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

Novas distribuições em análise de sobrevivência envolvendo composição e correlação dentre as causas competitivas

Vitor Alex Alves de Marchi

Orientador: Prof. Dr. Francisco Louzada Neto

UFSCar - São Carlos
Setembro/2015

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

Novas distribuições em análise de sobrevivência envolvendo composição e correlação dentre as causas competitivas

Vitor Alex Alves de Marchi

Orientador: Prof. Dr. Francisco Louzada Neto

Trabalho apresentado ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar para Defesa, como parte dos requisitos para obtenção do título de Doutor em Estatística.

UFSCar - São Carlos
Setembro/2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária/UFSCar

M317nd

Marchi, Vitor Alex Alves de.
Novas distribuições em análise de sobrevivência envolvendo composição e correlação dentre as causas competitivas / Vitor Alex Alves de Marchi. -- São Carlos : UFSCar, 2015.
132 f.

Tese (Doutorado) -- Universidade Federal de São Carlos, 2015.

1. Análise de sobrevivência. 2. Riscos Latentes. 3. Séries de potências. I. Título.

CDD: 519.9 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Tese de Doutorado do candidato Vitor Alex Alves de Marchi, realizada em 14/08/2015:

Prof. Dr. Francisco Louzada Neto
USP

Prof. Dr. Adriano Polpo de Campos
UFSCar

Prof. Dr. Adriano Kamimura Suzuki
USP

Prof. Dr. Jorge Luis Bazán Guzmán
USP

Prof. Dr. Gustavo Henrique de Araujo Pereira
UFSCar

Agradecimentos

Agradeço aos meus pais, Pedro e Rosilda, que me criaram e formaram meu caráter, tanto escolar quanto moral, e que me incentivaram na árdua tarefa de estudar, e a minha mulher Sara pela compreensão e estímulo durante os estudos.

Ao Prof. Dr. Francisco Louzada Neto pelo acolhimento para orientação, sugestões e toda a disposição para discussão do trabalho desenvolvido bem como sua compreensão em vários momentos difíceis que se fez presente.

Ao Prof. Dr. Francisco Antonio Rojas Rojas pela experiência adquirida para organização e orientação de trabalhos e pelo vínculo de amizade construído durante os anos.

Ao Prof. Dr. Marinho Gomes de Andrade pelas várias consultas e o forte vínculo de amizade estabelecidos desde o início do doutorado.

A todos os meus amigos e professores do Departamento de Estatística da UFSCar, que sempre estiveram presentes contribuindo com críticas e sugestões para o aprimoramento deste trabalho. Dentre estes se destacando, os já mencionados, o Prof. Dr. Carlos Diniz, o Dr. Hugo Henrique Kegler dos Santos, a Prof. Dra. Katiane Silva Conceição, Prof Dr. Luis Ernesto Bueno Salazar e Prof. Dr. Márcio Luis Lanfredi Viola que acabaram formando uma segunda família em São Carlos.

Aos membros da banca, Prof. Dr. Adriano Polpo de Campos, Prof. Dr. Adriano Kamimura Suzuki, Prof. Jorge Luis Bazán, Prof. Gustavo Henrique de Araújo Pereira e o Prof. Dr. Francisco Louzada Neto pela disponibilidade do seu tempo para a leitura e críticas das quais destacaram falhas e contribuições importantes para o presente trabalho.

Aos amigos de São Carlos e aos que já não fazem mais parte desta cidade pois seguiram suas vidas, e sem citar nomes para não esquecer de nenhum, deixo somente menção àqueles que convivi, dividi e compartilhei experiências tanto na convivência diária como esporádica e aos que continuo a dividir os momentos de minha vida com longas viagens me ensinando e apoiando o lado pedagógico da vida. Sim, todos vocês permanecerão presentes!

Agradeço também à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio concedido durante o doutorado.

Resumo

Nesta tese apresentamos novas distribuições para o estudo do tempo de vida com foco no cenário de riscos latentes com interpretações das construções dos modelos direcionados com os estudos da carcinogênese. Várias propriedades das distribuições são apresentadas e discutidas bem como a viabilidade destes novos modelos frente a vários modelos na literatura. Alguns dos modelos são estendidos para modelos de longa duração, que são modelos que permitem que os tempos de vida sejam infinitos ou em outras palavras que existem indivíduos não suscetíveis (ou curados) ao evento de interesse.

Pela complexidade de vários modelos obtidos foi necessário a implementação de rotinas de maximização para o controle de variáveis que não são oferecidos nas rotinas de maximização em programas tradicionais e esta rotina está disponibilizada em anexo.

Palavras-chave: Análise de Sobrevivência, Risco Latentes, Séries de Potência generalizada com parâmetro inflacionado.

Abstract

In this thesis, we construct distribution functions for analysis of lifetimes with the focus in scenes of latent risks inspired in models of the carcinogenesis process. Some properties of these distribution functions are presented and discussed as well as the viability front the models in literature. In some cases the models were describes as long-time model where a part of population presents infinite lifetime, i.e, on the population in study some objects are not susceptible (or immunes) for the event of interest.

Some models present complexity in the maximization and was necessary the implementation of routines of maximization that allows us the control some variables that are not offered in routines in tradicional programs. The implemented routine is available in appendix.

Keywords: Survival Analysis, Latents Risks, Inflated-parameter family of Generalized Power Series.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Muitos autores tem demonstrado interesse em desenvolver técnicas para a acelerada disseminação na literatura para a geração de distribuições com suporte não negativo nos últimos anos, como por exemplo, Marshall e Olkin (1997) que propuseram a mistura de distribuições com a geométrica, Gupta e Kundu (1999) que propuseram a distribuição Exponencial generalizada e (Barreto-Souza et al., 2011) que com o uso da distribuição Weibull para o tempo de vida generaliza Adamidis e Loukas (1998). O grande interesse para distribuições com o suporte em dados positivos partem especialmente do campo de análise de sobrevivência ou de confiabilidade, com desenvolvimento para dados médicos ou industriais, respectivamente.

O uso dos estimadores não paramétricos como em Kaplan e Meier (1958) era muito comum por não precisar assumir a distribuição dos dados, porém estas abordagens carecem de um ponto de vista atrativo, como a interpretação para valores de covariáveis, e assim surgiu a necessidade de assumir formas características para o conjunto de dados em análise. Na prática, podemos assumir principalmente comportamentos da distribuição dos dados, como por exemplo as formas da função de risco possuindo formas decrescentes, crescentes, unimodais ou com determinadas formas em algum conjunto do suporte da distribuição, bem como uma não simetria e/ou dispersão dos dados como assimetrias nas caudas da distribuição.

A análise de sobrevivência tem recebido um foco importante na área estatística e devido alguns modelos terem se tornado mais complexos e seu avanço se encontra ligado à capacidade dos computadores juntamente com técnicas de precisão adotadas. Para entender como os modelos podem se tornar complexos basta pensar em modelos utilizados na medicina ou em dados médicos, em que cada vez mais se conhece o comportamento das doenças, devido ao avanço das pesquisas nas suas respectivas áreas, que são posteriormente incluídos em modelos e as vezes sob diferentes pontos de vista. Veja por exemplo Rodrigues et al. (2009b) e Borges

et al. (2012).

Os dados em análise de sobrevivência referem-se ao tempo até a ocorrência de um evento de interesse. O tempo é contado desde o momento em que é iniciado a sua observação até a ocorrência da ausência de seu funcionamento, a sua destruição ou o momento em que não se é mais possível observar o objeto em estudo. Quando observado, o tempo é denominado tempo de falha e quando não é possível continuar observando o objeto em estudo devido a qualquer fator externo a ele, bem como o tempo ou o fator financeiro, dentre outros é caracterizado o tempo de censura. Com a investigação dos dados, pode-se estudar algumas características especiais, como por exemplo, com o passar do tempo o objeto pode apresentar uma menor ou maior chance de falha sobre determinadas características. Por maior chance de falha, podemos pensar por exemplo em componentes eletrônicos que podem queimar rapidamente devido a má fabricação e também a objetos que trabalham com peças móveis gerando desgastes. Além destes pontos, podemos encontrar objetos de estudos imunes ao evento de interesse e, quando há a presença destes objetos devemos utilizar modelos de longa duração que nada mais são modelos que incluem a possibilidade do tempo de vida para o evento de interesse ser infinito para a causa em estudo.

Na análise de sobrevivência, as distribuições associadas a variável resposta são em geral as distribuições Exponencial, Weibull, Log-Normal e Gama, sendo que a distribuição Exponencial é tida como a distribuição inicial na análise dos dados. Por esta apresentar várias restrições, como função de risco constante e função densidade de probabilidade decrescente o desenvolvimento de novas distribuições se tornou um foco de atenção dentre os pesquisadores das áreas de confiabilidade.

Seja Y uma variável aleatória não negativa que representa o tempo de falha de um elemento, e seja C uma variável aleatória independente de Y, que representa o tempo de censura associado a esse elemento. Assim, o dado observado é representado por $y = \min(Y, C)$, e δ é o indicador de censura dado por

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

Seja Y uma variável aleatória, não negativa contínua, que representa o tempo de vida de um indivíduo proveniente de uma dada população homogênea. A função densidade de probabilidade (f.d.p.), $f(y)$, é definida por

$$f(y) = \lim_{\Delta y \to 0} \frac{P(y \leq Y \leq y + \Delta y)}{\Delta y} = \frac{dF(y)}{dy},$$

em que $F(y) = P(Y \leq y)$ é a função distribuição de Y, sendo
$f(y) \geq 0, \quad \int_0^\infty f(y)dy = 1$ e $F(y) = \int_0^y f(y)dy$.

Uma importante relação com a função distribuição na análise de tempo de vida é a função de sobrevivência dada pela relação $S(y) = 1 - F(y)$, pois esta representa a probabilidade do objeto em estudo sobreviver além do tempo y.

A função de risco, $h(y)$, representa o risco instantâneo no instante y condicional à sobrevivência até o tempo y, e é definida por

$$h(y) = \lim_{\Delta y \to 0} \frac{P(y \leq Y \leq y + \Delta y | Y \geq y)}{\Delta y} = \frac{f(y)}{S(y)}. \tag{1.0.1}$$

Por meio da função de risco pode-se determinar qual é o melhor momento de substituição de um componente eletrônico, visto que a partir de um ponto o risco pode ser somente crescente, bem como pode-se determinar um período de atenção e/ou manutenção maior de um paciente ou máquina até um determinado momento até o risco se tornar decrescente ao longo do tempo.

Nos últimos anos tem crescido muito a generalização e/ou a modificação de algumas distribuições utilizadas em análise de sobrevivência. Exemplos podem ser vistos em Adamidis e Loukas (1998), que propuseram uma variação da função de distribuição Exponencial, a distribuição Geométrica Exponencial (EG), com função de risco decrescente, Gupta e Kundu (1999), que propõe a função distribuição Exponencial generalizada, com funções de risco crescente e decrescente, Kus (2007), que propôs uma outra modificação da função de distribuição Exponencial com função de risco decrescente, Barreto-Souza e Cribari-Neto (2009), generalizou a função distribuição proposta por Kus (2007) com a adição de um parâmetro de potência, o qual possibilitou funções de risco crescente, decrescente ou unimodal e Barriga et al. (2011) que propuseram a função distribuição Exponenciada Exponencial complementar fazendo a exponenciação da distribuição proposta em Smith e Bain (1975).

Uma forma empírica de determinar o comportamento da função risco é pelo gráfico do tempo total em teste (TTT), o qual é definido para amostras cuja as observações são não censuradas por $G(r/n) = \frac{\sum_{i=1}^{r} Y_{i:n} + (n-r)Y_{r:n}}{\sum_{i=1}^{n} Y_{i:n}}$, em que $r = 1, \ldots, n$ e $Y_{r:n}$ é a estatística de ordem r da amostra. A versão para amostras com censuras pode ser encontrada em Nachlas

(2005). Alguns autores também utilizam o gráfico $G(r/n^*)$ para os valores não censurados da amostra.

A Figura 1.1 apresenta as curvas do gráfico TTT e estas podem representar em particular quatro formatos típicas da função de risco das distribuições. Se a função de risco tem forma crescente (A); forma decrescente (B); forma unimodal (C); forma de banheira (D) e com risco constante (E). O comportamento da forma crescente é vista de tal forma que a curva TTT se encontra acima da linha diagonal e a forma decrescente abaixo da linha diagonal.

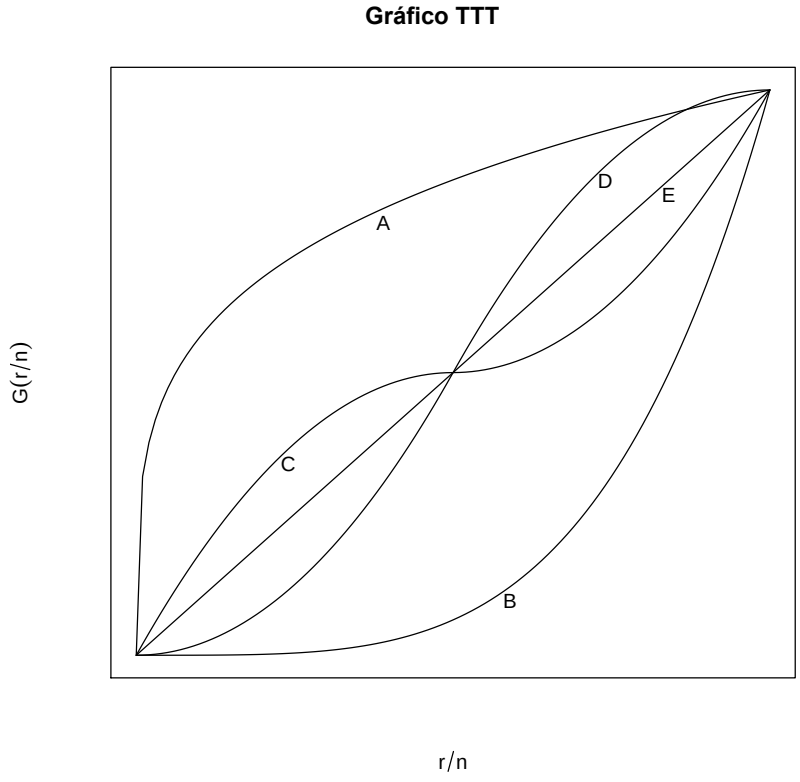

Figura 1.1: Formas da função de risco pela curva TTT.

A Figura 1.2 mostra uma função onde não há fração de cura (gráfico a direita) e uma amostra que há a evidência de uma parte da população ser imune ou ter uma fração de cura (gráfico a esquerda). A evidência de longa duração dos objetos em estudo é baseado no comportamento da curva da função sobrevivência, em que a longa duração é suposta caso a curva se estabilize em algum valor maior que 0. Uma estimativa para a função de sobrevivência não paramétrica bastante difundida é o estimador de Kaplan-Meier (K-M) (Kaplan e Meier, 1958).

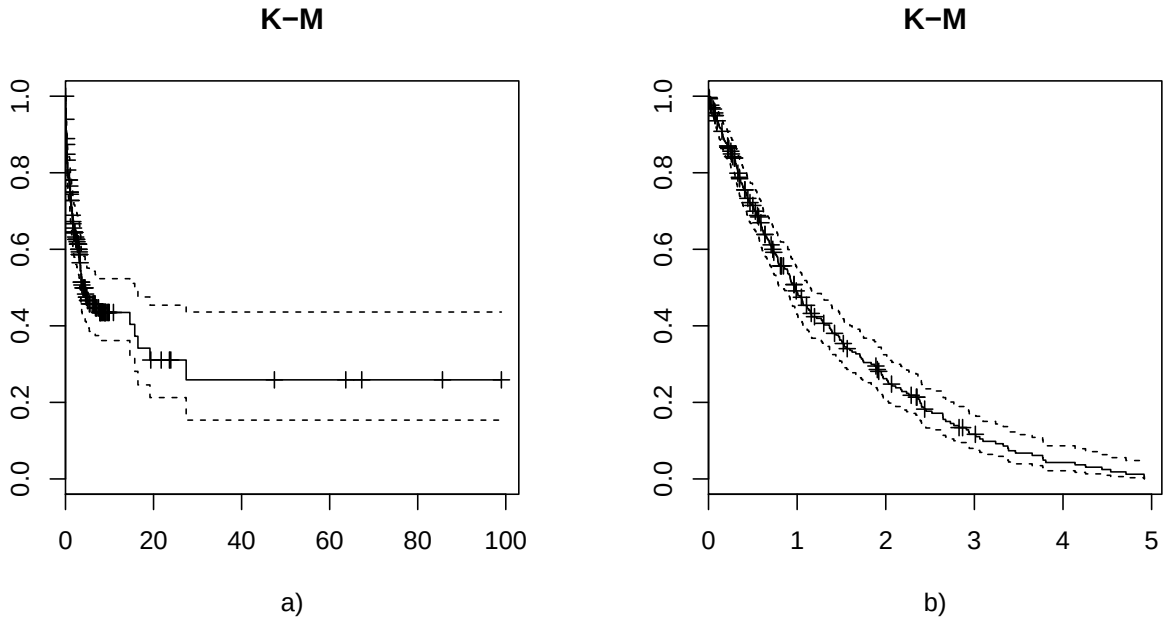

Figura 1.2: a) O estimador K-M mostra evidência de haver uma imunidade ou fração de cura na população; b) o estimador K-M não mostra evidência de fração de cura ou imunes.

1.1 Motivação e objetivos da tese

O foco principal é a obtenção de distribuições com diferentes formas para a função de risco e estas distribuições são obtidas com a mistura de distribuições dos riscos em cenários de riscos competitivos e riscos complementares e sempre que propício são apontados suas interpretações práticas e de construção.

Para o trabalho desenvolvido, considere o processo de carcinogênese, onde existem várias células (fatores de risco) que podem vir a se tornar cancerígenas mas que por algum motivo ao ter a célula ativada, seja pela regeneração ou pelo próprio sistema imunológico eliminar a célula cancerígena, ela pode não se tornar uma célula que provocará câncer (promovida). Neste processo da detecção do tumor não é possível determinar quantos fatores de risco existiam no paciente nem como a informação se o paciente em estudo morreu devido a primeira célula se tornar cancerígena, somente após todas as células se tornarem cancerígenas ou algum número intermediário. Este trabalho, dentro da análise desenvolvida na literatura, se concentra em modelos de mistura com ativação latente, onde não se é distinguido o fator causador (fator de risco) e também não existe a informação de quantos fatores de riscos existem no objeto em estudo e temos o desenvolvimento de distribuições assumindo que toda célula ativada se

torna uma célula promovida e ainda modelos, apresentados a partir do Capítulo 5, que além de trabalharem com ativações latentes incorporam a probabilidade de uma célula ativada ser promovida, bem como a correlação entre as células ativadas, ou seja, os modelos são capazes de responder a pergunta: "Se uma célula teve seu núcleo danificado, qual a probabilidade desta célula não se tornar um tumor?", com base no fato: "uma célula provoca o tumor somente se teve seu núcleo danificado por algum motivo".

Neste trabalho nos concentramos em utilizar os termos relacionados a análise de sobrevivência, porém, sua analogia com dados de confiabilidade para um determinado sistema na indústria pode ser feita de forma direta.

1.1.1 Fatores de risco e a ativação

Em cenários de estudo do tempo de vida com risco competitivos geralmente não é observado qual fator de risco ocasionou a falha, também conhecidos como riscos latentes, seja porque a informação não foi verificada por simplesmente não ser possível ser especificada ou por a verdadeira informação não ser detectável, o qual ocorre por exemplo em um objeto em que no final de sua vida tem-se a sua destruição. Entre vários cenários é comumente assumido que o primeiro dos riscos (primeiro a falhar) é responsável pela falha do objeto. Este tipo de cenário está relacionado intrinsecamente com sistemas em série em sua interpretação. Por outro lado pode-se assumir que a falha do objeto é ocasionada pela falha do último fator de risco, ou por alguma quantidade intermediária. Quando é assumido que depois do último fator em risco ocasionar a falha o sistema é intrinsecamente ligado a sistemas paralelos. Para entender a relação do sistema em série/paralelo com a ativação dos fatores de risco é aconselhável ao leitor tem em mente a representação da Figura 1.3. É importante ressaltar que cada fator de risco m está associado a um tempo de vida X_m.

Um ponto crucial para determinar a função do tempo de vida do sistema é saber quantos riscos estão associados, no entanto, neste trabalho não assumimos um valor fixo por questões de tornar o modelo mais parecido com o conhecimento atual da área de estudo na prática (área médica ou de confiança industrial). Assumindo que os riscos são latentes vamos atribuir uma probabilidade para cada quantidade de risco e consequentemente determinamos a função conjunta e a distribuição marginal da ativação os riscos (mínimo ou máximo), que é a distribuição de interesse. Em outras palavras, temos M riscos competitivos, desconhecidos com uma função de distribuição (f.d.) $P(M = m)$, $m = 0, 1, 2, \ldots$ e a função de distribuição de ordem dado $M = m$, $F(Y = y|M = m)$, em que geralmente $Y = \min\{X_1, \ldots, X_m\}$ (ou

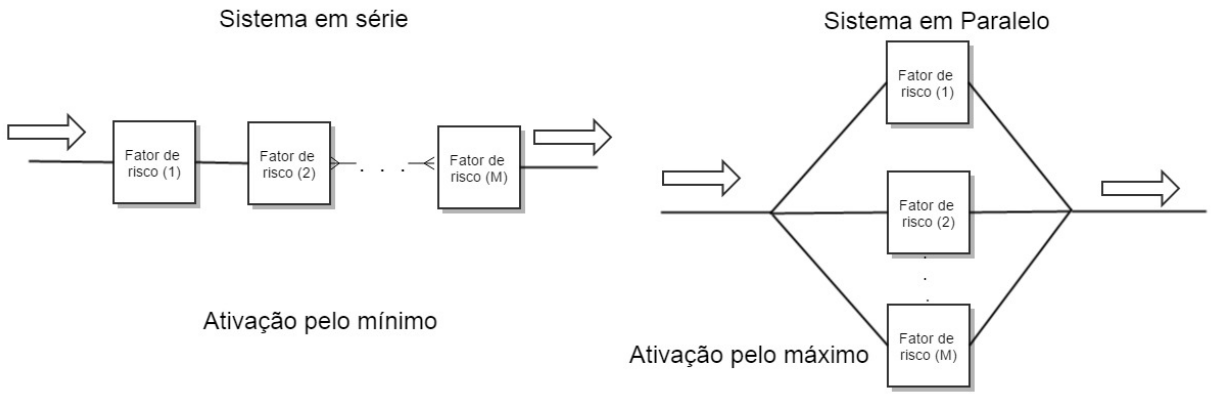

Figura 1.3: Esquematização de um sistema em série e em pararelo.

$Y = \max\{X_1, \ldots, X_m\}$) e portanto $F(Y = y) = \sum_m F(Y = y|M = m)P(M = m)$.

Várias distribuições foram mostradas na literatura utilizando riscos competitivos e o leitor interessado nesta ampla abordagem pode consultar mais detalhes em Louzada-Neto (1999), Lawless (2003), Crowder et al. (1991), Cox e Oakes (1984), Adamidis e Loukas (1998), Gupta e Kundu (2001), Gupta e Kundu (1999), Nadarajah e Kotz (2006), Kus (2007), Barreto-Souza e Cribari-Neto (2009), Louzada-Neto et al. (2011) e Louzada et al. (2011).

1.2 Organização da Tese

No Capítulo 2 apresentamos duas distribuições com suporte positivo para dados em cenário de risco latentes, denominadas distribuições E2G e CE2G, no qual buscamos uma alternativa menos complicada de expressar as distribuições de sobrevivência aos modelos recentemente apresentados na literatura e mantendo o foco em apresentar modelos que possuem variações nas suas respectivas funções de risco. Resultados das características das distribuições são apresentados analiticamente, bem como estudo de simulação e aplicação dos modelos a dados reais. No Capítulo 3 fazemos uma extensão do modelo CE2G para sua forma logarítima, podendo assim realizar a inferência estatística aproveitando vários resultados dos modelos de regressão Normal, porém com a adição de observações censuradas. Alguns resultados dos modelos de tempo de vida para modelos de log-locação-escala (ou tempo de vida acelerado) são estendidos diretamente para a distribuição Log-CE2G e seus resultados são apresentados numa aplicação em dados reais. No Capítulo 4 é estudado a viabilidade das distribuições IGPS (Kolev et al., 2000) com o intuito de substituir as distribuições Série de Potências (SP) normalmente

utilizadas em cenário de riscos latentes bem como para verificar se estes produzem modelos com funções de riscos diferenciadas. Neste capítulo apresentamos as distribuições como alternativas também para os modelos de regressão inflacionados devido a flexibilidade das distribuições IGPS. A viabilidade do modelo é estudada com aplicações em dados reais e a análise do modelo de regressão é checada também com medidas já conhecidas para modelos de dados de contagem. No Capítulo 5 apresentamos a distribuição CoPE, que é a aplicação de um caso particular da distribuição IGPS, juntamente com vários resultados analíticos dos momentos, da função característica, da função de risco entre outros. Neste capítulo também fazemos a interpretação de seu uso no estudo de carcinogênese como uma alternativa a modelos que consideram a capacidade de uma célula iniciada não se desenvolver (ser promovida). Também é verificado sua viabilidade com aplicações em conjuntos de dados reais e suas contribuições do ponto de vista de interpretação nos dados. No Capítulo 6, nos reservamos a apresentar os resultados das demais distribuições com o uso da distribuição IGPS e suas relações com os resultados apresentados no Capítulo 5 como propostas futuras pois carecem de análises mais detalhadas. Durante a apresentação dos resultados mostramos a aplicação das distribuições em três conjuntos de dados. Ainda, no Capítulo 6 apresentamos as funções de longa duração dos modelos com o uso da distribuição IGPS por meio da aplicação da unificação apresentada por Rodrigues et al. (2009a). A extensão dos modelos são realizados de tal forma que não é incluído mais um parâmetro no modelo (o parâmetro de longa duração), exceto pelo modelo L-CoLSE, e este parâmetro de longa duração é obtido de forma natural nas distribuições IGPS como poderá ser visto no decorrer do capítulo. No final do Capítulo 6 apresentamos uma aplicação dos modelos de longa duração a um conjunto de dados que não possui na literatura a estimação por meio de modelos paramétricos que consideram o termo de longa duração, e como observado nos comentários, os modelos de longa duração obtidos se apresentam adequados.

Devido a complexidade de maximizar as funções de log-verossimilhança apresentadas nos Capítulos 5 e 6, foram implementadas rotinas de maximização no *software* Matlab (MATLAB, 2010) pela necessidade de controle de parâmetros do tamanho do passo (valor "a" na rotina do Anexo E) do processo BFGS. Para maiores detalhes do algorítmo recomenda-se a leitura do Anexo E. Também, no Anexo E, apresentamos as funções implementadas para a distribuição CoPE, porém para as demais funções basta a alteração da função de logverossimilhança e consequentemente das funções de densidade e de sobrevivência associadas ao modelo escolhido. Ainda, no que se diz respeito ao uso da rotina implementada, ao escolher a distribuição a ser estimada, deve-se tomar valores diferenciados para o tamanho do passo e/ou

para a precisão da matriz de derivadas e das derivadas numéricas para a rotina não apresentar problemas numéricos, como por exemplo o cálculo de valores log(0) ou para não retornar o próprio ponto inicial.

A Figura 1.4 apresenta a estrutura para obter as distribuições desenvolvidas neste trabalho (sublinhadas) e a comparação com os principais resultados da literatura. Esta figura tem como objetivo auxiliar durante a leitura inicial a identificar rapidamente qual a estrutura dos riscos adotada bem como o cenário do tema motivacional (risco competitivo ou complementar na carcinogênese) e seus relacionamentos com técnicas já desenvolvidas e deve ser consultada sempre que houver alguma dúvida. As distribuições desenvolvidas no Capítulo 2 são as distribuições CEG e CE2G. A distribuição desenvolvida no Capítulo 5 utiliza o parâmetro "rho" da distribuição IGPS, que é indicada pela letra inicial "I", e com valores de "rho" igual a zero indica-se que a distribuição não apresenta este parâmetro e portanto recaem nas suas respectivas distribuições sem a inflação. A anotação "A serem desenvolvidas" representa o fato de resultados para as distribuições mencionadas no centro da figura, bem como demais distribuições, não terem sido encontradas na literatura e portanto suas obtenções se encontram em aberto para estudo.

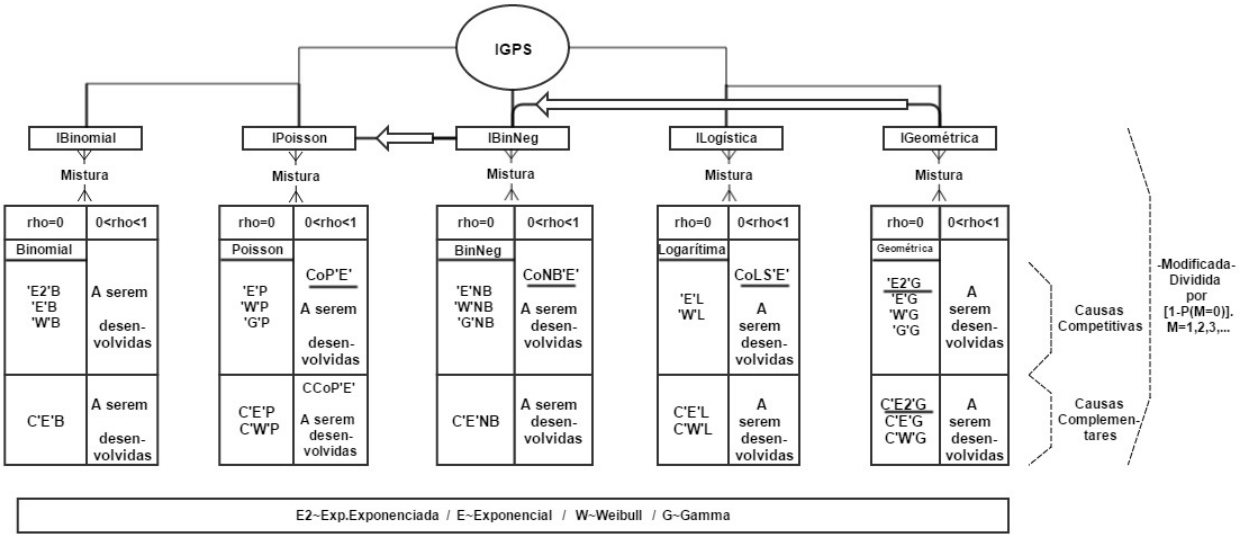

Figura 1.4: Desenho esquemático dos resultados do trabalho em comparação com os principais resultados da literatura.

Os resultados apresentados no Capítulo 2 foram condensados nos artigos Louzada et al. (2014) e Marchi et al. (2013). Também os resultados apresentados no Capítulo 3 foram utilizados na elaboração do artigo Louzada e Marchi (2015) já aceito para publicação. Os resultados

dos capítulos 4 e 5 foram condensados em relatórios técnicos e submetidos para publicação.

Capítulo 2

Distribuição E2G e CE2G

Neste capítulo abordaremos algumas novas distribuições que podem ajudar a aumentar o conhecimento sobre determinadas características dos dados em análise de sobrevivência pela flexibilidade da função de risco. Existem vários modelos de construção para as distribuições de sobrevivência como citados na introdução, porém por enquanto usaremos a restrição de $M = 1, 2, \ldots$. O fato de $M > 0$ implica diretamente que o modelo não apresenta fração de cura.

Este capítulo se baseia inteiramente no modelo de mistura com distribuições Geométricas em causas competitivas e complementares. Podemos notar que estes modelos são casos particulares dos modelos de Marshall e Olkin (1997), porém o método a qual é construído engloba a construção de distribuições no cenário de risco competitivos e complementares.

2.1 Distribuições no cenário de riscos latentes

Sejam $X_i, \ldots, X_m$ variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), com função distribuição de probabilidade (f.d.p.) $f(x|\phi)$. A variável de estatística de ordem $X_{r:m}$, $r = 1, \ldots, m$, é dada por

$$f_{r,m}(x|M = m, \phi) = \frac{m!}{(r-1)!(m-r)!} f_\phi(x)[F_\phi(x)]^{r-1}[1 - F_\phi(x)]^{m-r},$$

em que $f_\phi = f(x|\phi)$.

Considere $P(M = m)$ uma distribuição Série de Potência modificada em zero dada por $P(m) = \frac{\sum_{m=1}^{\infty} a_m \theta^m}{C(\theta)}$, em que $C(\theta) = \sum_{m=1}^{\infty} a_m \theta^m$ com $\theta \in (0, s)$. A Tabela 2.1 mostra alguns casos particulares para esta distribuição. Por exemplo, se $a_m = (m!)^{-1}$, temos $C(\theta) = e^\theta - 1$ a qual caracteriza a função de Poisson com parâmetro θ.

É importante notar que quando é utilizada a função de distribuição Geométrica, e é adotado $Y = \min\{X_1, \ldots, X_n\}$ ou $Y = \max\{X_1, \ldots, X_n\}$ a fórmula da função para o tempo de vida resultante é encontrada em Marshall e Olkin (1997) e o resultado que apresentamos na Seção 2.1.2 é um caso particular, restrito para o valor do parâmetro da Geométrica com $\theta \in [0, 1]$ e que pode ser estendido para valores com suporte positivo, no entanto o foco do trabalho é manter as características iniciais de forma a ter uma construção e explicação no objeto em estudo.

2.1.1 Escolha da distribuição dos riscos latentes como alternativa a Weibull e Gama

A distribuição Geométrica para M é abordado em Adamidis e Loukas (1998), juntamente com o tempo de vida $F(y)$ seguindo uma distribuição Exponencial para o tempo de vida com a regra $Y = \min\{X_1, \ldots, X_n\}$. Posteriormente Louzada et al. (2011) utilizaram as mesmas distribuições, porém, considerando $Y = \max\{X_1, \ldots, X_n\}$. Em ambos os casos o desenvolvimento foi motivado pela forma da função de risco, que no primeiro caso é decrescente e no segundo caso é crescente.

Outros autores como Barreto-Souza et al. (2011) generaliza Adamidis e Loukas (1998) com o uso da distribuição do tempo de vida $F(y)$ seguindo uma distribuição Weibull. As formas de sua função de risco são possuém diferentes formas. Gupta e Kundu (1999), propõem a distribuição Exponencial generalizada, com funções de risco crescente e decrescente, Kus (2007) propôs outra modificação na função exponencial e tem função de risco decrescente, Barreto-Souza e Cribari-Neto (2009) generalizaram a função proposta em Kus (2007) com a inclusão do parâmetro de potência, atribuindo assim a forma crescente, decrescente e unimodal para a função de risco and Barriga et al. (2011) propuseram a distribuição Exponencial Complementar Exponenciada utilizando a distribuição proposta por Smith e Bain (1975) dentre outros.

Como o leitor pode perceber, a função de distribuição Exponencial está presente em

Tabela 2.1: Principais casos particulares da distribuição Série de Potência

Distribuição	a_m	$A(\theta)$	s
Poisson	$m!^{-1}$	$e^{\theta} - 1$	∞
Geométrica	1	$\theta(1-\theta)^{-1}$	1
Logarítima	m^{-1}	$-\log(1-\theta)$	1

vários modelos como caso particular por sua grande importância no estudo do tempo de vida. Dentre as funções com suporte positivo, que estudam o tempo de vida, temos duas de dois parâmetros que se destacam, a Weibull e a Gama, sendo que a segunda possui uma forma não fechada para o cálculo da função Gamma dependendo do valor do parâmetro.

Neste capítulo utilizaremos uma distribuição que sempre possui forma fechada para o calculo da f.d.p da função distribuição e da função de sobrevivência, a distribuição Exponencial exponenciada (Gupta e Kundu, 2001), e tem como caso particular a distribuição Exponencial. A função distribuição da Exponencial exponenciada é dada por $\left(1-e^{-\lambda x)}\right)^{\alpha}$, em que $\lambda, \alpha > 0$. Segundo Gupta e Kundu (2001), além da distribuição Gama ter suas desvantagens por ser calculada de forma numérica, também pode-se encontrar desvantagens na distribuição Weibull (Gorski, 1968) e por este motivo escolhemos como distribuição de base no processo de composição a distribuição Exponencial exponenciada.

2.1.2 Classe de distribuições no cenário de riscos latentes usando a Geométrica

Mostramos a seguir a distribuição de falha para os três casos de ativação: falha por ativação do primeiro fator de risco, falha por ativação de um fator de risco aleatório e falha por ativação do último fator de risco.

Considere $F_r(y)$ e $S_r(y)$ a função distribuição e de sobrevivência dos riscos, respectivamente. Portanto obtemos as distribuições de:

- Falha por ativação do primeiro fator de risco.

  Utilizando Marshall e Olkin (1997) obtemos a função distribuição $F(y)$ dada por

$$F(y) = \frac{1 - S_r(y)}{1 - (1-\theta)(1 - S_r(y))}.$$

- Falha por ativação do último fator de risco.

  Utilizando Marshall e Olkin (1997) obtemos a função distribuição $F(y)$ dada por

$$F(y) = \frac{\theta F_r(y)}{1 - (1-\theta)F_r(y)}.$$

- Falha por ativação de um fator de risco aleatório.

Neste caso, para a obtenção da função distribuição de falha, determinamos que a probabilidade do fator r, $r = 1, \ldots, m$, falhar é $\frac{1}{m}$, ou seja, a chance da falha ser causada pelo fator de risco r é igual para o fator $r-1$ por exemplo, e assim

$$\begin{aligned} f(y) &= \sum_{m=1}^{\infty}\sum_{r=1}^{m} f(y|M=m, R=r)P(R=r|M=m)P(M=m) \\ &= \sum_{m=1}^{\infty}\sum_{r=1}^{m} \frac{(m-1)!}{(r-1)!(m-r)!} f_\phi(y)[F_\phi(y)]^{r-1}[1-F_\phi(y)]^{m-r}P(M=m) \\ &= \sum_{m=1}^{\infty} f_\phi(y)P(M=m) \\ &= f_\phi(y)\sum_{m=1}^{\infty} P(M=m) \\ &= f_\phi(y). \end{aligned}$$

Para obter a expressão da segunda linha, utilizamos o fato de que $P(R = r|M = m) = 1/m$ e para obter a terceira linha basta observar que a expressão é a expansão binomial com o fator $r-1$.

Portanto, atribuindo a mesma probabilidade para cada fator de risco de ser o causador da falha, obtemos que a própria função distribuição dos riscos é a função distribuição do tempo de falha.

2.2 Distribuição E2G

Seja Y uma variável aleatória denotando o tempo de vida de um componente. A variável aleatória Y tem distribuição E2G com parâmetros $\lambda > 0$, $\alpha > 0$ e $0 < \theta < 1$ se sua f.d.p é dada por

$$f(y) = \frac{\alpha\lambda\theta e^{-\lambda y}(1-e^{-\lambda y})^{\alpha-1}}{\left[1-(1-\theta)(1-(1-e^{-\lambda y})^{\alpha})\right]^2}, \tag{2.2.1}$$

em que λ é o parâmetro de escala, com α e θ sendo parâmetros de forma. A distribuição EG (Adamidis e Loukas, 1998) é obtida da distribuição E2G atribuindo o valor $\alpha = 1$. A Figura 2.2 mostra a f.d.p. da distribuição E2G (parte superior) e a função de risco (parte inferior) desta distribuição para $\theta = 0.01, 0.5, 0.99$ e $\alpha = 0.3, 1, 10$ com $\lambda = 1$ fixo.

A função de sobrevivência da distribuição E2G é dada por

$$S(y) = \frac{\theta(1-(1-e^{-\lambda y})^{\alpha})}{1-(1-\theta)(1-(1-e^{-\lambda y})^{\alpha})}, \tag{2.2.2}$$

em que $\alpha > 0$, $\theta \in (0, 1)$ e $\lambda > 0$.

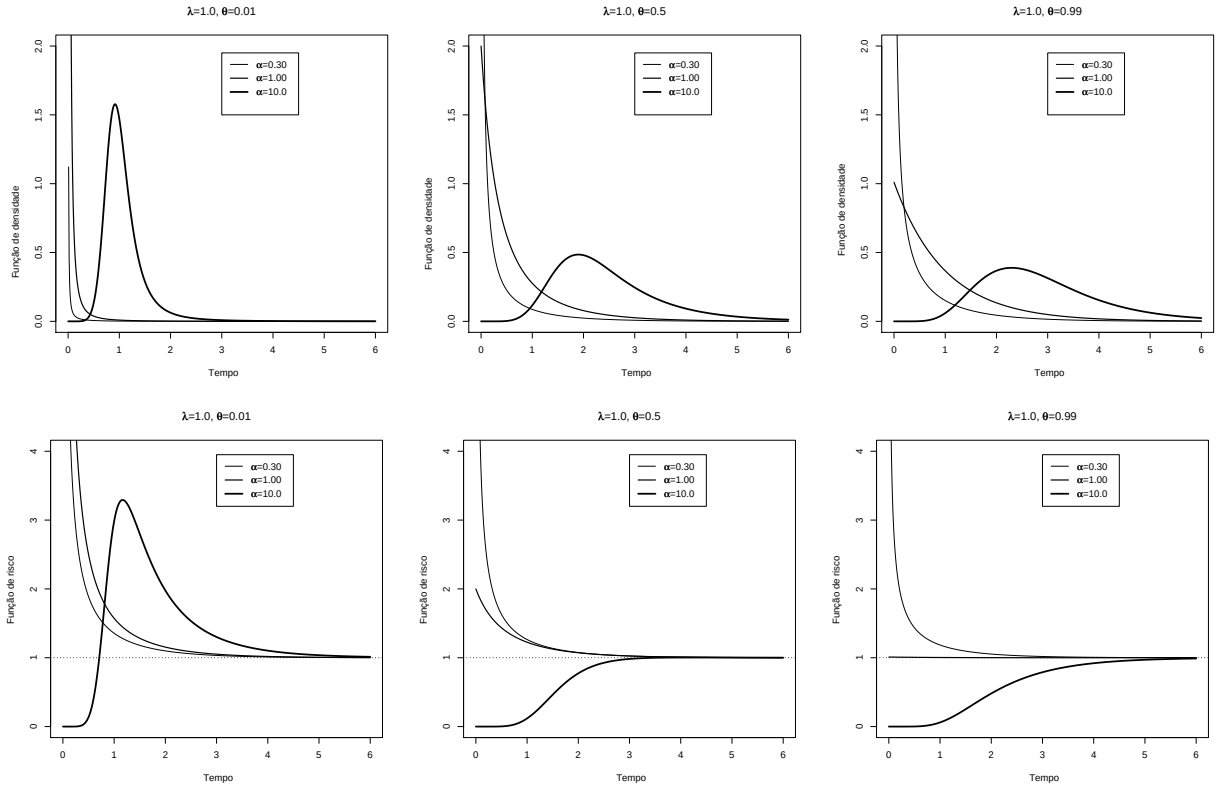

Figura 2.1: Superior: F.d.p. da distribuição E2G. Inferior: Funções de risco da distribuição E2G.

Da equação (2.2.2), a função de risco, de acordo com a relação $h(y) = f(y)/S(y)$, é dada por

$$h(y) = \frac{\alpha\lambda e^{-\lambda y}(1-e^{-\lambda y})^{\alpha-1}}{(1-(1-e^{-\lambda y})^{\alpha})\left[1-(1-\theta)\left(1-(1-e^{-\lambda y})^{\alpha}\right)\right]}. \tag{2.2.3}$$

Para $\alpha < 1$, temos que $h(y)$ não é finita, caso contrário temos que para $y = 0$ seu valor é finito e é dado por $h(0) = \lambda/\theta$ se $\alpha = 1$ e $h(0) = 0$ se $\alpha > 1$. Quando $y \to \infty$, a função tende a se estabilizar e portanto $\lim_{y\to\infty} h(y) = \lambda$.

A função de risco (2.2.3) pode se crescente, decrescente ou unimodal como é observado na Figura 2.2 (parte inferior), para os valores de $\theta = 0.01, 0.5, 0.99$ e $\alpha = 0.3, 1, 10$.

O $p-$ésimo quantil da distribuição E2G, ou a função inversa da função distribuição $F(q) = p$, é dada por

$$Q(p) = F^{-1}(p) = -\frac{\ln(1-(\frac{p\theta}{1-p+p\theta})^{1/\alpha})}{\lambda}, \tag{2.2.4}$$

em que $F(y) = 1 - S(y)$ é a função distribuição de Y.

De forma geral, podemos mostrar que para qualquer distribuição exponenciada, digamos $F_*^{\alpha}(y)$ com f.d.p. $\frac{d(F_*^{\alpha}(y))}{dy}$, no cenário de riscos competitivos latentes com a distribuição

Geométrica, a f.d.p. é dada por $f(y) = \dfrac{\theta\alpha f_*(y)F_*^{\alpha-1}(y)}{[1-(1-\theta)(1-F_*^{\alpha}(y))]^2}$ com função de sobrevivência

$$S(y) = \frac{\theta}{1-(1-\theta)(1-F_*^{\alpha}(y))}(1-F_*^{\alpha}(y)) = \theta_y(1-F_*^{\alpha}(y)), \tag{2.2.5}$$

em que $\theta_y = \frac{\theta}{1-(1-\theta)(1-F_*^{\alpha}(y))}$.

Na equação (2.2.5), temos $\theta_y = 1$ se $y = 0$ e $\theta_y = \theta$ quando $y \to \infty$. Desta forma, com a interpretação que a função de distribuição da Geométrica apresenta nesta construção, quando $y \to \infty$, certamente todos os "M" fatores de risco vão apresentar falha, e podemos então estimar $\hat{M} = 1/\theta$ (Esperança da distribuição Geométrica com parâmetro θ). Agora, θ_y poderia se utilizado para estimar o número de fatores de risco que falharam até o instante de tempo y^* pela fórmula $\hat{M}_{y^*} = 1/\theta_{y^*}$.

Para a função de risco, temos que

$$h(y) = \frac{\frac{d(F_*^{\alpha}(y))}{dy}}{(1-F_*^{\alpha}(y))}\frac{\theta_y}{\theta} = H_{F_*^{\alpha}}H_{\theta}, \tag{2.2.6}$$

em que $H_{F_*^{\alpha}}$ é a função de risco da distribuição $F_*(y)$ exponenciada (F_*^{α}) e $H_{\theta} = \theta_y/\theta$.

Observe que, $H_{\theta} = 1/\theta$ se $y = 0$ e $H_{\theta} = 1$ quando $y \to \infty$. Quando $\theta \to 0$, se $H_{F^{\alpha}} \to 0$ quando $y \to 0$ e $H_{F^{\alpha}} \to c$ quando $y \to \infty$, é fácil observar que $h(y)$ possui forma unimodal.

2.2.1 Formas da função de risco

A distribuição E2G possui forma de risco unimodal, crescente ou decrescente dependendo dos valores dos parâmetros de θ e α, como pode ser visto na parte inferior da Figura 2.2.

Pelo motivo da função de risco em (2.2.3) ser muito complexa, as formas da função de risco são obtidas numericamente. O que de fato, foi verificado numericamente, usando o *software* Maple MapleSoft (2012) e as condições do Teorema de Glaser (Glaser, 1980). Desta forma, consideramos os vários pontos da região paramétrica que caracteriza a forma da função de risco, escolhendo aqueles que caracterizam a forma desejada e então estudamos suas propriedades numericamente.

Seja $\eta(y) = -f'(y)/f(y)$ em que $f'(y)$ é a primeira derivada da f.d.p. (2.2.1). Portanto,

$$\eta(y) = \lambda - \frac{\lambda e^{-\lambda x}(\alpha-1)}{1-e^{-\lambda x}} + 2\frac{\alpha\lambda e^{-\lambda x}(1-\theta)(1-e^{-\lambda x})^{\alpha-1}}{1-(1-\theta)(1-e^{-\lambda x})}.$$

A primeira derivada de $\eta(y)$ é dada por

$$\begin{aligned}\eta'(y) &= \frac{(\alpha-1)\,\lambda^2 \mathrm{e}^{-\lambda x}}{1-e^{-\lambda x}} + \frac{(\alpha-1)\,\lambda^2 e^{-2\lambda x}}{(1-e^{-\lambda x})^2} - 2\frac{\alpha\lambda^2 e^{-\lambda x}\,(1-\theta)\left(1-e^{-\lambda x}\right)^{\alpha-1}}{1-(1-\theta)\,(1-(1-e^{-\lambda x})^{\alpha})} \\ &+ 2\frac{\alpha(\alpha-1)\lambda^2 e^{-2\lambda x}\,(1-\theta)\left(1-e^{-\lambda x}\right)^{\alpha-2}}{1-(1-\theta)\,(1-(1-e^{-\lambda x})^{\alpha})} - 2\frac{\alpha^2\lambda^2\left(e^{-\lambda x}\right)^2(1-\theta)^2\left(1-e^{-\lambda x}\right)^{2(\alpha-1)}}{(1-(1-\theta)\,(1-(1-e^{-\lambda x})^{\alpha}))^2}.\end{aligned} \tag{2.2.7}$$

Se $0 < a \leq 1$, $\eta'(y) < 0$ para $y > 0$, pelo Teorema de Glaser (Glaser, 1980) concluímos diretamente que a forma da função de risco (2.2.3) é decrescente.

Para verificar que a função de risco pode ser crescente, por propriedade, nós consideramos o ponto do espaço paramétrico $(\alpha, \lambda, \theta) = (10, 1, 0.99)$. Portanto é necessário verificar que $\eta'(t) > 0$ para $t > 0$, o que de fato ocorre. De Glaser (1980), podemos concluir que a função de risco é crescente. Numericamente, observamos que para valores de $\theta > 0.5$ a função de risco possui a forma crescente. Para exemplo, as combinações de (α, θ) próximos da região vizinha e iguais a $(10, 0.55)$, $(7.5, 0.57)$. $(6, 0.59)$, $(5, 0.6)$, $(3, 0.69)$, $(2.5, 0.8)$ e $(1.125, 0.9)$ possuem a forma da função de risco crescente.

Para verificar a forma unimodal da função de risco, consideramos o ponto $(\alpha, \lambda, \theta) = (10, 1, 0.01)$ e obtemos o ponto $t_0 = 1.3398$, de tal forma que $\eta'(t_0) = 0$, $\eta'(t) > 0$ para $t \in (0, t_0)$ (por exemplo $\eta'(0.1) = 899.2505$), e $\eta'(t) < 0$ para $t \in (t_0, \infty)$ (por examplo $\eta'(2) = -1.6464$) e consequentemente $\lim_{y\to 0} f(y) = 0$. Portanto, podemos concluir que a função de risco é unimodal.

2.2.2 Momentos, momentos de vida residual e momentos da estatística de ordem

Algumas características e particularidades de uma distribuição podem ser estudadas por meio dos momentos, como por exemplo a variância. As expressões matemáticas da esperança, variância e o $r-$ésimo momento em torno da origem de Y podem ser obtidos usando a expressão:

$$E[Y^r] = r\int_0^{\infty} y^{r-1} S(y) dy. \tag{2.2.8}$$

A expressão geral para o r-ésimo momento $\mu_r^{'} = E(Y^r)$ da variável Y, com f.d.p. dada por (2.2.1) pode ser obtida analiticamente considerando a expansão binomial em série, dada por

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \frac{(r)_k}{k!} x^k, \tag{2.2.9}$$

em que $(r)_k$ é o símbolo Pochhammer, dado por $(r)_k = r(r+1)\cdots(r+k-1)$ e se $|x| < 1$ a série converge, e ainda

$$(-r)_k = (-1)^k (r-k+1)_k. \tag{2.2.10}$$

Proposição 2.2.1 Para a variável aleatória Y com distribuição E2G, temos que, a função para o r-ésimo momento é dada por

$$\mu_r^{'} = \frac{\theta r!}{\lambda^r} \sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \frac{(-1)^{l+m}(j-l+2)_l(\alpha l+k-m+1)_m(1-\theta)^j}{l!m!(m+1)^r}.$$

Prova 2.2.1 De (2.2.2) e (2.2.8), usando a expressão (2.2.9), obtemos

$$\begin{aligned}
\mu_r^{'} =& r\int_0^{\infty} y^{r-1}S(y)dy \\
=& r\theta\int_0^{\infty} y^{r-1}\frac{(1-(1-e^{-\lambda y})^{\alpha})}{1-(1-\theta)(1-(1-e^{-\lambda y})^{\alpha})}dy \\
=& r\theta\int_0^{1}\left(-\frac{\ln(1-z)}{\lambda}\right)^{r-1}\frac{(1-z^{\alpha})}{[1-(1-\theta)(1-z^{\alpha})](1-z)\lambda}dz \\
=& \frac{r\theta(-1)^{r-1}}{\lambda^r}\sum_{k=0}^{\infty}\frac{(1)_k}{k!}\int_0^{1}(\ln(1-z))^{r-1}\frac{z^k(1-z^{\alpha})}{1-(1-\theta)(1-z^{\alpha})}dz \\
=& \frac{r\theta(-1)^{r-1}}{\lambda^r}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\frac{(1)_j}{j!}(1-\theta)^j\int_0^{1}(\ln(1-z))^{r-1}z^k(1-z^{\alpha})^{j+1}dz \\
=& \frac{r\theta(-1)^{r-1}}{\lambda^r}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\frac{(-(j+1))_l}{l!}(1-\theta)^j\int_0^{1}(\ln(1-z))^{r-1}z^{\alpha l+k}dz \\
=& \frac{r\theta}{\lambda^r}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\frac{(-(j+1))_l}{l!}(1-\theta)^j\int_0^{\infty}u^{r-1}(1-e^{-u})^{\alpha l+k}e^{-u}du \\
=& \frac{r\theta}{\lambda^r}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}\frac{(-(j+1))_l}{l!}\frac{(-(\alpha l+k))_m}{m!}(1-\theta)^j\int_0^{\infty}u^{r-1}e^{-u(m+1)}du
\end{aligned}$$

em que a última igualdade segue da integração da função Gama e da relação em (2.2.10), o que completa a prova.

