)}$	0,3632	0,0054	0,2880	0,0045	0,1418	0,0107
	$p_0^{(1)}$	0,3978	0,0041	0,4034	0,0061	0,2585	0,0149
$LGGG_{al}$	$p_0^{(0)}$	0,3394	0,0028	0,3095	0,0025	0,2749	0,0104
	$p_0^{(1)}$	0,4154	0,0030	0,3989	0,0052	0,3873	0,0149

TABELA 6.2: Resultado baseado nas 500 simulações de Monte Carlo para os três esquemas de ativação. Média das médias (MC Média) e média dos EQM (MC EQM) das estimativas *a posteriori* dos parâmetros dos modelos.

		Modelo ajustado					
Modelo	Fração de	Primeira ativação		Última ativação		Ativação aleatória	
verdadeiro	curados	MC Média	MC EQM	MC Média	MC EQM	MC Mean	MC EQM
	μ	1,0123	0,0448	-0,1279	4,6576	1,0123	1,0395
	σ	1,5269	0,0386	1,6906	0,4972	1,5269	0,4104
$LGGG_{mn}$	λ	0,6217	0,0986	0,4993	0,6291	0,6217	0,7000
	β_0	-2,7248	0,0512	-0,7071	0,0564	-2,7248	46,1135
	β_1	1,2948	0,0758	0,2044	0,2310	1,2948	31,2000
	μ	2,5368	1,0789	1,7293	0,1605	2,5368	0,3105
	σ	0,9049	0,1558	1,1025	0,0260	0,9049	0,0994
$LGGG_{mx}$	λ	1,0384	0,0868	0,8581	0,2786	1,0384	0,3333
	β_0	-3,3195	0,1027	-0,9571	0,1499	-3,3195	96,8260
	β_1	1,2529	0,2015	0,5327	0,1789	1,2529	75,0917
	μ	1,9067	0,3821	0,9426	1,1883	1,9067	0,0439
	σ	1,2025	0,0528	1,3998	0,1760	1,2025	0,1778
$LGGG_{al}$	λ	0,8557	0,0737	0,7146	0,3503	0,8557	0,5007
	β_0	-2,8022	0,0562	-0,8006	0,0571	-2,8022	40,3028
	β_1	0,9245	0,0901	0,3656	0,1364	0,9245	31,1166

6.5.2 Influência das observações discrepantes

O intuito é examinar o comportamento das medidas de diagnóstico apresentadas no modelo $LGGG_{mn}$ perante perturbação de observações censuradas e não censuradas.

TABELA 6.3: Porcentagem de amostras em que o modelo original foi indicado como a melhor opção de ajuste entre os comparados, de acordo com o critério $LPML$.

	Modelo ajustado		
Modelo verdadeiro	Primeira ativação	Última ativação	Ativação aleatória
$LGGG_{mn}$	75,2	15,2	9,6
$LGGG_{mx}$	22,4	60,0	17,6
$LGGG_{al}$	40,4	27,2	32,4

Para tanto, consideramos uma amostra simulada de tamanho 200, na qual os tempos $W_i = w_i$, $i = 1, 2, \cdots, 200$, variam de $w_{(1)} = 0,005$ à $w_{(200)} = 134,631$, com mediana $\widetilde{w} = 3,433$, média $\overline{w} = 12,608$ e desvio padrão $s = 21,717$. As observações selecionadas foram:

- Não censuradas: $w_{67} = 10,967$; $w_{107} = 0,106$ e $w_{197} = 15,766$;
- Censuradas: $w_{51} = 9,499$; $w_{101} = 3,648$ e $w_{151} = 10,697$.

A perturbação foi introduzida fazendo-se $w_i' = w_i + 4s$, em que s é o desvio padrão do conjunto de dados. Para criar a influência na amostra perturbamos inicialmente uma observação e prosseguimos até a perturbação de todas. Especificamos alguns conjuntos de perturbação para analisarmos. A seguir apresentamos a denominação do conjunto e observações perturbadas: A: nenhuma observação; - B: w_{67}; - C: w_{107}; - D: w_{197}; - E: w_{107} e w_{197}; - F: w_{51} , w_{101} e w_{151}.

Ajustamos os modelos $LGGG_{mx}$ e $LGGG_{al}$, além do modelo do qual obteve-se os dados - $LGGG_{mn}$. A parte computacional do MCMC é tal qual apresentada na Seção 6.5.1, bem como a verificação de convergência.

A influência da perturbação de dados censurados na fração de curados é praticamente nula (Tabela 6.4). A perturbação, seja de uma ou mais observações não censuradas da amostra, interfere na estimação da proporção de curados, algumas mais que outras.

A perturbação de dados censurados foi a que menos afetou as estimativas. Isso já era esperado, pois presume-se que, sendo o dado censurado há possibilidade de seu valor ser infinito, e esta propriedade foi observada nas simulações.

TABELA 6.4: Média e desvio padrão (SD) para a fração de cura estimada para o conjunto de dados considerando os modelos $LGGG$ para a primeira, última e ativação aleatória, em confronto com perturbação nos dados.

	$LGGG_{mn}$				$LGGG_{mx}$				$LGGG_{al}$			
	$p_0^{(0)}$		$p_0^{(1)}$		$p_0^{(0)}$		$p_0^{(1)}$		$p_0^{(0)}$		$p_0^{(1)}$	
Conjunto	Média	SD	Média	SD	Média	SD	Média	SD	Média	SD	Média	SD
A	0,325	0,044	0,387	0,046	0,329	0,049	0,360	0,048	0,305	0,052	0,380	0,051
B	0,296	0,048	0,355	0,050	0,306	0,050	0,339	0,052	0,278	0,057	0,353	0,057
C	0,303	0,048	0,350	0,051	0,299	0,049	0,345	0,051	0,280	0,055	0,358	0,050
D	0,298	0,048	0,353	0,052	0,307	0,050	0,341	0,051	0,267	0,084	0,340	0,082
E	0,286	0,048	0,332	0,051	0,285	0,053	0,330	0,054	0,205	0,011	0,278	0,125
F	0,331	0,045	0,388	0,046	0,331	0,049	0,368	0,047	0,310	0,151	0,382	0,050

Em contrapartida, quando perturbamos três dados houve problema na convergência do modelo $LGGG_{al}$ e para os demais a estimativa dos percentuais foi de 0,27 e 0,31 para p_0^1 e p_0^0, respectivamente, sendo o caso que apresentou as maiores diferenças entre verdadeiro e estimado.

Na Tabela 6.5 estão reportadas as estimativas da simulação de Monte Carlo dos critérios DIC, $EAIC$, $EBIC$ e $LPML$ para cada um dos conjuntos analisados.

TABELA 6.5: Comparação do ajuste dos modelos $LGGG_{mn}$, $LGGG_{al}$ e $LGGG_{mx}$ para os seis conjuntos considerados, utilizando os critérios bayesianos DIC, EAIC, EBIC e LPML.

	Primeira				Última				Aleatória			
Conjunto	DIC	EAIC	EBIC	LPML	DIC	EAIC	EBIC	LPML	DIC	EAIC	EBIC	LPML
A	736,5	742,2	758.7	-368,2	739,8	744,8	761,2	-369,7	738,1	743,2	759,6	-369,0
B	757,5	762,8	779.3	-379,9	755,5	760,6	777,1	-378,8	757,4	762,4	778,9	-380,0
C	764,5	769,6	786.1	-383,7	761,6	766,7	783,2	-382,0	763,4	768,5	785,0	-383,0
D	756,9	762,1	778.6	-380,4	754,2	759,3	775,8	-378,6	754,1	762,8	778,7	-382,2
E	773,4	778,7	795.2	-387,3	772,5	777,6	794,1	-387,2	759,0	781,5	798,0	-390,0
F	737,7	743,4	759.9	-368,7	740,9	746,0	762,4	-370,3	739,3	744,3	760,8	-369,5

Observa-se que o conjunto que apresenta os melhores critérios de ajuste é o A - sem observações perturbadas - e, dentre os demais, o que tem seus critérios mais próximos ao de A é o conjunto F - perturbação das observações censuradas -, indicando que valores elevados para dados censurados não interferem ou pouco interferem na qualidade de ajuste.

Consideramos a amostra *a posteriori* dos parâmetros da $LGGG_{mn}$ e calculamos as medidas de divergência ψ relatadas na Seção 6.4. Os resultados (Tabela 6.6) retratam

que os pontos w_{67}, w_{107} e w_{197} não são considerados influentes antes de perturbação (conjunto A), todavia, ao serem perturbados, (conjuntos B, C, D e F), suas medidas de divergência se elevam, tornando-os influentes pelos critérios considerados. As observações w_{51}, w_{101} e w_{151} que são censuradas (conjunto F) não apresentam valores de divergência que indiquem a influência mesmo após a perturbação, corroborando com o já afirmado em relação à qualidade de ajuste e estimação da fração de curados.

TABELA 6.6: Medida de divergência ψ para os dados simulados, considerando o ajuste do modelo $LGGG_{mn}$.

	Conjuntos / observação analisada										
Medida de	A			B	C	D	E		F		
Divergência	w_{67}	w_{107}	w_{197}	w_{67}	w_{107}	w_{197}	w_{107}	w_{197}	w_{51}	w_{101}	w_{151}
$d_{\text{K-L}}$	0,0256	0,0206	0,1243	1,4145	1,8332	2,2614	0,3720	0,5003	0,0073	0,0073	0,0097
d_{J}	0,0519	0,0416	0,2623	3,0177	4,1853	5,8155	0.8074	1.0938	0,0146	0,0146	0,0196
d_{L_1}	0,0907	0,0802	0,2008	0,6442	0,7194	0,7822	0.3557	0.4120	0,0472	0,0472	0,0554
d_{χ^2}	0,0558	0,0440	0,3556	14,0920	39,7656	188,481	1,5048	2,3686	0,0150	0,0150	0,0202

Nas Figuras 6.1 e 6.2 estão graficadas as quatro medidas de divergência para os conjuntos A e B, respectivamente. Claramente, podemos ver que todas as medidas apresentam um bom desempenho para a identificação dos casos influentes (Figura 6.2), fornecendo maiores valores de ψ.

