Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

SIMONE DEMEIS BRAGUIM

Modelo Hierárquico Log-Logístico: Uma Aplicação no estudo da Linezolida

Maringá - Paraná
Junho de 2015

SIMONE DEMEIS BRAGUIM

Modelo Hierárquico Log-Logístico: Uma Aplicação no estudo da Linezolida

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística como requisito para obtenção do título de Mestre em Bioestatística.
Orientador: Prof. Dr. Carlos Aparecido dos Santos

Maringá - Paraná
Junho de 2015

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Central - UEM, Maringá, PR, Brasil)

B813m

Braguim, Simone Demeis
Modelo hierárquico Log-logístico : uma aplicação no estudo da Linezolida / Simone Demeis Braguim. -- Maringá, 2015.
54 f. : il., figs., tabs.

Orientador: Prof. Dr. Carlos Aparecido dos Santos.
Dissertação (mestrado) - Universidade Estadual de Maringá, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2015.

1. Análise de sobrevivência (Biometria). 2. Análise bayesiana. 3. Modelo hierárquico Log-logístico. I. Santos, Carlos Aparecido dos, orient. II. Universidade Estadual de Maringá. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 21.ed. 519.542

AMMA-003005

SIMONE DEMEIS BRAGUIM

Modelo Hierárquico Log-Logístico: Uma Aplicação no estudo da Linezolida

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística como requisito para obtenção do título de Mestre em Bioestatística.
Orientador: Prof. Dr. Carlos Aparecido dos Santos

Trabalho aprovado.
Maringá - Paraná, 26 de Junho de 2015:

Carlos Aparecido dos Santos
Orientador

Prof. Dr. Vanderly Janeiro
Membro 2

Prof. Dr. Walter Moreira Lima
Membro 3

Maringá - Paraná
Junho de 2015

Dedico este trabalho à minha mãe, Daliria Rodrigues Demeis, por todo apoio e carinho.

Agradecimentos

A Deus, o qual me deu sabedoria e força para superar todas as tribulações durante esta etapa.

Ao meu querido orientador, Professor Dr. Carlos Ap. dos Santos, pelo exemplo de pessoa, dedição, carinho, paciência e por ter me conduzido no mestrado com a calma necessária para me ajudar a transpor os momentos difícies. Minha eterna gratidão.

Aos meus queridos filhos, Rodrigo Demeis Braguim e Maíza Demeis Braguim, por terem sido a minha luz e minha alegria.

As minhas queridas amigas: Daiana Aldrovandre e Elisangela D. Vaz, com as quais percebi que as boas amizades são a família que Deus nos permite escolher.

Aos professores do programa de Pós-Graduação Mestrado em Bioestatística, em especial ao prof. Dr. Vanderly Janeiro, coordenação e secretários, com os quais aprendi que nada na vida conquistamos sozinhos. Sempre precisamos de outras pessoas para alcançar os nossos objetivos.

A todos do Departamento de Estatística, pelos auxílios permanentes.

“ Quem entende faz, quem compreende ensina. ”

(Autor: Desconhecido)

Resumo

Neste trabalho é proposto a utilização do Modelo Log-Logístico no estudo que trata de verificar o tempo de sobrevivência dos pacientes após a utilização do antibiótico Linezolida, assim como os efeitos que são causados por esta droga. O modelo Log-Logístico foi uma alternativa ao modelo Weibull, o qual é bastante utilizado em análise de sobrevivência. Neste contexto, técnicas de inferência bayesiana foram empregadas para estimação dos parâmetros. Os dados foram obtidos de pacientes sob tratamento entre os anos de 2008 a 2010, no Hospital Universitário de Maringá (HUM), sendo a variável resposta o tempo de internamento dos pacientes na UTI (unidade de terapia intensiva). Para estimação dos parâmetros geraram-se três cadeias usando, para isto, métodos MCMC. Além disso, consideraram-se alguns gráficos e testes para análise do ajuste e da convergência destas cadeias.

Palavras-chaves: Análise de sobrevivência, análise Bayesiana, densidade Log-Logístico.

Abstract

This paper proposes Log-Logistic Modelapplication in the study that checks the survival time of patients after the use of the antibiotic linezolid, as well as the effects that are caused by this drug. The Log-Logistic Model was an alternative to the Weibull model, which is widely used in survival analysis. In this context, Bayesian inference techniques were employed to estimate the parameters. The data that were obtained from patients treated over the period of 2008 to 2010 at Hospital Universitário de Maringá (HUM), with variable reactions the length of stay of patients in the intensive care unit (ICT). In order to estimate the parameter, it was generated three chains using for so, MCMC methods. In addition, it was considered some charts and tests to analyze the adjustment and convergence of these chains.

Keywords Survival analysis, Bayesian analysis, Log-Logistic density.

Lista de ilustrações

Lista de tabelas

Sumário

Sessão 1

Introdução

1.1 Introdução

Descoberta na década de 1990, liberou-se a Linezolida para uso em alguns países a partir de 2000, sendo que, no Brasil, esta droga começou a ser utilizada apenas em 2007. Por ser um antibiótico sintético de amplo aspecto e, em geral, bacteriostático, seu uso por curto período de tempo é considerado seguro. Entretanto, o uso da droga de forma indiscriminada vem preocupando os pesquisadores da área médica, já que não se conhece todos dos possíveis efeitos a longo prazo.

Neste estudo considera-se a variável resposta t o tempo de internamento dos pacientes na UTI após o início do uso desta droga, com o intuito de estudar o tempo de uso deste antibiótico. Além dos efeitos causados por este, foram coletados dados de 148 pacientes sob tratamento entre os anos de 2008 a 2010, no Hospital Universitário de Maringá (HUM). E para descrever o comportamento desta, foi proposto o modelo hierárquico Log-Logístico, o qual possibilita a modelagem da incerteza nos hiperparâmetros e induz uma decomposição da distribuição *a priori* em dois níveis: o primeiro nível é o da distribuição amostral dos dados; e o segundo nível, a distribuição *a priori* para o parâmetro θ. Esta decomposição da distribuição *a priori* é, geralmente, justificada pela dificuldade de se quantificar exatamente a informação da distribuição e pelo interesse em incorporar a incerteza decorrente sobre os hiperparâmetros.

O fato de que o Modelo Bayesiano frente à inferência estatística exigir a atribuição de distribuições de probabilidades a quantidades desconhecidas, sejam elas observáveis (como os dados) ou não (como os parâmetros de Modelos amostrais), motivou a realização do presente trabalho. No caso em questão, um parâmetro desconhecido (θ), ao ser analisado pelo Modelo Bayesiano, este permitirá reduzir o seu desconhecimento. Além disso, a intensidade da incerteza a respeito de θ pode assumir diferentes graus. Para os Baye-

sianos, cada problema *é único* e com diferentes graus de incerteza. São representados por meio de Modelos probabilísticos para o parâmetro desconhecido θ. Em geral, graus de conhecimento variam de problema para problema e de pesquisador para pesquisador, conforme (MURTEIRA, 1990).

Neste sentido, é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre θ e, desta forma, especificando Modelos distintos. Sendo assim, não existe nenhuma distinção entre quantidades observáveis e os parâmetros de um Modelo estatístico; todos são considerados quantidades aleatórias.

Na última década, observaram-se os enormes avanços tecnológicos, os quais têm permitido o armazenamento de grandes volumes de informações em diversas áreas do conhecimento. Em especial, a área biológica, que está se beneficiando destes avanços de forma significativa.

De acordo com (CARDOSO F. F. E ROSA, 2009), os Modelos Hierárquicos Bayesianos proporcionam uma metodologia geral e flexível para ajustar a complexidade de fatores genéticos e ambientais que afetam o desempenho em características biológicas complexas, considerando-se o conhecimento *a priori* e a informação contida nos dados.

Em particular, a metodologia bayesiana fornece um caminho natural para obtenção das distribuições *a posteriori*, as quais são muito mais informativas do que simples estimativas pontuais. Desta forma, pode-se também obter intervalos de credibilidade para as principais características, tais como a média, mediana, desvio padrão, quantiles etc.

Paralelo ao crescimento dos métodos bayesianos, têm-se as pesquisas relacionadas com o estudo do tempo de sobrevivência/confiabilidade, as quais têm sido apresentadas como uma vertente crescente, e seu desenvolvimento pode ser quantitativamente medido pelo número de textos e artigos produzidos na área nos últimos dez anos conforme (COLOSIMO E. A. E GIOLO, 2006).

Análise de sobrevivência busca estudar o tempo até a ocorrência de determinado evento de interesse. Entende-se por confiabilidade o estudo relacionado a dados de engenharia, atuaria, dentre outros, e, por estudo de sobrevivência, o estudo do tempo relacionado a experimentos clínicos. Os autores (LEE E. E WANG, 2003), (LAWLESS, 1982) e (KLEIN J. P. E MOESCHBERGER, 1997) apresentaram, de forma geral e sintetizada, os principais conceitos aplicados à análise de tempos de vida e confiabilidade, sendo uma das referências mais citadas na área. (TOMAZELLA V. L. D., 2013) apresenta um estudo relacionado ao tempo de vida, com uma covariável t, proposto por (FEIGL P. E ZELEN, 1965).

Em suma, para os Bayesianos a distribuição *a posteriori* incorpora, por meio do Teorema de Bayes, toda a informação disponível sobre o parâmetro (informação inicial + informação da experiência ou da amostra).

Desta forma, os Bayesianos defendem que a informação inicial ou *a priori*, ou seja, anterior em relação à experiência é demasiadamente importante para ser ignorada ou tratada *ad hoc*. Pode traduzir-se formalmente por uma distribuição de probabilidade.

1.2 Objetivo

Este trabalho tem por objetivos:

- Explorar as distribuições Normal e half-Cauchy como distribuições *a priori* para as quantidades desconhecidas do modelo Log-Logístico.
- Utilizar modelos hierárquicos bayesianos, com dois níveis de hierarquia.
- Obter o ajuste do modelo desenvolvido considerando-se dados simulados.

1.3 Organização da dissertação

O desenvolvimento e a aplicação de técnicas para análise de dados relacionados à Estrutura Hierárquica vêm ganhando um certo destaque na última década. No qual tem sido proposto e conseguido um avanço significativo, o que se deve em boa parte ao grande avanço computacional. Nesse contexto, faz-se necessária uma breve revisão dos principais conceitos, apresentados abaixo.

Este trabalho está dividido em 5 capítulos, da seguinte forma: no capítulo 1 é apresentado a Introdução; no capítulo 2 são apresentados, de forma sucinta, os conhecimentos necessários para o entendimento do presente Modelo, por meio da Revisão de Literatura. No capítulo 3 são apresentados Materiais e Métodos. No capítulo 4, Resultados e Discussão do Modelo Log-Logístico; e, finalmente, no capítulo 5, Conclusão, perspectivas futuras para continuidade do trabalho e Referências.

Sessão 2

Revisão de Literatura

2.1 Paradigma Bayesiano

A inferência Bayesiana modela a incerteza relativa aos parâmetros baseando-se em evidências *a priori*, por meio de distribuições de probabilidade conhecidas como distribuições *a priori*. Necessita-se de duas fontes de informações para se fazer inferência sobre uma quantidade desconhecida: a informação sobre ela presente nos dados, expressa pela função de verossimilhança e seu conhecimento prévio, modelado por meio da distribuição *a priori*, $\pi(\theta)$, (MURTEIRA B. E TURKMAN, 2003). Combinando-se essas duas informações e utilizando o Teorema de Bayes obtém-se a distribuição *a posteriori*.

Apresentam-se a seguir alguns conceitos gerais de análise Bayesiana, que é uma alternativa para a modelagem de dados em situações geralmente mais complexas.

Na inferência clássica, um dos objetivos costuma-se determinar que generalizações podem-se fazer (se algumas são possíveis) sobre a população de interesse, a partir da amostra colhida da mesma. Deseja-se estimar o parâmetro θ, com $\theta \in \Theta$ um escalar ou vetor desconhecido, mas fixo. No modelo Bayesiano o parâmetro θ é um escalar ou vetor aleatório (não observável), no qual toda incerteza deve ser quantificada em termos de probabilidade. Então, a inferência Bayesiana se baseia em probabilidades subjetivas ou credibilidades *a posteriori* associadas com diferentes valores dos parâmetros θ e condicionadas pelo particular valor de t observado.

A quantidade θ serve como indexador da família de possíveis distribuições para as observações t, representando características de interesse que se deseja conhecer para ter uma descrição completa do processo.

A estimação paramétrica na inferência Bayesiana pode encontrar problemas com modelos mais complexos, que podem resultar em distribuições *a posteriori* de difícil tratamento analítico, como, por exemplo, a presença de um grande número de parâmetros

a ser estimados, a dificuldade na obtenção das densidades marginais de forma analítica, ou ainda distribuições *a priori* e *a posteriori* que não são conjugadas. Nesses casos, os métodos analíticos de aproximação não são indicados, sendo necessários métodos de aproximação numérica para estimação dos parâmetros de interesse. Os métodos analíticos de aproximação são mais eficientes computacionalmente, porém são baseados em resultados assintóticos e na suposição de normalidade. Também, são mais difíceis para programar à medida que o número de parâmetros aumenta.

