Modelo de Mistura Padrão com Tempos de Vida Exponenciais Ponderados

Bruno Pauka Gouveia

UFSCar - São Carlos/SP

Março/2010

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

Modelo de Mistura Padrão com Tempos de Vida Exponenciais Ponderados

Bruno Pauka Gouveia

Prof. Dr. Josemar Rodrigues

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

UFSCar - São Carlos/SP
Março/2010

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

G719mm

Gouveia, Bruno Pauka.
Modelo de mistura padrão com tempos de vida exponenciais ponderados / Bruno Pauka Gouveia. -- São Carlos : UFSCar, 2010.
60 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2010.

1. Análise de sobrevivência. 2. Fração de cura. 3. Software – GAMLSS - estatística. I. Título.

CDD: 519.9 (20ª)

Bruno Pauka Gouveia

Modelo de mistura padrão com tempos de vida exponenciais ponderados

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 05 de março de 2010.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Josemar Rodrigues (DEs-UFSCar/ Orientador)

1° Examinador _____
Prof. Dr. Mário de Castro Andrade Filho (ICMC-USP)

2° Examinador _____
Profa. Dra. Reiko Aoki (ICMC-USP)

Agradecimentos

Agradeço ao Prof. Josemar pela orientação e ao Prof. Mário de Castro pela ajuda nos mistérios do GAMLSS.

Agradeço à minha família, aos meus amigos pelo apoio e aqueles que direta ou indiretamente contribuíram para este trabalho

Finalmente, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio concedido para este trabalho.

Resumo

Neste trabalho apresentamos brevemente os conceitos que definem a análise de sobrevivência de longa duração. Dedicamo-nos exclusivamente ao modelo de mistura padrão de Boag (1949) e Berkson & Gage (1952), sendo que nos preocupamos com sua formulação, apresentamos a função probabilidade de imunes, que é derivada do próprio modelo e investigamos a questão da identificabilidade. Motivados pela possibilidade de que um planejamento experimental leve a uma seleção viciada da amostra, estudamos as distribuições ponderadas de probabilidade, mais especificamente a família das distribuições exponenciais ponderadas e suas propriedades. Estudamos duas distribuições pertencentes a essa família, a distribuição exponencial *length biased* e a distribuição beta exponencial. Fazendo uso do pacote GAMLSS em R, realizamos alguns estudos de simulação com o intuito de evidenciar o erro cometido quando se ignora a possibilidade de que a amostra seja proveniente de uma distribuição ponderada.

Palavras-chave: Modelos de longa duração, Modelo de mistura padrão, Distribuições de probabilidade ponderadas, Distribuição exponencial *length biased*, Distribuição beta exponencial, GAMLSS.

Abstract

In this work, we briefly introduce the concepts of long-term survival analysis. We dedicated ourselves exclusively to the standard mixture cure model from Boag (1949) and Berkson & Gage (1952), showing its deduction and presenting the imunes probability function, which is taken from the model itself and we investigated the identifiability issues of the mixture model. Motivated by the possibility that a experiment design can lead to a biased sample selection, we studied the weighted probability distributions, more specifically the weighted exponential distributions family and its properties. We studied two distributions that belong to this family; namely, the length biased exponential distribution and the beta exponential distribution. Using the GAMLSS package in R, we made some simulation studies intending to evidence the bias that occur when the possibility of a weighted sample is ignored.

Keywords: Long-term models, Standard mixture cure models, Weighted probability distributions, Exponential distribution, Length biased exponential distribution, Beta exponential distribution, GAMLSS.

Sumário

Capítulo 1

Introdução

Modelos para dados de sobrevivência com fração de cura vêm ganhando destaque em análise de sobrevivência e confiabilidade. Esses modelos admitem que na população observada existem elementos que não são suscetíveis ao evento de interesse. A proporção desses elementos *imunes* chamamos de fração de cura.

Numa situação menos geral, consideremos um grupo de pessoas com uma determinada doença e que foram submetidas a um certo tratamento. Nesse caso o evento de interesse é a morte de um paciente devido à doença em questão, mas será que alguns indivíduos nesse grupo não podem estar curados?

É baseado em situações como essa que Boag (1949), Berkson & Gage (1952), Yakovlev & Tsodikov (1996), Chen et al. (1999), Rodrigues et al. (2009) e outros propuseram modelos cada vez mais complexos e com o intuito de explicar melhor o mecanismo biológico envolvido.

Nesse cenário em que existem imunes na população de doentes, ao realizarmos a amostragem para um experimento podemos cometer um erro de planejamento ao considerar que a amostra obtida tem as mesmas características da população da qual ela se originou. No caso clínico podemos dizer que pacientes com tempos de vida mais longos terão mais chance de serem incluídos na amostra. Isso sugere que amostra obtida não é uma amostra aleatória simples, mas sim uma amostra viciada. Esse fato é que nos motivou a introduzir o conceito das variáveis aleatórias ponderadas no âmbito dos modelos de longa duração.

Variáveis aleatórias ponderadas surgem naturalmente quando a própria população impõe restrições, normalmente desconhecidas, que impedem uma amostragem casual simples. Introduzimos uma função de peso para contornar essas restrições. Dessa forma, o mecanismo de seleção transforma uma variável aleatória na sua versão ponderada. As variáveis ponderadas podem também ser provenientes de um planejamento experimental deficiente. Navarro et al. (2001) apresentam um exemplo em que se pretendia estimar o tempo médio de permanência de turistas no Marrocos. Nesse estudo foram selecionados turistas em hotéis e turistas que passavam por postos nas fronteiras. Percebeu-se que o tempo médio dos turistas que ficavam nos hotéis era duas vezes maior do que os que eram abordados nas fronteiras. Utilizamos uma variável ponderada na modelagem a fim de levar em consideração esse vício. Desse modo, pode-se fazer um planejamento contando com essa possibilidade, tirando proveito de uma amostra ponderada.

Bayarri & DeGroot (1992) apresentam um exemplo em que se desejava estudar a população de criminosos de um país. Seria muito caro e talvez impossível obter uma amostra aleatória dessa população. Então, considerou-se preferível estudar a população de criminosos já presos, resultando numa amostra ponderada. Cnann (1985) diz que um problema comum em estudos com humanos é que os indivíduos são observados após o diagnóstico da doença, que é quando eles entram no estudo. Desse modo, os indivíduos com menor tempo de vida, aqueles que morreram antes de iniciar o estudo, nunca são observados. Nessa situação a amostragem é viciada, proveniente de uma variável aleatória ponderada.

A proposta desse trabalho foi de juntar esses dois tópicos, construir o modelo de mistura padrão com uma distribuição exponencial ponderada para o tempo de vida dos indivíduos em risco e verificar o impacto nas estimativas quando admitimos que o tempo de vida segue uma distribuição exponencial simples, mas a amostra vem de uma variável ponderada.

1.1 Organização dos Capítulos

No capítulo 2 apresentamos uma breve revisão dos conceitos da análise de sobrevivência de longa duração e exploramos mais detalhadamente a formulação do modelo

de mistura padrão. Confrontamos algumas vantagens e desvantagens desse modelo, abordamos a questão da identificabilidade e uma função que retrata a probabilidade de cura dos indivíduos com relação ao tempo.

No capítulo 3 apresentamos a definição das variáveis aleatórias ponderadas enfatizando a família das distribuições exponenciais ponderadas e algumas de suas propriedades. Dedicamo-nos a duas distribuições dessa família, a saber, a distribuição exponencial *length biased* e a distribuição beta exponencial.

No capítulo 4 construímos os modelos de mistura padrão exponenciais ponderados ao admitirmos as distribuições vistas no capítulo 3 para o tempo de vida. Utilizamos o pacote GAMLSS em R (Rigby & Stasinopoulos, 2007) para realizar algumas simulações. O objetivo destas é mostrar alguns exemplos de como se comportam os estimadores dos parâmetros dos modelos e comparar os resultados quando ignoramos a ponderação na amostra.

Por último, temos um apêndice sobre o pacote GAMLSS. Explicamos brevemente os modelos GAMLSS e como funciona o pacote desenvolvido em linguagem R para ajustar esses modelos.

Capítulo 2

Análise de Sobrevivência de Longa Duração

A análise de sobrevivência é uma área da Estatística amplamente desenvolvida. Nela estamos interessados na previsão de tempos de vida de indivíduos sob um determinado risco de falha, na comparação entre dois tratamentos e na identificação de fatores de risco de uma certa doença, entre outros.

Diversos estudos em análise de sobrevivência são efetuados utilizando modelos estatísticos. Modelos estes que podem ser de natureza paramétrica, não-paramétrica ou semi-paramétrica. Os modelos não-paramétricos dependem exclusivamente dos dados. Já nos modelos paramétricos impõe-se que os dados seguem uma distribuição de probabilidade que envolvem parâmetros a serem estimados.

Neste texto trataremos de modelos paramétricos, pois deles podem ser construídas as funções de verossimilhança com as quais podemos aplicar tanto a teoria frequentista como a teoria bayesiana.

Os conjuntos de dados característicos da análise de sobrevivência são compostos por tempos de vida e informações pertinentes a cada indivíduo. Mas nem sempre esse tempo de vida é observado, já que existe a possibilidade de pacientes desistirem do experimento. Casos como este, ou quando o paciente morre de alguma causa que não é a de interesse do estudo, caracterizam observações censuradas. É importante enfatizar que, mesmo sendo incompletas, essas observações censuradas trazem alguma informação

e não devem ser descartadas.

Os mecanismos de censura mais comuns são:

- **Censura tipo 1.** A censura do tipo 1 acontece quando no planejamento, o tempo de vida limite é estabelecido. Assim, as observações com tempos de vida maiores do que ou iguais a esse tempo limite serão classificadas como censuradas.
- **Censura tipo 2.** A censura do tipo 2 acontece quando a duração do estudo é determinada por um número estipulado de falhas. Assim, atingido esse número de falhas, os indivíduos restantes terão seus tempos de vida censurados.
- **Censura aleatória.** Esse tipo de censura acontece quando o acompanhamento do indivíduo é interrompido por uma causa que não seja a de interesse. Por exemplo, o paciente abandona o tratamento, o paciente morre devido a uma outra causa ou ele é removido para um tratamento intensivo.

Independentemente de qual tipo seja, a censura pode ainda ser classificada como informativa ou não-informativa. A censura não-informativa é independente do risco em estudo. Já na censura informativa a perda do indivíduo está associada ao risco em estudo, como por exemplo o indivíduo ser retirado do tratamento devido a complicações de saúde.

Desse modo, os dados na análise de sobrevivência podem ser escritos como um par (t_i, δ_i), em que t_i corresponde ao tempo de vida ou ao tempo de censura e δ_i é o indicador de censura, tal que

$$\delta_i = \begin{cases} 1, \text{ se } t_i \text{ é um tempo de falha,} \\ 0, \text{ se } t_i \text{ é um tempo de censura,} \end{cases}$$

$i = 1, 2, ..., n$.

Seja T uma variável aleatória não negativa, usualmente contínua, representando o tempo de vida e $f(t)$ a função de densidade dessa variável, tal que

$$\int_0^\infty f(u)du = 1.$$

A função de distribuição da variável T é definida como

$$F(t) = P[T \leq t] = \int_0^t f(u)du.$$

Podemos especificar a variável aleatória T também através das funções de sobrevivência e de risco.

Definição 2.1 A função de sobrevivência $S(t)$ é definida como sendo a probabilidade de que o tempo de vida T de um indivíduo seja maior do que um instante de tempo t. Ou seja, é a probabilidade de um indivíduo sobreviver até um tempo t determinado, dada por

$$S(t) = P[T > t] = \int_t^\infty f(u)du = 1 - F(t).$$

A função de sobrevivência apresenta as seguintes propriedades:

- $S(0) = 1$,
- $\lim_{t\to\infty} S(t) = 0$,
- $S(t)$ é decrescente.

As funções de sobrevivência e de densidade se relacionam por meio de

$$f(t) = -\frac{d}{dt}S(t). \tag{2.1}$$

Definição 2.2 A função de risco é definida como

$$h(t) = \lim_{\Delta t\to 0} \frac{P[t \leq T \leq t + \Delta t | T \geq t]}{\Delta t}.$$

É interpretada como uma taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t.

Colosimo & Giolo (2006) afirmam que a função de risco dá mais informações sobre o problema do que a função de sobrevivência, pois existem casos em que diferentes funções de sobrevivência podem ter formas semelhantes, enquanto que as respectivas funções de risco podem diferir completamente, destacando a importância da função de risco na modelagem.

A função de risco também pode ser expressa como

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\log S(t). \tag{2.2}$$

Uma forma de estimar a curva da sobrevivência foi proposta por Kaplan & Meier (1958), o chamado estimador produto limite ou estimador de Kaplan-Meier (EKM). Este método, não-paramétrico, consiste em uma adaptação da função de sobrevivência empírica, que considera tantos intervalos de tempo quantos forem o número de tempos de falha distintos. Os limites dos intervalos de tempo são os tempos de falha da amostra.

Considere

- $t_{(1)} < t_{(2)} < ... < t_{(k)}$, os k tempos de falha distintos e ordenados,
- d_j o número de falhas em $t_{(j)}$, $j = 1, 2, ..., k$,
- n_j o número de indivíduos sob risco em $t_{(j)}$, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a $t_{(j)}$

Definimos o estimador de Kaplan-Meier como

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} \frac{n_j - d_j}{n_j} = \prod_{j:t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

O estimador de Kaplan-Meier inclui censuras, elas afetam as contagens n_j. Cada termo do produto é visto como uma estimativa da probabilidade condicional de se estar vivo no instante $t_{(j)}$ dado que se sobreviveu até o instante $t_{(j-1)}$.

Vejamos um exemplo de uma curva de sobrevivência estimada pelo EKM. O conjunto de dados, presente em Kersey et al. (1987), é referente aos tempos (em anos) de recorrência de leucemia após dois tipos de transplantes, alogênico (Grupo 1) e autogênico (Grupo 2). O conjunto de dados é composto por 46 indivíduos no Grupo 1, com 13 observações censurados e 44 indivíduos no Grupo 2, com 9 observações censuradas. Os dados estão reproduzidos em Maller & Zhou (1996, p. 92). A figura 2.1 mostra a curva de sobrevivência estimada usando o EKM para os dois grupos.

Observando o gráfico notamos que após um certo tempo não ocorrem mais mortes e a curva se estabiliza num patamar. Isso sugere que os indivíduos censurados ao final do estudo possam ser imunes ao risco em questão ou então que eles foram curados durante o processo. Tendo em vista esse fenômeno precisamos de um modelo estatístico que comporte essa possibilidade de cura.