6.6 Dados de melanoma maligno

Ajustamos as três distribuições $LGGGcr$ ao conjunto de dados $T12$ utilizando distribuições *a priori's* independentes: $\gamma_j \sim N(0, 10^4)$ $j = 0, 1, 2$, $\mu \sim N(0, 10^4)$, $\sigma \sim G(1, 1)$ e $\lambda \sim N(0, 1)$ e o algoritmo de Metropolis-Hasting. Foram geradas 200.000 amostras de MCMC da distribuição *a posteriori* e descartadas 50% delas como *burn-in*, adotou-se salto de tamanho 20 para reduzir a autocorrelação e obter melhores resultados de convergência, resultando em uma amostra de tamanho 5.000.

A comparação de qualidade de ajuste dos modelos foi realizada via critérios DIC, $EAIC$, $EBIC$ e $LPML$, disponíveis na Tabela 6.7, que indicam o modelo $LGGG_{mn}$ como o que apresenta o melhor ajuste. Adotamos esse modelo para as futuras análises nesse conjunto de dados.

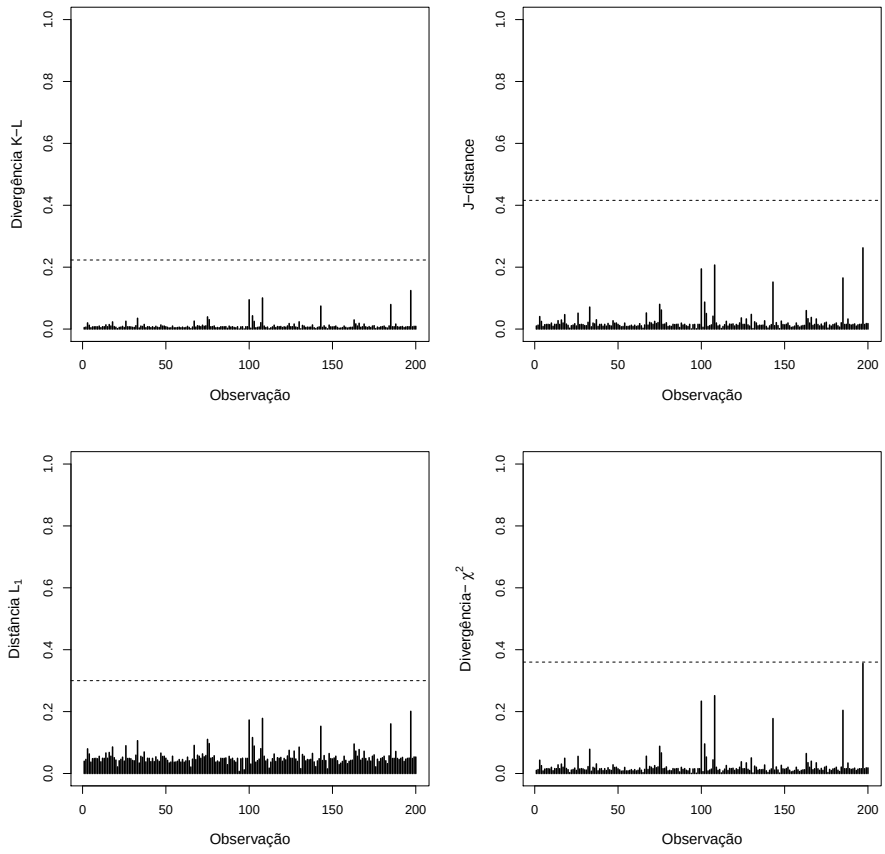

FIGURA 6.1: Medida de divergência ψ para os dados simulados sem nenhuma perturbação - conjunto A.

TABELA 6.7: Critérios bayesiano de avaliação de ajustes.

	Critérios			
Distribuição	LPLM	DIC	EAIC	EBIC
$LGGG_{mn}$	-211,09	420,97	428,50	448,44
$LGGG_{mx}$	-222,34	445,27	451,25	471,19
$LGGG_{al}$	-217,70	432,98	441,07	461,01

Os valores da média, mediana, desvio padrão e intervalo de 95% HPD *a posteriori* estão relatados na Tabela 6.8 e dão indicativos que as covariáveis possuem efeito na redução da fração de curados.

Considerando a amostra da distribuição *a posteriori* para os parâmetros do modelo $LGGG_{mn}$, as medidas de divergência ψ foram calculadas. A Figura 6.3 dispõem o gráfico das quatro medidas de divergência e observa-se que os casos 5 e 171 (w_5 e w_{171}) são

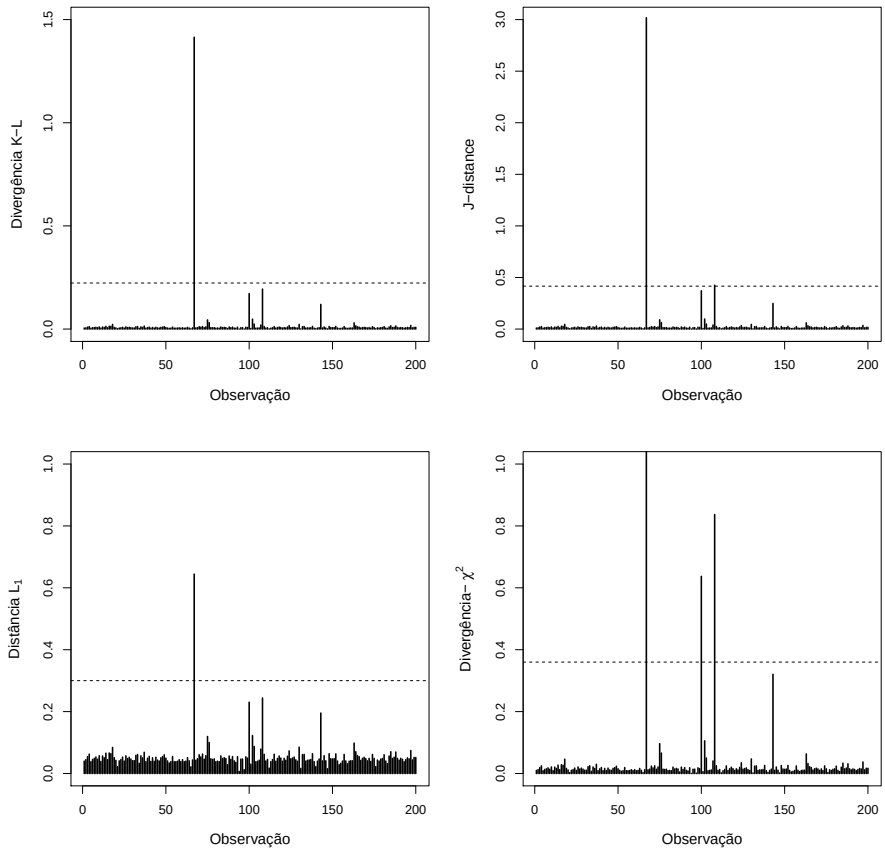

FIGURA 6.2: Medida de divergência ψ para os dados simulados, com a perturbação da observação w_{67} - conjunto B.

TABELA 6.8: Resumo estatístico da amostra *a posteriori* referente ao modelo $LGGG_{mn}$.

				Intervalo HPD (95%)	
Parâmetro	média	Mediana	desvio padrão	LI	LS
μ	2,025	1,951	0,346	1,530	2,899
σ	0,780	0,694	0,395	0,261	1,790
λ	0,595	0,592	0,941	-1,385	2,542
$\beta_{\text{intercepto}}$	1,546	1,598	0,499	0,418	2,411
$\beta_{\text{ulceração}}$	-1,491	-1,598	0,365	-2,190	-0,786
$\beta_{\text{espessura}}$	-0,175	-0,175	0,053	-0,279	-0,068

possíveis candidatos a observações influentes na distribuição *a posteriori*. O procedimento é retirar essas observações e avaliar o impacto da retirada das mesmas nas estimativas, utilizando a medida de alteração relativa (AR), em que a AR (em porcentagem) de cada

parâmetro estimado é definida por

$$AR_{\vartheta_j} = \left| \frac{(\widehat{\vartheta}_j - \widehat{\vartheta}_{j(I)})}{\widehat{\vartheta}_j} \right| \times 100\%,$$

sendo que $\widehat{\vartheta}_{j(I)}$ denota a média *a posteriori* de ϑ_j, com $j = 1, ..., 5$, após o conjunto $I = \{5, 171\}$ de observações ser removido.