■

A estatística de ordem, entre outras, são ferramentas fundamentais para inferência, principalmente no setor industrial e de saúde. Seja, $Y_1, \ldots, Y_n$ uma amostra aleatória independente da distribuição E2G e $Y_{1:n}, \ldots, Y_{n:n}$ denotando a correspondente estatística de ordem. Então, a f.d.p. $f_{i:n}(y)$ da i-ésima estatística de ordem $Y_{i:n}$ é dada por

$$f_{i:n}(y) = \frac{n!}{(i-1)!(n-i)!}F(y)^{i-1}(1-F(y))^{n-i}f(y).$$

O r-ésimo momento da i-ésima estatística de ordem $Y_{i:n}$, de acordo com (Barakat e Abdelkader, 2004), pode ser representada por

$$E\left[Y_{i:n}^{r}\right] = r\sum_{p=n-i+1}^{n}(-1)^{p-n+i-1}\binom{p-1}{n-i}\binom{n}{p}\int_{0}^{\infty}y^{r-1}\left[S(y)\right]^{p}dy. \tag{2.2.11}$$

Proposição 2.2.2 Para a variável aleatória Y com distribuição E2G, temos que, o r-ésimo momento da i-ésima estatística de ordem é dado por

$$\begin{aligned} E\left[Y_{i:n}^{r}\right] = & \frac{r!}{\lambda^{r}}\sum_{p=n-i+1}^{n}\sum_{l=0}^{j+p}\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}(-1)^{p-n+l+m+i+2r-3}\binom{p-1}{n-i}\theta^{p} \\ \times & \binom{n}{p}\frac{(p)_{j}(p+j-l+1)_{l}(1-\theta)^{j}(\alpha l+k-m+1)_{m}}{j!l!m!(m+1)^{r}}. \end{aligned}$$

Prova 2.2.2 Da equação (2.2.11), usando as expressões (2.2.2) e (2.2.9), e procedendo de maneira similar à prova da Proposição 2.2.1, obtém-se o resultado. ■

Dado que não houve falha até o momento t, o a distribuição do tempo de vida residual de uma variável aleatória Y com distribuição E2G, tem a função de sobrevivência dada por

$$S_{t}(y) = \Pr[Y > y+t|Y > t] = \left(\frac{1-(1-e^{-\lambda(y+t)})^{\alpha}}{1-(1-e^{-\lambda y})^{\alpha}}\right)\left(\frac{1-(1-\theta)(1-(1-e^{-\lambda y})^{\alpha})}{1-(1-\theta)(1-(1-e^{-\lambda(y+t)})^{\alpha})}\right).$$

A função média do tempo de vida residual de uma função contínua com função de sobrevivência $S(y)$ é dada por

$$\mu(t) = E(Y-t|Y>t) = \frac{1}{S(t)}\int_{t}^{\infty}S(u)du. \tag{2.2.12}$$

Proposição 2.2.3 Para a variável aleatória Y com distribuição E2G, temos que, a média do tempo de vida residual é dada por

$$\mu(t) = \frac{1}{\lambda}\left(\frac{1-(1-\theta)(1-(1-e^{-\lambda t})^{\alpha})}{1-(1-e^{-\lambda t})^{\alpha}}\right)\sum_{k=0}^{\infty}\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\frac{(1-\theta)^{i}(-1)^{j}(i-j+2)_{j}}{j!}\left(\frac{1-(1-e^{\lambda t})^{k+\alpha j+1}}{k+\alpha j+1}\right).$$

Prova 2.2.3 Substituindo $S(y)$ dada por (2.2.2) na equação (2.2.12), obtemos

$$\begin{aligned} \frac{1}{S(t)}\int_{t}^{\infty}S(u)du &= \frac{1-(1-\theta)(1-(1-e^{-\lambda t})^{\alpha})}{1-(1-e^{-\lambda t})^{\alpha}}\int_{t}^{\infty}\frac{1-(1-e^{-\lambda u})^{\alpha}}{1-(1-\theta)(1-(1-e^{-\lambda u})^{\alpha})}du \\ &= \frac{1}{\lambda}\frac{1-(1-\theta)(1-(1-e^{-\lambda t})^{\alpha})}{1-(1-e^{-\lambda t})^{\alpha}}\int_{1-e^{-\lambda t}}^{1}\frac{1-x^{\alpha}}{(1-(1-\theta)(1-x^{\alpha}))(1-x)}dx. \end{aligned}$$

Usando agora (2.2.9) e procedendo de maneira similar a prova da Proposição 2.2.1, obtém-se o resultado. ■

2.2.3 Inferência

A seguir consideramos o método de máxima verossimilhança para a estimação dos parâmetros e a matriz de Informação observada para a obtenção do erro padrão assintótico dos estimadores. Assumindo que os tempos de vida observados na amostra são independentes e identicamente distribuídos e são independentes do mecanismo de censura, os estimadores de máxima verossimilhança (MLE) dos parâmetros são obtidos diretamente da maximização da função log-verossimilhança $\ell(.)$, dada por

$$\begin{aligned}\ell(\theta,\lambda,\alpha) &= n\ln(\theta)+\ln(\alpha\lambda)\sum_{i=1}^{n}\delta_i-\lambda\sum_{i=1}^{n}\delta_i y_i+(\alpha-1)\sum_{i=1}^{n}\delta_i\ln(1-e^{-\lambda y_i})+ \qquad (2.2.13)\\ &+ \sum_{i=1}^{n}(1-\delta_i)\ln\left(1-(1-e^{-\lambda y_i})^{\alpha}\right)-\sum_{i=1}^{n}(1+\delta_i)\ln\left(1-(1-\theta)(1-(1-e^{-\lambda y_i})^{\alpha})\right),\end{aligned}$$

em que δ_i é o indicador de censura, que é igual a 0 ou 1 se a observação é censurada ou observada, respectivamente.

Os estimados MLEs, denotados por $\hat{\alpha}$, $\hat{\lambda}$ e $\hat{\theta}$, são as soluções simultâneas das seguintes equações, as quais serão resolvidas considerando a rotina *optim* do *Software* R Core Team (2012),

$$\frac{\sum_{i=1}^{n}\delta_i}{\hat{\alpha}}+\sum_{i=1}^{n}\delta_i\ln(1-e^{-\hat{\lambda}y_i})=\sum_{i=1}^{n}\frac{L_i^{\hat{\alpha}}\ln(L_i)(1-\delta_i)}{R_i}+\sum_{i=1}^{n}\frac{(1+\delta_i)(1-\hat{\theta})L_i^{\hat{\alpha}}\ln(L_i)}{T_i},$$

$$\frac{\sum_{i=1}^{n}\delta_i}{\hat{\lambda}}+(\hat{\alpha}-1)\sum_{i=1}^{n}\frac{\delta_i X_i}{L_i}=\sum_{i=1}^{n}\delta_i y_i+\sum_{i=1}^{n}\frac{\hat{\alpha}(1-\delta_i)L_i^{\hat{\alpha}}X_i}{L_i R_i}+\sum_{i=1}^{n}\frac{\hat{\alpha}(1+\delta_i)(1-\hat{\theta})L_i^{\hat{\alpha}}X_i}{L_i T_i}$$

e

$$\frac{n}{\hat{\theta}}=\sum_{i=1}^{n}\frac{(1+\delta_i)R_i}{T_i},$$

em que $L_i=1-e^{-\hat{\lambda}y_i}$, $R_i=1-L_i^{\hat{\alpha}}$, $T_i=1-(1-\hat{\theta})R_i$, e $X_i=y_i e^{-\hat{\lambda}y_i}$.

Visto que a matriz de informação não possui forma fechada, vamos considerar a matriz de informação observada para a estimação dos intervalos de confiança do vetor (α,θ,λ). Os elementos da matriz de informação observada são dados por

$$I_F(\alpha,\theta,\lambda)=\left.\begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\theta} & I_{\alpha\lambda}\\ I_{\theta\alpha} & I_{\theta\theta} & I_{\theta\lambda}\\ I_{\lambda\alpha} & I_{\lambda\theta} & I_{\lambda\lambda}\end{pmatrix}\right|_{(\alpha,\theta,\lambda)=(\hat{\alpha},\hat{\theta},\hat{\lambda})}, \qquad (2.2.14)$$

em que os elementos da matriz $I_F(\alpha,\theta,\lambda)$ estão apresentados no Apêndice A.

Sob certas condições de regularidade para a função de verossimilhança e com o tamanho da amostra grande, a distribuição assintótica para $(\hat{\alpha}, \hat{\theta}, \hat{\lambda})$, é uma distribuição Normal 3-variada com média zero e matriz de variância e covariância $I_F^{-1}(\alpha, \theta, \lambda)$.

Performance dos MLEs

O estudo a seguir é baseado na geração de 1,000 amostras com tamanhos $n = 20$, 30, 50, 100 e 200 da distribuição E2G com quatro diferentes vetores de parâmetros. Uma vez que os parâmetros não estão limitados, consideramos as restrições dos parâmetros por meio de transformações. Para o parâmetro θ foi considerado a transformação $\theta = e^{\theta^*}/(1 + e^{\theta^*})$, em que $\theta^* \in \mathbb{R}$, e para α e λ foi considerado a transformação exponencial. Para o cálculo de suas variâncias usamos o método Delta. Os valores iniciais para todos os casos foram considerados de forma a partir do ponto zero da reta real, ou seja, $(\alpha, \lambda, \theta) = (1, 1, 0.5)$.

Os resultados obtidos estão condensados na Tabela 2.2, a qual mostra a média dos 1,000 MLEs $(\mathrm{Av}(\hat{\alpha}, \hat{\lambda}, \hat{\theta}))$, juntamente com a cobertura para intervalos assintóticos de 95% de confiança $(\mathrm{C}(\alpha, \lambda, \theta))$, e também suas variâncias $\mathrm{Var}(\hat{\alpha})$, $\mathrm{Var}(\hat{\lambda})$ e $\mathrm{Var}(\hat{\theta})$. Os resultados sugerem que os MLEs se comportam adequadamente com a teoria. É importante notar que para valores pequenos de θ, veja $\theta = 0.25$, a estimação se torna próxima somente para amostras de tamanho $n = 200$ e sua cobertura ainda se encontra abaixo do valor especificado de 95% e a Variância dos estimadores MLE de λ só são consideráveis aceitáveis em amostras de tamanho $n = 200$, porém em gereal, as variâncias dos MLEs decrescem quando o tamanho da amostra aumenta e a probabilidade de cobertura Empírica ficam próximas ao valor nominal escolhido, principalmente quando o tamanho da amostra aumenta.

2.2.4 Aplicação

Para verificar que a distribuição E2G pode se uma distribuição competitiva do ponto de vista de ajuste, comparamos o seu ajuste com algumas distribuições de tempo de vida usuais em três conjuntos de dados extraídos da literatura. Para a escolha do conjunto de dados foi designado que um dos conjuntos apresente forma crescente, outro com forma unimodal e um terceiro com forma decrescente para a função de risco.

As seguintes distribuições de tempo de vida foram consideradas para o ajustes seguindo todas os mesmos procedimentos de estimação: a função de distribuição Exponencial, com f.d.p. dada por $f(x) = \lambda e^{-\lambda x}$, a função de distribuição de Weibull com f.d.p. dada por $f(x) = \frac{\theta}{\lambda}\left(\frac{x}{\lambda}\right)^{\theta-1} e^{-(x/\lambda)^{\theta}}$, a função de distribuição Gama com f.d.p. dada por $f(x) = \frac{1}{\lambda^{\theta}\Gamma(\theta)} x^{\theta-1} e^{-x/\lambda}$, a função distribuição EG (Adamidis e Loukas, 1998) com f.d.p. dada por $f(x) = \lambda(1 - (1 -$

Tabela 2.2: Média dos MLES, suas coberturas e variâncias da simulação da distribuição E2G para dados simulados.

n	$(\alpha, \lambda, \theta)$	Av$(\hat{\alpha}, \hat{\lambda}, \hat{\theta})$	C$(\alpha, \lambda, \theta)$	Informação Observada		
				Var$(\hat{\alpha})$	Var$(\hat{\lambda})$	Var$(\hat{\theta})$
20	(2.0,2.0,0.6)	(2.2344,2.0555,0.7837)	(0.9070,0.9960,0.9990)	1.2235	1.4662	1.3260
	(0.5,4.0,0.25)	(0.4915,6.5998,0.5500)	(0.9540,0.9360,0.6840)	0.0578	33.541	0.4829
	(2.0,1.5,0.75)	(2.4027,1.4952,0.8172)	(0.8790,0.9950,0.9990)	1.4991	0.7892	1.3336
	(4.0,0.5,0.8)	(4.7712,0.4817,0.8353)	(0.8730,0.9850,0.9990)	7.7977	0.0652	1.8246
30	(2.0,2.0,0.6)	(2.1794,2.0616,0.7702)	(0.9170,0.9920,0.9990)	0.8647	1.2180	0.9342
	(0.5,4.0,0.25)	(0.4792,5.7292,0.4960)	(0.9480,0.9460,0.7130)	0.0522	22.216	0.3467
	(2.0,1.5,0.75)	(2.2323,1.4664,0.8005)	(0.9020,0.9950,0.9990)	0.6987	0.4571	0.8887
	(4.0,0.5,0.8)	(4.5377,0.4901,0.8580)	(0.8850,0.9780,0.9980)	4.7082	0.0744	2.0561
50	(2.0,2.0,0.6)	(2.0611,2.0498,0.7608)	(0.9400,0.9880,0.9930)	0.5847	0.9912	0.7268
	(0.5,4.0,0.25)	(0.4889,5.3752,0.4205)	(0.9300,0.9290,0.7840)	0.0259	12.651	0.1658
	(2.0,1.5,0.75)	(2.1296,1.4718,0.8049)	(0.9290,0.9830,0.9970)	0.4093	0.2968	0.5931
	(4.0,0.5,0.8)	(4.2875,0.4892,0.8533)	(0.9190,0.9820,0.9990)	1.7010	0.0280	0.7649
100	(2.0,2.0,0.6)	(2.0135,2.0062,0.7071)	(0.9640,0.9880,0.9470)	0.2054	0.3932	0.3227
	(0.5,4.0,0.25)	(0.4974,4.5422,0.3271)	(0.9370,0.9580,0.8660)	0.0155	6.4245	0.0610
	(2.0,1.5,0.75)	(2.0466,1.4750,0.7988)	(0.9700,0.9760,0.9940)	0.2549	0.2090	0.4467
	(4.0,0.5,0.8)	(3.8986,0.4609,0.7550)	(0.9760,0.9620,0.9920)	1.0539	0.0231	0.4404
200	(2.0,2.0,0.6)	(2.0022,2.0127,0.6765)	(0.9660,0.9770,0.9150)	0.1024	0.1988	0.1572
	(0.5,4.0,0.25)	(0.4958,4.2346,0.2895)	(0.9600,0.9850,0.9130)	0.0067	2.8561	0.0291
	(2.0,1.5,0.75)	(2.0421,1.4910,0.7828)	(0.9640,0.9770,0.9930)	0.1147	0.0998	0.1859
	(4.0,0.5,0.8)	(4.0063,0.4963,0.8572)	(0.9730,0.9860,0.9960)	0.4365	0.0098	0.2766

$\theta)e^{-\lambda x})^{-1}$, a função de distribuição Weibull Modificada (MW) (Lai et al., 2003) com f.d.p. dada por $f(x) = \alpha x^{\theta-1}(\theta + \lambda x)e^{\lambda x}e^{-\alpha x^{\theta}\exp\{\lambda x\}}$ e a função distribuição Exponencial Generalizada-Poisson (GEP) (Barreto-Souza e Cribari-Neto, 2009) com f.d.p. dada por $f(x) = \frac{\alpha\beta\lambda}{(1-e^{-\lambda})^{\alpha}}$ $(1 - e^{-\lambda+\lambda\exp(-\beta x)})^{\alpha-1}\ e^{-\lambda-\beta x+\lambda\exp(-\beta x)}$.

O primeiro conjunto de dados, $T1$, consiste em 65 tempos de vida observados em paciente com câncer de mama tratados no período de 1929 à 1938, disponibilizados em Boag (1949).

O segundo conjunto de dados, $T2$, consiste em tempos de vida observados de pacientes com câncer do ducto biliar, o qual fazem parte do tratamento o estudo de qual combinação do tratamento com radiação (R0Rx) e a droga 5-fluorouracil (5-FU) prolonga o tempo de vida. O conjunto de dados foi extraído de Fleming et al. (1980) e consiste no tempo de vida, em dias, para o grupo de controle.

O terceiro conjunto de dados, $T3$, consiste em números de sucessíveis falhas do sistema de ar condicionado de cada membro em uma frota de 13 Aeronaves Boeings 720. O conjunto combinado possui 214 observações e foram consideradas em Adamidis e Loukas (1998). Este conjunto de dados foi analisado primeiramente em Proschan (1963) e discutido posteriormente em Dahiya e Gurland (1972), Gleser (1989), Kus (2007) e Barreto-Souza et al. (2011).

Primeiramente, para identificar a forma da função de risco vamos considerar o método baseado na curva do gráfico TTT de Aarset (1987) já apresentado no Capítulo 1. Pelas imagens à esquerda da Figura 2.2.4, podemos identificar que o conjunto de dados T1 possui forma crescente, o conjunto T2 possui forma unimodal e o conjunto T3 possui forma decrescente para a função de risco.

Na Tabela 2.3 disponibilizamos os valores de $-\max\ell(.)$ e AIC para todas as distribuições. Ambos os critérios oferecem evidências a favor da distribuição E2G para T2 e T3, e para T1 os valores são similares para as distribuições E2G, Weibull e Gama. Estes fatos indicam que a distribuição E2G pode ser uma distribuição de interesses práticos competitiva na análise de sobrevivência. O lado direito da Figura 2.2.4 mostra as curvas de sobrevivência ajustadas e sobrepostas com o estimador K-M.

Os estimadores MLEs (e seus correspondentes desvios padrões em parênteses) do parâmetro α, θ e $\lambda(\times 1000)$ da distribuição E2G são, 2.012(0.719), 0.582(0.551) e 41.11(9.98), respectivamente para $T1$, 1.978(2.377), 0.021(0.0489) e 0.85(0.188) para $T2$, e 1.238(0.273), 0.314(0.141) e 6.90(0.72) para $T3$.

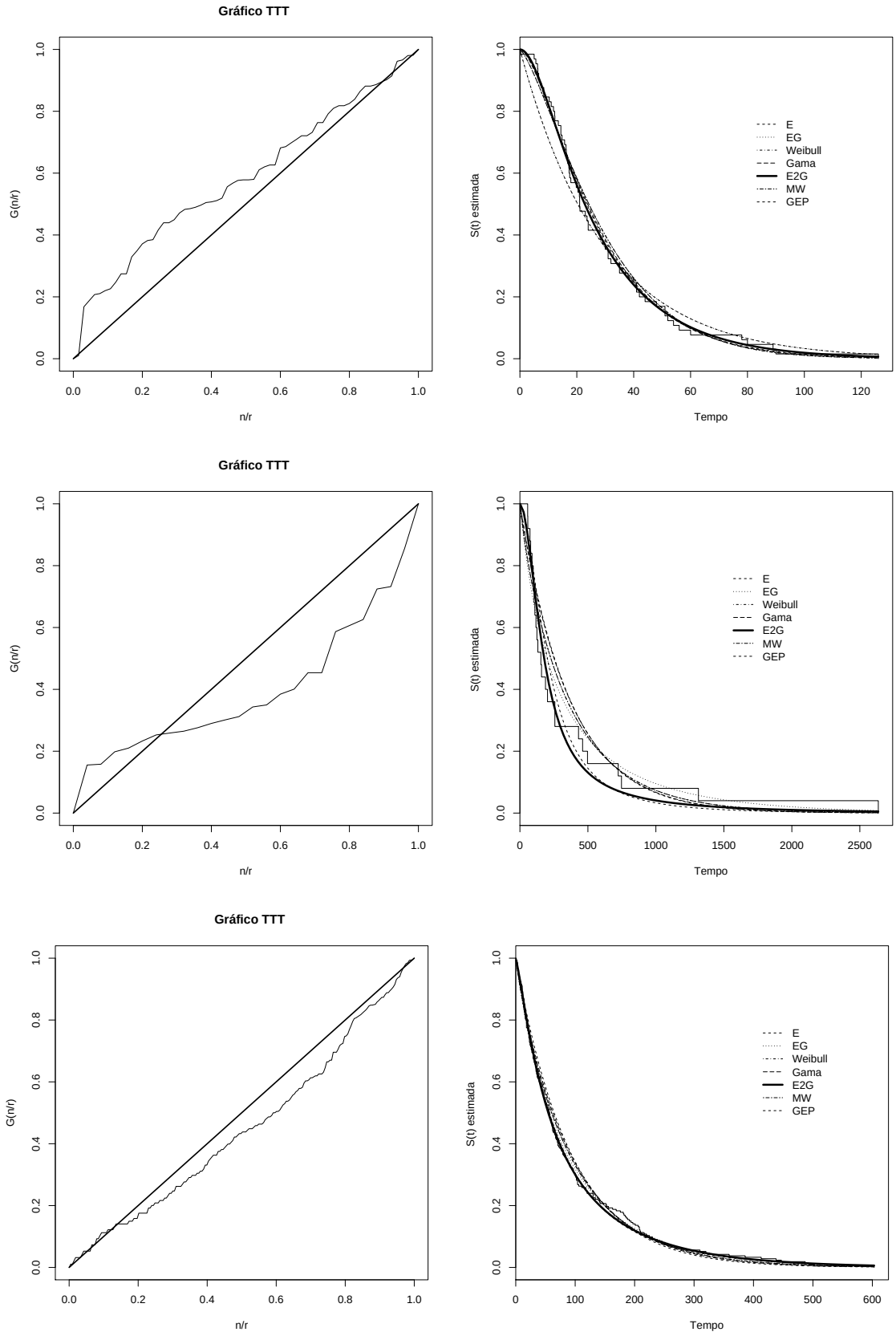

Figura 2.2: Esquerda: Gráfico TTT empírico. Direita: K-M e Curvas de sobrevivência ajustadas. Superior: Dados de $T1$. Meio: Dados de $T2$. Inferior: Dados de $T3$.

Tabela 2.3: Valores de $-\max \ell(.)$ e AIC para os dados T1, T2 e T3.

		E	EG	Weibull	Gama	E2G	MW	GEP
T1	$-\max \ell(.)$	284.6012	284.6012	280.6973	280.0386	279.7394	280.6974	279.9055
	AIC	571.2024	573.2024	565.3946	564.0772	565.4788	567.3947	565.8110
T2	$-\max \ell(.)$	172.6231	171.2032	172.2865	172.6161	167.3942	172.2865	168.7415
	AIC	347.2463	346.4064	348.5730	349.2322	340.7883	350.5731	343.4831
T3	$-\max \ell(.)$	1178.766	1175.925	1177.585	1178.291	1174.260	1177.585	1174.785
	AIC	2359.532	2355.849	2359.170	2360.582	2354.520	2361.170	2355.570

2.3 Distribuição CE2G

A seguir apresentamos a distribuição CE2G que é a composição da distribuição Exponencial exponenciada com a distribuição Geométrica no cenário de riscos complementares (latentes), o que origina a primeira letra da nomenclatura "C". Após apresentar a distribuição, detalharemos a motivação de sua construção. É importante notar aqui, que a distribuição apresentada E2G foi derivada com riscos competitivos, ou seja, a ativação pelo primeiro fator de risco e a distribuição CE2G será derivada pelos riscos complementares que é por meio da falha do último fator de risco.

Seja Y uma variável aleatória não negativa, representando o tempo de vida de um componente em uma determinada população. A variável aleatória Y tem distribuição CE2G, com parâmetros $\lambda > 0$, $\alpha > 0$ e $0 < \theta < 1$ se sua f.d.p. é dada por

$$f(y) = \frac{\alpha\lambda\theta e^{-\lambda y}(1-e^{-\lambda y})^{\alpha-1}}{\left[1-(1-\theta)(1-e^{-\lambda y})^{\alpha}\right]^2}, y > 0, \tag{2.3.1}$$

em que λ é o parâmetro de escala, α e β são parâmetros de forma. A Figura 2.3 (parte superior) mostra a f.d.p. da distribuição CE2G para $\lambda = 1$, $\theta = 0.05, 0.5, 0.95$ e $\alpha = 0.3, 1.0, 3$. Na Figura 2.3 podemos notar que a f.d.p. pode ser decrescente ou unimodal.

A função de sobrevivência da distribuição CE2G é dada por

$$S(y) = \frac{1-(1-e^{-\lambda y})^{\alpha}}{1-(1-\theta)(1-e^{-\lambda y})^{\alpha}}, y > 0, \tag{2.3.2}$$

em que, $\alpha > 0$, $\theta \in (0,1)$ e $\lambda > 0$.

De (2.3.2) e (2.3.1), a função de risco de acordo com a relação $h(y) = \frac{f(y)}{S(y)}$, é dada por

$$h(y) = \frac{\alpha\lambda\theta e^{-\lambda y}(1-e^{-\lambda y})^{\alpha-1}}{[1-(1-e^{-\lambda y})^{\alpha}[[1-(1-\theta)\,(1-e^{-\lambda y})^{\alpha}]}. \tag{2.3.3}$$

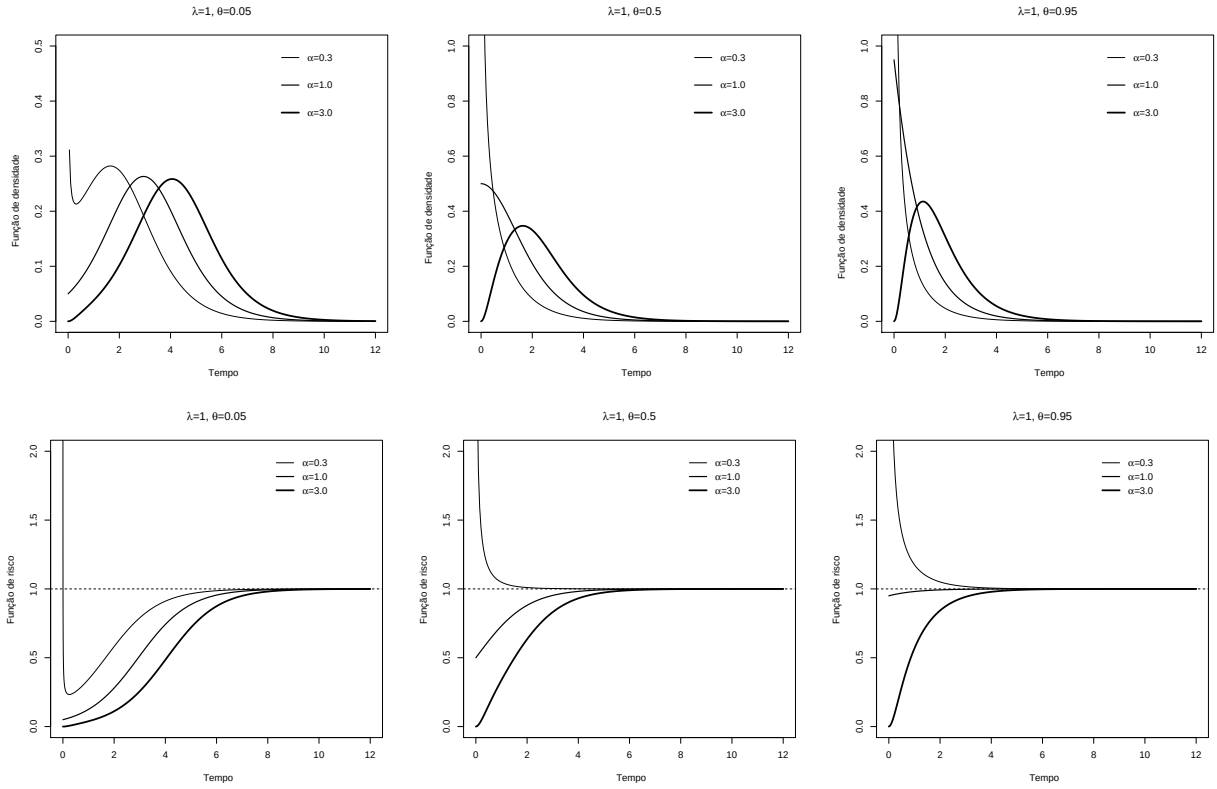

Figura 2.3: Superior: Função densidade de probabilidade da distribuição CE2G. Inferior: Função de risco da distribuição CE2G.

Temos que a função de risco $h(y)$ em (2.3.3) não é finita se $\alpha < 1$ no valor $y = 0$ e é dada por $h(0) = \lambda\theta$ se $\alpha = 1$ ou $h(0) = 0$ se $\alpha > 1$. Temos ainda que quando $y \to \infty$, $h(\infty) = \lambda$. A função de risco dada em (2.3.3) pode ser crescente, decrescente ou em forma de banheira, como observado na Figura 2.3 (parte inferior), que mostra a função de risco para valores de $\lambda = 1$, $\theta = 0.05, 0.5, 0.95$ e $\alpha = 0.3, 1.0, 3$.

A função quantil da distribuição CE2G, é dada por

$$Q(p) = F^{-1}(p) = -\frac{\ln(1 - (\frac{p}{\theta(1-p)+p})^{1/\alpha})}{\lambda}, \tag{2.3.4}$$

em que $F(y) = 1 - S(y)$ é a função distribuição de Y.

Considere que no estudo de confiabilidade é possível observar somente o componente de vida quando todos os componentes em risco falharem. Além disso, considere as ocasiões que a informação sobre o risco que provocou a falha do componente em análise não pode ser observada. Desta forma se constitui o cenário de problemas de Riscos Complementares (CR ou RC), o qual é estudado em várias áreas e possui uma vasta literatura a respeito deste assunto. Para maiores informações recomenda-se a leitura de Lawless (2003), Crowder et al. (1991) e Cox e Oakes (1984).

Neste contexto, nosso modelo é obtido da seguinte maneira. Seja M a variável aleatória denotando o número de riscos para ocasionar a falha com $m = 1, 2, ...$ e considere que M tem distribuição Geométrica dada por

$$P(M = m) = \theta(1-\theta)^{m-1}, \tag{2.3.5}$$

em que $0 < \theta < 1$ e $m = 1, 2, ...$, considere também t_i, $i = 1, 2, 3, ...$ variáveis aleatórias dos tempos de vida para cada risco i, isto é, o tempo até o evento do i-ésimo fator de risco complementar e de Gupta e Kundu (2001), como a CEG, usamos T_i com distribuição Exponencial exponenciada de parâmetros λ e α, dados por

$$f(t_i; \lambda, \alpha) = \alpha g(t_i; \lambda) G(t_i; \lambda) = \alpha\lambda \exp\{-\lambda t_i\}(1 - \exp\{-\lambda t_i\})^{\alpha-1}, \tag{2.3.6}$$

em que $g(.)$ e $G(.)$ são a f.d.p. e a função distribuição da distribuição proposta por Gupta e Kundu (2001) com parâmetro λ.

No cenário de risco complementares latentes, o número de causas M e o tempo de vida t_j associado a um risco particular não são observados, porém o tempo de vida Y sobre todos os risco é usualmente observado. Portanto, será observado somente a variável aleatória dada por

$$Y = \max\{T_1, T_2 \ldots, T_M\}. \tag{2.3.7}$$

Proposição 2.3.1 Se a variável aleatória Y é definida como em (2.3.7), então, considerando (2.3.5) e (2.3.6), Y é distribuida de acordo com a distribuição CE2G, com f.d.p. dada por (2.3.1).

Prova 2.3.1 A função de distribuição condicional de (2.3.7) dado $M = m$ é dada por $f(y|M = m, \lambda, \alpha) = m\alpha\lambda e^{-\lambda y}(1 - e^{-\lambda y})^{\alpha-1}[(1 - e^{-\lambda y})^{\alpha}]^{m-1}; t > 0,\ m = 1, \ldots.$ Portanto, a função de distribuição margina de Y é dada por

$$\begin{aligned}
f(y) &= \sum_{m=1}^{\infty} m\alpha\lambda e^{-\lambda y}(1 - e^{-\lambda y})^{\alpha-1}\left[(1 - e^{-\lambda y})^{\alpha}\right]^{m-1} \times \theta(1-\theta)^{m-1} \\
&= \theta\alpha\lambda e^{-\lambda y}(1 - e^{-\lambda y})^{\alpha-1} \sum_{m=1}^{\infty} m\left[(1 - e^{-\lambda y})^{\alpha}(1-\theta)\right]^{m-1} \\
&= \theta\alpha\lambda e^{-\lambda y}(1 - e^{-\lambda y})^{\alpha-1} \sum_{m=1}^{\infty} \frac{\left[(1 - e^{-\lambda y})^{\alpha}(1-\theta)\right]^{m-1}}{1 - (1 - e^{-\lambda y})^{\alpha}(1-\theta)} \\
&= \theta\alpha\lambda e^{-\lambda y}(1 - e^{-\lambda y})^{\alpha-1}\left[\frac{1}{1 - (1-\theta)(1 - e^{-\lambda y})^{\alpha}}\right]^2.
\end{aligned}$$

■

Uma outra maneira para demonstrar seria trabalhar diretamente com a fórmula do máximo encontrada em Marshall e Olkin (1997) utilizando a função dada em (2.3.7). Porém com o objetivo de dar mais um sentido na construção das distribuições, optamos aqui por usar o argumento da motivação de sua construção.

A distribuição CE2G também foi obtida por Bidram et al. (2012), a qual é chamada de "*New Generalized Exponential Geometric*" (NGEG), onde são apresentados resultados semelhantes aos descritos aqui.

2.3.1 Propriedades da distribuição CE2G

Muitas das características de uma distribuição podem ser estudadas por meio dos momentos, como a média, a variância, a curtose e a assimetria, por exemplo. No entanto, é interessante obter uma expressão geral para o i-ésimo momento dada pela fórmula $\mu_r^{'} = E(Y^r)$. Para a distribuição CE2G é difícil de se obter uma fórmula analítica para os momentos e portanto vamos resumir aqui somente a média e a variância, ambas utilizando a função característica.

A função geradora de momentos de uma variável aleatória Y com f.d.p. dada por (2.3.1) pode ser obtida analiticamente, se considerarmos a expressão dada em Gradshteyn e Ryzhik (2007) pela equação 1.6 na página 329:

$$\int_0^1 z^{p-1}(1-z)^{n-1}(1+bz^m)^l dz = \Gamma(n)\sum_{k=0}^{\infty}\binom{l}{k}\frac{(b)^k\Gamma(p+km)}{\Gamma(p+n+km)}. \tag{2.3.8}$$

Para qualquer valor real t, seja $\Phi_i(t)$ a função característica de Y, isto é, $\Phi_i(t) = E\left[e^{itY}\right]$, em que i denota a unidade imaginária. Com esta notação podemos estabelecer o resultado a seguir.

Proposição 2.3.2 Para uma variável aleatória Y com distribuição CE2G, temos que, sua função característica é dada por

$$\Phi_i(t) = \alpha\theta\Gamma\left(1-\frac{it}{\lambda}\right)\sum_{k=0}^{\infty}\binom{-2}{k}\frac{\Gamma\left(\alpha[k+1]\right)(\theta-1)^k}{\Gamma\left(\alpha[k+1]+1-\frac{it}{\lambda}\right)}, \tag{2.3.9}$$

em que $i = \sqrt{-1}$.

Prova 2.3.2

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

em que a última igualdade segue da mudança de variável $z = 1 - e^{-\lambda y}$.

Comparando a última integral com (2.3.8), relacionamos: $n = 1 - \frac{it}{\lambda}$, $b = \theta - 1$, $m = \alpha = p$ e $l = -2$. Desta forma, utilizando (2.3.8) com os valores encontrados a prova é finalizada.

Proposição 2.3.3 A variável aleatória Y com distribuição CE2G, tem a média e a variância dadas, respectivamente, por

$$E(Y) = \frac{\theta}{\lambda} \sum_{k=0}^{\infty} \binom{-2}{k} \frac{(\theta-1)^k}{(k+1)} \left[\Psi\left(0, \alpha[k+1]+1\right) - \Psi\left(0,1\right)\right]$$

e

$$\begin{aligned}Var(Y) &= \frac{\theta}{\lambda^2}\Bigg\{ \sum_{k=0}^{\infty} \Bigg[\binom{-2}{k} \frac{(\theta-1)^k}{(k+1)} \\ &- \Bigg(\Psi(0,1)^2 + \frac{\pi^2}{6} + \Psi(0,\alpha[k+1]+1)\Big[\Psi(0,\alpha[k+1]+1) - 2\Psi(0,1)\Big] \\ &- \Psi\left(1,\alpha[k+1]+1\right) \Bigg)\Bigg] - \theta\Big[\sum_{k=0}^{\infty} \binom{-2}{k} \frac{(\theta-1)^k}{(k+1)} \left(\Psi\left(0,\alpha[k+1]+1\right) - \Psi\left(0,1\right)\right)\Big]^2 \Bigg\},\end{aligned}$$

em que $\Psi(n,z) = \dfrac{d^{n+1}}{dz^{n+1}} \ln(\Gamma(z))$ é conhecida como a função Psi-Gama.

Prova 2.3.3 O primeiro resultado é obtido pela relação $E(Y) = \left.\frac{\Phi'_Y(t)}{i}\right|_{t=0}$. Da literatura, $E(Y^2) = \left.\frac{\Phi''_Y(t)}{i^2}\right|_{t=0}$ e $Var(Y) = E(Y^2) - [E(Y)]^2$, e portanto com um pouco de manipulação algébrica se obtém o resultado. ■

A *Skewness* é uma medida de assimetria da função de distribuição que pode assumir valores positivos ou negativos. Quantitativamente, uma medida de assimetria negativa indica que a cauda da distribuição a esquerda, na f.d.p., é mais longa que a cauda a direita e a concentração da massa da f.d.p. se encontra a direita da média. Uma medida positiva de assimetria, indica que a cauda a direita é mais longa que a cauda a esquerda e a concentração da massa da f.d.p. está a esquerda da média.

A medida de assimetria *skewness* da variável aleatória Y, denotada por γ_1, é dada pela fórmula do terceiro momento central normalizada:

$$\gamma_1 = \frac{E[(Y-\mu)^3]}{(E[(Y-\mu)^2])^{3/2}} = \frac{E(Y^3) - 3E(Y^2)E(Y) + 3E^2(Y)E(Y) - E^3(Y)}{[E(Y^2) - E^2(Y)]^{3/2}}. \tag{2.3.10}$$

A Curtose é uma medida do "pico" de uma função distribuição. De uma maneira similar ao conceito de assimetria, a curtose é um indicador da forma da função distribuição e é comum em prática é utilizada para fazer comparação com a forma da distribuição Normal padrão.

Uma medida comum da curtose, originalmente de Karl Pearson, denotada por γ_2, é baseada na versão do quarto momento e é dada por

$$\gamma_2 = \frac{E[(Y-\mu)^4]}{(E[(Y-\mu)^2])^2} = \frac{E(Y^4) - 4E(Y^3)E(Y) + 6E(Y^2)E^2(Y) - 3E^4(Y)}{[E(Y^2) - E^2(Y)]^2}. \tag{2.3.11}$$

As expressões algébricas para a curtose e assimetria da distribuição CE2G são extensas e de simples manipulação algébrica por envolver os momentos (calculados pela função característica). Devido a este fato apresentamos na Figura 2.4 a curtose (γ_2) e a *skewness* (γ_1) da distribuição CE2G para α com $\lambda = 1$, $\theta = 0.1, 0.5, 0.9$ e para θ com $\lambda = 1$, $\alpha = 0.3, 1.0, 3$.

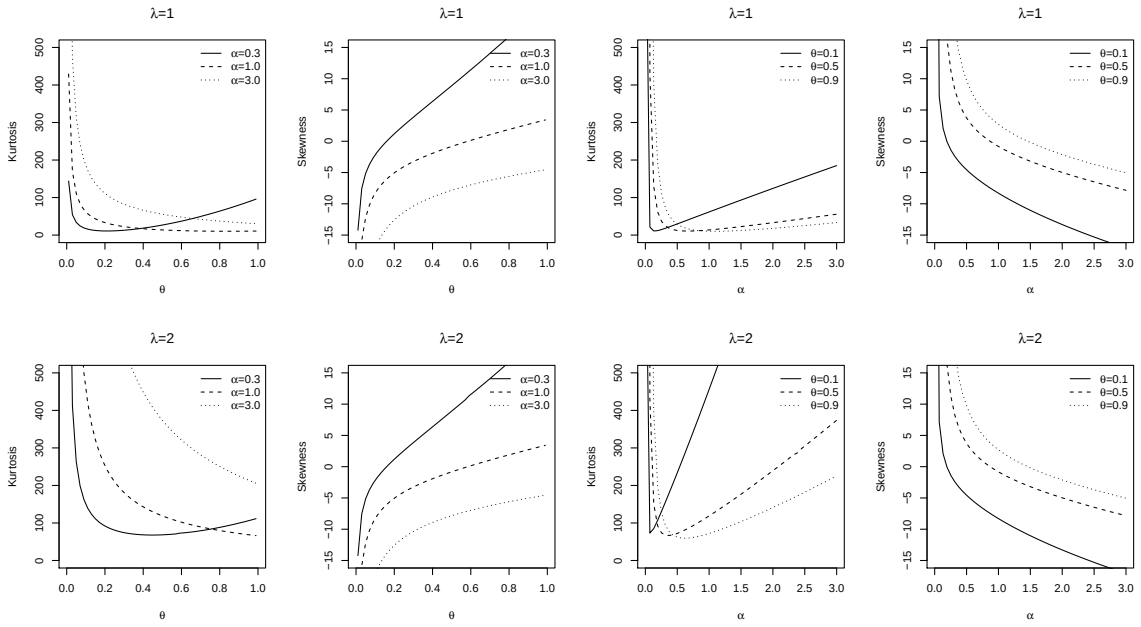

Figura 2.4: Superior: Curtose e *skewness* da distribuição CE2G para $\lambda = 1$. Inferior: Curtose e *skewness* da distribuição CE2G para $\lambda = 2$.

2.3.2 Estatísticas de ordem

Estatísticas de ordem são dentre várias, uma ferramenta fundamental na estatística não paramétrica e inferência. Seja $Y_1, \ldots, Y_n$ uma amostra aleatória de distribuições CE2G em que $Y_{1:n}, \ldots, Y_{n:n}$ são as estatísticas de ordem respectivas. Então, a f.d.p. $f_{i:n}(y)$ da i-ésima estatística de ordem $Y_{i:n}$ é dada pela relação

$$f_{i:n}(x) = \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1}(1-F(y))^{n-k} f(y).$$

O r-ésimo momento da i-ésima estatística de ordem, pode ser obtido diretamente do resultado encontrado em Barakat e Abdelkader (2004):

$$E\left[Y_{i:n}^r\right] = r \sum_{p=n-i+1}^{n} (-1)^{p-n+i-1} \binom{p-1}{n-i} \binom{n}{p} \int_0^\infty y^{r-1} \left[S(y)\right]^p dy. \tag{2.3.12}$$

Considere também a expansão binomial dada por

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \frac{(r)_k}{k!} x^k, \tag{2.3.13}$$

em que $(r)_k$ é o símbolo Pochhammer, dado por $(r)_k = r(r+1)\cdots(r+k-1)$ e se $|x| < 1$ a série em (2.3.13) converge. Considere ainda a relação do símbolo de Pochhammer dada por

$$(-r)_k = (-1)^k (r-k+1)_k. \tag{2.3.14}$$

Proposição 2.3.4 Para a variável aleatória Y com distribuição CE2G, temos que, o r-ésimo momento da i-ésima estatística de ordem é dada por

$$\begin{aligned} E\left[Y_{i:n}^r\right] = \; & \frac{r!}{\lambda^r} \sum_{p=n-i+1}^{n} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{p} \sum_{m=0}^{\infty} (-1)^{p-n+i+r+m+l-2} \binom{p-1}{n-i} \binom{n}{p} \\ \times \; & \frac{(1-\theta)^j (p)_j (p-l+1)_l (\alpha(j+l)+k-m+1)_m}{j!l!m!(m+1)^r}. \end{aligned}$$

Prova 2.3.4 De (2.3.2) e (2.3.13), temos que

$$\begin{aligned} \int_0^\infty y^{r-1} \left[S(y)\right]^p dy &= \int_0^\infty y^{r-1} \left(\frac{1-(1-e^{-\lambda y})^\alpha}{1-(1-\theta)(1-e^{-\lambda y})^\alpha} \right)^p dy \\ &= \frac{(-1)^{r-1}}{\lambda^r} \int_0^1 \frac{\ln^{r-1}(1-x)}{(1-x)} \left(\frac{1-x^\alpha}{1-(1-\theta)x^\alpha} \right)^p dx \\ &= \frac{(-1)^{r-1}}{\lambda^r} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{p} \frac{(1-\theta)^j (p)_j (-p)_l}{j!l!} \int_0^1 x^{\alpha(j+l)+k} \ln^{r-1}(1-x) dx \; (*) \\ &= \frac{(-1)^{r-1}}{\lambda^r} \\ &\quad \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{l=0}^{p} \sum_{m=0}^{\infty} \frac{(1-\theta)^j (p)_j (-p)_l}{j!l!} \frac{(-[\alpha(j+l)+k])_m}{m!} \frac{(r-1)!}{(m+1)^r}. \end{aligned} \tag{2.3.15}$$

Substituindo a equação (2.3.15) em (2.3.12) juntamente com a propriedade (2.3.14), segue o resultado.

(*) Usando a mudança de variável $\ln(1-x) = -u$ e a expansão em (2.3.13), o resultado obtido é o núcleo da distribuição Gama.

■

2.3.3 Entropia

A entropia de uma variável aleatória Y é a medida da variação da incerteza. Uma medida popular de entropia é a de Rényi (Rényi, 1961).

Seja a variável aleatória Y, então a medida de entropia de Rényi é dada por

$$\gamma(\rho) = \frac{1}{1-\rho}\log\left(\int f^{\rho}(y)dy\right), \tag{2.3.16}$$

em que $\rho > 0$ e $\rho \neq 1$.

Proposição 2.3.5 Se a variável aleatória Y tem distribuição CE2G, então a medida de entropia de Rényi é dada por

$$\gamma(\rho) = \frac{1}{1-\rho}\log\left((\alpha\theta)^{\rho}\lambda^{\rho-1}\sum_{k=0}^{\infty}\left[\frac{(1-\theta)^k(2\rho)_k\Gamma(\rho(\alpha-1)+k\alpha+1)\Gamma(\rho)}{k!\Gamma(\alpha(\rho+k)+1)}\right]\right).$$

Prova 2.3.5 De (2.3.16), obtemos que

$$\begin{aligned}
\int f^{\rho}(y)dy &= \int_0^{\infty}\frac{(\alpha\lambda\theta)^{\rho}e^{-\lambda\rho y}(1-e^{-\lambda y})^{\rho(\alpha-1)}}{\left[1-(1-\theta)(1-e^{-\lambda y})^{\alpha}\right]^{2\rho}}dy \\
&= (\alpha\lambda\theta)^{\rho}\int_0^{\infty}\sum_{k=0}^{\infty}\left[e^{-\lambda\rho y}(1-e^{-\lambda y})^{\rho(\alpha-1)+k\alpha}(1-\theta)^k\frac{(2\rho)_k}{k!}\right]dy \\
&= (\alpha\theta)^{\rho}\int_0^{\infty}\sum_{k=0}^{\infty}\left[(1-e^{-\lambda y})^{\rho(\alpha-1)+k\alpha}(1-\theta)^k\frac{(2\rho)_k}{k!}(\lambda e^{-\lambda y})^{\rho-1}\right]\lambda e^{-\lambda y}dy \\
&= (\alpha\theta)^{\rho}\lambda^{\rho-1}\sum_{k=0}^{\infty}\left[(1-\theta)^k\frac{(2\rho)_k}{k!}\int_0^{\infty}u^{\rho(\alpha-1)+k\alpha}(1-u)^{\rho-1}du\right] \\
&= (\alpha\theta)^{\rho}\lambda^{\rho-1}\sum_{k=0}^{\infty}\left[(1-\theta)^k\frac{(2\rho)_k}{k!}\frac{\Gamma(\rho(\alpha-1)+k\alpha+1)\Gamma(\rho)}{\Gamma(\alpha(\rho+k)+1)}\right]. \qquad (2.3.17)
\end{aligned}$$

Portanto, usando (2.3.17) em $\gamma(\rho)$ segue o resultado. ■

2.3.4 Confiabilidade

No contexto de confiabilidade (*reliability*), o modelo de *stress-strength* descreve a vida de um componente cuja a variável aleatória de robustez (*strength*) Y é submetida a um estresse

(*stress*) aleatório X. O componente falha no instante que o estresse aplicado exceder a robustez, tendo o componente funcionando satisfatoriamente sempre que $Y > X$. Portanto $R = P(X < Y)$ é a medida de confiabilidade do componente. Na área de modelos de *stress-strength* há uma concentração de trabalho para a estimação da confiabilidade R quando Y e X são variáveis aleatórias independentes pertencentes a mesma família de distribuição univariadas.

Proposição 2.3.6 Se a variável aleatória Y tiver distribuição CE2G, então, a confiabilidade $R = P(X < Y)$ para X e Y independentes e identicamente distribuidas (i.i.d.) é dada por

$$\theta^2 \sum_{k=0}^{\infty} \frac{(1-\theta)^k (3)_k}{k!(k+2)}.$$

Prova 2.3.6 Para X e Y i.i.d. com distribuição CE2G em que X é a variável de estresse e Y a de robustez, a confiabilidade $R = P(X < Y)$ é dada por

$$\begin{aligned}
R &= \int_0^{\infty}\int_0^{y} \frac{\alpha\lambda\theta e^{-\lambda x}(1-e^{-\lambda x})^{\alpha-1}}{[1-(1-\theta)(1-e^{-\lambda x})^{\alpha}]^2} \frac{\alpha\lambda\theta e^{-\lambda y}(1-e^{-\lambda y})^{\alpha-1}}{[1-(1-\theta)(1-e^{-\lambda y})^{\alpha}]^2} dxdy \\
&= \int_0^{\infty} \frac{\theta(1-e^{-\lambda y})^{\alpha}}{[1-(1-\theta)(1-e^{-\lambda y})^{\alpha}]} \frac{\alpha\lambda\theta e^{-\lambda y}(1-e^{-\lambda y})^{\alpha-1}}{[1-(1-\theta)(1-e^{-\lambda y})^{\alpha}]^2} dy \\
&= \sum_{k=0}^{\infty} \theta^2\alpha\lambda \frac{(3)_k}{k!}(1-\theta)^k \int_0^{\infty} (1-e^{-\lambda y})^{\alpha(k+2)-1} e^{-\lambda y} dy \\
&= \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \theta^2\alpha\lambda \frac{(3)_k(1-\alpha(k+2))_j}{k!j!}(1-\theta)^k \int_0^{\infty} e^{-\lambda(j+1)y} dy \\
&= \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \theta^2\alpha \frac{(3)_k(1-\alpha(k+2))_j}{k!j!(j+1)}(1-\theta)^k \\
&= \sum_{k=0}^{\infty} \theta^2 \frac{(3)_k}{k!(k+2)}(1-\theta)^k.
\end{aligned}$$

O que completa a prova. ■

2.3.5 Distribuição de vida residual

Dado que não ocorreu a falha até o instante de tempo t, a função distribuição do tempo de vida residual de uma variável aleatória Y, com distribuição CE2G, tem função de sobrevivência dada por

$$S_t(y) = \Pr[Y > y + t | Y > t] = \left(\frac{1-(1-e^{-\lambda(y+t)})^{\alpha}}{1-(1-e^{-\lambda t})^{\alpha}}\right)\left(\frac{1-(1-\theta)(1-e^{-\lambda t})^{\alpha}}{1-(1-\theta)(1-e^{-\lambda(y+t)})^{\alpha}}\right).$$

A média do tempo de vida residual de uma função distribuição é dada por

$$\mu(t) = E(X - t|X > t) = \frac{1}{S(t)} \int_t^\infty S(u)du. \tag{2.3.18}$$

Proposição 2.3.7 Para a variável aleatória Y com distribuição CE2G, temos que a média do tempo de vida residual é dada por

$$\mu(t) = \frac{1}{\lambda} \left(\frac{1 - (1-\theta)(1-e^{-\lambda t})^\alpha}{1 - (1 - e^{-\lambda t})^\alpha} \right) \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{1} \frac{(1-\theta)^i (-1)_j}{j!} \left(\frac{1 - (1 - e^{\lambda t})^{\alpha(i+j)+k+1}}{\alpha(i+j)+k+1} \right).$$

Prova 2.3.7 De (2.3.18) e utilizando $S(y)$ dada por (2.3.2), obtemos

$$\begin{aligned} \frac{1}{S(t)} \int_t^\infty S(u)du &= \frac{1 - (1-\theta)(1-e^{-\lambda t})^\alpha}{1 - (1 - e^{-\lambda t})^\alpha} \int_t^\infty \frac{1 - (1 - e^{-\lambda u})^\alpha}{1 - (1-\theta)(1 - e^{-\lambda u})^\alpha)} du \\ &= \frac{1}{\lambda} \frac{1 - (1-\theta)(1-e^{-\lambda t})}{1 - (1 - e^{-\lambda t})^\alpha} \int_{1-e^{-\lambda t}}^1 \frac{1 - x^\alpha}{(1 - x^\alpha(1-\theta))(1-x)} dx. \end{aligned}$$

Usando a expansão dada em (2.3.13) de maneira similar à prova da Proposição 2.3.4 e o uso da relação (2.3.14) segue o resultado. ■

2.3.6 Inferência

Assumindo que os tempos de vida do objeto em estudo são identicamente distribuídos e independentes entre si e entre o tempo de censura, os estimadores de verossimilhança MLEs dos parâmetros são obtidos diretamente da maximização da função log-verossimilhança dada por

$$\begin{aligned} \ell(\theta, \lambda, \alpha) &= \ln(\alpha\theta\lambda) \sum_{i=1}^n \delta_i - \lambda \sum_{i=1}^n \delta_i y_i + (\alpha - 1) \sum_{i=1}^n \delta_i \ln(1 - e^{-\lambda y_i}) + \\ &+ \sum_{i=1}^n (1 - \delta_i) \ln\left(1 - (1 - e^{-\lambda y_i})^\alpha\right) - \sum_{i=1}^n (1 + \delta_i) \ln\left(1 - (1-\theta)(1 - e^{-\lambda y_i})^\alpha\right), \end{aligned} \tag{2.3.19}$$

em que δ_i é o indicador da censura, o qual é igual a 0 ou 1, indicando se o dado foi censurado ou observado, respectivamente. A vantagem, como já mencionado deste método, é que este procedimento é obtido diretamente da maximização da função log-verossimilhança por meio das rotinas já implementadas em vários pacotes computacionais. Consideraremos aqui a rotina *optim* do Software (R Core Team, 2012).

A inferência para amostras com tamanho consideradas grandes dos parâmetros são baseadas nos valores dos MLEs e de seus erros padrões estimados. Para o cálculo dos desvios padrões dos estimadores de máxima verossimilhança do $(\alpha, \theta, \lambda)$ nós consideramos a matriz de

informação observada de Fisher, a qual é dada por

$$I_F(\alpha,\theta,\lambda) = \left(\begin{array}{ccc} I_{\alpha\alpha} & I_{\alpha\theta} & I_{\alpha\lambda} \\ I_{\theta\alpha} & I_{\theta\theta} & I_{\theta\lambda} \\ I_{\lambda\alpha} & I_{\lambda\theta} & I_{\lambda\lambda} \end{array} \right)\Bigg|_{(\alpha,\theta,\lambda)=(\hat{\alpha},\hat{\theta},\hat{\lambda})}, \tag{2.3.20}$$

em que os elementos da matriz $I_F(\alpha,\theta,\lambda)$ são apresentados no anexo B.