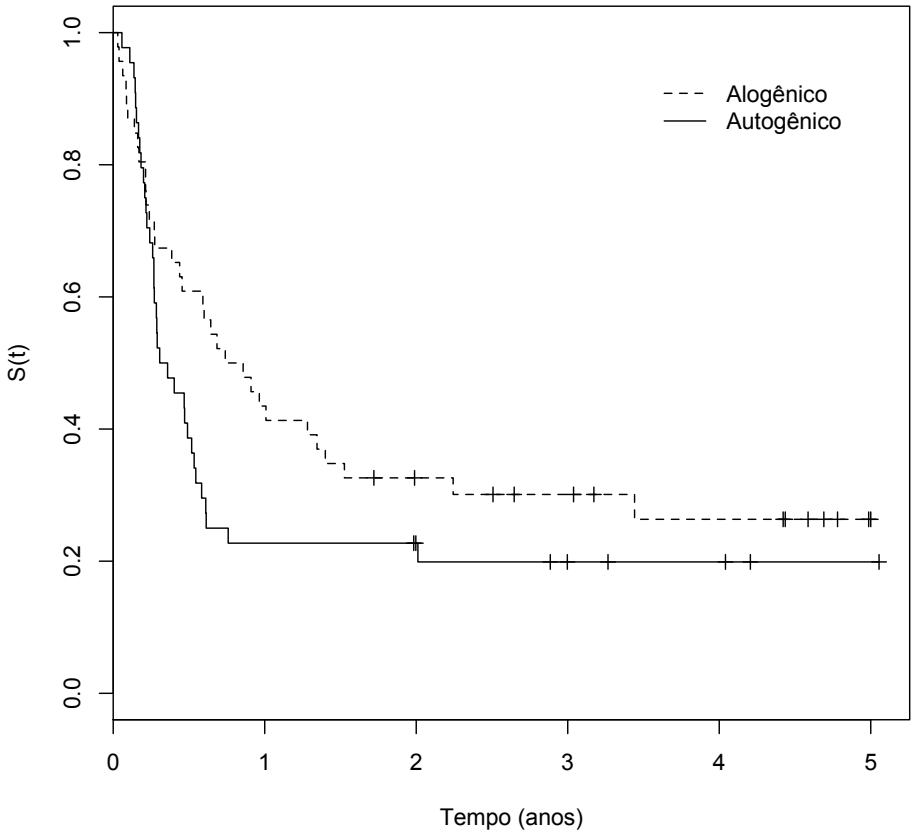

FIGURA 2.1: Estimativa de Kaplan-Meier de acordo com o tipo de transplante.

2.1 Modelos de Longa Duração

Os modelos de longa duração surgiram como uma maneira de superar as limitações dos modelos tradicionais, já que nessa nova classe de modelos admite-se que uma parte da população é imune ao evento de interesse. Alguns modelos já são consagrados em análise de sobrevivência. Boag (1949) e Berkson & Gage (1952) propuseram um modelo de mistura atribuindo uma variável binária para separar os indivíduos imunes dos suscetíveis. Posteriormente surgiram modelos mais complexos que tentam explicar melhor o mecanismo biológico envolvido, entre eles destacamos o modelo de tempo de promoção de Yakovlev & Tsodikov (1996) e o modelo unificado de Rodrigues et al. (2008).

Nos modelos de longa duração podemos classificar os indivíduos em duas categorias, os suscetíveis e os imunes. Os suscetíveis são aqueles sob risco de falha e os imunes são os que não estão sujeitos ao evento de interesse (falha), são os que em teoria

possuem tempos de falha infinitos, levando em consideração o risco de interesse. Assim, os indivíduos imunes sempre serão censurados no final do experimento e não poderemos fazer distinção deles dos que foram censurados no processo.

Com esse novo cenário precisamos reescrever as funções de sobrevivência e de risco, pois para indivíduos imunes a função de sobrevivência não será igual a 0 quando o tempo tende a infinito. Dessa forma acrescentamos um parâmetro que corresponderá à fração de curados. Primeiramente, a função de densidade da variável aleatória T deve ser flexibilizada de forma que

$$\int_0^\infty f(u)du = p \leq 1.$$

Agora podemos definir uma nova função de sobrevivência imprópria, que denotaremos por S^*.

Definição 2.3 A função de sobrevivência imprópria é dada por

$$S^*(t) = 1 - p + \int_t^\infty f(u)du, \qquad t \geq 0,$$

com as seguintes propriedades:

- Se $p = 1$, então $S^*(t) = S(t)$,
- $S^*(0) = 1$,
- $S^*(t)$ é decrescente,
- $\lim_{t\to\infty} S^*(t) = 1 - p$.

Notemos que a última propriedade concorda com a idéia da curva de sobrevivência se estabilizar a partir de um certo instante de tempo. Assim, o ponto em que a curva se estabiliza é justamente a probabilidade de imunes (fração de cura) da população.

Podemos estender as relações (2.1) e (2.2) para as funções impróprias e definir as funções densidade e de risco impróprias, dadas por

$$f^*(t) = -\frac{d}{dt}S^*(t)$$

e

$$h^*(t) = \frac{f^*(t)}{S^*(t)} = -\frac{d}{dt}\log S^*(t).$$

2.1.1 Modelo de Mistura Padrão

Um dos modelos mais utilizados na análise de sobrevivência é o modelo de mistura de padrão de Boag (1949) e Berkson & Gage (1952). Esse modelo é derivado da atribuição de uma variável de Bernoulli B_i, não observável, aos indivíduos suscetíveis e imunes da amostra. Nessa atribuição, um indivíduo é dito suscetível se $B_i = 1$, com $P[B_i = 1] = p$ e $B_i = 0$ se o indivíduo for imune, com $P[B_i = 0] = 1 - p$. A probabilidade $1 - p$ é chamada fração de cura.

Como temos duas subpopulações na amostra, teremos duas funções de distribuição acumuladas. Indivíduos suscetíveis ($B_i = 1$) possuem função de distribuição acumulada $F(t)$ própria. Os indivíduos imunes terão função de distribuição acumulada degenerada em 0, pois em teoria seus tempos de vida são infinitos, isto é,

$$P[T \leq t|B_i = 1] = F(t)$$

e

$$P[T \leq t|B_i = 0] = 0.$$

Assim, para a população como um todo teremos uma mistura das duas subpopulações e a função de sobrevivência $S^*(t)$ será imprópria e da forma

$$\begin{aligned} S^*(t) &= P[T > t] = P[T > t|B_i = 0]P[B_i = 0] + P[T > t|B_i = 1]P[B_i = 1] \\ &= 1 - p + pS(t), \end{aligned} \tag{2.3}$$

em que $S(t)$ é a função de sobrevivência do grupo suscetível e $1 - p$ é a fração de cura.

Dessa forma, o modelo de mistura padrão pode ser visto como um modelo de mistura das duas subpopulações envolvidas, e ele é caracterizado pelas funções de sobrevivência em (2.3) e pela função densidade

$$f^*(t) = -\frac{d}{dt}S^*(t) = pf(t), \tag{2.4}$$

em que $f(t)$ é a função densidade própria relativa ao grupo dos suscetíveis.

Notemos que as relações mostradas acima satisfazem as propriedades das funções de densidade e sobrevivência impróprias que caracterizam os modelos de longa duração.

A função de densidade $f^*(t)$ é uma função imprópria, pois

$$\int_0^\infty f^*(t)dt = \int_0^\infty pf(t)dt = p\int_0^\infty f(t)dt = p \leq 1.$$

Para a função de sobrevivência, a verificação das propriedades na definição 2.3 é imediata. Portanto, as funções que caracterizam o modelo de mistura padrão são funções impróprias e o modelo é de longa duração.

Fazendo uma analogia à construção da função verossimilhança para dados censurados com censura não-informativa como em Colosimo & Giolo (2006), consideramos um conjunto de dados observados $\boldsymbol{t} = (t_1, t_2, ..., t_n)$ e construímos a função verossimilhança usando as funções de densidade e sobrevivência impróprias, cuja expressão é

$$L(\theta, p; \boldsymbol{t}) \propto \prod_{i=1}^{n} \{f^*(t_i; \boldsymbol{\theta})\}^{\delta_i} \{S^*(t_i; \boldsymbol{\theta})\}^{1-\delta_i} = \prod_{i=1}^{n} \{pf(t_i; \boldsymbol{\theta})\}^{\delta_i} \{1 - p + pS(t_i; \boldsymbol{\theta})\}^{1-\delta_i}, \quad (2.5)$$

em que $\boldsymbol{\theta}$ é um possível vetor de parâmetros para a distribuição dos tempos de vida e δ_i é um indicador de censura.

É importante notar que a função verossimilhança em (2.5) não exprime a distribuição conjunta dos dados. Essa função verossimilhança é marginalizada em relação ao número de causas de falha, que não é observável. A justificativa de que essa função de verossimilhança é marginalizada é um teorema presente em Rodrigues et al. (2008). Esse teorema é formulado para o modelo unificado, proposto pelos autores, que modela M causas competidoras para a falha.

Como o modelo de mistura padrão é um caso particular do modelo unificado, pois temos somente uma causa de falha, o teorema justifica essa função de verossimilhança composta pelas funções impróprias. A versão para o modelo de mistura padrão desse teorema está apresentada a seguir.

Teorema 2.1 A função de verossimilhança em (2.5) é uma função de verossimilhança marginal.

Demonstração A verossimilhança do modelo unificado é dada por

$$L(\boldsymbol{\theta}, p; \boldsymbol{t}) \propto \sum_{\boldsymbol{m}} \prod_{i=1}^{n} \{S(t_i; \boldsymbol{\theta})\}^{m_i - \delta_i} \{m_i f(t_i; \boldsymbol{\theta})\}^{\delta_i} P(M_i = m_i),$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros da distribuição dos tempos de vida do grupo suscetível, M_i é o número de causas as quais o indivíduo i está exposto e t_i é o tempo de vida observado, censurado ou não.

Como temos somente uma causa de falha, pois trabalhamos com uma variável Bernoulli, teremos que M_i assumirá os valores 0 e 1 e consequentemente, como vimos anteriormente, $P(M_i = 0) = 1 - p$ e $P(M_i = 1) = p$. A demonstração segue considerando as seguintes situações:

- δ_i=0:

$$\begin{aligned} L(\boldsymbol{\theta}, p; \boldsymbol{t}) &\propto \sum_{\boldsymbol{m}} \prod_{i=1}^{n} \{S(t_i; \boldsymbol{\theta})\}^{m_i} P(M_i = m_i) = \prod_{i=1}^{n} \sum_{\boldsymbol{m}} \{S(t_i; \boldsymbol{\theta})\}^{m_i} P(M_i = m_i) \\ &= \prod_{i=1}^{n} \{S(t_i; \boldsymbol{\theta})\}^{0} P(M_i = 0) + \{S(t_i; \boldsymbol{\theta})\}^{1} P(M_i = 1) \\ &= \prod_{i=1}^{n} 1 - p + pS(t_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} S^*(t_i; \boldsymbol{\theta}). \end{aligned}$$

- δ_i=1:

$$\begin{aligned} L(\boldsymbol{\theta}, p; \boldsymbol{t}) &= \sum_{\boldsymbol{m}} \prod_{i=1}^{n} \{S(t_i; \boldsymbol{\theta})\}^{m_i - 1} m_i f(t_i; \boldsymbol{\theta}) P(M_i = m_i) \\ &= \prod_{i=1}^{n} \sum_{\boldsymbol{m}} \{S(t_i; \boldsymbol{\theta})\}^{m_i - 1} m_i f(t_i; \boldsymbol{\theta}) P(M_i = m_i) \\ &= \prod_{i=1}^{n} \{S(t_i; \boldsymbol{\theta})\}^{-1} 0 f(t_i; \boldsymbol{\theta}) P(M_i = 0) + \{S(t_i; \boldsymbol{\theta})\}^{0} 1 f(t_i; \boldsymbol{\theta}) P(M_i = 1) \\ &= \prod_{i=1}^{n} p f(t_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} f^*(t_i; \boldsymbol{\theta}). \end{aligned}$$

Combinando as duas situações obtemos o resultado enunciado. ■

2.2 Função Probabilidade de Imunes

Vimos que quando existem indivíduos imunes na população, a curva de sobrevivência atinge um patamar constante após um longo período de tempo e os indivíduos que ainda estão vivos ao final do estudo são dados como censurados. Mas entre esses

indivíduos censurados ao final do experimento, não poderia acontecer mais uma falha no próximo instante de tempo? Fica aqui a dúvida de que entre os indivíduos censurados ao final do experimento possam existir indivíduos que não são imunes.

Assim, suponha que um indivíduo teve seu tempo de falha censurado. Qual a probabilidade de que esse indivíduo seja curado ou imune? Para calcular essa probabilidade nos baseamos na variável B_i com distribuição de Bernoulli que divide a população em imunes e suscetíveis.

Dado que um indivíduo i sobreviveu até o instante de tempo $t > 0$, denotamos a probabilidade de que ele seja imune por

$$p(t) = P[B_i = 0|T > t].$$

Pelo Teorema de Bayes,

$$p(t) = \frac{P[B_i = 0, T > t]}{P[T > t]} = \frac{P[T > t|B_i = 0]P[B_i = 0]}{P[T > t]}.$$

Sabemos que $P[T \leq t|B_i = 0] = 0$; então, $P[T > t|B_i = 0] = 1$. Juntando isso ao fato de que $P[B_i = 0] = 1 - p$ e levando em conta (2.3), temos

$$p(t) = \frac{1-p}{1-pF(t)} = \frac{1-p}{S^*(t)}. \tag{2.6}$$

Chamamos essa função de *função probabilidade de imunes* e podemos interpretá-la como uma medida de plausibilidade de que o indivíduo esteja curado, ou seja, à medida que o tempo passa, maior é a probabilidade de que um indivíduo em risco seja imune ou curado.

A função probabilidade de imunes é crescente em t, de um valor $1 - p$ quando $t = 0$, que corresponde a nenhuma informação a respeito da imunidade do indivíduo além da taxa de cura, a um valor 1 quando $t \to \infty$, que corresponde à certeza de imunidade do indivíduo se o seu tempo de vida for bem grande.

Podemos utilizar essa função como uma medida de eficiência entre dois tratamentos, já que quanto mais rápido for o crescimento da curva, mais eficaz é o tratamento no sentido de que a certeza de que os indivíduos estejam curados é maior em menos tempo. Assim, o melhor tratamento será aquele que tiver uma curva mais alta e mais inclinada.

A determinação da função probabilidade de imunes pode ser feita de forma paramétrica ou não-paramétrica.