O procedimento Bayesiano é baseado no teorema de Bayes, ou seja, dados os tempos de falha $\mathbf{t}' = (t_1, t_2, \ldots, t_n)$ obtidos de um modelo paramétrico $f(\mathbf{t}|\theta)$, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, temos:

$$p\left(\mathbf{t}|\theta\right) = \frac{f\left(\theta|\mathbf{t}\right)\pi\left(\theta\right)}{\int f\left(\theta|\mathbf{t}\right)\pi\left(\theta\right)d\theta}, \tag{2.1}$$

em que $p(\theta|\mathbf{t})$ é a distribuição *a posteriori* de θ quando $\mathbf{T} = \mathbf{t}$, combinando a função de verossimilhança $f(\theta|\mathbf{t})$ com *a priori* $\pi(\theta)$.

Destaca-se que nos interessa algum componente de θ, digamos θ_i $(i = 1, ..., k)$; obtemos a distribuição marginal integrando a distribuição conjunta $p(\theta|\mathbf{t})$, ou seja,

$$p\left(t|\theta_i\right) = \int_{\Theta_{-i}} f\left(\theta|\mathbf{t}\right)\pi\left(\theta\right)d\theta_{-i}, \tag{2.2}$$

na qual o subscrito $(-i)$ implica a integração de todos os componente exceto θ_i.

Observa-se que os cálculos de (2.2) nem sempre são fáceis, havendo, muitas vezes, necessidade de recorrer a procedimentos numéricos. Esses procedimentos baseiam-se em amostragem que englobem métodos de integração simples de Monte Carlo, métodos de reamostragem por importância e métodos de Monte Carlo via Cadeias de Markov (MCMC). Estes são mais simples para implementação e não apresentam restrições quanto ao número de parâmetros a serem estimados.

Entre os métodos MCMC, os mais utilizados são o amostrador de Gibbs (*Gibbs Sampling*) e Metropolis-Hastings. Ver por exemplo, (MURTEIRA B. E TURKMAN, 2003). É importante observar que o uso destes algoritmos e métodos MCMC, em geral, é necessário se a geração não iterativa da distribuição, da qual se deseja obter uma amostra, for muito complicada ou custosa.

2.2 Distribuição *a priori*

Segundo (BOX G. E. P. E TIAO, 1992), a distribuição *a priori* resume a informação relativa ao parâmetro θ, desconhecido antes da realização do experimento, isto é, modela a incerteza relativa ao parâmetro antes da observação dos dados. Tal distribuição desem-

penha um papel importante da inferência Bayesiana, pois o uso de métodos Bayesianos tem se tornado uma alternativa poderosa na análise de dados por permitir ao pesquisador utilizar-se de informação/conhecimento *a priori*, usando-se, para isso, uma distribuição probabilística do conhecimento prévio.

Assim, a estatística via métodos Bayesianos vem ganhando espaço e se consolidando, tornando-se uma alternativa simples e viável a partir do implemento de computadores cada vez mais velozes, com sua fundamentação teórica baseada na famosa fórmula de Bayes. Ver por exemplo (BERGER, 1985) e (BOX G. E. P. E TIAO, 1973).

De acordo com (PAULINO D. E TURKMAN, 2003), a informação *a priori* que se pretende incorporar na análise é a informação fornecida por um especialista ou pesquisador, tais como dados históricos do problema ou de experimentos análogos.

Para os Bayesianos cada problema *é único* e tem um contexto real próprio onde θ é uma quantidade significativa acerca da qual existem, em geral, graus de conhecimento que variam de problema para problema e de pesquisador para pesquisador. Assim, a distribuição de probabilidade que capta essa variabilidade é baseada na informação *a priori* e, a distribuição *a priori* é de natureza subjetiva, ou seja, específica de um dado problema e de um dado pesquisador.

Desta forma, usam-se neste trabalho, para estimar os parâmetros μ e β, as distribuições *a priori* Normal e half-Cauchy, de acordo com (MURTEIRA B. E TURKMAN, 2003), respectivamente apresentadas abaixo. Assim, tem-se que a função densidade de probabilidade (f.d.p.) é dada por:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{1}{2}\left(\frac{t-\tau}{\sigma}\right)^2\right]}. \tag{2.3}$$

Apresenta-se sua função de distribuição acumulada,

$$F(t) = \frac{2}{\sqrt{2\pi}\sigma} \int_{-\infty}^{t} e^{\left[-\frac{1}{2}\left(\frac{t-\tau}{\sigma}\right)\right]} dt. \tag{2.4}$$

Pode-se simular a distribuição half-Cauchy, $T \sim HC(\mu = 0, \theta = 1)$ como chute inicial, como razão entre duas normais independentes. Sua função densidade de probabilidade é,

$$f(t) = \frac{1}{\pi(1+t^2)}, \qquad -\infty < t < \infty. \tag{2.5}$$

2.3 Distribuição *a posteriori*

Em análise Bayesiana, inferências são realizadas diretamente da distribuição *a posteriori*, baseando-se apenas nos dados disponíveis e na distribuição *a priori*. A partir da

posteriori podem-se obter estimativas pontuais, obtidas para resumir as características de tal distribuição, como também estimativas intervalares e inferências probabilísticas.

Entretanto, baseia-se em amostras de tamanho finito (pequeno). O teorema de Bayes propicia soluções precisas para o problema de amostras de tamanho finito, pois, para cada conjunto de dados, pequeno ou grande, existe uma distribuição *a posteriori* exata para realização de inferências.

Para um valor fixo de t, a função $L(\theta, t) = \pi(t|\theta)$ fornece a plausibilidade ou verossimilhança de cada um dos possíveis valores de θ enquanto $\pi(\theta)$ é chamada distribuição *a priori* de θ. Estas duas fontes de informação, *a priori* e verossimilhança, são combinadas levando à distribuição *a posteriori* de θ, dada por $\pi(\theta|t)$. Assim, a forma usual do teorema de Bayes é:

$$\pi(\theta|t) \propto L(\theta, t)\pi(\theta).$$

A distribuição *a posteriori* conjunta de todos os parâmetros do Modelo é obtida pelo produto de todas as densidades especificadas.

2.4 Distribuições Hierárquicas

Distribuições *a priori* hierárquicas: ao invés de especificar a distribuição *a priori* como uma função simples, elas consistem em decompor a especificação da distribuição *a priori* em níveis, conforme (PAULINO D. E TURKMAN, 2003). A distribuição *a priori* de um parâmetro θ depende dos valores dos hiperparâmetros. Além disso, ao invés de fixar valores para os hiperparâmetros, pode-se especificar uma distribuição *a priori* $\pi(\theta)$, completando, assim, o segundo nível na hierarquia.

Segundo (GELMAN, 2006), teoricamente não há limitação quanto ao número de níveis, mas devido à complexidade resultante, as distribuições *a priori* hierárquicas são especificadas geralmente em dois ou três níveis. Devido à dificuldade de interpretação dos hiperparâmetros em níveis mais altos, na prática é comum especificar distribuições *a priori* subjetivas para estes níveis. Ainda, de acordo com (BERNARDO J. M E SMITH, 1994), a metodologia Hierárquica modela a incerteza nos hiperparâmetros por meio de novas distribuições, induzindo-se assim, a uma decomposição da distribuição *a priori* em níveis.

O Modelo Bayesiano deste trabalho foi definido por meio de uma hierarquia de dois níveis, sendo o primeiro nível da distribuição amostral dos dados e o segundo, a distribuição *a priori* para o parâmetro θ.

O Modelo hierárquico Bayesiano, $\{\pi(t|\theta), \pi(\theta)\}$, onde $\pi(\theta)$ pode ser decomposto nas condicionais. Esta decomposição da distribuição *a priori* é, geralmente, justificada pela dificuldade de se quantificar exatamente a informação dessa distribuição e pelo interesse

em incorporar a incerteza decorrente sobre os hiperparâmetros, segundo (MURTEIRA B. E TURKMAN, 2003).

De forma geral, a distribuição *a priori* do primeiro nível é uma distribuição conjugada natural do modelo amostral por motivos de flexibilidade computacional.

Desta forma, verifica-se que a distribuição *a posteriori* de θ é:

$$\pi(t|\theta) = \pi(t_1, \ldots, t_k|\theta_1, \ldots, \theta_k) = \prod_{i=1}^{k} \pi(t_i|\theta_i),$$

$$\pi(\theta|\phi) = \pi(\theta|\phi) = \prod_{i=1}^{k} \pi(\theta_i|\phi),$$

$$\pi(\phi)$$

Assim, se θ é um parâmetro discreto, designado por $\pi(\theta)$ a função de probabilidade *a priori*, tem-se que $\pi(\theta)$ exprime o grau de credibilidade, que se atribui ao particular θ considerado. Se θ é um parâmetro contínuo, caso mais frequente e do presente trabalho, designado por $\pi(\theta)$ a função densidade de probabilidade *a priori*, tem-se que $\pi(\theta)d\theta$ exprime o grau de credibilidade que o mesmo indivíduo atribui ao intervalo $(\theta, \theta + d\theta)$.

Dado $t_1, \ldots, t_k$ que estão disponíveis a partir de k observações diferentes, mas de fontes relacionadas. Mas, por causa da relacionalidade das k observações, o parâmetros $\theta_1, \ldots, \theta_k$ são julgados por si mesmos permutáveis. O segundo ou terceiro níveis de hierarquia proporciona assim *a priori* de θ da forma de representação de mistura de distribuições familiares.

$$\pi(\theta) = \pi(\theta_1, \ldots, \theta_k) = \int \prod_{i=1}^{k} \pi(\theta_i|\phi)\pi(\phi)d\phi$$

Manipulações de probabilidade envolvendo teorema de Bayes, de forma geral, fornecem as inferências *a posteriori* necessárias,

$$\pi(\theta_i|t) = \int \pi(\theta_i|\phi, t)\pi(\phi|t)d\phi, \tag{2.6}$$

onde

$$\pi(\theta_i|\phi, t) \propto \pi(t|\theta_i)\pi(\theta_i|\phi)$$

$$\pi(\phi|t) \propto \pi(t|\phi)\pi(\phi)$$

e

$$\pi(t|\phi) = \int \pi(t|\theta)\pi(\theta|\phi)d\theta.$$

2.5 Métodos computacionais

Segundo (BOX G. E. P. E TIAO, 1992), para inferir em relação a qualquer parâmetro unidimensional do vetor $\boldsymbol{\theta}$, a distribuição conjunta *a posteriori* dos parâmetros (multidimensional) deve ser integrada em relação a todos os outros parâmetros que a constituem, ou seja, deve-se procurar obter a distribuição marginal de cada um dos parâmetros. Geralmente existem dificuldades para a obtenção de uma forma analítica para a distribuição marginal. Sendo que essas dificuldades devem-se à complexidade das distribuições conjuntas obtidas ou devido à dimensão do parâmetro θ em estudo.

2.5.1 Método de Monte Carlo via Cadeias de Markov

O Método de Monte Carlo é baseado na simulação de Cadeias de Markov (MCMC), cuja distribuição estacionária é a distribuição *a posteriori* de interesse, sem a qual não é possível gerar uma amostra para fazer inferência via Monte Carlo.

A ideia básica por meio do Método é transformar o problema estático num problema dinâmico, construindo um processo estocástico temporal, artificial, que seja fácil de simular e que convirja para a distribuição original. Em geral, este processo é temporal.

Uma Cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia, depende somente do estado atual e dos dados, não da história passada da cadeia, segundo (MURTEIRA B. E TURKMAN, 2003). As primeiras iterações são influenciadas pelo estado inicial e são descartadas. Esse período é conhecido como aquecimento da cadeia (*burn-in*). Além disso, considera-se uma dependência entre as observações subsequentes da cadeia e, para diminuir a alta correlação existente entre os valores amostrais, deve-se considerar um espaçamento entre as iterações armazenadas, digamos k iterações. Esse valor é denominado como *lag*.

A ideia dos métodos MCMC é obter uma amostra da distribuição conjunta dos parâmetros de interesse, por meio de um processo iterativo, onde, ao final de cada ciclo de atualizações, os valores gerados são considerados amostras aleatórias da distribuição de probabilidade conjunta. Baseiam-se na construção de uma Cadeia de Markov homogênea, que tem como distribuição limite a distribuição que se pretende simular. Neste trabalho a distribuição Log-Logístico.

Têm-se dois métodos que geram variáveis independentes identicamente distribuídos (iid): o algoritmo Metropolis-Hastings e o algoritmo Gibbs, que se apresentam a seguir.

2.5.2 Método Metropolis-Hastings

O algoritmo de Metropolis-Hastings gera uma amostra da distribuição conjunta *a posteriori* $\pi(\theta|t)$, a partir das distribuições condicionais completas com formas analíticas não conhecidas na literatura. Tal algoritmo usa a ideia de que um valor é gerado de uma distribuição proposta ou candidata (c), θ^c, e este valor é aceito com uma dada probabilidade, conforme (HASTINGS, 1970) e (METROPOLIS N. E ROSEMBLUT, 1953).

O algoritmo de Metropolis-Hastings está estruturado nos seguintes passos:

(1) Inicialize o contador de iterações $t = 0$ e especifique os valores iniciais $\theta^0 = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$ para os parâmetros.

(2) Gere um valor θ^c da distribuição proposta $q(.|\theta_1)$, usualmente denominada de *função de importância*.