Assumindo as condições de regularidade para os parâmetros α, θ e λ no espaço paramétrico, a distribuição assintótica de $(\hat{\alpha},\hat{\theta},\hat{\lambda})$, quando $n \to \infty$, é uma Normal 3-variada com média zero e matriz de variância-covariância $I_F^{-1}(\alpha,\theta,\lambda)$.

Utilizamos aqui os critérios para comparação de modelos já existentes na literatura, como por exemplo os utilizados em Barreto-Souza et al. (2011), Adamidis et al. (2005), Barreto-Souza e Cribari-Neto (2009) e Mudholkar et al. (1995). Usamos o Critério de Informação Akaike (AIC) e o Critério de Informação Bayesiana (BIC), que são dados respectivamente, por $-2\ell(.)+2q$ e $-2\ell(.)+q\log(n)$, em que $\ell(.)$ é a log-verossimilhança calculada no valor do vetor MLEs, q é o número de parâmetros a serem estimados e n é o tamanho da amostra. Portanto o melhor modelo ajustado é aquele que apresentar o menor valor de AIC e BIC.

2.3.7 Estudo de simulação

Para verificar que os estimadores MLEs do processo de estimação apresentam as propriedades assintóticas foi feito um estudo baseado na geração de 100 amostras da distribuição CE2G com seis diferentes conjuntos de parâmetros para $n = 20$, 50, 100, 200, 500 e 1000. Para o método de maximização adotado, é necessário que nenhum parâmetro fique restrito nos valores reais, isto é, o parâmetro pode assumir valores no intervalo $(-\infty,\infty)$, e portanto é considerado transformações nos parâmetros de tal forma que, para $\theta = e^{\theta^*}/(1+e^{\theta^*})$, com $\theta^* \in \mathbb{R}$, e para α e λ são consideradas transformações exponenciais. Baseado na literatura, podemos obter diretamente os valores dos MLEs sem transformações e suas variâncias são obtidas por meio do Método Delta. Para os valores iniciais foi utilizado $(\alpha,\lambda,\theta) = (1,1,0.5)$.

Os resultados obtidos se encontram na Tabela 2.4, a qual apresenta a média dos estimadores MLEs $(\mathrm{Av}(\hat{\alpha},\hat{\lambda},\hat{\theta}))$, juntamente com a probabilidade de cobertura de 95% para o intervalo $(\mathrm{C}(\alpha,\lambda,\theta))$, o vício, o erro médio quadrático (MSE), e seus desvios padrões $(\mathrm{Sd}(\hat{\alpha},\hat{\lambda},\hat{\theta}))$. Os resultados apresentados sugerem que os estimadores MLEs apresentam as propriedades assintóticas adequadamente. Os valores dos desvios decrescem quando o tamanho da amostra aumenta, bem como os valores das probabilidades de cobertura vão se aproximando dos níveis

escolhidos conforme o tamanho da amostra aumenta. É importante notar que para alguns valores de θ, veja $\theta = 0.75$ e $\theta = 0.80$, a sua cobertura ainda se encontra abaixo do valor especificado de 95% para amostras menores que $n = 200$, no entanto, para $\theta = 0.90$, a cobertura se encontra adequada para amostra com $n = 100$ e para valores a partir de $n = 200$ o MSE parece ter decaimento proporcional constante. De modo geral, o parâmetro θ parece ser o único que não atingi o comportamento da distribuição assintótica para n menor que 200.

2.3.8 Aplicações

Considerando diferentes distribuições da literatura e utilizando três conjuntos de dados extraídos da literatura, vamos mostrar que a distribuição CE2G apresentada pode obter um bom ajuste frente outras apresentadas.

O primeiro conjunto de dados, $T4$, consiste em observações de 143 crianças contaminadas com o vírus HIV do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto entre os anos de 1986 até 2001 para a soro-reversão (Perdona e Louzada-Neto, 2010). Soro-reversão pode acontecer em crianças que nascem de mães infectadas pelo vírus HIV.

O segundo conjunto de dados, $T5$, apresenta os tempos de vida em horas de 417 lâmpadas incandescentes foscas de 40W de 110 volts extraídas em 42 semanas para o controle de qualidade (Davis, 1952). O tempo de sobrevivência, em dias, para o grupo de controle das lâmpadas são obtidos do conjunto original.

O terceiro conjunto de dados, $T6$, apresenta os tempos de sobrevivência para ratos de laboratórios, que são expostos a uma dose fixada de radiação durante o período de vida de 5 a 6 semanas. A causa da morte para cada rato foi determinada após a autópsia como uma das três possibilidades: linfoma timico (C1), sarcoma de células retículares (C2), outras causa (C3) (Hoel, 1972). Considere aqui as causas C3 para o grupo de controle.

Primeiramente, para identificar o formato da curva de risco do tempo de vida, vamos considerar o gráfico TTT de Aarset (1987). Como pode ser visto esquerda da Figura 2.3.8, as curvas TTT são côncavas para $T4$, $T5$, e $T6$, indicando assim que as funções de risco são todas crescentes.

Para comparar o ajuste da distribuição CE2G vamos utilizar distribuições comuns na literatura para a análise do tempo de vida. São elas a distribuição Exponencial, a distribuição Exponencial Exponenciada (EE), com f.d.p. dada por $f(x) = \alpha * \lambda e^{-\lambda x}(1 - e^{-\lambda x})^{\alpha-1}$, a distribuição EG (Adamidis e Loukas, 1998), a distribuição Weibull, a distribuição Gama, a distribuição Weibull Modificada (MW), a distribuição Exponencial-Poisson generalizada (GEP)

Tabela 2.4: Média dos MLEs, desvio padrão, cobertura, vício e MSE da distribuição CE2G para dados simulados.

	n	Av$(\hat{\alpha},\hat{\lambda},\hat{\theta})$	Sd$(\hat{\alpha},\hat{\lambda},\hat{\theta})$	Vício	MSE	C(α,λ,θ)
$(\alpha,\lambda,\theta)=(1.48,3.10,0.75)$	20	(1.5716,3.4497,0.7522)	(0.7890,1.1204,0.3327)	(0.0916,0.3497,0.0022)	(0.6247,1.3651,0.1096)	(0.990,0.990,0.800)
	50	(1.4902,3.4026,0.7145)	(0.4478,0.7103,0.3066)	(0.0102,0.3026,-0.0355)	(0.1987,0.5911,0.0943)	(0.990,0.990,0.860)
	100	(1.4765,3.2589,0.7233)	(0.2683,0.4964,0.2494)	(-0.0035,0.1589,-0.0267)	(0.0713,0.2692,0.0623)	(0.990,0.990,0.910)
	200	(1.4798,3.1846,0.7379)	(0.2090,0.3846,0.2176)	(-0.0002,0.0846,-0.0121)	(0.0433,0.1536,0.0470)	(0.990,0.990,0.970)
	500	(1.4725,3.1617,0.7361)	(0.1584,0.2977,0.1811)	(-0.0075,0.0617,-0.0139)	(0.0249,0.0916,0.0326)	(0.990,0.990,0.990)
	1000	(1.5020,3.1116,0.7697)	(0.1061,0.1832,0.1321)	(0.0220 ,0.0116,0.0197)	(0.0116,0.0334,0.0177)	(0.990,0.990,0.920)
$(\alpha,\lambda,\theta)=(1.25,2.63,0.24)$	20	(1.6389,2.7783,0.4016)	(1.0305,0.8411,0.3342)	(0.3889,0.1483,0.1616)	(1.2026,0.7224,0.1367)	(0.990,0.990,0.990)
	50	(1.4826,2.7004,0.3459)	(0.7378,0.5976,0.2589)	(0.2326,0.0704,0.1059)	(0.5930,0.3586,0.0776)	(0.990,0.990,0.990)
	100	(1.3892,2.6563,0.3046)	(0.5549,0.3699,0.1893)	(0.1392,0.0263,0.0646)	(0.3242,0.1362,0.0396)	(0.990,0.990,0.990)
	200	(1.2869,2.6143,0.2729)	(0.3339,0.2520,0.1229)	(0.0369,-0.0157,0.0329)	(0.1117,0.0631,0.0160)	(0.990,0.990,0.990)
	500	(1.2609,2.6029,0.2497)	(0.1980,0.1444,0.0632)	(0.0109,-0.0271,0.0097)	(0.0389,0.0214,0.0041)	(0.990,0.990,0.990)
	1000	(1.2696,2.6243,0.2479)	(0.1621,0.1123,0.0517)	(0.0196,-0.0057,0.0079)	(0.0264,0.0125,0.0027)	(0.990,0.990,0.990)
$(\alpha,\lambda,\theta)=(0.25,0.63,0.20)$	20	(0.3852,0.6554,0.4163)	(0.2658,0.2378,0.3376)	(0.1352,0.0254,0.2163)	(0.0882,0.0566,0.1596)	(0.920,0.990,0.990)
	50	(0.2809,0.6400,0.2641)	(0.1264,0.1368,0.1973)	(0.0309,0.0100,0.0641)	(0.0168,0.0186,0.0427)	(0.990,0.990,0.990)
	100	(0.2935,0.6064,0.2841)	(0.1162,0.0931,0.1732)	(0.0435,-0.0236,0.0841)	(0.0152,0.0091,0.0368)	(0.990,0.990,0.990)
	200	(0.2657,0.6354,0.2246)	(0.0810,0.0744,0.1009)	(0.0157,0.0054,0.0246)	(0.0067,0.0055,0.0107)	(0.990,0.990,0.990)
	500	(0.2569,0.6388,0.2078)	(0.0429,0.0492,0.0537)	(0.0069,0.0088,0.0078)	(0.0019,0.0025,0.0029)	(0.990,0.990,0.990)
	1000	(0.2536,0.6313,0.2044)	(0.0307,0.0303,0.0339)	(0.0036,0.0013,0.0044)	(0.0009,0.0009,0.0012)	(0.990,0.990,0.990)
$(\alpha,\lambda,\theta)=(0.30,0.60,0.90)$	20	(0.3258,0.7817,0.8033)	(0.1165,0.3750,0.2751)	(0.0258,0.1817,-0.0967)	(0.0141,0.1723,0.0843)	(0.990,0.990,0.800)
	50	(0.2813,0.6879,0.7639)	(0.0658,0.2013,0.2639)	(-0.0187,0.0879,-0.1361)	(0.0046,0.0479,0.0875)	(0.990,0.990,0.850)
	100	(0.2869,0.6535,0.8123)	(0.0489,0.1406,0.2222)	(-0.0131,0.0535,-0.0877)	(0.0025,0.0224,0.0566)	(0.990,0.990,0.930)
	200	(0.2905,0.6325,0.8364)	(0.0343,0.0921,0.1553)	(-0.0095,0.0325,-0.0636)	(0.0013,0.0095,0.0279)	(0.990,0.990,0.970)
	500	(0.3007,0.6117,0.8884)	(0.0219,0.0647,0.1214)	(0.0007,0.0117,-0.0116)	(0.0005,0.0043,0.0147)	(0.990,0.990,0.970)
	1000	(0.2970,0.6053,0.8821)	(0.0184,0.0455,0.1003)	(-0.0030,0.0053,-0.0179)	(0.0003,0.0021,0.0103)	(0.990,0.990,0.980)
$(\alpha,\lambda,\theta)=(0.50,2.00,0.40)$	20	(0.5748,2.3413,0.4948)	(0.2790,0.8066,0.3586)	(0.0748,0.3413,0.0948)	(0.0826,0.7606,0.1363)	(0.990,0.990,0.990)
	50	(0.6019,2.0303,0.5348)	(0.2218,0.4461,0.2941)	(0.1019,0.0303,0.1348)	(0.0591,0.1979,0.1038)	(0.990,0.990,0.990)
	100	(0.5100,2.0592,0.4423)	(0.1622,0.3178,0.2465)	(0.0100,0.0592,0.0423)	(0.0262,0.1035,0.0620)	(0.990,0.990,0.990)
	200	(0.5307,2.0009,0.4503)	(0.1091,0.2491,0.1864)	(0.0307,0.0009,0.0503)	(0.0127,0.0614,0.0369)	(0.990,0.990,0.990)
	500	(0.5045,1.9954,0.4194)	(0.0727,0.1594,0.1154)	(0.0045,-0.0046,0.0194)	(0.0053,0.0252,0.0136)	(0.990,0.990,0.990)
	1000	(0.5051,2.0072,0.4034)	(0.0493,0.1002,0.0598)	(0.0051,0.0072,0.0034)	(0.0024,0.0100,0.0036)	(0.990,0.990,0.980)
$(\alpha,\lambda,\theta)=(2.00,0.25,0.80)$	20	(2.1599,0.3199,0.6131)	(1.0176,0.1112,0.3449)	(0.1599,0.0699,-0.1869)	(1.0508,0.0171,0.1527)	(0.990,0.990,0.790)
	50	(2.0826,0.2743,0.7193)	(0.5220,0.0528,0.2874)	(0.0826,0.0243,-0.0807)	(0.2766,0.0033,0.0883)	(0.990,0.990,0.880)
	100	(1.9984,0.2629,0.7519)	(0.4419,0.0418,0.2711)	(-0.0016,0.0129,-0.0481)	(0.1933,0.0019,0.0751)	(0.990,0.990,0.870)
	200	(2.0322,0.2569,0.7808)	(0.3046,0.0272,0.2050)	(0.0322,0.0069,-0.0192)	(0.0929,0.0008,0.0420)	(0.990,0.990,0.970)
	500	(1.9945,0.2552,0.7849)	(0.1613,0.0218,0.1783)	(-0.0055,0.0052,-0.0151)	(0.0258,0.0005,0.0317)	(0.990,0.990,0.920)
	1000	(1.9659,0.2526,0.7774)	(0.1358,0.0160,0.1496)	(-0.0341,0.0026,-0.0226)	(0.0194,0.0003,0.0227)	(0.990,0.990,0.960)

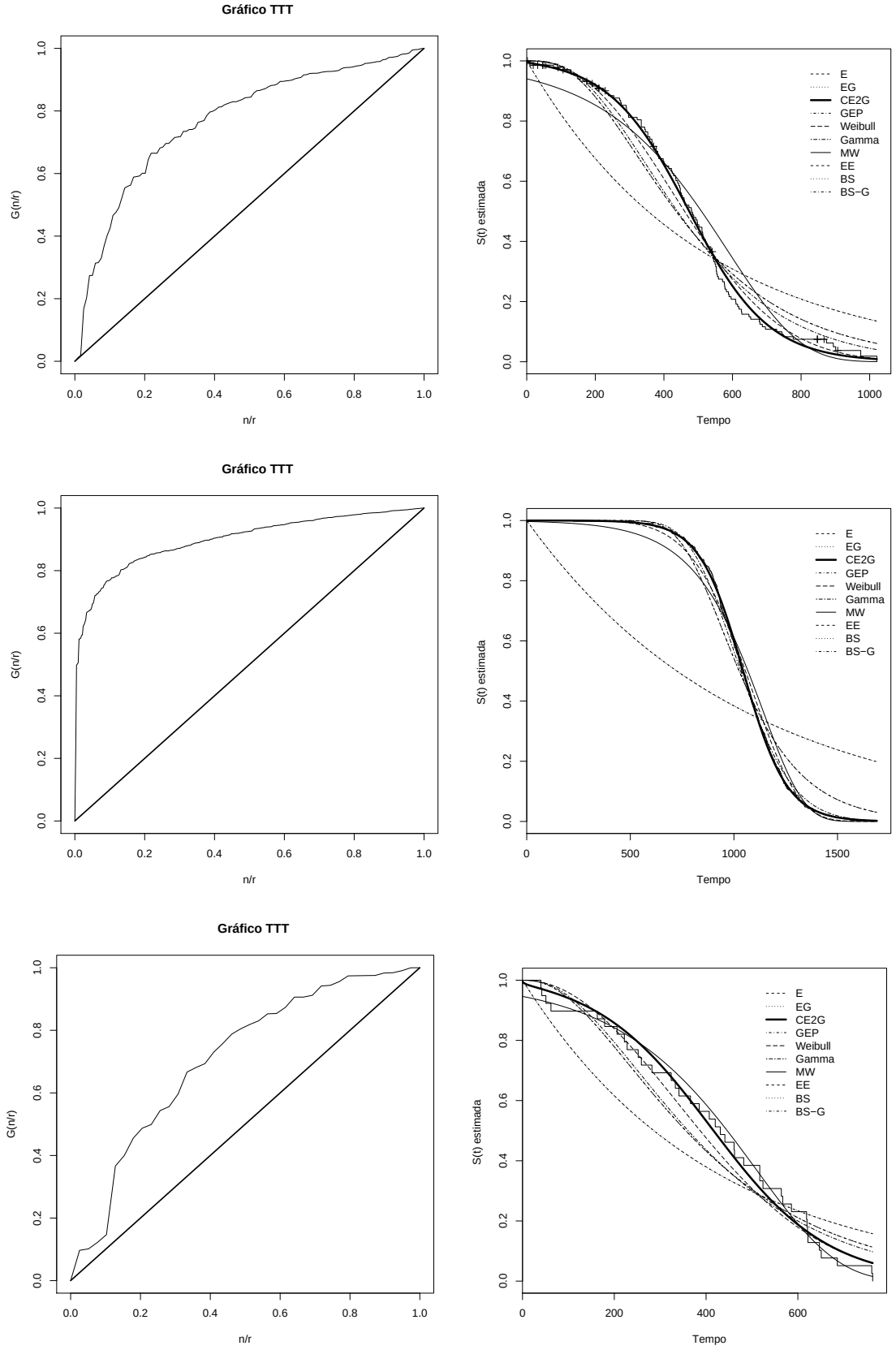

Figura 2.5: Esquerda: Gráfico TTT empírico para $T4$, $T5$ e $T6$, respectivamente. Direita: Modelos ajustados para $T4$, $T5$ e $T6$, respectivamente.

(Barreto-Souza e Cribari-Neto, 2009), , a distribuição Birnbaum-Saunders generalizada (BS-G) (Birnbaum e Saunders, 1969), com f.d.p. dada por $f(y) = \frac{\sqrt{(y-\mu)/\beta}+\sqrt{\beta/(x-\mu)}}{2\alpha(x-\mu)}\phi\Big([\sqrt{(y-\mu)/\beta} - \sqrt{\beta/(x-\mu)}]/\alpha\Big)$, em que $\phi(.)$ é a f.d.p. da distribuição Normal padrão e a distribuição Birnbaum-Saunders (BS) é obtida atribuindo o valor $\mu = 0$ na distribuição BS-G.

Tabela 2.5: Valores do AIC e BIC para todas as distribuições ajustadas em T4, T5 e T6.

		E	EE	EG	Weibull	Gama	CE2G	MW	GEP	BS	BS-G
T4	AIC	1723.7	1657.2	1725.8	1630.5	1649.4	1616.0	1660.0	1659.3	1919.7	1708.5
	BIC	1726.7	1663.2	1731.7	1636.5	1655.3	1624.9	1668.9	1668.2	1925.6	1717.3
T5	AIC	6649.8	5703.2	6651.8	5599.0	5605.9	5571.0	5664.7	5705.3	5648.3	5601.3
	BIC	6653.9	5711.3	6659.9	5607.1	5613.8	5583.1	5676.8	5717.4	5656.3	5613.4
T6	AIC	549.8	538.2	551.8	530.3	536.5	530.6	530.7	540.3	550.8	534.0
	BIC	551.5	541.6	555.2	533.7	539.8	535.6	535.7	545.3	554.1	539.0

A Tabela 2.5 apresenta os valores dos critérios AIC e BIC para todas as distribuições mencionadas. Destes valores concluímos que há evidência a favor da distribuição CE2G para os conjuntos de dados $T4$ e $T5$ em todos os dois critérios de comparação. Para o conjunto $T6$ a distribuição CE2G mostra ajuste similar em valores às distribuições Weibull e MW, indicando assim que é um alternativa viável para distribuições de tempo de vida, ainda mais que esta apresenta apelo construtivo.

A direita da Figura 2.3.8 apresenta as curvas de ajuste das distribuições mencionada para $T4$, $T5$ e $T6$. Os resultados apresentados na Tabela 2.5 são suportados pelo estimador de KM quando comparado com os ajustes das distribuições. Os estimadores MLEs (e seus desvios padrões em parênteses) dos estimadores α, $\theta(\times 1000)$ e $\lambda(\times 10000)$ da distribuição CE2G são, respectivamente, 3.7469 (0.5688), 41.4860 (9.7659) e 17536.46 (7.1814) para $T4$, 5.1765 (19.4159), 0.2625 (0.9915) e 94.6676(3.8720) para $T5$, e 0.0018180 (0.9818), 0.0698 (0.3770) e 78.7704 (11.5084) para $T6$.

2.4 Conclusão

Neste capítulo foram propostas duas novas distribuições de tempo de vida: a distribuição E2G e a CE2G, ambas no cenário de risco latentes, onde a E2G é desenvolvida no cenário com riscos competitivos e a CE2G com riscos complementares.

A distribuição E2G generaliza a distribuição EG proposta por (Adamidis e Loukas, 1998) e possui função de risco crescente, decrescente e unimodal. As propriedades da distri-

buição E2G foram discutidas, incluindo provas analíticas da sua f.d.p., da função de sobrevivência, da função de risco, dos momentos, do r-ésimo momento da i-ésima estatística de ordem e a vida residual média. Os estimadores de máxima verossimilhança foram obtidos diretamente da maximização da log-verossimilhança. Na aplicação do modelo em dados reais foi apresentado conjuntos de dados que mostram a importância do modelo E2G frente a outros modelos por sua interpretação.

A distribuição CE2G possui função de risco crescente, decrescente e em forma de banheira. As propriedades da distribuição CE2G foram discutidas, incluindo provas analíticas de sua f.d.p., da função de sobrevivência, da função de risco, dos momentos, do r-ésimo momento da i-ésima estatística de ordem, a vida residual média, uma medida de entropia e uma medida de confiabilidade. Os estimadores de máxima verossimilhança foram obtidos diretamente da maximização da log-verossimilhança. Na aplicação do modelo em dados reais foi apresentado conjuntos de dados que mostram a importância do modelo CE2G frente a outros modelos por sua interpretação.

Neste capítulo não verificamos que o modelo é identificável, porém no próximo capítulo abordamos a identificabilidade de um modelo derivado pela transformação logarítima e assim a verificação da identificabilidade dos modelos E2G e CE2G se tornam apenas um problema algébrico por apresentar resultados bem próximos para as funções envolvidas na identificabilidade.

Os resultados deste capítulo, foram apresentados a duas revistas e encontram-se publicados em Louzada et al. (2014) e Marchi et al. (2013).

Capítulo 3

Modelo de regressão Log-CE2G

Na análise de sobrevivência é comum o estudo de uma determinada relação particular entre variáveis características do objeto em estudo com uma determinada resposta. Uma relação usual é a relação linear, a qual costuma ser denominada modelos de vida logarítima, os quais mantém algumas propriedades da regressão para modelos Normais com a vantagem de poder ser considerado dados censurados. Tais modelos são especialmente úteis no estudos de tempos de vida pois pela relação particular entre as variáveis dependentes e a resposta, a resposta pode variar entre indivíduos que possuem níveis diferentes para os valores das covariáveis.

Muitos autores tem sugerido o uso de modelos baseados na transformação logarítima das distribuições de tempo de vida. Franco (1984) propôs o uso da distribuição log-Logística para a distribuição do tempo até a falha em análise binária. Smith e Hammond (1988) propuseram que a distribuição residual da média tem a distribuição log-Gamma. Leiva et al. (2007) o uso da distribuição Birbaum-Saunders na presença de dados censurados. A distribuição Birbaun-Saunders é uma distribuição importante originada para o estudo do tempo de vida com o fator de fadiga. Oliveira et al. (2008) propuseram o uso da distribuição BurrXII com dados censurados como altenativa a distribuição log-Logística.

A importância dessas distribuições também se deve ao fato que os modelos possuem funções de risco não-monótonas, o que são comuns em análise de tempos de vida e confiabilidade. Muitos autores têm sugerido distribuições com funções de risco não-monótonas. Cancho et al. (2009) apresentaram a distribuição log-Exponenciada-Weibull com fração de cura. Santana et al. (2012) propuseram a distribuição Kumaraswamy-log-Logística que têm como caso particular a distribuição log-Logística e a log-BurXII. Lemonte et al. (2013) propuseram a distribuição exponenciada de Kumaraswamy baseada na função Beta generalizada.

Neste capítulo propomos o modelo de regressão de log-locação-escala para tempos

de vida usando a distribuição Exponencial exponenciada complementar-geométrica (CE2G) de Marchi et al. (2013), determinada de modelo de regressão log-CE2G. A distribuição CE2G também foi obtida por Bidram et al. (2012), a qual é chamada de "*New Generalized Exponential Geometric*" (NGEG). A inferência será implementada de forma direta da maximização da log-verossimilhança e será determinada uma análise de resíduos.

3.1 Modelo de regressão log-CE2G

Muitos autores se baseiam no modelo de risco proporcional de Cox (Cox, 1972). Por outro lado, em muitas aplicações práticas, os tempos de vida são afetados por variáveis, comumente chamadas de covariáveis, como o nível de colesterol, a pressão sanguínea entre outras. Portanto, é importante explorar a relação entre o tempo de vida e as variáveis explicativas tendo o comportamento do modelo de regressão uma abordagem intuitiva sobre o tempo de vida.

Seja T uma variável aleatória com distribuição CE2G. Então a transformação $Y = \sigma \log(T)$ tem distribuição log-CE2G. A função de distribuição e a f.d.p. da distribuição log-CE2G são dadas por

$$F(y) = \frac{\theta(1-\exp\{-e^{\frac{y-\mu}{\sigma}}\})^{\alpha}}{1-(1-\theta)(1-\exp\{-e^{\frac{y-\mu}{\sigma}}\})^{\alpha}} \text{ e} \tag{3.1.1}$$

$$f(y) = \frac{\alpha\theta \exp\{\frac{y-\mu}{\sigma} - e^{\frac{y-\mu}{\sigma}}\}(1-\exp\{-e^{\frac{y-\mu}{\sigma}}\})^{\alpha-1}}{\sigma\left[1-(1-\theta)(1-\exp\{-e^{\frac{y-\mu}{\sigma}}\})^{\alpha}\right]^2}, \tag{3.1.2}$$

respectivamente, em que $\lambda = \exp\{-\mu/\sigma\}$, $-\infty < y < \infty$, $\sigma > 0$, $\alpha > 0$ e $-\infty < \mu < \infty$. A Figura 3.1 apresenta a f.d.p. e a função de risco para alguns valores desta distribuição.

3.2 Formas da função de risco da distribuição log-CE2G

A transformação $Y = \sigma \log(T)$, onde $T \sim$ CE2G afeta diretamente as formas da f.d.p. e da função de risco da distribuição CE2G para a log-CE2G, portanto, a regressão é utilizada nos valores logarítmicos transformados dos dados de resposta. A forma da função de risco muda como consequência da transformação logarítima e a distribuição log-CE2G apresenta função de risco crescente.

As Figuras 3.1 e 3.2 apresentam as funções de densidade de probabilidade e de risco para as suas correspondentes distribuições log-CE2G e CE2G assumindo os parâmetros $\theta = 0.1, 0, 3$

e $\alpha = 0.5, 1.0, 2.0$. Para motivos de comparação, utilizamos na distribuição log-CE2G os valores $\mu = 0$ e $\sigma = 1$, o que corresponde em utilizar o valor $\lambda = 1$ na distribuição CE2G. Podemos ver da Figura 3.1 que a f.d.p. da distribuição log-CE2G é unimodal e que a função de risco é estritamente crescente pela Figura 3.2 que a f.d.p. da distribuição CE2G é unimodal e que a função de risco é crescente, decrescente ou possui forma de banheira. A forma da função de risco da distribuição log-CE2G é de difícil demonstração e portanto optamos pelo método gráfico para mostrar seu comportamento para alguns pontos do espaço paramétrico por meio da sua derivada. Após obter a derivada da função de risco, fica-se sugestivo por meio de gráficos, omitidos aqui, que seu valor será sempre positivo, ou seja, a função de risco é crescente embora que o esperado fosse que a distribuição log-CE2G mantivesse várias formas para a função de risco.

3.3 Momentos e função característica

A função geradora de momentos da variável Y com distribuição log-CE2G pode ser obtida analiticamente e sua função característica é dada por

$$\Phi_i(t) = \alpha\theta\lambda' \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \binom{-2}{j} \binom{\alpha(j+1)-1}{k} (1-\theta)^j (-1)^{j+k} \frac{\Gamma(t\sigma+1)}{[\lambda'(k+1)]^{t\sigma+1}}, \tag{3.3.1}$$

em que $\lambda' = e^{-\mu/\sigma}$.

Os momentos da distribuição log-CE2G podem ser obtidos diretamente da relação $\Phi^{(r)}(t)\big|_{t=0} = E(Y^r)$, em que $\Phi^{(r)}$ é a r-ésima derivada de $\Phi(t)$ em t. Os momentos não são apresentados aqui pois a derivada da função característica pode ser obtida diretamente por *software* computacional e não é visualmente atrativa.

3.4 Identificabilidade do modelo

A condição de identificabilidade, isto é, a existência de uma representação única para cada uma das classes dos modelos considerados, é de grande importância do ponto de vista estatístico para a correta estimação dos parâmetros envolvidos. O Theorema 2 demonstrado em Teicher (1963) apresenta condições suficientes para que seja garantida a identificabilidade em distribuições univariadas. Note que a Equação (5) da prova do teorema é satisfeita para qualquer "k" e ainda $\sum c_i = \sum P(M = m) = 1$. Para termos as condições apresentadas no teorema, ou seja, ser um modelo univariado, basta a fixação de alguns parâmetros.

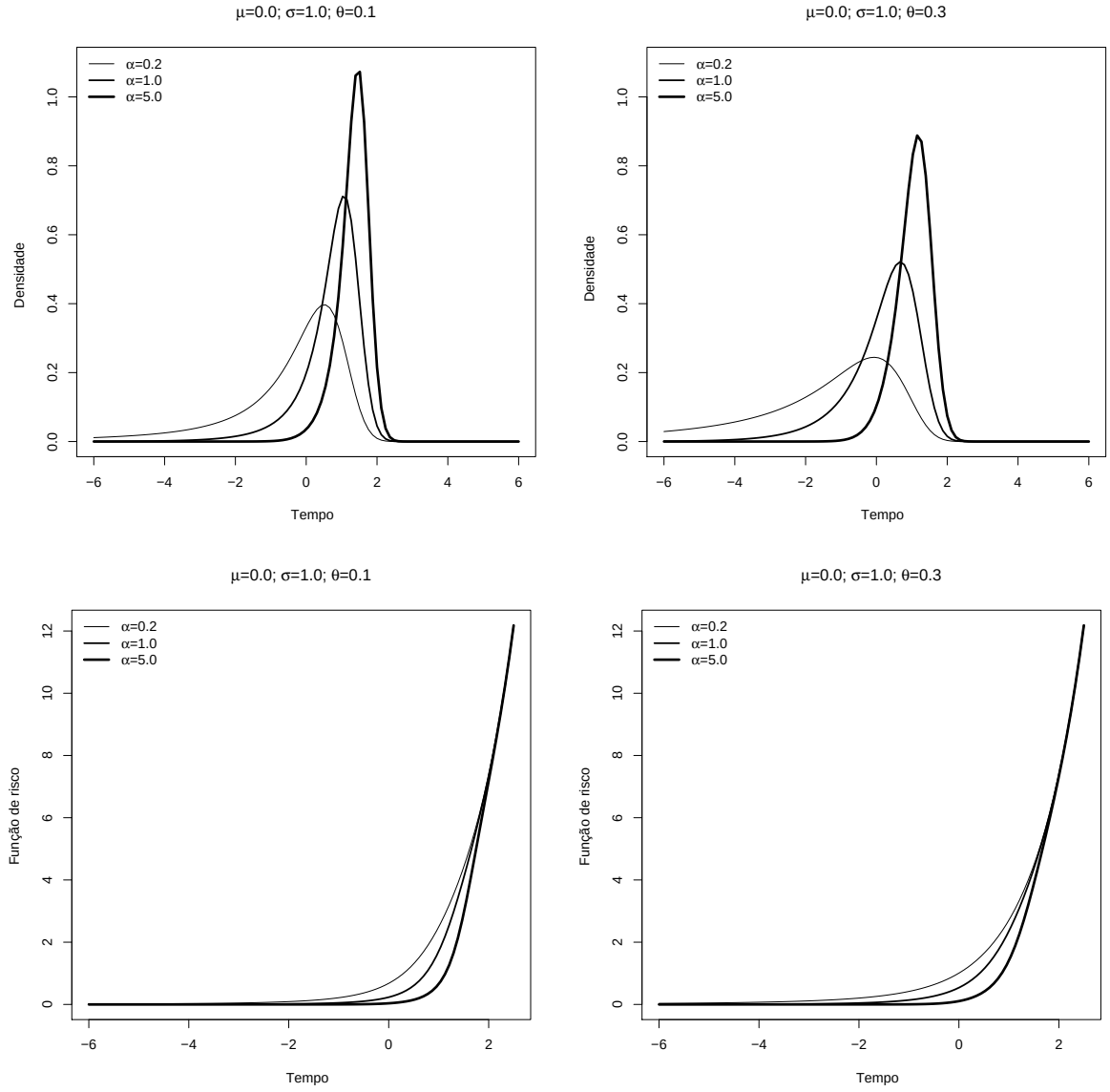

Figura 3.1: Painel Superior: F.d.p. da distribuição log-CE2G. Painel Inferior: Função de risco da distribuição log-CE2G. Para os valores fixados $\mu = 0$ e $\sigma = 1$.

Considere os valores fixados tal que $\mu_1 > \mu_2$, $\alpha_1 > \alpha_2$, e $\sigma_2 > \sigma_1$. Portanto, temos que $F_1(y) < F_2(y)$, em que

$$F_1(y) = \frac{\theta(1-\exp\{-e^{\frac{y-\mu_1}{\sigma}}\})^{\alpha_1}}{1-(1-\theta)(1-\exp\{-e^{\frac{y-\mu_1}{\sigma}}\})^{\alpha_1}} \text{ e } F_2(y) = \frac{\theta(1-\exp\{-e^{\frac{y-\mu_2}{\sigma}}\})^{\alpha_2}}{1-(1-\theta)(1-\exp\{-e^{\frac{y-\mu_2}{\sigma}}\})^{\alpha_2}}.$$

Da equação (3.3.1), obtemos a função característica de $\Phi_{i1}(t)$ e $\Phi_{i2}(t)$ para $F_1(y)$ e $F_2(y)$, respectivamente. Podemos verificar que

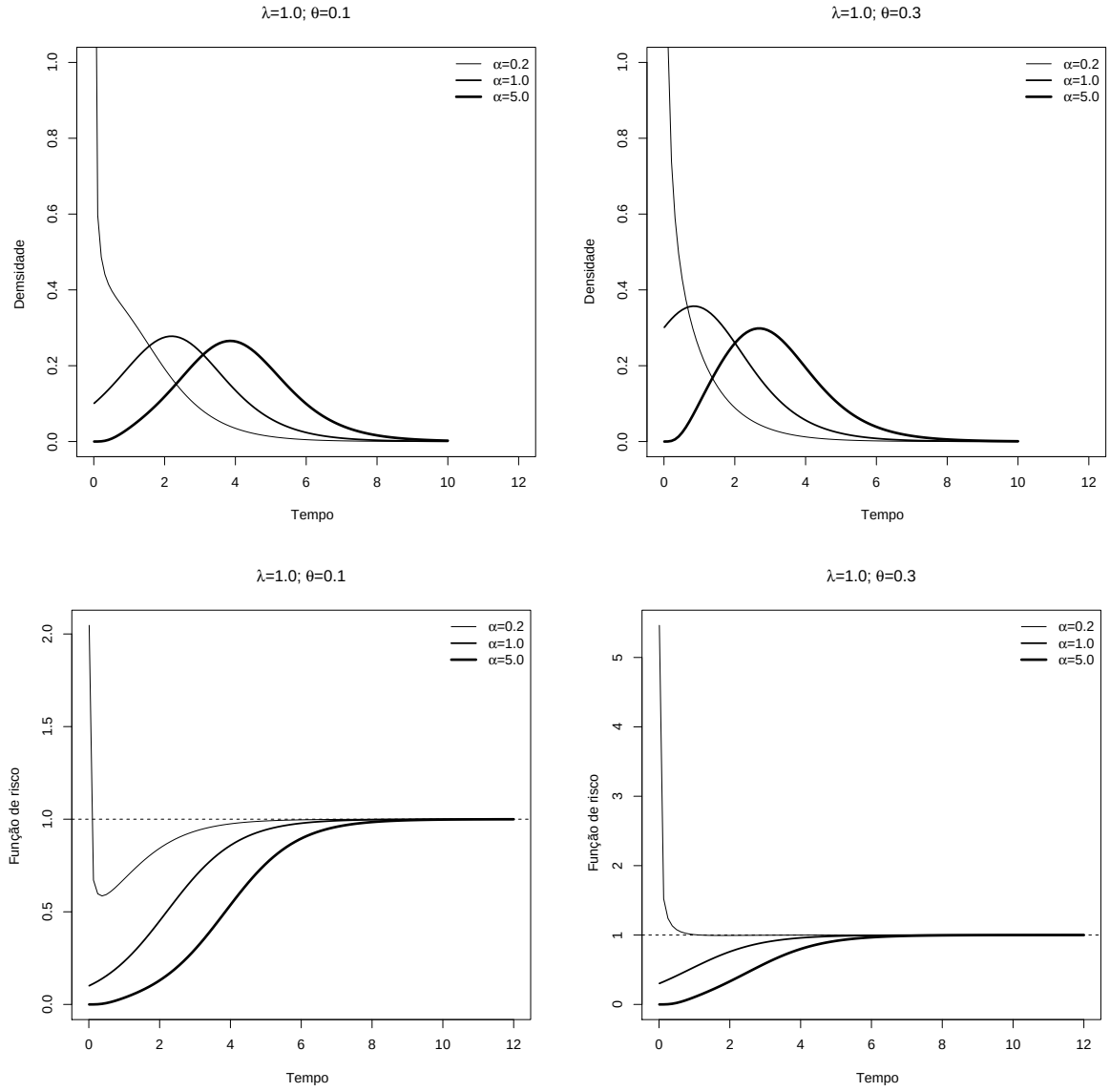

Figura 3.2: Painel Superior: F.d.p. da distribuição CE2G. Painel Inferior: Função de risco da distribuição CE2G. Para o valor fixado $\lambda = 1$.

$$\lim_{t\to -1/\sigma_1^+} \Phi_{i1}(t) = \infty, \quad \lim_{t\to -1/\sigma_1^-} \Phi_{i1}(t) = -\infty \text{ e}$$

$$\lim_{t\to -1/\sigma_1} \Phi_{i2}(t) = \alpha_2\theta_2\lambda_2' \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \binom{-2}{j}\binom{\alpha_2(j+1)-1}{k}(1-\theta)^j(-1)^{j+k}\frac{\Gamma(1-\sigma_2/\sigma_1)}{[\lambda_2(k+1)]^{1-\sigma_2/\sigma_1}},$$

em que $\lambda_2' = e^{-\mu_2/\sigma_2}$.

Portanto, $\lim_{t\to -1/\sigma_1} \dfrac{\Phi_{i2}(t)}{\Phi_{i1}(t)} = 0$, isto é, pelo teorema 2 em Teicher (1963) o modelo log-CE2G é identificável.

3.5 Função log-CE2G padronizada

A distribuição padronizada,$F_0(y)$ e a f.d.p., $f_0(y)$, padronizadas de $Y \sim$log-CE2G ($\mu = 0$ e $\sigma = 1$) são dadas por

$$F_0(y) = \frac{\theta(1-\exp\{-e^y\})^\alpha}{1-(1-\theta)(1-\exp\{-e^y\})^\alpha} \text{ e} \quad (3.5.1)$$

$$f_0(y) = \frac{\alpha\theta\exp\{y-e^y\}(1-\exp\{-e^y\})^{\alpha-1}}{[1-(1-\theta)(1-\exp\{-e^y\})^\alpha]^2}, \quad (3.5.2)$$

respectivamente.

Considerando uma amostra $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ de $(Y, \mathbf{X})$, em que $\mathbf{x}_i$ são as covariáveis associadas a i-ésima resposta y_i com $\mathbf{X} = (X_1, \ldots, X_p)$. Cada resposta é definida por $y_i = \min\{\log(t_i), \log(c_i)\}$, em que $\log(t_i)$ e $\log(c_i)$ é o logarítimo do tempo de vida y_i e os logarítimo do tempo de censura, c_i, respectivamente. Vamos assumir que os tempos de vida e o tempo de censura são independentes com censura não informativa.

Um modelo linear de regressão para a variável resposta y_i baseada na distribuição log-CE2G é da forma

$$y_i = \mathbf{x}_i^t\beta + \sigma z_i, \quad i = 1, \ldots, n, \quad (3.5.3)$$

em que z_i tem distribuição dada pela equação (3.5.1) e (3.5.2) com o vetor $\beta = (\beta_0, \ldots, \beta_p)$ o vetor de parâmetros a ser estimado. Na equação (3.5.3) o parâmetro de locação é $\mu_i = \mathbf{x}_i^t\beta$ e portanto $\mu = (\mu_1, \ldots, \mu_n)^t$ é o vetor de locação para o modelo log-CE2G com estrutura linear $\mu = \mathbf{X}\beta$. O modelo de regressão log-CEG é definido por $\alpha = 1$.

A log-verossimilhança para o modelo log-CE2G de parâmetros $(\alpha, \beta, \theta, \sigma)$ é dada por

$$\begin{aligned}\ell(\alpha,\beta,\theta,\sigma) &= k\log\left(\frac{\alpha\theta}{\sigma}\right) + \sum_{i\in F}\left(\frac{y_i-\mathbf{x}_i^t\beta}{\sigma} - e^{\frac{y_i-\mathbf{x}_i^t\beta}{\sigma}}\right) + (\alpha-1)\sum_{i\in F}\log\left(1-\exp\left[-e^{\frac{y_i-\mathbf{x}_i^t\beta}{\sigma}}\right]\right) \\ &-2\sum_{i\in F}\log\left(1-(1-\theta)\left(1-\exp\left[-e^{\frac{y_i-\mathbf{x}_i^t\beta}{\sigma}}\right]\right)^\alpha\right) + \sum_{i\in C}\log\left(1-\left(1-\exp\left[-e^{\frac{y_i-\mathbf{x}_i^t\beta}{\sigma}}\right]\right)^\alpha\right) \\ &-\sum_{i\in C}\log\left(1-(1-\theta)\left(1-\exp\left[-e^{\frac{y_i-\mathbf{x}_i^t\beta}{\sigma}}\right]\right)^\alpha\right),\end{aligned}$$

em que k é o número observado de itens que falharam, F é o conjunto de objetos que tiverem o tempo de vida observados e C é o conjunto de objetos que não se foi observado o tempo de falha, ou seja, os objetos censurados.

A inferência dos parâmetros é obtida por meio da estimação por máxima verossimilhança baseados no comportamento de amostras suficientemente grandes. Para este cenário

os estimadores MLEs de $\hat{\alpha}$, $\hat{\beta}$, $\hat{\theta}$ e $\hat{\sigma}$ e seus desvios padrões são obtidos diretamente da maximização da função de log-verossimilhança $\ell(\alpha, \beta, \theta, \sigma)$ pela rotina *optim* do *Software* R Core Team (2012).

3.6 Máximo local da função de log-verossimilhança

A prova analítica da unicidade dos estimadores de máxima verossimilhança não foi obtida por dificuldades em provar suas condições. De fato, a prova consiste em analizar a definição da matriz Hessiana ser negativa definida, o que é complicado para este caso. Por esta razão apresentamos a seguir um gráfico onde é possível ver o máximo da função de log-verossimilhança dos parâmetros comparados dois a dois. Considere o caso onde é observado $y = 2$. Do procedimento da rotina *optim* do *software* R Core Team (2012), é obtido que $\hat{\alpha} = 10.963, \hat{\mu} = 1.998, \hat{\sigma} = 0.001$ e $\hat{\theta} = 0.4837$.

A Figura 3.3 apresenta os gráficos de contorno com níveis especificados da função de log-verossimilhança para parâmetros dois a dois e os respectivos valores dos MLE's . É possível notar que os valores dos MLE's obtidos estão numa região de máximo.

3.7 Resíduos e observações influentes

Em alguns casos pode ser de interesse conhecer quais observações na amostra mais influenciam na reta de regressão. Se esta observação influente é ou não uma medida errônea que foi observada no processo pode ser de interesse estudá-la para compreender e explicar o processo real que gerou tal observação. Pois pode ocorrer uma pequena porção de resultados na amostra que possuem um processo diferente dos demais elementos.

3.7.1 Influência global

Para modelos cuja as covariáveis explicam o tempo de vida por meio da relação linear da forma $\mathbf{x}'\beta$, isto é, $\mu(\mathbf{x}) = \mathbf{x}'\beta$, os valores *leverege*

$$h_{ii} = \mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i,$$

em que X é uma matiz $n \times p$ que a i-ésima linha é o vetor $\mathbf{x}_i$, é tido como uma boa medida em potencial para determinar a influência de $\mathbf{x}_i$ (Lawless, 2003, pg. 288). Segundo Kutner et al. (2005), valores de h_{ii} maiores que duas vezes a média dos valores leverege ($\frac{2p}{n}$ em modelos de

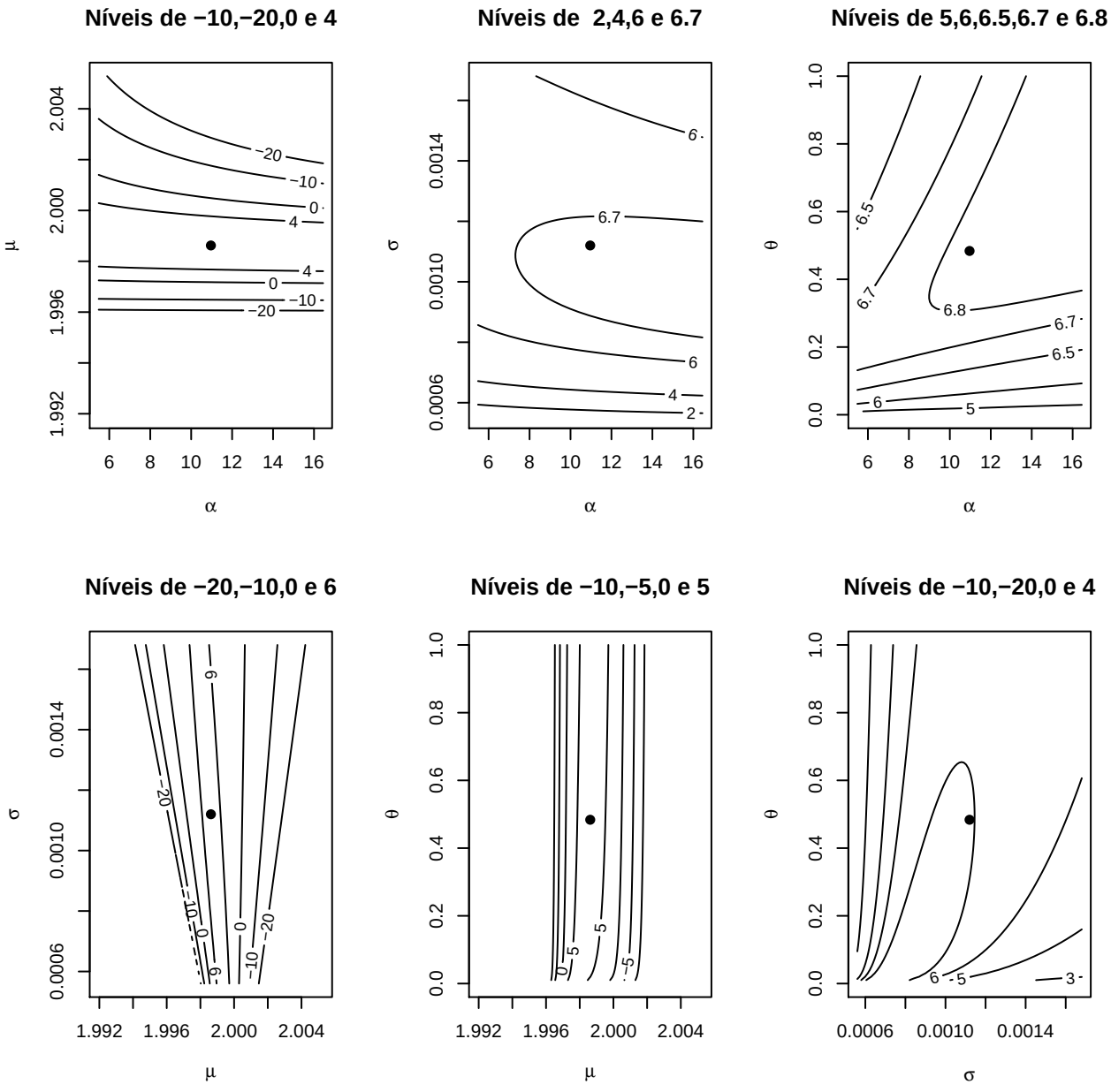

Figura 3.3: Gráfico de contorno com níveis especificados da função de log-verossimilhança para parâmetros dois a dois.

regressão linear) são considerados indicadores de casos outiliers em $\mathbf{x}_i$, em que p é o número de covariáveis na função de regressão incluindo o interceptor.

Uma outra medida para detectar pontos de influência é a abordagem de deleção de caso (*case-deletion*), que consiste em retirar um grupo de observações, digamos g, e refazer o ajuste do modelo para comparar os valores de $\hat{\theta}$ e $\hat{\theta}_{-g}$, em que $\hat{\theta}$ é a estimativa dos MLEs sem nenhuma amostra retirada e $\hat{\theta}_{-g}$ é a estimativa dos MLEs para o grupo de observações g retirado. Quando o grupo g consiste somente na retirada da i-ésima observação, a medida de influência global recebe o nome de estatística de Likehood-Drop (LD) e é calculada por

$$LD_i(\theta) = 2\ell(\hat{\theta}_{-i}) - 2\ell(\hat{\theta}), \tag{3.7.1}$$

a qual é comparada ao valor do quantil χ^2_p da distribuição Qui-Quadrado, em que p é a dimensão do vetor θ (Lawless, 2003, pg. 287).

Uma aproximação correspondente para o LD_i pode ser obtida por meio da expansão de segunda ordem da série de Taylor para $\ell(\theta)$, dada por

$$LD_i = (\hat{\theta}_{-i} - \hat{\theta})^t \mathbf{I}(\hat{\theta})^{-1} (\hat{\theta}_{-i} - \hat{\theta}), \tag{3.7.2}$$

em que $\mathbf{I}(\hat{\theta})$ é a matriz de Informação. A aproximação na equação (3.7.2) é boa o suficiente para indicar quais observações têm grande impacto quando retiradas (Lawless, 2003). A medida de influência individual nos parâmetros $\psi(\theta)$ é calculada por

$$\Delta\psi_{-i} = \frac{\hat{\psi} - \hat{\psi}_{-i}}{se(\hat{\psi})}, \tag{3.7.3}$$

em que $se(\hat{\psi})$ é a estimativa do desvio padrão de ψ.

3.8 Análise de resíduos

Para verificar se o modelo está bem ajustado ao conjunto de dados é comum na literatura o uso de métodos gráficos para a análise de resíduos. Neste tipo de análise, os resíduos têm que apresentar certas características para que o modelo se apresente adequado.

O ajuste dos modelos de locação-escala ou dos equivalentes modelos de tempo de vida acelerados (AFT), pode ser verificado pelos resíduos padronizados dados pela fórmula $\hat{z}_i = \frac{(y_i - \hat{\mu}_i)}{\hat{\sigma}}$, $i = 1, \ldots, n$. Para se verificar o bom ajuste, espera-se que estes resíduos se comportem como independentes e identicamente distribuídos (i.i.d.) com função de sobrevivência $S_0(z) = 1 - F_0(z)$, pelo menos para um n razoavelmente grande (Lawless, 2003). Portanto o gráfico PP-plot dos resíduos $\hat{z}_i$ pode ser utilizado para validar o ajuste.

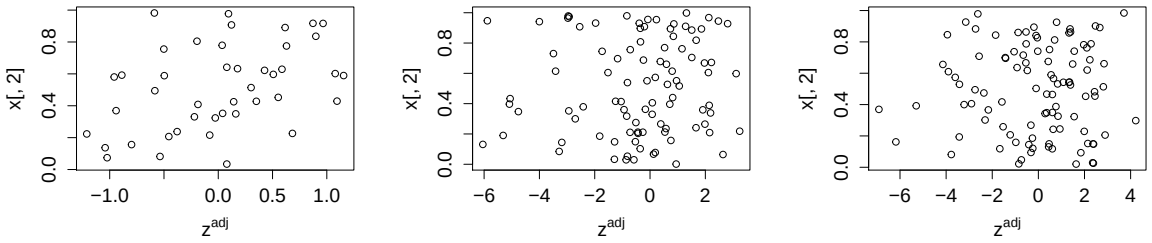

Figura 3.4: Resíduos ajustados para verificar a homecedasticidade nos dados simulados.

É conhecido que o valor $\hat{z}_i$ para uma observação censurada apresenta valor menor que o valor residual verdadeiro. Assim, em modelos de locação-escala o resíduo ajustado pode ser obtido por

$$\hat{z}_i^{adj} = \delta_i \hat{z}_i + (1-\delta_i) E(Z_i | Z_i \geq \hat{z}_i),$$

em que $E(Z_i|Z_i \geq \hat{z}_i) = \int_{\hat{z}_i}^{\infty} \frac{f_0(x)}{1-F_0(x)} dx$. Desta forma, os gráficos dos valores $\hat{z}_i^{adj}$ contra as covariáveis ou outros fatores como os valores ajustados $\hat{y}_i$ podem ser utilizados para checar se σ têm valor médio constante (não é heterocedástico). É importante notar, entretanto, que a distribuição dos resíduos ajustados depende do tempo de censura e pode ser muito diferente da distribuição de Z (Lawless, 2003).