Na abordagem não-paramétrica utilizamos o estimador de Kaplan-Meier $\hat{F}_{ekm}(t)$ para a função de distribuição da população $F^*(t)$ e segundo Maller & Zhou (1996) e Balka et al. (2009) consideramos como estimativa não-paramétrica da fração de cura $\hat{p}_{ekm} = 1 - \min(\hat{S}_{ekm}(t))$. Assim, a formulação não paramétrica da função probabilidade de imunes fica da seguinte forma:

$$\hat{p}_{ekm}(t; p, \boldsymbol{\theta}) = \frac{1 - \hat{p}_{ekm}}{1 - \hat{F}_{ekm}(t; \boldsymbol{\theta})}.$$

Na abordagem paramétrica, impomos um modelo estatístico paramétrico para a função de distribuição da população. Então, utilizamos a teoria da verossimilhança para estimar os parâmetros desse modelo e com essas estimativas calculamos a função probabilidade de imunes. Assim, a formulação paramétrica da função probabilidade de imunes fica da seguinte forma:

$$\hat{p}(t) = \frac{1 - \hat{p}}{1 - \hat{F}^*(t)},$$

em que $\hat{p}$ é o estimador de máxima verossimilhança de p e $\hat{F}^*(t; \boldsymbol{\theta})$ é o estimador de máxima verossimilhança de $F^*(t; \boldsymbol{\theta})$ obtido pela propriedade da invariância.

2.2.1 Exemplo

Consideremos o conjunto de dados referente ao tempo de recorrência de leucemia que usamos anteriormente. Inicialmente calculamos a função probabilidade de imunes de forma não-paramétrica e depois impondo que os tempos seguem uma distribuição exponencial com parâmetro λ.

Para esse exemplo, observamos na figura 2.2 que a curto prazo (em torno de seis meses) o grupo 1 (alogênico) tem maior probabilidade de cura. Por outro lado, a partir de seis meses, o grupo 2 apresenta maior probabilidade de cura, ou seja, a longo prazo o grupo 2 apresenta uma probabilidade de cura maior que o grupo 1.

Assim, os gráficos sugerem que a curto prazo o transplante alogênico (grupo 1) é mais eficaz, mas a curva do transplante autogênico (grupo 2) cresce mais acentuadamente, indicando um tratamento mais eficaz a longo prazo.

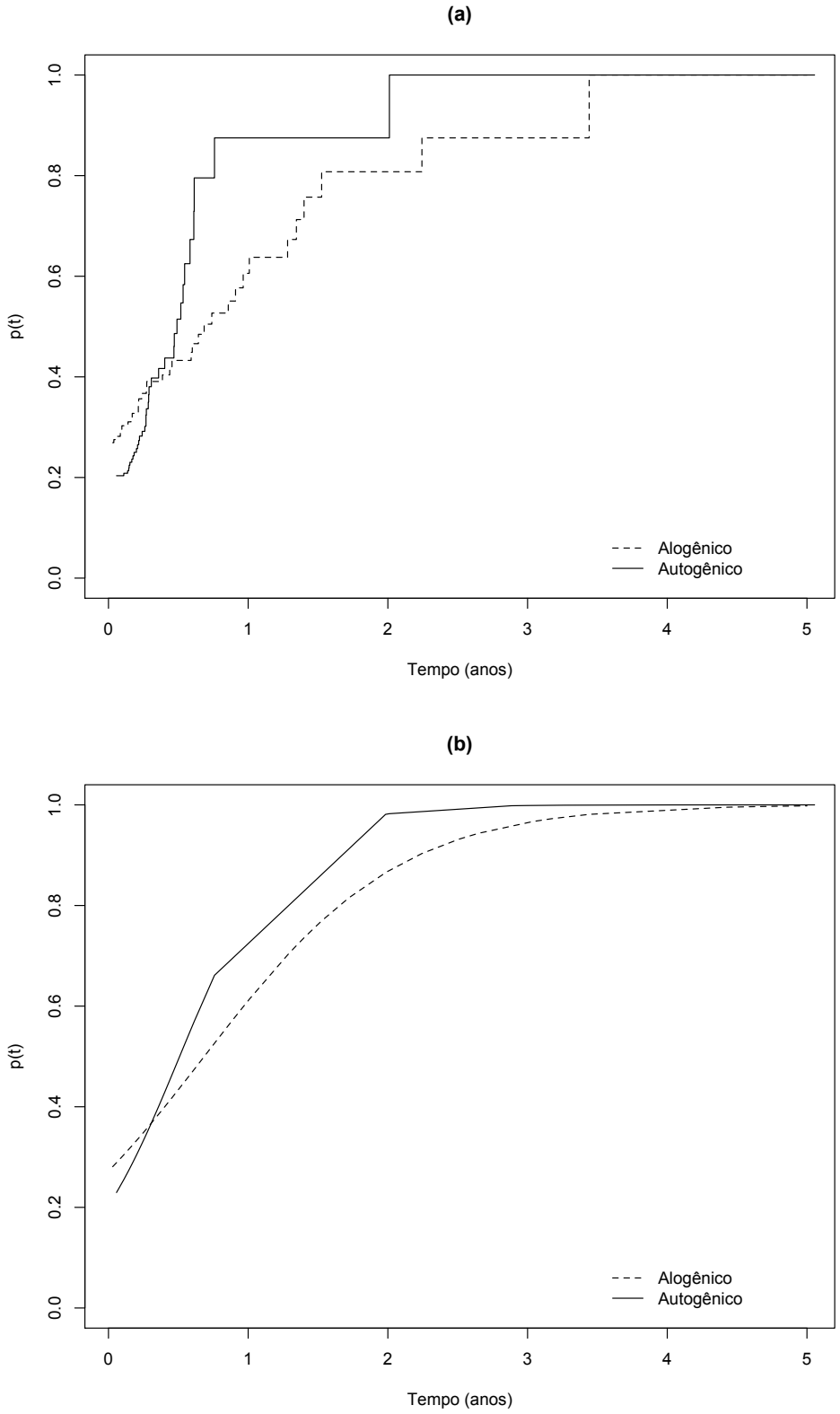

FIGURA 2.2: Funções probabilidade de imunes (a) estimação não-paramétrica e (b) estimação paramétrica

2.3 Vantagens e Desvantagens do Modelo de Mistura Padrão

O modelo de mistura padrão foi um dos primeiros modelos paramétricos em análise de sobrevivência de longa duração. Mas como qualquer modelo estatístico, ele tem vantagens e desvantagens.

A atribuição da variável Bernoulli à condição dos pacientes implica numa fácil formulação do modelo, dependendo somente de algumas noções de probabilidade. Por outro lado, esse modelo não exprime o mecanismo biológico envolvido, algo que, por exemplo, já não acontece no modelo de tempo de promoção de Yakovlev & Tsodikov (1996) em que se modela o número de causas que competem para a ocorrência do evento de interesse.

Uma outra característica do modelo de mistura padrão é que ele não é um modelo de riscos proporcionais. Basta observar a função de risco do modelo,

$$h^*(t) = -\frac{pf(t)}{1-p+pS(t)}.$$

Note que se considerarmos $f(t)$ como a função de risco base, comum a todos os indivíduos, o resto da expressão ainda depende de t através da função de sobrevivência $S(t)$.

Segundo Ibrahim et al. (2001) uma desvantagem, do ponto de vista bayesiano, é que quando associamos covariáveis ao parâmetro p, ao utilizarmos distribuições *a priori* não informativas para os coeficientes da regressão, obtemos distribuições *a posteriori* impróprias.

2.4 Identificabilidade do Modelo

Uma questão importante a ser levantada é sobre a identificabilidade do modelo de mistura. Li et al. (2001) fornece a seguinte definição de identificabilidade.

Definição 2.4 A classe dos modelos de mistura H é identificável se para quaisquer dois membros de H dados por $S_1^*(t;\theta_1,x) = 1-p_1(x)+p_1(x)S_1(t;\theta_1|B_i=1,x)$ e $S_2^*(t;\theta_2,x) = 1-p_2(x)+p_2(x)S_2(t;\theta_2|B_i=1,x)$, então $S_1^*(t;\theta_1,x) = S_2^*(t;\theta_2,x)$ se, e somente se,

$p_1(x) = p_2(x)$ e $S_1(t;\theta_1|B_i = 1, x) = S_2(t;\theta_2|B_i = 1, x)$, para todo vetor de covariáveis x e para quase todo $t \in (0, \infty)$.

No cenário em que trabalhamos, a função de sobrevivência $S(t)$ é especificada por uma distribuição paramétrica e p é um parâmetro constante, ou seja, ele não depende de covariáveis. Então, apresentamos o teorema a seguir, presente em Li et al. (2001), que abrange um caso mais geral em que p é uma função de um vetor de covariáveis $\boldsymbol{x}$.

Teorema 2.2 O modelo de mistura dado por (2.4) no qual especificamos uma distribuição paramétrica para $S(t)$ é identificável independentemente se $p(x)$ é paramétrica ou não.

Demonstração Pela definição acima, supondo que $S_1(t;\theta_1|B_i = 1, x) = S_2(t;\theta_2|B_i = 1, x)$ e $p_1(x) = p_2(x)$ a implicação que $S_1^*(t;\theta_1, x) = S_2^*(t;\theta_2, x)$ é imediata. Precisamos mostrar a outra implicação. Suponhamos que $S_1^*(t;\theta_1, x) = S_2^*(t;\theta_2, x)$. Por (2.3) segue que

$$\frac{p_1(x)}{p_2(x)} = \frac{1 - S_2(t;\theta_2|Y = 1)}{1 - S_1(t;\theta_1|Y = 1)} = c.$$

Como o lado esquerdo depende somente de x e o lado direito somente de t, essa razão é igual a uma constante positiva c que não depende nem de x nem de t. Dessa forma, $S_2(t;\theta_2|Y = 1)$ não pertence à mesma família paramétrica que $S_1(t;\theta_1|Y = 1)$ se $c \neq 1$ a não ser que $\theta_1 = \theta_2$.

Então, temos $c = 1$, o que implica em $p_1(x) = p_2(x)$ e $S_1(t;\theta_1|Y = 1) = S_2(t;\theta_2|Y = 1)$ e portanto, o modelo é identificável pela definição 2.4. ■

Como foi dito anteriormente, esse teorema considera p como uma função de um vetor de covariáveis x e para o caso em que p é um parâmetro constante uma demonstração análoga pode ser efetuada. Logo, podemos incluir como uma vantagem do modelo de mistura padrão sua identificabilidade no caso em que a função de sobrevivência $S(t)$ é especificada através de uma distribuição de probabilidade.

Capítulo 3

Distribuições de Probabilidades Ponderadas

Suponha que a realização x de uma v.a. X, com uma função densidade $f(x;\theta)$ seja incluída no registro de um pesquisador com probabilidade proporcional a $w(x;\theta,\phi)$, uma função de peso não-negativa, que dependa dos dados e que pode ou não depender dos parâmetros da função densidade $f(x;\theta)$ ou de um outro parâmetro perturbador ϕ. Essa observação registrada não foi obtida de X mas sim de uma v.a. ponderada X_{pond}.

Distribuições ponderadas ocorrem naturalmente em contextos em que a probabilidade de uma observação ser incluída na amostra seja multiplicada por uma função de peso não-negativa. Segundo Rao (1985), quando temos uma amostra em mãos, nem sempre sabemos a qual população ela pertence. Com isso em mente, podemos questionar a possibilidade de que essa amostra tenha sido retirada de uma população por um mecanismo ponderado.

Amostras desse tipo podem ter características de sobredispersão, subdispersão ou um vício de seleção. Isso acontece devido à influência da função de peso. Distribuições de probabilidades ponderadas são utilizadas na tentativa de criar um modelo mais apropriado para dados obtidos de populações sem um referencial.

Artigos recentes apresentam modelos ponderados. Del Aguila et al. (2001) apresentam uma versão ponderada para amostragem viciada de diversos modelos estatísticos, entre eles os modelos normal, binomial negativa e Poisson. Kokonendiji et

al.(2008) propõem um modelo Poisson ponderado com um mecanismo de identificação de sobredispersão ou subdispersão.

Seja X uma variável aleatória não negativa com função densidade $f(x;\theta)$. Segundo Gupta & Kirmani (1990), definimos a variável ponderada X_{pond}, associada à variável X, com função de peso $w(x;\theta,\phi)$, através da função densidade ponderada

$$f_{pond}(x;\theta,\phi) = \frac{w(x;\theta,\phi)f(x;\theta)}{E(w(X;\theta,\phi))}, \tag{3.1}$$

em que $E(w(X;\theta,\phi))$ é a esperança em relação à função densidade $f(x;\theta)$. Como em Rodrigues (2008) a função de peso pode ser escrita na forma exponencial, $w(x;\theta,\phi) = e^{rt(x;\theta,\phi)}$, com $t(x;\theta,\phi)$ uma função dos dados que pode ou não depender dos parâmetros do modelo ou de um parâmetro pertubador e r uma constante conhecida. Nessa nova função densidade ponderada, a função densidade de origem $f(x;\theta)$ assume o papel de uma distribuição "pai".

Com a função densidade ponderada em mãos podemos definir, como em Gupta & Kirmani (1990), as funções de sobrevivência e de risco ponderadas, dadas por

$$S_{pond}(x) = \frac{E(w(X;\theta,\phi)|X > x)}{E(w(X;\theta,\phi))}S(x;\theta) \tag{3.2}$$

e

$$h_{pond}(x) = \frac{w(x;\theta,\phi)}{E(w(X;\theta,\phi)|X > x)}h(x;\theta). \tag{3.3}$$

Observemos que a ponderação funciona como um fator de correção da função densidade original, no sentido em que a ponderação pode ampliar ou reduzir a contribuição das caudas da distribuição "pai".

Antes de vermos alguns resultados sobre densidades ponderadas, apresentamos algumas definições que serão úteis na compreensão de alguns resultados presentes nesse capítulo.

Definição 3.1 Uma v.a. X é estocasticamente maior que uma v.a. Y ($X >_{st} Y$) se para qualquer a, $P(X \geq a) \geq P(Y \geq a)$.

Definição 3.2 Seja X uma v.a. com primeiro e segundo momentos finitos. O coeficiente de variação (CV) de X é dado por $CV = \frac{\sqrt{Var(X)}}{E(X)}$.

Definição 3.3 Sejam X e Y duas v.a. com respectivas informações de Fisher $I_X(\theta)$ e $I_Y(\theta)$. A v.a. X é dita mais informativa que a v.a. Y ($X >_F Y$) se, e somente se, $I_X(\theta) > I_Y(\theta)$.

Os seguintes resultados, presentes em Gupta & Kirmani (1990) relacionam a v.a ponderada com a sua v.a. "pai".