(3) Calcule a probabilidade de aceitação $\alpha(\theta_1, \theta^c)$, onde,

$$\alpha(\theta_1, \theta^c) = min\left\{1, \frac{\pi(\theta^c|\theta_2, \ldots, \theta_p)q(\theta_1|\theta^c)}{\pi(\theta_1|\theta_2, \ldots, \theta_p q(\theta^c|\theta_1))}\right\}.$$

(4) Gere um valor u, a partir de uma distribuição $U(0, 1)$.

(5) Se $u < \alpha$, então aceite o valor candidato e faça $\theta_1^{(t+1)} = \theta^c$. Caso contrário, rejeite e faça $\theta_1^{(t+1)} = \theta^t$.

(6) Incremente o contador de t para $t + 1$ e volte ao passo (2) até atingir a convergência.

Neste trabalho, faz-se uso do algoritmo Metropolis-Hastings, o qual gera variáveis correlacionadas a partir de uma Cadeia de Markov.

2.5.3 Método de Gibbs

O algoritmo Gibbs é um caso especial do Método Metropolis-Hastings. Observam-se dois tipos de amostrador do algoritmo Gibbs:

(1) Gibbs bi-etapa:

- O método Gibbs cria uma Cadeia de Markov a partir de uma distribuição conjunta.
- Se duas variáveis X e Y têm densidade conjunta, então $f(x, y)$.
- Com densidades condicionais $f_{y|x}$ e $f_{x|y}$.
- Gerando-se Cadeias de Markov (X_t, Y_t).

2.6 Intervalos de credibilidade HPD

Para se obter os intervalos de credibilidade de mais alta densidade (*hight probability density* - HPD), defini-se uma função de densidade, a qual se baseia no fato de que os limites do intervalo terão o mesmo valor de densidade e faz-se então a procura destes limites para o nível desejado.

A definição para um intervalo de credibilidade é:

$$R(t) = \{\theta : \pi(\theta|t) \geq c_\gamma\},$$

em que c_γ é a maior constante tal que:

$$P(\theta \in R(t)|t) = \int_{R(t)} \pi(\theta|t)dt \geq \gamma.$$

Destaca-se que, para o Modelo Log-Logístico, o intervalo HPD corresponde àquele em que *a posteriori* apresenta o mesmo valor nos limites, ou seja,

$$\pi(R_1(t)|t) = \pi(R_2(t)|t).$$

Em geral, não se obtem analiticamente e, assim, deve-se empregar métodos numéricos para construí-los.

No presente trabalho, tem-se:

$$\theta|t \sim \text{Gama}(\lambda, \phi)$$

Observa-se que *a posteriori* é assimétrica e unimodal. Desta forma, obtem-se um intervalo de credibilidade, HPD(λ, γ), onde se utiliza os quantis da distribuição Gama, da seguinte forma, para $\gamma = 0,95$,

$$P(\lambda \leq R_1|t) = \frac{1-\gamma}{2} \qquad e \qquad P(\lambda \geq R_2|t) = \frac{1-\gamma}{2}.$$

2.7 Análise Residual

Análise Residual de dados completos (sem censura) trata, geralmente, sobre o enfoque de regressão.

Pode-se definir os resíduos de várias maneiras, mas a ideia fundamental é que se um Modelo para a distribuição de T dado t, especificado em termos do parâmetro θ, aplicando-se a dados independentes (t_i, x_i), $i = 1, ...n$.

Na maioria das configurações buscam-se resíduos $e_1, \ldots, e_n$ os quais são aproximadamente independentes e identicamente distribuídos (iid) quando o Modelo está correto.

Outras considerações:

- Obtem-se a construção da verossimilhança da mesma forma que outras distribuições.
- Equações não podem-se resolvidas analiticamente.
- Necessita-se de métodos iterativos de solução, por exemplo, método de Newton-Raphson.

Para revelar o impacto destes pontos detectados, as seguintes medidas RC_{θ_j} podem ser calculadas como sendo:

$$\text{RC}_{\theta_j} = \left| \frac{\hat{\theta}_j - \hat{\theta}_{j(I)}}{\hat{\theta}_j} \right| \times 100\%, \qquad j = 1, \ldots, p+1,$$

com $\hat{\theta}_{j(I)}$ denotando as estimativas de verossimilhança de θ_j após o conjunto I de observações ter sido removido. Desta forma, temos as seguintes medidas:

$$\text{TRC} = \sum_{i=1}^{n_p} \left| \text{RC}_{\theta_j} \right|, \quad \text{MRC} = \max_j \left| \text{RC}_{\theta_j} \right| \quad \text{and} \quad LD_{(I)}(\boldsymbol{\theta}) = 2\{l(\hat{\boldsymbol{\theta}} - l(\hat{\boldsymbol{\theta}}_I)\},$$

no qual TRC representa a mudança relativa total, MRC o máximo desta mudança e LD a diferença de verossimilhanças com $n_p = 2$ (número de parâmetros e hiperparâmetros) e $\hat{\boldsymbol{\theta}}^0$ denota as estimativas *a posteriori* de $\boldsymbol{\theta}$ após o conjunto I de observações ser removido.

2.8 Dados de Sobrevivência

Paralelo ao crescimento dos métodos Bayesianos, têm-se as pesquisas relacionadas com o estudo do tempo de sobrevivência/confiabilidade, a qual apresenta-se como uma vertente crescente, e seu desenvolvimento pode-se mensurar pelo número de textos e artigos produzidos na área nos últimos dez anos conforme (COLOSIMO E. A. E GIOLO, 2006).

Destaca-se que, na análise de dados de sobrevivência e confiabilidade, o Modelo mais utilizado é a distribuição Weibull, a qual possui várias propriedades. Dentre elas, destaca-se que a função de taxa de falha pode modelar diferentes formas (constante, crescente, decrescente, unimodal e forma de U), mas limitou-se na função de risco,

observando-se como monótona, constante para $\beta = 1$, crescente para $\beta > 1$, e decrescente para $\beta < 1$, sendo que β é o parâmetro que caracteriza a forma da distribuição dos tempos de sobrevivência.

Pode-se definir análise de sobrevivência como a análise do tempo até a ocorrência de uma dado evento, onde este tempo é denominado tempo de falha, o qual pode ser o tempo até a morte do paciente, bem como até a cura ou recidiva de uma doença.

Observa-se que alguns eventos de interesse não são terminais e podem ocorrer, para o mesmo indivíduo, mais de uma vez, sendo denominado eventos recorrentes. Este tipo de dados surge naturalmente em estudos longitudinais nas áreas de estudos clínicos, demografia, criminologia, confiabilidade industrial e produção.

Entende-se por confiabilidade o estudo relacionado a dados de engenharia atuarias, dentre outros citados a cima e, por estudo de sobrevivência, o estudo do tempo relacionado a experimentos clínicos.

Os autores (LEE E. E WANG, 2003), (LAWLESS, 1982), (KLEIN J. P. E MOESCHBERGER, 1997), apresentaram, de forma geral e sintetizada, os principais conceitos aplicados à análise de tempos de vida e confiabilidade, sendo uma das referência mais citadas na área. Conforme (TOMAZELLA V. L. D., 2013) apresentou um estudo relacionado ao tempo de vida, com uma covariável t, proposto por (FEIGL P. E ZELEN, 1965).

Apresentam-se a seguir, alguns conceitos que são importantes em análise de sobrevivência.

Destaca-se que os modelos de sobrevivência apresentam-se como classe de modelos estocásticos usados para analisar características e fatores associados ao tempo até ocorrência do evento de interesse (falha ou desfecho).

Observa-se que os principais elementos da análise de sobrevivência são:

- estruturas de causalidade fundamentadas em raciocínio epidemiológico e evidências empíricas;
- associação de efeitos principais e de interação a cada variável;
- possibilidade de afirmações probabilísticas para descrição e representação do conhecimento;
- variabilidade associada a predições produzidas pelo modelo, ou seja, pelos intervalos de credibilidade (HPD) para as estimativas.

Considere-se uma quantidade de interesse desconhecida θ (tipicamente não observável), conforme (MURTEIRA B. E TURKMAN, 2003). A informação que se dispõe sobre θ,

resumida probabilisticamente através da distribuição *a priori* de θ, $\pi(\theta)$, pode ser aumentada observando-se uma quantidade aleatória t relacionada com θ. A distribuição amostral $\pi(t|\theta)$ define esta relação. A ideia é que, após observar $T = t$, a quantidade de informação sobre θ aumenta e este acrescimo é bastante intuitivo e o Teorema de Bayes é a regra de atualização utilizada para quantificar este aumento de informação:

$$\begin{aligned} \pi(\theta|t) &= \frac{\pi(\theta,t)}{\pi(t)} \\ &= \frac{\pi(t|\theta)\pi(\theta)}{\pi(t)} \\ &= \frac{\pi(t|\theta)\pi(\theta)}{\int \pi(\theta,t)d\theta}. \end{aligned} \tag{2.7}$$

Trata-se de uma proposição extremamente simples, conforme (BOX G. E. P. E TIAO, 1973). Nota-se que $\frac{1}{\pi(t)}$, que não depende de θ, funciona como uma constante normalizadora de $\pi(\theta|t)$.

Para um valor fixo de t, a função $L(\theta,t) = \pi(t|\theta)$ fornece a verossimilhança de cada um dos possíveis valores de θ, enquanto $\pi(\theta)$ é chamada de distribuição *a priori* de θ. Estas duas fontes de informações, *a priori* e verossimilhança, são combinadas levando à distribuição *a posteriori* de θ, $\pi(\theta|t)$.

Assim, a forma usual do Teorema de Bayes é dado pela expressão:

$$\pi(\theta|t) \propto l(\theta;t)\pi(\theta). \tag{2.8}$$

em que $\propto$ indica proporcionalidade.

Em palavras obtem-se:

distribuição *a posteriori* $\propto$ verossimilhanca x distribuição *a priori*.

Observa-se que, ao omitir o termo $\pi(t)$, a igualdade em (1.1) transforma-se em uma proporcionalidade. Sendo que esta forma simplificada do Teorema de Bayes será útil em problemas que envolvam estimação de parâmetros, visto que o denominador é apenas uma constante normalizadora. Em outras situações, como seleção de Modelos, este termo tem um papel fundamental.

Desta forma, verifica-se que a função de verossimilhança tem importante papel na fórmula de Bayes, pois representa o meio através do qual os dados, t, transformam o conhecimento *a priori* sobre θ, ou seja, a verossimilhança pode interpretar-se como expressão da informação sobre θ fornecida pelos dados t.

Em análise de sobrevivência as unidades de estudo são usualmente indivíduos, as quais essencialmente formam-se pelos tempos de vida das mesmas até a ocorrência do evento de interesse.

Os dados de sobrevivência incorporam tanto os tempos de sobrevivência como o conjunto de outras variáveis observáveis, que podem estar relacionadas com esses tempos. Destaca-se que estas variáveis são denominadas covariáveis ou variáveis explicativas. Observa-se que, quando os tempos de sobrevivência estão relacionados com essa covariáveis, diz-se que a população é heterogênea; caso contrário, ela é homogênea, conforme (COLOSIMO E. A. E GIOLO, 2006).

Observa-se que as covariáveis normalmente são medidas uma única vez ao longo do tempo de estudo e, por conta disso, são fixas. Em alguns casos, onde as covariáveis não são fixas, estas variam em função do tempo, modificando-se durante o período de observação.

Tem-se uma variável aleatória contínua e não negativa denotando o tempo de sobrevivência $(T \geq 0)$; têm-se então as funções de densidade de probabilidade $f(t)$, a de sobrevivência $S(t)$ e a de risco $h(t)$, que são apresentadas posteriormente mais detalhadamente. Utilizam-se, na prática, estas funções, com o objetivo de descrever os aspectos apresentados pelo conjunto de dados.

A função densidade de probabilidade (fdp) é expressa como o limite da probabilidade de um indivíduo vir a experimentar o evento de interesse no intervalo de tempo $[t, t + dt)$ por unidade de tempos, ou seja, expressa por:

$$f(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T \leq t + \triangle t)}{\triangle t}. \tag{2.9}$$

Segundo (LAWLESS, 1982), a função dos tempos de sobrevivência, $S(t)$, é a função que descreve a forma distribucional dos tempos de sobrevivência através da probabilidade de um indivíduo não falhar (ou do evento de interesse não ocorrer), pelo menos até um instante de tempo t.

A função de risco também é conhecida como taxa de mortalidade condicional. Além disso, através da função de risco podem-se caracterizar classes especiais de distribuições de tempos de sobrevivência, de acordo com o seu comportamento em relação ao tempo. A função de risco pode ser constante, crescente, decrescente ou mesmo não monótona.

As funções de sobrevivência, de densidade e de risco são matematicamente equivalentes, (COLOSIMO E. A. E GIOLO, 2006). Uma vez definida qualquer uma delas, respeitando suas propriedades, obtêm-se as demais por consequência.

A função de sobrevivência é uma das principais funções usadas para descrever a variável aleatória “tempo” e é definida como sendo a probabilidade de um indivíduo não

falhar (ou de o evento de interesse não ocorrer) até um determinado tempo t, ou seja, a probabilidade de uma observação sobreviver ao tempo t, ver (LEE E. E WANG, 2003). Em termos probabilísticos, isto é escrito como:

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= 1 - P(T < t) \\ &= 1 - \int_0^t f(u)\, du, \end{aligned} \tag{2.10}$$

em que $f(\cdot)$ é a função densidade de probabilidade.