Para o modelo de regressão log-CE2G, a média residual tem valor $E(Z_i|Z_i \geq \hat{z}_i)$ definido por

$$\begin{aligned} E(Z_i|Z_i \geq z_i) &= \frac{\alpha \mu_i \theta}{\alpha+1} {}_2F_1\left(\left[2, \frac{\alpha+1}{\alpha}\right], \left[\frac{2(\alpha+1)}{\alpha}\right], (1-\theta)(1-\exp(-e^{z_i}))^{\alpha}\right) \\ &- \alpha\theta\sigma \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \frac{(k+1)(1-\theta)^k (\alpha(k+1))_j}{(j+1)!} \left(z_i \exp\{-(j+1)e^{z_i}\} + \Gamma(0,(j+1)e^{z_i})\right), \end{aligned}$$

em que ${}_2F_1([a,b],[c],z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)k!} z^k$ é a função hipergeométrica, $(a)_j = (a)(a+1)\cdots(a+j-1)$ é o símbolo de Pochhammer e $\Gamma(a,b) = \int_b^{\infty} e^{-t} t^{a-1} dt$ é a função gama incompleta (superior). A Figura 3.4 apresenta situações com três níveis de censura (40%, 20% e 0%) em amostras de tamanho $n = 100$ com uma variável $X \sim U(0,1)$. Podemos observar que o modelo apresenta características desejadas na covariável $x[,2]$ (X), ou seja, não aparenta ser resíduos com presença de dados heteroscedásticos.

3.9 Estudo de simulação

Para verificar as propriedades assintóticas dos estimadores MLEs, faremos um estudo de simulação que consiste em 1000 amostras geradas da distribuição log-CE2G, usando a equação (3.5.3), para três diferentes conjuntos de parâmetros e com três covariáveis, considerando $n =$ 30, 60 e 100. Considerando que a distribuição possui parâmetros restritos no espaço real, o parâmetro θ é obtido por meio da transformação $\theta = e^{\theta^*}/(1 + e^{\theta^*})$, em que $\theta^* \in \mathbb{R}$, e para α e σ consideramos a transformação exponencial. Para o cálculo da variância de cada parâmetro, foi utilizado o método Delta. Devido a escolha dos parâmetros o algorítimo de maximização adotado (rotina *optim* do *software* R Core Team (2012)) é necessário a escolha do valor inicial de σ maior que e^2 para que a rotina se inicie.

A Tabela 3.1 apresenta os resultados da simulação. É apresentado a média dos 1000 MLEs (Av), a probabilidade de cobertura para intervalos de confiança construídos com 95% de confiança (C), e os desvios padrões (Sd). Os resultados sugerem que os estimadores MLEs se comportam adequadamente: Conforme o tamanho da amostra aumenta a variância decresce e a probabilidade de cobertura se aproxima do valor estabelecido. É importante ressaltar que aproximadamente 10% das amostras apresentaram problemas no processo de estimação, como por exemplo erro computacional de funções pois os parâmetros estavam no limite do espaço paramétrico ou a matriz Hessiana não era positiva definida e portanto foram eliminadas e uma nova amostra foi gerada até completar 1000 simulações. Ainda é necessário cuidados na estimação de amostras de tamanho pequeno, como é possível ver na primeira e terceira configuração, pelo motivo que a cobertura dos parâmetros podem ser altamente influenciados como pode ser verificado para os parâmetros α e β_4. Ainda, em alguns casos pode-se notar que a probabilidade de cobertura ainda não atingiu os valores especificados de 95% de confiança para amostras de tamanho 100.

3.10 Influência dos valores iniciais

Os valores iniciais escolhidos podem influenciar nos resultados da maximização do modelo pois o métodos *BFGS* é utilizado para maximização direta da função de log-verossimilhança. Devido a esta preocupação analisamos os resultados dos procedimentos de maximização para amostras com $(\alpha, \theta, \sigma, \beta_0, \beta_1, \beta_2, \beta_3) = (2.5, 0.7, 2.0, 2.0, -1.0, 0.005, -0.1)$ para nível de censura de 10% e tamanho de amostra $n = 100$. Os valores iniciais foram gerados da distribuição Nor-

Tabela 3.1: Valor médio dos MLEs, média dos desvios padrões e probabilidade de cobertura do modelo log-CE2G para dados simulados.

	$\hat{\alpha}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\alpha}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\alpha}$	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Valor Verdadeiro	2.5	0.7	2.0	2.0	-1.0	0.005	-0.1	0.5	0.3	1.2	4.0	3.0	-0.3	2.0	1.0	0.3	0.5	3.0	1.0	4.0	1.0
								$n = 30$													
Av	3.343	0.729	1.517	1.846	-0.991	0.004	-0.092	0.383	0.296	0.671	4.460	3.054	-0.316	2.001	0.900	0.393	0.257	3.299	0.976	3.977	0.999
Sd	4.863	0.095	0.941	6.139	1.098	0.027	0.079	0.718	0.040	0.574	1.581	1.431	0.226	0.100	1.802	0.117	0.251	0.453	0.283	0.165	0.034
C (95%)	0.850	0.999	0.907	0.865	0.853	0.965	0.851	0.999	0.999	0.491	0.528	0.521	0.531	0.539	0.842	0.999	0.547	0.501	0.577	0.530	0.608
								$n = 60$													
Av	2.077	0.746	1.461	2.632	-1.006	-0.002	-0.095	0.480	0.295	1.001	4.132	3.089	-0.315	1.997	1.089	0.379	0.401	3.126	0.999	3.991	1.001
Sd	2.887	0.094	0.677	4.614	0.787	0.026	0.057	0.485	0.040	0.428	1.044	0.879	0.127	0.066	1.344	0.118	0.203	0.386	0.180	0.107	0.023
C (95%)	0.834	0.999	0.979	0.952	0.903	0.999	0.934	0.999	0.999	0.896	0.878	0.837	0.862	0.809	0.933	0.998	0.948	0.894	0.886	0.828	0.828
								$n = 100$													
Av	3.009	0.726	1.832	1.718	-0.978	0.006	-0.095	0.476	0.296	1.078	4.092	3.030	-0.307	2.002	0.985	0.367	0.425	3.115	0.998	3.989	0.999
Sd	3.393	0.098	0.699	4.017	0.585	0.024	0.046	0.306	0.040	0.322	0.764	0.640	0.092	0.051	0.935	0.111	0.148	0.287	0.138	0.084	0.018
C (95%)	0.762	0.999	0.955	0.956	0.934	0.996	0.934	0.999	0.999	0.972	0.962	0.897	0.930	0.868	0.985	0.999	0.999	0.967	0.917	0.845	0.847

mal com média zero e desvio padrão 2, N(0,2), com a ressalva de que para os parâmetros α e γ foi feita uma transformação exponencial (e^u) e para o parâmetro θ foi realizada transformação $e^u/(1+e^u)$ em que u é o resultado da N(0,2). para o parâmetro θA Tabela 3.2 apresenta os valores iniciais usados na rotina *BFGS* e os valores dos MLE's da função de log-verossimilhança. A rotina de maximização é completada em aproximadamente 5 segundos em um computador com processador Intel Core i5-4200Y, com 4GB de memória Ram e sistema operacional Windows de 64 bits e em cada resultado as estimativas apresentam estar bem próximas dos valores verdadeiros. É importante mencionar que apesar dos valores estimados para β_3 parecem não variar pois foi utilizado o arredondamento com 4 casas decimais.

Tabela 3.2: Estimativas da amostra com 10% de censura e $n = 100$ para o modelo log-CE2G para influência de valores iniciais.

Os Valores dos parâmetros verdadeiros são $(\alpha, \theta, \sigma, \beta_0, \beta_1, \beta_2, \beta_3) = (2.5, 0.7, 2.0, 5.0, -1.0, 0.01, -0.1)$														
Valores Iniciais							Estimativas dos MLE's							
0.352	0.052	61.313	-3.677	-0.835	3.359	-0.659	2.337	0.740	1.958	5.595	-0.764	0.014	-0.151
0.865	0.368	2.465	-1.506	1.641	2.742	-0.474	2.167	0.807	1.871	5.874	-0.751	0.014	-0.152
3.281	0.001	10.044	2.004	1.038	0.728	-0.146	2.251	0.781	1.913	5.378	-0.752	0.015	-0.147
0.093	0.972	11.822	0.651	2.540	-0.681	1.353	2.119	0.840	1.841	5.960	-0.755	0.013	-0.152
0.527	0.563	5.135	3.012	-0.835	-1.153	1.029	2.310	0.708	1.957	5.533	-0.768	0.014	-0.150
0.619	0.296	2.601	0.439	-0.795	-1.070	0.883	2.095	0.855	1.826	5.941	-0.757	0.015	-0.152
0.009	0.851	3.725	-1.050	-2.191	-1.850	1.251	2.548	0.729	2.039	5.067	-0.752	0.016	-0.147
0.135	0.324	82.599	2.554	2.594	-3.248	0.039	2.189	0.804	1.879	5.753	-0.745	0.014	-0.151
0.938	0.411	35.766	-1.326	-1.693	-2.424	2.007	2.384	0.772	1.966	5.181	-0.752	0.015	-0.146
1.804	0.418	58.674	0.398	-1.690	1.871	1.412	2.693	0.658	2.115	5.027	-0.736	0.015	-0.148

3.11 Aplicação em dados de câncer de pulmão avançado

Como aplicação a dados reais, comparamos o ajuste da distribuição log-CE2G com outras distribuições de tempo de vida logarítima de três parâmetros em um conjunto de dados extraído da literatura: a distribuição log-EEP, obtida da equação (3.5.3) quando z_i tem distribuição Exponencial Exponenciada Poisson (EEP) com f.d.p. dada por $f(x) = \alpha\beta\lambda exp(-\beta x)[1-exp(-\beta x)]^{\alpha-1}exp(-\lambda[1-exp(-\beta x)])^{\alpha}[1-exp(-\lambda)]^{-1}$, em que $\alpha, \beta, \lambda > 0$; a distribuição log-MW, considerando a distribuição Weibull modificada (MW) para z_i com f.d.p. dada por $f(x) = \alpha x^{\theta-1}(\theta+\lambda x)e^{\lambda x}e^{-\alpha x^{\theta}\exp\{\lambda x\}}$, em que $\alpha, \theta \geq 0$ e $\lambda > 0$; e a distribuição log-GEP, considerando a distribuição Exponencial-Poisson generalizada para z_i (Barreto-Souza e Cribari-Neto, 2009).

Este conjunto de dados, chamado também de $T7$, apresenta uma amostra de 40 pacientes com cancêr de pulmão avançado e foi extraído de Lawless (2003). O principal motivo do estudo é comparar os efeitos de dois tratamentos quimioterápicos para o prolongamento do tempo de vida. Todos os pacientes nesta amostra receberam tratamento anterior e foram direcionados aleatoriamente a dois novos tratamentos, denominados como "padrão" e "teste". Além disso, cada paciente recebeu uma nota de Karnofsky, ou simplesmente um estado de desenvolvimento (PS). A idade do paciente e o número de meses desde o diagnóstico da presença do câncer até o momento de participarem do estudo (diag) também foram observados.

A Figura 3.5 apresenta o gráfico TTT-plot para os tempos de vida y. Podemos determinar que a função de risco dos tempos de vida apresenta forma de banheira, isto é, a distribuição log-CE2G é adequada para este conjunto de dados. Observe que o gráfico TTT-plot é utilizado no conjunto de dados original, ou seja, sem fazer a transformação logarítima nos resultados y_i, $i = 1, \ldots, n$.

As funções de log-verossimilhanças dos estimadores MLEs obtidos são 58.0220, 58.0284, 59.3023 e 58.0247 para os modelos log-CE2G, log-EEP, log-MW e log-GEP, respectivamente. Em todos os modelos foi considerado a relação

$$\mu = \beta_0 + \beta_1 I(Tratamento = teste) + \beta_2 PS + \beta_3 Idade + \beta_4 diag.$$

A Tabela 3.3 mostra os valores dos MLEs e seus respectivos desvios padrões (Sd) para o modelo log-CE2G. O parâmetro β_2 estimado pode ser interpretado como o incremento no log-tempo de vida em sobrevivência de 0.612 a cada 10 pontos acrescentados no nível de PS, isto é, um decrescimento de 0.49 dias no tempo de sobrevivência. A mesma conclusão pode ser feita

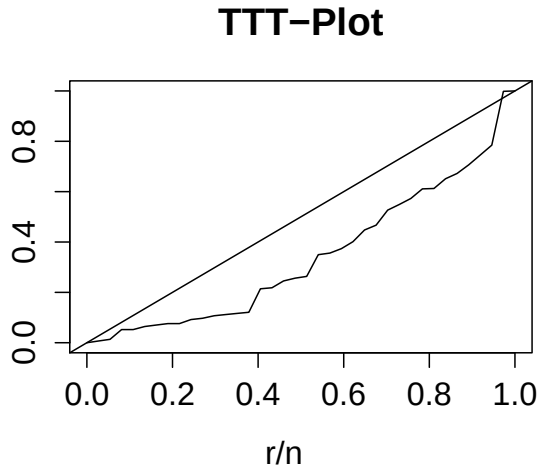

Figura 3.5: Gráfico TTT-Plot para os tempos de vida.

diretamente para os outros parâmetros em μ_i. Resultados e conclusões semelhantes são obtidas em Lawless (2003), pg. 303. Ainda, para valores de $\hat{\alpha} > 1$ o modelo apresenta fragilidade, o que é o caso do modelo ajustado. O valor $\hat{\alpha} \approx 6$ é comparado como um sistema em série de seis componentes, cada um tendo uma distribuição Exponencial para os fatores de risco. A Figura 3.6 apresenta os gráficos de probabilidade (figura à esquerda) e de quantil (figura à direita) para o modelo log-CE2G ajustado em T7, indicando que o modelo log-CE2G é adequado para os dados em análise. Na Figura 3.7 observamos os gráficos dos resíduos ajustados contra as covariáveis afim de checar que os resíduos não apresentam evidência contra as características assumidas para o modelo log-CE2G, pois para diferente valores aos quais foram plotados a variância parece não apresentar tendência. Ainda os valores acima de zero (0) para os resíduos ajustados já é esperado pois para $\mu = 0$ os valores são concentrados com maior massa acima do valor zero como visto na Figura 3.1 além do que, os resíduos ajustados são maiores que os resíduos calculados como já mencionado na Seção 3.8.

Para a análise de influência global calculamos os valores *leverage*, h_{ii}, os quais são apresentados na Figura 3.8, onde a linha horizontal é representada pelo valor $2p/n$ que separa os possíveis casos influentes.

A Tabela 3.4 apresenta os valores para o *case-deletion*, $LD_i(\theta)$, considerando os pontos (conjunto G) apontados na Figura 3.8 usando a equação (3.7.1) e também a influência, LD_i, nos parâmetros $\hat{\psi} = \hat{\alpha}, \hat{\theta}, \hat{\sigma}, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ e $\hat{\beta}_4$ utilizando a Equação (3.7.3). Da Tabela 3.4 podemos concluir que individualmente o conjunto de pontos G não são observações influenciáveis quando comparado o valor LD_i com o valor $\chi^2_{(8,0.95)} = 15.5073$, porém para $\Delta\psi_{-i}$ nota-se que todas

Tabela 3.3: Valores dos parâmetros do modelo log-CE2G ajustado para T7.

	Parâmetros	Estimado	Sd	Z
Modelo	β_0(Int.)	-2.2447	0.3664	-6.1251
	β_1(Tratamento)	0.1926	0.3578	0.5384
	β_2(PS)	0.0612	0.0082	7.4162
	β_3(Idade)	0.0094	0.0140	0.6999
	β_4(diag)	0.0014	0.0123	0.1517
	σ	2.3175	0.3464	6.6901
	α	6.4741	0.0557	116.0864
	θ	0.6107	0.0419	14.5751

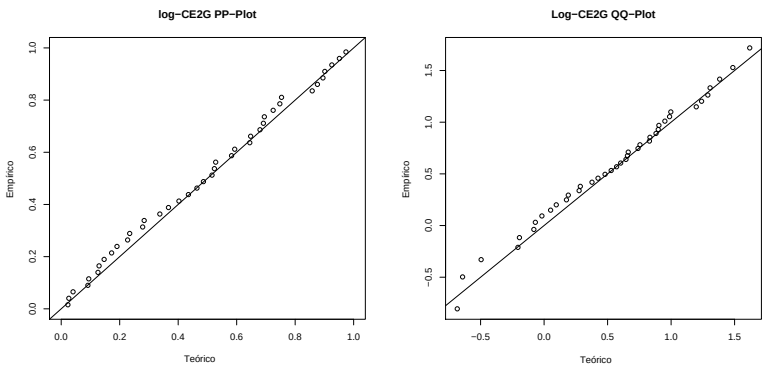

Figura 3.6: Esquerda: Gráfico de Probabilidades (PP-plot) para o modelo log-CE2G. Direita: Gráfico de Quantil-Quantil (QQ-Plot) para o modelo log-CE2G.

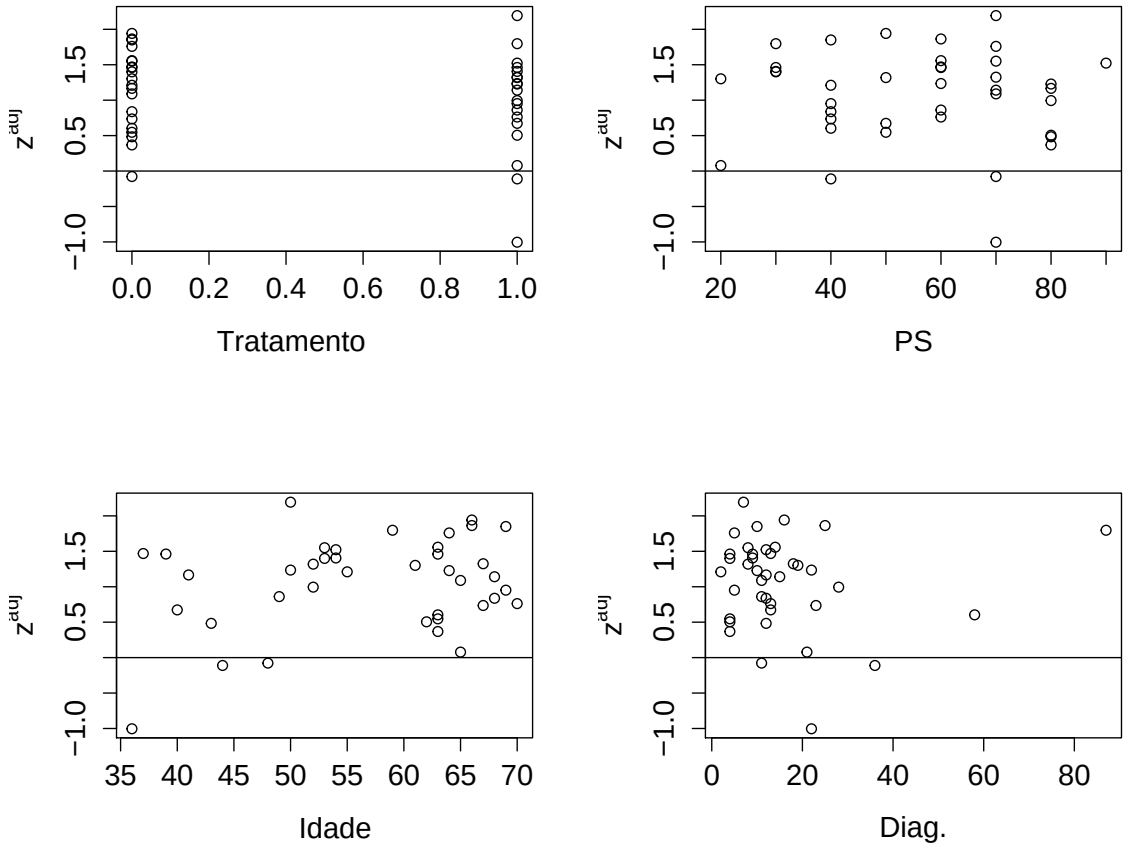

Figura 3.7: Resíduos ajustados contra as covariáveis do modelo log-CE2G para os dados de câncer de pulmão avançado.

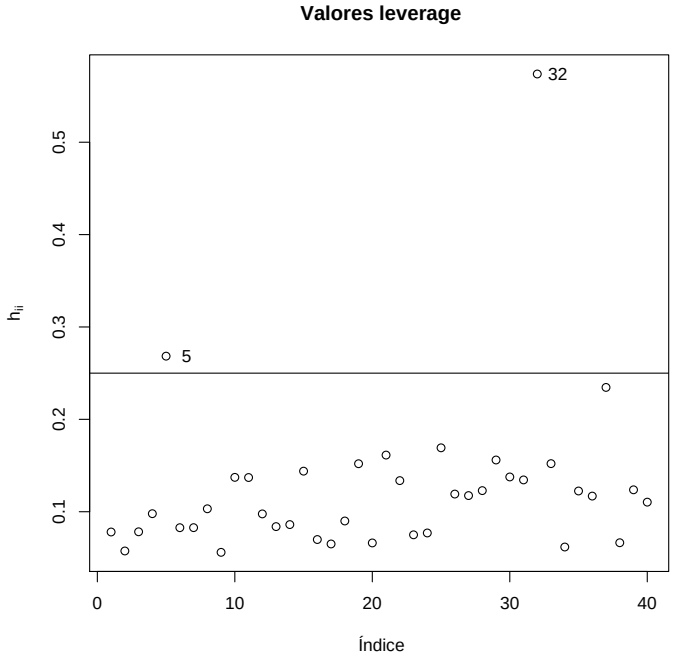

Figura 3.8: Valores *leverage*, h_{ii}, para a amostra de câncer de pulmão avançado no modelo log-CE2G.

as observações em G são influenciáveis ao menos em um dos parâmetros. Quando observado mais atentamente o conjunto G, notamos que há uma grande discrepância nas medidas de y contra as demais do mesmo nível em PS. Portanto, já que a variável PS é significante ao nível $\alpha = 0.05$ os indivíduos que compõem o conjunto G devem ser monitorados com uma atenção maior durante o estudo.

Tabela 3.4: Valores de LD para o conjunto $G = \{5, 32\}$ e a medida de Influência $\Delta\psi$.

i	LD_{-i}	$\Delta\psi_{-i}$							
		β_0(Int.)	β_1(Tratamento)	β_2(PS)	β_3(Idade)	β_4(diag)	σ	α	θ
5	3.7855	-14.002	-0.44	-0.05	-0.01	0.58	5.18	825.22	-11.61
32	5.9523	-12.500	-0.29	0.26	0.07	-2.16	4.83	944.42	-10.00

3.12 Conclusão

O modelo de regressão log-CE2G com presença de dados censurados foi desenvolvido com base na distribuição CE2G que apresenta funções de risco crescente, decrescente e em forma de banheira, o que pode tornar mais atraente para princípios de aplicação prática. Discutimos a aplicação de análise de resíduos e pontos influentes no modelo de regressão log-CE2G. O método

de estimação apresentado é a aplicação direta da máxima verossimilhança e não apresenta problemas práticos. Todo o conteúdo apresentado foi aplicado em um conjunto de dados reais, o qual indica que o modelo pode ser útil em situações práticas.

Para obter uma comparação do modelo de regressão log-CE2G com os modelos usuais de tempo de vida logarítima, recorremos ao conjunto de dados de pacientes com câncer de pulmão avançado, e ajustamos os seguintes modelos: a distribuição de valor extremo (log-Weibull) com f.d.p. dada por $f(y) = \sigma^{-1}\exp\left((y-\mu)/\sigma - exp(-(y-\mu)/\sigma\right)$, a distribuição log-Normal com f.d.p. dada por $f(y) = \sqrt{(2\pi\sigma)^{-1}\exp\left(-(y-\mu)/(2\sigma)\right)}$ e a distribuição log-Logística com f.d.p. dada por $f(y) = \sigma^{-1}\exp\left(-(y-\mu)/\sigma\right)/(1+\exp\left(-(y-\mu)/\sigma\right))^2$. As funções de log-verossimilhança para os MLE's apresentam os valores 58.0220, 58.2198, 58.8156 e 58.7242 para os modelos log-CE2G, log-Normal, log-Weibull e log-Logistico, respectivamente, assumindo a relação $\mu = \beta_0 + \beta_1 I(Treatmento = teste) + \beta_2 PS + \beta_3 Idade + \beta_4 diag$. Estes valores indicam que o modelo log-CE2G é um modelo que também pode ser adotado, assim como os demais no ajuste de dados e em particular para o conjunto de dados apresentado.

Os resultados deste capítulo foram submetidos e já aceito para publicação na revista *Communications in Statistics – Theory and Methods* (Louzada e Marchi, 2015).

Capítulo 4

Modelos de regressão alternativos para distribuições discretas com Inflação de Zeros

A modelagem para dados discretos está presente em muitas áreas da ciência como seguros, saúde pública, epidemiologia e psicologia, tendo como principal modelo desenvolvido o modelo de regressão de Poisson que tem como característica que o valor da média é igual ao da variância, ou seja, uma suposição para a análise prática não muito conveniente pois muitos conjuntos de dados podem apresentar superdispersão (ou subdispersão). Para este caso, o modelo de regressão de Poisson subestima a dispersão das amostras observadas. O motivo da superdispersão ainda varia devido a muitas situações particulares. A inflação de zeros, que é uma manifestação de superdispersão, significa que os resultados tidos como zero são maiores que o esperado. Este acontecimento pode ser de interesse desde que a incidência de zeros frequentemente está associada a casos práticos especiais do objeto e pode ser interessante descobrir as causas destes zeros em excesso observados. Por exemplo, Ridout et al. (2001) sugeriram que, em monitoramento da quantidade de doenças que provoquem lesões em plantas, uma planta pode não ter lesões por causa de sua resistência à doença, ou simplesmente porque a doença não afeta a planta em estudo. Portanto a existência de inflação de zeros é um incentivo para desenvolver modelos que se adequam ao conjunto de dados. A ideia básica atrás dos modelos derivados da inflação de zeros (ZI) é a mistura de uma distribuição degenerada em zero com uma distribuição como a Poisson, Binomial, Binomial Negativa, entre outros que têm suporte nos números inteiros não negativos.

O primeiro conceito de uma distribuição de inflação de zeros foi originada do trabalho

de Cohen (1963) que misturou distribuições de Poisson. Mullahy (1986) discutiu o problema de inflação de zeros em econometria. Lambert (1992) introduziu a regressão na distribuição de Poisson Inflacionada de Zeros (ZIP) e ilustrou os resultados em uma análise de dados relacionados aos defeitos de produção. Hall (2000) fez uma extensão do modelo de Lambert e a metodologia para situações limitadas superiormente na contagem, que como resultado obteve a regressão na distribuição Binomial Inflacionada de Zeros (ZIB). Outro modelo bem popular é o modelo de regressão na distribuição Binomial Negativa Inflacionada de Zeros (ZINB), a qual é definida similarmente pela mudança da distribuição Poisson pela Binomial Negativa na obtenção do modelo ZIP. O modelo ZINB está bem discutido em Ridout et al. (2001), onde foi desenvolvido uma estatística para testar entre os modelos de regressão ZIP e ZINB. Mwalili et al. (2008) ilustraram como o modelo de regressão ZINB pode ser utilizado corretamente para a má especificação do modelo.

A distribuição de Série de Potência Generalizada com parâmetro de Inflação (IGPS) foi introduzida por Kolev et al. (2000) para modelar dados de contagem que apresentam originalmente superdispersão. Esta distribuição é uma extensão da distribuição Série de Potência (Gupta, 1974; Consul, 1990) pela inclusão de um parâmetro adicional ρ. A interpretação natural deste parâmetro em termos é a proporção de "inflação de zeros" e como coeficiente de correlação. A distribuição IGPS também pode ser avaliada no contexto de fração de cura em modelos de sobrevivência, veja Borges et al. (2012). Entretanto, não temos conhecimento do estudo para determinar o comportamento da distribuição IGPS quando utilizada para modelos de contagem de dados com excesso de zeros. Portanto, propomos um modelo de regressão para dados de contagem com excesso de zeros baseado na distribuição IGPS, a qual possui como diferencial o fato de não assumir um modelo de mistura de duas distribuições, o que é usualmente feito em modelos de regressão com inflação de zeros.

4.1 Construção do modelo de regressão

4.2 A distribuição IGPS

A distribuição IGPS de variável Y tem a f.p. dada por

$$\mathbb{P}[Y=y;\theta,\rho]=\frac{1}{g(\theta)}\sum_{y_1,y_2,\ldots} y_n\big[\theta(1-\rho)\big]^{\sum_{i=1}^{\infty} y_i}\rho^{\sum_{i=2}^{\infty}(i-1)y_i},\quad y=0,1,2,\ldots,\ \rho\in[0,1), \quad (4.2.1)$$

em que y_n depende somente de y, $g(\theta) = \sum_{y=0}^{\infty} a_y \theta^y$ é positiva, finita e diferenciável com $\theta \in (0, s)$ (s pode ser ∞) de tal forma que $g(\theta)$ é finita, e o somatório é feito sobre todo o conjunto de inteiros não negativos $y_1, y_2, \ldots$ de tal forma que $\sum_{i=1}^{\infty} i y_i = y$. Para mais detalhes da distribuição IGPS, o leitor por consultar Kolev et al. (2000) e Minkova (2002). No desenvolvimento deste capítulo, o parâmetro ρ representará a proporção adicional de zeros. Altos valores de ρ indicam uma alta proporção de zeros, enquanto $\rho \to 0$ implica em menor adição de zeros. É interessante notar que quando $\rho = 0$ a distribuição IGPS se reduz a distribuição de Série de Potência Generalizada (Gupta, 1974; Consul, 1990). A Tabela 4.1, extraída de Borges et al. (2012), apresenta as escolhas de a_y, $g(\theta)$ e o parâmetro θ correspondentes aos casos especiais da distribuição IGPS, nomeadas por Parâmetro Inflacionado de Poisson (IPP), Binomial Negativa (IPNB), Binomial (IPB) e Série Logarítima (IPLS).

Tabela 4.1: Escolhas de a_y, $g(\theta)$ e o parâmetro θ dos casos particulares da distribuição IGPS.

Distribuição	a_y	$g(\theta)$	θ	s
IPP	$\frac{1}{y_1! y_2! \cdots}$	e^{θ}	η	∞
IPB	$\begin{pmatrix} m \\ m - y_1 - y_2 - \cdots, y_1, y_2, \ldots \end{pmatrix}$	$(1+\theta)^m$	$\frac{\pi}{1-\pi}$	1
IPNB	$\frac{\Gamma\left(\phi^{-1} + \sum_{i=1}^{\infty} y_i\right)}{\Gamma(\phi^{-1})\left[\sum_{i=1}^{\infty} y_i\right]!}$	$(1-\theta)^{-\phi^{-1}}$	$\frac{\phi\eta}{1+\phi\eta}$	∞
IPLS	$\frac{(-1+y_1+y_2+\cdots)!}{y_1! y_2! \cdots}$	$-\log(1-\theta)$	$1-\pi$	1

4.3 Modelos de regressão IPP e IPNB

Visto que os modelos de regressão ZIP e ZINB são atualmente os modelos mais utilizados para análise de dados de contagem com excesso de zeros, voltamos a atenção aos modelos alternativos IPP e IPNB. Desta forma, para determinar o modelo de regressão IPP assumimos que a variável resposta Y_i $(i = 1, \ldots, n)$ tem distribuição IPP com f.d. dada por

$$\mathbb{P}[Y_i = y_i] = \begin{cases} e^{-\lambda_i}, & \text{se } y_i = 0 \\ \lambda_i(1-\rho_i)\rho_i^{y_i - 1} e^{-\lambda_i/\rho_i} \; {}_1F_1\left[y_i + 1; 2; \frac{\lambda_i(1-\rho_i)}{\rho_i}\right], & \text{se } y_i = 1, 2, \ldots \end{cases}, \quad (4.3.1)$$

em que os parâmetros $\lambda_i > 0$ e $\rho_i \in [0,1)$ são dependentes de vetores de variáveis explicativas (covariáveis) $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{k_1 i})$ e $\mathbf{z}_i = (1, z_{1i}, \ldots, z_{k_2 i})$, respectivamente, e ${}_pF_q[a_1; \ldots; a_q; b_1, \ldots, b_p; , z)$ é a função hipergeométrica, definida por

$$ {}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; w) = \sum_{k=0}^{\infty} \frac{w^k \prod_{j=1}^{p} \Gamma(a_j + k)\Gamma^{-1}(a_j)}{\Gamma(k+1) \prod_{j=1}^{q} \Gamma(b_j + k)\Gamma^{-1}(b_j)}, $$

em que $\Gamma(\cdot)$ é a função Gama. A função hipergeométrica é facilmente calculada e está disponível em várias plataformas de *software* como o R Core Team (2012), MapleSoft (2012), entre outros. Como em Lambert (1992), propomos relacionar λ_i com as covariáveis $\mathbf{x}_i$ pela função de ligação logarítima e ρ as covariáveis $\mathbf{z}_i$ pela função de ligação logística, isto é

$$ \log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\beta} \quad \text{e} \quad \log\left(\frac{\rho_i}{1-\rho_i}\right) = \mathbf{z}_i'\boldsymbol{\gamma}, \tag{4.3.2} $$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k_1})$ e $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{k_2})$ são parâmetros desconhecidos. Desta forma, as equações (4.3.1) e (4.3.2) definem o modelo de regressão IPP.

Observação 4.3.1 A média e a variância da distribuição IPP são $\mathbb{E}[Y_i] = \frac{\lambda_i}{1-\rho_i}$ e $\mathbb{V}[Y_i] = \frac{\lambda_i(1+\rho_i)}{(1-\rho_i)^2}$, respectivamente. Logo, temos que $\mathbb{V}[Y_i] > \mathbb{E}[Y_i]$ se $\rho_i > 0$, o que caracteriza a super-dispersão.

Observação 4.3.2 A distribuição IPP ao contrário das usuais distribuições de inflação de zeros não usa a mistura de distribuições e segue que o parâmetro ρ interfere na proporção de zeros no processo de Poisson usual. Além disso, o parâmetro ρ é uma medida de dependência entre o valor de contagem resultante. Para maiores detalhes recomenda-se a leitura de Kolev et al. (2000).

Considere agora, como alternativa ao modelo de regressão ZINB, o modelo de regressão IPNB obtido pela afirmação que a variável resposta Y_i tem distribuição IPNB com f.d.p. dada por

$$ Pr[Y_i = y_i] = \begin{cases} (1+\phi)^{-\frac{\lambda_i}{\phi}}, & \text{se } y_i = 0 \\ \lambda_i(1-\rho_i)\rho_i^{y_i + \lambda_i/\phi}(\phi+\rho_i)^{-\frac{\phi+\lambda_i}{\phi}} \; {}_2F_1\left[y_i+1; \frac{\phi+\lambda_i}{\phi}; 2; \frac{\phi(1-\rho_i)}{\phi+\rho_i}\right], & \text{se } y_i = 1, 2, \ldots \end{cases} \tag{4.3.3} $$

em que $\lambda_i > 0$ e $\rho_i \in [0,1)$ dependendo de vetores de covariáveis $\mathbf{x}_i$ e $\mathbf{z}_i$, respectivamente, com funções de ligação (4.3.2), e ϕ^{-1} ($\phi > 0$) como parâmetro de dispersão o qual assumimos que não depende de covariáveis.

Observação 4.3.3 Para a distribuição IPNB, consideramos a reparametrização $r = \frac{\lambda}{\phi}$ e $\pi = \frac{1}{1+\phi}$, com $\phi, \lambda > 0$ na distribuição *Inflated-Parameter Negative Binomial* em Minkova (2002). A média e a variância da distribuição IPNB são $\mathbb{E}[Y_i] = \frac{\lambda_i}{1-\rho_i}$ e $\mathbb{V}[Y_i] = \frac{\lambda_i(1+\rho_i+\phi)}{(1-\rho_i)^2}$, respectivamente. Esta distribuição se reduz a distribuição IPP no limite $\phi \to 0$.

4.4 Inferência

Para a inferência dos parâmetros vamos considerar a estimação de máxima verossimilhança e também determinaremos as expressões para a matriz de informação de Fisher observada. Assumindo que $y_1, \ldots, y_n$ são observações independentes da variável aleatória Y com f.d.p. (4.3.1) ou f.d.p. (4.3.3), os estimadores de máxima verossimilhança dos parâmetros para as distribuições IPP e IPNB, respectivamente, são obtidas pela maximização direta das funções de log-verossimilhança, dadas por

$$\begin{aligned}\ell(\theta) = \ell(\beta,\gamma) &= -\sum_{i=1}^{n} e^{\mathbf{x}_i'\beta} - \sum_{y_i>0} e^{\mathbf{x}_i'\beta-\mathbf{z}_i'\gamma} + \sum_{y_i>0}(\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma) - \sum_{y_i>0} y_i\left(\log\left(1+e^{\mathbf{z}_i'\gamma}\right) - \mathbf{z}_i'\gamma\right) \\ &+ \sum_{y_i>0}\log\left({}_1F_1\left[y_i+1;2;e^{\mathbf{x}_i'\beta-\mathbf{z}_i'\gamma}\right]\right)\end{aligned}$$

e

$$\begin{aligned}\ell(\theta) = \ell(\beta,\gamma,\phi) &= -\sum_{y_i=0}\phi^{-1}\log(1+\phi)e^{\mathbf{x}_i'\beta} + \sum_{y_i>0}\mathbf{x}_i'\beta + \sum_{y_i>0}\phi^{-1}e^{\mathbf{x}_i'\beta}\mathbf{z}_i'\gamma - \sum_{y_i>0} y_i\log(1+e^{\mathbf{z}_i'\gamma}) \\ &- \sum_{y_i>0}\left(1+\phi^{-1}e^{\mathbf{x}_i'\beta}\right)\log\left((1+\phi)e^{\mathbf{z}_i'\gamma}+\phi\right) \\ &+ \sum_{y_i>0}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right).\end{aligned}$$

Para amostras suficientemente grandes, os estimadores MLEs ($\hat{\theta}$) das distribuições IPP e IPNB devem ser assintoticamente distribuídos por uma distribuição Normal multivariada com vetor de médias θ e matriz de covariância igual à inversa da matriz de informação observada $\mathbf{I}$ segundo o Teorema Central do Limite. Os elementos da matriz $\mathbf{I}$ são dadas no Apêndice C.

Para comparar os resultados obtidos pelas diferentes distribuições ajustadas podemos usar o valor da log-verossimilhança calculada nos estimadores MLEs. A distribuição com o melhor ajuste corresponde ao maior valor encontrado. Além disso, pode ser de interesse determinar se dois modelos são equivalente sob condições gerais. Um teste baseado no teste da razão da verossimilhança que caracteriza a distribuição assintótica da estatística da razão de verossimilhança sob condições gerais é o teste de Vuong (1989). Por condições gerais é entendido que

os modelos podem ser encaixados, não encaixados, ou sobrepostos, ou ambos e ainda somente um, ou nenhum dos modelos competitivos podem conter a verdadeira estrutura de geração dos dados observados. Modelos sobrepostos são utilizados em economia e modelam agentes que vivem num espaço finito de tempo suficientemente longo para se sobrepor com pelo menos um período da vida de outro agente em que um agente é um tomador de decisões.

4.4.1 Teste de Voung

Para testar o modelo f contra o modelo g ($f \times g$), vamos utilizar o teste de Voung para distribuições assintóticas da estatística da razão de verossimilhança (LR) sob condições gerais, o qual é dado por

$$T_{fg} = \frac{1}{\omega\sqrt{n}}(l_f(\hat{\alpha}) - l_g(\hat{\psi})), \tag{4.4.1}$$

em que $l_f(\hat{\alpha})$ é a log-verossimilhança de f calculada em seus estimadores MLEs $\hat{\alpha}$ com

$$\omega^2 = \frac{1}{n}\sum_{i=1}^{n}(l_f(y_i|\hat{\alpha}) - l_g(y_i|\hat{\psi}))^2 - \left(\frac{1}{n}\sum_{i=1}^{n}[l_f(y_i|\hat{\alpha}) - l_g(y_i|\hat{\psi})]\right)^2.$$

Para modelos não encaixados, a estatística de teste T_{fg} converge em distribuição para uma distribuição Normal padrão. Seja c o valor de nível crítico para algum nível de significância especificado, então se a estatística de teste é maior que o valor c, rejeita-se a hipótese nula em favor do modelo f ser melhor que o modelo g. Se a estatística de teste tiver valor menor que -c, rejeitamos a hipótese nula em favor do modelo g ser melhor que o modelo f. Finalmente, se $|T_{fg}| \leq c$ a hipótese nula não é rejeitada e não podemos distinguir entre os dois modelos qual obteve o melhor ajuste.

4.5 Estudo de simulação

Neste estudo de simulação vamos analisar o comportamento assintótico dos estimadores MLEs dos modelos propostos. O estudo é baseado em 100 amostras geradas com três diferentes conjuntos de parâmetros, de tamanho amostrais $n = 30$, 50 e 100 e com covariáveis em cada um dos parâmetros λ e ρ. Para o calculo da variância dos parâmetros foi utilizado a inversa da matriz de informação $\mathbf{I}$. Os valores da variável X_{i1} foram gerados a partir da distribuição Gama com média 5/9 e variância $5/9^2$ e para Z_{i1} os valores foram gerados da distribuição Gama com média 2/8 e variância $2/8^2$. Estes valores foram escolhidos arbitrariamente de forma que

a média se mantivesse menor que 1 na distribuição Gama. A distribuição Gama foi preferida por representar uma medida contínua positiva, o que é frequentemente observado na área de sobrevivência.

4.5.1 Distribuição IPP

As amostras de tamanho n para a distribuição IPP foram geradas usando um conjunto de três configurações para os parâmetros na transformação $\log(\lambda)$ e para o parâmetro de inflação ρ.

A Tabela 4.2 apresenta os valores da média dos 100 MLEs (Av), seus desvios padrões médios (Sd), a probabilidade de cobertura para intervalos 95% de confiança (C), e o erro quadrático médio (MSE). Os resultados sugerem que os estimadores MLEs se comportam adequadamente, ou seja, os desvios padrões dos estimadores MLEs decrescem quando o tamanho da amostra aumenta e a probabilidade de cobertura são próximos ao valor escolhido de 95%, se apresentando mais próximos com o aumento do tamanho da amostra.

Tabela 4.2: Média dos estimadores MLEs da IPP, seus Desvios Padrões, Cobertura e MSE.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
Valor verdadeiro	1.0	-0.4	0.5	1.5	-0.4	1.1	-0.5	1.8	1.0	0.3	-1.0	0.5
$n = 30$												
Av	1.1334	-0.4819	0.3646	1.0683	-0.4392	1.1604	-0.5907	1.3214	1.0318	0.2755	-1.2236	0.0695
Sd	0.4512	0.6417	0.5193	1.2305	0.4316	0.6262	0.8156	2.4635	0.2865	0.4649	0.9655	2.4957
C(95%)	0.8800	0.9000	0.9600	0.9400	0.9400	0.9700	0.9700	0.9600	0.9800	0.9500	0.9800	0.9900
MSE	0.2193	0.4144	0.2853	1.6853	0.1859	0.3919	0.6668	6.2372	0.0823	0.2146	0.9729	6.3514
$n = 50$												
Av	1.0406	-0.4328	0.4223	1.3706	-0.3495	1.0700	-0.6789	1.8322	0.9994	0.3572	-1.2325	0.4228
Sd	0.3236	0.5149	0.3914	1.4205	0.4573	0.7236	0.8057	2.1142	0.2520	0.3997	0.7773	2.2492
C(95%)	0.9300	0.9300	0.9400	0.9400	0.9200	0.9400	0.9200	0.9600	0.9400	0.9200	0.9800	0.9900
MSE	0.1053	0.2636	0.1577	2.0143	0.2096	0.5193	0.6747	4.4264	0.0629	0.1614	0.6523	5.0144
$n = 100$												
Av	1.0178	-0.4133	0.4512	1.5236	-0.4206	1.1312	-0.6039	1.8894	1.0183	0.3114	-1.1482	0.3062
Sd	0.2630	0.4470	0.2275	0.6652	0.2950	0.4358	0.3369	0.8496	0.1462	0.2244	0.5421	1.4041
C(95%)	0.9100	0.9000	0.9700	0.9800	0.9000	0.9300	0.9800	0.9500	0.9400	0.9400	0.9800	0.9800
MSE	0.0688	0.1980	0.0536	0.4386	0.0866	0.1890	0.1231	0.7226	0.0215	0.0500	0.3128	1.9895

4.5.2 Distribuição IPNB

As amostras de tamanho n para a distribuição IPP foram geradas usando um conjunto de três configurações para os parâmetros na transformação $\log(\lambda)$, para o parâmetro de inflação ρ e para ϕ.

A Tabela 4.3, assim como a Tabela 4.2, apresenta os valores da média dos 100 MLEs (Av), seus desvios padrões médios (Sd), a probabilidade de cobertura para intervalos 95% de confiança (C), e o erro quadrático médio (MSE). Os resultados sugerem que os estimadores MLEs se comportam adequadamente, ou seja, os desvios padrões dos estimadores MLEs decrescem quando o tamanho da amostra aumenta e a probabilidade de cobertura são próximos ao valor escolhido de 95%, se apresentando mais próximos com o aumento do tamanho da amostra.

Tabela 4.3: Média dos estimadores MLEs da IPNB, seus Desvios Padrões, Cobertura e MSE.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\phi}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\phi}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\phi}$
Valor Verdadeiro	1.5	-0.4	1.0	1.4	1.0	-0.5	1.4	-0.3	0.2	0.6	0.9	-0.3	0.7	-0.2	1.4
							$n = 30$								
Av	1.4658	-0.3688	0.8005	1.8578	1.2872	-0.5853	1.2319	-0.1257	-0.1998	0.2187	0.7045	-0.2751	0.8548	-0.3472	1.1210
Sd	0.6440	0.6786	0.9683	1.9795	2.3860	0.6257	0.8776	0.8721	2.1812	0.5733	0.8963	1.4361	1.0717	2.3488	2.4069
C(95%)	0.9800	0.9500	0.9900	0.9700	0.9500	0.9400	0.8800	0.9600	0.9800	0.7700	0.9300	0.9200	0.9300	0.9600	0.8400
MSE	0.6822	0.7063	1.6015	5.4677	11.6132	0.3949	0.7908	0.7834	4.8699	0.4708	0.8334	2.0425	1.1611	5.4832	5.8133
							$n = 50$								
Av	1.6490	-0.4465	0.7111	1.3686	2.2056	-0.5813	1.3507	-0.2544	0.1388	0.4230	0.7685	-0.2757	0.7507	-0.4894	1.5477
Sd	0.6572	0.7375	0.9817	1.0458	3.2648	0.4246	0.5583	1.0786	1.9165	0.7574	0.5702	0.6625	0.9404	1.4164	2.4175
C(95%)	0.9300	0.8800	0.9900	0.9500	0.9200	0.9900	0.9400	0.9200	0.9600	0.8400	0.9700	0.9400	0.9200	0.9800	0.8100
MSE	0.4498	0.5406	1.0376	1.0837	12.0061	0.1851	0.3111	1.1539	3.6400	0.5992	0.3392	0.4351	0.8781	2.0698	5.8075
							$n = 100$								
Av	1.5712	-0.4086	0.8110	1.5070	1.7876	-0.5931	1.3638	-0.1999	0.1473	0.3728	0.8022	-0.2868	0.7512	-0.3733	1.4967
Sd	0.4626	0.3332	0.8019	0.8073	2.3714	0.3415	0.3717	0.7684	1.3004	0.6956	0.4718	0.5222	0.7969	0.9141	1.8958
C(95%)	0.9800	0.9300	0.9800	0.9700	0.9600	0.9900	0.9600	0.9400	0.9800	0.9200	0.9800	0.9100	0.9500	0.9700	0.8500
MSE	0.2169	0.1100	0.6723	0.6567	6.1875	0.1241	0.1381	0.5946	1.6770	0.5307	0.2300	0.2701	0.6313	0.8572	3.5675

4.5.3 Diagnóstico de má especificação do modelo

Podemos detectar que o modelo não é adequado aos dados examinando os resultados da estimação juntamente com os dados. O objetivo é verificar que as suposições na qual a análise está baseada são pouco precisas, e portanto, detectar e corrigir alguma má especificação existente no modelo.

4.5.4 Diagnóstico de má especificação do modelo por gráficos

Se a variável Y_i satisfaz a condição $E(Y_i) = f(x_i, \theta)$, então os valores $f(x_i, \hat{\theta})$ estimam o valor esperado de Y_i. Para detectar uma má escolha da função de regressão, podemos utilizar o gráfico dos valores estimados por $f(x_i, \hat{\theta})$ e os observados y_i contra os valores x_i ou somente o valor das estimativas contra os valores observados.

Considere os resíduos dados por $\epsilon_i = Y_i - f(x_i, \theta)$, para $i = 1, \ldots, n$. Para modelos de dados de contagem nenhum resíduo tem média zero, variância constante e função de dis-

tribuição simétrica. Temos também que ϵ são heterocedásticos e uma forma de corrigir esta característica é dividir os resíduos pelo desvio padrão, denominados aqui por erros padrões, $e_i = \frac{Y_i - f(x_i,\theta)}{\sigma_i}$, para $i = 1, \ldots, n$. Portanto, uma ideia natural e comum, é estimar os erros e os erros padrões, e então estudar o seus comportamentos graficamente para verificar a heterocedasticidade.

Os resíduos

$$\hat{e}_i = \frac{y_i - f(x_i, \hat{\theta})}{\hat{\sigma}_i}, \text{ para } i = 1, \ldots, n \tag{4.5.1}$$

são estimativas para os erros padronizados. A má especificação do modelo é usualmente detectada pela examinação dos resíduos gráficos. Para detectar a má escolha do modelo de regressão vamos utilizar o gráfico dos resíduos $\hat{e}_i$ contra as covariáveis x_i.

4.5.5 Diagnóstico de má especificação do modelo por testes

No diagnóstico anterior para determinar a má especificação do modelo estávamos focados em determinar a capacidade do modelo obsevando individualmente os resultados do ajuste. Vamos agora considerar toda a capacidade do modelo de apresentar um bom ajuste em uma única medida.

Uma medida muito utilizada para medir a bondade do ajuste (*goodness of fit*) para algum modelo discreto Y_i com média $f(x_i, \theta)$ e variância σ_i^2 é a estatística de Pearson, dada por

$$P = \sum_{i=1}^{n} \frac{(y_i - f(x_i, \hat{\theta}))^2}{\hat{\sigma}_i^2}.$$

Se a média e a variância estão corretamente especificada, então $E(P) = n$, pelo motivo de P ter distribuição aproximada de uma χ^2 (Qui-Quadrado), porém esta afirmação é verdadeira somente em casos especiais de dados agrupados com múltiplas observações para cada $f(x_i, \theta)$. Na prática o valor P é comparado com uma χ^2 com $n - k$ graus de liberdade, refletindo uma correção devido a estimação por $f(x_i, \theta)$, em que k é o número de variáveis regressoras incluindo o intercepto na função $f(x_i, \theta)$.

Para os modelos paramétricos, uma diagnóstico rápido porém não muito eficiente é comparar os valores estimados por suas probabilidades com a sua frequência observada, em que a distribuição de frequências é calculada como a médias sobre todas as observações dos valores das probabilidades estimados pelo ajuste. Para comparar as probabilidades ajustadas com as observadas podemos utilizar a estatística de teste Qui-Quadrado.

A estatística do teste (T_χ) conhecida por teste de bondade do ajuste Qui-Quadrado (*Chi Square Goodness-of-Fit test*), é dada por

$$T_\chi = \sum_{j=1}^{J} \frac{(n\bar{p}_j - n\hat{p}_j)^2}{n\hat{p}_j}, \tag{4.5.2}$$

em que J é o número de cédulas (grupos de observações), $\bar{p}_j$ é a proporção de amostras que $y = j$ e $\hat{p}_j = n^{-1}\sum_{i=1}^{n} p_{ij}$, com p_{ij} representando as probabilidades estimadas para cada observação i que está na j-ésima cédula. Para aplicar o teste, cada cédula de valores esperados tem que ter um número superior a 5 observações e assim será aceitável utilizar o teste de Qui-Quadrado. Se uma cédula contiver um número menor que 5, deverá ser agrupados cédulas até que cada uma delas contenham 5 ou mais observações. Na prática, o valor T_χ é comparado com o valor de χ^2_{J-1}.

4.6 Aplicação

Com aplicação a dados reais, vamos comparar as distribuições propostas IPP e IPNB com as distribuições usuais ZIP e ZINB-2 (em que o 2 indica uma parametrização da Binomial Negativa) em dois conjuntos de dados extraídos da literatura.

O primeiro conjunto de dados, denominado por BAP ou $T8$, foi extraído de Ridout et al. (2001) e registra o número de raízes produzidas por 270 aplicações de micropropagação em maçãs colunar Trajan. O interesse, ou seja, a variável resposta, é o número de raízes produzidas pelas aplicações de micropropagação. As aplicações foram feitas de 8 ou 16 horas de foto-período em um sistema de cultura que utiliza um ou quatro diferentes concentrações de cytokinin BAP na cultura. Para nosso modelo de regressão vamos considerar somente que existem valores distintos de λ e ρ, para cada uma das quatro diferentes concentrações de cytokinin BAP para os dois foto-períodos também, considerando o valor BAP=2.2 como valor de referência para as variáveis dummy criadas para BAP e para o foto-período o valor 8 como referência.

Usando o teste de superdispersão de Ridout et al. (2001) com $c = 1$ nos dados BAP, temos que $T = 4.04$ e portanto esperamos que os modelos com superdispersão tenham bons ajustes para os dados BAP. A Figura 4.1 mostra forte evidência de inflação de zeros como esperado, pela formulação e resultado do teste.

A Figura 4.2 apresenta as diferenças entre as proporções estimadas menos a proporção empírica dos resultados para os dados BAP utilizando os modelos de regressão IPP, ZIP, ZINB e IPNB. Podemos notar que os modelos IPNB e ZIP aparentemente apresentam os melhores

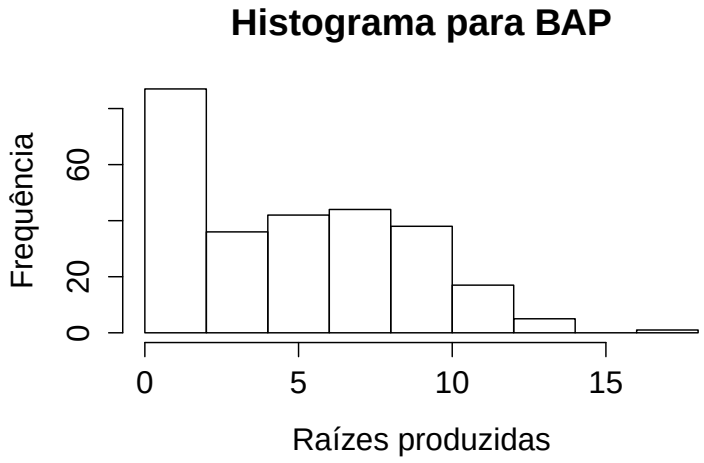

Figura 4.1: Histograma do número de raízes produzidas em BAP (T8).

ajuste no conjunto de dados BAP, e temos que o modelo IPNB aparentemente é o melhor ajustado.

A Tabela 4.4 apresenta os valores para a estatística T_χ para (4.5.2) em cada um dos modelos já mencionados aqui. Pela Tabela 4.4 podemos concluir que o modelo IPNB é o mais adequado para o conjunto de dados BAP. Para calcular (4.5.2) foram criadas 13 cédulas, agrupando valores de $y \geq 13$ em uma mesma cédula.