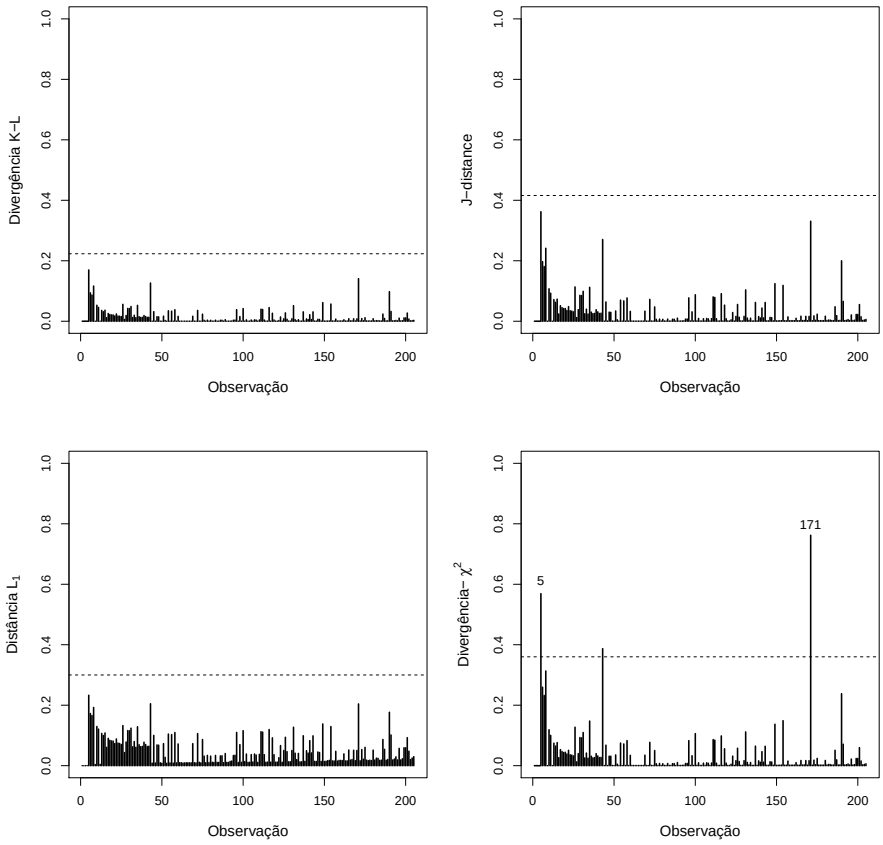

FIGURA 6.3: Gráficos das medidas de divergência ψ para o conjunto de dados $T12$.

A Tabela 6.9 apresenta a AR sofrida pelas estimativas com a retirada de um dos pontos possivelmente influentes e de ambos ao mesmo tempo. O intervalo HPD de 95% para as novas estimativas estão nos parênteses e notamos que há pequenas mudanças nas inferências para os coeficientes. Particularmente, λ é não significativo a 5%, nesses casos considerados tal qual no original.

Voltamos nossa atenção para a importância das covariáveis na fração de curados p_0. A Tabela 6.10 apresenta o resumo da amostra *a posteriori* para a fração de curados estratificada pelo tamanho do tumor, sendo os tamanhos considerados iguais a 0,64, 1,94 e 6,63 mm, os quais correspondem aos percentis 10, 50 e 90.

TABELA 6.9: ARs (em %), intervalo HPD de 95% e critério de comparação de ajustes mediante retirada de ponto influentes, modelo $LGGG_{mn}$.

	Observação retirada					
	w_5		w_{171}		w_5 e w_{171}	
Parâmetro	AR	HPD	AR	HPD	AR	HPD
μ	3,672	(1,498; 2,709)	7,409	(1,454; 2,624)	7,210	(1,453; 2,647)
σ	2,195	(0,270; 1,608)	16,597	(0,262; 1,473)	10,374	(0,264; 1,617)
λ	13,039	(-1,285; 2,463)	44,062	(-0,885;2,458)	9,839	(-1,128; 2,315)
γ_1	0,738	(0,504; 2,403)	12,424	(0,728; 2,499)	6,014	(0,597; 2,454)
γ_2	0,507	(-2,227; -0,806)	0,395	(-2,191; -0,804)	0,121	(-2,218; -0,804)
γ_3	9,185	(-0,263; -0,052)	1,714	(-0,283; -0,077)	10,106	(-0,267; -0,048)
DIC		417,57		411,51		408,21
LPML		-209,10		-206,33		-204,81

TABELA 6.10: Resumo das medidas *a posteriori* da fração de curados estratificadas pelo tamanho do tumor.

Dados	Tamanho do tumor	Câncer	Média	Mediana	Desvio Padrão	Intervalo 95% HPD
Completo (D)	0,64	Ausente	0,797	0,815	0,082	(0,575, 0,906)
		Presente	0,488	0,501	0,115	(0,233, 0,682)
	1,94	Ausente	0,759	0,779	0,090	(0,525, 0,882)
		Presente	0,434	0,445	0,107	(0,202, 0,615)
	6,63	Ausente	0,590	0,605	0,120	(0,323, 0,786)
		Presente	0,259	0,259	0,083	(0,102, 0,423)
$D - \{5, 171\}$	0,64	Ausente	0,814	0,830	0,072	(0,622, 0,911)
		Presente	0,513	0,528	0,107	(0,269, 0,692)
	1,94	Ausente	0,783	0,799	0,078	(0,578, 0,891)
		Presente	0,464	0,478	0,100	(0,234, 0,629)
	6,63	Ausente	0,647	0,660	0,108	(0,390, 0,820)
		Presente	0,307	0,308	0,084	(0,131, 0,463)

Claramente a retirada das observações influentes - w_5 e w_{171} - conduz ao aumento da fração de curados. Por este aspecto, pontos influentes podem fornecer resultados equivocados quanto à proporção de curados na população e a identificação de ponto influente pode ser feita via metodologia aqui apresentada.

As estimativas pontuais apresentadas para dos parâmetros do modelo LGGG diferem dos valores que remetem o modelo a um de seus casos particulares. Pelos histogramas das amostras da distribuição *a posteriori* dos parâmetros λ e σ (Figura 6.4) percebemos que os valores dos parâmetros que particularizam o o modelo utilizado em alguns casos estão contidos no intervalo de valores possíveis para o parâmetro, mas os valores da amostra não se limitam aos valores que conduzem aos casos particulares, possibilitando dúvidas entre quais dos casos particulares seria o mais adequado. Mediante tal impasse, julgamos que o modelo geral LGGG é o mais adequado para os ajustes.

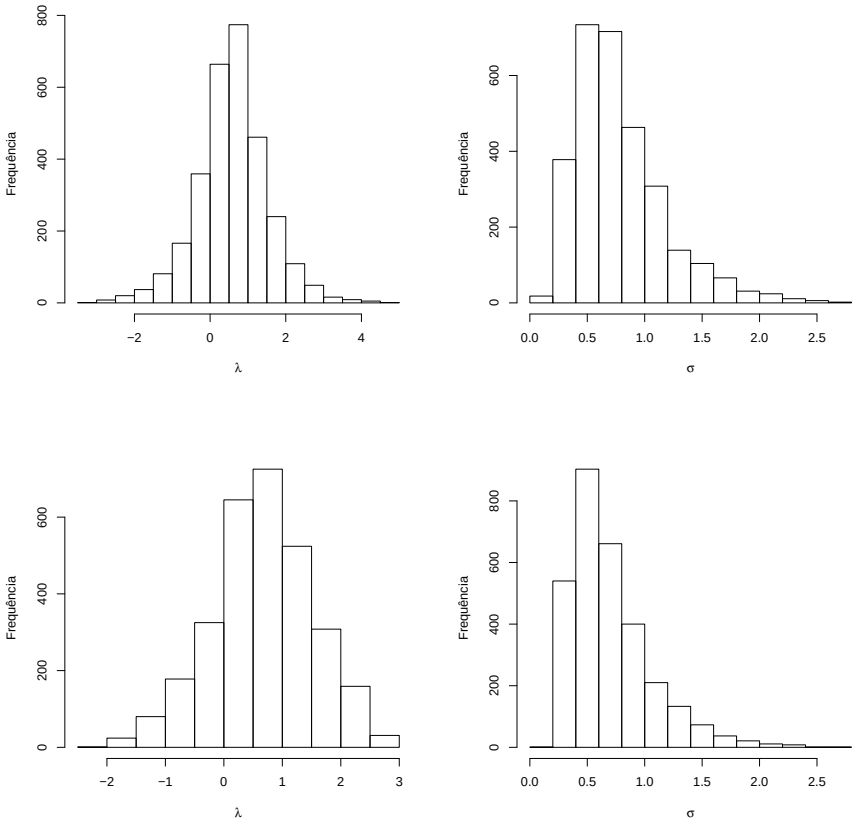

FIGURA 6.4: Histogramas das amostras da distribuição *a posteriori* para os parâmetros λ e σ, respectivamente, para o conjunto completo D (superior) e para o conjunto D-{5,171} (inferior).

As funções densidades marginais *a posteriori* aproximadas pelo histograma, considerando as 3000 observações amostrais geradas, em que pode-se visualizar a convergência das cadeias e a sequência dessas amostras, tanto para o conjunto D quanto para o conjunto $D - \{5, 171\}$ estão disponíveis nas Figuras A.7 e A.8 do Apêndice.

6.7 Comentários

O método de diagnóstico de influência sob o ponto de vista bayesiano, com base na divergência ψ, que considera a distribuição *a posteriori* dos parâmetros do modelo proposto, mostra-se eficaz para identificar perturbações em observações da amostra, e verificamos que essas perturbações influenciam na qualidade de ajuste bem como no valor da percentual de curados. Ele tem vários casos particulares, e assim acreditamos que a ferramenta bayesiana desenvolvida para detectar casos influentes nas distribuilções *LGGGcr* é uma significante contribuição.

Na aplicação ao conjunto de dados de melanoma, identificamos o modelo $LGGG_{mn}$ como o melhor ajuste. Observou-se que a probabilidade de convergência diminui rapidamente em pacientes com tumores mais espessos, e que a fração de curados é menor para os pacientes com ulceração. A interpretação do papel da covariável foi facilitada pela parametrização utilizada - na fração de cura. A distribuição *a priori* utilizada para λ possui caráter informativo e contempla o espaço paramétrico dos principais casos particulares do modelo LGGG. Ao considerar uma variância maior, a convergência também foi obtida com valor médio da distribuição marginal *a posteriori* bem próximo do apresentado, porém com grande oneração de tempo computacional.

Capítulo 7

Comentários finais

Neste capítulo apresentamos uma síntese dos modelos apresentados bem como uma estrutura geral do conceito utilizado e perspectivas para novas distribuições utilizando a mesma metodologia tanto quanto com a inclusão de novos elementos na modelagem.