Alternativamente, (2.10) pode ser escrita na forma:

$$S(t) = 1 - F(t), \tag{2.11}$$

em que $F(t)$ é a probabilidade de um indivíduo não sobreviver ao tempo t. Assim, a partir de (2.11), temos $S(t) + F(t) = 1$.

Pelas propriedades da função de sobrevivência e da função densidade acumulada, temos que:

$$\begin{aligned} \lim_{t\to 0} S(t) &= 1 \quad \text{e} \quad \lim_{t\to \infty} S(t) = 0, \\ \lim_{t\to 0} F(t) &= 0 \quad \text{e} \quad \lim_{t\to \infty} F(t) = 1. \end{aligned}$$

Tendo em vista estas propriedades, como $F(t)$ é uma função o monótona crescente e $S(t)$ uma função monótona decrescente, ou não crescente, conforme (LEE E. E WANG, 2003).

A função de sobrevivência pode ser obtida também a partir das relações:

$$f(t) = \frac{d}{dt} F(t) \quad \text{ou} \quad f(t) = -S'(t). \tag{2.12}$$

Observa-se que o $100(1-p)\,\%$ percentil da variável aleatória T é definido como o valor de t_p tal que $P(T \leq t_p) = p$.

Destaca-se que a função de risco fornece a taxa instantânea de falha dado que o indivíduo sobreviveu até um determinado tempo t. Esta função fornece a probabilidade de o indivíduo falhar no intervalo de tempo $t + \Delta t$ com $\Delta t \to 0$. Ou seja,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|\, T \geq t)}{\Delta t}. \tag{2.13}$$

Devido à sua interpretação, a função de risco (2.13) tem sido preferida por muitos autores para descrever o comportamento do tempo de sobrevivência. A função de risco descreve como a probabilidade instantânea de falha (taxa de falha) se modifica com o passar do tempo. Ela é também conhecida como taxa de falha instantânea, força de mortalidade e taxa de mortalidade condicional (LAWLESS, 1982).

Além disso, através da função de risco podemos caracterizar classes especiais de distribuições de tempo de sobrevivência, de acordo com o seu comportamento como função do tempo. A função de risco pode ser constante, crescente, decrescente ou mesmo não monótona.

A probabilidade de uma falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência como sendo:

$$S(t_1) - S(t_2). \tag{2.14}$$

A taxa de falha ou risco, no intervalo $[t_1, t_2)$ é definida como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, dividida pelo comprimento do intervalo. Algebricamente:

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1) S(t_1)}. \tag{2.15}$$

De forma geral, redefinindo o intervalo como sendo $[t, t + \Delta t)$, temos a partir da função (2.15):

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{2.16}$$

A partir da função de risco, dada em (2.13), temos que:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)} \\ &= \lim_{\Delta t \to 0} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t}\right] \frac{1}{S(t)} \\ &= \frac{d}{dt} F(t) \frac{1}{S(t)} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \tag{2.17}$$

em que $f(t)$ é a função densidade da variável aleatória T.

Usando a relação apresentada em (2.12) temos que (2.17) pode ainda ser escrita como sendo:

$$h\left(t\right)=-\frac{S'\left(t\right)}{S\left(t\right)}=-\frac{d}{dt}\log\left[S\left(t\right)\right]. \tag{2.18}$$

A partir da integração de (2.18) temos que:

$$S\left(t\right)=\exp\left[-\int_0^t h\left(u\right)du\right]. \tag{2.19}$$

onde $\int_0^t h\left(u\right)du$ é a função de risco acumulada e é finita para algum tempo $t>0$.

Destaca-se que a principal característica de dados de sobrevivência é a observação parcial da resposta, denominada censura. Podem-se citar quatro tipos:

- Censura do tipo I: ocorre quando o estudo termina em um tempo preestabelecido e alguns dos tempos de sobrevivência dos pacientes não puderam ser observados, tendo seus tempos censurados à direita.
- Censura do tipo II: ao invés do tempo final do estudo ser preestabelecido, o estudo termina após um determinado número de pacientes experimentar o evento de interesse. Neste caso, se não há perda acidental, as observações censuradas são de comprimento igual a das observações não censuradas.
- Censura aleatória: diferentemente das outras censuras, ela é mais difícil do experimentador ter controle. Geralmente ocorre quando o paciente abandona o estudo sem ter experimentado o evento de interesse.
- Censura intervalar: ocorre quando não se conhece o tempo exato em que ocorreu o evento de interesse, mas, sim, que ele ocorreu dentro de um intervalo de tempo especificado.

2.8.1 Análise de Influência

Uma das formas para se verificar o bom ajuste de um modelo é através de análise de resíduos e influência, segundo (COOK, 1986), conforme exposto a seguir.

A primeira ferramenta para acessar a sensibilidade das medidas é a partir de uma análise de influência global. Para isso, fizemos um estudo de caso deleção no qual o efeito de cada observação nas estimativas foi medido. A primeira medida usada foi a distância generalizada de Cook, a qual é definida por $\boldsymbol{\theta}_i=(\alpha_i,\beta_i,\lambda_i)$ e $\hat{\boldsymbol{\theta}}=(\hat{\alpha},\hat{\beta},\hat{\lambda})$ e é dada por:

$$CD_i(\boldsymbol{\theta})=\left[\ \boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}\ \right]^T\left[-\ddot{L}(\boldsymbol{\theta})\right]\left[\ \boldsymbol{\theta}_i-\hat{\boldsymbol{\theta}}\ \right], \tag{2.20}$$

em que $\ddot{L}(\boldsymbol{\theta})$ pode ser aproximada através da estimativa da matriz de variância e covariância. Outra forma de medir a influência global de um ponto é a partir da distância de verossimilhanças dada por:

$$LD_i(\boldsymbol{\theta}) = 2\left\{l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta}_i)\right\}. \tag{2.21}$$

2.8.2 Análise de resíduos

Análise de resíduos tem a finalidade de avaliar a adequação da distribuição proposta para a variável resposta. Segundo (COLOSIMO E. A. E GIOLO, 2006), os gráficos de resíduos *martingale*, ou *deviance*, versus os tempos fornecem uma forma de verificar a adequação do modelo ajustado, bem como auxiliam na detecção de observações atípicas.

Neste trabalho, os resíduos utilizados foram: *martingale* e *deviance*. Os resíduos *martingale* foram inicialmente introduzidos para processos de contagem e depois utilizados para modelos de regressão paramétricos, sobrevivência, etc. Este resíduo é definidos por:

$$\hat{m}_i = \delta_i - \hat{e}_i, \tag{2.22}$$

em que δ_i é a variável indicadora de falha e $\hat{e}_i$ são os resíduos de Cox-Snell.

Já os resíduos *deviance* são uma tentativa de tornar os resíduos *martingale* mais simétricos em torno de zero, o que, em geral, facilita a detecção de pontos atípicos. São definidos por:

$$\hat{d}_i = \textbf{sinal}(\hat{m}_i)\left[-2(\hat{m}_i) + \delta_i \log(\delta_i - \hat{m}_i)\right]^{\frac{1}{2}}. \tag{2.23}$$

Destacamos que se o modelo for apropriado, esse resíduos devem apresentar-se com um comportamento aleatório em torno de zero.

Sessão 3

Materiais e Métodos

Embora a distribuição Weibull, proposta por W. Weibull (1953), seja amplamente utilizada na análise de dados de sobrevivência e/ou confiabilidade, este modelo apresenta grande limitação na forma da função de risco. Enquanto o modelo Weibull se adequa à situações com taxa de falha monótona (crescente, decrescente ou constante), conforme (FABRI F. V., 2012), não é incomum que o comportamento aleatório das observações experimentais dos fenômenos da vida real apresentem taxa de falha unimodal, conforme (COLOSIMO E. A. E GIOLO, 2006). Em diversas situações reais a função de risco muda de direção após algum valor máximo, como o caso dos tempos de vida de pacientes transplantados ou o tempo até a falha de componentes de placas eletrônicas, por exemplo.

Desta forma, nesta seção, apresentamos o modelo Log-Logístico. Esta distribuição acaba sendo uma boa alternativa ao modelo Weibull por ser simples, mais flexível quanto à forma da taxa de falha e ainda com parâmetros-chave totalmente interpretáveis.

3.1 Modelo Log-Logístico

Neste trabalho foi proposto um estudo do antibiótico linezolida, com dados reais coletados no Hospital Universitário de Maringá (HU) com 148 pacientes internados na Unidade de Terapia Intensiva (UTI) com doenças graves, onde a variável de interesse t é o tempo de internação do paciente na UTI após o início do uso desta droga, por meio do modelo Log-Logístico apresentado a seguir.

Suponha X uma variável aleatória não negativa com distribuição logística, sua função densidade de probabilidade é definida por:

$$f(x) = \frac{\sigma^{-1} e^{\frac{(x-\mu)}{\sigma}}}{\left[1 + e^{\frac{(x-\mu)}{\sigma}}\right]^2}, \tag{3.1}$$

com $\mu = \log x$, $\sigma = \beta^{-1}$, $-\infty < \mu < \infty$ e $\sigma > 0$.

Considere a seguinte transformação: $T = \log X$. Desta forma a variável aleatória T, que, neste caso, representa o tempo de vida de um indivíduo, é log-logisticamente distribuída com função densidade de probabilidade dada por:

$$f(t) = \frac{e^{\mu}\beta t^{\beta-1}}{[1 + e^{\mu}t^{\beta}]^2}, \tag{3.2}$$

em que $-\infty < \mu < \infty$ e $\sigma > 0$.

Apesar de existirem diferentes parametrizações para o modelo Log-Logístico, utiliza-se esta parametrização, pois esta nos permite a interpretação dos parâmetros de forma mais simplificada. Assim, a função de distribuição é dada por:

$$F(t) = \frac{e^{\mu}t^{\beta}}{1 + e^{\mu}t^{\beta}}. \tag{3.3}$$

É importante ressaltar que os parâmetros μ e β são, respectivamente, parâmetros de escala e forma. O parâmetro μ também representa, nesta parametrização, a mediana da distribuição.

A figura 1 apresenta as curvas de densidade para diferentes combinações de valores dos parâmetros de escala e forma, respectivamente, $LL(\mu, \beta)$, onde pode-se observar um comportamento para valores diferentes atribuídos a μ e β, onde posteriormente escolheram-se os chutes iniciais.

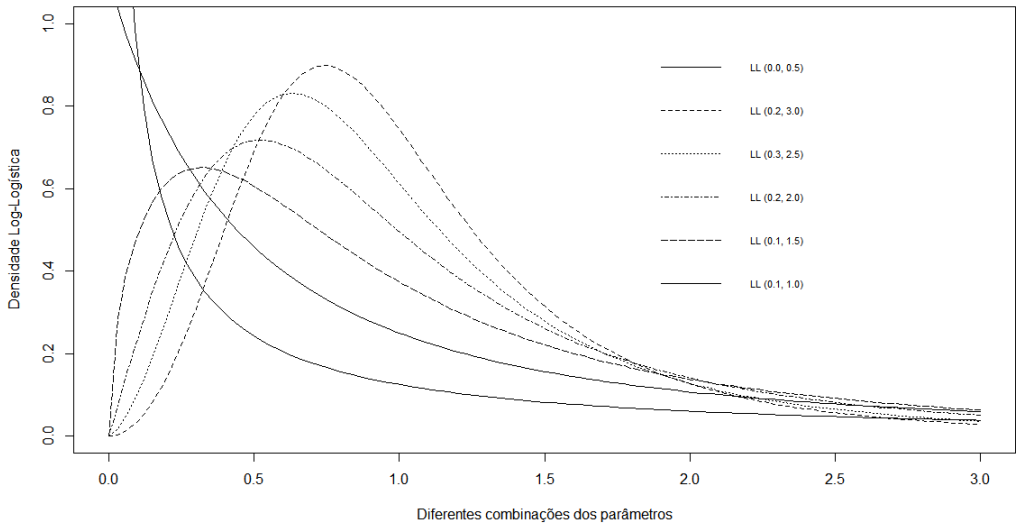

Figura 1 – Exemplo de curvas de densidade Log-Logístico para diferentes combinações dos parâmetros μ e β.

A distribuição Log-Logístico é uma distribuição de probabilidade contínua para uma

variável aleatória positiva, neste caso, é o tempo (t) até a morte do paciente ou algum outro evento de interesse. Ela é usada no estudo de eventos de vida cuja intensidade, primeiro aumenta e depois diminui, ou seja, em que a função de risco tem forma côncava, onde inicialmente cresce para o máximo no tempo e então decresce para zero quando o tempo se aproxima do infinito.

3.1.1 Funções de Risco e Sobrevivência

A vantagem do Modelo Log-Logístico é apresentar uma expressão simples para a função de risco e para função de sobrevivência, respectivamente, indicadas nas expressões abaixo:

$$h(t) = \frac{e^{\mu}\beta t^{\beta-1}}{1 + e^{\mu}t^{\beta}} \tag{3.4}$$

em que $t > 0$, $-\infty < \mu < \infty$ e $\beta > 0$.