Tabela 4.4: Teste de bondade do ajuste Qui-Quadrado para os modelos IPP, ZIP, IPNB e ZINB nos dados BAP.

Modelo	IPP	ZIP	IPNB	ZINB
T_χ	49.0527	40.5669	11.0346	38.0875

A Figura 4.3 apresenta os erros padrões estimados contra as covariáveis para a análise da má especificação dos modelos. Podemos observar que as médias dos erros, nos modelos ZIP e ZINB, mudam para a covariável Fotoperíodo (quarta coluna) e a variância parece permanecer constante os modelos ZINB e IPNB para as covariáveis.

Do fato que menos a logverossimilhança dos modelos ZIP, IPP, IPNB e ZINB sã 627.9028, 620.9083, 620.9083 and 620.0298 vamos calcular o teste de Voung para determinar qual modelo se aproxima mais do modelo verdadeiro para o conjunto de dados BAP.

A Tabela 4.5 apresenta os valores T_{fg} do valore (4.4.1) do teste de Voung sobre a hipótese nula. Pela Tabela (4.4.1) podemos concluir que não é possível dintinguir qual modelo é melhor ajustado pois como o teste é baseado numa distribuição normal padrão e tomando $c =$

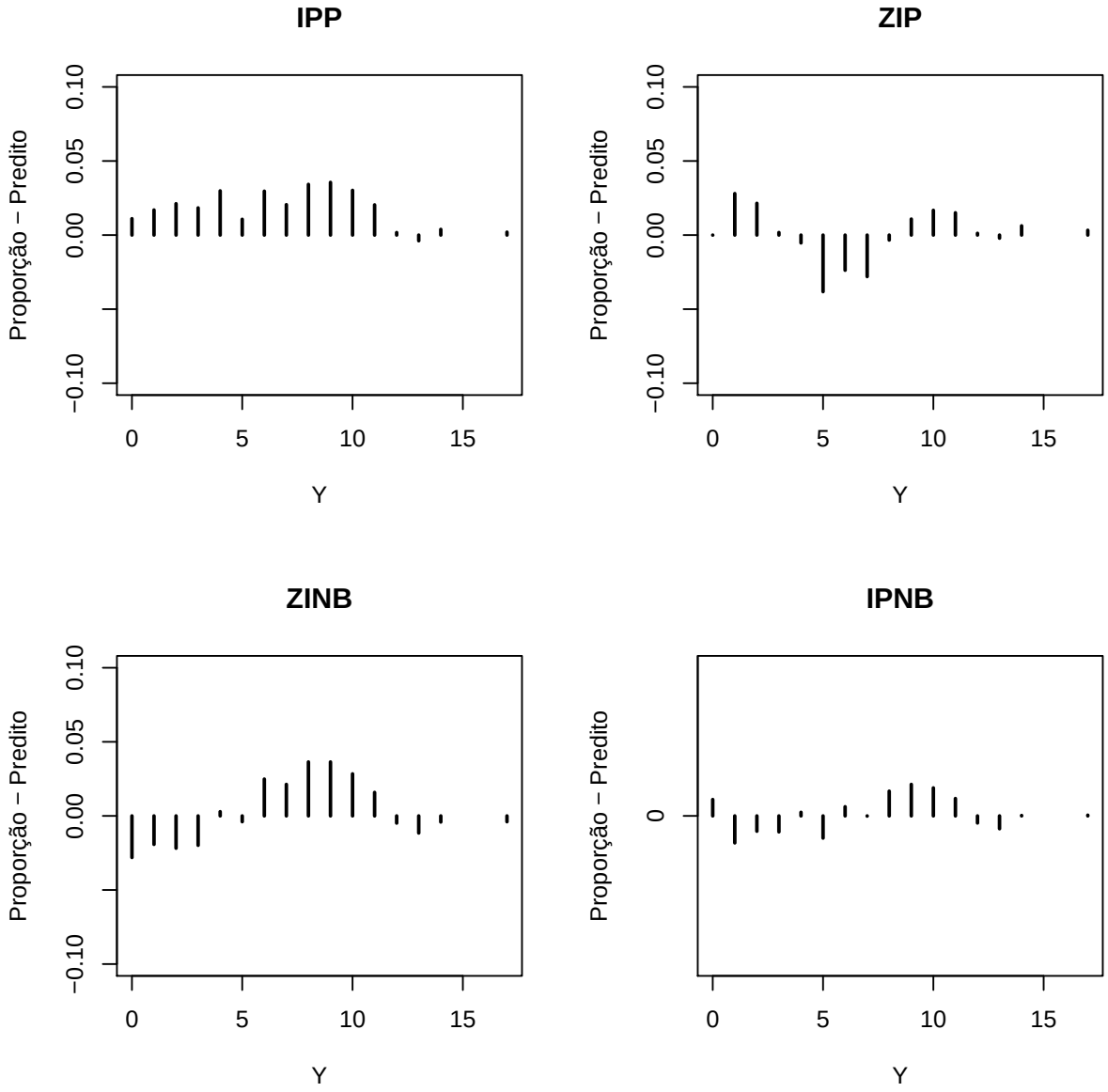

Figura 4.2: Diferenças entre as proporções estimadas menos a proporção empírica dos resultados para os dados BAP utilizando os modelos IPP, ZIP, ZINB e IPNB.

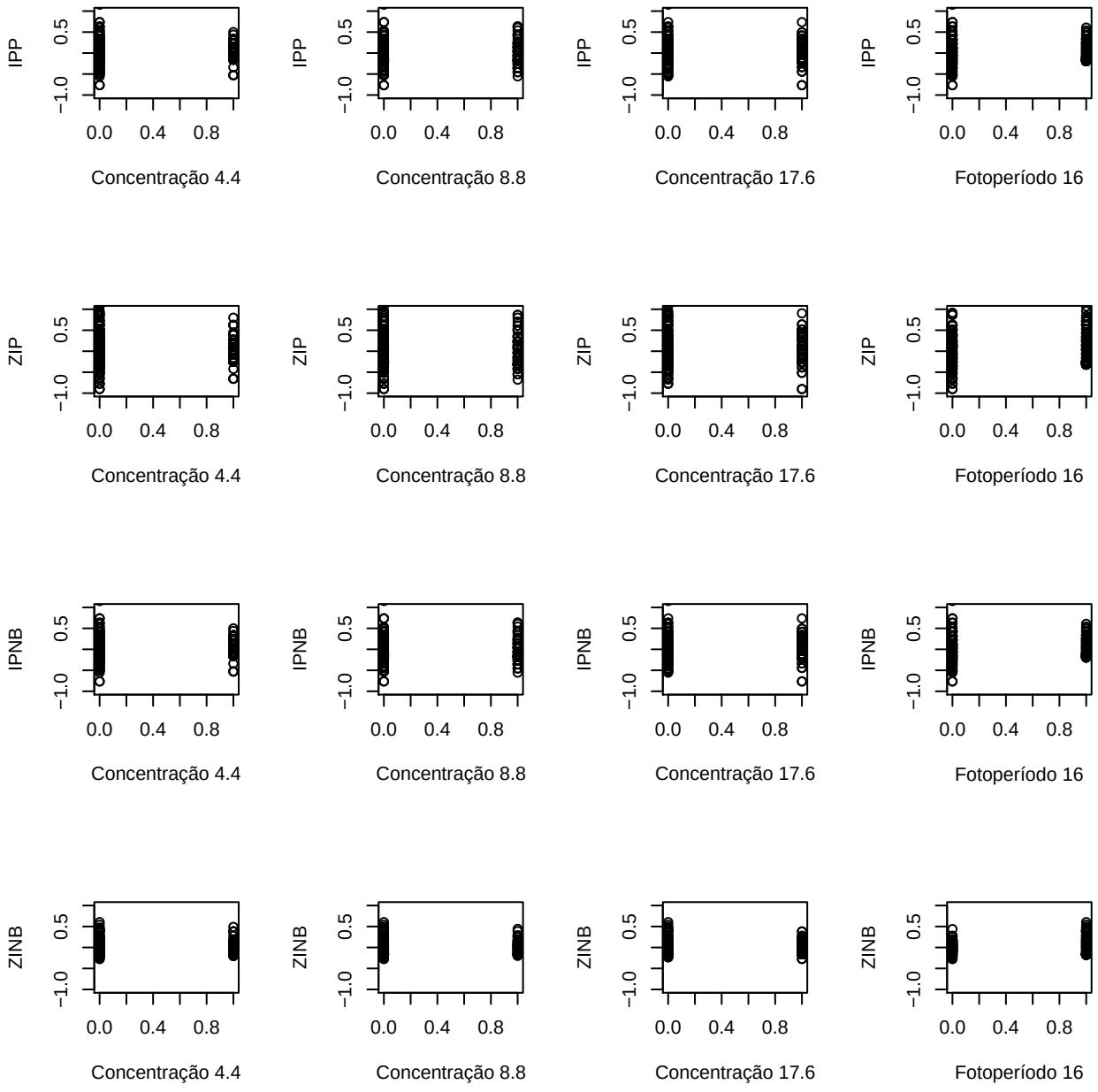

Figura 4.3: Resíduos padronizados para os modelos IPP, ZIP, IPNB e ZINB *versus* covariáveis para os dados BAP.

1.96 para a região de rejeição o teste conclui que todos os modelos tiveram ajustes semelhantes pois todos os valores em módulo são menores que c. É importante notar que para usar o teste de Voung, neste caso, devemos considerar que os modelos encaixados não são equivalentes. Note ainda que se escolhermos $c = 0.5$, ou seja, a região de rejeição é de meio desvio padrão da normal, então o teste indicaria que (IPP e IPBN)>ZIP e IPBN>ZINB, onde > indica um melhor ajuste.

Tabela 4.5: Valores de T_{fg} do teste Voung para os modelos IPP, ZIP, ZINB e IPNB para BAP.

	Teste de modelo $f \times g$					
	IPP × ZIP	IPP × ZINB	IPP × IPNB	IPNB × ZIP	IPNB × ZINB	ZINB × ZIP
T_{fg}	0.7321	-0.1195	0.0940	0.7321	0.74922	-0.1195

A Tabela 4.6 apresenta as estimativas dos parâmetros dos quatro modelos para os dados BAP. Note que os valores estimados para as distribuições IPP e IPBN aparentemente são os mesmos, o que pode ser explicado pelo valor considerado pequeno, porem significante, de ϕ e o arredondamento na quarta casa decimal. Ainda, note que mesmo tendo valores tão próximos os dois modelos são bem distintos quanto a estimação da proporção (Veja Figura 4.2).

Tabela 4.6: Modelos Inflacionados ajustados no conjunto BAP (T8).

		IPP		IPNB		ZIP		ZINB	
	Par.	Est.	Var	Est.	Var	Est.	Var	Est.	Var
conc.	β_0	1.4867	0.0189	1.4867	0.0190	1.8848	0.0039	1.8901	0.0059
4.4	β_1	0.3651	0.0213	0.3651	0.0214	0.1627	0.0068	0.1562	0.0104
8.8	β_2	0.3898	0.0194	0.3898	0.0194	0.1080	0.0062	0.0998	0.0093
17.6	β_3	0.3338	0.0192	0.3338	0.0192	0.07518	0.0063	0.0584	0.0096
photo.	β_4	-2.0340	0.0213	-2.0340	0.0213	-0.2767	0.0037	-0.2786	0.0055
conc.	γ_0	-1.1245	0.2116	-1.1245	0.2124	-4.2492	0.6610	-4.4492	0.9596
4.4	γ_1	-0.4872	0.1817	-0.4871	0.1823	0.1342	0.2544	0.1413	0.2603
8.8	γ_2	-0.6876	0.1609	-0.6876	0.1618	-0.3980	0.2632	-0.3984	0.2701
17.6	γ_3	-0.7469	0.1666	-0.7469	0.1667	0.0906	0.2242	0.0603	0.2339
photo.	γ_4	2.6203	0.1956	2.6203	0.1951	4.1784	0.5946	4.3723	0.9070
ϕ		-	-	$1.8\ 10^{-6}$	$9.2\ 10^{-9}$	-	-	2.5694	0.1114

Legenda: Par=Parâmetros, Est=Estimativa, Var=Variância, conc=Concentração e photo=Fotoperíodo

Levando em consideração todos os resultados das Tabelas 4.4 e 4.5, e das Figuras 4.2 e 4.3, concluímos que o modelo de regressão que melhor se ajusta no conjunto de dados BAP é o modelo IPNB mesmo não apresentando a maior logverossimilhança.

O segundo conjunto de dados utilizado, denominado BA2004 ou $T9$, obtidos em SEI (2004) é composto por registros do ano de 2004 nos municípios do estado da Bahia, Brasil. O objetivo é explicar os casos de dengue sobre certas condições. Foram registrados 132 casos onde não houve ocorrências de dengue, ou seja, $y = 0$. A variável resposta são os casos de dengue com covariáveis X_1: casos de AIDS, X_2: casos de lectospirose, X_3: casos de tuberculose pulmonar, X_4: densidade populacional e X_5: IDH, para ambos λ e ϕ. Foi calculado o valor $T = 2 \times 10^{25}$ o qual indica que os dados apresentam superdispersão e que são corroborados pela Figura 4.4.

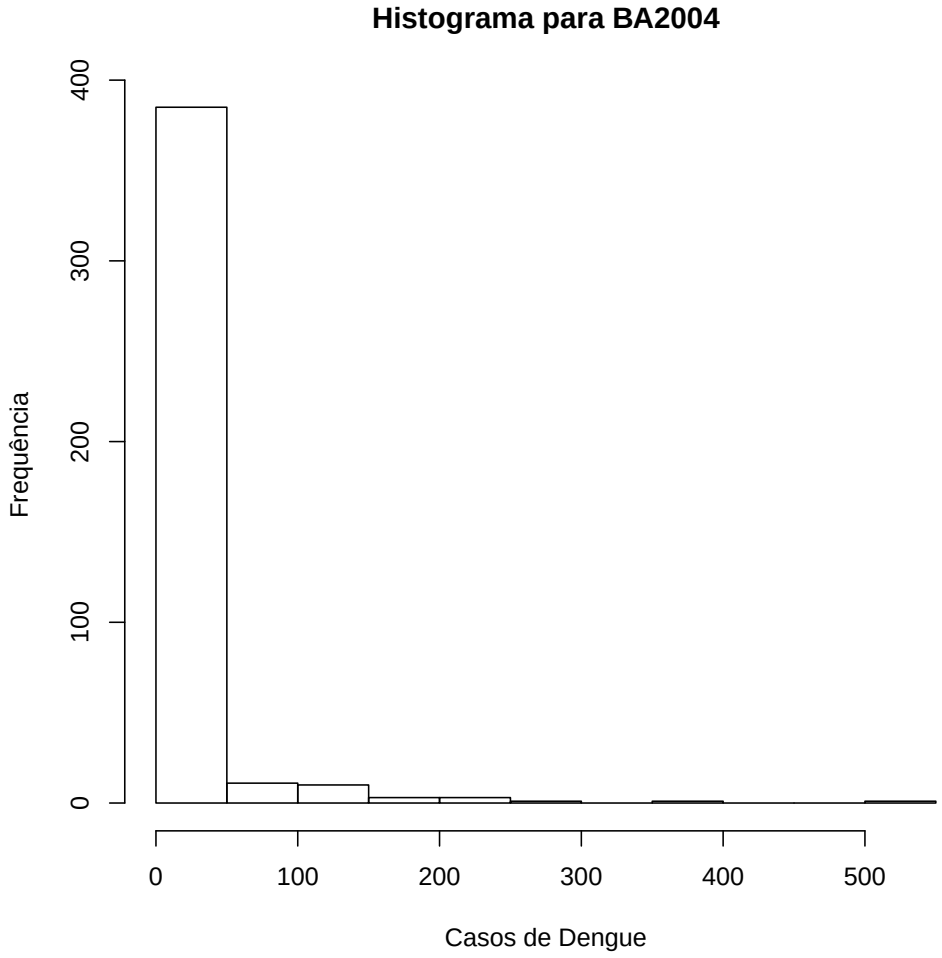

Figura 4.4: Histograma para os casos de dengue no estado da Bahia em 2004.

A Tabela 4.7 apresenta os valores dos MLEs, os seus desvios padrões (Sd), para os modelos IPNB e IPP temos $Sd \times 10^{-6}$, e a log-verossimilhança ($\ell(.)$) para os modelos ZIP, IPP, ZINB e IPNB. Podemos observar que os modelos IPNB e ZINB aparentam ter o melhor ajuste no conjunto BA2004.

A Tabela 4.8 apresenta os valores para a estatística de teste T_χ da equação (4.5.2) para

Tabela 4.7: Modelos Inflacionados ajustados no conjunto BA2004 (T9).

	β	β_0	β_1	β_2	β_3	β_4	β_5		
	γ	γ_0	γ_1	γ_2	γ_3	γ_4	γ_5	ϕ	$-\ell(.)$
IPP	β	-1.1234	0.0227	-0.0010	-0.0028	0.0011	1.7097		
	Sd	1.0488	0.0331	0.0547	0.0054	0.0005	1.6795		
	γ	-5.8666	0.0646	0.0812	-0.0135	-0.0011	13.2774	-	1363.36
	Sd	1.0960	0.0237	0.0494	0.0038	0.0005	1.7493		
ZIP	β	-3.9722	0.0791	0.0684	-0.0137	-0.0001	10.9741		
	Sd	0.0005	0.0022	0.0010	0.0003	0.00005	0.0001		
	γ	-0.9871	-0.4074	-0.4791	-0.0544	-0.0029	1.3660	-	6199.29
	Sd	0.0010	0.0008	0.0005	0.0007	0.0005	0.0001		
ZINB	β	1.0070	0.0733	0.1334	-0.0040	0.0001	2.1969		
	Sd	1.8234	0.0287	0.0726	0.0069	0.0003	2.9313		
	γ	-0.7096	0.0070	-0.0561	-0.4287	-0.9604	-1.7457	3.5427	1288.03
	Sd	109.19	5.9929	382.96	4.7560	4.5505	171.33	0.0316	
IPNB	β	1.1279	0.0109	0.0082	-0.0021	0.0009	1.4567		
	Sd	0.0382	0.05713	0.0482	0.0012	0.0370	0.0290		
	γ	-4.7449	0.1601	0.1741	-0.0298	-0.0008	6.6443	25.55727	1291.11
	Sd	0.0782	0.0170	0.0736	0.0323	0.0470	0.0582	0.0111	

cada um dos modelos já mencionados. Podemos concluir que o modelo IPNB é o modelo mais adequado para o conjunto de dados BA2004. Para calcular os valores da equação (4.5.2) foram criadas 18 cédulas para satisfazer a condição de haver pelo menos 5 elementos por cédula. A a Tabela 4.9 apresenta a estatística de teste de Voung, o que conclui que o melhor ajuste é atribuído ou ao modelo IPNB pois segundo o teste se ajusta melhor que o modelo ZIP e IPP tendo o modelo destacado melhor ajuste para o modelo IPP, e entre o modelo ZIP e ZINB o teste não distingue qual tem o melhor ajuste. Observe que a conclusão é baseada em múltiplas comparações pois quando é feito o teste para o modelo ZINB contra os demais não se é possível apontar qual apresenta melhor ajuste, mesmo o modelo ZINB tendo a maior logverossimilhança.

Tabela 4.8: Teste de bondade do ajuste Qui-Quadrado para os modelos IPP, ZIP, IPNB e ZINB nos dados BA2004.

Modelo	IPP	ZIP	IPNB	ZINB
T_χ	60937	2802289	9362	85132

Tabela 4.9: Valores de T_{fg} do teste Voung para os modelos IPP, ZIP, ZINB e IPNB para BA2004.

	Teste de modelo $f \times g$					
	IPP × ZIP	IPP × ZINB	IPP × IPNB	IPNB × ZIP	IPNB × ZINB	ZINB × ZIP
T_{fg}	4.8718	-1.2×10^{-11}	-4.7670	4.8889	6.1×10^{-10}	-3.8×10^{-13}

A Figura 4.5 apresenta as diferenças entre a proporção estimada de valores menos a proporção de valores empíricos na amostra. A Figura 4.5 apresenta as mesmas proporções no eixo vertical para não afetar as conclusões. Pela Figura 4.5 podemos observar que os modelos IPNB, ZIP e IPP se mostram mais adequados, com melhor medidas de previsão para o modelo IPNB.

Considerando os resultados das Tabelas 4.7, 4.8 e 4.9 e da Figura 4.5 concluímos que o melhor modelo de regressão inflacionado apresentado para o conjunto BA2004 é o modelo IPNB.

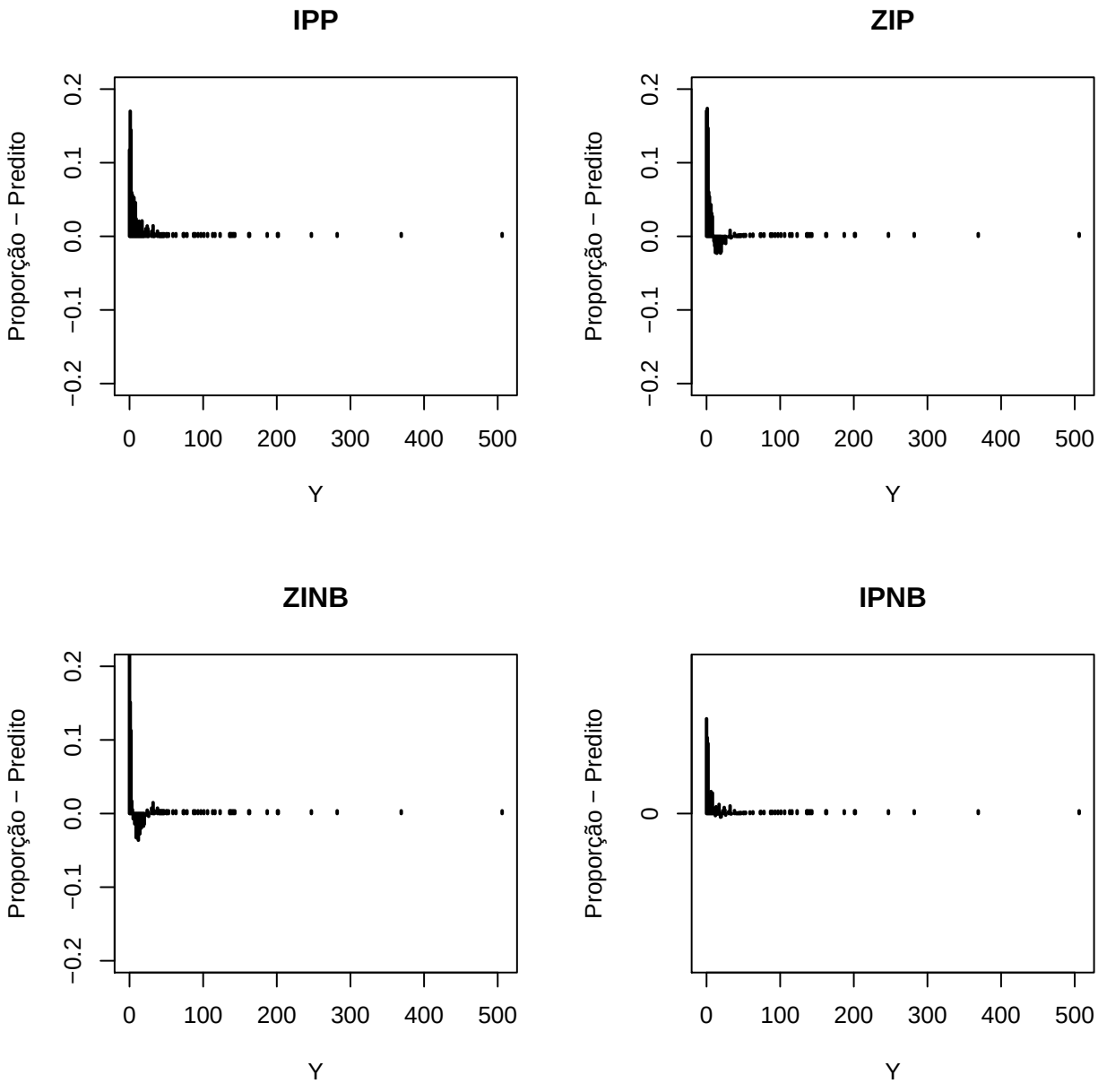

Figura 4.5: Diferenças entre as proporções estimadas menos a proporção empírica dos resultados para os dados BA2004 utilizando os modelos IPP, ZIP, ZINB e IPNB.

4.7 Conclusão

As distribuições IPP e IPNB são alternativas aos tradicionais modelos inflacionados ZIP e ZINB. A estimação dos modelos IPP e IPNB é um pouco complicada com necessidade de mais tempo computacional e requer alguns cuidados devido a presença da função hipergeométrica, porém os problemas de estimação que aparecem na rotina *optim* do *software* R Core Team (2012), podem ser resolvidos simplesmente pelo controle da função gradiente (comando *ndeps*). Também foram implementados no *software* Matlab (MATLAB, 2010) uma rotina BFGS em que é possível controlar outros parâmetros na estimação e assim obter os resultados mais rapidamente. A estimação é feita maximizando diretamente a função log-verossimilhança. Os testes bem como o desvio padrão dos estimadores são obtidos baseados na distribuição assintótica. Também foi avaliado testes de má especificação do modelo em que foi constatado que o parâmetro de inflação influencia corretamente na escolha entre os modelos com e sem inflação de zeros.

Esta seção também é utilizada para motivar a construção do próximo capítulo, visto que apresentaremos uma relação interessante que o parâmetro de inflação tem quando é discutido o processo de carcinogênese na formulação do modelo.

Os resultados apresentados neste capítulo estão em forma de relatório e foram enviados para uma revista e encontrando-se em processo de revisão até o momento e podem ser requisitados ao autor a qualquer momento.

Capítulo 5

Distribuição Poisson Correlacionada-Exponencial com riscos latentes competitivos (CoPE)

Como observado nos capítulos anteriores é comum o uso das distribuições discretas para se fazer mistura de distribuições, a fim de trabalhar com tempo de vidas latentes. Neste capítulo, vamos utilizar a distribuição dada em Kolev et al. (2000) e mencionar o ganho de interpretação do ponto de vista no estudo de carcinogênese.

Vimos novas classes de distribuições baseadas na distribuição Exponencial introduzidas a alguns anos nos Capítulos 2 e 3. A distribuição de Adamidis e Loukas (1998) que propuseram a distribuição EG que é a composição da distribuição Exponencial com a distribuição Geométrica, Kus (2007), fez uma modificação da distribuição Exponencial e sua composição com a distribuição Poisson Truncada (PE), Barreto-Souza e Cribari-Neto (2009) que generalizaram a distribuição proposta por Kus (2007) pela inclusão de um parâmetro de potência e Louzada et al. (2011) que propuseram a distribuição Exponencial Geométrica complementar. O processo de composição tem uma interpretação prática para o tempo de vida parecido com o processo de exponenciação e para uma discussão mais ampla a respeito, o leitor interessado pode buscar mais informações em Marshall e Olkin (1997). Cancho et al. (2011) propôs a distribuição Exponencial de potência complementar pela exponenciação da distribuição de potência Exponencial proposta por Smith e Bain (1975).

Neste capítulo, vamos propor uma nova distribuição de tempo de vida, obtida pela composição da distribuição de Poisson com parâmetro de inflação (IPP) encontrada em Kolev et al. (2000) com uma distribuição Exponencial padrão. Desta forma vamos nos referir a

esta nova distribuição como distribuição de Poisson-Exponencial Correlacionada (CoPE). A distribuição CoPE pode ser vista como um mecanismo de interpretação prática como uma rede de ligações nos riscos latentes e é uma distribuição composta que pode acomodar a correlação dentre as variáveis latentes aleatórias ao utilizar os conceitos encontrados em Minkova (2002).

A construção da distribuição CoPE está baseada no cenário de riscos latente competitivos (Louzada-Neto, 1999), no sentido de que a informação sobre o fator que foi responsável pela falha do objeto de estudo foi causada pela falha do tempo mínimo de todos os riscos em observação, porém os riscos não estavam correlacionados. Em muitas ocasiões a informação não está disponível ou é impossível de distinguir qual a verdadeira causa que implicou na falha do objeto mesmo por um *expert*. Ainda, a verdadeira causa pode se mascarada por algum motivo. Em confiabilidade, os componentes ainda podem ser totalmente destruídos no experimento e ainda em sistemas com manutenção urgente o sistema terá componentes substituídos sem ser identificado a verdadeira causa.

5.1 Distribuição CoPE

Antes de prosseguir com a construção da distribuição CoPE, vamos apresentar a função $H([a_1,\ldots,a_q],[b_1,\ldots,b_p],z)$ que é a função hipergeométrica generalizada, ${}_qF_p([a_1,\ldots,a_q],[b_1,\ldots,b_p],z)$, formalmente definida pela série

$$H([a_1,\ldots,a_q],[b_1,\ldots,b_p],z)=\sum_{k=0}^{\infty}\frac{\prod_{j=1}^{q}(a_j)_k}{\prod_{j=1}^{q}(b_j)_k}\frac{z^k}{k!},$$

em que $(a)_k=a(a+1)\cdots(a+k-1)$ é o símbolo de Pochhammer.

Seja a variável aleatória não negativa Y denotando o tempo de vida de um componente em uma determinada população. A variável aleatória Y terá distribuição CoPE com parâmetros $\lambda>0$, $\theta>0$ e $0<\rho<1$ se sua f.d.p. é dada por

$$f(y)=\lambda K\sum_{m=1}^{\infty}m\rho^m(e^{-\lambda x})^m H([m+1],[2],U), \tag{5.1.1}$$

em que $U=\theta(1-\rho)/\rho$ e $K=Ue^{-\theta/\rho}/(1-e^{-\theta})$.

O parâmetro λ é um parâmetro de escala, θ é o parâmetro da distribuição de Poisson, o que em nosso caso é o parâmetro dos riscos, e o parâmetro ρ é o mais atrativo pois nos permite fazer a associação entre os tempos de vida dos riscos (causas da falha) como veremos na seção

seguinte utilizando as propriedades encontradas por Minkova (2002). O parâmetro ρ pode ser visto como um parâmetro de correlação ou de probabilidade.

A distribuição CoPE tem sua moda em $y = 0$. Para todos os valores dos parâmetros no espaço paramétrico a f.d.p. da distribuição CoPE é estritamente decrescente em y e tende a zero quando $y \to \infty$, isto é, $f'(y) < 0$ para $y \in (0, \infty)$. Se o parâmetro $\rho = 0$ em na construção da distribuição IGPS dada em Kolev et al. (2000), ou seja quando $\rho \to 0$ em (5.1.1), obtemos a distribuição Exponencial-Poisson proposta por Kus (2007).

A Figura 5.1 apresenta a função f.d.p. da distribuição CoPE para λ=0.01, $\theta = 1$ e $\rho = 0.1, 0.5, 0.9$.

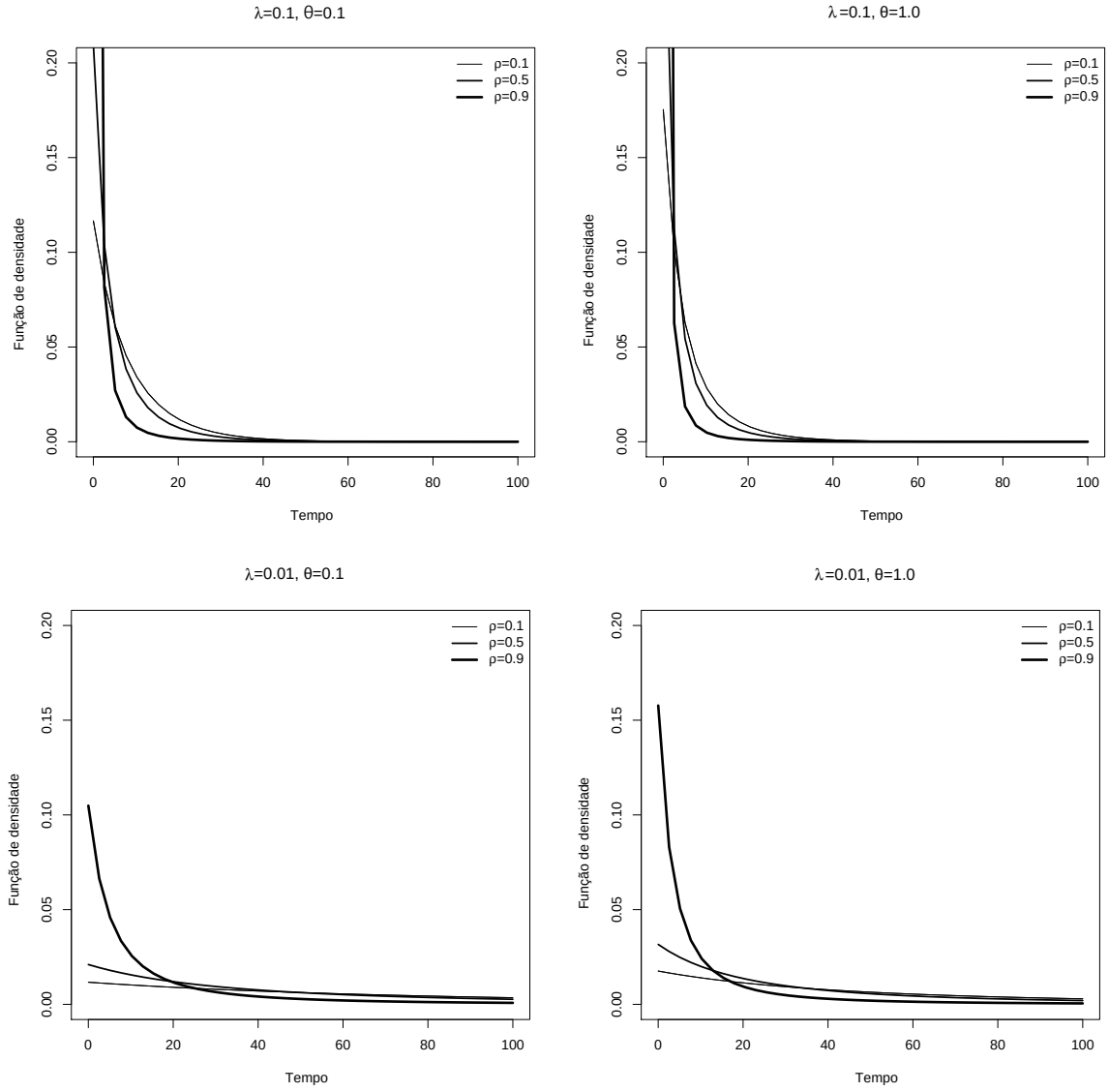

Figura 5.1: Função de densidade de probabilidade da distribuição CoPE para λ=0.1,0.01, $\theta = 1$,0.1. e $\rho = 0.1, 0.5, 0.9$.

A função de sobrevivência da distribuição CoPE é dada por,

$$S(y) = K \sum_{m=1}^{\infty} \rho^m [e^{-\lambda x}]^m H([m+1],[2],U), \tag{5.1.2}$$

em que U e K são dados em (5.1.1), para $\rho \in (0,1)$ e $\lambda, \theta > 0$.

A função quantil ou a inversa da função distribuição $F(x_p) = p$, não pode ser obtida explicitamente, porém o p-ésimo quantil pode ser obtido resolvendo a seguinte equação numericamente

$$\frac{1-p}{K} = \sum_{m=1}^{\infty} \rho^m (e^{-\lambda x_p})^m H([m+1],[2],U),$$

em que U e K são dados em (5.1.1).

5.1.1 Desenvolvimento da CoPE

Considere o cenário clássico de riscos competitivos associados a cenários de tempo de vida onde os riscos não são observados, a menos que observamos somente o tempo de vida mínimo entre todos os riscos. Em estudos de confiabilidade, podemos observar somente o componente com o tempo mínimo de um sistema com uma série de riscos, isto é, observamos somente o valor de cada componente como o tempo de vida mínimo sobre todos os motivos para falha e a causa da falha. Problemas de riscos competitivos é visto em áreas como a carcinogênese e existe uma ampla literatura de técnicas estatísticas desenvolvidas recentemente, como por exemplo, Borges et al. (2012) e Cobre et al. (2013). Lembramos novamente que para uma introdução, os interessados podem ler Lawless (2003), Crowder et al. (1991) e Cox e Oakes (1984). A dificuldade consiste dos riscos serem latentes no sentido que nenhuma informação sobre eles é obtida, ou seja, qual provocou a falha do componente, que tem o seu tempo de vida observado. Este cenário é chamado de riscos competitivo latente (Louzada-Neto, 1999). Em muitas ocasiões a informação não está disponível ou é impossível de que a verdadeira causa da falha seja especificada. Na área de confiabilidade, uma placa pode ser totalmente destruída no experimento e a verdadeira causa pode ser mascarada. Em estudos médicos, um paciente pode morrer e a verdadeira causa da morte pode ser atribuída a inúmeros fatores de risco. Como dito anteriormente, existe uma literatura bem difundida com os procedimentos estatísticos para a análise de tempos de vida na presença de riscos competitivos latentes e os interessados podem ler Adamidis e Loukas (1998), Cancho et al. (2011), Marchi et al. (2013), entre outros.

Uma característica das metodologias citadas é o fato de que as causas competitivas latentes são assumidas independentes, e pode ser um problema pois um sistema pode ser sobrecarregado pela falha de algum componente. Sendo assim o processo se torna analiticamente

mais atrativo, porém não incorpora a possibilidade de considerar a possível dependência entre as causas latente, o que pode não ser verdade sempre em situações práticas. Levando em consideração estes problemas, em confiabilidade, os componente podem ter uma *motherboard* conectada a outras, criando assim uma estrutura de dependência. Em estudos médicos, as múltiplas causas podem estar ligadas e dependentes entre si e assim ocasionar a morte.

Neste contexto, vamos considerar o fato da falha pela primeira causa e uma função de probabilidade com parâmetro de dependências entre as causas de falha. Portanto, o modelo CoPE é obtido como se segue.

Seja $M = 1, 2, \ldots$ uma variável aleatória representando o número de causas da falha com distribuição IPP. De acordo com Minkova (2002), temos

$$P(M = m|\theta, \rho) = \frac{e^{-\theta}}{1 - e^{-\theta}} \sum_{i=1}^{m} \begin{pmatrix} m-1 \\ i-1 \end{pmatrix} \frac{[\theta(1-\rho)]^i \rho^{m-i}}{i!}, \tag{5.1.3}$$

em que $0 < \rho < 1$, $\theta > 0$ e $m = 1, 2, ...$ Segundo Minkova (2002), a equação (5.1.3) é uma mistura de uma distribuição Binomial Negativa com parâmetros m e ρ com uma distribuição de Poisson com parâmetro θ, isto é, a Equação (5.1.3) é a probabilidade de que seja necessário m repetições até que o i-ésimo sucesso ocorra, em que a probabilidade de cada sucesso é $1 - \rho$ e m é um resultado da variável de Poisson truncada em zero. Portanto, o parâmetro ρ pode ser visto como uma medida de associação entre as causas de falha. Se $\rho = 0$, a equação (5.1.3) se ressume a distribuição de Poisson usual.

Considere também t_i realizações de uma variável aleatória de tempo de vida, isto é, o tempo até o evento causado pela i-ésima causa (risco) competitiva latente e T_i com distribuição Exponencial com parâmetro λ e sua f.d.p. dada por

$$f(t_i; \lambda) = \lambda e^{-\lambda t_i}, i = 1, 2, 3, \ldots \tag{5.1.4}$$

No cenário de riscos competitivos latentes, o número de causa M e o tempo de vida t_i associado a uma causa particular não é observada (latente). Assumimos que o tempo de vida mínimo Y sobre todas as causa é observado. Além disso, vamos considerar que essas variáveis são independentes de M. Portanto, observamos somente a variável aleatória dada por

$$Y = \min\{T_1, T_2 \ldots, T_M\}. \tag{5.1.5}$$

No campo de carcinogênese, de um ponto de vista microscópico, como discutido em Cobre et al. (2013), M pode ser visto como o número de células promovidas/ativadas com

número de células iniciadas I. Portanto as células promovidas/ativadas, com células iniciadas como latentes, a soma na equação (5.1.3) fornece a distribuição marginal de M pela relação,

$$\begin{aligned} P(M=m|\theta,\rho) &= \sum_{i=1}^{\infty} P(M=m|I=i,\theta,\rho)P(I=i|\theta) \\ &= \sum_{i=1}^{m} P(M=m,I=i|\theta,\rho). \end{aligned}$$

A equação (5.1.3) inclui a possibilidade das células iniciadas não serem promovidas/ativadas, o que é esperado pois células iniciadas podem ser mortas ou regeneradas pelo próprio corpo ou por um tratamento quimioterápido. Neste modelo assumimos que pelo menos uma célula iniciada é promovida/ativada. Ainda, o modelo inclui a possibilidade que mais de uma célula promovida/ativada e/ou iniciada ocorram.

A seguir mostraremos que a variável aleatória Y tem f.d.p. dada por (5.1.1).

Teorema 5.1.1 Se uma variável aleatória é definida por (5.1.5), então considerando (5.1.4) e (5.1.3), Y tem distribuição CoPE, com f.d.p. dada por (5.1.1).

Prova 5.1.1 A f.d.p. condicional de (5.1.5) dado $M=m$ é dada por

$$f(y|M=m,\lambda) = m\lambda e^{-\lambda y}(e^{-\lambda y})^{m-1}; t>0,\ m=1,\ldots$$

Portanto, a f.d.p. marginal de Y é dada por

$$\begin{aligned} f(y|\lambda,\theta,\rho) &= \sum_{m=1}^{\infty} \frac{m\lambda e^{-\lambda x}(e^{-\lambda x})^{m-1}e^{-\theta}}{1-e^{-\theta}} \sum_{i=1}^{m} \left(\begin{array}{c} m-1 \\ i-1 \end{array} \right) \frac{[\theta(1-\rho)]^i \rho^{m-i}}{i!} \\ &= \sum_{m=1}^{\infty} \frac{m(e^{-\lambda x})^m \lambda \rho^m e^{-\frac{\theta}{\rho}}\theta(1-\rho)}{\rho(1-e^{-\theta})} H\left([m+1],[2],\frac{\theta(1-\rho)}{\rho}\right) \\ &= \underbrace{\lambda\, e^{-\frac{\theta}{\rho}} \frac{\theta(1-\rho)}{\rho(1-e^{-\theta})}}_{K} \sum_{m=1}^{\infty} m\rho^m (e^{-\lambda x})^m H\left([m+1],[2],\frac{\overbrace{\theta(1-\rho)}^{U}}{\rho}\right) \\ &= \lambda K \sum_{m=1}^{\infty} m\rho^m (e^{-\lambda x})^m H([m+1],[2],U). \end{aligned}$$

O que completa a prova. ■

5.1.2 Função de risco e função de risco reversa

De (5.1.1) e (5.1.2), a função de risco, de acordo com a relação $h(y) = f(y)/S(y)$, é dada por

$$h(y) = \frac{\lambda \sum_{m=1}^{\infty} m\rho^m (e^{-\lambda x})^m H([m+1],[2],U)}{\sum_{m=1}^{\infty} \rho^m [e^{-\lambda x}]^m H([m+1],[2],U)}, \tag{5.1.6}$$

em que $U = \frac{\theta(1-\rho)}{\rho}$ e $K = U\frac{e^{-\frac{\theta}{\rho}}}{(1-e^{-\theta})}$. O valor inicial é sempre finito porém não é dado por uma expressão fechada. A função de risco é decrescente e no infinito tende ao valor $h(\infty) = \lambda$, pelo fato que a equação (5.1.6) é decrescente e $h(y)/\lambda > 1$.

A função de risco (5.1.6) é decrescente como mostra a Figura 5.2 para valores de $\rho = 0.1, 0.5, 0.9$, $\lambda = 0.1, 0.01$ e $\theta = 1$.

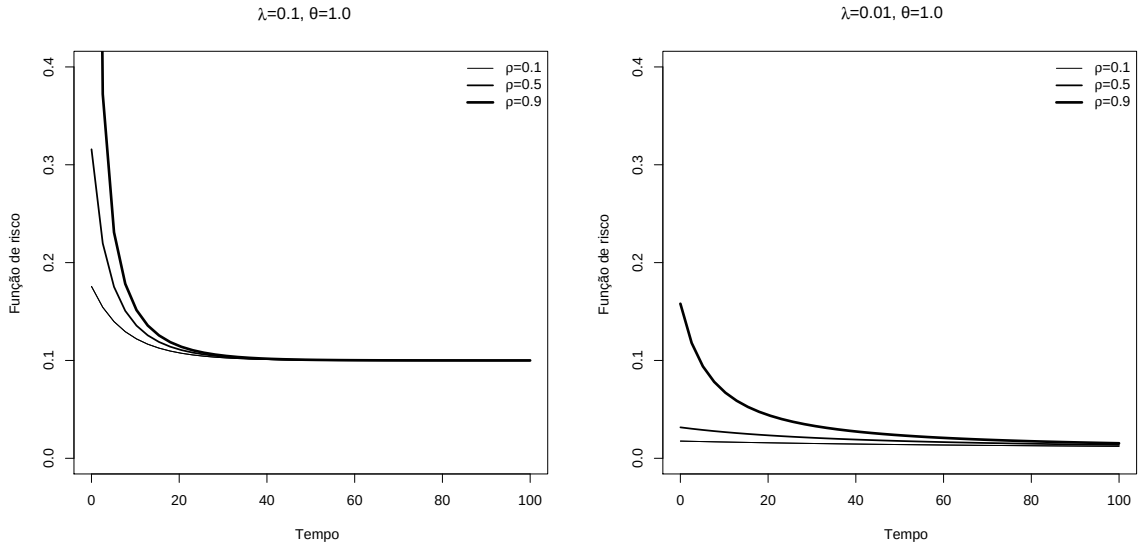

Figura 5.2: Função de risco da distribuição CoPE para λ=0.1,0.01, $\theta = 1$ e $\rho = 0.01, 0.5, 0.99$.

Proposição 5.1.1 Se uma variável aleatória Y tem distribuição CoPE, então a função de risco é estritamente decrescente.

Prova 5.1.2 Seja a função de risco $h(y)$ dada pela equação (5.1.6). Assim

$$h'(x) = \frac{f'(x)S(x) + f^2(x)}{S^2(x)}. \tag{5.1.7}$$

Agora

$$\begin{aligned} h'(x) \leq (\geq)0 \quad &\Rightarrow \quad f'(x)S(x) + f^2(x) \leq (\geq)0 \Rightarrow \\ 0 \ \geq (\leq) \ &-(\lambda K)^2 \Bigg(\sum_{m=1}^{\infty} m^2 (e^{-\lambda x})^m \rho^m H([m+1],[2],U) \\ &\times \sum_{m=1}^{\infty} (e^{-\lambda x})^m \rho^m H([m+1],[2],U) \\ &- \left[\sum_{m=1}^{\infty} m (e^{-\lambda x})^m \rho^m H([m+1],[2],U) \right]^2 \Bigg). \end{aligned} \tag{5.1.8}$$

Considerando $p = 2$, $q = 2$ $(1/q + 1/p = 1)$, $a_k = m\left((e^{-\lambda x})^m \rho^m H([m+1],[2],U)\right)^{1/2}$ e $b_k = \left((e^{-\lambda x})^m \rho^m H([m+1],[2],U)\right)^{1/2}$ na desigualdade de Holder, dada por

$$\sum_{k=1}^{\infty} a_k b_k \leq \left(\sum_{k=1}^{\infty} |a_k|^p\right)^{1/p} \left(\sum_{k=1}^{\infty} |b_k|^q\right)^{1/q}.$$

Temos que

$$\left[\sum_{m=1}^{\infty} m(e^{-\lambda x})^m \rho^m H([m+1],[2],U)\right]^2 \leq \sum_{m=1}^{\infty} m^2 (e^{-\lambda x})^m \rho^m H([m+1],[2],U) \times \sum_{m=1}^{\infty} (e^{-\lambda x})^m \rho^m H([m+1],[2],U).$$

Comparando com (5.1.8), concluímos que $h'(y) < 0$, isto é, a função de risco é decrescente. ■

De maneira análoga podemos chegar na mesma conclusão para a função de risco residual reversa. A função de risco reversa pode ser definida como a variável aleatória condicional $t - X|X \leq t$, que denota o tempo decorrido desde o tempo de falha do componente dado que seu tempo de vida é menor ou igual a t. Esta variável aleatória pode ser vista também como o tempo de inatividade (ou tempo desde a falha). Para mais detalhes, recomenda-se ler Nanda et al. (2003) e Kundu e Nanda (2010). Usando (5.1.1) e (5.1.2), a função de risco reversa é dada por

$$r(x) = \frac{f(x)}{F(x)} = \frac{\lambda \sum_{m=1}^{\infty} m\rho^m [e^{-\lambda x}]^m H([m+1],[2],U)}{\sum_{m=1}^{\infty} \rho^m (1 - [e^{-\lambda x}]^m) H([m+1],[2],U)}. \tag{5.1.9}$$

Proposição 5.1.2 Se a variável aleatória Y tem distribuição CoPE então a função de risco reversa é estritamente decrescente.

Prova 5.1.3 Seja $r(y)$, a função de risco reversa dada pela equação (5.1.9). Assim,

$$r'(x) = \frac{f'(x)F(x) - f^2(x)}{F^2(x)}.$$

É fácil ver que $r(x)$ é decrescente pois temos que $f'(x) < 0$. ■

5.1.3 Algumas propriedades

Como já dito, algumas das características e medidas de uma distribuição podem ser estudadas pelos seus momentos, como a média e a variância. Uma expressão bem conhecida para os r-ésimos momentos de uma variável aleatória não negativa X podem se calculados pela fórmula

$$E[X^r] = r \int_0^{\infty} x^{r-1} S(x) dx. \tag{5.1.10}$$

Proposição 5.1.3 Para uma variável aleatória Y com distribuição CoPE, o r-ésimo momento é dado por

$$\mu_r^{'} = \frac{r!K}{\lambda^r}\sum_{m=1}^{\infty}\frac{\rho^m}{m^r}H([m+1],[2],U).$$

Prova 5.1.4 De (5.1.2) and (5.1.10), temos que

$$\begin{aligned} E[X^r] &= r\int_0^{\infty} x^{r-1}S(x)dx \\ &= Kr\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\int_0^{\infty} x^{r-1}e^{-\lambda x m}dx \\ &= Kr\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\frac{\Gamma(r)}{(m\lambda)^r} \\ &= \frac{r\Gamma(r)K}{\lambda^r}\sum_{m=1}^{\infty}\frac{\rho^m}{m^r}H([m+1],[2],U), \end{aligned}$$

completando a prova. ■

Proposição 5.1.4 Para uma variável Y com distribuição CoPE, a função característica é dada por

$$\Phi_i(t) = \lambda K\sum_{m=1}^{\infty}\frac{m\rho^m}{\lambda m - it}F([m+1],[2],\frac{\theta(1-\rho)}{\rho}).$$

Prova 5.1.5 O resultado segue diretamente da integração pela definição da função característica. ■

Se não ocorrer a falha antes do tempo t, a distribuição de vida residual de uma variável aleatória X, tem função de sobrevivência dada por

$$S_t(x) = \Pr[X > x+t|X > t] = \frac{S(x+t)}{S(t)},\ \text{com } x \geq 0.$$

O r-ésimo momento do tempo de vida residual de uma variável não negativa contínua com função de sobrevivência $S(x)$ é dado por

$$\mu(t) = E((X-t)^r|X > t) = \frac{1}{S(t)}\int_t^{\infty} r(x-t)^{r-1}S(x)dx. \tag{5.1.11}$$

Proposição 5.1.5 Para uma variável aleatória Y com distribuição CoPE, o r-ésimo momento do tempo de vida residual é dado por

$$\mu_r(t) = \frac{K}{S(t)}\frac{r!}{\lambda^r}\sum_{m=1}^{\infty}\frac{e^{-\lambda m t}}{m^r}\rho^m H([m+1],[2],U).$$

Prova 5.1.6 Da equação (5.1.11) e usando $S(y)$ dada por (5.1.2), temos que

$$\begin{aligned} E((X-t)^r|X>t) &= \frac{1}{S(t)}\int_t^\infty r(x-t)^{r-1}S(u)du \\ &= \frac{Kr}{S(t)}\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\int_t^\infty (x-t)^{r-1}e^{-\lambda x m}dx \\ &= \frac{Kr}{S(t)}\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\frac{\Gamma(r)e^{-\lambda t m}}{(\lambda m)^r} \\ &= \frac{Kr\Gamma(r)}{\lambda_e^r S(t)}\sum_{m=1}^{\infty}\frac{\rho^m e^{-\lambda t m}}{m^r}H([m+1],[2],U). \end{aligned}$$

■

O r-ésimo momento do tempo de vida residual reversa da variável não negativa X pode ser obtida pela fórmula dada por

$$m_r(t) = E((t-X)^r|X\leq t) = \frac{1}{F(t)}\int_0^t r(t-x)^{r-1}F(x)dx. \tag{5.1.12}$$

Proposição 5.1.6 Para uma variável aleatória Y com distribuição CoPE, o r-ésimo momento de vida residual reversa é dado por

$$m_r(t) = \frac{Kt^r}{F(t)}\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\left(1-e^{-m\lambda t}H([r],[1+r],m\lambda t)\right).$$

Prova 5.1.7 Da equação (5.1.12) e utilizando a expressão de $F(y)$ temos que

$$\begin{aligned} E((t-X)^r|X\leq t) &= \frac{1}{F(t)}\int_t^\infty r(t-x)^{r-1}F(x)dx \\ &= \frac{Kr}{F(t)}\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\int_t^\infty (t-x)^{r-1}(1-e^{\lambda x m})dx \\ &= \frac{Kr}{F(t)}\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\left(\frac{t^r}{r}-\frac{t^r e^{-\lambda m t}H([r],[1+r],\lambda m t)}{r}\right) \\ &= \frac{Kr}{F(t)}\sum_{m=1}^{\infty}\rho^m H([m+1],[2],U)\frac{t^r}{r}\left(1-e^{-m\lambda t}H([r],[1+r],m\lambda t)\right). \end{aligned}$$

■

5.1.4 Curvas de Bonferroni e Lorenz

As curvas de Bonferroni e Lorenz, bem como os indicadores de Bonferroni e Gini, tem aplicação não somente na economia para o estudo de renda e pobreza, mas também em campos como a confiabilidade, demografia, seguros e medicina como medidas de concentração.

As curvas de Bonferroni também são capazes de indicar os parâmetros de escala. As curvas de Bonferroni e Lorenz são curvas definidas, respectivamente por

$$B(p) = \frac{1}{p\mu}\int_0^p F^{-1}(x)dx, \text{ e} \quad L(p) = \frac{1}{\mu}\int_0^p F^{-1}(x)dx,$$

em que $\mu = E(X)$ e $F(q) = p$.