Teorema 3.1 Seja X uma v.a. não negativa e X_{pond} a v.a. ponderada associada.

- Se $w(x;\theta,\phi)$ é monótona crescente, $h_{pond}(x;\theta) \leq h(x;\theta)$. Então, $S_{pond}(x;\theta) \geq S(x;\theta)$ ($X \leq_{st} X_{pond}$), para todo x.
- Se $w(x;\theta,\phi)$ é monótona decrescente, $h_{pond}(x;\theta) \geq h(x;\theta)$. Então, $S_{pond}(x;\theta) \leq S(x;\theta)$ ($X \geq_{st} X_{pond}$), para todo x.

Corolário 3.1 Seja X uma v.a. não negativa e X_{pond} a v.a. ponderada associada com função de peso $w(x;\theta,\phi)$. São válidas:

- $E(X_{pond}) > E(X)$ se $\mathrm{Cov}(X, w(X)) > 0$.
- $E(X_{pond}) < E(X)$ se $\mathrm{Cov}(X, w(X)) < 0$.

3.1 Distribuições Exponenciais Ponderadas

A distribuição exponencial é bastante utilizada na modelagem de tempos de vida, apesar da restrição de sua função de risco ser constante. Com o intuito de se livrar dessa restrição e aproveitar as propriedades das densidades ponderadas, utilizamos a distribuição exponencial como distribuição "pai". Desse modo definimos a família das distribuições exponenciais ponderadas as distribuições da seguinte forma

$$f(x;\theta,\phi) = \frac{w(x;\theta,\phi)\theta e^{-\theta x}}{E(w(X;\theta,\phi))}.$$

Rodrigues (2008) apresenta uma série de resultados sobre a família das exponenciais ponderadas, entre eles destacamos os seguintes.

Teorema 3.2 Sejam X_{pond} uma v.a. com distribuição exponencial ponderada e X a correspondente v.a. com distribuição exponencial com parâmetro θ. Então, $X_{pond} >_{st} X$ ($X_{pond} <_{st} X$) se $t(x;\theta,\phi)$ é monótona crescente (monótona decrescente), com $r > 0$.

Teorema 3.3 Seja X_{pond} uma v.a. com distribuição exponencial ponderada. Se $t(x;\theta,\phi)$ é uma função estritamente convexa e $r > 0$ ($r < 0$), então $CV(X_{pond}) > 1$ ($CV(X_{pond}) < 1$).

Teorema 3.4 Seja X_{pond} uma v.a. com distribuição exponencial ponderada e X a correspondente v.a. com distribuição exponencial com parâmetro θ. Então, temos que $X_{pond} >_F X$.

Demonstração A informação de Fisher da v.a. ponderada X_{pond} pode ser escrita como

$$I_{X_{pond}}(\theta) = I_X(\theta) + \overbrace{\frac{d^2}{d\theta^2}\log E_\theta(w(X;\theta,\phi))}^{(1)} - \overbrace{\frac{d^2}{d\theta^2}E_\theta(\log w(X;\theta,\phi))}^{(2)}.$$

Pela desigualdade de Jensen, (1) é maior do que ou igual a (2) e portanto, $I_{X_{pond}}(\theta) \geq I_X(\theta)$. ■

Neste trabalho, apresentaremos duas distribuições exponenciais ponderadas, a *length biased* e a beta exponencial.

3.2 Distribuição Exponencial *Length Biased*

Em diversas situações é razoável pensar que quanto maior for o tempo de vida de um paciente, maior é a chance de ele ser selecionado para uma amostra. Esse fato aponta a possibilidade de uma amostragem viciada e uma distribuição ponderada pode ser empregada a fim de compensar esse vício na amostra.

Tomaremos $w(x;\theta,\phi) = x$ como função de peso, ou seja, tomamos uma função de peso igual à magnitude da observação. Chamaremos essa distribuição ponderada de exponencial *length biased* (elb).

Para escrevermos a função densidade e a função de sobrevivência dessa nova variável ponderada, devemos calcular as esperanças em (3.1) e (3.2). Iniciamos com

$$E(w(X;\theta,\phi)) = E(X) = \frac{1}{\theta}.$$

A esperança condicional em (3.2) e (3.3) e será calculada em relação à distribuição exponencial truncada, que é dada por

$$f(s;x) = \begin{cases} \frac{\theta e^{-\theta s}}{e^{-\theta x}} & \text{se} \quad s \geq x, \\ 0 & \text{se} \quad s < x. \end{cases}$$

Calculando a esperança condicional,

$$E(w(X)|X > x) \;=\; E(X|X > x) = \int_x^\infty \frac{s\theta e^{-\theta s}}{e^{-\theta x}} ds.$$

Resolvendo essa integral por partes obtemos

$$E(w(X)|X > x) \;=\; E(X|X > x) = x + \frac{1}{\theta}.$$

Assim, a função densidade, a função de sobrevivência e a função de risco da v.a. exponencial *length biased* são dadas por

$$f_{elb}(x;\theta) \;=\; x\theta^2 e^{-\theta x}, \tag{3.4}$$

$$S_{elb}(x;\theta) \;=\; e^{-\theta x}(1+\theta x), \tag{3.5}$$

$$h_{elb}(x;\theta) \;=\; \frac{x\theta}{x+\frac{1}{\theta}}. \tag{3.6}$$

Comparando as funções S_{elb} e h_{elb} com as respectivas funções de sobrevivência $S(x;\theta)$ e risco $h(x;\theta)$ da v.a. exponencial, vemos que

$$S_{elb}(x;\theta) \;\geq\; S(x;\theta)$$

e

$$h_{elb}(x;\theta) \;\leq\; h(x;\theta).$$

Observemos que a função densidade $f_{elb}(x;\theta)$ corresponde a uma distribuição $Gama(2,\theta)$. Mais geralmente, poderíamos utilizar $w(x) = x^\phi$ $(\phi > 0)$ como função de peso, o que resultaria numa distribuição $Gama(\phi+1,\theta)$. Entretanto, na literatura as citações à função de peso *length biased* são sempre da forma como utilizamos aqui.

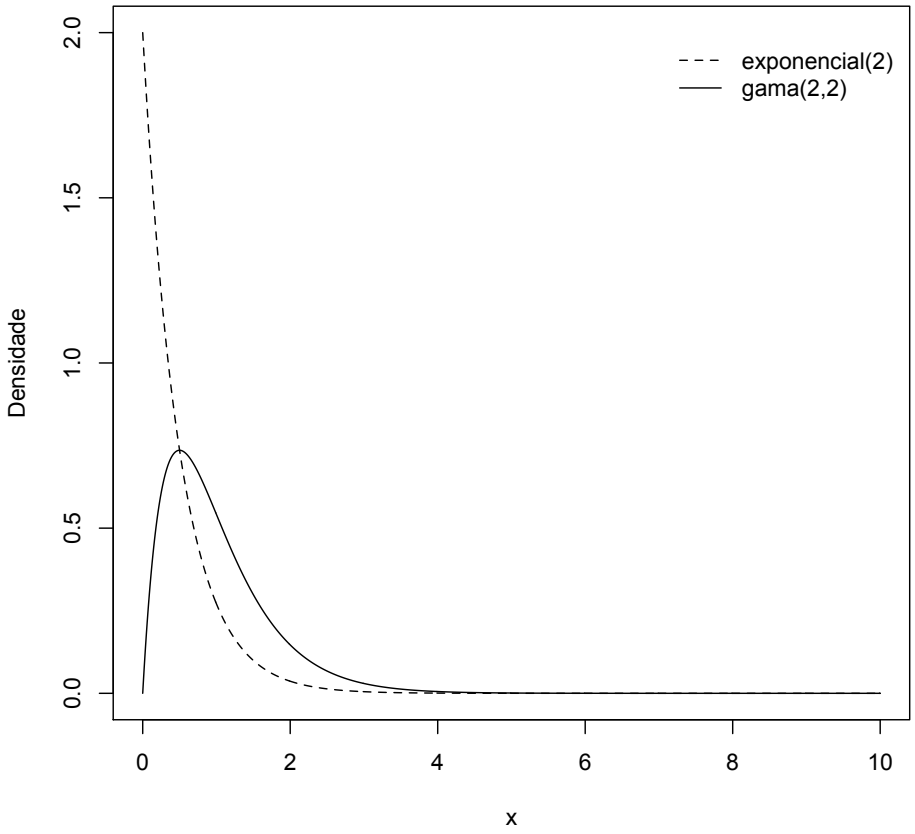

FIGURA 3.1: Comparação de uma distribuição exponencial com uma exponencial *length biased*

Vejamos graficamente o impacto da ponderação sobre distribuição exponencial "pai". Na figura 3.1 notamos que na distribuição exponencial temos uma grande concentração de massa em torno da origem. Com a ponderação, essa massa foi dispersada e a cauda da distribuição ponderada ficou mais pesada.

Como mencionado no início desse capítulo, a função de peso de uma v.a. ponderada pode ser escrita na forma exponencial, $w(x;\theta,\phi) = e^{rt(x;\theta,\phi)}$. No caso da distribuição exponencial *length biased*, a função de peso escrita na forma exponencial é dada por

$$w(x) = x = e^{\log x}.$$

Logo, $r = 1$ e $t(x;\theta;\phi) = \log x$. Agora, a partir da função de peso tentamos relacionar a distribuição ponderada com a distribuição "pai"utilizando os teoremas da seção 2.1. Ao olharmos para a estatística $\log x$, vemos que ela é uma função monótona crescente. Como

$r = 1$, pelo teorema 3.2, a distribuição exponencial *length biased* é estocasticamente maior do que a distribuição exponencial "pai".

Vejamos agora o comportamento do coeficiente de variação da distribuição exponencial *length biased*. Como $r > 0$, segundo o teorema 3.3 o CV da distribuição exponencial *length biased* seria maior do que 1 se a estatística $t(x;\theta,\phi)$ fosse estritamente convexa. Mas nesse caso $t(x;\theta,\phi) = \log x$ não é convexa, pois sua 2^a derivada não é positiva em todo o seu domínio. Portanto, o teorema 3.3 não se aplica.

Entretanto, podemos analisar quais fatores influenciam o CV dessa distribuição. O CV da distribuição exponencial *length biased*, o qual denotaremos por CV_{elb}, pela definição (3.2) é dado por

$$CV_{elb} = \frac{\sqrt{Var(X_{elb})}}{E(X_{elb})} = \frac{\frac{\sqrt{2}}{\theta}}{\frac{2}{\theta}} = \frac{\sqrt{2}}{2}. \tag{3.7}$$

CV_{elb} é menor do que 1, isso significa que a distribuição exponencial *length biased* é menos dispersa que a distribuição exponencial.

Vejamos como se comporta a função de risco da distribuição exponencial *length biased*. Pela figura 3.2 vemos que a exponencial *length biased* só comporta funções de risco crescentes. Isso era esperado, pois ela é uma distribuição $Gama(2,\theta)$, que tem risco crescente quando o parâmetro de forma é maior do que 1, e nesse caso temos o valor 2.

3.3 Distribuição Beta Exponencial

Tendo em vista a frequente utilização da distribuição exponencial em análise de sobrevivência e da confiabilidade, Nadarajah & Kotz (2006) propuseram a distribuição beta exponencial como uma forma de generalizar a distribuição exponencial. Além de conter a distribuição exponencial como caso particular, a distribuição beta exponencial é mais flexível em se tratando da forma de suas funções de densidade e risco.

A formulação dessa distribuição vem da possibilidade de se derivar uma nova classe de distribuições a partir da função de distribuição acumulada (f.d.a.), $G(x)$, de uma variável aleatória. A f.d.a. dessa nova distribuição generalizada é dada por

$$F(x) = I_{G(x)}(a,b) = \frac{B_{G(x)}(a,b)}{B(a,b)}, \tag{3.8}$$

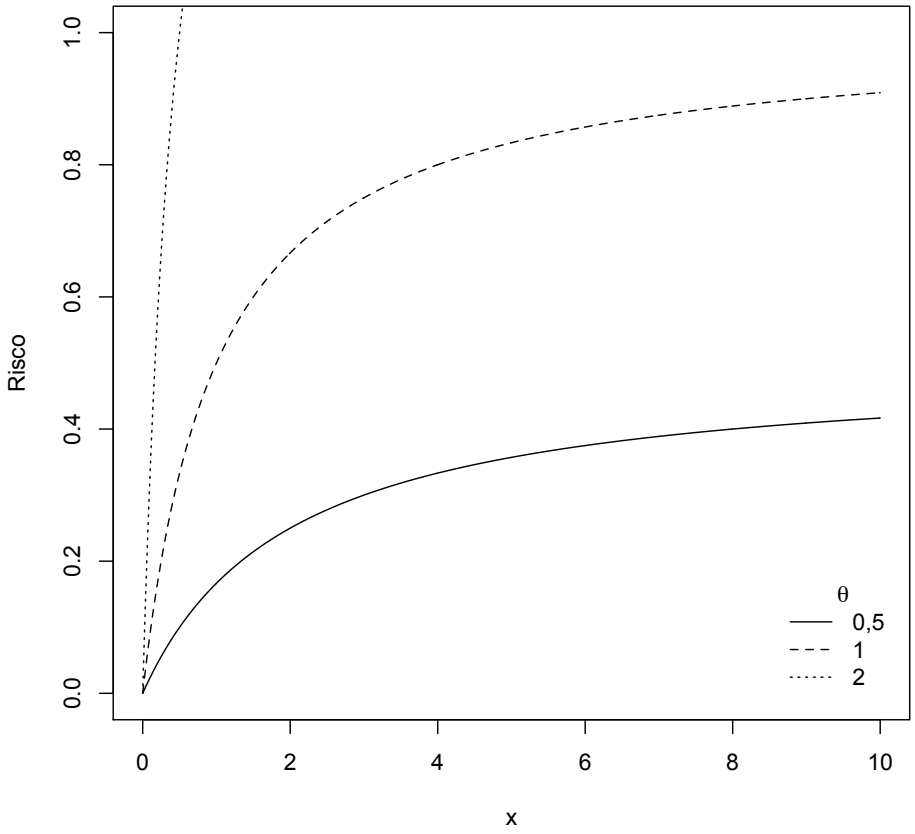

FIGURA 3.2: Função de risco da distribuição exponencial *length biased* para diferentes valores de θ

em que a e b são parâmetros, ambos maiores do que zero, e $B_{G(x)}(a,b)$ denota a função beta incompleta, definida por

$$B_y(a,b) = \int_0^y \omega^{a-1}(1-\omega)^{b-1}d\omega$$

Para construir a distribuição beta exponencial, utilizamos a f.d.a. da distribuição exponencial com parâmetro θ na expressão (3.8). Dessa forma as funções de densidade e de risco são dadas por

$$f(x) = \frac{\theta}{B(a,b)}e^{-b\theta x}(1-e^{-\theta x})^{a-1}$$

e

$$h(x) = \frac{\theta e^{-b\theta x}(1-e^{-\theta x})^{a-1}}{B_{e^{-\theta x}}(b,a)},$$

com $a > 0$, $b > 0$, $\theta > 0$ e $x > 0$.