7.1 Modelos desenvolvidos

O Capítulo 2 apresenta a formulação geral considerando a distribuição de probabilidade Geométrica tanto para a população dos suscetíveis quanto para a que contempla também os indivíduos imunes da população, nos diferentes esquemas de ativação. Comparações entre as funções de sobrevivências para os diferentes esquemas de ativação são estabelecidas.

No Capítulo 3 apresentamos a distribuição inédita EG_{mx} denominada de CEG no artigo intitulado *The Complementary Exponential Geometric Distribution: Model, Properties, and a Comparison with its Counterpart*, publicado em 2011 na revista *Computational Statistics & Data Analysis* (Louzada *et al.*, 2011). Sua concepção está baseada na ativação pelo máximo dos tempos e tendo a distribuição de probabilidade Exponencial para os tempos de ativação de cada uma das CRs. Apresentamos sua aplicabilidade em conjuntos de dados da engenharia e da área da saúde. Comparamos a distribuição EG_{mx} em termos de comportamento das funções de risco, densidade e sobrevivência e

também quanto ao ajustes em conjuntos de dados, no contexto da inferência clássica, com a distribuição EG_{mn}. A distribuição EG_{mn} considera a ativação do mínimo dos tempos, ela também é apresentada no capítulo, porém foi proposta por Adamidis & Loukas (1998).

No Capítulo 4 agregamos a informação de longa duração às distribuições EG_{mx} e EG_{mn}. Utilizando a metodologia de Berkson e Gage (1952) que resultou nas distribuições LEG_{mx} e LEG_{mn}, respectivamente para a ativação do máximo e do mínimo dos tempos. Comparações das funções que descrevem esses modelos foram apresentadas. Dados de câncer foram bem acomodados pelo modelo LEG_{mn}, cujos ajustes foram comparados aos obtidos pelos modelos de longa duração BG das distribuições Exponencial e Weibull. O modelo LEG_{mx} foi ajustado para dados de câncer e finanças. Todos os ajustes apresentaram resultados satisfatórios e pelos critérios de comparação de ajustes considerados, o LEG_{mx} prevaleceu aos modelos de longa duração BG das distribuições Exponencial, Weibull e Log-Logístico. Consideramos a abordagem clássica nos ajustes para as duas distribuições propostas. Os resultados deste capítulo estão sumarizados nos artigos publicados em 2012 *A new long term lifetime distribution induced by a latent complementary risk framework* pela revista *Journal of Applied Statistics* (Louzada *et al.*, 2012) e *A new long-term survival distribution for cancer data* na revista *Journal of Data Science* (Roman *et al.*, 2012), com a ressalva que no segundo artigo consideramos a distribuição Geométrica com probabilidade de sucesso dada por $1-\theta$ ao invés de θ como nesta tese, cuja diferenças em termos inferenciais se restringe à obtenção de valores complementares para tal parâmetro.

As distribuições GGG_{mx}, $LGGG_{mn}$, $LGGG_{al}$ e $LGGG_{mx}$ são as inovações apresentadas no Capítulo 5. Suas formulações consideram que os tempos de ativação de cada uma das CRs segue a distribuição Gama Generalizada de três parâmetros (Stacy, 1962; Farewell & Prentice, 1977) e diferentes formas de ativação do evento de interesse. Apresentamos comparações entre a distribuição GGG_{mx} obtida considerando o cenário de ativação pelo máximo dos tempos com a distribuição GGG_{mn} que representa o cenário de ativação pelo mínimo (proposta com outra parametrização por Ortega *et al.* (2011)). As distribuições $LGGG_{mn}$, $LGGG_{al}$ e $LGGG_{mx}$ são distribuições de longa duração obtidas por um mecanismo distinto à metodologia BG mas que possui uma relação com o mesmo, sem, todavia, ter um parâmetro distinto para representar a proporção de curados. Comportamento das funções de sobrevivência e comparações entre elas são apresentadas.

Considerando a inferência clássica, os modelos foram ajustados a conjuntos de dados de câncer e comparados com o ajuste da distribuição de longa duração BG para a distribuição Weibull. Os critérios de comparações de ajustes são favoráveis aos modelos por nós propostos.

A inferência bayesiana foi utilizada no Capítulo 6. Considerando as distribuições $LGGG_{mn}$, $LGGG_{al}$ e $LGGG_{mx}$ simulações para avaliar propriedades de ajustes e influência de observações discrepantes são realizadas, proporcionando resultados satisfatórios. Ajustes dos modelos à dados câncer e análise de influência de observações perturbadas na proporção de cura também são apresentadas.

7.2 Perspectivas futuras

Focamos a tese na construção de modelos de sobrevivência considerando diferentes forma de ativação do evento de interesse; de forma geral, considerando p_m a função massa de probabilidade da v.a. M que denota a quantidade de CR e $f_o(t_i)$ a função densidade de probabilidade da v.a. T_i que corresponde ao tempo até a ocorrência da *i-ésima* CR, sendo M desconhecido. A função densidade de probabilidade da v.a. $X = \min\{T_1, T_2, \cdots, T_M\}$ e $Y = \max\{T_1, T_2, \cdots, T_M\}$ são dadas respectivamente por

$$f(x) = \frac{f_o(x)}{S_o(x)} \sum_{m=1}^{\infty} m[S_o(x)]^m p_m \qquad \text{e} \qquad f(y) = \frac{f_o(y)}{F_o(y)} \sum_{m=1}^{\infty} m[F_o(y)]^m p_m, \tag{7.1}$$

em que $\sum_{m=1}^{\infty} p_m = 1 - p_o$ com $p_o = P(M = 0)$ e $\sum_{m=1}^{\infty} p_m = 1$, respectivamente para o modelo de longa duração (grupo dos imunes e suscetíveis) e para o modelo sem longa duração (grupo dos suscetíveis).

Nosso objetivo voltou-se para a função massa de probabilidade p_m correspondente a distribuição de probabilidade Geométrica e duas opções para $f_o(t_i)$, que resultaram em sete distribuições inéditas, nominalmente, EG_{mx}, GGG_{mx}, LEG_{mx}, LEG_{mn},$LGGG_{mx}$, $LGGG_{mn}$ e $LGGG_{al}$ e as últimas cinco acomodando dados de longa duração.

Todavia, com outras especificações para p_m e $f_o(t_i)$ são obtidas outras distribuições de probabilidade. Por exemplo, sendo p_m correspondente à distribuição de probabilidade Poisson truncada no zero com parâmetro ϕ, a f.d.p. para o mínimo e o máximo

dos tempos, no contexto sem longa duração é

$$f(x) = \frac{\phi f_o(x) e^{-\phi F_o(x)}}{1 - e^{-\phi}} \qquad \text{e} \qquad f(y) = \frac{\phi f_o(y) e^{\phi S_o(y)}}{1 - e^{-\phi}}. \tag{7.2}$$

E para o mínimo dos tempos tem-se função distribuição acumulada, função de sobrevivência e função de risco, respectivamente

$$F(x) = \frac{\phi \left[1 - e^{-\phi F_o(x)}\right]}{1 - e^{-\phi}}, \;\; S(x) = 1 - \frac{\phi \left[1 - e^{-\phi F_o(x)}\right]}{1 - e^{-\phi}} \;\text{ e }\; h(x) = \frac{\phi f_o(x) e^{-\phi F_o(x)}}{1 - e^{-\phi} - \phi \left[1 - e^{-\phi F_o(x)}\right]}.$$

Para o máximo dos tempos tem-se função distribuição acumulada, função de sobrevivência e função de risco

$$F(y) = \frac{e^{-\phi S_o(y)} - e^{-\phi}}{1 - e^{-\phi}}, \;\; S(y) = \frac{1 - e^{-\phi S_o(y)}}{1 - e^{-\phi}} \;\text{ e }\; h(y) = \frac{\phi f_o(y) e^{-\phi S_o(y)}}{1 - e^{-\phi S_o(y)}}.$$

Quando considerada p_m correspondente a distribuição de Poisson com suporte em $\{0, 1, 2, \cdots\}$ temos os modelos de longa duração cujas f.d.p. para a ativação pelo mínimo e máximo dos tempos correspondem a

$$f_{pop}(x) = \phi f_o(x) e^{-\phi F_o(x)} \qquad \text{e} \qquad f_{pop}(y) = \phi f_o(y) e^{\phi S_o(y)}. \tag{7.3}$$

Para o mínimo dos tempos tem-se função de sobrevivência e função de risco dadas por

$$S_{pop}(x) = e^{-\phi F_o(x)} \;\text{ e }\; h_{pop}(x) = \phi f_o(x),$$

respectivamente. Enquanto que para o máximo dos tempos a função de sobrevivência e função de risco são, respectivamente,

$$S_{pop}(y) = 1 + e^{-\phi} e^{-\phi S_o(y)} \;\text{ e }\; h_{pop}(y) = \frac{\phi f_o(y) e^{-\phi S_o(y)}}{1 + e^{-\phi} e^{-\phi S_o(y)}}.$$

As misturas de distribuições apresentadas na (7.1), ou então com p_m específico tal qual apresentado no Capítulo 2 e nas expressões das f.d.p. dadas em (7.2) e (7.3), permitem obter várias distribuições de probabilidade para os tempos observados W, utilizando outras distribuições de probabilidade para os tempos T_i. Várias distribuições nesse enfoque já foram obtidas, mas as opções ainda não foram esgotadas.

As possibilidades não se limitam na obtenção de novas distribuições, a inclusão de estruturas de correlação entre as CR podem ser consideradas bem como uma forma

de tratar eventos recorrentes. Possíveis vantagens com a exponenciação da função de sobrevivência ou função distribuição acumulada podem ser avaliadas. Também a influência local exercida pela perturbação de observações da amostra aleatória no contexto da inferência clássica é algo pendente e possível de ser explorado.