Os indicadores de Bonferroni e Gini são definidos por

$$B = 1 - \int_0^1 B(p)dp, \text{ e} \quad G = 1 - 2\int_0^1 L(p)dp, \text{ respectivamente.}$$

Proposição 5.1.7 Para uma variável aleatória Y com distribuição CoPE, as curvas de Bonferroni e Lorenz são dadas por

$$B(p) = \frac{K}{\lambda\mu p}\sum_{m=1}^{\infty}\frac{\rho^m}{m}H([m+1],[2],U)\left(1 - e^{-\lambda m F^{-1}(p)}[1+\lambda m F^{-1}(p)]\right) \quad (5.1.13)$$

e

$$L(p) = \frac{K}{\lambda\mu}\sum_{m=1}^{\infty}\frac{\rho^m}{m}H([m+1],[2],U)\left(1 - e^{-\lambda m F^{-1}(p)}[1+\lambda m F^{-1}(p)]\right), \quad (5.1.14)$$

respectivamente.

Prova 5.1.8 O resultado é obtido por uma simples manipulação algébrica usando o fato que $\int_a^b F^{-1}(x)dx = \int_{F^{-1}(a)}^{F^{-1}(b)} xf(x)dx$. ■

Proposição 5.1.8 Para uma variável aleatória Y com distribuição CoPE, o indicador de Gini é dado por

$$\begin{aligned} G &= 1 - 2\int_0^1 L(p)dp \\ &= 1 - 2\frac{K}{\lambda\mu}\sum_{m=1}^{\infty}\left[\frac{\rho^m}{m}H([m+1],[2],U)\right. \\ &\times \left.\left(1 - K\sum_{j=1}^{\infty}\frac{\rho^j}{(j+m)}H([j+1],[2],U) - mK\sum_{j=1}^{\infty}\frac{\rho^j}{(j+m)^2}H([j+1],[2],U)\right)\right], \end{aligned}$$

em que $\mu = E(Y)$.

Prova 5.1.9 Após a mudança de variável $p = F(x)$ em $\int L(p)dp$, com um pouco de manipulação algébrica a prova é finalizada. ■

Para o indicador de Bonferroni não encontramos uma forma fechada, porém pode ser estimado por uma integração de Monte Carlo. Para R pontos simulados , $u_1, u_2, \ldots, u_R$, de uma distribuição Uniforme, temos que $B \approx 1 - \left(\sum_{i=1}^{R} B(u_i)/R\right)$.

A Figura 5.3 apresenta as curvas de Bonferroni e Lorenz para $\lambda = 1$, $\theta = 0.4, 2, 10$ e $\rho = 0.3, 0.5, 0.8$. As curvas de Bonferroni e Lorenz são afetadas pelo parâmetro θ e ρ, o que é indicativo que λ é um parâmetro de escala pelas propriedades destas curvas. Aqui salientamos que não apresentamos os gráficos para $\lambda \neq 1$ pois são idênticos aos demais e por isso chega-se a conclusão que λ é parâmetro de escala.

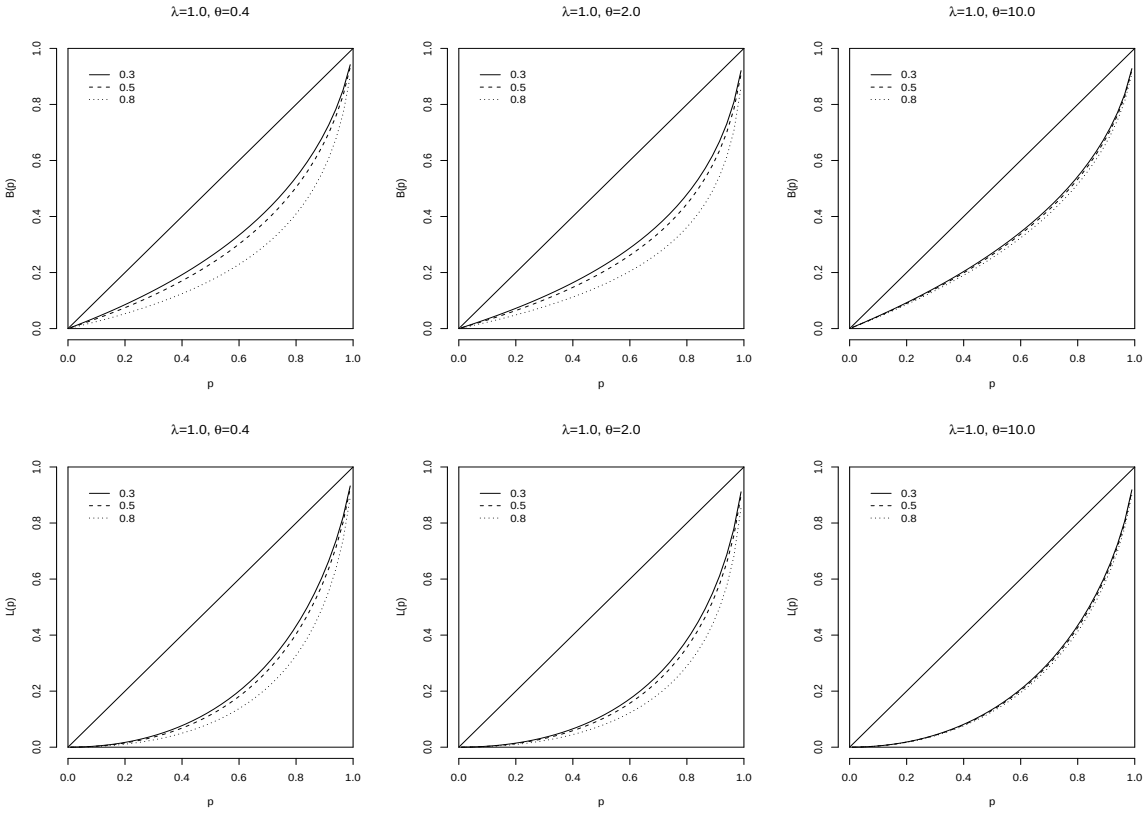

Figura 5.3: As curvas de Bonferroni $B(p)$ e Lorenz $L(p)$ para a distribuição CoPE.

O tempo total ponderado (*scaled total time*) de uma função distribuição F, de acordo com Pundir e Arora (2005) é definido por

$$S_F[F(t)] = \frac{1}{\mu}\int_0^t S(u)du.$$

Proposição 5.1.9 Para uma variável aleatória Y com distribuição CoPE, o tempo total ponderado é dado por

$$S_F[F(t)] = \frac{K}{\lambda\mu}\sum_{m=1}^{\infty}\frac{\rho^m}{m}(1-e^{-\lambda tm})H([m+1],[2],U). \tag{5.1.15}$$

Prova 5.1.10 O resultado é obtido diretamente pela integral de $\int S(u)du$. ■

O tempo total em teste acumulado modificado de uma função distribuição F, C_F, segundo Pundir e Arora (2005), é obtido por meio da relação $C_F = 1 - G$.

5.2 Inferência

A seguir vamos considerar a estimação pelo método de máxima verossimilhança e apresentar as expressões associadas a matriz de informação de Fisher observada.

Assumindo que os tempos de vida são independentes e identicamente distribuídos e independentes do mecanismo de censura, os estimadores MLEs dos parâmetros são obtidos pela maximização da função de log-verossimilhança dada por

$$\begin{aligned}\ell(\lambda,\theta,\rho) &= \ln(\lambda)\sum_{i=1}^{n}(\delta_i) + n\left(\ln(\theta) - \frac{\theta}{\rho} + \ln(1-\rho) - \ln(\rho) - \ln(1-e^{-\theta})\right) \\ &+ \sum_{i=1}^{n}\delta_i \ln\left[\frac{f(y_i)}{\lambda K}\right] + \sum_{i=1}^{n}(1-\delta_i)\ln\left[\frac{S(y_i)}{K}\right], \qquad (5.2.1)\end{aligned}$$

em que δ_i é o indicador de censura, sendo igual a 0 ou 1 se o dado é censurado ou observado, respectivamente. Pelas Equações (5.1.1) e (5.1.2) temos $f(y)$ e $S(y)$, respectivamente. Ainda, $U = \frac{\theta(1-\rho)}{\rho}$ e $K = U\frac{e^{-\frac{\theta}{\rho}}}{(1-e^{-\theta})}$.

Segue pelo método de máxima verossimilhança que os estimadores MLEs, digamos $\hat{\lambda}$, $\hat{\theta}$ e $\hat{\rho}$, são as soluções simultâneas das seguintes equações

$$\begin{aligned}\frac{\sum_{i=1}^{n}\delta_i}{\lambda} &= \sum_{i=1}^{n}\left[\delta_i\frac{\sum_{m=1}^{\infty}m^2\rho^m x_i e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])}{\sum_{m=1}^{\infty}m\rho^m e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])}\right. \\ &+ \left.(1-\delta_i)\frac{\sum_{m=1}^{\infty}m\rho^m x_i e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])}{\sum_{m=1}^{\infty}\rho^m e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])}\right],\end{aligned}$$

$$\begin{aligned}n\left(-\frac{1}{\rho}+\frac{1}{\theta}-\frac{e^{-\theta}}{1-e^{-\theta}}\right) &= -\frac{(1-\rho)}{2\rho}\left\{\sum_{i=1}^{n}\left[\delta_i\frac{\sum_{m=1}^{\infty}(m+1)m\rho^m e^{-\lambda x_i m}H([m+2],[3],[\theta(1-\rho)/\rho])}{\sum_{m=1}^{\infty}m\rho^m e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])}\right.\right. \\ &+ \left.\left.(1-\delta_i)\frac{\sum_{m=1}^{\infty}(m+1)\rho^m e^{-\lambda x_i m}H([m+2],[3],[\theta(1-\rho)/\rho])}{\sum_{m=1}^{\infty}\rho^m e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])}\right]\right\}\end{aligned}$$

e

$$\begin{aligned}n\left(\frac{\theta}{\rho^2}+\frac{1}{\rho(1-\rho)}\right) &= \frac{e^{\frac{\theta}{\rho}}\theta(1-\rho)}{\rho(1-e^{\theta})}\sum_{i=1}^{n}\left\{\delta_i\Big[\frac{\lambda\theta}{2\rho^2}\sum_{m=1}^{\infty}(m+1)m\rho^m e^{-\lambda x_i m}H([m+2],[3],[\theta(1-\rho)/\rho])\right. \\ &+ \left.\sum_{m=1}^{\infty}m^2\rho^{m-1}e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])\Big]/f(x_i)\right. \\ &+ (1-\delta_i)\Big[\frac{\theta}{2\rho^2}\sum_{m=1}^{\infty}(m+1)\rho^m e^{-\lambda x_i m}H([m+2],[3],[\theta(1-\rho)/\rho]) \\ &+ \left.\sum_{m=1}^{\infty}m\rho^{m-1}e^{-\lambda x_i m}H([m+1],[2],[\theta(1-\rho)/\rho])\Big]/S(x_i)\right\}.\end{aligned}$$

Como pode ser observado, os estimadores são analiticamente intratáveis, porém o método de máxima verossimilhança permite utilizar procedimentos que usam pacotes computacionais

com algorítimos que encontram os valores que maximizam a função de logverossimilhança diretamente. Para obter as estimativas de máxima verossimilhança, foi desenvolvida uma rotina *BFGS* para o controle de variáveis do algorítimo que não podem ser alterados devido a pertencerem pacotes computacionais já implementados. O detalhamento da rotina se encontra no Apêndice E. Para amostras consideradas grandes, a inferência para os parâmetros MLEs e seus erros padrões são baseados nos valores da matriz de informação de Fisher inversa. Visto que uma forma fechada para a matriz de informação não é viável, vamos considerar a matriz de informação de Fisher observada para determinar os intervalos de estimação de (λ, θ, ρ), dados por

$$I_F(\lambda,\theta,\rho) = - \left(\begin{array}{ccc} I_{\lambda\lambda} & I_{\lambda\theta} & I_{\lambda\rho} \\ I_{\theta\lambda} & I_{\theta\theta} & I_{\theta\rho} \\ I_{\rho\lambda} & I_{\rho\theta} & I_{\rho\rho} \end{array} \right) \Bigg|_{(\lambda,\theta,\rho)=(\hat{\lambda},\hat{\theta},\hat{\rho})}, \tag{5.2.2}$$

em que os elementos da matriz $I_F(\lambda, \theta, \rho)$ estão apresentados no Apêndice D.

Sob condições de regularidade para os parâmetros λ, θ e ρ no espaço paramétrico e pelo Teorema Central do Limite a distribuição assintótica de $(\hat{\lambda} - \lambda, \hat{\theta} - \theta, \hat{\rho} - \rho)$, quando $n \to \infty$, é uma Normal trivariada com vetor de média nulo e matriz de variância e covariância $I_F^{-1}(\lambda, \theta, \rho)$.

Como já mencionado, diferentes distribuições podem ser comparadas utilizando os valores de $-\max \ell(.)$, em que $\ell(.)$ é a log-verossimilhança calculada nos valores obtidos MLEs de sua respectiva logverossimilhança, e também pelo critério de Informação Akaike (AIC) que penaliza o super-ajuste e tenta minimizar a diferença de Kullbak-Labler dado entre a verdadeira distribuição e a distribuição calculada como candidata. O critério de AIC é dado por $-2\ell(.)+2q$, em que $\ell(.)$ é definida como anteriormente, q é o número de parâmetros estimados e n é o tamanho da amosta. A melhor distribuição é escolhida pelo menor valor de $-\max \ell(.)$ e AIC. Além destes valores, também podemos utilizar o teste de Kolmogorov-Smirnov (KS *test*) que compara a distribuição estimada com a distribuição amostral ou com uma distribuição de referência.

5.3 Estudo de simulação

O estudo a seguir é baseado em 1000 gerações de conjunto de dados da distribuição CoPE com diferentes conjuntos de parâmetros para $n = 15$, 30 e 50. Os tamanhos das amostra foram escolhidos desta forma pelos cálculos serem comprometidos pelo termo da soma das

funções hipergeométricas o qual se gera um tempo computacional superior a meses para a conclusão da rotina de simulação. As restrições dos parâmetros α e λ foram obtidos pela transformação exponencial e para o parâmetro ρ por meio da transformação $e^{\rho^*}/(1+e^{\rho^*})$, em que $\rho^* \in \mathbb{R}$. O ponto $(-3.5, -1.5, 0)$ foi determinado como valor inicial de todas as simulações devido ao fato de sempre iniciar o processo de estimação ao apresentar valor finito para a logverossimilhança, entretanto em alguns testes, os quais não apresentamos aqui, notou-se que o valor da convergência não era influenciado por esta escolha.

A Tabela 5.1 apresenta as médias dos estimadores MLEs (Av), seus desvios padrões (Sd) e o erro quadrático médio (MSE), juntamente com a probabilidade de cobertura para níveis especificados de 95% de confiança dos parâmetros. Os valores de MSE bem como os valores Sd decrescem quando o tamanho da amostra aumenta. Além disso, a probabilidade de cobertura se aproximam dos valores determinados quando o tamanho da amostra aumenta.

Tabela 5.1: Média dos estimadores MLEs, seus desvios padrões, MSE e coberturas da distribuição CoPE para dados simulados.

Parâmetros	θ	λ	ρ	θ	λ	ρ	θ	λ	ρ
Valores	0.001	0.5	0.6	0.1	0.005	0.1	0.001	0.3	0.4
				$n = 15$					
Av	0.0008	0.1149	0.3631	0.0894	0.0161	0.2614	0.0012	0.2874	0.3994
Sd	0.1023	0.2248	0.3596	0.0335	0.0338	0.1991	0.0899	0.2805	0.2501
MSE	0.0012	0.2253	0.3798	0.0016	0.0022	0.2021	0.0004	0.3069	0.2753
CP (95%)	0.9200	0.7100	0.9800	0.9200	0.8800	0.8400	0.8800	0.9400	0.7200
				$n = 30$					
Av	0.0008	0.2156	0.3015	0.0913	0.0059	0.2204	0.0012	0.2573	0.3968
Sd	0.1023	0.3010	0.3270	0.0169	0.0246	0.1558	0.0896	0.1996	0.1971
MSE	0.0009	0.2118	0.2468	0.0012	0.0023	0.1516	0.0003	0.2258	0.2237
CP (95%)	0.9800	0.7400	0.9800	0.9200	0.9600	0.8800	0.9000	0.9400	0.7600
				$n = 50$					
Av	0.0010	0.6273	0.6481	0.0925	0.0020	0.1958	0.0012	0.2665	0.4119
Sd	0.0003	0.2594	0.1173	0.0172	0.0101	0.1364	0.0896	0.1829	0.1994
MSE	0.0006	0.1742	0.2062	0.0013	0.0015	0.1356	0.0002	0.2036	0.2119
CP (95%)	0.9900	0.8900	0.9400	0.9300	0.9200	0.9400	0.9200	0.9400	0.9100

5.4 Aplicações em dados simulados e reais

A seguir, vamos ilustrar o ajuste da distribuição CoPE a seis conjuntos de dados, dos quais três são conjuntos de dados gerados para propósito ilustrativo e três foram extraídos da literatura.

5.4.1 Conjunto de dados ilustrativos

Suponha que temos três conjuntos de dados, digamos $I1$, $I2$ e $I3$, gerados da distribuição CoPE com uma correlação de $\rho = 0.01$ ($\lambda = 0.2, \theta = 0.06$) para $I1$, $\rho = 0.7$ ($\lambda = 0.04, \theta = 0.8$) para $I2$ e $\rho = 0.8$ ($\lambda = 0.05, \theta = 0.06$) para $I3$. Desta forma, esperamos um melhor ajuste da distribuição PE para o conjunto $I1$ e um melhor ajuste para a distribuição CoPE para os dados $I2$ e $I3$, pois quando $\rho \to 0$ a distribuição CoPE é a distribuição PE.

A Tabela 5.2 apresenta os resultados para os ajustes obtidos para as distribuições PE e CoPE. A presença da correlação é adequadamente capturada pela distribuição CoPE. A Figura 5.4.1 apresenta as funções de sobrevivência para os conjuntos de dados $I1$, $I2$ e $I3$ para as distribuições PE e CoPE.

Tabela 5.2: Parâmetros estimados para as distribuições PE e CoPE para os dados ilustrativos I1, I2 e I3.

		λ	θ	ρ	$-\max \ell(.)$	AIC
Valores verdadeiros		0.20	0.06	0.01		
I1	CoPE	0.3223	0.0419	0.0209	63.2672	132.5344
	PE	0.3310	0.0017	–	63.1676	130.3352
Valores verdadeiros		0.04	0.80	0.70		
I2	CoPE	0.0160	0.8940	0.7042	121.7000	249.4000
	PE	0.0004	0.0425	–	124.7484	253.4968
Valores verdadeiros		0.05	0.06	0.80		
I3	CoPE	0.0603	0.8698	0.7566	78.3140	162.6280
	PE	0.0002	0.1575	–	85.4351	174.8702

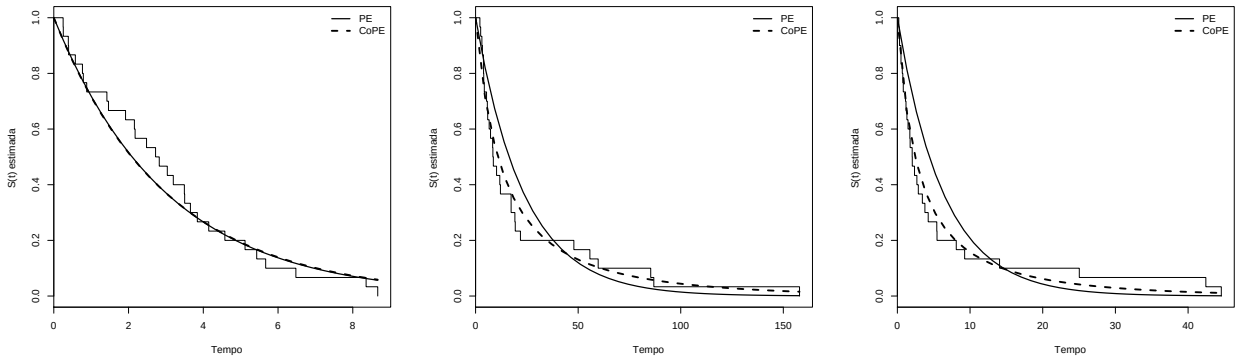

Figura 5.4: A curva KM e as curvas de sobrevivência ajustadas de PE e CoPE. Esquerda: para os dados $I1$. Meio: para $I2$. Direita: para $I3$.

5.4.2 Aplicação em dados reais

O conjunto de dados, denominado $T10$, é composto por tempos de sobrevivência de 20 pacientes de câncer no pulmão com terapia já realizada e o conjunto de dados, denominado $T11$, por tempos de sobrevivência de 44 pacientes de câncer no pulmão sem terapia já realizada. Ambos os conjuntos $T10$ e $T11$ foram extraídos de Prentice (1973). O terceiro conjunto de dados, denominado $T12$, é composto por tempos de remissão, em semanas, de um grupo de 30 pacientes com leucemia que receberam tratamentos similares e foi extraído de Lawless (2003).

Os fatores competindo para causar a falha não são observados, no entanto, pode-se especular alguma quantidade para os possíveis fatores em risco dos conjuntos $T10$, $T11$ e $T12$. Por exemplo, para $T10$ e $T11$ podemos considerar o tempo de exposição ao tabaco, fatores genéticos, amianto e poluição do ar, tempo que é fumante passivo, e exposição a inúmeras substâncias. Para $T12$ podemos considerar a pré-disposição genética, íons radioativos artificiais, vírus T-linfotrópico humano, vírus da imunodeficiência humana, benzeno e alguns petroquímicos, agentes quimioterápicos alquileno usados em tratamentos anteriores, transmissão fetal, sindrome de down, entre outros. É claro também que as causas acima podem ser correlacionadas.

A Tabela 5.3 apresenta os valores $-\max \ell(.)$, AIC e o p valor do teste KS para as distribuições PE e CoPE ajustadas. A distribuição CoPE apresenta ajuste melhor ou equivalente aos ajustes de outras distribuições comuns no estudo de tempo de vida, como pode ser visto comparando os resultados da Tabela 5.3 com os da Tabela 5.4.

Observamos que para $T10$ e $T12$ os valores para o parâmetro de correlação ρ são iguais a 0.4749 e 0.5136, respectivamente, indicando correlação moderada dentre as causas

competitivas latente. Para $T11$ o parâmetro de correlação é dado por 0.0132, considerado de baixa correlação dentre as causas competitivas latentes. Assim, é apropriado dizer que se é observado que a distribuição CoPE aparenta apresentar ajuste apropriado em dados com presença de correlação nas causas competitivas. Esta afirmação pode ser suportada pelo valor de p do teste KS como todos os valores de AIC apresentados, em particularmente pela Tabela 5.3.

A Figura 5.5 apresenta as distribuições PE e CoPE ajustadas nos conjuntos de dados $T10$, $T11$ e $T12$.

Tabela 5.3: Parâmetros estimados das distribuições PE e CoPE para os conjuntos T10, T11 e T12.

		λ	θ	ρ	$-\max \ell(.)$	AIC	p valor KS
T10	CoPE	0.0052	0.8669	0.4749	111.4978	228.9956	0.7640
	PE	0.0094	0.1981	–	113.1932	230.3864	0.4710
T11	CoPE	0.0081	0.0153	0.0132	255.3605	516.7210	0.9252
	PE	0.0082	0.0011	–	255.3605	514.7211	0.9216
T12	CoPE	0.0373	0.7358	0.5136	99.1935	204.3870	0.9312
	PE	0.0328	0.5388	–	110.3611	224.7221	0.8714

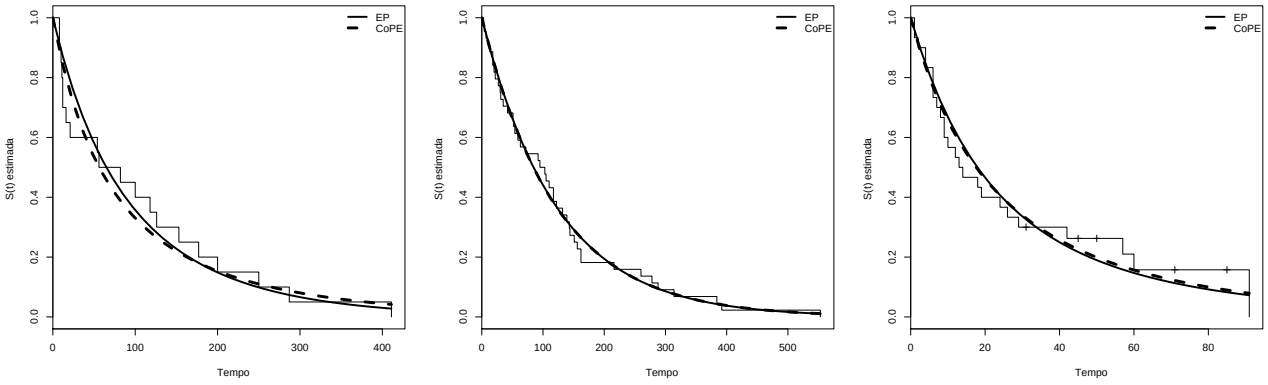

Figura 5.5: Curva KM e curvas das distribuições ajustadas PE e CoPE. Esquerda: para o conjunto T10. Meio: para T11. Direira: para T12.

Além dos valores apresentados se pensarmos no campo da carcinogênese somente, pela equação (5.1.3), $1-\rho$ é a probabilidade de cada célula iniciada se tornar uma célula ativada, e portanto, para $T10$ e $T12$ temos as probabilidades 0.5251 e 0.4864, respectivamente, enquanto para $T11$ é igual a 0.9868.

Para ilustração ajustamos algumas distribuições de tempo de vida comuns na literatura e comparamos com o ajuste da distribuição CoPE. Para os critérios de comparação consideramos o valor de AIC e o teste KS. As distribuições utilizadas para comparação são a distribuição Weibull, a distribuição Gama, a distribuição EL (Bakouch et al., 2011) com f.d.p. $f(x) = [\lambda\ (1+\lambda+\lambda x)^{\alpha-1}/(1+\lambda)^{\alpha}]\ (\beta\,(1+\lambda+\lambda x)(\lambda x)^{\beta-1} - \alpha)^{-(\lambda x)^{\beta}}$ e a distribuição Exponencial-Poisson generalizada (GEP) de Barreto-Souza e Cribari-Neto (2009) com f.d.p. $f(x) = [\alpha\beta\lambda/(1-e^{-\lambda})^{\alpha}]\ (1 - exp(-\lambda + \lambda \exp(-\beta x)))^{\alpha-1}\ exp(-\lambda - \beta x + \lambda\ \exp(-\beta x))$. Os resultados para os ajustes são apresentados na Tabela 5.4, os quais foram comparados com os valores da Tabela 5.3, e estes suportam a evidência que a distribuição CoPE é uma boa alternativa de ajuste para os dados $T10$ e $T12$ como visto tanto pelo valor p da estatística KS quanto pelo valor AIC, sendo que para $T12$, o valor de AIC é muito menor do que os demais.

Tabela 5.4: Valores de AIC e p valor do teste KS para as distribuições EL, Weibull, Gama e GEP ajustados para T10, T11 e T12.

		EL	Weibull	Gama	GEP	PE	CoPE
T10	p valor KS	0.5700	0.5710	0.5452	0.5604	0.4710	0.7640
	AIC	231.8538	229.8548	229.9062	231.8282	230.3864	228.9956
T11	p valor KS	0.9184	0.9181	0.9208	0.9394	0.9216	0.9252
	AIC	516.7121	515.7390	515.7410	516.6836	514.7211	516.7210
T12	p valor KS	0.8709	0.8252	0.6736	0.8702	0.8714	0.9312
	AIC	224.8123	223.3568	223.9217	224.5522	224.7221	204.3870

5.5 Conclusão

Neste capítulo foi proposta uma nova distribuição de tempo de vida inspirada no processo de carcinogênese, a distribuição CoPE, que tem um parâmetro que captura a dependência entre os riscos (causas) competitivas latente. Foi apresentada a sua construção baseada no cenário de riscos competitivos correlacionados em que se é observado somente o tempo associado com o tempo mínimo de falha dos riscos competitivos e portanto somente o tempo de falha do objeto em estudo é observado. Além disso, no estudo de carcinogênese com riscos latentes, a distribuição tem interpretações práticas do parâmetro que envolve a probabilidade de células iniciadas de se tornarem células promovidas.

As propriedades da distribuição proposta foram apresentadas, incluindo a prova formal

de sua f.d.p., e expressões para o cálculo da função de sobrevivência e função de risco, dos momentos, da média residual, as curvas de Bonferroni e Lorenz, os indicadores de Bonferroni e Gini, o tempo total ponderado e o tempo total em teste acumulado transformado.

Os estimadores de máxima verossimilhança foram obtidos diretamente da função de logverossimilhança por meio de rotina BFGS implementada no *software* Matlab MATLAB (2010). A importância prática desta nova distribuição em conjuntos de dados fictícios e reais, em que a distribuição CoPE se mostrou adequada e em alguns casos com melhor ajuste. Particularmente, a distribuição CoPE apresenta um tempo computacional e cuidados bem maiores, cerca de 5 minutos para o ajuste em um conjunto com 100 amostras, comparado aos demais modelos em um computador com processador Intel Core i5-4200Y, com 4GB de memória Ram e sistema operacional Windows de 64 bits, no entanto ela apresentou ajustes considerados melhores aos dados em que a presença da correlação entre os fatores de risco são coerentes, bem como no conjunto de dados em que os pacientes apresentavam menor probabilidade de desenvolver o câncer por já terem participado de seções para tratamento. Este melhor ajuste já se era esperado pois o parâmetro ρ tem como interpretação a probabilidade de células iniciadas se tornarem células cancerígenas frente aos demais modelos apresentados.

Capítulo 6

Considerações finais e propostas futuras

Nos capítulos 2, 3 e 5 foram apresentados vários modelos para estudo do tempo de vida, um estudo em particular para modelos de regressão para dados discretos no Capítulo 4 que tem uma importante relação na construção de modelos com base no processo de carcinogênese e que foi mencionado no Capítulo 5 a sua interpretação no cenário de riscos competitivos latentes.

A principal contribuição deste trabalho, depois de desenvolver modelos com várias formas da função de risco, foi a construção da distribuição com a série de potência generalizada com parâmetro-inflacionado (IGPS) encontrada em Kolev et al. (2000) conjuntamente com uma rotina de maximização capaz de fazer o processo de estimação onde foi obtido resultados mais completos para a distribuição CoPE, a qual utiliza a IGPS com o caso particular de Poisson. Portanto, deixamos neste capítulo os principais resultados para as outras distribuições, os quais ainda precisam ser melhor analisados e discutidos.

Neste capítulo apresentamos, assim como na construção da distribuição CoPE, a composição como mistura da distribuição IGPS Poisson com a distribuição Exponencial para os riscos no sistema de ativação dado pelo máximo e em seguida mostramos as propriedades como consequência direta da distribuição CoPE para as outras distribuições IGPS no cenário de riscos competitivos de ativação pelo mínimo. Tais resultados são apresentados aqui como proposta futura pois ainda carecem de estudos mais detalhados, principalmente no que se diz respeito ao aperfeiçoamento do algorítimo de maximização em relação ao tempo necessário para se fazer a simulação bem como a consistência dos resultados, pois esta rotina ainda não apresenta tempo e nem retorna resultados compatíveis para tais estudos.

Para estas distribuições envolvendo a distribuição IGPS na estrutura de modelagem,

muitos softwares não são capazes de fazer o processo de maximização da logverossimilhança pela apresentação de somas infinitas nas funções, no entanto uma rotina de estimação quase-newton (BFGS) apresentada tornou possível controlar mais parâmetros do que nas rotinas de maximização amplamente utilizadas no *software* R Core Team (2012) e MATLAB (2010). Logo, tornamos possível a estimação dos parâmetros das distribuições com um custo (tempo) computacional razoável. O tempo de estimação para amostras de tamanho 100, por exemplo, apresentam em torno de 5 minutos em um computador com processador Intel Core i5-4200Y, com 4GB de memória Ram e sistema operacional Windows de 64 bits, quando em cuidados para verificar cada passo do algorítimo de estimação afim de evitar erros computacionais.

6.1 Distribuição CCoPE

Assim como mencionado no Capítulo 2, podemos construir distribuições com a série de potência somente alterando a escolha da distribuição para os riscos latentes. A seguir vamos apresentar a distribuição Complementar-CoPE (CCoPE), que é interpretada no cenário de risco latente e também as distribuições resultantes sobre cada escolha da distribuição para as causas/riscos. Aqui não daremos a mesma atenção encontrada no capítulo anterior, pois os resultados para simulação são incompatíveis com o nível de tempo requerido.

Assim como no Capítulo 2, no cenário de risco latentes podemos ter a falha do objeto em estudo pelo tempo da última ativação, ou seja, a distribuição do tempo de vida dado a quantidade de causas M é dada por $Y = \max\{T_1, T_2 \ldots, T_M\}$. Considere T_i, $i = 1, 2, 3, ...$ variáveis aleatórias dos tempos de vida, para cada risco i, tendo distribuição Exponencial, isto é, o tempo até o evento do i-ésimo fator de risco complementar (T_i) tem distribuição Exponencial com parâmetro λ, considere também que a distribuição para o número de causas M é dada por 4.3.1.

Seja a variável aleatória não negativa Y denotando o tempo de vida de um componente em uma determinada população. A variável aleatória Y terá distribuição CCoPE com parâmetros $\lambda > 0$, $\theta > 0$ e $0 < \rho < 1$ se sua f.d.p. é dada por algumas das seguintes equações

$$\begin{aligned} f(y) &= K\sum_{m=1}^{\infty} m(1-e^{-\lambda y})^{m-1}\lambda e^{-\lambda y}\rho^m H([m+1],[2],U) \\ &= \lambda e^{-\lambda y} K\sum_{m=1}^{\infty} m(1-e^{-\lambda y})^{m-1}\rho^m H([m+1],[2],U] \\ &= \frac{\lambda}{e^{\lambda y}-1}\sum_{m=1}^{\infty} m(1-e^{-\lambda y})^{m}\rho^m H([m+1],[2],U) \end{aligned}$$

em que $U = \theta(1-\rho)/\rho$ e $K = Ue^{-\theta/\rho}/(1-e^{-\theta})$.

A Figura 6.1 apresenta a f.d.p. da distribuição CCoPE para $\rho = 0.1, 0.5, 0.9$, λ=0.01 e $\theta = 1$.

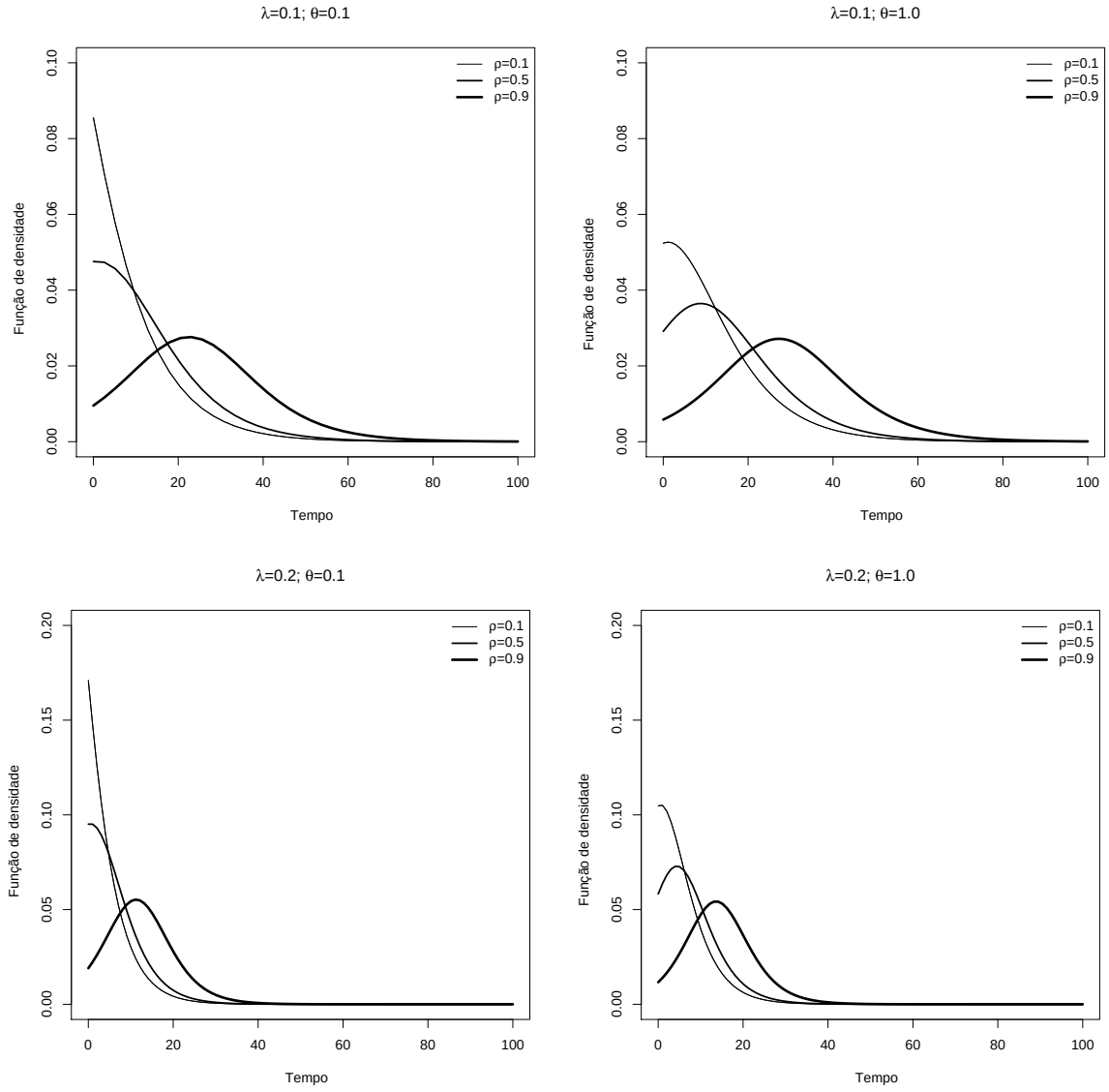

Figura 6.1: Função de densidade de probabilidade da distribuição CCoPE para $\rho = 0.1, 0.5, 0.9$, $\lambda = 0.1, 0.01$ e $\theta = 1, 0.1$.

A função de distribuição da distribuição CCoPE é dada por

$$\begin{aligned} F(y) &= K\sum_{m=1}^{\infty} m\rho^m H([m+1],[2],U)\int_0^y (1-e^{-\lambda y})^{m-1}\lambda e^{-\lambda y}dy \qquad (6.1.1)\\ &= K\sum_{m=1}^{\infty} m\rho^m H([m+1],[2],U)\frac{(1-e^{-\lambda y})^m}{m}\\ &= K\sum_{m=1}^{\infty}(1-e^{-\lambda y})^m\rho^m H([m+1],[2],U)\\ &= K\sum_{m=1}^{\infty}\sum_{i=0}^{m}\left(\begin{array}{c} m \\ i \end{array}\right) e^{-\lambda y i}\rho^m H([m+1],[2],U). \end{aligned}$$

em que U e K são dados em (6.1.1), para $\rho \in (0,1)$ e $\lambda, \theta > 0$.

A função de sobrevivência que adotamos para os cálculos no decorrer do trabalho é

$$S(y) = 1 - K\sum_{m=1}^{\infty}\rho^m(1-e^{-\lambda y})^m H([m+1],[2],U), \qquad (6.1.2)$$

para que se assemelhe as equações da distribuição CoPE já apresentada no capítulo anterior e também por apresentar uma facilidade maior para os cálculos das medidas particulares desta distribuição.

A função quantil, ou a inversa da função distribuição $F(x_p) = p$, assim como na distribuição CoPE, não é obtida algebricamente porém o p-ésimo quantil pode ser obtido resolvendo a seguinte equação numericamente

$$\frac{p}{K} = \sum_{m=1}^{\infty}\rho^m(1-e^{-\lambda x_p})^m H([m+1],[2],U),$$

em que U e K são dados em (6.1.1).

6.1.1 Função de risco, risco reversa, momentos, momentos residuais e momentos reversos

De (6.1.1) e (6.1.2), a função de risco e a função de risco reversa, de acordo com a relação $h(y) = f(y)/S(y)$ e $r(y) = f(y)/F(y)$ são dadas por

$$h(y) = \frac{\lambda e^{-\lambda y}K\sum_{m=1}^{\infty} m(1-e^{-\lambda y})^{m-1}\rho^m H([m+1],[2],U)}{1-K\sum_{m=1}^{\infty}\rho^m(1-e^{-\lambda y})^m H([m+1],[2],U)} \qquad (6.1.3)$$

e (6.1.4)

$$r(y) = \lambda e^{-\lambda y}\frac{\sum_{m=1}^{\infty} m(1-e^{-\lambda y})^{m-1}\rho^m H([m+1],[2],U)}{\sum_{m=1}^{\infty}\rho^m(1-e^{-\lambda y})^m H([m+1],[2],U)}, \text{ respectivamente.} \qquad (6.1.5)$$

O valor inicial é finito para a função de risco e infinito para a função de risco reverso, e tende a zero para a função de risco, porém não é dado por uma expressão fechada. Este

valor pode ser verificado por *softwares* computacionais, como o Maple (MapleSoft, 2012). A função de risco é unimodal, a função de risco reversa apresenta em um subspaço de forma unimodal também, e no infinito tendem ao valor $h(\infty) = 0$ e $r(\infty) = 0$. Não é possível analisar o comportamento anterior mencionado como no caso da CoPE, no entanto usando L'Hospital encontramos o fato $h(\infty) = 0$ para a função de risco e para a função de risco reversa basta observar o termo $(e^{\lambda y} - 1)^{-1}$.

A função de risco (6.1.3) e (6.1.5) são unimodais como mostra a Figura 6.2 para valores de $\rho = 0.5, 0.9$, $\lambda = 0.5, 0.1$ e $\theta = 10, 12$. Em particular, a função (6.1.5) apresenta a unimodalidade em um subespaço enquanto em outro é estritamente decrescente.

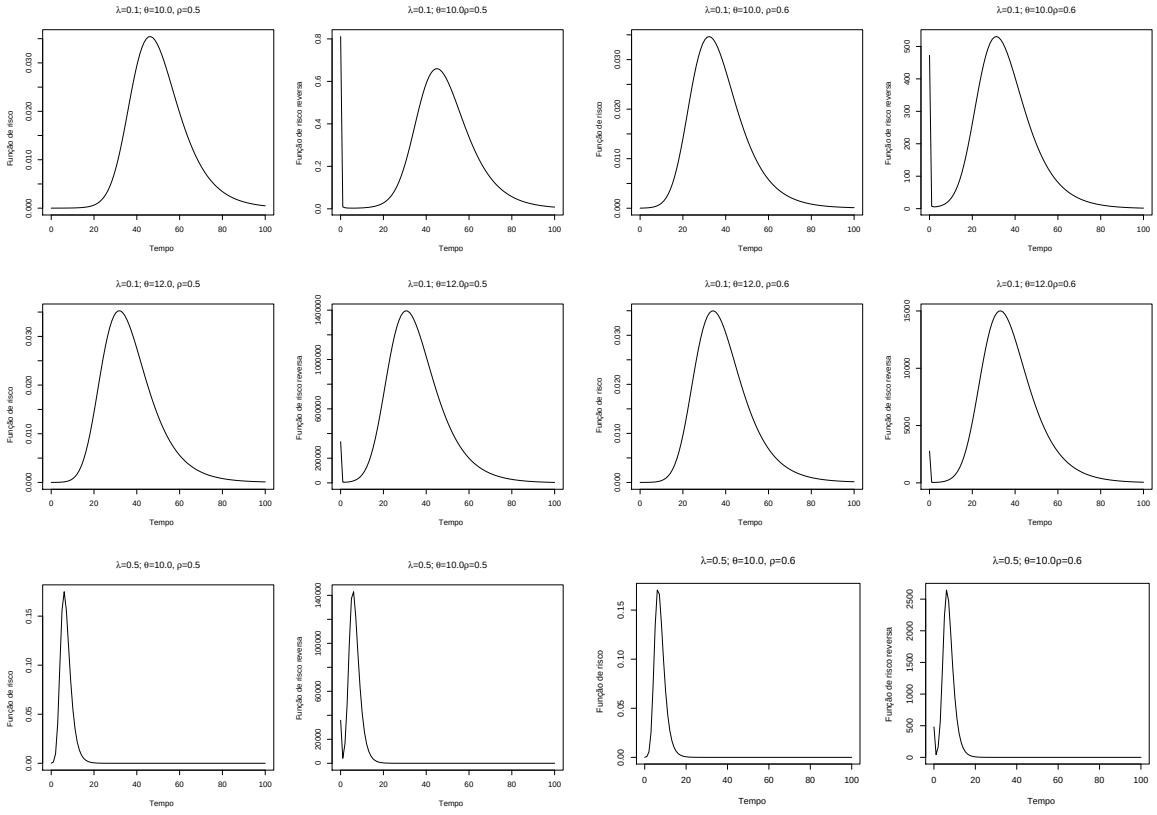

Figura 6.2: Função de risco da distribuição CCoPE para $\rho = 0.5, 0.6$, λ=0.1,0.5 e $\theta = 10, 12$.

Proposição 6.1.1 Para uma variável aleatória Y com distribuição CCoPE, o r-ésimo momento é dado por

$$\mu_r^{'} = \frac{r!K}{\lambda^r} \sum_{m=1}^{\infty} \sum_{i=0}^{m-1} \binom{m-1}{i} (-1)^i \frac{m\rho^m}{(1+i)^{r+1}} H([m+1], [2], U).$$

Prova 6.1.1 De (6.1.1), temos que

$$\begin{aligned} E(Y^r) &= \int_0^\infty y^r f(y)dy \\ &= K\lambda \sum_{m=1}^{\infty} m\rho^m H([m+1],[2],U) \int_0^\infty y^r e^{-\lambda y}(1-e^{-\lambda y})^{m-1}dy \\ &= K\lambda \sum_{m=1}^{\infty}\sum_{i=0}^{m-1} m\rho^m H([m+1],[2],U)\binom{m-1}{i}(-1)^i \int_0^\infty y^r e^{-\lambda(1+i)y}dy. \end{aligned}$$

A última integral pode ser resolvida comparando com a função Gama e assim a prova é completada. ■

Proposição 6.1.2 Para uma variável Y com distribuição CCoPE, a função característica é dada por

$$\Phi_i(t) = \lambda K \sum_{m=1}^{\infty}\sum_{j=0}^{m-1}\binom{m-1}{j}(-1)^j \frac{m\rho^m}{\lambda(1+j)-it} H([m+1],[2],\frac{\theta(1-\rho)}{\rho}).$$

Prova 6.1.2 O resultado segue diretamente expansão binomial do termo $(1-e^{\lambda y})^{m-1}$ e da integração do fator exponencial quando aplicada a definição da função característica. ■

Proposição 6.1.3 Para uma variável aleatória Y com distribuição CCoPE, o r-ésimo momento do tempo de vida residual é dado por

$$\mu_r(t) = \frac{K}{\lambda^r S(t)} \sum_{m=1}^{\infty}\sum_{i=0}^{m-1}\binom{m-1}{i}(-1)^i \frac{m\rho^m H([m+1],[2],U)}{(1+i)^{r+1}} \gamma(r+1, t(\lambda(1+i))),$$

em que $\gamma(a,x)$ é a função Gama incompleta inferior.

Prova 6.1.3 De (6.1.1) e da função de vida residual obtida pela relação $\frac{d\left(\frac{S(x)}{S(t)}\right)}{dx}$,$x \geq t$, temos que

$$\begin{aligned} E(Y^r) &= \frac{1}{S(t)}\int_t^\infty y^r f(y)dy \\ &= \frac{K\lambda}{S(t)} \sum_{m=1}^{\infty}\sum_{i=1}^{m-1}\binom{m-1}{i}(-1)^i m\rho^m H([m+1],[2],U) \int_t^\infty y^{r+1-1} e^{-\lambda(1+i)y}dy. \end{aligned}$$

A última integral pode ser resolvida com a transformação $z = y\lambda(1+i)$ e comparando com a função Gama incompleta inferior completa a prova. ■

Proposição 6.1.4 Para uma variável aleatória Y com distribuição CCoPE, o r-ésimo momento de vida residual reversa é dado por

$$\begin{aligned} m_r(t) &= \frac{Kt^r}{(r+1)F(t)} \sum_{m=1}^{\infty}\sum_{i=1}^{m}\binom{m}{i}(-1)^i \rho^m e^{\lambda it} H([m+1],[2],U)\Big(rH([1+r],[2+r],\lambda it) \\ &+ H([r],[r+2],\lambda it)\Big). \end{aligned}$$

Prova 6.1.4 Aplicando a expansão binomial no termo $(1-e^{\lambda y})^m$ na Equação (5.1.12) a integral é resolvida com o uso do *software* Maple (MapleSoft, 2012). ■

6.1.2 Curvas de Bonferroni e Lorenz

Como definifo no capítulo anterior, as curvas de Bonferroni e Lorenz são curvas definidas, respectivamente, por

$$B(p) = \frac{1}{p\mu}\int_0^p F^{-1}(x)dx \text{ e } \quad L(p) = \frac{1}{\mu}\int_0^p F^{-1}(x)dx,$$

em que $\mu = E(X)$ e $F(q) = p$.

Os indicadores de Bonferroni e Gini são definidos, respectivamente, por

$$B = 1 - \int_0^1 B(p)dp \text{ e } \quad G = 1 - 2\int_0^1 L(p)dp.$$

Proposição 6.1.5 Para uma variável aleatória Y com distribuição CCoPE, as curvas de Bonferroni e Lorenz são respectivamente, dadas por

$$\begin{aligned} B(p) &= \frac{K}{\lambda\mu p}\sum_{m=1}^{\infty}\sum_{i=0}^{m-1}\binom{m-1}{i}\frac{m\rho^m}{(i+1)^2}(-1)^i H([m+1],[2],U) \\ &\times \left(1 - e^{-\lambda(i+1)F^{-1}(p)}[1+\lambda(i+1)F^{-1}(p)]\right) \end{aligned} \tag{6.1.6}$$

e

$$\begin{aligned} L(p) &= \frac{K}{\lambda\mu}\sum_{m=1}^{\infty}\sum_{i=0}^{m-1}\binom{m-1}{i}\frac{m\rho^m}{(i+1)^2}(-1)^i H([m+1],[2],U) \\ &\times \left(1 - e^{-\lambda(i+1)F^{-1}(p)}[1+\lambda(i+1)F^{-1}(p)]\right). \end{aligned} \tag{6.1.7}$$

Prova 6.1.5 O resultado é obtido por uma simples manipulação algébrica usando o fato que $\int_a^b F^{-1}(x)dx = \int_{F^{-1}(a)}^{F^{-1}(b)} xf(x)dx$ e como resultado da Gama incompleta inferior com forma fechada com o parâmetro 2. ■

A Figura 6.3 apresenta as curvas de Bonferroni e Lorenz para $\lambda = 1$, $\theta = 0.4, 1, 2$ e $\rho = 0.3, 0.5, 0.8$. As curvas de Bonferroni e Lorenz são afetadas pelo parâmetro θ e ρ, o que é indicativo que λ é um parâmetro de escala pelas propriedades destas curvas. Salientamos que não apresentamos os gráficos para $\lambda \neq 1$ pois são idênticos aos demais e por isso chega-se a conclusão que λ é parâmetro de escala.

Para o indicador de Bonferroni e Gini não encontramos uma forma fechada, porém podem ser estimados por uma integração de Monte Carlo. Para R pontos simulados, $u_1, u_2, \ldots, u_R$, de uma distribuição Uniforme, temos que $B \approx 1-\left(\sum_{i=1}^R B(u_i)/R\right)$ e $G \approx 1-2\left(\sum_{i=1}^R G(u_i)/R\right)$.

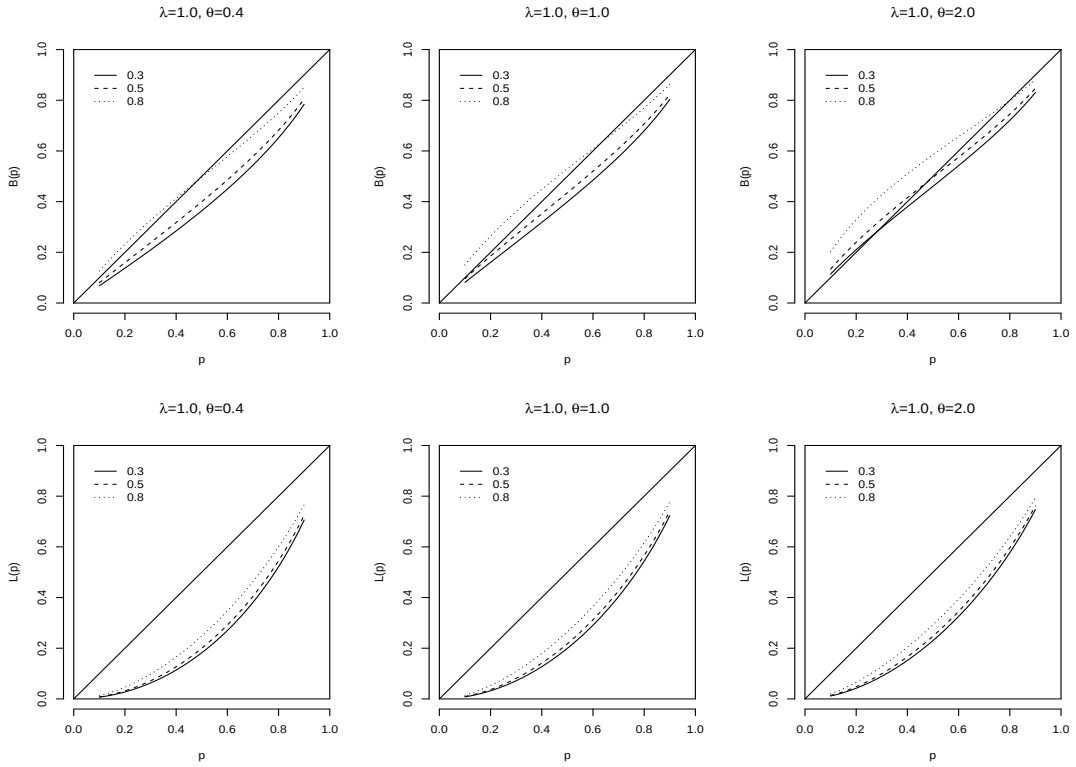

Figura 6.3: Curvas de Bonferooni $B(p)$ e Lorenz $L(p)$ para a distribuição CCoPE.