Outras distribuições podem ser geradas da mesma forma, apenas mudando a f.d.a a ser inserida em (3.8). Nadarajah & Kotz (2006) citam que a distribuição beta exponencial é, nesta classe de distribuições, uma das mais tratáveis, diferentemente da distribuição beta Weibull que, apesar de ser mais geral, não possui expressões fechadas para os seus momentos.

Notemos que a distribuição beta exponencial admite, como casos particulares, as distribuições:

- exponencial com parâmetro θ, quando $a = 1$ e $b = 1$,
- exponencial com parâmetro $b\theta$, quando $a = 1$ e
- exponencial exponenciada quando $b = 1$.

Como resultado do acréscimo dos parâmetros a e b, as funções de densidade e risco da distribuição beta exponencial apresentam várias formas, como mostra a figura 3.3. Nesses gráficos estamos assumindo que $b = \theta = 1$

Como os gráficos sugerem, o parâmetro a é o único responsável pela mudança na forma da distribuição. Esse fato pode ser confirmado através do estudo do comportamento das derivadas da função de densidade. As derivadas de primeira e segunda ordem do logaritmo da função de densidade são dadas por

$$\frac{d}{dx}\log f(x) = \frac{(a-1)\theta e^{-\theta x}}{1-e^{-\theta x}} - b\theta \tag{3.9}$$

e

$$\frac{d^2}{dx^2}\log f(x) = \frac{(1-a)\theta^2 e^{-\theta x}}{(1-e^{-\theta x})^2} \tag{3.10}$$

Analisamos o comportamento de (3.9) e (3.10) para $a < 1$ e para $a > 1$. Para $a < 1$, a expressão em (3.10) é positiva; portanto, a expressão em (3.9) é crescente, com $\log f'(0) = -\infty$ e $\log f'(\infty) = -b\theta$, implicando que a função $f(x)$ é decrescente.

Para o caso em que $a > 1$, a expressão em (3.10) é negativa, o que torna (3.9) decrescente, com $\log' f(0) = \infty$ e $\log' f(\infty) = -b\theta$, o que implica a existência de um único

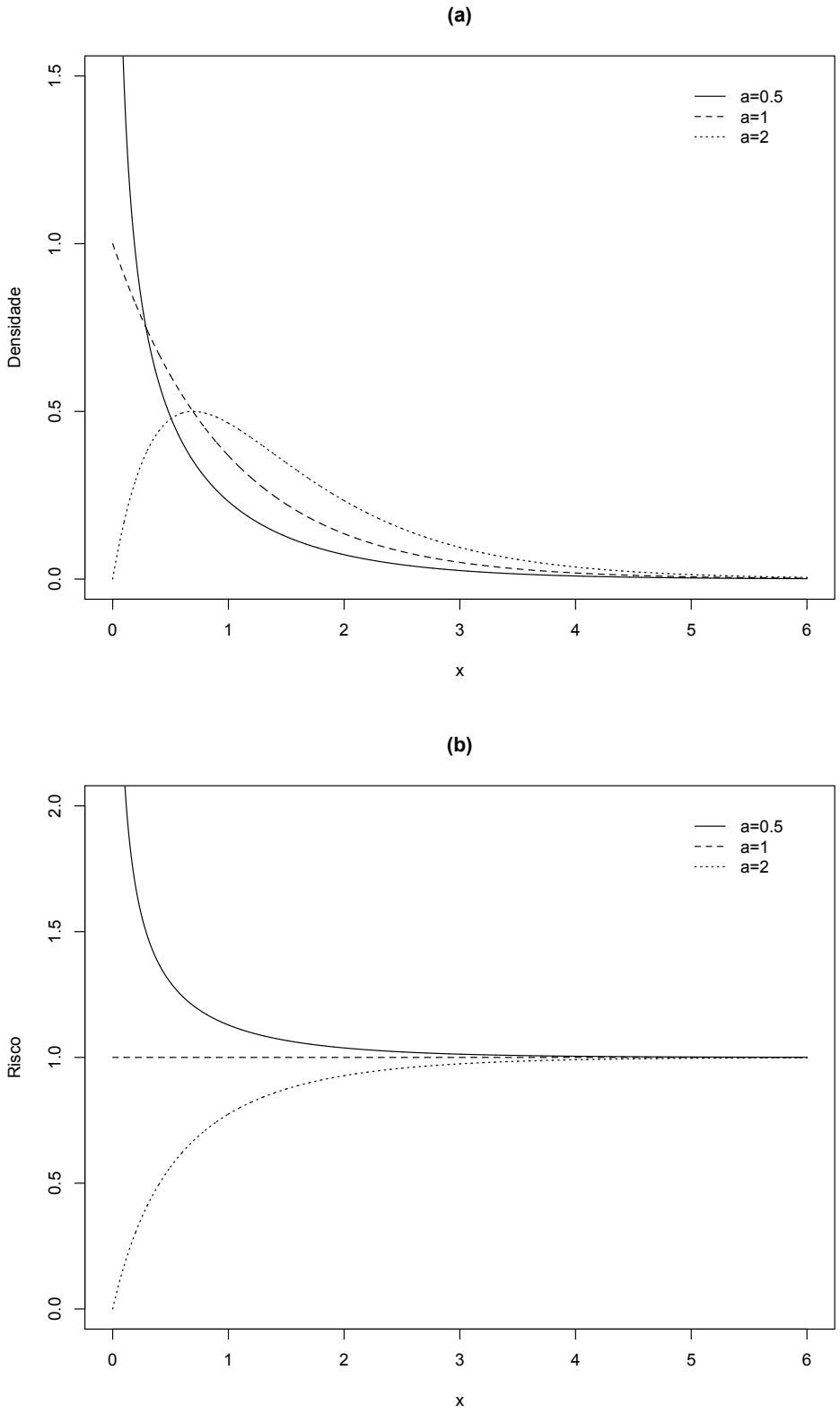

FIGURA 3.3: Funçõe densidade(a) e risco (b) da distribuição beta exponencial para diferentes valores de a

ponto x_0 tal que $f'(x_0) = 0$, ou seja, $f(x)$, tem uma única moda. Assim, $f(x)$ é crescente para $x < x_0$ e decrescente para $x > x_0$, em que x_0 é o ponto que anula (3.9).

Para encerrar a caracterização da distribuição beta exponencial, apresentaremos as expressões para a função geradora de momentos, a função característica, a esperança e a variância. A função geradora de momentos é definida como M(t)=E(e^{tX}), e para a distribuição beta exponencial é dada por

$$M(t) = \frac{\theta}{B(a,b)} \int_0^\infty e^{(t-b\theta)x}(1-e^{-\theta x})^{a-1}dx.$$

Substituindo $y = e^{-\theta x}$, a integral se simplifica de forma que

$$M(t) = \frac{B(b-\frac{t}{\theta},a)}{B(a,b)}. \tag{3.11}$$

A função característica de uma v.a. X, definida por $\Phi(t) = E(e^{itX})$, que para a distribuição beta exponencial se torna

$$\Phi(t) = \frac{B(b-\frac{it}{\theta},a)}{B(a,b)}.$$

A partir de (3.11) obtemos as expressões da esperança e da variância, dadas por

$$E(X) = \frac{\Psi(a+b) - \Psi(b)}{\theta} \tag{3.12}$$

e

$$Var(X) = \frac{\Psi'(b) - \Psi'(a+b)}{\theta^2} \tag{3.13}$$

em que $\Psi(x) = \frac{d}{dx}\log\Gamma(x)$, também conhecida como função digama e Ψ' é a função trigama .

Analisando a complexidade das funções que caracterizam a distribuição beta exponencial, que vantagens ela leva sobre, por exemplo, a distribuição Weibull como uma generalização da distribuição exponencial? Um ganho em se trabalhar com a distribuição beta exponencial é que essa distribuição faz parte da família das distribuições exponenciais ponderadas, o que nos fornece algumas informações sobre ela de antemão.

Teorema 3.5 A distribuição beta exponencial com função densidade

$$f(x;a,b,\theta) = \frac{\theta}{B(a,b)}e^{-b\theta x}(1-e^{-\theta x})^{a-1}.$$

pertence à família das distribuições exponenciais ponderadas.

Demonstração A função de densidade beta exponencial, pode ser escrita como

$$f(x;a,b,\theta) = \underbrace{\frac{(1-e^{-\theta x})^{a-1}}{bB(a,b)}}_{\frac{w(x)}{E(w(x))}} b\theta e^{-b\theta x}.$$

Como o denominador $bB(a,b)$ é constante, ele fará o papel da esperança da função de peso. Já o numerador fará o papel da função de peso, $w(x) = (1-e^{-\theta x})^{a-1}$. Então, a partir dessa função de peso, calculamos a sua esperança.

$$E(w(X)) = \int_0^\infty w(x)g(x)dx = \int_0^\infty (1-e^{-\theta x})^{a-1} b\theta e^{-b\theta x} dx.$$

Fazendo a mudança de variável $\tau = e^{-\theta x}$, temos

$$E(w(X)) = \int_1^0 -\frac{(1-\tau)^{a-1} b\theta \tau^b}{\theta\tau} d\tau = b\int_0^1 (1-\tau)^{a-1}\tau^{b-1} d\tau = bB(a,b). \qquad (3.14)$$

Portanto, a distribuição beta exponencial pertence à classe das distribuições exponenciais ponderadas, com função de peso $w(x) = (1-e^{-\theta x})^{a-1}$. ■

Analisamos a 1^a derivada da função de peso da distribuição beta exponencial, dada por

$$\frac{d}{dx}w(x) = (a-1)\theta e^{-\theta x}(1-e^{-\theta x})^{a-2}.$$

O parâmetro a é o único responsável pela mudança no sinal da derivada. Assim, quando $a > 1$ temos que a derivada da função de peso é positiva, implicando que a função de peso é crescente. Por outro lado, quando $0 > a > 1$ a derivada da função de peso é negativa, implicando que a função de peso é decrescente.

Com a função de peso explicitada, podemos escrevê-la na forma exponencial

$$w(x) = (1-e^{-\frac{\theta}{b}x})^{a-1} = e^{\log((1-e^{-\frac{\theta}{b}x})^{a-1})} = e^{(a-1)\log(1-e^{-\frac{\theta}{b}x})}.$$

Dessa forma temos $r = a-1$ e $t(x;\theta,\phi) = \log(1-e^{-\frac{\theta}{b}x})$. Para qualquer valor de r, $t(x;\theta,\phi)$ é uma função crescente e portanto, pelo teorema (3.2), se $r > 0$ temos que a distribuição beta exponencial é estocasticamente maior que a distribuição exponencial "pai"com parâmetro $\theta = b\theta$.

Com relação ao coeficiente de variação, assim como aconteceu com a distribuição exponencial *length biased*, a função $t(x;\theta,\phi)$ não é estritamente convexa. Então, calculamos o CV da distribuição beta exponencial e tentamos explicitar os casos em que ele é

maior ou menor do que 1. Juntando a definição 3.2 com (3.12) e (3.13) temos

$$CV_{be} = \frac{(\Psi'(b) - \Psi'(a+b))^{\frac{1}{2}}}{\Psi(a+b) - \Psi(b)}.$$

Vejamos a tabela 3.1 onde apresentamos o valor do CV_{be} para alguns valores de a e b.

TABELA 3.1: CV_{be} para alguns valores de parâmetros

a	b	CV_{be}
1	1	1
0,01	1	9,4312
2	1	0,7453
1	0,01	1
1	2	1
1	1	1
1	1	1
0,01	0,01	1,7314
2	2	0,7211

Calculando o CV_{be} para vários valores dos parâmetros a e b constatamos numericamente que se a for menor do que 1, o valor de CV_{be} é maior que 1. Isto implica que o desvio padrão é maior que a esperança, caracterizando sobredispersão. Por outro lado, se a for maior que 1, o valor de CV_{be} é menor do que 1 e teremos que o desvio padrão será menor que a esperança, caracterizando subdispersão. Vale ressaltar que para o caso em que $a = 1$, temos que $CV_{be} = 1$.

Concluímos então que a distribuição beta exponencial acomoda situações em que temos sobredispersão ($CV > 1$) e subdispersão ($CV < 1$).

Capítulo 4

Modelo de Mistura Padrão Exponencial Ponderado

Neste capítulo formulamos o modelo de mistura padrão utilizando as distribuições exponenciais ponderadas vistas no capítulo anterior para, em um cenário de longa duração, modelar o tempo de vida de indivíduos sob risco.

Inicialmente definimos o modelo de mistura padrão exponencial como um modelo base para a comparação com os modelos ponderados. Verificamos também o comportamento da função probabilidade de imunes, apresentada na seção 2.2, em cada um dos modelos.

Realizamos simulações com dois propósitos. O primeiro foi de verificar o erro de estimação cometido quando ignoramos a possibilidade de ponderação e utilizamos um modelo não ponderado. E segundo, analisamos o comportamento dos estimadores em cada um dos três modelos para diferentes tamanhos de amostra.

Para a estimação dos parâmetros envolvidos utilizaremos o pacote GAMLSS em R (Rigby & Stasinopoulos, 2007). Este pacote oferece recursos que serão úteis para verificar a qualidade do ajuste dos modelos aos conjuntos de dados. Mais detalhes sobre o GAMLSS podem ser encontrados no Apêndice A.

4.1 Modelo de Mistura Exponencial

O modelo de mistura padrão exponencial é definido admitindo-se que o tempo de vida dos indivíduos em risco segue uma distribuição exponencial com parâmetro θ. Recorrendo às equações em (2.3) e (2.4), basta substituir no modelo de mistura padrão as funções de densidade e sobrevivência da distribuição exponencial. Desse modo, o modelo de mistura padrão exponencial é dado por

$$S_{exp}(t) = 1 - p + pe^{-\theta t} \tag{4.1}$$

e

$$f_{exp}(t) = p\theta e^{-\theta t},$$

de modo que a função de probabilidade de imunes para o modelo exponencial é dada por

$$p_{exp}(t) = \frac{1-p}{1-p+pe^{-\theta t}}. \tag{4.2}$$

O comportamento dessa função depende de p e θ. O valor de p altera somente o ponto em que a curva começa. Já o valor de θ altera a velocidade de crescimento da curva. No gráfico 4.1 consideramos o valor $p = 0,25$. Nele vemos que o valor de θ altera somente a velocidade de crescimento da curva.