No contexto de ajuste bayesiano a possível identificação da forma das distribuições condicionais completas que permitiriam a utilização do método de Metropolis-Hastings com passos de Gibbs, ou da log-concavidade das mesmas o que permitiria a utilização do método de rejeição adaptativo devem ser considerados futuramente. Além disso, verificar se a distribuição composta permite a utilização de distribuições *a prioris* impróprias no parâmetro de proporção de curados também deve ser considerada em estudos vindouros.

Nosso enfoque foi totalmente paramétrico. Entretanto, considerar estruturas não paramétricas é uma possibilidade e que pode ser considerada futuramente. Tem-se, por exemplo, o modelo modelo exponencial por partes que segundo Ibrahim *et al.* (2001), parte de sua importância se deve ao fato de acomodar funções de taxa de falha com diversas formas, não havendo a necessidade de impor restrições quanto à forma da função de risco para obtermos um ajuste apropriado do modelo aos dados.

Referências

Aarset, M. V. (1985). The null distribution for a test of constant versus "bathtub" failure rate. *Scandinavian Journal of Statistics*, **12**(1), 55–61.

Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, **2**(1), 106–108.

Abd El-Fatah, I. M., Mead, M. E. & Semary, H. E. (2011). The Applications of the Modified Generalized Gamma Distribution in Inventory Control. *International Journal of Contemporary Mathematical Sciences*, **6**(35), 1699–1712.

Abramowitz, M. & Stegun, I. (1972). *Handbook of Mathematical Functions*. Dover Publications, New York.

Adamidis, K. & Loukas, S. (1998). A Lifetime Distribution with Decreasing Failure Rate. *Statistics & Probability Letters*, **39**, 35–42.

Allenby, G. M., Leone, R. P. & Jen, L. (1999). A Dynamic Model of Purchase Timing with Application to Direct Marketing. *Journal of the American Statistical Association*, **94**(446), 365–374.

Allison, P. D. (1995). *Survival Analysis Using SAS: A Pratical Guide*. SAS Institute Inc., Cary, North Carolina.

Barreto-Souza, W., Morais, A. L. & Cordeiro, G. M. (2010). The Weibull-geometric distribution. *Journal of Statistical Computation and Simulation*, **81**(5), 645–657.

Berkson, J. & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, **47**(259), 501–515.

Boag, J. W. (1949). Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society*, **11**(1), 15–53.

Brasil (2011). *Estimativa 2012: Incidência de câncer no Brasil*. Ministério da Saúde e Instituto Nacional do Câncer José Alencar Gomes da Silva.

Brooks, S. P. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde (2002). *Journal of the Royal Statistical Society B*, **64**, 616–618.

Cancho, V. G. & Bolfarine, H. (2001). Modeling the presence of imnunes by using the Exponentiated-Weibull model. *Journal of Applied Statistical Science*, **28**(6), 659–671.

Cancho, V. G., Ortega, E. M. M. & Bolfarine, H. (2009). The log-exponentiated-Weibull regression models with cure rate: Local influence and residual analysis. *Journal of Data Science (Print)*, **7**, 433–458.

Cancho, V. G., Louzada, F. & Barriga, G. D. C. (2010a). The Poisson-exponential lifetime distribution. *Computational Statistics & Data Analysis*, **55**, 677–686. doi:10.1016/j.csda.2010.05.033.

Cancho, V. G., Ortega, E. M. M. & Paula, G. A. (2010b). On estimation and influence diagnostics for log-Birnbaum-Saunders student-t regression models: Full Bayesian analysis. *Journal of Statistical Planning and Inference (Print)*, **140**(9), 2486–2496.

Cancho, V. G., Dey, D. K., Lachos, V. H. & Andrade, M. G. (2011). Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. *Computational Statistics & Data Analysis*, **55**(1), 588–602.

Cancho, V. G., Louzada, F. & Barriga, G. D. C. (2012). The geometric Birnbaum–Saunders regression model with cure rate. *Journal of Statistical Planning and Inference*, **142**, 993–1000.

Carlin, B. P. & Louis, T. A. (2001). *Bayes and Empirical Bayes Methods for Data Analysis*. Test in Statistical Series. Chapman & Hall/CRC, New York, second edition.

Carvalho, A. P., Dutra, L. C. & Tonelli, E. (2003). Vacinação contra influenza em crianças infectadas pelo hiv: alterações imunológicas e na carga viral. *Jornal de Pediatria*, **79**(1), 29–40.

Chahkandi, M. & Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate. *Computational Statistics & Data Analysis*, **53**(12), 4433–4440.

Chen, M. H. & Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, **57**(1), 43–52.

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B (Methodological)*, **48**(2), 133–169.

Cook, R. D. & Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman & Hall, New York.

Cooner, F., Banerjee, S. & McBean, A. M. (2006). Modelling geographically referenced survival data with a cure fraction. *Statistical Methods in Medical Research*, **15**(4), 307–324.

Cooner, F., Banerjee, S., Carlin, B. P. & Sinha, D. (2007). Flexible cure rate modeling under latent activation schemes. *Journal of the American Statistical Association*, **102**(478), 560–572.

Cordeiro, G, M., Ortega, E. M. M. & Silva, G. O. (2011). The exponential generalized gamma distribution with application to lifetime data. *Journal of Statistical Computational and Simulation*, **81**(7), 827–842.

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. J. & Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics*, **5**(1), 329–359.

Cox, C., Chu, H., Schneider, M. & Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in medicine*, **26**(23), 4352–4374.

Cox, D. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

Cox, D. R. (1961). Tests of Separate Families of Hypotheses. *Proc. Fourth Berkeley Symposion on Mathematics Statistics and Probability*, **1**, 105–123.

Crowder, M., Kimber, A., Smith, R. & Sweeting, T. (1991). *Statistical Analysis of Reliability Data*. Chapman and Hall, London.

Dahiya, R. C. & Gurland, J. (1972). Goodness of fit tests for the gamma and exponential distributions. *Technometrics*, **14**(3), 791–801.

Dey, D. K. & Birmiwal, L. R. (1994). Robust Bayesian analysis using divergence measures. *Statistics & Probability Letters*, **20**(4), 287–294.

Farewell, V. T. (1982). The use mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**(4), 1041–1046.

Farewell, V. T. & Prentice, R. L. (1977). A study of distributional shape in life testing. *Technometrics*, **19**(1), 69–75.

Feller, W. (1965). *An Introduction to Probability Theory and Its Applications - Volume II*. John Wiley & Sons, Inc., New York.

Gamerman, D. & Lopes, H. F. (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Boca Raton, second edition.

Geisser, S. & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**(365), 153–160.

Gelfand, A. E., Dey, D. K. & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods (with discussion). *Bayesian Statistics*, **1**(4), 7–167.

Geweke, J. (1992). Evaluation the accuracy of sampling-based approaches to the calculation of posterior moments. assessment of local influence. *Bayesian statistics*, **1**(4), 169–193.

Ghitany, M. E. & Maller, R. A. (1992). Asymptotic results for exponential mixture models with long with long-term survivors. *Statistics*, **23**, 321–336.

Ghitany, M. E., Maller, R. A. & Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of multivariate Analysis*, **49**(2), 218–241.

Glaser, R. E. (1980). Bathtub and related failure rate characterization. *Journal of the American Statistical Association*, **75**(731), 667–672.

Gleser, L. J. (1989). The gamma distribution as a mixture of exponential distributions. *American Statistician*, **43**(2), 115–117.

Goetghebeur, E. & Ryan, L. (1995). A modified log rank test for competing risks with missing failure type. *Biometrika*, **77**, 207–211.

Greenhouse, J. B. & Wolf, R. A. (1984). A competing risk deviation of a mixture model for the analysis of survival data. *Communications in Statistics. Theory and Methods*, **13**, 3133–3154.

Gupta, R. & Kundu, D. (1999). Generalized Exponential Distributions. *Australian and New Zealand Journal of Statistics*, **41**(2), 173–188.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York.

Jodrá, P. (2008). On a Connection Between the Polylogarithm Function and the Bass Diffusion Model. *Proceedings of the Royal Society (A)*, **464**, 3081–3088.

Johnson, N. L., Kemp, A. W. & Kolz, S. (2005). *Univariate Discrete Distributions - third editions*. Willey Interscience, New Jersey - USA.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York, NY, second edition.

Kao, J. H. K. (1958). A new life-quality measure for electron tubes. *IRE Transactions on Reliability and Quality Control*, **13**, 15–22.

Kaplan, E. L. & Meier, P. (1958). Nonparametic estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(282), 457–481.

Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B. & Ramsay, N. K. C. (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, **317**(8), 461–467.

Kim, S., Chen, M. H. & Dey, D. K. (2011). A new threshold regression model for survival data with a cure fraction. *Lifetime Data Analysis*, **17**, 101–122.

Kus, C. (2007). A new lifetime distribution. *Computation Statistical & Data Analysis*, **51**, 4497–4509.

Lawless, J. F. (1980). Inference in the Generalized Gamma and Log Gamma Distributions. *Technometrics,*, **22**(3), 409–419.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, volume second edition. Wiley, New York, NY.

Li, C. S., Taylor, Jeremy, M. G. . & Sy, J. P. (2001). Identifiability of cure models. *Statistics and Probability Letters*, **54**, 389–395.

Louzada, F., Roman, M. & Cancho, V. G. (2011). The Complementary Exponential Geometric Distribution: Model, Properties, and a Comparison with its Counterpart. *Computational Statistics & Data Analysis*, **55**, 2516–2524.

Louzada, F., Cancho, V. G., Roman, M. & Leite, J. G. (2012). A new long term lifetime distribution induced by a latent complementary risk framework. *Journal of Applied Statistics*, **online**, 1–14.

Louzada-Neto, F. (1999). Poly-hazard regression models for lifetime data. *Biometrics*, **55**, 1121–1125.

Lu, K. & Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, **57**(4), 1191–1197.

Lu, K. & Tsiatis, A. A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis*, **11**(1), 29–40.