Particularmente, para a distribuição Log-Logístico a função de risco e sobrevivência são definidas, respectivamente, por:

$$h(t) = \frac{e^{\mu}\beta t^{\beta-1}}{1 + e^{\mu}t^{\beta}} \tag{3.5}$$

e

$$S(t) = \frac{1}{1 + e^{\mu}t^{\beta}}. \tag{3.6}$$

com $-\infty < \mu < \infty$ e $\beta > 0$, ver figuras 2. Nas figuras 3, apresentamos as curvas de risco e sobrevivência da variável aleatória T, ou seja, de uma variável aleatória com distribuição logística.

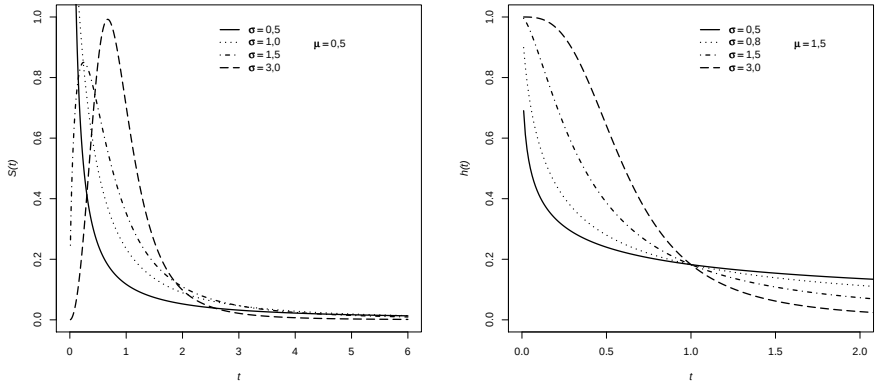

Figura 2 – Exemplo de curvas de sobrevivência e risco Log-Logístico para diferentes combinações dos parâmetros.

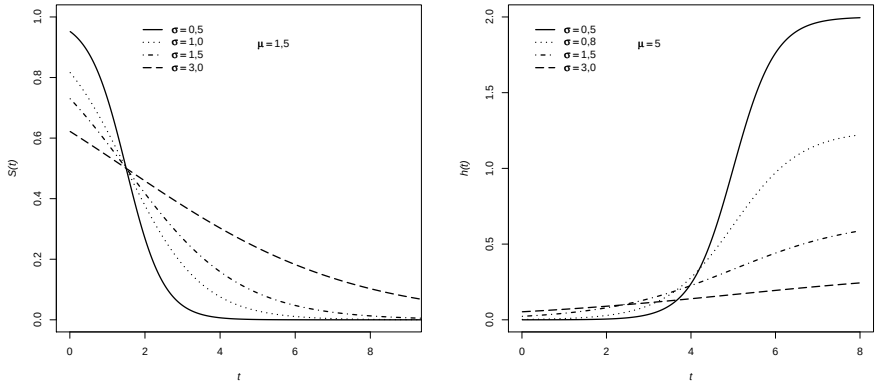

Figura 3 – Exemplo de curvas de sobrevivência e risco Log-Logístico para diferentes combinações dos parâmetros.

3.1.2 Função Geradora de Momentos da Log-Logístico

Os momentos de $T = logT$ são encontrados por meio da Função Geradora de Momentos (FGM).

Sabe-se que a primeira derivada da função distribuição Log-Logístico, dada anteriormente, é o momento de ordem 1, dado por:

$$\begin{aligned}
M'_T(t) &= \frac{e^{\mu}\beta(\beta-1)t^{\beta-2}(1+e^{\mu}t^{\beta})^2 - e^{\mu}\beta t^{\beta-1}2(1+e^{\mu}t^{\beta})e^{\mu}\beta t^{\beta-1}}{(1+e^{\mu}t^{\beta})^4} \\
&= \frac{(1+e^{\mu}t^{\beta})e^{\mu}\beta[(\beta-1)t^{\beta-2}(1+e^{\mu}t^{\beta}) - t^{\beta-1}2e^{\mu}t^{\beta-1}\beta]}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta[(\beta-1)t^{\beta-2}(1+e^{\mu}t^{\beta}) - 2e^{\mu}\beta t^{2(\beta-1)}]}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta(\beta-1)t^{\beta-2}(1+e^{\mu}t^{\beta})}{(1+e^{\mu}t^{\beta})^3} - \frac{2(e^{\mu})^2\beta^2 t^{2(\beta-1)}}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta(\beta-1)t^{\beta-1}t^{-1}}{(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2 t^{2(\beta-2)}}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta(\beta-1)t^{\beta-1}t^{-1}}{(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2 t^{2(\beta-1)t^{-1}}}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta(\beta-1)t^{\beta-1}}{t(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2 t^{(\beta-1)}t^{\beta}}{t(1+e^{\mu}t^{\beta})^3}.
\end{aligned}$$

A segunda derivada é:

$$
\begin{aligned}
M''_T(t) &= \frac{e^{\mu}\beta(\beta-1)t^{\beta-2}(1+e^{\mu}t^{\beta})}{(1+e^{\mu}t^{\beta})^3} - \frac{2(e^{\mu})^2\beta^2t^{2(\beta-1)}}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta(\beta-1)t^{\beta-1}t^{-1}}{(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2t^{2(\beta-1)}t^{-1}}{(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta(\beta-1)t^{\beta-1}}{t(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2t^{(\beta-1)}t^{\beta}}{t(1+e^{\mu}t^{\beta})^3} \\
&= \frac{e^{\mu}\beta t^(\beta-1)(\beta-1)^2}{t^2(1+e^{\mu}t^{\beta})^2} - \frac{e^{\mu}\beta t^(\beta-1)(\beta-1)}{t^2(1+e^{\mu}t^{\beta})^2} - \frac{4(e^{\mu})^2\beta^2t^(\beta-1)(\beta-1)t^{\beta}}{(t^2(1+e^{\mu}t^{\beta})^3} + \\
&\quad \frac{6(e^{\mu})^3\beta^3t^(\beta-1)(t^{beta})^2}{(1+e^{\mu}t^{\beta})^4t^2} - \frac{2(e^{\mu})^2\beta^3t^(\beta-1)t^{\beta}}{(1+e^{\mu}t^{\beta})^3t^2} + \frac{2(e^{\mu})^2\beta^2t^(\beta-1)t^{\beta}}{(1+e^{\mu}t^{\beta})^3t^2}.
\end{aligned}
$$

Logo, a média da distribuição Log-Logístico é:

$$
E(T) = \frac{e^{\mu}\beta(\beta-1)t^{\beta-1}}{t(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2t^{(\beta-1)}t^{\beta}}{t(1+e^{\mu}t^{\beta})^3}.
$$

Observa-se que a variância é o momento central de ordem 2, a qual obtem-se por meio da diferença entre o momento de ordem 2 e o quadrado do momento de ordem 1, dada por:

$$
\begin{aligned}
V(T) &= E(T^2) - [E(T)]^2 \\
&= M''(0) - [M'(0)]^2 \\
&= \frac{e^{\mu}\beta t^(\beta-1)(\beta-1)^2}{t^2(1+e^{\mu}t^{\beta})^2} - \frac{e^{\mu}\beta t^(\beta-1)(\beta-1)}{t^2(1+e^{\mu}t^{\beta})^2} - \frac{4(e^{\mu})^2\beta^2t^(\beta-1)(\beta-1)t^{\beta}}{(t^2(1+e^{\mu}t^{\beta})^3} + \\
&\quad \frac{6(e^{\mu})^3\beta^3t^(\beta-1)(t^{beta})^2}{(1+e^{\mu}t^{\beta})^4t^2} - \frac{2(e^{\mu})^2\beta^3t^(\beta-1)t^{\beta}}{(1+e^{\mu}t^{\beta})^3t^2} + \frac{2(e^{\mu})^2\beta^2t^(\beta-1)t^{\beta}}{(1+e^{\mu}t^{\beta})^3t^2} - \\
&\quad \left[\frac{e^{\mu}\beta(\beta-1)t^{\beta-1}}{t(1+e^{\mu}t^{\beta})^2} - \frac{2(e^{\mu})^2\beta^2t^{(\beta-1)}t^{\beta}}{t(1+e^{\mu}t^{\beta})^3}\right]^2.
\end{aligned}
$$

3.1.3 Quantil

O quantil é dado pelo inverso da função de distribuição:

$$
q_{t_i} = \frac{(1+e^{\mu}t^{\beta})}{e^{\mu}\beta t^{\beta-1}}.
$$

3.1.4 Verossimilhança

O Método de Máxima Verossimilhança é um método muito usado na estimação de parâmetros em análise de sobrevivência, do qual tem-se interesse no presente estudo.

Suponha que $(t_1, \ldots, t_n)$ é uma amostra aleatória de tempos de sobrevivência provenientes de uma distribuição Log-Logístico com parâmetros μ e β e que é associado a cada t_i . Desta forma, a função de verossimilhança é escrita na forma:

$$L(\mu, \beta|\boldsymbol{t}) = \prod_{i=1}^{n} \frac{e^{\mu}\beta t_i^{\beta-1}}{(1 + e^{\mu}t_i^{\beta})^2} \tag{3.7}$$

Aplicando logaritmo em 3.7, o logaritmo da função de verossimilhança é escrito como:

$$l(\mu, \beta|\boldsymbol{t}) = \mu + log(\beta) + (\beta - 1)log(t) - 2log(1 + e^{\mu}t^{\beta}) \tag{3.8}$$

A Função Escore de T, sendo a derivada de primeira ordem da função de log-verossimilhança, que é denotada por $U(T, \theta)$ é dada por:

$$U(T|\theta) = \frac{dlogL(T|\theta)}{d\theta} \tag{3.9}$$

A partir da Função Score é possível determinar o valor (ou expressão) que torna a verossimilhança máxima. Para encontrarmos o Método de Máxima Verossimilhança basta igualarmos a zero a Função Score.

$$\frac{\partial}{\partial\mu} l(\mu, \beta|\boldsymbol{t}) = -\frac{-1 + e^{\mu}t^{\beta}}{1 + e^{\mu}t^{\beta}} \tag{3.10}$$

$$\frac{\partial}{\partial\beta} l(\mu, \beta|\boldsymbol{t}) = -\frac{-1 - e^{\mu}t^{\beta} - \beta\log(t) + e^{\mu}t^{\beta}\log(t)\beta}{\beta(1 + e^{\mu}t^{\beta})}. \tag{3.11}$$

A função Score foi a causa da mudança do objetivo inicial do trabalho, pois não foi possível obter solução analítica.

3.2 Distribuição *a posteriori* do Modelo Hierárquico Log-Logístico

Supor uma distribuição a *a priori* para os parâmetros do Modelo é sempre uma tarefa difícil e muitas vezes sem nenhum suporte teórico para tal. Neste trabalho em especial, faz-se uso de uma *a priori* Normal para o parâmetro de escala, e half-Cauchy para o parâmetro de forma da distribuição Log-Logístico, parâmetro este ligado a discrepância do Modelo, como apresentado por (GELMAN, 2006).

Desta forma, a modelagem Hierárquica fornece uma abordagem poderosa e flexível para a representação de crenças sobre estruturas estendidas de dados observáveis, e utiliza-se cada vez mais esta metodologia em modelagem e análise estatística.

Construíu-se um modelo hierárquico em dois níveis, sendo que o primeiro nível de hierarquia dos dados refere-se aos parâmetros μ e β a serem estimados pelas distribuições *a priori* Normal e half-Cauchy:

$$\mu \sim N(\tau, \sigma^2), \qquad \mu \in \Re;$$

e

$$\beta \sim HC(\theta), \qquad \beta > 0.$$

ou seja,

$$\pi(\mu) = \frac{1}{\sqrt{2}\pi\sigma^2} \exp\left\{-\frac{(\mu-\tau)^2}{2\sigma^2}\right\}$$

com $\tau \in \mathbb{R}$, $\sigma^2 > 0$, conhecidos; e ainda:

$$\pi(\beta|\theta) = \frac{2\theta}{\pi(t^2+\theta^2)}$$

no qual $\pi(\beta|\theta)$ depende de β (hiperparâmetro), $\theta > 0$.

Assim, no primeiro nível temos a primeira *a posteriori*:

$$\begin{aligned} \pi(\mu, \beta|t) &\propto L(\mu, \beta) * \pi(\mu) * \pi(\beta|\theta) \\ &\propto \frac{(e^{\mu}\beta)^n e^{\mu}\beta \left(\prod_{i=1}^{n} \frac{t_i^{(\beta-1)}}{\left(1+e^{\mu}t_i^{\beta}\right)^2}\right) \sqrt{2} e^{-\left(\frac{(t-\tau)^2}{\sigma^2}\right)}\theta}{\sqrt{\pi\sigma^2}\pi(t^2+\theta^2)} \end{aligned}$$

No segundo nível tem-se o interesse de estimar θ à partir da *posteriori*, onde $\theta \sim Gama(\phi, \lambda)$, ϕ e λ conhecidos, ou seja, tem-se a segunda *posteriori* $\pi(t|\mu, \beta, \tau, \sigma^2, \theta, \lambda, \phi)$, onde μ e β são os parâmetros e τ, σ^2, θ são os hiperparâmetros, os quais não são fixados mas, estima-se a partir de uma distribuição, dada por $\theta \sim Gama(\lambda, \phi)$.