4.2 Modelo de Mistura Exponencial *Length Biased*

O modelo de mistura padrão exponencial *length biased* é definido admitindo-se que o tempo de vida dos indivíduos em risco segue uma distribuição exponencial *length biased*. Seguindo o procedimento anterior, basta substituir no modelo de mistura padrão as funções de densidade e sobrevivência da distribuição exponencial *length biased* encontradas em (3.4) e (3.5).

Assim, o modelo de mistura padrão exponencial *length biased* é dado por

$$S_{elb}(t) = 1 - p + pe^{-\theta t}(1 + \theta t) \tag{4.3}$$

e

$$f_{elb}(t) = pt\theta^2 e^{-\theta t}.$$

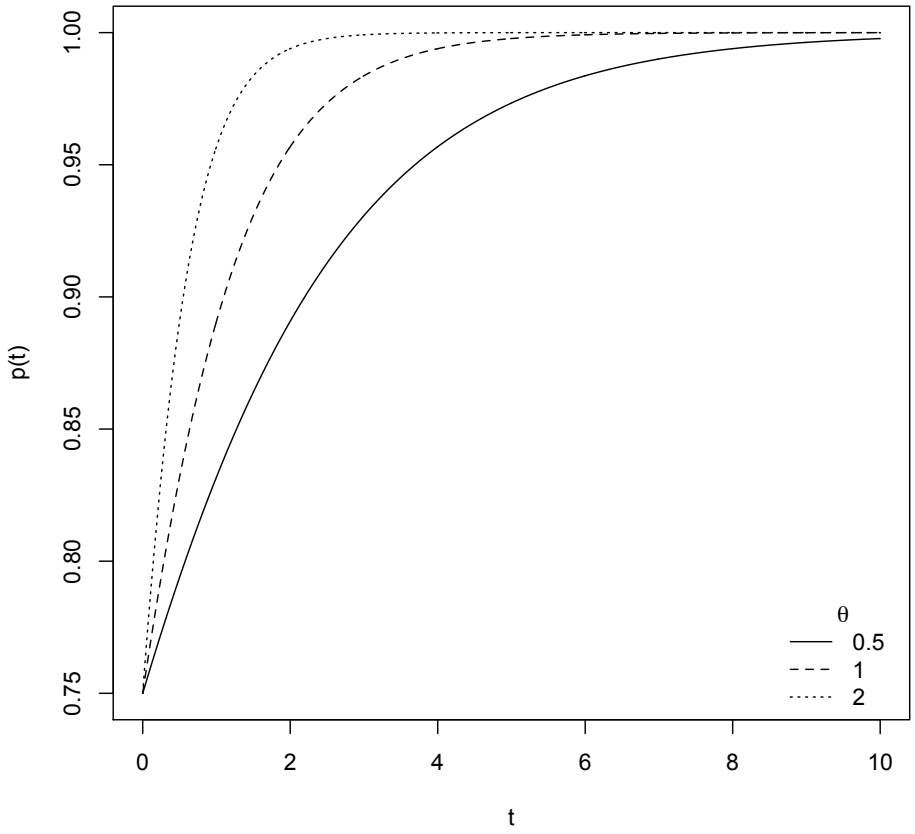

FIGURA 4.1: Função probabilidade de imunes do modelo exponencial para diferentes valores de θ.

Comparando (4.1) e (4.3) podemos ver que o modelo exponencial *length biased* é uma versão corrigida do modelo exponencial. Essa correção faz com que o modelo exponencial *length biased* se aproxime do modelo exponencial à medida que a taxa θ cresce. Isso ocorre pois quando θ cresce, o termo $e^{-\theta t}$ decresce muito mais rapidamente do que a correção $(1+\theta t)$ cresce. Uma outra relação interessante entre esses dois modelos é a seguinte:

$$S_{elb}(t) \geq S_{exp}(t). \tag{4.4}$$

Ilustremos graficamente os fatos citados acima. Na figura 4.2 temos as duas funções de sobrevivência com os parâmetros $p = 0,65$ e $\theta = 2$ e em seguida as mesmas funções, mas com o parâmetro $\theta = 20$.

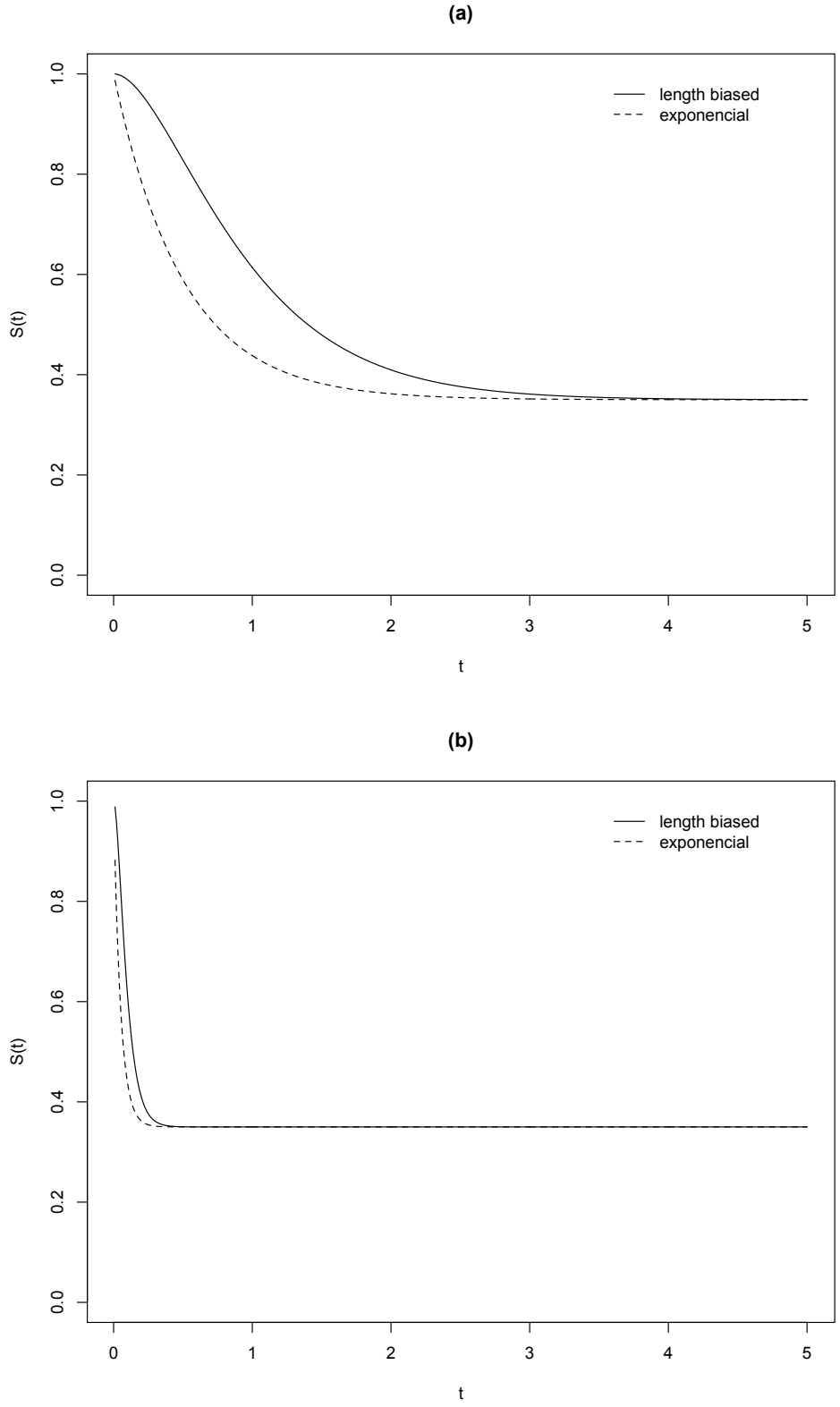

FIGURA 4.2: Funções de sobrevivência do modelo exponencial e exponencial *length biased*: (a) $\theta = 2$ e (b) $\theta = 20$

A função probabilidade de imunes do modelo exponencial *length biased* é dada por

$$p_{elb}(t) = \frac{1-p}{1-p+pe^{-\theta t}(1+\theta t)}. \tag{4.5}$$

Ilustramos na figura 4.3 como essa função se comporta quando variamos o parâmetro θ. Nesse gráfico consideramos $p = 0,25$.

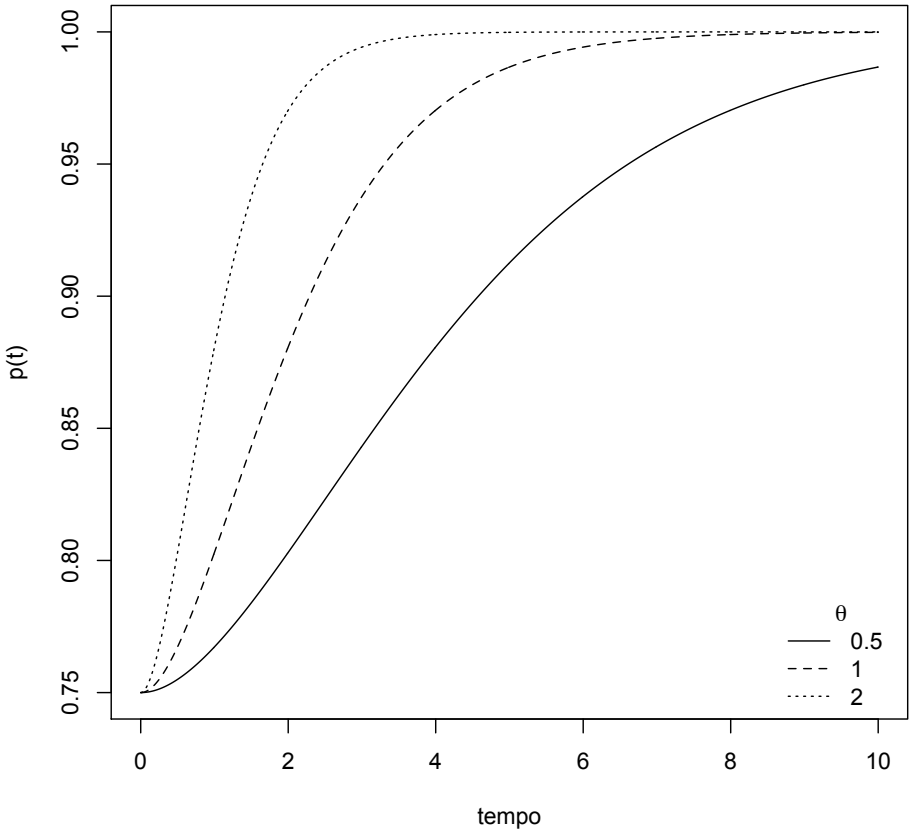

FIGURA 4.3: Função probabilidade de imunes do modelo exponencial *length biased* para diferentes valores de θ.

Pelo gráfico vemos que assim como acontece com o modelo exponencial, o valor de θ controla a velocidade do crescimento da curva. Comparando esse gráfico com o referente ao modelo exponencial, notamos que para o mesmo valor de θ a curva da função probabilidade de imunes cresce mais rapidamente nele do que no modelo *length biased*. Esse fato é esperado, visto que partindo de (4.4) e observando a estrutura da função probabilidade de imunes temos $p_{exp}(t) \geq p_{elb}(t)$.

4.3 Modelo de Mistura Beta Exponencial

De forma análoga aos modelos anteriores, definimos o modelo de mistura padrão beta exponencial ao admitirmos que o tempo de vida dos indivíduos em risco segue uma distribuição beta exponencial. As funções de sobrevivência e densidade do modelo beta exponencial, com parâmetros θ, a e b são as seguintes:

$$S_{be}(t) = 1 - p + p\frac{B_{1-e^{-\theta t}}(a,b)}{B(a,b)} \tag{4.6}$$

e

$$f_{be}(t) = \frac{p\theta e^{-b\theta t}(1-e^{-\theta t})^{a-1}}{B(a,b)},$$

lembrando que para $a = 1$ e $b = 1$, o modelo se reduz ao modelo exponencial com parâmetro θ. A função probabilidade de imunes do modelo beta exponencial é dada pela expressão

$$p_{be}(t) = \frac{1-p}{1-p+p\frac{B_{1-e^{-\theta t}}(a,b)}{B(a,b)}}. \tag{4.7}$$

A figura 4.4 mostra como essa função se comporta para alguns valores de a e b, quando fixamos os valores $\theta = 1$ e $p = 0,25$.

4.4 Estudo de Simulação

Tendo em vista a possibilidade de utilizar funções densidades ponderadas no modelo de mistura padrão exponencial e sob a premissa de que as observações obtidas podem não ser provenientes da v.a. em estudo e sim de uma versão modificada da mesma, surge a questão dos estimadores. Como eles se comportam quando ignoramos a ocorrência de uma v.a. ponderada?