6.2 Simulação

Como mencionado no início do capítulo, a rotina de maximização ainda não apresenta rendimento satisfatório e portanto para motivos de ilustração o estudo a seguir é baseado em 50 gerações de conjunto de dados da distribuição CCoPE com diferentes conjuntos de parâmetros para $n = 15$, 30 e 50 para avaliar a rotina de maximização desenvolvida. Os tamanhos das amostras foram escolhidos desta forma pelos cálculos computacionais serem comprometidos pelo termo da soma das funções hipergeométricas. As restrições dos parâmetros α e λ foram obtidos pela transformação exponencial e para o parâmetro ρ por meio da transformação $e^{\rho^*}/(1+e^{\rho^*})$, onde $\rho^* \in \mathbb{R}$.

A Tabela 6.1 apresenta as médias dos estimadores MLEs (Av), suas variâncias (Var) e o erro quadrático médio (MSE), juntamente com a probabilidade de cobertura para níveis especificados de 95% de confiança dos parâmetros. Em geral, o processo de estimação não apresentou bons resultados mas os valores do MSE bem como os valores Var decrescem quando o tamanho da amostra aumenta. É importante observar que a probabilidade de cobertura ainda não está próxima dos valores determinados, porém quando o tamanho da amostra aumenta o valor da cobertura tende a aumentar. Ainda, baseados em outros estudos de simulação, não

apresentados aqui pois a rotina não apresenta estabilidade numérica, os valores estimados de θ, bem como os suas respectivas variâncias e MSE, são afetados drasticamente conforme os valores de ρ aumentam. Também é importante mencionar que a escolha para a representação da função de sobrevivência, ou pela relação $1 - F(y)$ dada pela Equação 6.1.1 influenciam diretamente numa melhoria dos resultados. Para a tabela em questão foi utilizado a representação da segunda linha da Equação 6.1.1 pois foi a que apresentou resultados menos conflitantes no sentido do tamanho da amostra com a cobertura, a variância e o MSE.

Tabela 6.1: Média dos estimadores MLEs, suas variâncias, MSE e coberturas da distribuição CCoPE para dados Simulados.

Parâmetros	θ	λ	ρ	θ	λ	ρ	θ	λ	ρ
Valores	0.5	0.5	0.2	0.3	0.1	0.8	0.3	0.1	0.5
				$n = 15$					
Av	0.7174	0.5206	0.2229	0.5049	0.5998	0.1269	0.4557	0.5582	0.1680
Var	0.2342	0.1941	0.0382	1.0967	1.5028	0.0329	0.1573	0.4330	0.0345
MSE	0.2767	0.1906	0.0380	1.1167	1.7226	0.4853	0.1784	0.6343	0.1440
CP (95%)	0.1000	0.6200	0.6000	0.8163	0.6327	0.3673	0.7163	0.7755	0.5714
				$n = 30$					
Av	0.5920	0.5527	0.1027	0.2897	0.5527	0.1146	0.3345	0.7806	0.1820
Var	0.5071	0.6305	0.0202	0.0708	0.5595	0.0250	0.2356	0.9655	0.0404
MSE	0.5054	0.6207	0.0292	0.0695	0.7532	0.4943	0.2321	1.4093	0.1407
CP(95%)	0.2000	0.4800	0.3800	0.9574	0.9149	0.3404	0.7292	0.9167	0.6875
				$n = 50$					
Av	0.3654	1.4629	0.1447	0.1624	1.3511	0.1768	0.1911	1.5058	0.2180
Var	0.2694	2.1864	0.0251	0.0166	5.0330	0.0493	0.0182	3.7835	0.0403
MSE	0.2822	3.0699	0.0277	0.0353	6.4975	0.4367	0.0297	5.6840	0.1190
CP (95%)	0.3000	0.8000	0.4800	0.7872	0.9149	0.5106	0.7778	0.9222	0.8667

A estabilidade numérica mencionada se resume ao fato que sem o acompanhamento humano da rotina de maximização, ocorrem saltos representativos nos pontos estimados durante a rotina e estes saltos deveriam ser controlados pela mudança dos parâmetros como descrito no Anexo E, porém o acompanhamento para cada amostra é inviável. Os parâmetros para o tamanho do passo da rotina apresentada no Anexo E são $a = 0.08$, $rhoo = 0.5$ e $c = 0.00001$ e foram escolhidos observando o comportamento em amostras simuladas de tamanho $n = 30$ e $\rho = 0.2$.

6.3 Distribuição CoNBE e sua comparação com a CoPE

Considerando o mesmo procedimento adotado na Seção 5.1.1, podemos atribuir outras distribuições para o número de causas competitivas latentes que consideram o parâmetro de inflação ρ, como alternativas às distribuições CoPE e CCoPE propostas anteriormente.

A seguir apresentamos a distribuição CoNBE, que consiste em considerar que o número de causas latentes, no cenário de risco competitivos, apresenta distribuição INB (Binomial Negativa com parâmetro inflacionado) dada pela Equação (ii) da Proposição 1 de Minkova (2002). A distribuição CoNBE é obtida seguindo os mesmos procedimentos da Seção (5.1.1), ou seja, considerando que os tempos de vida de cada risco tem distribuição Exponencial com parâmetro λ e é observado o tempo de falha do objeto em estudo que é ocasionado pelo tempo de vida da primeira falha latente.

Seja a variável aleatória não negativa Y denotando o tempo de vida de um componente em uma determinada população. A variável aleatória Y terá distribuição CoNBE com parâmetros $\lambda > 0$, $r > 0$, $0 < \theta < 1$ e $0 < \rho < 1$ se sua f.d.p. é dada por

$$f(y) \;=\; K\lambda \sum_{m=1}^{\infty} m\rho^m e^{-\lambda y m} H([m+1, r+1], 2, U), \tag{6.3.1}$$

em que $U = \frac{(1-\theta)(1-\rho)}{1-\theta(1-\rho)}$ e $K = \frac{r\theta^r \rho^r U^{r+1}}{(1-\theta^r)(1-\theta)^r(1-\rho)^r}$.

Pela Equação (6.3.1), o leitor pode perceber a semelhança com a equação (5.1.1). Desta forma todas os resultados apresentados no capítulo da distribuição CoPE, menos as da matriz de informação, podem ser replicadas para a distribuição CoNBE, simplesmente fazendo a troca dos termos K, U pelos correspondentes e da hipergeométrica generalizada $H([m+1], [2], U)$ pela correspondente $H([m+1, r+1], [2], U)$.

A Figura 6.3 apresenta a f.d.p. da distribuição CoNBE para $\rho = 0.1, 0.5, 0.9$, λ=0.5,0.05, $\theta = 0.5, 0.9$ e $r = 1, 3$.

A função de risco da CoNBE é decrescente como mostra a Figura 6.5 para valores de $\rho = 0.5, 0.9$, $\lambda = 0.05$ e $\theta = 0.9$ e $r = 2$.

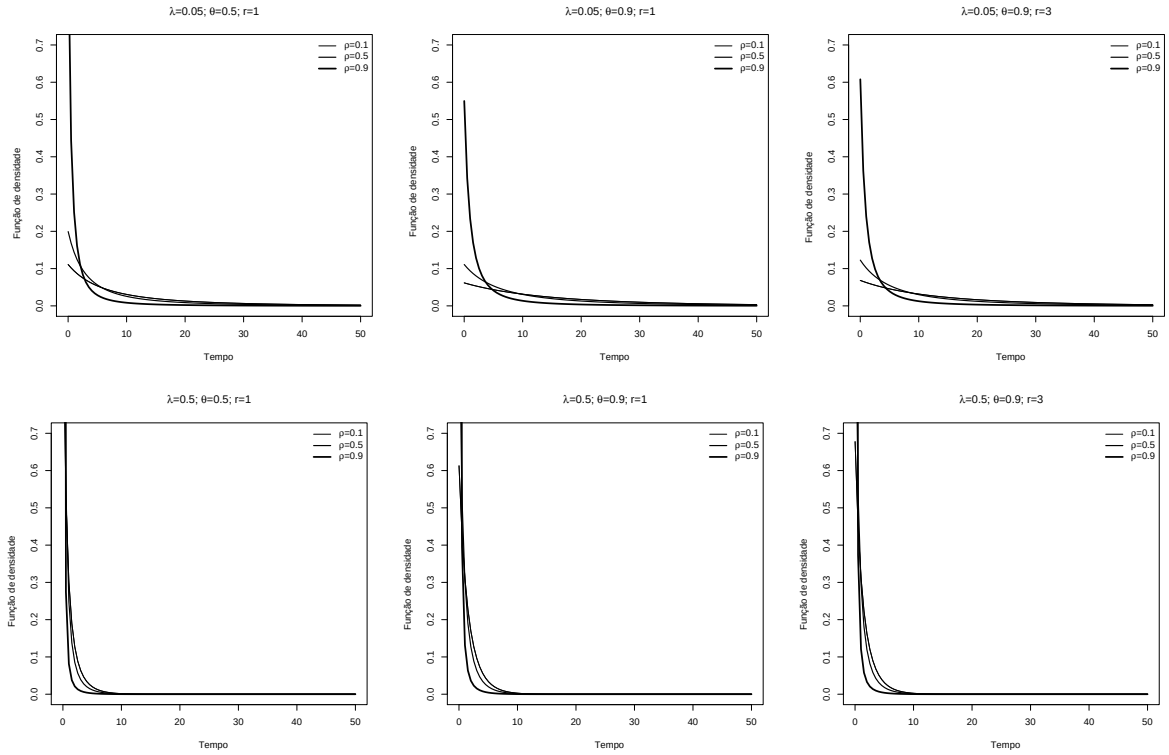

Figura 6.4: Função de densidade de probabilidade da distribuição CoNBE para $\rho = 0.1, 0.5, 0.9$, λ=0.5,0.05, $\theta = 0.5, 0.9$ e $r = 1, 3$.

6.4 Distribuição CoLSE

A seguir apresentamos a distribuição CoLSE, que consiste em considerar que o número de causas latentes, no cenário de risco competitivos, apresenta distribuição ILS (Série Logarítima com parâmetro inflacionado) dada pela Equação (iv) da Proposição 1 de Minkova (2002). A distribuição CoLSE é obtida seguindo os mesmos procedimentos da Seção 5.1.1, ou simplesmente, com os mesmos passos da seção (6.3).

Seja a variável aleatória não negativa Y denotando o tempo de vida de um componente em uma determinada população. A variável aleatória Y terá distribuição CoLSE com parâmetros $\lambda > 0$, $0 < \theta < 1$ e $0 < \rho < 1$ se sua f.d.p. é dada por

$$f(y) = -\frac{\lambda}{\ln(\theta)} \frac{1-\rho-\theta(1-\rho)}{(1-e^{-\lambda y}\rho)(e^{\lambda y}-1+\theta(1-\rho))}. \tag{6.4.1}$$

A função de sobrevivência da distribuição CoLSE é dada por uma das seguintes equações

$$\begin{aligned} S(y) &= 1 + \frac{1}{\ln(\theta)} \sum_{m=1}^{\infty} \frac{(1-e^{-\lambda y m})([1-\theta(1-\rho)]^m - \rho^m)}{m} \qquad (6.4.2) \\ &= 1 + \frac{1}{\ln(\theta)} \left[-\ln(\theta(1-\rho)) + \ln(1-\rho) + \ln(1-e^{-\lambda y}[1+\theta(1-\rho)]) - \ln(1-\rho e^{-\lambda y}) \right]. \end{aligned}$$

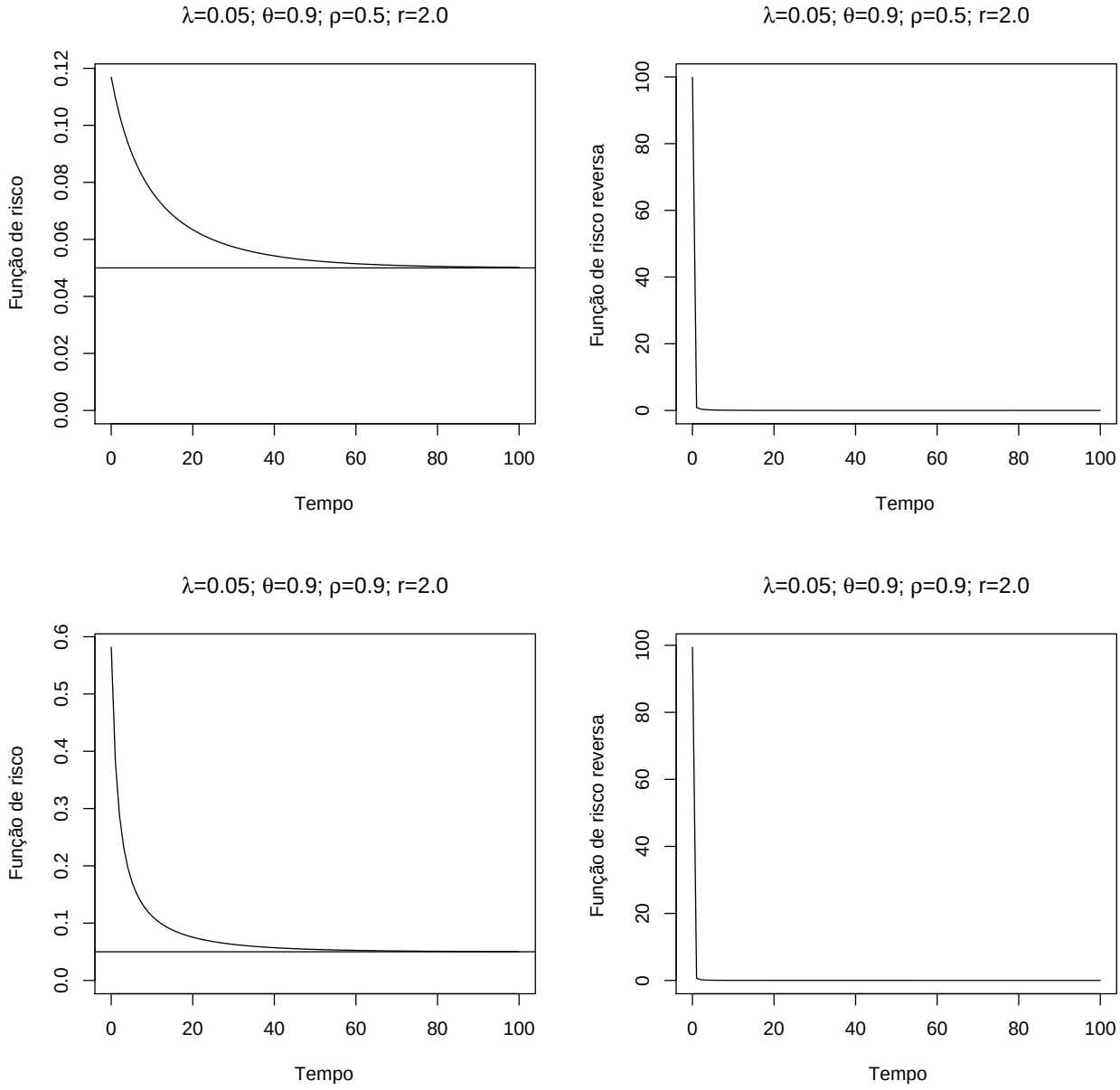

Figura 6.5: Função de risco e risco reverso da distribuição CoNBE para $\rho = 0.5, 0.9$, $\lambda = 0.05$ e $\theta = 0.9$ e $r = 2$.

A Figura 6.4 apresenta a f.d.p. da distribuição CoLSE para $\rho = 0.1, 0.5, 0.9$, λ=0.1,0.5 e $\theta = 0.5, 0.9$.

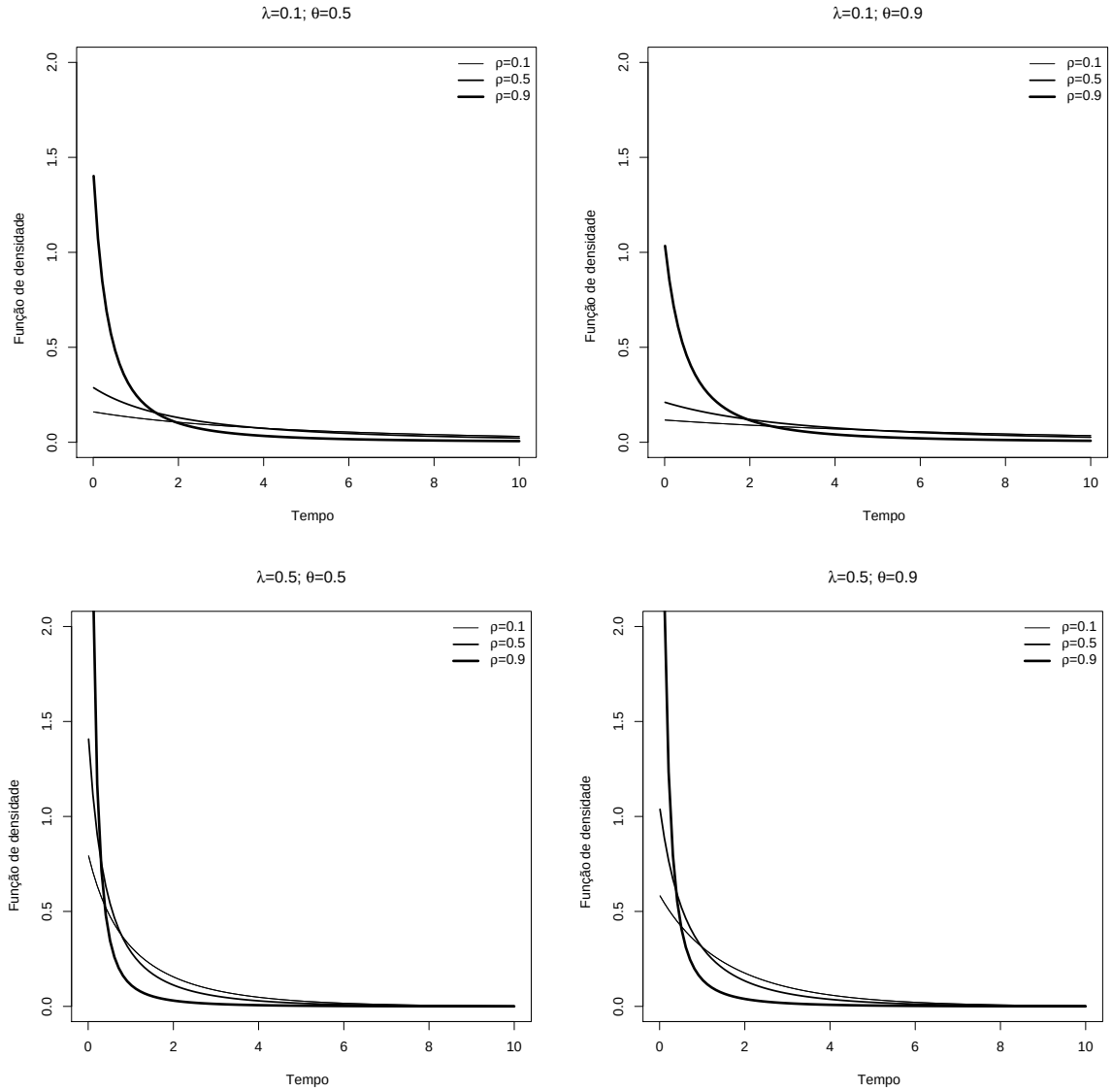

Figura 6.6: Função de densidade de probabilidade da distribuição CoLSE para $\rho = 0.1, 0.5, 0.9$, λ=0.1,0.5 e $\theta = 0.5, 0.9$.

6.4.1 Função de risco

De (6.4.1) e (6.4.2), a função de risco, de acordo com a relação $h(y) = f(y)/S(y)$, é dada por

$$h(y) = \frac{-\frac{\lambda}{\ln(\theta)} \frac{1-\rho-\theta(1-\rho)}{(1-e^{-\lambda y}\rho)(e^{\lambda y}-1+\theta(1-\rho))}}{1 + \frac{1}{\ln(\theta)} \left[-\ln(\theta(1-\rho)) + \ln(1-\rho) + \ln(1-e^{-\lambda y}[1+\theta(1-\rho)]) - \ln(1-\rho e^{-\lambda y})\right]} \quad (6.4.3)$$

Proposição 6.4.1 Se uma variável aleatória Y tem distribuição CoLSE, então a função de risco é estritamente decrescente.

Prova 6.4.1 Definindo $\eta(y) = -f'(y)/f(y)$ em que $f'(y)$ é a primeira derivada da f.d.p. (6.4.1). Desta forma temos que

$$\eta'(y) = \frac{\lambda^2 e^{-\lambda y}(\theta(1-\rho)-1)}{[1+e^{-\lambda y}(\theta(1-\rho)-1)]^2} - \frac{\lambda\rho e^{\lambda y}}{(e^{\lambda y}-\rho)^2}. \tag{6.4.4}$$

Como $(\theta(1-\rho)-1) < 0$, temos $\eta'(y) < 0$ para $y > 0$ e pelo Teorema de Glaser (Glaser, 1980) concluímos diretamente que a forma da função de risco (2.2.3) é decrescente.

■

O valor inicial é sempre finito é dado por uma expressão fechada. A função de risco é decrescente e no infinito tende ao valor $h(\infty) = \lambda$, que pode ser comprovado pela regra de L'Hôpital. Aplicando a regra, a equação recai sobre uma função semelhante a $\eta(y)$ em que é fácil notar o resultado $h(\infty) = \lambda$.

A função de risco (6.4.3) é decrescente como mostra a Figura 6.7 para valores de $\rho = 0.5, 0.8$, $\lambda = 1.0, 0.1, 1.5$ e $\theta = 0.3, 0.5$.

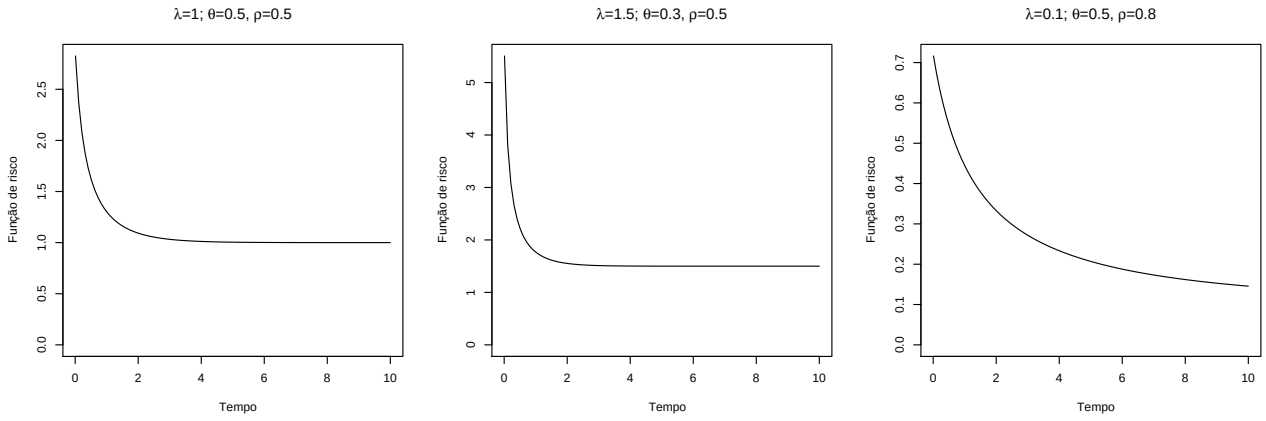

Figura 6.7: Função de risco da distribuição CoLSE para $\rho = 0.5, 0.8$, $\lambda = 1.0, 0.1, 1.5$ e $\theta = 0.3, 0.5$.

6.4.2 Função de risco reversa

De (6.4.1) e (6.4.2), a função de risco, de acordo com a relação $r(y) = f(y)/F(y)$, é dada por

$$r(y) = \frac{\lambda[1-\rho-\theta(1-\rho)][(1-e^{-\lambda y}\rho)(e^{\lambda y}-1+\theta(1-\rho))]^{-1}}{-\ln(\theta(1-\rho))+\ln(1-\rho)+\ln(1-e^{-\lambda y}[1+\theta(1-\rho)])-\ln(1-\rho e^{-\lambda y})}. \tag{6.4.5}$$

Proposição 6.4.2 Se uma variável aleatória Y tem distribuição CoLSE, então a função de risco reversa é estritamente decrescente.

Prova 6.4.2 A prova é obtida diretamente pela Proposição B.11 do livro Marshall e Olkin (2007, pg. 14). ■

6.5 Aplicação em dados reais

Considere novamente os conjuntos de dados $T10$, $T11$ e $T12$ apresentados no Capítulo 5, com uma mudança em $T10$ e $T11$ a qual dividimos o tempo de sobrevivência por 10 e chamados aqui de $TM10$ e $TM11$. Esta mudança foi feita para que não precisássemos escolher um ponto inicial diferente para a maximização da distribuição CCoPE, visto que ela tem o termo $e^{-\lambda y}$ na divisão e começando com $\lambda = 1$ no algorítimo, o computador faria uma divisão por zero, pois para valores de $y > 375$ o programa retorna que $e^{-y} = 0$.

A Tabela 6.2 apresenta as estimativas das distribuições CoPE, CCoPE, CoNBE e CoLSE para os dados $TM10$, $TM11$ e $T12$, juntamente com seus desvios padrões em parênteses e os valores $-\ell(.)$ e AIC.

Tabela 6.2: Parâmetros estimados das distribuições CoPE, CCoPE, CoNBE e CoLSE para os dados TM10, TM11 e T12.

		λ	θ	ρ	r	$-\max \ell(.)$	AIC
TM10	CoPE	0.064(0.018*)	0.649(0.132*)	0.518(0.053*)	-	64.919	135.838
	CCoPE	0.117(0.034*)	0.704(0.232*)	0.515(0.079*)	-	79.395	164.790
	CoNBE	0.031(0.426*)	0.525(0.488*)	0.480(0.789*)	0.907(0.003)	67.733	143.466
	CoLSE	0.057(0.088*)	0.501(0.506*)	0.483(0.777*)	-	67.827	141.654
TM11	CoPE	0.042(0.214*)	0.628(0.595*)	0.520(0.111*)	-	152.816	311.632
	CCoPE	0.103(0.041*)	0.735(0.194*)	0.515(0.065*)	-	181.628	369.256
	CoNBE	0.021(0.004)	0.525(0.892*)	0.480(0.001)	0.908(0.005)	159.519	327.038
	CoLSE	0.034(0.006)	0.502(0.506*)	0.483(0.001)	-	165.569	337.192
T12	CoPE	0.037(0.002*)	0.735(0.002)	0.514(0.560*)	-	99.1935	204.386
	CCoPE	0.060(0.029*)	0.554(0.136*)	0.642(0.056*)	-	117.047	240.094
	CoNBE	0.003(0.677*)	0.517(0.512*)	0.487(0.790*)	0.937(0.002)	121.377	250.754
	CoLSE	0.011(0.014)	0.985(3.480)	0.784(0.175)	-	108.903	223.806

$*\times 10^{-3}$

Segundo os critérios de $-\ell(.)$ e AIC, o modelo CoPE é o que melhor se ajusta nos três conjuntos de dados $TM10$, $TM11$ e $T12$. As estimativas dos desvios padrões foi obtido da

matriz hessiana do algorítimo, ou seja, da matriz de informação de Fisher observada de cada modelo.

6.6 Comentários

Até o momento, apresentamos as funções de densidade e de sobrevivência da distribuição CCoPE, CoNBE e CoLSE e algumas propriedades como o momento, momento residual para a distribuição CCoPE. A relação da distribuição CoNBE e da CoPE é apresentada e todos os resultados apresentados no Capítulo 5 se aplicam a CoNBE, menos a matriz de informação de Fisher observada. Para a distribuição CoLSE somente foi apresentado as formas de sua função de risco pois a expansão binomial não pode ser aplicada como nos demais casos. Por fim apresentamos as estimativas das distribuições aqui apresentadas em três conjuntos de dados e se comparadas com as estimativas obtidas nas Tabelas 5.3 e 5.4 podemos confirmar que é viável a aplicação das distribuições e ainda melhoramentos na rotina de maximização devem ser desenvolvidos para não precisar de acompanhamento humano devido aos valores característicos de cada amostra. As estimativas foram obtidas por meio da rotina BFGS implementada no *software* MATLAB (2010), para as distribuições CCoPE e CoNBE e pelo *software* R Core Team (2012) para a distribuição CoLSE.

6.7 Propriedades diretas para longa duração das distribuições com parâmetro de inflação ρ

Como já dito anteriormente, a análise de sobrevivência é considerada em muitas áreas como saúde, ciência financeira e confiabilidade em industrias. Em análises anteriores direcionamos os estudos aos casos de riscos competitivos latentes e a distribuição das causas determinavam pelo menos a presença de um fator de risco. Entretanto em alguns casos, a população em estudo apresenta indivíduos que não são suscetíveis ao evento de interesse, ou seja, não apresentam fatores de riscos. Em outras palavras, a distribuição das causas é degenerada em zero. Muitos autores trabalham com o modelo de mistura, como por exemplo Boag (1949) e Berkson e Gage (1952). O leitor que querer aprofundar no tema também pode ler Maller e Zhou (1996). Mais recentemente Rodrigues et al. (2009a) fizeram a unificação dos modelos com mistura com os modelos gerados a partir da função geradora $A(s) = a_0 + a_1 p_1 + a_2 p_2 + \ldots$, para a função de sobrevivência s dos riscos competitivos latentes com estrutura determinada

pela primeiro falha dos riscos.

Para os modelos de longa duração, utilizaremos as distribuições IGPS e estas que serão apresentadas serão representadas com a adição a letra "L-" na inicial do nome das respectivas distribuições. Ainda, para as distribuições, CoPE, CoNBE e CoLSE usaremos para suas f.d.p. a nomenclatura $f(y)$ e $S(y)$ para a f.d.p e a função de sobrevivência, e as distinguiremos quando necessário.

6.8 Modelos de mistura

Para explicar melhor os modelos de mistura, e consequentemente ajudar a entender a extensão direta dos resultados, considere que os indivíduos suscetíveis ao evento de risco são mencionados por ER (em risco), ou imunes, enquanto os indivíduos não suscetíveis ao evento são mencionados por FR (fora de risco - ou imunes), ou seja, indivíduos FR tem tempo de vida infinitos (considerando somente sobre o evento de interesse). A cada parte da população existe uma porcentagem de indivíduos ER, e portanto uma probabilidade de um indivíduo fazer parte destas população.

O modelo de tempo de vida de longa duração, usando os componentes de mistura são construídos da seguinte forma. Seja a variável aleatória Y representando o tempo até a ocorrência do evento de interesse, e p a probabilidade de um indivíduo pertencer ao grupo FR. Considerando que existe a possibilidade de um indivíduo não ser suscetível ao evento de interesse, a função de sobrevivência imprópria é dada por

$$S(y) = pS_{FR}(y) + (1-p)S_{ER}(y),$$

em que os termos $S_{FR}(y)$ and $S_{ER}(y)$ são as funções de sobrevivência dos indivíduos FR e ER, respectivamente. Aqui vale lembrar que $S_{FR}(y) = 1$ e portanto o modelo de mistura é dado por

$$S(y) = p + (1-p)S_{ER}(y). \tag{6.8.1}$$

A equação (6.8.1) é popularmente conhecida como Modelo de Mistura de Boag (1949), ou de Berkson e Gage (1952).

6.9 Modelos pela função geradora

O modelo de longa duração por meio da função geradora de sequência introduzida por Feller (1968) é apresentada por Rodrigues et al. (2009a), relacionando este modelo com o

modelo de mistura de Boag.

Segundo Rodrigues et al. (2009a), dada uma função de sobrevivência própria, $S(y)$, então a função de sobrevivência de longa duração é dada pela relação

$$S_p(y) = A(S(y)) = \sum_{m=0}^{\infty} p_m [S(y)]^m. \tag{6.9.1}$$

Para mais detalhes, recomenda-se ler Rodrigues et al. (2009a) e Rodrigues et al. (2009b).

6.9.1 Relação do modelo de mistura com a função geradora

Rodrigues et al. (2009a) estabelece a relação da distribuição de Boag com a função geradora de Feller (1968). Abaixo mostramos a relação, que é parte essencial para estabelecermos as relações das distribuições dos Capítulos 5 e 6.1.

Pela equação (6.9.1), temos a seguinte expressão

$$S_p(y) = \sum_{m=0}^{\infty} p_m [S(y)]^m = p_0 + \sum_{m=1}^{\infty} p_m [S(y)]^m. \tag{6.9.2}$$

Novamente, recorrendo a equação (7) em Rodrigues et al. (2009a), temos a relação com o modelo de mistura dada por

$$S_p(y) = p_0 + (1 - p_0) S^*(y), \tag{6.9.3}$$

em que $S^*(y)$ é uma função de sobrevivência própria.

A seguir apresentaremos as funções de longa duração para os modelos encontrados nos Capítulos 5 e 6.1, O valor p_0 é obtido diretamente das distribuições IGPS dadas em Kolev et al. (2000), isto é, não haverá a adição de parâmetros das distribuições já apresentadas, a menos da CoLSE, havendo assim o aproveitamento de vários resultados.

6.10 Modelo de longa duração com as distribuições IGPS com risco latentes exponenciais

Nesta seção apresentaremos os modelos de longa duração para as distribuições CoPE, CoNBE e CoLSE. Primeiramente será apresentado a equação (6.9.3) em sua forma geral para os modelos mencionados e suas relações diretas para os momentos e suas derivações, como os momentos de vida residual.

6.10.1 Relação entre os modelos com risco latentes sem e com longa duração

De acordo com a equação (6.9.3), considerando os riscos competitivos latentes exponenciais, temos que

$$S_p(y) = P(M = 0) + (1 - P(M = 0)) \sum_{m=1}^{\infty} P(M = m)[S(y)]^m, \tag{6.10.1}$$

em que M tem distribuição IGPS como em Minkova (2002) e $S(y) = e^{-\lambda y}$. Observamos aqui que o valor $P(M = 0)$ é obtido diretamente da distribuição IGPS e o termo $(1 - P(M = 0)) \sum_{m=1}^{\infty} P(M = m)[S(y)]^m$ é calculado simplesmente pela multiplicação de $(1 - P(M = 0))$ nas distribuições CoPE e CoNBE. Para a distribuição CoLSE é necessário atribuir um novo parâmetro γ, tal que $P(M = 0) = \gamma$.

Como consequência direta do resultado observado em (6.10.1) a f.d.p. de $S_p(y)$, digamos $f_p(y)$, se resume a multiplicação $(1 - P(M = 0))f(y)$, em que $f(y)$ pode ser a f.d.p. da CoPE ou da CoNBE.

Observação 6.10.1 Os momentos para as distribuições de longa duração $f_p(y)$, para as distribuições CoPE, CoNBE e CoLSE, são obtidos pela relação direta

$$\mu'_{p;r} = (1 - P(M = 0))\mu'_r,$$

em que μ'_r é o momento das distribuições CoPE, CoNBE e CoLSE, respectivamente.

Como consequência do resultado acima, podemos por exemplo obter as relações do momento e da variância por

$$\begin{aligned} E_p(Y) &= (1 - P(M = 0))\mu_1 \\ &\text{e} \\ Var_p(Y) &= (1 - P(M = 0))[(\mu_2 - \sigma^2)P(M = 0) + \sigma^2], \end{aligned}$$

em que $\mu_1 = E(X)$, $\mu_2 = E(X^2)$ e $\sigma^2 = Var(X)$ com $X \sim$ CoPE/CoNBE/CoLSE.

Proposição 6.10.1 O momento residual para as distribuições de longa duração $f_p(y)$, das distribuições CoPE, CoNBE e CoLSE, são obtidos pela relação

$$\mu^t_{r;p} = \frac{S(t)(1 - P(M = 0))}{S_p(t)}\mu^t_r,$$

em que μ^t_r é o momento residual de $f(y)$.

Prova 6.10.1 Como mencionado anteriormente, distribuição de vida residual de uma variável aleatória X, tem função de sobrevivência dada por $\frac{S(x)}{S(t)}$, com $x > t$ e portando tem f.d.p. dada por $\frac{f(x)}{S(t)}$, com $x > t$. Por meio desta relação temos

$$\begin{aligned}
\mu_{r;p}^{t} &= \int_t^{\infty} (y-t)^r \frac{f_p(y)}{S_p(t)} dy \\
&= \int_t^{\infty} \frac{(1-P(M=0))(y-t)^r f(y)}{S_p(t)} \\
&= \frac{(1-P(M=0))S(t)}{S_p(t)} \int_t^{\infty} (y-t)^r \frac{f(y)}{S(t)} \\
&= \frac{(1-P(M=0))S(t)}{S_p(t)} \mu_r^t.
\end{aligned}$$

■

Proposição 6.10.2 O momento residual reverso para as distribuições de longa duração $f_p(y)$, das distribuições CoPE, CoNBE e CoLSE, são obtidos pela relação

$$m_{r;p}^{t} = \frac{F(t)(1-P(M=0))}{F_p(t)} m_r^t,$$

em que m_r^t é o momento residual reverso de $f(y)$.

Prova 6.10.2 A distribuição de vida residual reversa de uma variável aleatória X, tem função de distribuição dada por $\frac{F(x)}{F(t)}$, com $x > t$. Portando tem f.d.p. dada por $\frac{f(x)}{F(t)}$, com $x > t$. Por meio desta relação temos que

$$\begin{aligned}
\mu_{r;p}^{t} &= \int_0^{t} (t-y)^r \frac{f_p(y)}{F_p(t)} dy \\
&= \int_0^{t} \frac{(1-P(M=0))(t-y)^r f(y)}{F_p(t)} \\
&= \frac{(1-P(M=0))F(t)}{F_p(t)} \int_0^{t} (t-y)^r \frac{f(y)}{F(t)} \\
&= \frac{(1-P(M=0))F(t)}{F_p(t)} m_r^t.
\end{aligned}$$

■

6.10.2 Distribuição L-CoPE

Pela relação da Equação (6.10.1) e pelos resultados de Minkova (2002), temos que a distribuição de longa duração com riscos competitivos latentes exponenciais gerada pela distribuição correlacionada Poisson, chamada de L-CoPE, tem função de sobrevivência e f.d.p. dada, respectivamente por

$$S_p(y) = e^{-\theta} + K\sum_{m=1}^{\infty}\rho^m e^{-\lambda y m} H([m+1],[2],U) \quad (6.10.2)$$

e

$$f_p(y) = K\lambda\sum_{m=1}^{\infty} m\rho^m e^{-\lambda y m} H([m+1],[2],U),$$

em que $K = e^{-\theta/\rho}U$ e $U = \theta(1-\rho)/\rho$.

É interessante notar que a diferença entre a f.d.p. da distribuição L-CoPE para a CoPE difere somente no valor de K, que não é mais dividido por $1 - e^{-\theta} = 1 - P(M = 0)$.

A Figura 6.10.2 apresenta a função sobrevivência e a f.d.p. da distribuição L-CoPE para $\rho = 0.1, 0.5, 0.9$, $\lambda = 0.1, 0.5$ e $\theta = 0.5, 0.9$.

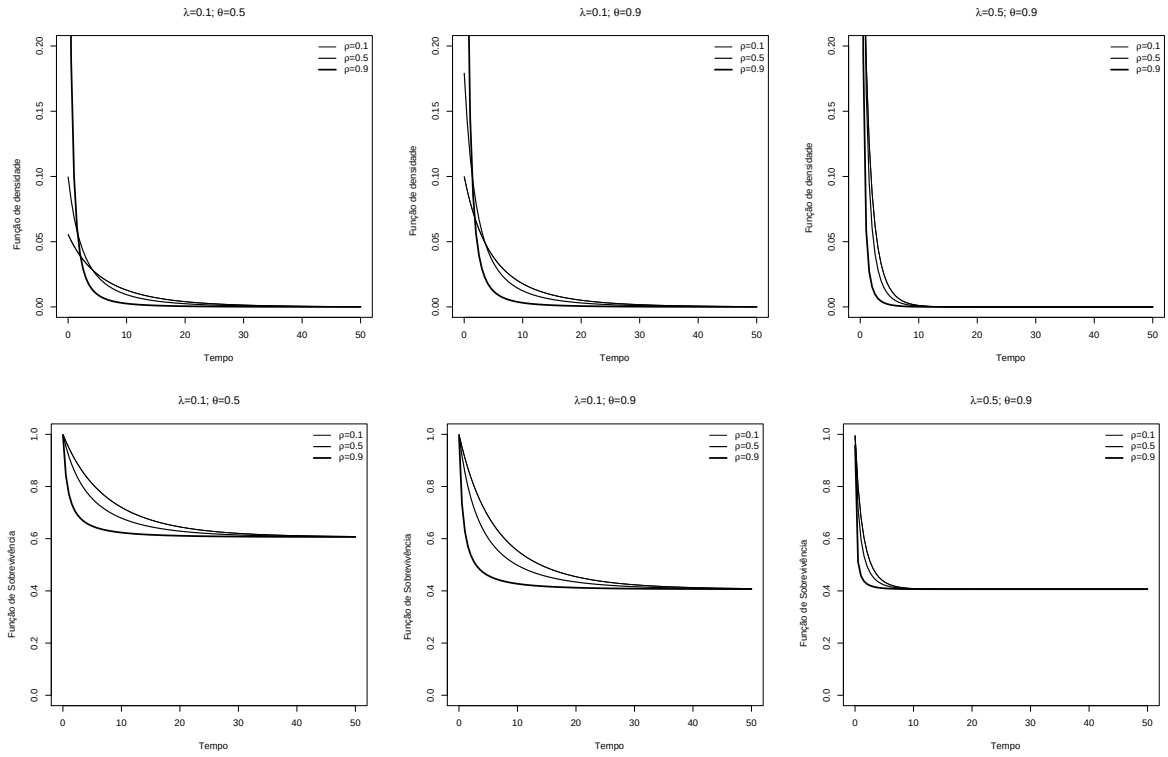

Figura 6.8: F.d.p. e função de sobrevivência da distribuição L-CoPE para $\rho = 0.1, 0.5, 0.9$ e λ=0.1,0.5, $\theta = 0.5, 0.9$.

6.10.3 Distribuição L-CoNBE

Pela relação da equação (6.10.1) e pelos resultados de Minkova (2002), temos que a distribuição de longa duração com riscos competitivos latentes exponenciais gerada pela distribuição correlacionada Binomial Negativa, chamada de L-CoNBE, tem função de sobrevivência e f.d.p. dada respectivamente, por

$$S_p(y) = \theta^r + K\sum_{m=1}^{\infty}\rho^m e^{-\lambda y m} H([m+1, r+1], [2], U) \quad (6.10.3)$$

e

$$f_p(y) = K\lambda\sum_{m=1}^{\infty} m\rho^m e^{-\lambda y m} H([m+1, r+1], [2], U),$$

em que $U = \frac{(1-\theta)(1-\rho)}{1-\theta(1-\rho)}$ e $K = \frac{r\theta^r\rho^r U^{r+1}}{(1-\theta)^r(1-\rho)^r}$.

É interessante notar que a diferença entre a f.d.p. da distribuição L-CoNBE para a CoNBE difere somente no valor de K, que não é mais dividido por $1-\theta^r = 1-P(M=0)$.

A Figura 6.10.3 apresenta a função sobrevivência e a f.d.p. da distribuição L-CoNBE para $\rho = 0.1, 0.5, 0.9$ e λ=0.1,0.5, $\theta = 0.9$ e $r = 1.5, 3.0$.

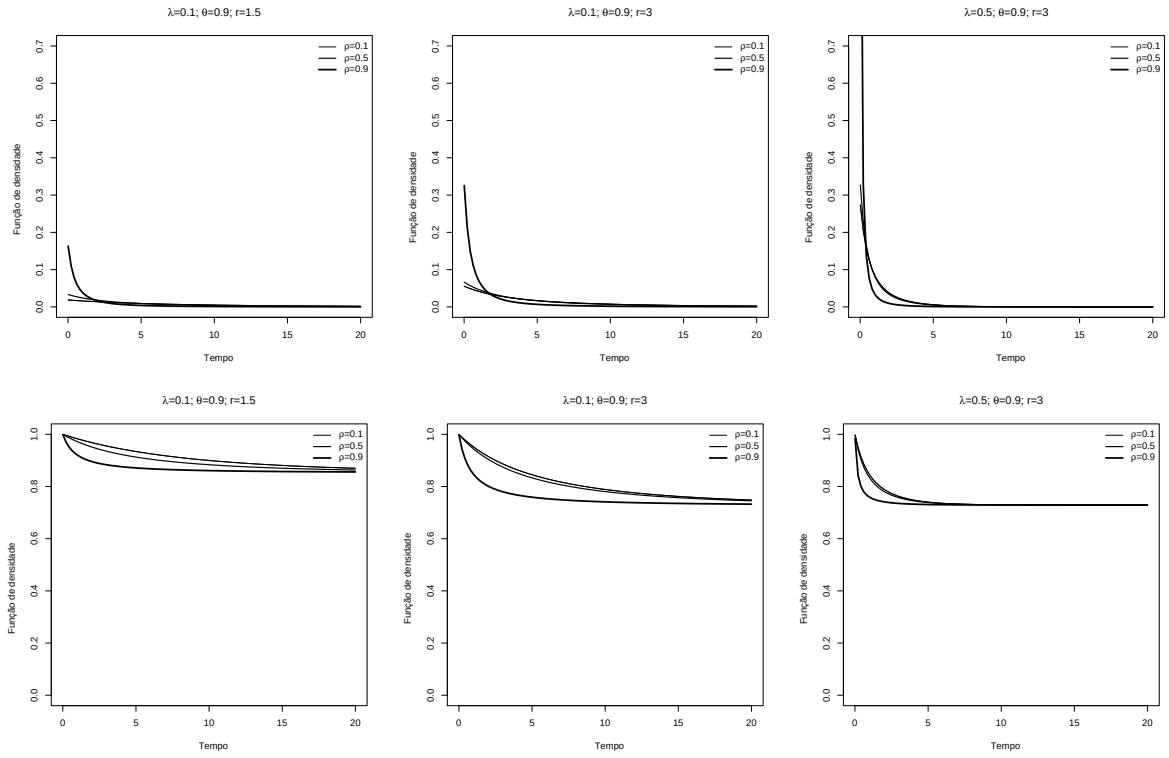

Figura 6.9: F.d.p. e função de sobrevivência da distribuição L-CoNBE para $\rho = 0.1, 0.5, 0.9$ e λ=0.1,0.5, $\theta = 0.9$ e $r = 1.5, 3.0$.

6.10.4 Distribuição L-CoLSE

Para a distribuição de longa duração com riscos competitivos latentes exponenciais gerada pela distribuição correlacionada Série Logarítima, chamada de L-CoLSE, devemos acrescentar um parâmetro γ, de forma que $P(M=0) = \gamma$. Pela relação da Equação (6.10.1) e pelos

resultados de Minkova (2002), a distribuição L-CoLSE tem função de sobrevivência e f.d.p. dada respectivamente, por

$$S_p(y) = \gamma + (1-\gamma)\left[1 + \frac{1}{\ln(\theta)}\sum_{m=1}^{\infty}\frac{(1-e^{-\lambda y m})([1-\theta(1-\rho)]^m - \rho^m)}{m}\right] \quad (6.10.4)$$

e

$$f_p(y) = -(1-\gamma)\frac{\lambda}{\ln(\theta)}\frac{1-\rho-\theta(1-\rho)}{(1-e^{-\lambda y}\rho)(e^{\lambda y}-1+\theta(1-\rho))}.$$

A Figura 6.10.4 apresenta a função Sobrevivência e a f.d.p. da distribuição L-CoLSE para $\rho = 0.1, 0.5, 0.9$ e λ=0.1,0.5, $\theta = 0.5, 0.9$ e $\gamma = 0.3$.

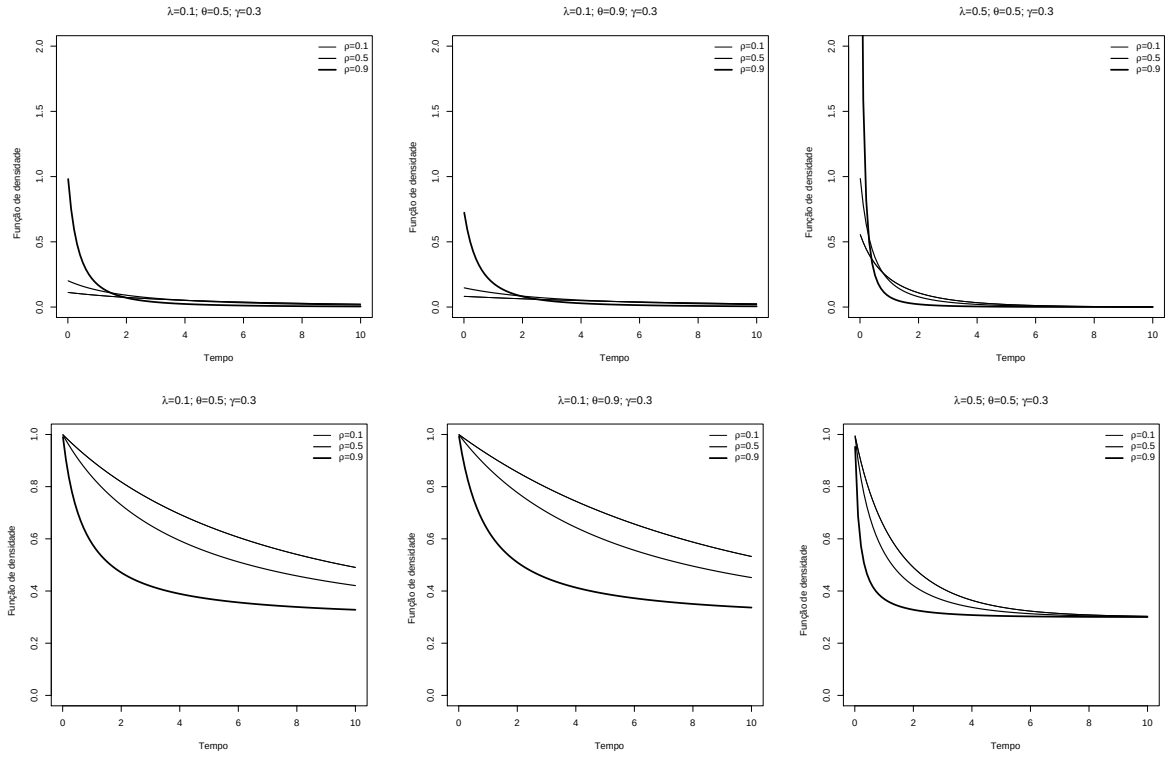

Figura 6.10: Função f.d.p. e de sobrevivência da distribuição L-CoLSE para $\rho = 0.1, 0.5, 0.9$ e λ=0.1,0.5, $\theta = 0.5, 0.9$ e $\gamma = 0.3$.

6.11 Forma da função de risco e risco reversa

Proposição 6.11.1 A função de risco e de risco reversa das distribuições de longa duração L-CoPE, L-CoNBE e L-CoLSE são estritamente decrescentes e tendem a zero quando $y \to \infty$.

Prova 6.11.1 Considere a relação $f_p(y) = (1-P(M=0))f(y)$. Portanto para as distribuições L-CoPE e L-CoNBE, a função de risco é obtida por meio da expressão $h_p(y) = f_p(y)/S_p(y)$.

Assim

$$h'_p(y) = [f'_p(y)S_p(y) + f_p^2(y)]/S_p^2(y). \tag{6.11.1}$$

Considerando o numerador de $h'_p(y)$, podemos obter a equação

$$\begin{aligned} f'_p(y)S_p(y) + f_p^2(y) &= (1 - P(M=0))f'(y)[P(M=0) + (1 - P(M=0))S(y)] \\ &+ (1 - P(M=0))^2 f^2(y) \\ &= (1 - P(M=0))f'(y)P(M=0) \\ &+ (1 - P(M=0))^2[f'(y)S(y) + f^2(y)]. \end{aligned} \tag{6.11.2}$$

Comparando a Equação (6.11.2) com a Proposição (5.1.1), concluímos que $h'_p(y) < 0$, pois $f'(y)S(y) + f^2(y) < 0$ e $f'(y) < 0$.

Para a distribuição L-CoLSE, considere $\eta_p(y) = f'_p(y)/f_p(y)$ do Teorema de Glaser (Glaser, 1980). Comparando $\eta_p(y)$ com a Equação (6.4.4), encontramos a relação $\eta_p(y) = \eta(y)$ e portanto a função de risco da distribuição L-CoLSE é decrescente.

Para concluir que $h(\infty) = 0$ é necessário que $\lim_{y\to\infty} f_p(y) = 0$, pois assim $\lim_{y\to\infty} \frac{f_p(y)}{S_p(y)} = 0$, desde que $\lim_{y\to\infty} S_p(y) = P(M = 0)$. Como $f'_p(y) < 0$, pois $f'(y) < 0$ e $\int_0^\infty f(y)dy = 1$, temos que $\lim_{y\to\infty} f_p(y) = 0$.

Para a função de risco reversa basta lembrar da Proposição B11 do livro Marshall e Olkin (2007), a qual indica se a função de risco é estritamente decrescente então a função de risco reversa é estritamente decrescente, o que conclui a prova. ■

6.12 Aplicação em dados reais

Para ilustrar os métodos apresentados no capítulo, considere o conjunto para tumor Aneuploide dos dados citados em Sickle-Santanello et al. (1988), chamados de $T13$, e que podem ser acessados em `http://www.stat.nus.edu.sg/~stachenz/tongue.txt`. Os dados apresentam o tempo de vida desde o diagnóstico até a morte do paciente em semanas. No estudo, 52 pacientes tem um perfil no DNA aneuploide (anomalia) e 28 pacientes um diploide (normal). Como pode ser visto na Figura 6.11, os dados apresentam longa duração e só encontramos um trabalho que faz uso de estimação paramétrica (Jain et al., 2014), porém em seu artigo a distribuição apresentada não possui longa duração.

A Tabela 6.3 apresenta as estimativas das distribuições L-CoPE, L-CoNBE e L-CoLSE para o conjunto de dados $T13$, juntamento com seus desvios padrões em parênteses e os valores $-\ell(.)$ e AIC.

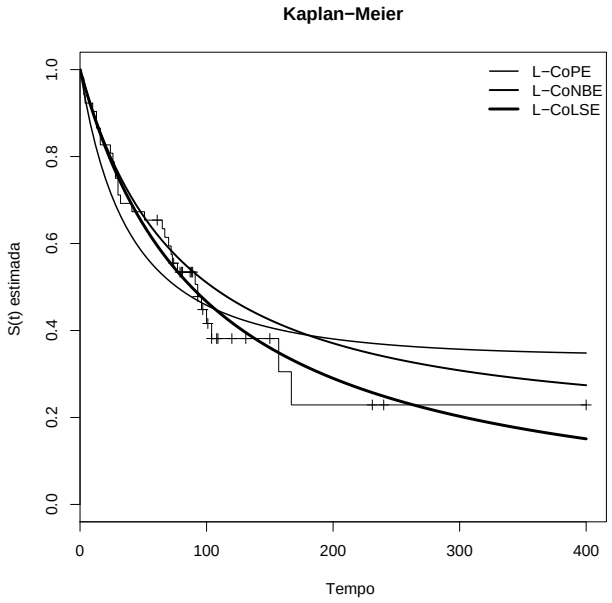

Figura 6.11: K-M e as estimativas dos modelos L-CoPE, L-CoNBE e L-CoLSE para os dados T13.