MacKenzie, G. (1996). Regression models for survival data. *Journal of the Royal Statistical Society D*, **45**(1), 21–34.

Maller, R. A. & Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data. *Biometrics*, **51**(4), 1197–1205.

Maller, R. A. & Zhou, X. (1996). *Survival Analysis with Long-Term Survivors*. Wiley, New York.

Marshall, A. W. & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the Exponential and Weibull families. *Biometrika,*, **84**(3), 641–652.

Mazucheli, J., Louzada, F. & Achcar, J. A. (2012). The polysurvival model with long-term survivors. *Revista Brasileira de Probabilidade e Estatística*, **26**, 313–324.

Mendonça, G. A. S. (1993). Câncer na população feminina brasileira. *Revistista Saúde Pública*, **17**(1), 68–75.

Minayo, M. C. d. S. (1988). Saúde-doença: uma concepção popular da etiologia. *Cad. Saúde Pública*, **4**(4), 363–381.

Mudholkar, G. S., Srivastava, S. K. & Kollia, G. D. (1996). A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data. **91**(436), 1575–1583.

Nelder, J. A. & Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal*, **7**, 308–313.

Ortega, E. M. M., Cancho, V. G. & Lachos, V. H. (2008). Assessing Influence in Survival Data with a Cure Fraction and Covariates. *Statistics and Operations Research Transactions*, **32**(2), 115–140.

Ortega, E. M. M., Cancho, V. G. & Paula, G. A. (2009). Generalized Log-Gamma Regression Models with Cure Fraction. *Lifetime Data Analysis*, **15**(1), 79–106.

Ortega, E. M. M., Cordeiro, G. M. & Pascoa, M. A. R. (2011). The Generalized Gamma Geometric Distribution. *Journal of Statistical Theory and Applications*, **10**, 433–454.

Papoila, A. L. (2011). Censura intervalar: Modelação de dados do estado atual. *Boletim Sociedade Portuguesa de Estatística*, pages 35–46.

Pascoa, M. A. R., Ortega, E. M. M. & Cordeiro, G. M. (2011). The Kumaraswamy Generalized Gamma Distribution with Application in Survival Analysis. *Statistical Methodology*, **8**(5), 411–433.

Peng, F. & Dey, D. K. (1995). Bayesian Analysis of Outlier Problems Using Divergence Measures. *The Canadian Journal of Statistics*, **23**(2), 199–213.

Perdoná, G. S. C. (2006). *Modelos de Risco Aplicados à Análise de Sobrevivência*. Tese doutorado, Instituto de Ciências Matemática e de Computação - Universidade de São Paulo, São Carlos–SP.

Perdoná, G. S. C. & Louzada-Neto, F. (2011). A General Hazard Model for Lifetime Data in the Presence of Cure Rate. *Journal of Applied Statistics*, **38**(8), 1395–1405.

Perperoglou, A., Keramopoullos, A. & Houwelingen, H. C. (2007). Approaches in modelling long-term survival: An application to breast cancer. *Statistics in Medicine*, **26**(13), 2666–2685.

Pons, O. & Lemdani, M. (2003). Estimation and test in long-term survival mixture models. *Computational Statistics & Data Analysis*, **41**, 465–479.

Proschan, F. (1963). Theorical explanation of observed decreasing failure rate. *Technometrics*, **5**(3), 375–383.

R Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiser, B., Guttman, I., Lin, D., Guess, M. & Usher, J. (1995). Bayesian inference for masked system lifetime data. *Applied Statistics*, **44**, 79–90.

Rinne, H. (2009). *The Weibull Distribuition: A handbook*. Chapman & Hall/CRC, Boca Raton, FL.

Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada, F. (2009a). On the unification of long-term survival models. *Statistics & Probability Letters*, **79**(6), 753–759.

Rodrigues, J., Cancho, V. G., de Castro, M. & Louzada, F. (2009b). On the unification of the long-term survival models. *Statistics & Probability Letters*, **79**, 753–759.

Rodrigues, J., Castro, M., Cancho, V. G. & Balakrishnan, N. (2009c). COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning Inference*, **139**, 3605–3611.

Rodrigues, J., de Castro, M., Cancho, V. G. & Balakrishnan, N. (2009d). COM–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning Inference*, **139**, 3605–3611.

Roman, M., Louzada, F., Cancho, V. G. & Leite, J. G. (2012). A new long-term survival distribution for cancer data. *Journal of Data Science,*, **10**, 241–258.

Scheike, T. (2009). *timereg package*. R package version 1.1-0. With contributions from T. Martinussen and J. Silver. R package version 1.1-6.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, **64**(4), 583–639.

Stacy, E. W. (1962). A generalization of the gamma distribution. *The Annals of Mathematical Statistics*, **33**(3), 1187–1192.

Tahmasbi, R. & Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate. *Computational Statistics & Data Analysis*, **52**(8), 3889–3901.

Tojeiro, C. A. V., & Louzada, F. (2012). A general threshold stress hybrid hazard model for lifetime data. *Statistical Papers*, **58**, 833–848.

Tojeiro, C. A. V., Roman, M. & Louzada, F. (2013). The complementary Weibull geometric distribution. *Journal of Statistical Computation and Simulation (Print)*, accepted for publication.

Tsodikov, A. D., Ibrahim, J. G. & Yakovlev, A. Y. (2003). Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models. *Journal of the American Statistical Association*, **98**(464), 1063–1078.

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(4), 739–750.

Yakovlev, A. Y. & Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of"permanent employment"in japan. *Journal of the American Statistical Association*, **87**(418), 284–292.

Apêndice A

Demonstrações e gráficos

Somatórios que necessitaram de desenvolvimento:

1) Considere $k \in (0,1)$, temos que

$$\sum_{m=1}^{\infty} mk^m = \frac{k}{1-k}\sum_{m=1}^{\infty} mk^{m-1}(1-k) = \frac{k}{1-k}E(M) = \frac{k}{(1-k)^2}, \tag{A.1}$$

em que M é uma v.a. com distribuição Geométrica com suporte em $\{1, 2, \cdots\}$ e probabilidade de sucesso $1-k$.

2)

$$\begin{aligned}
\sum_{j=0}^{\infty}\sum_{k=0}^{j} a_{jk} =& a_{00} + a_{10} + a_{11} + a_{20} + a_{21} + a_{22} + a_{30} + a_{31} + a_{32} + a_{33} + \cdots \\
=& a_{00} + a_{10} + a_{20} + a_{30} + a_{40} + a_{50} + \cdots \\
+& \quad a_{11} + a_{21} + a_{31} + a_{41} + a_{51} + \cdots \\
+& \quad a_{22} + a_{32} + a_{42} + a_{52} + \cdots \\
+& \quad a_{33} + a_{43} + a_{53} + \cdots \\
&\vdots \\
=& \sum_{k=0}^{\infty}\sum_{j=k}^{\infty} a_{jk}.
\end{aligned} \tag{A.2}$$

Considerando a definição de função de densidade para estatística de ordem, para a distribuição EG_{mn} temos

$$\begin{aligned} g_{k:n}(x) =& \frac{1}{B(k, n-k+1)} \frac{\lambda(1-\theta)e^{-\lambda x}}{(1-\theta e^{-\lambda x})^2} \left(\frac{1-e^{-\lambda x}}{1-\theta e^{-\lambda x}}\right)^{k-1} \left(\frac{(1-\theta)e^{-\lambda x}}{1-\theta e^{-\lambda x}}\right)^{n-k} \\ =& \frac{1}{B(k, n-k+1)} \frac{\lambda(1-\theta)e^{-\lambda x}}{(1-\theta e^{-\lambda x})^{n+1}} \left(1-e^{-\lambda x}\right)^{k-1} \left((1-\theta)e^{-\lambda x}\right)^{n-k} \end{aligned} \tag{A.3}$$

Usando a função de densidade para estatística de ordem para o máximo dos tempos, temos que a função densidade de ordem para a distribuição EG_{mx} é dada por

$$\begin{aligned} g_{k:n}(y) =& \frac{1}{B(k, n-k+1)} \frac{\lambda\theta e^{-\lambda y}}{(e^{-\lambda y}(1-\theta)+\theta)^2} \left(\frac{\theta(1-e^{-\lambda y})}{e^{-\lambda y}(1-\theta)+\theta}\right)^{k-1} \left(\frac{e^{-\lambda y}}{e^{-\lambda y}(1-\theta)+\theta}\right)^{n-k} \\ =& \frac{1}{B(k, n-k+1)} \frac{\lambda\theta^k \left(1-e^{-\lambda y}\right)^{k-1} \left(e^{-\lambda y}\right)^{n-k+1}}{(e^{-\lambda y}(1-\theta)+\theta)^{n+1}} \end{aligned} \tag{A.4}$$

Apresentamos alguns gráficos adicionais utilizados para formalizar conclusões contidas no texto.

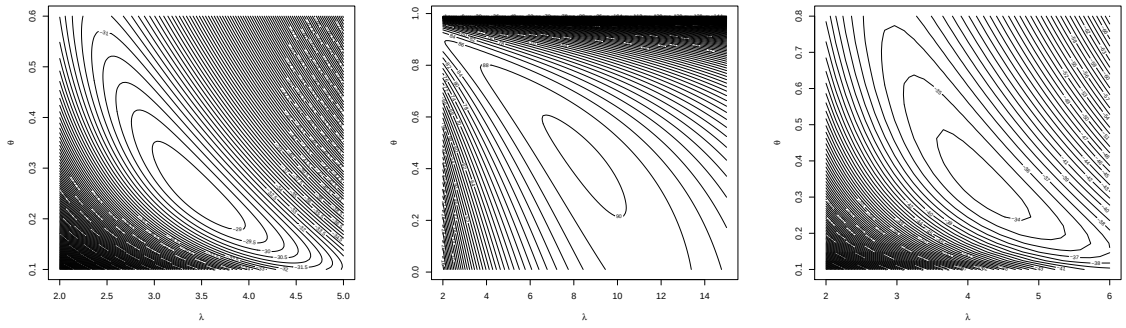

FIGURA A.1: Contorno da função log-verossimilhança para $\lambda = 3$, percentual de censura de 25% e tamanho amostral $n = 100$ das distribuições EG_{mx} (esquerda) com $\theta = 0,30$. Com $p = 0,30$ LEG_{mn} (centro) sendo $\theta = 0,20$ e LEG_{mx} (direita) sendo $\theta = 0,50$.