Para responder em parte a essa questão, realizamos um estudo de simulação que consistiu em gerar amostras das duas distribuições ponderadas apresentadas anteriormente e confrontar as estimativas dos parâmetros desses modelos com as estimativas do modelo exponencial. Inicialmente geramos 5000 amostras, com censuras, de um modelo exponencial *length biased*. Para esse modelo os parâmetros fixados foram $\theta = 2$

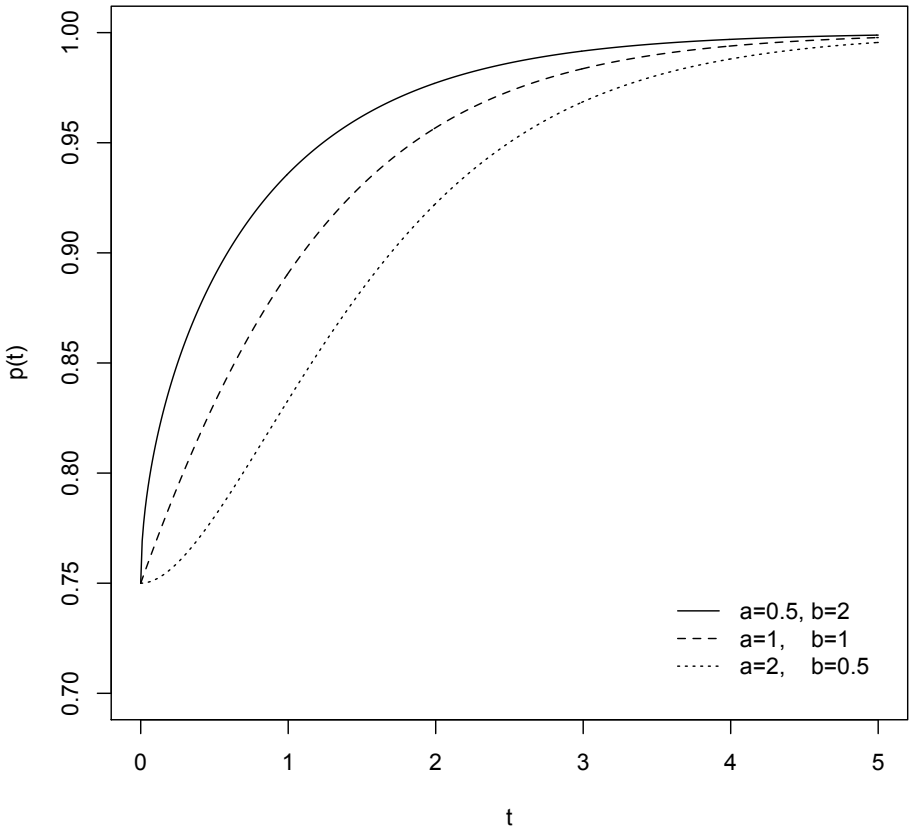

FIGURA 4.4: Função probabilidade de imunes do modelo beta exponencial para diferentes valores de (a, b).

e $p = 0,75$. Para cada amostra estimamos os parâmetros do modelo exponencial e do modelo exponencial *length biased*. Repetimos esse procedimento de geração e estimação para o modelo beta exponencial, em que utilizamos os parâmetros $a = b = \theta = 2$ e $p = 0,75$. Comparamos os modelos ponderados com o modelo exponencial através da média, do desvio padrão (DP) e do erro quadrático médio (EQM) das estimativas. Variamos o tamanho das amostras em $n = 50, 100$ e 200 para tentar avaliar como ele afeta estas propriedades dos estimadores. A estimação dos parâmetros foi efetuada utilizando o pacote GAMLSS em R (Rigby & Stasinopoulos, 2007). Para o modelo exponencial *length biased*, os resultados estão apresentados na tabela 4.1.

Os resultados mostram que as médias das estimativas não foram muito afetadas pelo aumento no tamanho da amostra, mesmo com uma amostra pequena elas já ficaram próximos dos valores verdadeiros. A estimativa do parâmetro θ no modelo exponencial foi

TABELA 4.1: Resultados da simulação com o modelo exponencial *length biased*.

Tamanho da amostra (n)	modelo		Média	DP	EQM
50	exponencial	p	0,7541	0,0669	0,0045
		θ	0,9751	0,1291	1,0670
	length biased	p	0,7499	0,0662	0,0044
		θ	2,0318	0,2560	0,0665
100	exponencial	p	0,7536	0,0476	0,0023
		θ	0,9670	0,0902	1,0751
	length biased	p	0,7497	0,0471	0,0022
		θ	2,0148	0,1791	0,0323
200	exponencial	p	0,7542	0,0333	0,0011
		θ	0,9641	0,0633	1,0772
	length biased	p	0,7504	0,0330	0,0011
		θ	2,0085	0,1254	0,0158

aproximadamente metade do valor da mesma no modelo exponencial *length biased*. Isso fez com que o valor do EQM para o modelo exponencial ficasse em torno de 1 enquanto que o EQM do modelo exponencial *length biased* se manteve próximo de zero e diminuiu quando aumentamos o tamanho da amostra, conforme o esperado. Comportamento contrário ao EQM de $\hat{\theta}$ no modelo exponencial que apresentou um leve crescimento. O comportamento dos desvios padrão de $\hat{\theta}$ foi de acordo com o esperado. Eles diminuíram com o aumento do tamanho amostral. Os valores para o modelo *length biased* foram aproximadamente o dobro dos valores do modelo exponencial.

Sobre o parâmetro p, o EQM e o DP dos dois modelos foram muito semelhantes e um fato curioso aconteceu, pois as médias das estimativas também foram semelhantes nos dois modelos. Esse fato é citado em Shao & Zhou (2004). Estes autores observaram que havia uma grande diferença no desvio, mas que as estimativas de p praticamente não diferiam. Eles justificaram esse acontecimento dizendo que a estimativa de p é predominantemente determinada pelos valores dos dados próximos da cauda direita, enquanto que a função de verossimilhança é influenciada por todo conjunto de dados mais igualmente, causando a diferença no desvio global.

Vejamos na tabela 4.2, os resultados da simulação envolvendo o modelo beta exponencial.

TABELA 4.2: Resultados da simulação com o modelo beta exponencial.

Tamanho da amostra (n)	modelo		Média	DP	EQM
50	exponencial	p	0,7483	0,0653	0,0042
		θ	2,3970	0,3032	0,2495
	beta exponencial	p	0,7471	0,0720	0,0051
		a	1,8465	0,4112	0,1926
		b	1,5392	0,1358	0,2307
		θ	2,2207	0,3075	0,1432
100	exponencial	p	0,7522	0,0451	0,0020
		θ	2,3724	0,2098	0,1827
	beta exponencial	p	0,7521	0,0459	0,0021
		a	1,8563	0,2882	0,1037
		b	1,5782	0,1059	0,1891
		θ	2,2129	0,2108	0,0897
200	exponencial	p	0,7511	0,0318	0,0010
		θ	2,3693	0,1459	0,1576
	beta exponencial	p	0,7510	0,0315	0,0009
		a	1,8815	0,2068	0,0568
		b	1,6134	0,0816	0,1561
		θ	2,2242	0,1465	0,0717

Os resultados da simulação mostram que as médias das estimativas não sofreram grandes alterações com o aumento do tamanho da amostra. Novamente a média das estimativas do parâmetro p para ambos os modelos ficaram muito próximas, assim como os valores de DP e de EQM de $\hat{p}$. Com relação aos outros parâmetros, eles se comportaram de forma inesperada. As médias não ficaram tão próximas dos valores verdadeiros, mesmo que ao aumentar o valor da amostra a aproximação tenha melhorado, indicando que talvez o aumento do tamanho amostral não tenha sido suficiente. Os valores de DP e de EQM dos estimadores dos parâmetros a, b e θ diminuíram à medida que o tamanho amostral

aumentou, conforme o esperado.

Em simulações preliminares foi constatado que o algoritmo de maximização do GAMLSS teve problemas de convergência no modelo beta exponencial. As estimativas dos parâmetros se apresentaram muito dependentes dos valores iniciais. Essa dependência não só afetou as estimativas, mas também o tempo gasto com o processamento, já que o algoritmo demorou muito mais para convergir. A solução encontrada foi sugerida em Rigby & Stasinopoulos (2005) nas réplicas aos comentários de outros pesquisadores. Os autores afirmam que problemas de convergência podem ocorrer quando o modelo que está sendo ajustado é muito inadequado para os dados ou quando os valores iniciais utilizados são extremamente pobres. Nesse segundo caso foi sugerido um ajuste preliminar com valores iniciais quaisquer e em seguida utilizar as estimativas desse primeiro ajuste como valores iniciais para um segundo ajuste. Essa estratégia apresentou melhoras nas estimativas, talvez o tamanho amostral utilizado não seja grande o suficiente para um ajuste mais preciso.

4.5 Exemplo - Dados de Leucemia

Para este exemplo aproveitamos o conjunto de dados que foi apresentado no capítulo 2. Os dados são referentes a um estudo sobre recorrência de leucemia em pacientes que foram submetidos a dois tipos de transplantes, alogêncio e autogênico. Utilizaremos o pacote GAMLSS para o ajuste dos três modelos abordados nesse capítulo: o modelo exponencial, o modelo exponencial *length biased* e o modelo beta exponencial. Inicialmente realizaremos a análise em cada grupo separadamente. Em seguida consideraremos o grupo como uma covariável ligada à fração de cura. Utilizando as ferramentas do GAMLSS compararemos o desempenho dos modelos.

4.5.1 Grupo 1 - Alogênico

O primeiro grupo consiste de 46 observações, sendo que dessas, 13 são censuradas. Ajustamos os três modelos e obtivemos as seguintes estimativas para os parâmetros de cada um. Calculamos os intervalos de confiança de 95% assintóticos e via *bootstrap* não-

paramétrico com 5000 replicações.

TABELA 4.3: Estimativas dos parâmetros e intervalo de confiaça - Grupo 1.

Modelo	Parâmetro	Estimativa	Intervalos de confiança de 95%	
			Assintótico	Bootstrap
exponencial	p	0,7278	(0,5726 ; 0,8421)	(0,5908 ; 0,8640)
	θ	1,4343	(0,9731 ; 2,1141)	(0,9151 ; 2,1914)
length biased	p	0,7214	(0,5682 ; 0,8333)	(0,5878 ; 0,8522)
	θ	2,9889	(2,3154 ; 3,8583)	(2,0395 ; 4,4617)
beta exponencial	p	0,7286	(0,5727 ; 0,8432)	(0,5905 ; 0,8631)
	a	0,9678	(0,5128 ; 1,8267)	(0,7103 ; 1,4791)
	b	0,9079	$(8{,}2\text{x}10^{-11}$; $9{,}9\text{x}10^{10})$	(0,7534 ; 1,2729)
	θ	1,5264	$(8{,}4\text{x}10^{-11}$; $2{,}7\text{x}10^{10})$	(0,9270 ; 2,4304)

Pela tabela 4.3 vemos que o valor das estimativas de p é praticamente o mesmo nos três modelos, esse fato era esperado como vimos anteriormente. Notemos a proximidade nas estimativas dos modelos exponencial e beta exponencial. Isso ocorre devido à estimativa do parâmetro a ser próxima de 1, que faz com que o modelo beta exponencial se aproxime do modelo exponencial. Os intervalos de confiaça confirmam esse fato, já que o valor 1 está presente no intervalo para o parâmetro a. Ainda sobre os intervalos de confiança, os intervalos assintóticos e os obtidos por *bootstrap* ficaram próximos, exceto pelos intervalos assintóticos para os parâmetros b e θ do modelo beta exponencial. Esses intervalos apresentaram amplitudes muito grandes, reflexo do alto valor para a variância dos estimadores.

TABELA 4.4: Medidas de qualidade do ajuste - Grupo 1.

Modelo	Desvio global	AIC	BIC
exponencial	92,45	96,45	100,11
length biased	105,61	109,61	113,26
beta exponencial	92,45	100,45	107,76

Na tabela 4.4 temos a confirmação da proximidade entre os modelos exponencial

e beta exponencial, pois as medidas de qualidade de ajuste para esses modelos foram bem próximas. Os valores do desvio global, AIC e BIC do modelo *length biased* foram maiores que os outros dois, indicando que esse modelo é o menos adequado para esses dados.

Os gráficos *wormplot* (Rigby & Stasinopoulos, 2007), gráficos QQ sem a tendência, indicam os modelos exponencial e beta exponencial como melhor ajustados, pois todos os pontos pertencem à região central delimitada pelas duas curvas tracejadas. Pela figura 4.5(b) vemos que para o modelo *length biased* temos pontos não pertencentes a essa região central, indicando que esse modelo não é adequado para esse conjunto de dados.

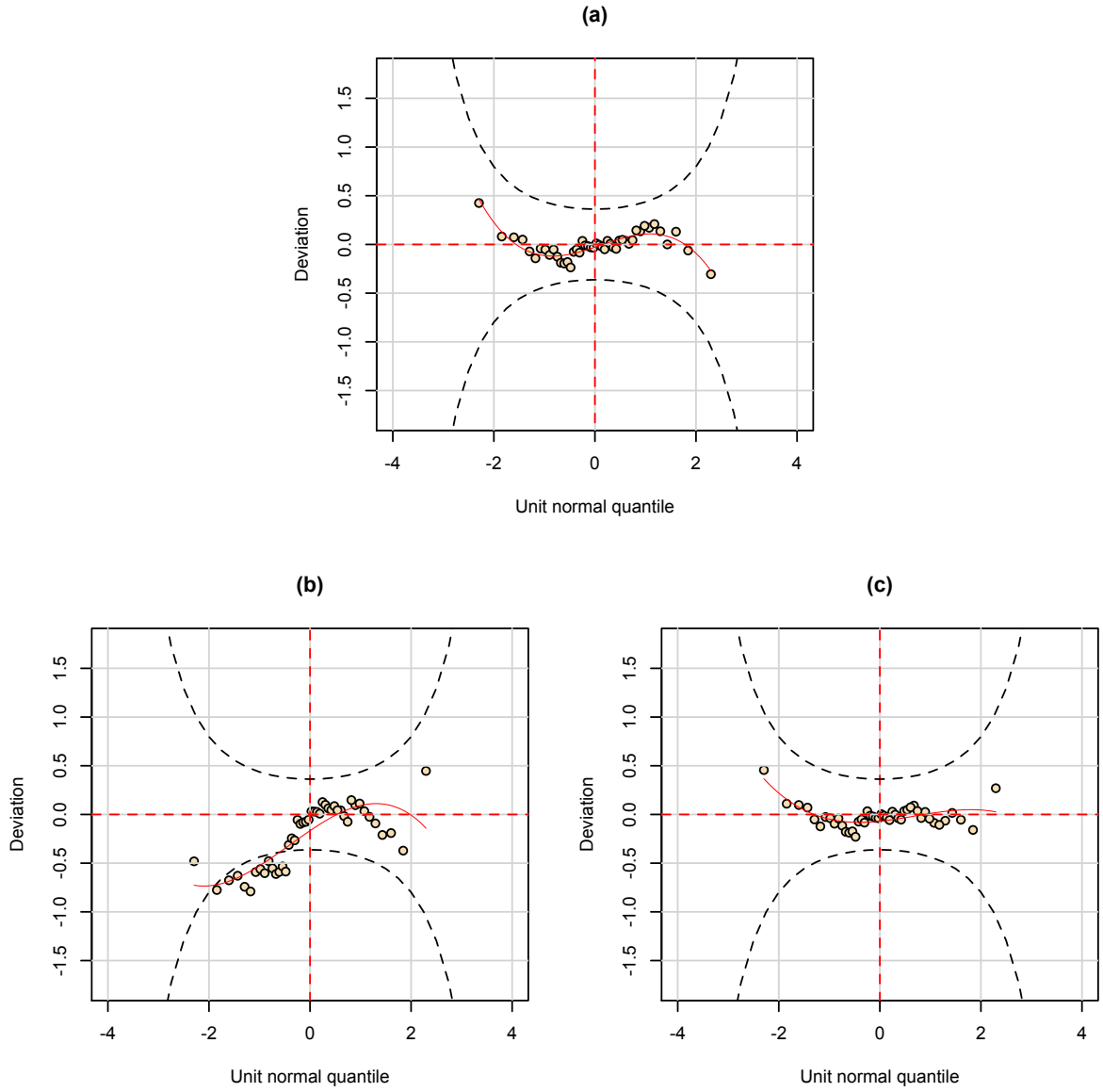

FIGURA 4.5: Gráficos *wormplot*: (a) exponencial (b) *length biased* e (c) beta exponencial.