Assim, obtem-se o parâmetro θ considerando-se como parâmetro de forma $\beta > 1$, onde:

$$\pi\left(t|\mu, \beta, \tau, \sigma^2, \theta, \lambda, \phi\right) \propto \frac{\phi^{\lambda} t^{(\phi-1)} e^{(-\frac{\lambda t\sigma^2 + t^2 - 2t\tau + \tau^2}{\sigma^2})} (e^{\mu}\beta)^n e^{\mu}\beta \left(\prod_{i=1}^{n} \frac{t_i^{\beta-1}}{(1+e^{\mu}t_i^{\beta})^2}\right) \sqrt{2}\theta}{\Gamma(\phi)\sqrt{\pi\sigma^2}\pi(t^2+\theta^2)}.$$

no qual não se fixam os valores dos hiperparâmetros τ, σ^2 e θ, mas estimam-se a partir da distribuição Gama,

$$\theta \sim Gama(\lambda, \phi), \qquad \lambda \qquad \text{e} \qquad \phi \qquad \text{conhecidos.}$$

Logo, a distribuição *a posteriori* do Modelo Hierárquico Log-Logístico é dada por:

$$\pi(\mu, \beta|t) \propto \frac{t^{\beta+\phi-2}}{[1 - e^{\mu}t^{\beta}]^2[t^2 + \theta^2]},$$

onde $t, \beta, \theta > 0$, e $\mu \in \mathbb{R}$.

Assim, obtem-se *a priori* $\pi(\theta)$, completando, assim, o segundo nível na hierarquia:

$$\pi(\theta) = \frac{\phi^{\lambda}t^{\phi-1}e^{-\lambda t}}{\Gamma(\phi)}.$$

Destaca-se que $\pi(\theta)$ é conhecido e não depende de nenhum outro hiperparâmetro.

Sessão 4

Resultados e Discussão

Na simulação é utilizado métodos numéricos, os resultados não são tão precisos quanto os resultados teóricos (obtidos por métodos analíticos), mas representam uma boa aproximação. Utilizou-se de simulação *bootstrap* para checar a qualidade dos ajustes para os parâmetros do modelo estudado. A importância da técnica *bootstrap* não está somente em avaliar as estimativas dos parâmetros, mas também em obter boas estimativas dos erros padrão da distribuição gerada pelas estimativas dos parâmetros nas iterações de reamostragem.

4.1 Simulação *Bootstrap*

A técnica *bootstrap* de reamostragem surgiu em meados de 1035. Divulgou-se em 1979 por Bradley Efron, após os avanços computacionais, como uma abordagem alternativa ao cálculo de intervalos de credibilidade, em circunstâncias em que outras técnicas não são aplicáveis. Essa técnica é especialmente útil para lidar com problemas estatísticos que envolvem amostras de tamanho pequeno e/ou estimadores cuja distribuição (exata ou assintótica) ainda não foi obtida.

Realizou-se um estudo de simulação *bootstrap* com o objetivo de analisar as propriedades frequentistas dos procedimentos de estimação do modelo hierárquico Log-Logístico. Para examinar as propriedades frequentistas construíram-se os intervalos de credibilidade (HPD) para todos os parâmetros e calculamos sua probabilidades de cobertura (PC). Os valores dos parâmetros foram escolhidos baseados em um experimento clínico, em que é analisada a eficácia de um tratamento. Ver seção 6.1. Para tal consideramos o conjunto de dados, com 148 observações, dos tempos de vida de pacientes tratados com a droga linezolida. Desta forma, por meio do *software* SAS, consideramos amostras de seis tamanhos diferentes que variam de 30 a 500, as quais reamostramos $B = 1000$ vezes.

Para todas as amostras e reamostras, o modelo Log-Logístico hierárquico foi estimado. Para tal, foram geradas cadeias MCMC de dimensão 11.000 com *burn-in* de 1.000 e salto 10, ou seja, uma cadeia final de tamanho 1.000. As estimativas dos parâmetros μ e β estão dispostos a tabela 1 a seguir.

Obtêm-se os valores do erro quadrático médio e a probabilidade de cobertura (PC) dos intervalos de credibilidade. Para todas as amostras calculamos os intervalos de credibilidade de 95% e verifica-se se continham os respectivos verdadeiros valores dos parâmetros. Os resultados da PC para diferentes tamanhos amostrais apresentam-se na tabela 1.

Tabela 1 – Resultado da simulação bootstrap para o modelo hierárquico Log-Logístico.

Amostra	Parâmetro	Estimativas	EQM	PC
$n = 30$	μ	$-5,9908$	$2,5525$	$0,665$
	β	$1,9058$	$0,4770$	$0,698$
$n = 60$	μ	$-5,8574$	$1,1747$	$0,799$
	β	$1,8144$	$0,1455$	$0,745$
$n = 100$	μ	$-5,8766$	$0,9230$	$0,937$
	β	$1,7593$	$0,2132$	$0,921$
$n = 150$	μ	$-5,8752$	$0,6347$	$0,945$
	β	$1,7063$	$0,1607$	$0,951$
$n = 300$	μ	$-5,8329$	$0,3781$	$0,950$
	β	$1,6361$	$0,0981$	$0,962$
$n = 500$	μ	$-5,8048$	$0,2661$	$0,943$
	β	$1,6168$	$0,0722$	$0,957$

Como esperado, os erros quadráticos médios diminuem conforme o aumento do tamanho amostral. Em contrapartida, as probabilidades de cobertura dos intervalos convergem, com o aumento do tamanho amostral, para o valor nominal de 95%.

4.2 Aplicação dados reais

Nesta seção demostra-se a utilidade do modelo hierárquico Log-Logístico através de um estudo realizado no Hospital Universitário de Maringá nos anos de 2008 a 2012. Para tal, 148 pacientes com enfermidades graves (em geral pacientes submetidos a tratamentos paleativos), foram observados durante tratamento com a droga Linezolida.

Desta forma, tem-se o interesse em observar o tempo até a cura e/ou morte dos pacientes submetidos ao tratamento, ou seja, a variável de interesse é o tempo t. Portanto, pacientes internados na UTI do HU que não tiveram uma resposta positiva ao tratamento com antibióticos tradicionamente usados, e, assim, foram submetidos ao tratamento com a nova droga e seus tempos de resposta, a este novo método, foram anotados.

Inicialmente, fez-se uma análise descritiva das covariáveis estudadas: sexo e idade. Observa-se que $59,4\%$ dos pacientes são do sexo masculino contra $40,6\%$ do sexo feminino (ver figura 4) e, o tempo médio observado de tratamento para pessoas do sexo feminino foi de $26,4$ dias contra $29,8$ dias para os homens. Observa-se a idade dos pacientes com médias de 51 e 53 anos, respectivamente, para pacientes dos sexos feminino e masculino.

Destacam-se os *outlines* observados entre os pacientes do sexo feminino e os pacientes do sexo masculino, ou seja, $3,4\%$, para ambos os sexos, sobreviveram mais de $36,3$ dias.

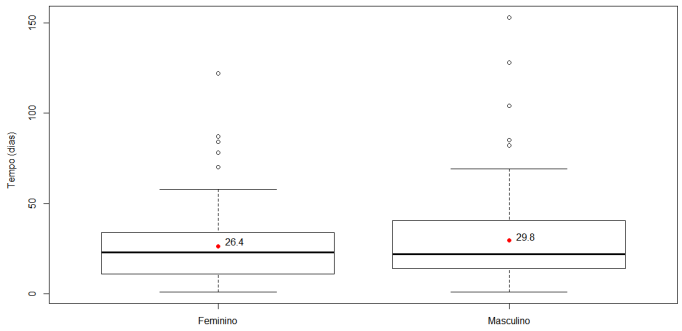

Figura 4 – Distribuição dos tempos observados por sexo.

O tempo médio de internação dos pacientes tratados com Linezolida na UTI foi de $28,4$ dias, com mediana igual a $22,5$ dias (o primeiro e o terceiro quartis iguais a $13,0$ e $36,3$ dias, respectivamente). Ainda com relação ao tempo sob tratamento, o mínimo observado foi 1 dia e o tempo máximo 153 dias, com um coeficiente de variação de $88,4\%$. Ver figura 5.

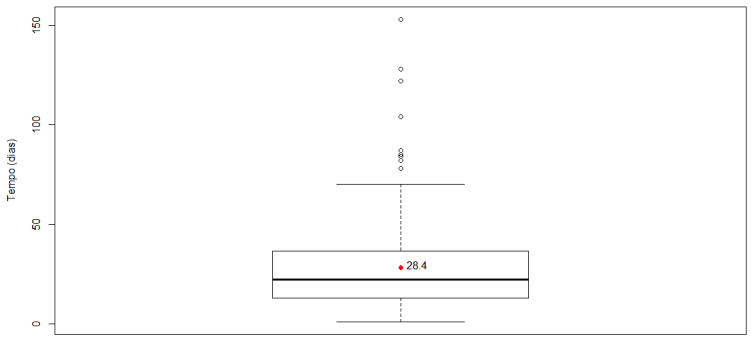

Figura 5 – Distribuição dos tempos observados (em dias).

Com relação à censura, $48,4\%$ foram a óbito e, se considerarmos o gênero dos

pacientes, $57,3\%$ eram do sexo masculino e $42,7\%$ do sexo feminino. A idade média dos pacientes que foram a óbito foi de 58 anos com um coeficiente de variação de $27,0\%$ contra $46,4$ anos para os pacientes que experimentaram a cura (coeficiente de variação de $37,2\%$).

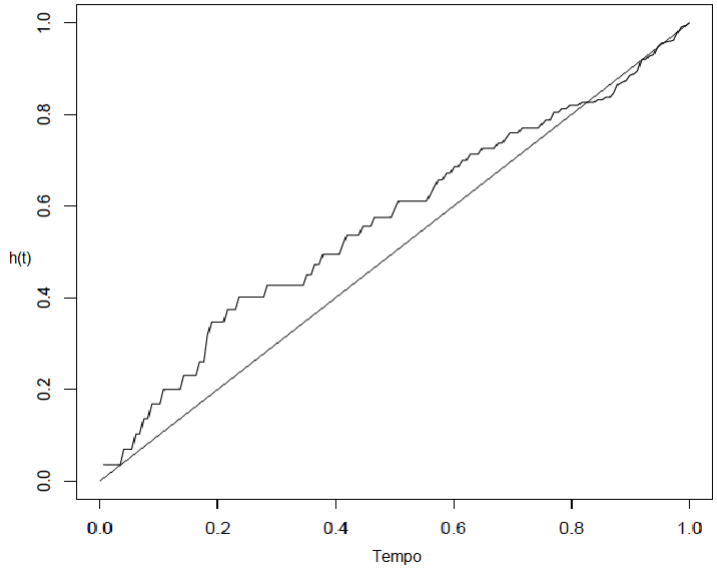

Figura 6 – TTT-Plot da função de risco dos tempos observados.

Na figura 6, apresentamos o TTT-Plot através do qual podemos observar um possível comportamento unimodal da taxa de falha, indicando o modelo Log-Logístico como uma boa alternativa ao ajuste dos tempos. Para estimação dos parâmetros do modelo hierárquico Log-Logístico consideramos o método MCMC onde três cadeias foram geradas de tamanho 11.000, com *burn-in* de 1.000 e salto de 10, resultando em cadeias de tamanho 1.000. A tabela 2 apresenta as estimativas dos parâmetros e hiperparâmetros do modelo proposto e seus respectivos intervalos de credibilidade 95% estão dispostos na tabela 3.

Tabela 2 – Médias *a posteriori* dos parâmetros do modelo hierárquico Log-Logístico.

Parâmetros	Média	Desvio Padrão	Percentis		
			25%	50%	75%
μ	-5.5879	0.419	-5.8768	-5.5812	-5,3012
β	1.8418	0.1269	1,7573	1,8411	1,9237
θ	101.3	98.95	31.2996	67.6162	142.6

Tabela 3 – Intervalos de credibilidade de 95%.

Parâmetros	Intervalos		Intervalos HPD	
μ	-6,426	-4,7726	-6,4659	-4,8235
β	1,5929	2,0944	1,5954	2,096
θ	3,4091	362,5	0,0905	296,6

A convergência das cadeias podem ser visualizadas por meio das figuras 7 que segue.