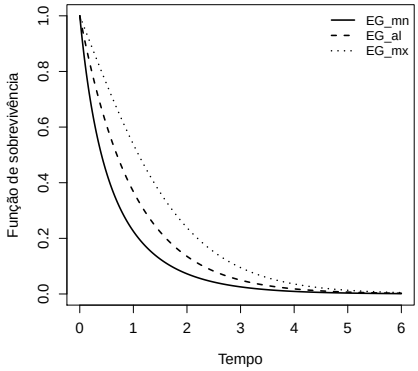

FIGURA A.2: Gráfico da função de sobrevivência dos modelos da família EG considerando os três esquemas de ativação, com $\lambda = 1$ e $\theta = 0,5$.

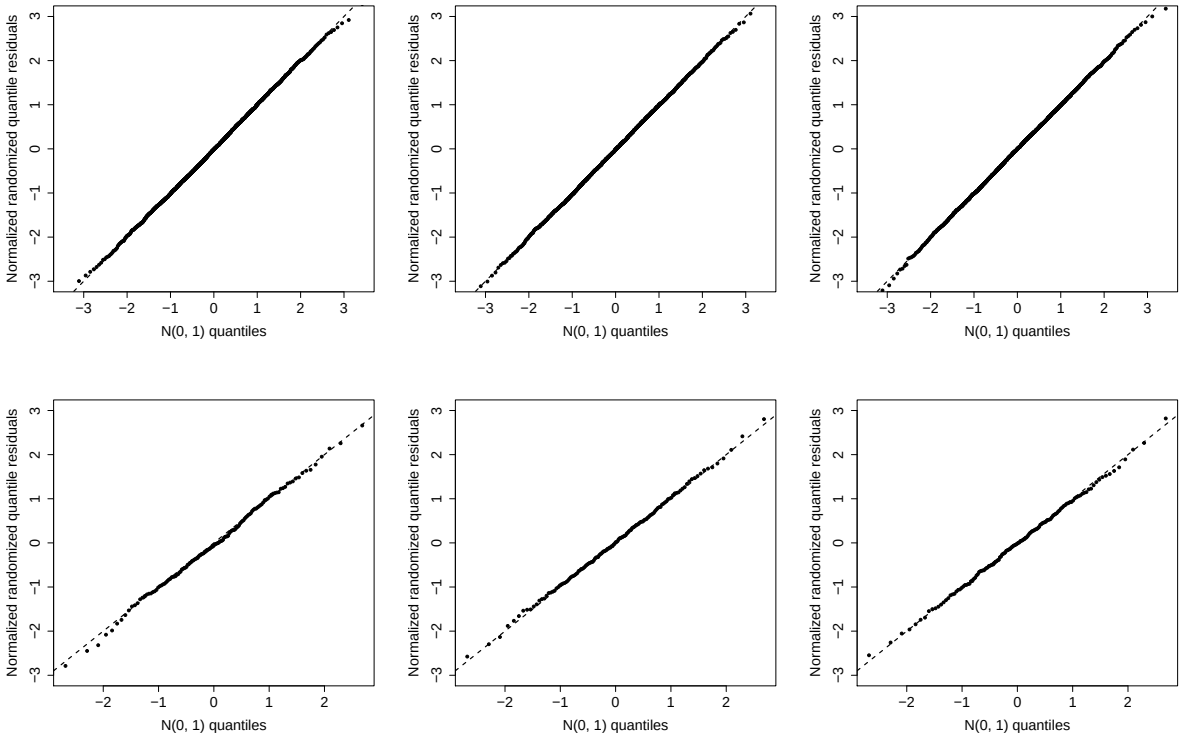

FIGURA A.3: Gráfico Quantil-Quantil dos resíduos normalizados para as parametrizações I, II e III, respectivamente. $T6$ (superior) e $T4$ (inferior).

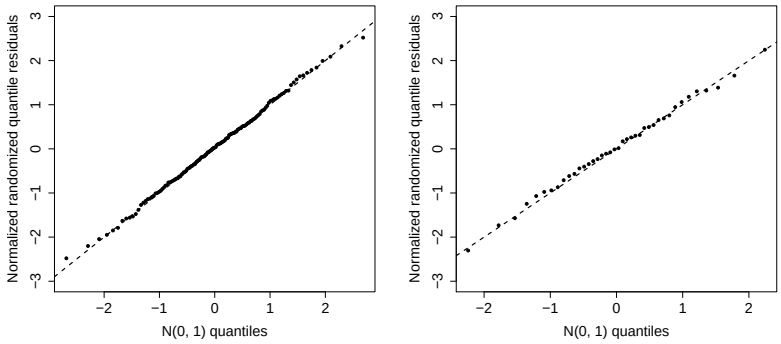

FIGURA A.4: Gráfico *QQnorm* dos modelos ajustados à $T6$ - $LGGG_{mn}$ (esquerda) - e à $T4$ - $LGGG_{mx}$ (direita)-.

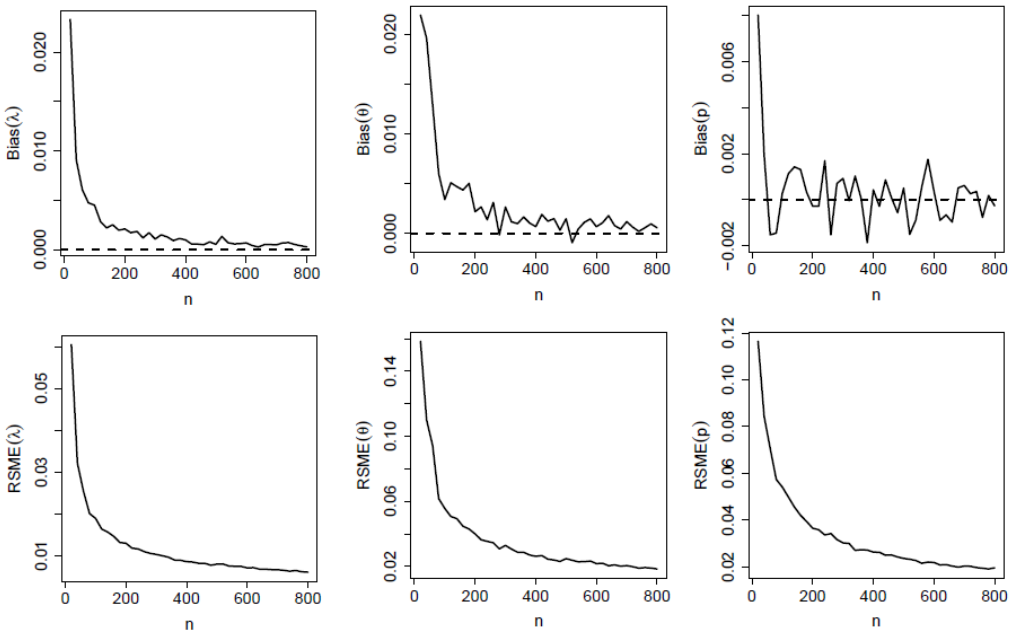

FIGURA A.5: Vício (superior) e EQM (inferior) dos EMVs dos parâmetros da LEG_{mx} pelo tamanho amostral $n = 20, 40, \ldots, 800$