Tabela 6.3: Parâmetros estimados das distribuições L-CoPE, L-CoNBE e L-CoLSE para o conjunto de dados T13.

		λ	θ	ρ	r (ou γ)	$-\max \ell(.)$	AIC
T13	L-CoPE	0.0087(0.0078*)	1.0721(0.0337)	0.4788(0.0078)	-	197.0928	400.1856
	L-CoNBE	0.0027(0.0008)	0.3684(0.0056)	0.2895(0.0138)	1.5795(0.0518)	182.7052	373.4104
	L-CoLSE	0.0013(0.9040*)	0.9683(0.4555*)	0.8758(0.0010)	0.0714*(0.0001*)	182.1985	372.3970

*$\times 10^{-3}$

Segundo os critérios de $-\ell(.)$ e AIC, o modelo L-CoLSE é o que melhor e ajusta no conjunto de dados $T13$. As estimativas dos desvios padrões foi obtido da matriz Hessiana do algorítimo, ou seja, da matriz de informação de Fisher observada de cada modelo.

6.13 Comentários

Foram apresentados os modelos de longa duração L-CoPE, L-CoNBE e L-CoLSE que são derivados dos modelos CoPE, CoNBE e CoLSE apresentados nos Capítulos 5 e 6.1 como uma extensão direta por meio do modelo unificado apresentado em Rodrigues et al. (2009a). Estes modelos são construídos com base nas distribuições de distribuição IGPS dada em Kolev et al. (2000) e têm a vantagem, em casos como o modelo L-CoPE e L-CoNBE de não acrescentar parâmetros aos modelos de longa duração. As propriedades dos momentos, exceto o modelo L-CoLSE, e as formas da função de risco de cada modelo foram diretamente relacionados com os resultados já encontrados nas Seções 6.1 à 6.6. Também, demonstramos a aplicação em um conjunto de dados reais apontados em Sickle-Santanello et al. (1988), o qual não parece ter nenhum modelo paramétrico na literatura que apresente longa duração para os dados de tumores Aneuploides, e para este conjunto de dados, o modelo L-CoLSE se mostra mais adequado. As estimativas foram obtidas por meio da rotina BFGS implementada no *software* Matlab (MATLAB, 2010), para as distribuições L-CoPE e L-CoNBE e pelo *software* (R Core Team, 2012) para a distribuição L-CoLSE.

6.14 Propostas de pesquisas futuras

Neste trabalho foram apresentados vários modelos para estudo do tempo de vida, sendo dois em particular para estudo em dados de contagem, o qual são baseados na estrutura de riscos competitivos e riscos complementares em que são apontados suas interpretações práticas e de construção sempre que propício. O foco principal era a obtenção de modelos com funções de riscos com várias formas, porém não foi observado em todos os casos os resultados esperados.

A Figura 6.12 apresenta a estrutura para obter as distribuições desenvolvidas neste trabalho (sublinhadas) e a comparação com os principais resultados da literatura. Esta figura tem como objetivo auxiliar a identificar rapidamente qual a estrutura dos riscos adotada bem como o cenário do tema motivacional (carcinogênese) e seus relacionamentos com técnicas já desenvolvidas. A anotação “A serem desenvolvidas” representa o fato de resultados para as

distribuições mencionadas no centro da figura, bem como demais distribuições, não terem sido encontradas na literatura e portanto suas obtenções se encontram em aberto para estudo.

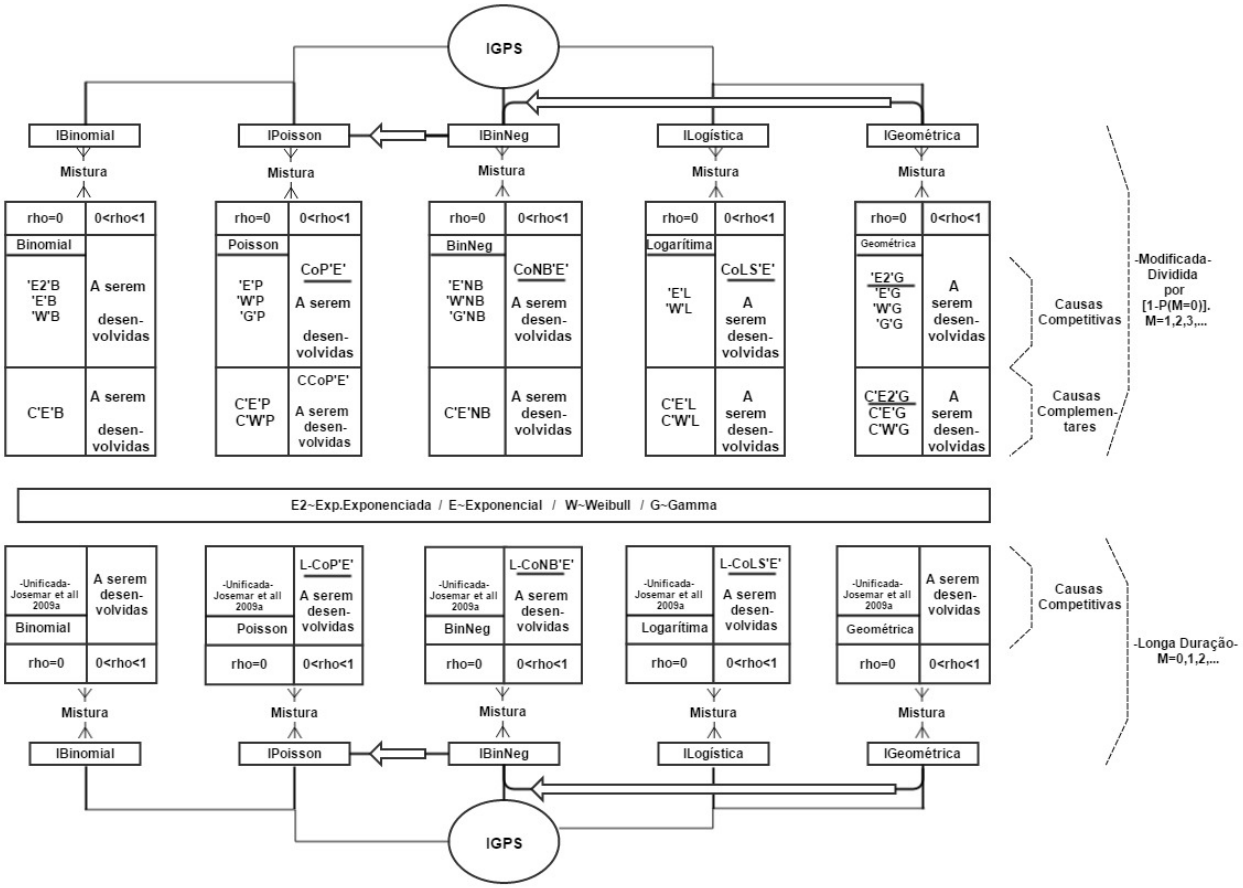

Figura 6.12: Desenho esquemático dos resultados do trabalho em comparação com os principais resultados da literatura.

Além de apresentar as novas distribuições, apresentamos várias características analíticas como os momentos, o momento residual reverso, a função característica, dentre outros. As provas analíticas das formas das funções de risco de cada modelo foram apresentadas, bem como a aplicação em dados reais de cada modelo desenvolvido. Em alguns casos os modelos se mostram boas alternativas aos modelos já consagrados na literatura, em que muitos casos até apresentando melhores resultados, de acordo com os critérios de seleção apresentados. Também, no estudo de carcinogênese as distribuições apresentadas nos Capítulos 5 e 6 possuem interpretação adicional sobre a correlação entre células iniciadas e ativadas.

Como ponto de partida, consideramos as distribuições E2G e CE2G que são construídas no cenário de risco competitivo e complementar, respectivamente, considerando o número de riscos vindos de uma distribuição geométrica com tempo de vida exponencial para cada risco

onde os riscos são independentes. Para tais distribuições não foram avaliadas as performances na presença de covariáveis. Portanto para uma análise mais completa é necessária uma abordagem com covariáveis, visto que foram desenvolvidas para estudos de tempo de vida baseadas na carcinogênese e este campo está repleto de dados com várias leituras de covariáveis para cada paciente.

O objetivo principal foi o desenvolvimento de funções de riscos com formas variadas, no entanto para alguns modelos apresentam a forma decrescente e/ou unimodal. Tais formas parecem ter sido obtidas pois foram utilizadas como função dos tempos de riscos a distribuição Exponencial e portanto o uso de outras distribuições como a Weibull seria uma escolha que poderia flexibilizar as formas das funções de risco.

Para as distribuições obtidas considerando o cenário onde os riscos são correlacionados, que foram apresentadas neste capítulo, vários estudos de simulação devem ser ainda desenvolvidos afim de encontrar resultados que corroborem para a teoria assintótica dos estimadores. Ainda, as distribuições obtidas devem ser comparadas com outras distribuições comumente utilizadas para cada caso respectivo. Tais resultados não foram apresentados aqui devido ao alto custo (tempo) computacional para geração de amostras e estimação. Portanto rotinas mais eficientes de geração devem ser implementadas para viabilizar o estudo bem como possíveis melhorias na rotina de maximização apresentada no Anexo E. Também, para estas distribuições é interessante o estudo para dados com covariáveis visto que os resultados apresentados são obtidos somente para dados sem presença de covariáveis.

Apesar dos esforços não foi possível verificar a identificabilidade do modelo apresentado no Capítulo 5 com o procedendo relatado no Capítulo 3, deixando assim mais um ponto para ser pesquisado para o a distribuição CoPE bem como para as demais distribuições que foram apresentadas neste capítulo.

Todo o trabalho é baseado em métodos de estimação clássicos que é a maximização da logverossimilhança e como visto o processo de estimação envolveu o desenvolvimento de uma rotina que ainda não é a ideal. Portanto outros métodos de estimação como o algorítimo E-M, métodos bayseanos e a ortogonalização dos parâmetros devem ser consideradas em estudos posteriores. Dentre estes métodos, na visão do autor, o algorítimo E-M deve resolver o problema na demora da maximização e portanto deve ser o primeiro método a ser implementado.

Referências Bibliográficas

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

Adamidis, K., Dimitrakopoulou, T., Loukas, S., 2005. On an extension of the exponencial-geometric distribution. Statistics Probability Latters, 73, 259–269.

Adamidis, K., Loukas, S., 1998. A lifetime distribution with decreasing failure rate. Statistics Probability Latters, 39, 35–42.

Bakouch, H. S., Al-Zahrani, B. M., Al-Shomrani, A. A., Marchi, V. A., Louzada, F., 2011. An extended lindley distribution. Journal of the Korean Statistical Society, 41, 75–85.

Barakat, H., Abdelkader, Y., 2004. Computing the moments of order statistics from nonidentical random variables. Statistical Methods and Applications, 12 (1), 15–26.

Barreto-Souza, W., Cribari-Neto, F., 2009. A generalization of the exponential-poisson distribution. Statistics and Probability Letters, 79, 2493–2500.

Barreto-Souza, W., de Morais, A. L., Cordeiro, G. M., 2011. The weibull-geometric distribution. Journal of Statistical Computation and Simulation, 81 (5), 645–657.

Barriga, G. D. C., Louzada-Neto, F., Cancho, V. G., 2011. The complementary exponential power lifetime model. Computational Statistics & Data Analysis, 55, 1250–1259.

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

Bidram, H., Behboodian, J., Towhidi, M., 2012. A new generalized exponential geometric distribution. Communications in Statistics - Theory and Methods,, 528–542.

Birnbaum, Z. W., Saunders, S. C., 1969. A new family of life distributions. Journal of Applied Probability, 6, 319–327.

Boag, J., 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society, 11, 15–53.

Borges, P., Rodrigues, J., Balakrishnan, N., 2012. Correlated destructive generalized power series cure rate models and associated inference with an application to a cutaneous melanoma data. Computational Statistics & Data Analysis, 56, 1703–1713.

Cancho, V. G., Louzada-Neto, F., Barriga, G. D. C., 2011. The poisson-exponential lifetime distribution. Computational Statistics & Data Analysis, 55, 677–686.

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, 7, 233–458.

Cobre, J., Castro Perdoná, G. S., Peria, F. M., Louzada, F., 2013. A mechanistic breast cancer survival modelling through the axillary lymph node chain. Statistics in Medicine, 32, 1536–1546.

Cohen, A., 1963. Estimation in mixtures of discrete distributions. Em: In Proceedings of the International Symposium on Discrete Distributions. Lecture Notes in Computer Science. Montreal, Quebec, páginas. 373–378.

Consul, P. C., 1990. New class of location-parameter discrete probability distributions and their characterizations. Communications in Statistics - Theory and Methods, 19, 4653–4666.

Cox, D., Oakes, D., 1984. Analysis of Survival Data. Chapman and Hall, Boca Raton, Florida.

Cox, D. R., 1972. Regression models and life-tables. Journal of the Royal Statistical Society. Series B, 34 (2), 187–220.

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

Dahiya, A., Gurland, J., 1972. Goodness of fit tests for the gamma and exponential distributions. Technometrics, 14, 791–801.

Davis, D., 1952. An analysis of some failure data. Journal of the American Statistical Association, 47 (258), 113–150.

Feller, W., 1968. An Introduction to Probability Theory and its Applications. Vol. I. Wiley, New York.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., Harrington, D. P., 1980. Modified kolmogorov-smirnov test procedures with application to arbitrarily right-censored data. Biometrics, 36 (4), 607–625.

Franco, M. A. P., 1984. A log logistic model for survival time with covariates. Biometrika, 71 (3), 621–623.

Glaser, R. E., 1980. Bathtub and related failure rate characterizations. Journal of the American Statistical Association, 75, 667–672.

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

Gorski, A. C., 1968. Beware of the weibull euphoria. IEEE Transactions, 17, 202–203.

Gradshteyn, I., Ryzhik, I., 2007. Table of Integrals, Series, and Products. Table of Integrals, Series, and Products, Elsevier.

Gupta, R., Kundu, D., 1999. Generalized exponential distributions. Australian and New Zealand Journal of Statistics, 41, 173–188.

Gupta, R. C., 1974. Modified power series distributions and some of its applications. Sankhyã: The Indian Journal of Statistics, Series B, 35, 288–298.

Gupta, R. D., Kundu, D., 2001. Exponentiated exponential family: an alternative to gamma and weibull distributions. Biometrical Journal, 43, 117–130.

Hall, D., 2000. Zero-inflated poisson an binomial regression with random effects: a case study. Biometrics, 56, 1030–1039.

Hoel, D. G., 1972. A representation of mortality data by competing risks. Biometrics, 28 (2), 475–488.

Jain, K., Singla, N., Sharma, S. K., 2014. The generalized inverse generalized weibull distribution and its properties. Journal of Probability, 2014, 11 páginas.

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

Kolev, N., Minkova, L., Neytchev, P., 2000. Inflated-parameter family of generalized power series distributions and their application in analysis of overdispersed insurance data. ARCH Research Clearing House, 2, 295–320.

Kundu, C., Nanda, A. K., 2010. Some reliability properties of the inactivity time. Communications in Statistics: Theory and Methods, 39, 899–911.

Kus, C., 2007. A new lifetime distribution. Computation Statist. Data Analysis, 51, 4497–4509.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., 2005. Applied Linear Statistical Models. McGraw-Hill, New York, NY.

Lai, C., Xie, M., Murthy, N., 2003. A modified weibull distribution. IEEE Transactions on Reliability, 52, 33–37.

Lambert, D., 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, 34 (1), 1–14.

Lawless, J. F., 2003. Statistical Models and Methods for Lifetime Data. Vol. second edition. Wiley, New York, NY.

Leiva, V., Barros, M., Paula, G. A., Galea, M., 2007. Influence diagnostics in log-birnbaum-saunders regression models with censored data. Computational Statistics & Data Analysis, 51 (12), 5694–5707.

Lemonte, A., Barreto-Souza, W., Cordeiro, G. M., 2013. The exponentiated kumaraswamy distribution and its log-transform. Brazilian Journal of Probability and Statistics, 27, 31–53.

Louzada, F., Marchi, V., Roman, M., 2014. The exponentiated exponential-geometric distribution: a distribution with decreasing, increasing and unimodal failure rate. Statistics, 48 (1), 167–181.

Louzada, F., Marchi, V. A. A., 2015. A log-location-scale lifetime distribution with censored data: regression modeling, residuals and global influence. Communications in Statistics - Theory and Methods,, 14 páginas.

Louzada, F., Roman, M., Cancho, V. G., 2011. The complementary exponential geometric distribution: Model, properties and a comparison with its counterpart. Computational Statistics and Data Analysis, 55, 2516–2524.

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

Louzada-Neto, F., Cancho, V. G., Barriga, G. D. C., 2011. The poisson-exponential distribution: a bayesian approach. Journal of Applied Statistics, 38, 1239–1248.

Maller, R., Zhou, X., 1996. Survival Analysis with Long-Term Survivors. John Wiley and Sons Chichester.

MapleSoft, 2012. Computer Algebra System v.13. Waterloo Maple, Waterloo, ON Canada.

Marchi, V., Louzada, F., Carpenter, J., 2013. The complementary exponentiated exponential geometric lifetime distribution. Journal of Probability and Statistics, 2013, 12 páginas.

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.

Marshall, A. W., Olkin, I., 2007. Life Distributions: Structure of Nonparametric, Semiparametric,and Parametric Families. Springer, New York, NY.

MATLAB, 2010. version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts.

Minkova, L. D., 2002. A generalization of the classical discrete distributions. Communications in Statistics - Theory and Methods, 31 (6), 871–888.

Mudholkar, G. S., Srivastava, D. K., Freimer, M., 1995. The exponentiated weibull family: A reanalysis of the bus-motor-failure data. Technometrics, 37, 436–445.

Mullahy, J., 1986. Specification and testing of some modified count data models. Journal of Econometrics, 33 (3), 341–365.

Mwalili, S., Lesaffre, E., Declerck, D., 2008. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Statistical Mehods in Medical Research, 17, 123–139.

Nachlas, J. A., 2005. Reliability Engineering: Probabilistic Models and Maintenance Methods. Taylor & Francys Group, Boca Raton, FL.

Nadarajah, S., Kotz, S., 2006. The exponentiated type distributions. Acta Applicandae Mathematicae, 92, 97–111.

Nanda, A. K., Singh, H., Misra, N., Paul, P., 2003. Reliability properties of reversed residual lifetime. Communications in Statistics, Theory and Methods, 392, 2031–2042.

Nocedal, J., Wright, S. J., 2006. Numerical Optimization, 2nd Edição. Springer, New York, NY.

Oliveira, Giovana ; Ortega, E. M. M., Cancho, V. G., Barreto, L., 2008. Log-Burr XII regression models with censored data. Computational Statistics & Data Analysis, 52, 3820–3842.

Perdona, G. S. C., Louzada-Neto, F., 2010. A general hazard model for lifetime data in the presence of cure rate. Journal of Applied Statistics, 38, 1395–1405.

Prentice, R. L., 1973. Exponential survivals with censuring and explanatory variable. Biometrika, 60, 279–288.

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

Pundir, S., Arora, S.and Jain, K., 2005. Bonferroni curve and the related statistical inference. Statistics and Probability Letters, 75, 140–150.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rényi, A., 1961. On Measures of Entropy and Information. Vol. 1. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Contributions to the Theory of Statistics, Berkeley, Calif.

Ridout, M., Hinde, J., Demétrio, C. G. B., 2001. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics, 57 (1), 219–223.

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

Rodrigues, J., de Castro, M., Cancho, V. G., Balakrishnan, N., 2009b. Com-poisson cure rate survival models and an application to a cutaneous melanoma data. Journal of Statistical Planning and Inference, 139 (10), 3605–3611.

Santana, T. V., Ortega, E. M., Cordeiro, G. M., Silva, G., 2012. The kumaraswamy-log-logistic distribution. Journal of Statistical Theory and Applications, 11, 265–292.

SEI, 2004. Superitendência de estudos econômicos e sociais da Bahia (banco de dados). www.sei.ba.gov.br, acessado em dezembro 2009.

Sickle-Santanello, B. J., Farrar, W. B., DeCenzo, J. F., Keyhani-Rofagha, S., Klein, J., Pearl, D., Laufman, H., O'Toole, R. V., 1988. Technical and Statistical Improvements for Flow Cytometric DNA Analysis of Paraffin-Embedded Tissue. Cytometry, 9, 594–599.

Smith, R. M., Bain, L. J., 1975. An exponential power life-testing distribution. Communications in Statistics, 4, 469–481.

Smith, S. P., Hammond, K., 1988. Rank regression with log gamma residuals. Biometrika, 75 (4), pp. 741–751.

Teicher, H., 1963. Identifiability of finite mixtures. The Annals of Mathematical Statistics, 32, 244–248.

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57 (2), 307–333.

Anexo A.

Neste anexo mostramos os valores dos elementos da matriz de informação de Fisher observado para a distribuição E2G. A matriz se refere a Equação (2.2.14). Desta forma, a partir da Equação (2.3.19), obtemos

$$
\begin{aligned}
I_{\alpha\alpha} &= \sum_{i=1}^{n}\left(\frac{c_i}{\alpha^2}+\frac{(1-c_i)L_i^{\alpha}\ln^2(L_i)}{R_i}+\frac{(1-c_i)L_i^{2\alpha}\ln^2(L_i)}{R_i^2}+\frac{(1+c_i)(1-\theta)L_i^{\alpha}\ln^2(L_i)}{T_i}-\frac{(1+c_i)(1-\theta)^2L_i^{2\alpha}\ln^2(L_i)}{T_i^2}\right),\\
I_{\alpha\theta} &= I_{\theta\alpha}=\sum_{i=1}^{n}\left(-\frac{(1+c_i)L_i^{\alpha}\ln(L_i)}{T_i}-\frac{(1+c_i)(1-\theta)L_i^{\alpha}\ln(L_i)R_i}{T_i^2}\right),\\
I_{\alpha\lambda} &= I_{\lambda\alpha}=\sum_{i=1}^{n}\left(-\frac{c_iX_i}{L_i}+\frac{\alpha(1-c_i)L_i^{\alpha}\ln(L_i)X_i}{L_iR_i}+\frac{(1-c_i)L_i^{\alpha}X_i}{L_iR_i}+\frac{\alpha(1-c_i)L_i^{2\alpha}\ln(L_i)X_i}{L_iR_i^2}\right.\\
&+ \left.\frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}\ln(L_i)X_i}{L_iT_i}-\frac{(1+c_i)(1-\theta)L_i^{\alpha}X_i}{L_iT_i}+\frac{\alpha(1+c_i)(1-\theta)^2L_i^{2\alpha}\ln(L_i)X_i}{L_iT_i^2}\right),\\
I_{\theta\theta} &= \frac{n}{\theta^2}-\sum_{i=1}^{n}\left(\frac{(1+c_i)R_i^2}{T_i^2}\right),\\
I_{\theta\lambda} &= I_{\lambda\theta}=\sum_{i=1}^{n}\left(-\frac{\alpha(1+c_i)L_i^{\alpha}X_i}{L_iT_i}-\frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}R_iX_i}{L_iT_i^2}\right),\\
I_{\lambda\lambda} &= \sum_{i=1}^{n}\left(\frac{c_i}{\lambda^2}+\frac{(\alpha-1)c_iy_iX_i}{L_i}+\frac{(\alpha-1)c_iX_i^2}{L_i^2}-\frac{\alpha(1-c_i)L_i^{\alpha}y_iX_i}{L_iR_i}-\frac{\alpha(1-c_i)L_i^{\alpha}X_i^2(1-\alpha)}{L_i^2R_i}\right.\\
&+ \left.\frac{\alpha^2(1-c_i)L_i^{2\alpha}X_i^2}{L_i^2R_i^2}-\frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}y_iX_i}{L_iT_i}+\frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}X_i^2(1-\alpha)}{L_i^2T_i}-\frac{\alpha^2(1+c_i)(1-\theta)^2L_i^{2\alpha}X_i^2}{L_i^2T_i^2}\right),
\end{aligned}
$$

em que $L_i = 1-e^{-\lambda y_i}$, $R_i = 1-L_i^{\alpha}$, $T_i = 1-(1-\theta)R_i$ e $X_i = y_ie^{-\lambda y_i}$.

Anexo B.

Neste anexo mostramos os valores dos elementos da matriz de informação de Fisher observado para a distribuição CE2G. A matriz se refere a Equação (2.3.20). Desta forma, a partir da Equação (2.3.19), obtemos

$$
\begin{aligned}
I_{\alpha\alpha} &= \sum_{i=1}^{n}\left(c_i/\alpha^2 + \frac{(1-c_i)L_i^{\alpha}\ln^2(L_i)}{R_i} + \frac{(1-c_i)L_i^{2\alpha}\ln^2(L_i)}{R_i^2} - \frac{(1+c_i)(1-\theta)L_i^{\alpha}\ln^2(L_i)}{T_i} - \frac{(1+c_i)(1-\theta)^2L_i^{2\alpha}\ln^2(L_i)}{T_i^2}\right),\\
I_{\alpha\theta} &= I_{\theta\alpha} = \sum_{i=1}^{n}\left(\frac{(1+c_i)L_i^{\alpha}\ln(L_i)}{T_i} + \frac{(1+c_i)(1-\theta)L_i^{2\alpha}\ln(L_i)}{T_i^2}\right),\\
I_{\alpha\lambda} &= I_{\lambda\alpha} = \sum_{i=1}^{n}\left(\frac{-c_iX_i}{L_i} + \frac{\alpha(1-c_i)L_i^{\alpha}\ln(L_i)X_i}{L_iR_i} + \frac{(1-c_i)L_i^{\alpha}X_i}{L_iR_i} + \frac{\alpha(1-c_i)L_i^{2\alpha}\ln(L_i)X_i}{L_iR_i^2}\right.\\
&- \left.\frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}\ln(L_i)X_i}{L_iT_i} - \frac{(1+c_i)(1-\theta)L_i^{\alpha}X_i}{L_iT_i} - \frac{\alpha(1+c_i)(1-\theta)^2L_i^{2\alpha}\ln(L_i)X_i}{L_iT_i^2}\right),\\
I_{\theta\theta} &= \sum_{i=1}^{n}\left(c_i/\theta^2 - \frac{(1+c_i)L_i^{2\alpha}}{T_i^2}\right),\\
I_{\theta\lambda} &= I_{\lambda\theta} = \sum_{i=1}^{n}\left(\frac{\alpha(1+c_i)L_i^{\alpha}X_i}{L_iT_i} + \frac{\alpha(1+c_i)(1-\theta)L_i^{2\alpha}X_i}{L_iT_i^2}\right),\\
I_{\lambda\lambda} &= \sum_{i=1}^{n}\left(\frac{c_i}{\lambda^2} + \frac{(\alpha-1)c_iy_iX_i}{L_i} + \frac{(\alpha-1)c_iX_i^2}{L_i^2} - \frac{\alpha(1-c_i)L_i^{\alpha}y_iX_i}{L_iR_i} - \frac{\alpha(1-c_i)L_i^{\alpha}X_i^2(1-\alpha)}{L_i^2R_i}\right.\\
&+ \left.\frac{\alpha^2(1-c_i)L_i^{2\alpha}X_i^2}{L_i^2R_i^2} + \frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}y_iX_i}{L_iT_i} + \frac{\alpha(1+c_i)(1-\theta)L_i^{\alpha}X_i^2(1-\alpha)}{L_i^2T_i} - \frac{\alpha^2(1+c_i)(1-\theta)^2L_i^{2\alpha}X_i^2}{L_i^2T_i^2}\right),
\end{aligned}
$$

em que $L_i = 1 - e^{-\lambda y_i}$, $R_i = 1 - L_i^{\alpha}$, $T_i = 1 - (1-\theta)L_i^{\alpha}$ e $X_i = y_ie^{-\lambda y_i}$.

Anexo C.

A matriz de observação de Fisher **i** correspondente ao modelo IPP com logverossimilhança $\ell(\beta, \gamma)$ da SeÃ§Ã£o 4.4 é

$$\mathbf{i} = \left[\begin{array}{cc} \frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_j \partial \beta_k} & \frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma_j \partial \beta_k} \\ . & \frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma_j \partial \gamma_k} \end{array} \right] \Bigg|_{(\beta,\gamma)=(\hat{\beta},\hat{\gamma})},$$

em que

$$\begin{aligned}
\frac{\partial^2 \ell(\beta,\gamma)}{\partial \beta_j \partial \beta_k} &= -\sum_{i=1}^{n} x_k^i x_j^i e^{\mathbf{x}_i'\beta} - \sum_{y_i>0} x_k^i x_j^i e^{\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma} + \sum_{y_i>0} P_{kj}^{\beta}; \\
\frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma_j \partial \gamma_k} &= -\sum_{y_i>0} z_k^i z_j^i e^{\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma} + \sum_{y_i>0} y_i z_k^i z_j^i \frac{e^{\mathbf{z}_i'\gamma}}{(1+e^{\mathbf{z}_i'\gamma})^2} + \sum_{y_i>0} P^{\gamma} kj; \\
\frac{\partial^2 \ell(\beta,\gamma)}{\partial \gamma_j \partial \beta_k} &= \sum_{y_i>0} x_k^i z_j^i e^{\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma} + \sum_{y_i>0} P_{kj}^{\beta\gamma};
\end{aligned}$$

onde $\mathbf{x}_i$, $\mathbf{z}_i$ são as observações para o i-ésimo elemento, r_k^i é o k-ésimo elemento de $\mathbf{r}_i$, $z_0^i = x_0^i = 1$ e

$$\begin{aligned}
P_{kj}^{\beta} &= \frac{\partial^2}{\partial \beta_k \partial \beta_j} \log\left({}_1F_1\left[y_i+1; 2; e^{\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma}\right]\right), \\
P_{kj}^{\gamma} &= \frac{\partial^2}{\partial \gamma_k \partial \gamma_j} \log\left({}_1F_1\left[y_i+1; 2; e^{\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma}\right]\right) \text{ e} \\
P_{kj}^{\beta\gamma} &= \frac{\partial^2}{\partial \beta_k \partial \gamma_j} \log\left({}_1F_1\left[y_i+1; 2; e^{\mathbf{x}_i'\beta - \mathbf{z}_i'\gamma}\right]\right)
\end{aligned}$$

são derivadas parciais numéricas. De Nocedal e Wright (2006), podemos aproximar a segunda derivada de $f(x)$ por

$$\frac{\partial^2}{\partial x_k \partial x_j} f(x) = \frac{f(x+\epsilon e_k + \epsilon e_j) - f(x+\epsilon e_k) - f(x+\epsilon e_j) + f(x)}{\epsilon^2} + O(\epsilon),$$

em que e_k é o vetor com 1 na k-ésima posição e 0 caso contrário e, ϵ é o erro.

A matriz de observação de Fisher $\mathbf{i}$ correspondente ao modelo IPNB com logverossimilhança $\ell(\beta,\gamma,\phi)$ da SeÃ§Ã£o 4.4 é

$$\mathbf{i} = \left[\begin{array}{ccc} \frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\beta_j\partial\beta_k} & \frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\gamma_j\partial\beta_k} & \frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\phi\partial\beta_k} \\ . & \frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\gamma_j\partial\gamma_k} & \frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\phi\partial\gamma_k} \\ . & . & \frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\phi^2} \end{array}\right]\Bigg|_{(\beta,\gamma,\phi)=(\hat\beta,\hat\gamma,\hat\phi)},$$

em que

$$\begin{aligned}
\frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\beta_j\partial\beta_k} &= -\sum_{y_i=0} x_k^i x_j^i e^{\mathbf{x}_i'\beta}\phi^{-1}\log(1+\phi) - \sum_{y_i>0} x_k^i x_j^i e^{\mathbf{x}_i'\beta}\mathbf{z}_i'\gamma\phi^{-1} \\
&- \sum_{y_i>0} x_k^i x_j^i e^{\mathbf{x}_i'\beta}\phi^{-1}\log((1+\phi)e^{\mathbf{z}_i'\gamma}+\phi) + \sum_{y_i>0} B_{kj}^{\beta}; \\
\frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\gamma_j\partial\beta_k} &= \sum_{y_i>0} x_k^i z_j^i \phi^{-1} e^{\mathbf{x}_i'\beta} - \sum_{y_i>0} \frac{x_k^i z_j^i \phi^{-1}(1+\phi)e^{\mathbf{x}_i'\beta}}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi} + \sum_{y_i>0} B_{jk}^{\gamma\beta}; \\
\frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\phi\partial\beta_k} &= \sum_{y_i=0} \frac{(1+\phi)\log(1+\phi)-\phi}{(1+\phi)\phi^2} x_k^i e^{\mathbf{x}_i'\beta} - \sum_{y_i>0} x_k^i \phi^{-2}\mathbf{z}_i'\gamma e^{\mathbf{x}_i'\beta} \\
&- \sum_{y_i>0} \frac{x_k^i e^{\mathbf{x}_i'\beta}\phi^{-1}(1+e^{\mathbf{z}_i'\gamma})}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi} + \sum_{y_i>0} x_k^i e^{\mathbf{x}_i'\beta}\phi^{-2}\log((1+\phi)e^{\mathbf{z}_i'\gamma}+\phi) + \sum_{y_i>0} B_k^{\beta\phi}; \\
\frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\gamma_j\partial\gamma_k} &= -\sum_{y_i>0} \frac{y_i z_k^i z_j^i e^{\mathbf{z}_i'\gamma}}{(1+e^{\mathbf{z}_i'\gamma})^2} - \sum_{y_i>0} \frac{z_k^i z_j^i (1+\phi^{-1}e^{\mathbf{x}_i'\beta})\phi(1+\phi)e^{\mathbf{z}_i'\gamma}}{[(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi]^2} + \sum_{y_i>0} B_{kj}^{\gamma}; \\
\frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\phi\partial\gamma_k} &= -\sum_{y_i>0} z_k^i \phi^{-2} e^{\mathbf{x}_i'\beta} - \sum_{y_i>0} \frac{z_k^i e^{\mathbf{z}_i'\gamma}(1+\phi^{-1}e^{\mathbf{x}_i'\beta})}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi} + \sum_{y_i>0} \frac{z_k^i \phi^{-2}(1+\phi)e^{\mathbf{z}_i'\gamma+e^{\mathbf{x}_i'\beta}}}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi} \\
&+ \sum_{y_i>0} \frac{z_k^i (1+\phi)e^{\mathbf{z}_i'\gamma}(1+e^{\mathbf{z}_i'\gamma})(1+\phi^{-1}e^{\mathbf{x}_i'\beta})}{[(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi]^2} + \sum_{y_i>0} B_k^{\gamma\phi}; \\
\frac{\partial^2\ell(\beta,\gamma,\phi)}{\partial\phi^2} &= -\sum_{y_i=0} \frac{(2\phi^2+4\phi+2)\log(1+\phi)-3\phi^2-2\phi}{(1+\phi)^2\phi^3} e^{\mathbf{x}_i'\beta} + \sum_{y_i>0} 2\phi^{-3}e^{\mathbf{x}_i'\beta}\mathbf{z}_i'\gamma \\
&+ \sum_{y_i>0} 2e^{\mathbf{x}_i'\beta}\phi^{-3}\log((1+\phi)e^{\mathbf{z}_i'\gamma}-\phi) - \sum_{y_i>0} \frac{2e^{\mathbf{x}_i'\beta}(e^{\mathbf{z}_i'\gamma}-1)}{\phi^2[(1+\phi)e^{\mathbf{z}_i'\gamma}-\phi]} \\
&- \sum_{y_i>0} \frac{e^{\mathbf{x}_i'\beta}(e^{\mathbf{z}_i'\gamma}-1)^2}{\phi[(1+\phi)e^{\mathbf{z}_i'\gamma}-\phi]} + \sum_{y_i>0} B^{\phi},
\end{aligned}$$

onde $\mathbf{x}_i$, $\mathbf{z}_i$ são as observações para o i-ésimo elemento, r_k^i é o k-ésimo elemento de $\mathbf{r}_i$, $z_0^i = x_0^i = 1$

e

$$
\begin{aligned}
B_{kj}^{\beta} &= \frac{\partial^2}{\partial\beta_k\partial\beta_j}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right);\\
B_{kj}^{\beta\gamma} &= \frac{\partial^2}{\partial\beta_k\partial\gamma_j}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right);\\
B_{k}^{\beta\phi} &= \frac{\partial^2}{\partial\phi\partial\beta_k}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right);\\
B_{kj}^{\gamma} &= \frac{\partial^2}{\partial\gamma_k\partial\gamma_j}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right);\\
B_{k}^{\gamma\phi} &= \frac{\partial^2}{\partial\phi\partial\gamma_k}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right);\\
B^{\phi} &= \frac{\partial^2}{\partial\phi^2}\log\left({}_2F_1\left[y_i+1;1+\phi^{-1}e^{\mathbf{x}_i'\beta};2;\frac{\phi}{(1+\phi)e^{\mathbf{z}_i'\gamma}+\phi}\right]\right)
\end{aligned}
$$

são as derivadas parciais numéricas.

Anexo D.

Os valores dos elementos da matriz de informação do modelo CoPE na Equação (5.2.1) com presença de censura de acordo com a equação (5.2.1) são

$$
\begin{aligned}
I_{\lambda\lambda} &= -\frac{\sum_{i=1}^{n} c_i}{\lambda^2} + \sum_{i=1}^{n}\left[c_i \frac{\left(\sum_{m=1}^{\infty} m^3 x_i^2 L_i\right)\left(\sum_{m=1}^{\infty} mL_i\right) - \left(\sum_{m=1}^{\infty} m^2 x_i L_i\right)^2}{\left(\sum_{m=1}^{\infty} mL_i\right)^2}\right.\\
&+ \left.(1-c_i)\frac{\left(\sum_{m=1}^{\infty} m^2 x_i^2 L_i\right)\left(\sum_{m=1}^{\infty} L_i\right) - \left(\sum_{m=1}^{\infty} mx_i L_i\right)^2}{\left(\sum_{m=1}^{\infty} L_i\right)^2}\right],\\
I_{\lambda\theta} &= I_{\theta\lambda} = -\sum_{i=1}^{n}\frac{(1-\rho)}{2\rho}\left[c_i \frac{\left(\sum_{m=1}^{\infty}(m+1)m^2 x_i U_i\right)\left(\sum_{m=1}^{\infty} mL_i\right) - \left(\sum_{m=1}^{\infty}(m+1)mU_i\right)\left(\sum_{m=1}^{\infty} m^2 x_i L_i\right)}{\left(\sum_{m=1}^{\infty} mL_i\right)^2}\right.\\
&+ \left.(1-c_i)\frac{\left(\sum_{m=1}^{\infty}(m+1)mx_i U_i\right)\left(\sum_{m=1}^{\infty} L_i\right) - \left(\sum_{m=1}^{\infty}(m+1)U_i\right)\left(\sum_{m=1}^{\infty} mx_i L_i\right)}{\left(\sum_{m=1}^{\infty} L_i\right)^2}\right],\\
I_{\lambda\rho} &= I_{\rho\lambda} = -\sum_{i=1}^{n} c_i \frac{\left(\sum_{m=1}^{\infty}\frac{m^3 L_i}{\rho} - \sum_{m=1}^{\infty}\frac{\theta}{2\rho^2}(m+1)m^2 x_i U_i\right)\sum_{m=1}^{\infty} mL_i - \left(\sum_{m=1}^{\infty}\frac{m^2 L_i}{\rho} + \sum_{m=1}^{\infty}\frac{\theta}{2\rho^2}(m+1)mU_i\right)\sum_{m=1}^{\infty} m^2 x_i L_i}{\left(\sum_{m=1}^{\infty} mL_i\right)^2}\\
&- \sum_{i=1}^{n}(1-c_i)\frac{\left(\sum_{m=1}^{\infty}\frac{m^2 L_i}{\rho} - \sum_{m=1}^{\infty}\frac{\theta}{2\rho^2}(m+1)mx_i U_i\right)\sum_{m=1}^{\infty} L_i + \left(\sum_{m=1}^{\infty}\frac{mL_i}{\rho} - \sum_{m=1}^{\infty}\frac{\theta}{2\rho^2}(m+1)U_i\right)\sum_{m=1}^{\infty} mx_i L_i}{\left(\sum_{m=1}^{\infty} L_i\right)^2},
\end{aligned}
$$

$$
\begin{aligned}
I_{\theta\rho} &= I_{\rho\theta} = -\frac{1}{2\rho^2}\left\{\sum_{i=1}^{n}\left[c_i\frac{\sum_{m=1}^{\infty}(m+1)mU_i}{\sum_{m=1}^{\infty} mL_i} + (1-c_i)\frac{\sum_{m=1}^{\infty}(m+1)U_i}{\sum_{m=1}^{\infty} L_i}\right]\right\} + \frac{n}{\rho^2}\\
&+ \frac{(1-\rho)}{2\rho}\left\{\sum_{i=1}^{n} c_i\left[\frac{\left(\sum_{m=1}^{\infty}\frac{(m+1)m^2}{\rho}U_i - \frac{\theta}{3\rho^2}\sum_{m=1}^{\infty} m(m+1)(m+2)P_i\right)\left(\sum_{m=1}^{\infty} mL_i\right)}{\left(\sum_{m=1}^{\infty} mL_i\right)^2}\right.\right.\\
&- \left.\frac{\left(\sum_{m=1}^{\infty}\frac{m^2}{\rho}L_i - \frac{\theta}{2\rho^2}\sum_{m=1}^{\infty} m(m+1)U_i\right)\left(\sum_{m=1}^{\infty} m(m+1)U_i\right)}{\left(\sum_{m=1}^{\infty} mL_i\right)^2}\right]\\
&+ \left.\sum_{i=1}^{n}\left[(1-c_i)\frac{\left(\sum_{m=1}^{\infty}\frac{(m+1)m}{\rho}U_i - \frac{\theta}{3\rho^2}\sum_{m=1}^{\infty}(m+1)(m+2)P_i\right)\left(\sum_{m=1}^{\infty} L_i\right) - \left(\sum_{m=1}^{\infty}\frac{m}{\rho}L_i - \frac{\theta}{2\rho^2}\sum_{m=1}^{\infty}(m+1)U_i\right)\left(\sum_{m=1}^{\infty}(m+1)U_i\right)}{\left(\sum_{m=1}^{\infty} L_i\right)^2}\right]\right\},\\
I_{\theta\theta} &= \frac{n}{\theta^2}\left(\frac{e^{-\theta}(2+\theta^2) - e^{-2\theta} - 1}{(1-e^{-\theta})^2}\right) + \left\{\sum_{i=1}^{n} c_i\frac{\frac{(1-\rho)^2}{6\rho^2}\left(\sum_{m=1}^{\infty} m(m+1)(m+2)P_i\right)\left(\sum_{m=1}^{\infty} mL_i\right) - \left(\sum_{m=1}^{\infty}\frac{(1-\rho)m(m+1)U_i}{\rho}\right)^2}{\left(\sum_{m=1}^{\infty} mL_i\right)^2}\right.\\
&+ \left.\sum_{i=1}^{n}(1-c_i)\frac{\left(\sum_{m=1}^{\infty}\frac{(1-\rho)^2(m+1)(m+2)P_i}{6\rho^2}\right)\left(\sum_{m=1}^{\infty} L_i\right) - \left(\sum_{m=1}^{\infty}\frac{(1-rho)m+1U_i}{rho}\right)^2}{\left(\sum_{m=1}^{\infty} L_i\right)^2}\right\},
\end{aligned}
$$

$$
\begin{aligned}
I_{\rho\rho} &= n\left(\frac{\rho + 4\rho\theta - 2\theta - 2\rho^2\theta - 2\rho^2}{\rho^3(1-\rho)^2}\right)\\
&- \sum_{i=1}^{n} c_i\left\{\left[-\frac{\theta}{\rho^3}\sum_{m=1}^{\infty} m(m+1)U_i + \frac{\theta}{2\rho^2}\left(\sum_{m=1}^{\infty}\frac{m^2(m+1)}{\rho}U_i - \frac{\theta}{3\rho^2}\sum_{m=1}^{\infty} m(m+1)(m+2)P_i\right) + \sum_{m=1}^{\infty}\frac{m^2}{\rho^2}L_i - \sum_{m=1}^{\infty}\frac{m^3}{\rho^2}L_i\right]\left[\sum_{m=1}^{\infty} mL_i\right]\right.\\
&+ \left.\left[\sum_{m=1}^{\infty}\frac{m^2}{\rho}L_i - \frac{\theta}{2\rho^2}\sum_{m=1}^{\infty} m(m+1)U_i\right]^2\right\}/\left(\sum_{m=1}^{\infty} mL_i\right)^2\\
&- \sum_{i=1}^{n}(1-c_i)\left\{\left[-\frac{\theta}{\rho^3}\sum_{m=1}^{\infty}(m+1)U_i + \frac{\theta}{2\rho^2}\left(\sum_{m=1}^{\infty}\frac{m(m+1)}{\rho}U_i - \frac{\theta}{3\rho^2}\sum_{m=1}^{\infty}(m+1)(m+2)P_i\right) + \sum_{m=1}^{\infty}\frac{m}{\rho^2}L_i - \sum_{m=1}^{\infty}\frac{m}{\rho^2}L_i\right]\left[\sum_{m=1}^{\infty} L_i\right]\right.\\
&+ \left.\left[\sum_{m=1}^{\infty}\frac{m}{\rho}L_i - \frac{\theta}{2\rho^2}\sum_{m=1}^{\infty}(m+1)U_i\right]^2\right\}/\left(\sum_{m=1}^{\infty} L_i\right)^2
\end{aligned}
$$

em que $L_i = \rho^m e^{-\lambda x_i m} H([m+1],[2],[\theta(1-\rho)/\rho])$, $U_i = \rho^m e^{-\lambda x_i m} H([m+2],[3],[\theta(1-\rho)/\rho])$ e $P_i = \rho^m e^{-\lambda x_i m} H([m+3],[4],[\theta(1-\rho)/\rho])$.

Anexo E.

Neste anexo será apresentada todas as funções e rotinas para a estimação da distribuição CoPE apresentada no Capítulo 5. As demais distribuições que usam as rotinas abaixo são obtidas somente se alterando as funções de sobrevivência e de densidade. Esta rotina foi necessária para o controle de parâmetros no processo BFGS que os softwares não apresentam por padrão. O método implementado segue as ideias apresentadas no livro de Nocedal e Wright (2006).

O processo BFGS é um método de estimação Quasi-Newton, ou seja, ele é baseado no processo de minimização de Newton com alteração no tamanho do passo e na estimação da matriz de derivadas. Entenda-se por passo o deslocamento entre o ponto de busca na iteração k para a iteração $k+1$. No método de Newton o tamanho do passo é 1, pois

$$\theta_{k+1} = \theta_k - \underbrace{1}_{passo} \cdot B \cdot grad,$$

em que $grad = \frac{\partial \ell(\theta)}{\partial \theta}|_{\theta_k}$, $B^{-1} = \frac{\partial^2 \ell(\theta)}{\partial \theta^2}|_{\theta_k}$.

Na rotina implementada o passo é reduzido até que se tenha atingido a condição

$$-\ell(\theta_{k+1}) > -\ell(\theta_k) + c \cdot a \cdot grad' \cdot (-B \cdot grad), \tag{6.14.1}$$

em que $grad = \frac{\partial \ell(\theta)}{\partial \theta}|_{\theta_k}$, $B^{-1} = \frac{\partial^2 \ell(\theta)}{\partial \theta^2}|_{\theta_k}$, e c e a são parâmetros a serem especificados por conveniência com $0 < a, c < 1$. Caso a condição não seja atingida o passo é diminuído multiplicando a por um valor $0 < \rho < 1$ a ser também definido por conveniência. O passo neste método é então o resultado a que satisfez a condição (6.14.1) e assim

$$\theta_{k+1} = \theta_k - \underbrace{a}_{passo} \cdot B \cdot grad.$$

Abaixo encontram-se as funções e as rotinas utilizadas no processo de minimização de $-\ell(\theta)$ do capítulo 5 para o *software* Matlab MATLAB (2010).

Função f.d.p.

```
function [res] = denslcope(x,c,beta)
erro=1e-10;
res=[];
hy=load('hy.m');
kk=length(hy(:,1));
for j=1:length(x(1,:))
soma=0;
parc=1;
```

```
m=1;
if(c(1,j)==1)
while(parc>erro&&m<=kk)
parc=exp(log(m)+m*-beta(1,1)*x(1,j)+ (m-1)*log(beta(3,1))+ log(hy(m,1)));
soma=soma+parc;
m=m+1;
end
if(m>kk&&parc>erro)
while(parc>erro&&m>kk)
hy(m,1)=hypergeom([m+1],[2],beta(1,1)*(1-beta(3,1))/beta(3,1));
parc=exp(log(m)+m*-beta(1,1)*x(1,j)+ (m)*log(beta(3,1))+ log(hy(m,1)));
soma=soma+parc;
m=m+1;
end
end
res(j,1)=real(beta(1,1)*exp(-beta(2,1)/beta(3,1))*(beta(2,1))*(1-beta(3,1))/beta(3,1)*soma);
kk=length(hy(:,1));
else
res(j,1)=1;
end
end
dlmwrite('hy.m',hy)
```

Função de Sobrevivência

```
function [res] = scope(x,c,beta)
erro=1e-10;
res=[];
hy=load('hy.m');
kk=length(hy(:,1));
for j=1:length(x(1,:))
soma=0;
parc=1;
m=1;
if(c(1,j)==0)
while(parc>erro&&m<=kk)
parc=exp(m*(-beta(1,1)*x(1,j))+ m*log(beta(3,1))+ log(hy(m,1)));
soma=soma+parc;
m=m+1;
end
if(m>kk&&parc>erro)
while(parc>erro&&m>kk)
hy(m,1)=hypergeom([m+1],[2],beta(2,1)*(1-beta(3,1))/beta(3,1));
parc=exp(m*-beta(1,1)*x(1,j)+ m*log(beta(3,1))+ log(hy(m,1)));
soma=soma+parc;
m=m+1;
end
end
res(j,1)=real(1-exp(-beta(2,1)/beta(3,1))*beta(2,1)*(1-beta(3,1))/(beta(3,1)*(1-exp(-beta(2,1))))*soma);
kk=length(hy(:,1));
else
res(j,1)=1;
end
end
dlmwrite('hy.m',hy)
```

Função de log-verossimilhança

```
function [res01] = verocope(beta)
x=load('xc.m');
c=x(2,:);
x=x(1,:);
n=length(x(1,:));
par(1,1)=exp(beta(1,1));
par(2,1)=exp(beta(2,1));
par(3,1)=exp(beta(3,1))/(1+exp(beta(3,1)));
hy=[];
```

```
hy(1,1)=hypergeom([1+1],[2],par(2,1)*(1-par(3,1))/par(3,1));
dlmwrite('hy.m',hy)
vero=sum(c*log(denscope(x,c,[par(1,1);par(2,1);par(3,1)]))+sum((1-c)*log(scope(x,c,[par(1,1);par(2,1);par(3,1)])));
res01=real(-vero);
```

Função gradiente

```
function [res] = gbetacope(pos,beta)
beta(1,1)=exp(beta(1,1));
beta(2,1)=exp(beta(2,1));
beta(3,1)=exp(beta(3,1))/(1+exp(beta(3,1)));
n=length(beta);
h=1e-7;
if(pos==1)
vero1=verocope([beta(1,1)+h;beta(2:n,1)]);
vero2=verocope([beta(1,1)-h;beta(2:n,1)]);
end
if(pos==n)
vero1=verocope([beta(1:(n-1),1);beta(n,1)+h]);
vero2=verocope([beta(1:(n-1),1);beta(n,1)-h]);
end
if(pos<n)&&(pos>1)
vero1=verocope([beta(1:(pos-1),1);beta(pos,1)+h;beta((pos+1):n,1)]);
vero2=verocope([beta(1:(pos-1),1);beta(pos,1)-h;beta((pos+1):n,1)]);
end
res=real((vero1-vero2)/(2*h));
```

Rotina BFGS

Obs: Esta rotina é chamada pelo nome no software Matlab, no caso pelo comando ">BFGScope".

```
beta=[0;0;0];

format long
k=length(beta);
grad=[];
for j=1:(k)
    grad(j,1)=gbetacope(j,beta);
end

B=1.e-5*eye(k);
errop=1e-8;
cont=1;
maxite=200;
parada=sqrt(sum(grad.^2));
if (parada<errop)
    pare=1;
else pare=0;
end

verobeta=verocope(beta)
while (pare==0) && (cont<maxite)

  a=0.1;    %parâmetros para controle
  rhoo=0.3; %valor rho a ser mulpitplicado o passo
  cc=0.0001;%valor c do método
  resulta=beta+a*(-B*grad);
  cont2=1;

verores=verocope(resulta)
if isnan(verores)
 verores=1.1*verocope(beta);
 beta=betaa
```

```
end
  while verores>verobeta+cc*a*grad'*(-B*grad) &&  cont2<100% && verores<verobeta+(1-cc)*a*grad'*(-B*grad)  % em alguns casos é necessária a condição
      a=a*rhoo;
      resulta=beta+a*(-B*grad);
      cont2=cont2+1
 verores=verocope(resulta);
 end
   betap=resulta;
  Dx=a*(-B*grad);
grada=[];
for j=1:(k)
    grada(j,1)=gbetacope(j,betap);
end
y=grada-grad;
if (sum(abs(y))>0)
Bp=(eye(k)-(y*Dx')/((y'*Dx)))'*B*(eye(k)-(y*Dx')/((y'*Dx)))+(Dx*Dx')/((y'*Dx));
parada=sqrt(sum(grad.^2));
if ((parada<errop) || (max(abs(beta-betap)./abs(betap))<1.e-50)|| (cont>15 && (verores/verobeta>2)) || cont>=maxite || abs(verores-verobeta)/verobeta<1.e-50)
    pare=1;
end
betaa=beta;
beta=betap(1:k);
grad=grada;
B=Bp;

cont=cont+1;
guardar=verobeta;
elseif (1)
        pare=1;
s = warning('off','all'); % turn all warnings off
if true,
    disp('O algoritimo parcial estabilizou, ou encerrou')
    warning('Here is a warning that doesn''t have an id.')
end
warning(s)
end
%if(mod(cont,10)==0)
    verobeta=verores
    cont
    beta
    [exp(beta(1,1)),exp(beta(2,1)),exp(beta(3,1))/(1+exp(beta(3,1)))]
    [[datestr(now,6),'-'],datestr(now,13)]
%end
end
```