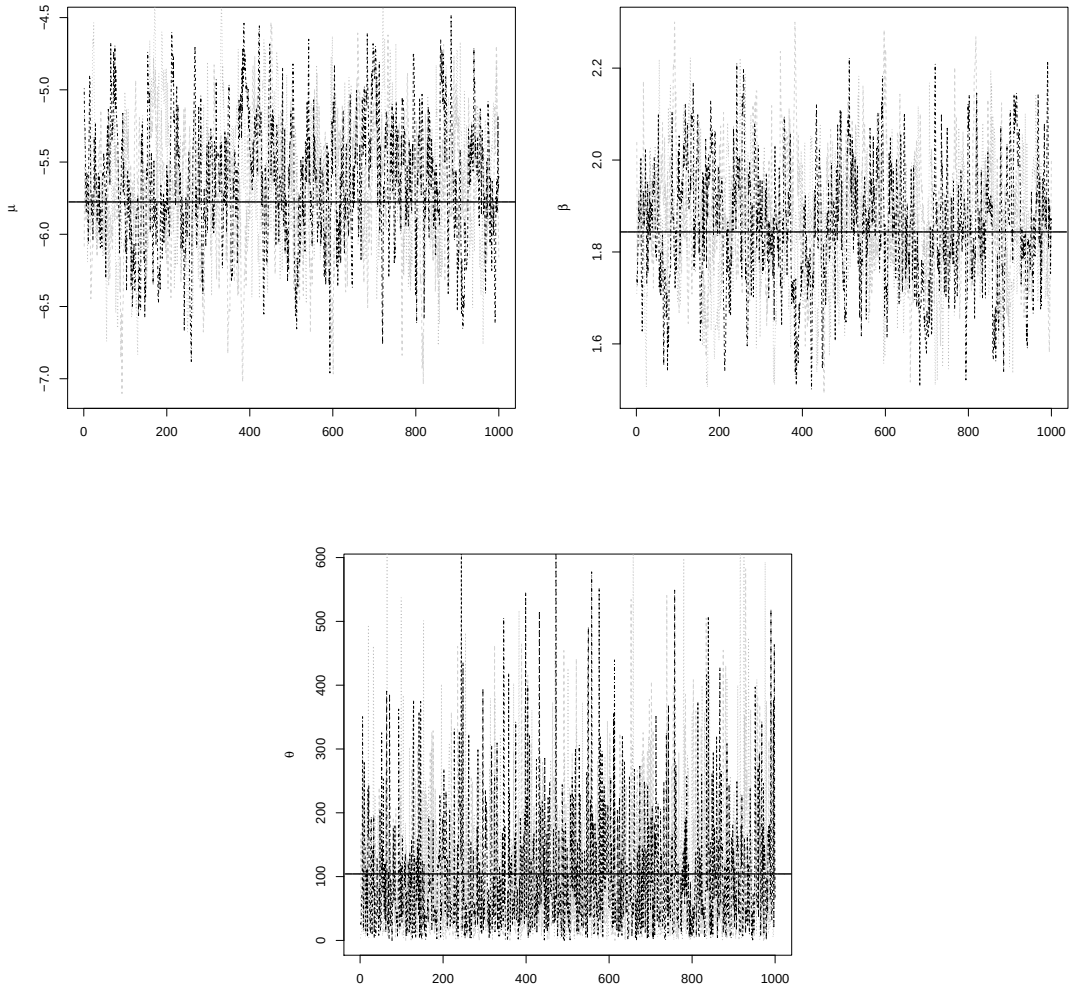

Figura 7 – Convergência dos parâmetros μ, β e θ.

Além disso, as figuras 8 apresentam a distribuição marginal *a posteriori* dos parâmetros e hiperparâmetros μ, β e θ, respectivamente.

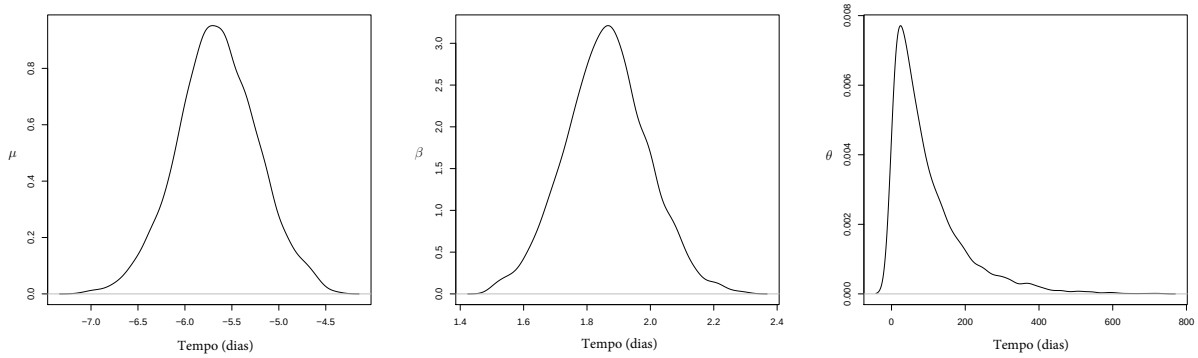

Figura 8 – Distribuição a posteriori marginal dos parâmetros μ, β e θ.

O ajuste do modelo Log-Logístico, com dados completos, pode ser visualizado nas figuras 9 e 10 a seguir. Na figura 11 observa-se o comportamento da taxa de falha estimada pelo modelo, bem como o tempo mais provável de cura do paciente tratado com a droga.

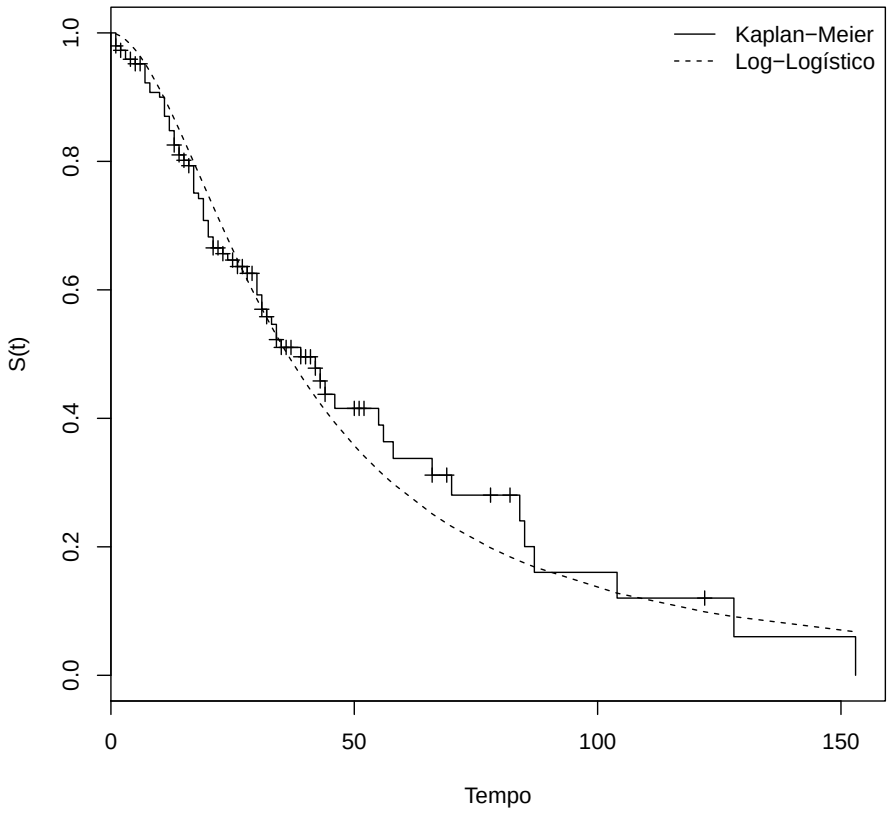

Figura 9 – A função de sobrevivência (empírica versus a estimada pelo modelo).

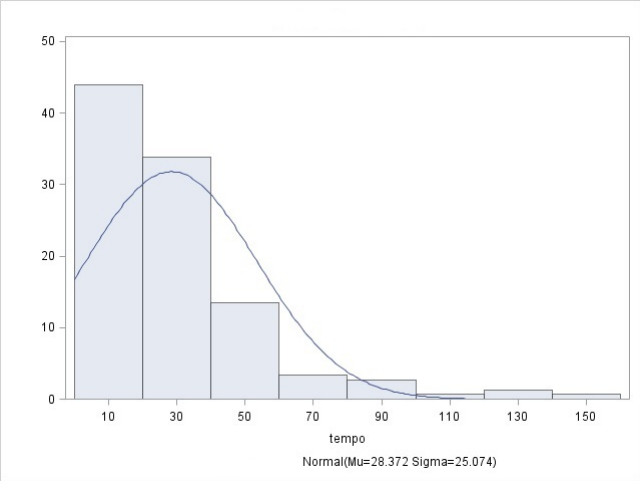

Figura 10 – Histograma dos tempos juntamente com a densidade estimada pelo chute inicial da distribuição Normal.

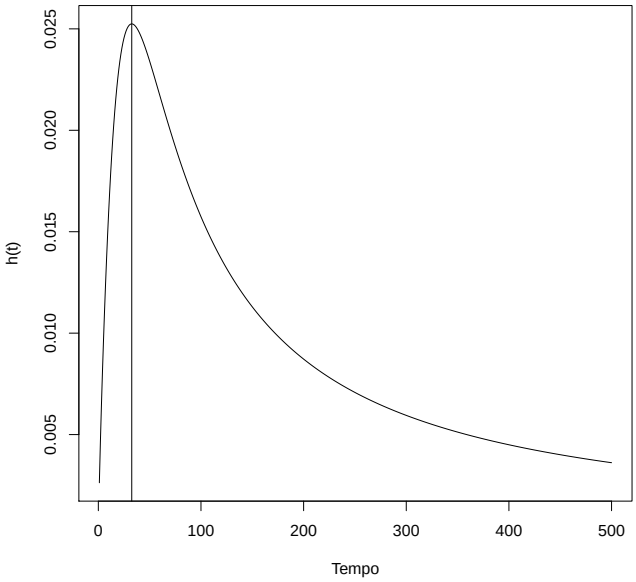

Figura 11 – Curva de risco juntamente com o tempo máximo estimado.

Verifica-se se o modelo Log-Logistico tem um bom ajuste por meio de análise de resíduos e influência. Assim, a primeira ferramenta para acessar a sensibilidade das medidas é a partir de uma análise de influência global. Para isso, faz-se um estudo de caso deleção, no qual mede-se o efeito de cada observação nas estimativas. Utilizou-se como primeira medida a distância generalizada de Cook.

As figuras 12 e 13 apresentam, respectivamente, as distâncias de Cook e verossimilhanças.

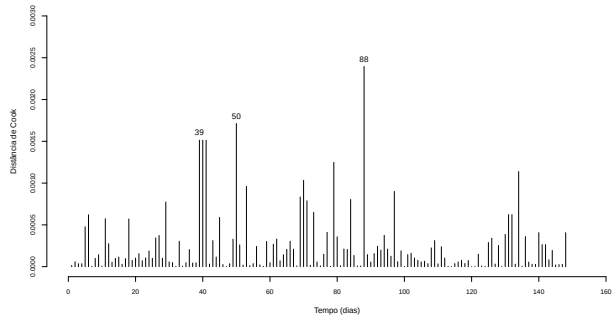

Figura 12 – Distância de Cook global.

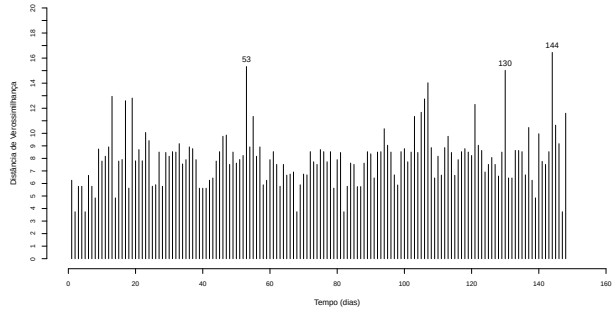

Figura 13 – Distância de verossimilhanca global.

Após esta análise global, alguns pontos foram identificados como possíveis pontos influentes. Realizou-se uma análise de influência local com o intuito de verificar a possível influência dos pontos identificados. Para tal, considerou-se uma perturbação na variável resposta t, tempo, e obtiveram-se novamente as distância de Cook e verossimilhanças, representadas nas figuras 14 e 15, respectivamente.

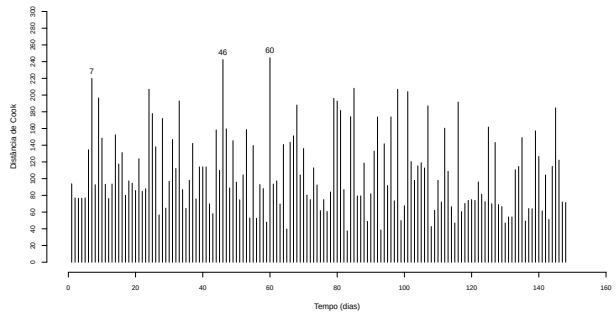

Figura 14 – Distância de Cook local.

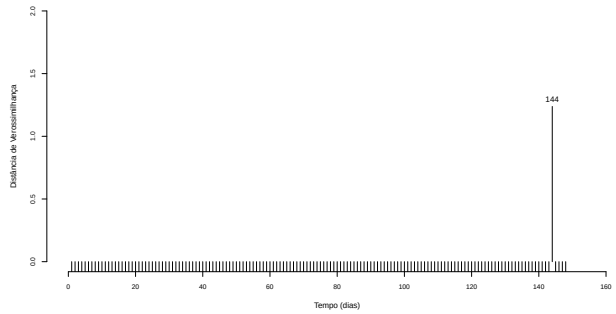

Figura 15 – Distância de verossimilhança local.

Assim, utilizando-se de análise de diagnóstico (influência global e local) e análise de resíduos (*deviance* e *martingale* consideram-se os pontos discriminados como possíveis pontos influentes. E, com o objetivo de analisar o impacto dessas observações nas estimativas dos parâmetros, realizamos uma reanálise dos dados reajustando o modelo sob algumas situações. Primeiramente, reajustamos o modelo eliminando individualmente cada observação. Depois, reajustamos o modelo eliminando duas observações combinadas e, finalmente, eliminamos todos esses pontos conjuntamente.