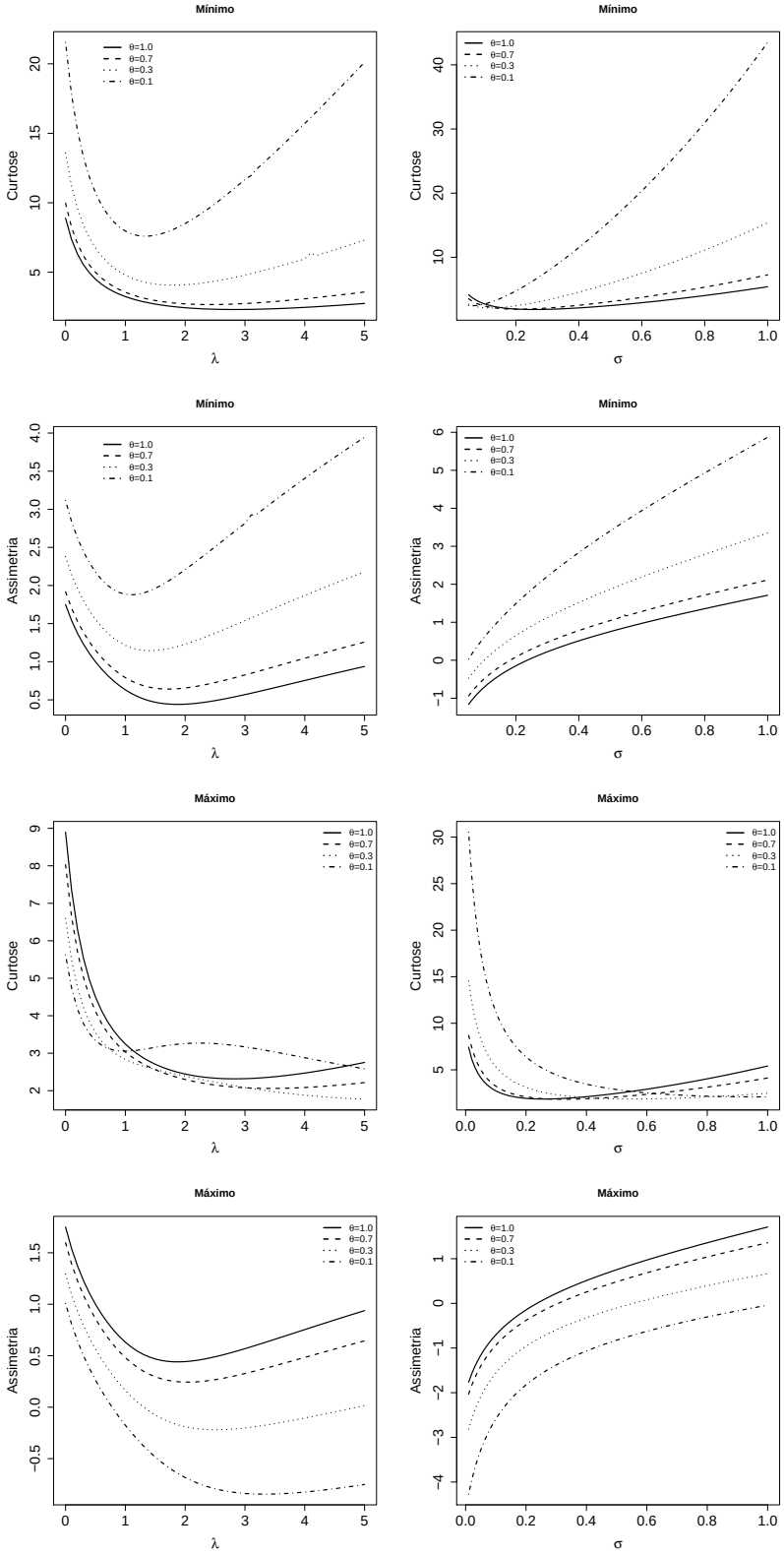

FIGURA A.6: Gráficos de assimetria e curtose com variação de θ para $\mu = -0.7$ e $\sigma = 0.5$ ou $\lambda = 4$, considerando a distribuição GGG_{mn} (Superior) GGG_{mx}(Inferior)

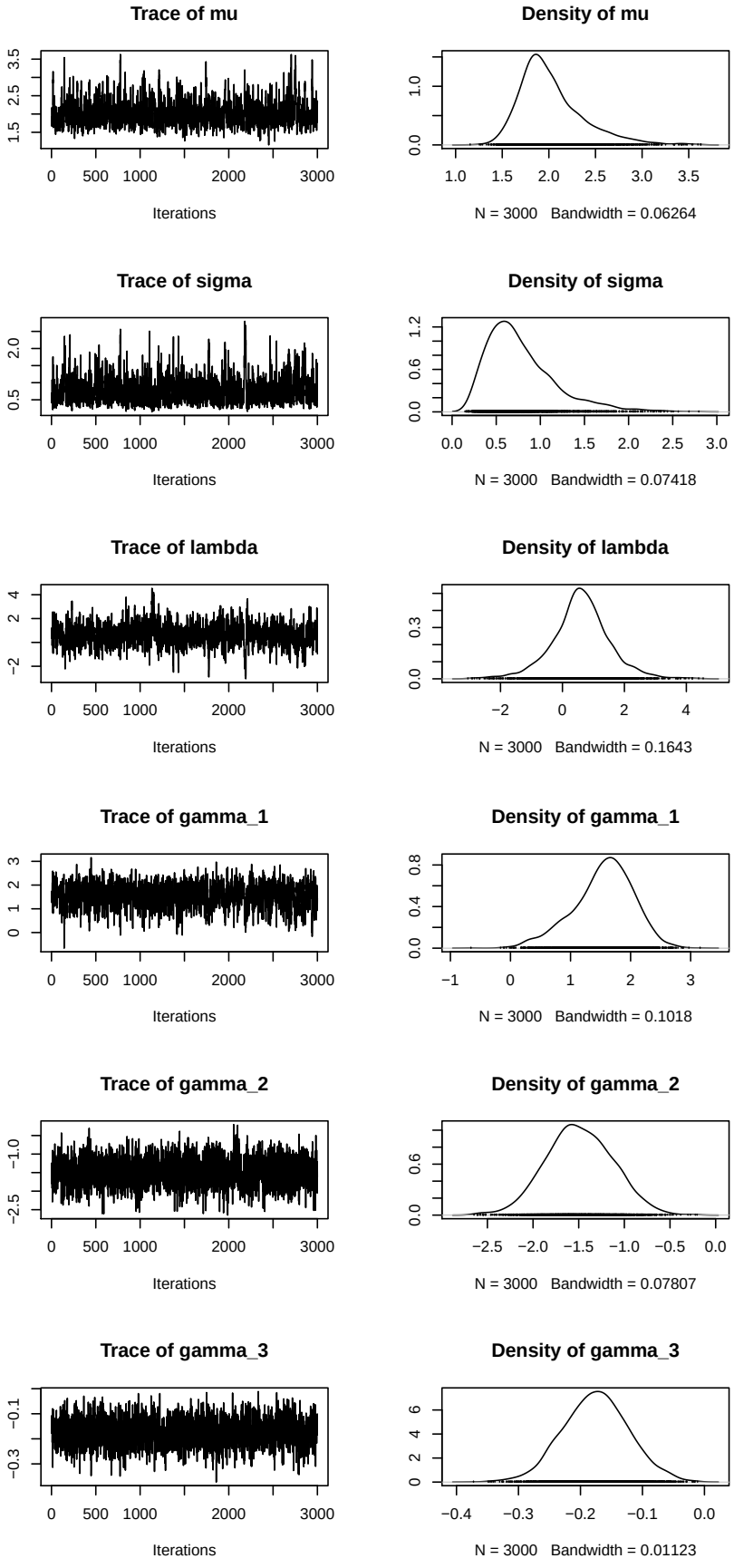

FIGURA A.7: Gráfico das densidades marginais *a posteriori* para os parâmetros do modelo $LGGG_{mn}$ no conjunto de dados reais completo.

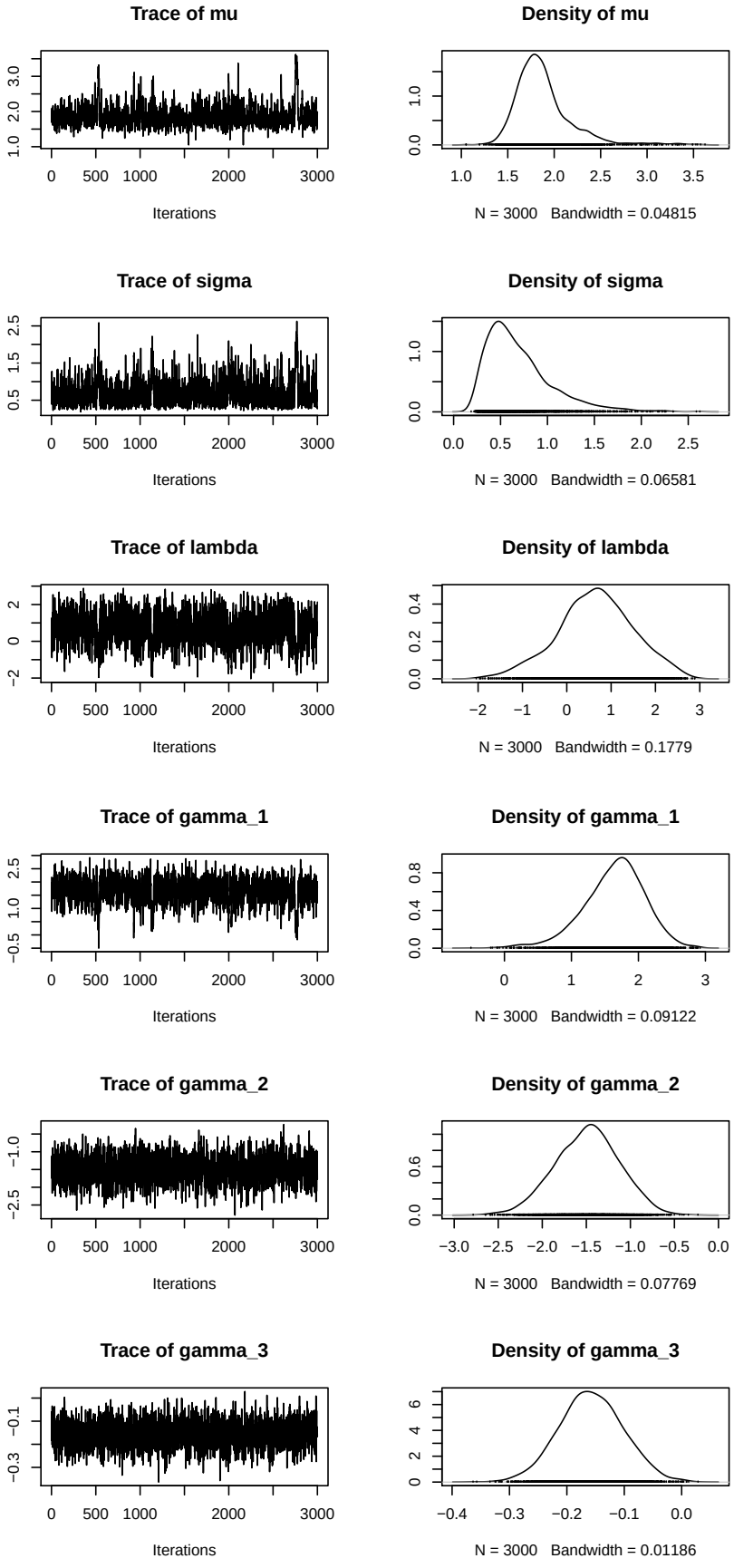

FIGURA A.8: Gráfico das densidades marginais *a posteriori* para os parâmetros do modelo $LGGG_{mn}$ no conjunto de dados reais com a retirada das observações 5 e 171.