Destaca-se que a observação $n = 144$ foi indicada como ponto influente na distância de verossimilhança tanto na distância de *Cook*.

Após a análise das estimativas dos parâmetros e hiperparâmetros de todos os possíveis pontos influentes, detectaram-se como pontos influentes $\#46$, $\#50$ e $\#144$.

Na tabela 4 realizamos uma análise do impacto do ponto influente identificado e retirado do modelo em estudo, com o intuito de verificar a possível influência deste nas estimativas. Para isso, apresentamos os valores de RC, MRC, TRC e LD para cada conjunto I de parâmetros removido.

Tabela 4 – Impacto dos pontos influentes identificados e retirados do modelo Log-Logístico.

Ponto removido	Parametro	RC	TRC	MRC	$LD_{(I)}$
$\{46\}$	μ	0.1080	0.1743	0.1080	8.9709
	β	0.0663			
$\{50\}$	μ	0.1121	0.1842	0.1121	8.5865
	β	0.0721			
$\{144\}$	μ	0.2010	0.2736	0.2010	17.1848
	β	0.0726			
$\{46\},\{50\}$	μ	0.2230	0.3770	0.2230	17.1848
	β	0.1540			
$\{46\},\{144\}$	μ	0.2510	0.3740	0.2510	26.2024
	β	0.1230			
$\{50\},\{144\}$	μ	0.1990	0.3340	0.1990	25.8400
	β	0.1350			
$\{46\},\{50\},\{144\}$	μ	0.3120	0.5300	0.3120	34.5536
	β	0.2180			

Analisando o conjunto de dados, observa-se que a observação $\#50$ foi a que obteve o menor tempo de sobrevivência sobre o eixo longitudinal, $\#2$ dias. A observação $\#46$ se refere ao indivíduo que sobreviveu 28 dias. E, a observação $\#144$ se refere ao indivíduo que teve o maior tempo de sobrevivência observado que foi de $\#153$ dias. Sendo o nível de influência destes valores conjuntamente, mudança relativa total (TRC) é de 0.5300.

4.2.1 Análise de resíduos

Análise de resíduos tem a finalidade de avaliar a adequação da distribuição proposta para a variável resposta. Segundo (COLOSIMO E. A. E GIOLO, 2006), os gráficos de resíduos *martingale*, ou *deviance*, versus os tempos fornecem uma forma de verificar a adequação do modelo ajustado, bem como auxiliam na detecção de observações atípicas.

Neste trabalho, os resíduos utilizados foram: *martingale* e *deviance*. Os resíduos *martingale* foram inicialmente introduzidos para processos de contagem e depois utilizados para modelos de regressão paramétricos, sobrevivência, etc. Este resíduo é definidos por:

$$\hat{m}_i = \delta_i - \hat{e}_i, \tag{4.1}$$

em que δ_i é a variável indicadora de falha e $\hat{e}_i$ são os resíduos de Cox-Snell.

Já os resíduos *deviance* são uma tentativa de tornar os resíduos *martingale* mais simétricos em torno de zero, o que em geral, facilita a detecção de pontos atípicos. São definidos por:

$$\hat{d}_i = \textbf{sinal}(\hat{m}_i)\left[-2(\hat{m}_i) + \delta_i \log(\delta_i - \hat{m}_i)\right]^{\frac{1}{2}}. \tag{4.2}$$

Destacamos que, se o modelo for apropriado, esses resíduos devem apresentar-se com um comportamento aleatório em torno de zero, como o caso do modelo proposto. Ver figuras 16 e 17 a seguir.

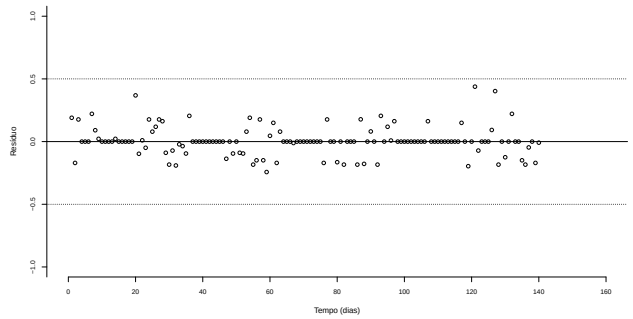

Figura 16 – Resíduos deviance modelo ajustado Log-Logístico.

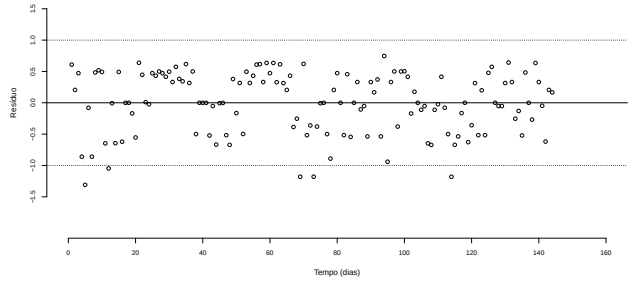

Figura 17 – Resíduos martingale modelo ajustado Log-Logístico.

4.3 Reanálise dos Dados

Após a análise de influência e resíduos, os possíveis pontos influentes identificados foram $\#46$, $\#50$, $\#144$ e retirados individualmente, e combinados entre si, como apresentado na tabela 4. Desta forma, o modelo hierárquico Log-Logístico foi reestimado e as médias *a posteriori* foram obtidas usando método MCMC. Assim sendo, 3 cadeias de tamanho 11.000 com *burn-in* de 1.000 e salto de 10 foram consideradas, gerando cadeias finais de tamanho 1.000. Os resultados estão dispostos nas tabelas 5 e 6 a seguir.

Tabela 5 – Médias a *a posteriori* dos parâmetros do modelo hierárquico Log-Logístico após reanálise.

Parâmetros	Estimativas	Erro Padrão	Intervalo de Confiança 95%	
			Inferior	Superior
μ	-5.9754	0.4266	-6.7918	-5.1375
β	1.7199	0.1378	1.4740	2.0141
θ	207.1	207.8000	5.4798	743.4000

Tabela 6 – Intervalos de credibilidade de 95%.

Parâmetros	Intervalo HPD	
	Inferior	Superior
μ	-6.8462	-5.2062
β	1.4658	2.0027
θ	0.1406	589.5000

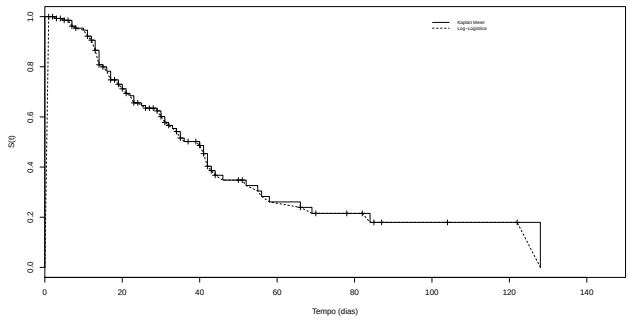

Figura 18 – Função de sobrevivência do modelo Log-Logístico conjuntamente com a curva de Kaplan-Meier.

Após a exclusão dos pontos considerados influentes $\#46$, $\#50$ e $\#144$, observamos que a maior mudança relativa total foi após a exclusão do ponto $\#144$, com $TRC = 0.1028$ e a menor mudança relativa total foi do ponto 46, com $TRC = 0.0664$. Consideramos que foi válido retirar os pontos considerados influentes, pois obtivemos estimativas menores para o erro padrão.

Observamos que o comportamento da taxa de falha estimada pelo modelo Log-Logístico apresentou-se adequada, como podemos observar na figura 18.

Os resultados apresentados nas tabelas 5 e 6 são as estimativas de máxima verossimilhança, na qual observamos que houve diminuição do erro padrão. Verificou-se a qualidade do ajuste por meio da figura 18 no qual plotou-se a função de sobrevivência de kaplan-Meier e a função estimada pelo modelo Log-Logístico, a qual mostra que o modelo está bem ajustado.

Sessão 5

Conclusão

Os resultados obtidos neste trabalho, com dados reais do tempo de sobrevivência após o internamento de pacientes na UTI, no Hospital Universitário de Maringá, sob tratamento com a substância ativa Linezolida nos levaram à conclusão de que o modelo Log-Logístico proposto apresenta vantagens em relação aos modelos frequentemente utilizados, como, por exemplo, o modelo Weibull. Sua utilização é, estatisticamente, de mais fácil aplicação e interpretação e, também, uma atraente alternativa já que os dados apresentam taxa de falha unimodal.

Para análise, fez-se o uso de metodologia bayesiana por meio de *prioris* hierárquicas. Para estimação dos parâmetros do modelo geraram-se três cadeias utilizando-se do método MCMC e, verificou-se a convergência destas através de análise gráfica e do teste de Gelman e Rubin. A qual trouxe resultados positivos indicando que seu uso traria benefícios aos pacientes, pois o tempo médio de uso do medicamento foi de $28,4$ dias, com mediana de $22,5$ dias e o tempo de internamento mais observado foi de $73,0$ dias. Pode-se observar por meio da curva de risco que o tempo estimado do paciente vir a óbito é de $32,5$ dias. Após este período, à medida que o tempo aumenta o risco decresce gradativamente.

Os resultados obtidos na reanálise dos dados nos mostraram que o modelo Log-Logístico proposto é bastante coerente com a teoria desenvolvida neste trabalho e, assim, foi relevante a análise de influência de alguns pontos detectados, demonstrando a sensibilidade do modelo a possíveis observações influentes.

E, finalmente, para verificação da qualidade de ajuste do modelo proposto, plotamos a função de sobrevivência de Kaplan-Meier e a função estimada pela distribuição Log-Logístico, a qual nos mostrou que o modelo está bem ajustado para modelar dados de tempos de vida, pois diminuiu significativamente o valor do erro quadrático médio.

5.1 Perspectivas Futuras

Dentre o leque de perspectivas futuras, destaca-se que nosso intuito é o de trabalhar com outros modelos. Também, trabalhar com conjunto de dados maior e fazer um estudo longitudinal, pois este é um método de pesquisa que visa analisar as variações nas caraterísticas dos mesmos elementos amostrais (máquinas, peças ou indivíduos) ao longo de um considerável período de tempo, frequentemente vários anos. Pois situações deste tipo são muito usadas na produção industrial e, em especial, na Medicina, no estudo de grande relevância, como é o caso da epidemiologia de doenças.

Referências

BERGER, J. O. *Teoria da Decisisão Estatística e Análise Bayesiana*. Berlin: Springer, 1985.

BERNARDO J. M E SMITH, A. F. M. *Bayesian Theory*. New York: John Wiley and Sons, 1994.

BOX G. E. P. E TIAO, G. C. *Inferência Bayesiana na análise estatística*. Addison-Wesley: Reading, 1973.

BOX G. E. P. E TIAO, G. C. *Bayesian inference in statistical analysis*. New York: Wiley, 1992. 588 p.

CARDOSO F. F. E ROSA, G. J. d. M. e. T. J. R. A. d. A. Modelos hierárquicos bayesiano para estimação robusta e análise de dados censurados em melhoramento animal. *Revista Brasileira de Zootecnia*, v. 38, p. 72–80, 2009.

COLOSIMO E. A. E GIOLO, S. R. *Análise de Sobrevivência aplicada. ABE - Projeto Fisher*. São Paulo: Editora Edgard Blucher, 2006. 408 p.

COOK, R. D. Assessment of Local Influence. *Journal of the Royal Statistical Society*, v. 48, p. 133–169, 1986.

FABRI F. V., e. a. Avaliação do consumo de antimicrobianos em hospital privado do norte do paraná. *In: III Encontro Paranaense de microbiologia, 2012, Londrina. Anais do III encontro paranaense de microbiologia*, n. v. CDroom, 2012.

FEIGL P. E ZELEN, M. *Estimation of exponencial survival probabilities with concomitant information.* R: Biometrics, 1965. v. 21. 826–838 p.

GELMAN, A. Prior distributions for variance parameters in hierarchical models. *Communications in Statistics. Theory and Methods*, v. 1, n. 3, p. 515–533, 2006.

HASTINGS, W. K. Monte Carlo Sampling methods using Markov chains and their application. *Biometrika*, v. 57, p. 97–109, 1970.

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

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. New York: John Wiley and Sons, 1982.

LEE E. E WANG, J. W. *Statistical Methods for Survival Data analysis*. Wiley: Interscience, 2003.

METROPOLIS N. E ROSEMBLUT, A. e. R. M. N. e. T. A. H. e. T. E. The 1988 Wald Memorial Lectures, The Present Position in Bayesian Statistics. *Journal of Chemical Physics*, v. 21, p. 1987–1092, 1953.

MURTEIRA, B. *Probabilidades e Estatística*. Lisboa: Editora McGraw-Hill, 1990. v. 1.

MURTEIRA B. E TURKMAN, M. A. A. e. P. C. D. *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 2003. v. 1.

PAULINO D. E TURKMAN, A. e. M. B. *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 2003.

TOMAZELLA V. L. D., G. V. G. e. L. F. Bayesian reference analysis for the Poisson-exponential lifetime distribution. *Chilean Journal of Statistics*, n. 4, p. 99–113, 2013.