Por último, analisamos as curvas de sobrevivência preditas por cada um dos modelos e comparamos estas com a curva estimada não parametricamente pelo estimador de Kaplan-Meier.

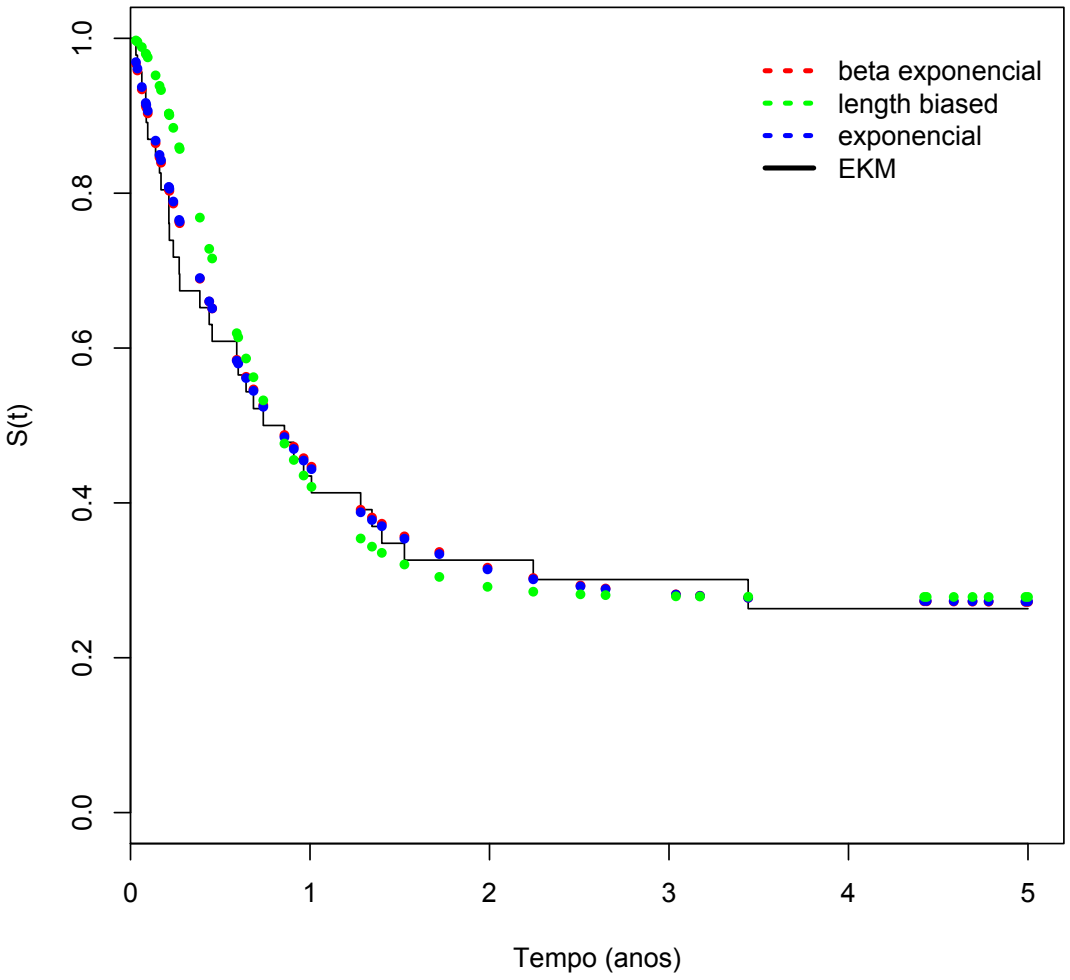

FIGURA 4.6: Curvas de sobrevivência preditas por cada um dos modelos para o grupo 1 e estimativa de Kaplan-Meier.

Pela figura 4.6 confirmamos que os modelos exponencial e beta exponencial são bem próximos e que o modelo *length biased* ajusta mal os dados, sobrestimando a curva no início passando a subestimá-la a partir de 1 ano até aproximadamente se igualar às demais após 4 anos. Concluímos que para o grupo 1 o modelo exponencial é mais indicado, já que apresenta menores valores de AIC e BIC que o modelo beta exponencial, além de ser um modelo muito mais simples que ele.

4.5.2 Grupo 2 - Autogênico

O segundo grupo consiste de 44 observações dessas quais nove são censuradas. Novamente ajustamos os três modelos e obtivemos as seguintes estimativas para os parâmetros:

TABELA 4.5: Estimativas dos parâmetros e intervalos de confiança - Grupo 2.

Modelo	Parâmetro	Estimativa	Intervalos de confiança de 95%	
			Assintótico	Bootstrap
exponencial	p	0,7966	(0,6474 ; 0,8931)	(0,6674 ; 0,9102)
	θ	2,6809	(1,8939 ; 3,7948)	(2,0225 ; 3,5137)
length biased	p	0,7972	(0,6492 ; 0,8931)	(0,6691 ; 0,9095)
	θ	5,3955	(4,2382 ; 6,8688)	(4,0908 ; 7,0458)
beta exponencial	p	0,7958	(0,6466 ; 0,8924)	(0,6363 ; 0,8872)
	a	2,5096	(1,8801 ; 3,3500)	(1,6005 ; 5,9622)
	b	1,5688	(1,5048 ; 1,6355)	(1,1125 ; 2,3946)
	θ	3,1700	(3,0474 ; 3,2976)	(2,2499 ; 2,3946)

Pela tabela 4.5 vemos que novamente os valores das estimativas de p ficaram bem próximos e que o modelo beta exponencial está bem diferenciado do modelo exponencial, já que os intervalos de confiança para o parâmetro a não contêm o valor 1.

TABELA 4.6: Medidas de qualidade do ajuste - Grupo 2.

Modelo	Desvio global	AIC	BIC
exponencial	45,01	49,01	52,58
length biased	34,51	38,51	42,08
beta exponencial	32,97	40,97	48,11

A tabela 4.6 mostra que o modelo exponencial apresenta os maiores valores para o desvio global, AIC e BIC, indicando que entre os três modelos analisados ele é o que pior se ajusta aos dados.

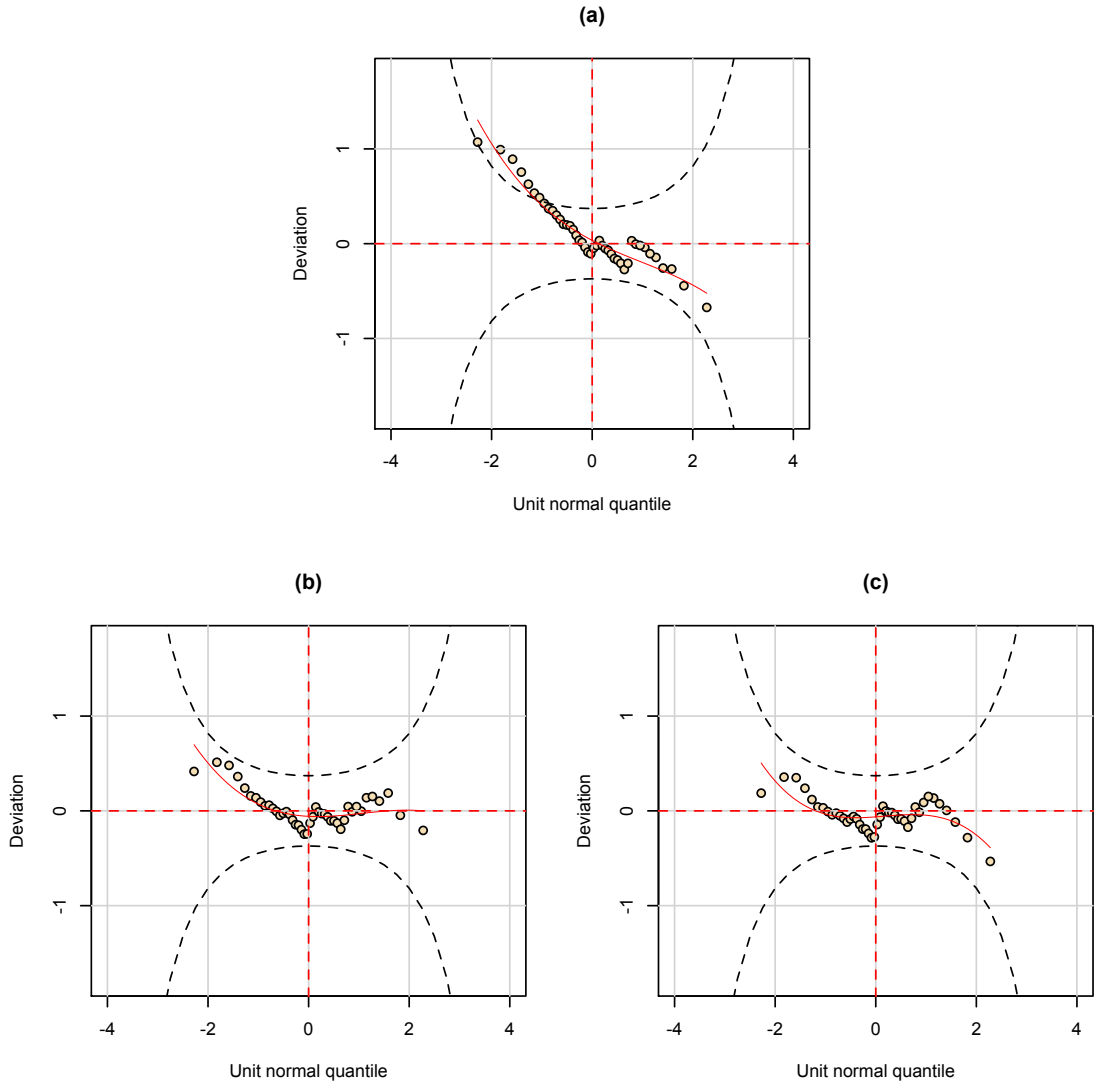

FIGURA 4.7: Gráficos *wormplot*: para o grupo 2 (a) exponencial (b) *length biased* e (c) beta exponencial

Pela figura 4.7(a) vemos que para o modelo exponencial temos pontos não pertencentes a essa região central, indicando que esse modelo não é adequado para esse conjunto de dados e pela figura 4.8 confirmamos esse modelo não é adequado, pois ele não segue a tendência do estimador de Kaplan-Meier. Os modelos *length biased* e beta exponencial produziram curvas de sobrevivência preditas muito parecidas. Concluímos então que para o grupo 2 o modelo *length biased* é o mais adequado já que apresenta menores valores de AIC e BIC que o modelo beta exponencial e ainda é um modelo mais simples, pois envolve dois parâmetros contra quatro do modelo beta exponencial.

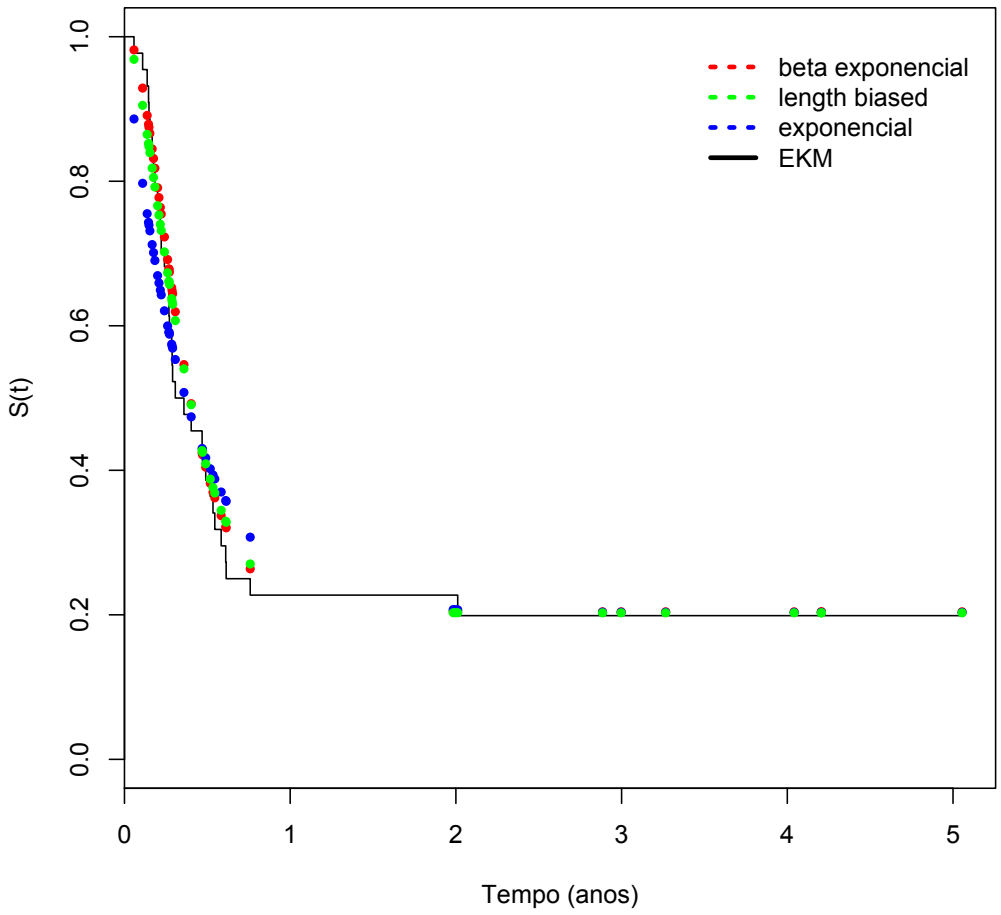

FIGURA 4.8: Curvas de sobrevivência preditas por cada um dos modelos para o grupo 2 e estimativa de Kaplan-Meier.

4.6 Comparação Entre os Tipos de Transplante

Vimos que para o conjunto de dados de leucemia o modelo beta exponencial obteve um desempenho muito próximo dos modelos que foram ditos melhor ajustados para cada grupo. Sendo assim, podemos utilizá-lo nos dois grupos, o que possibilita efetuarmos uma comparação da eficácia dos tratamentos os quais cada grupo foi submetido. Para isso utilizaremos a função probabilidade de imunes.

Para o modelo beta exponencial a função probabilidade de imunes foi obtida em (4.7). Calculamos a função para os dois grupos e o gráfico delas pode ser visto na figura 4.9.

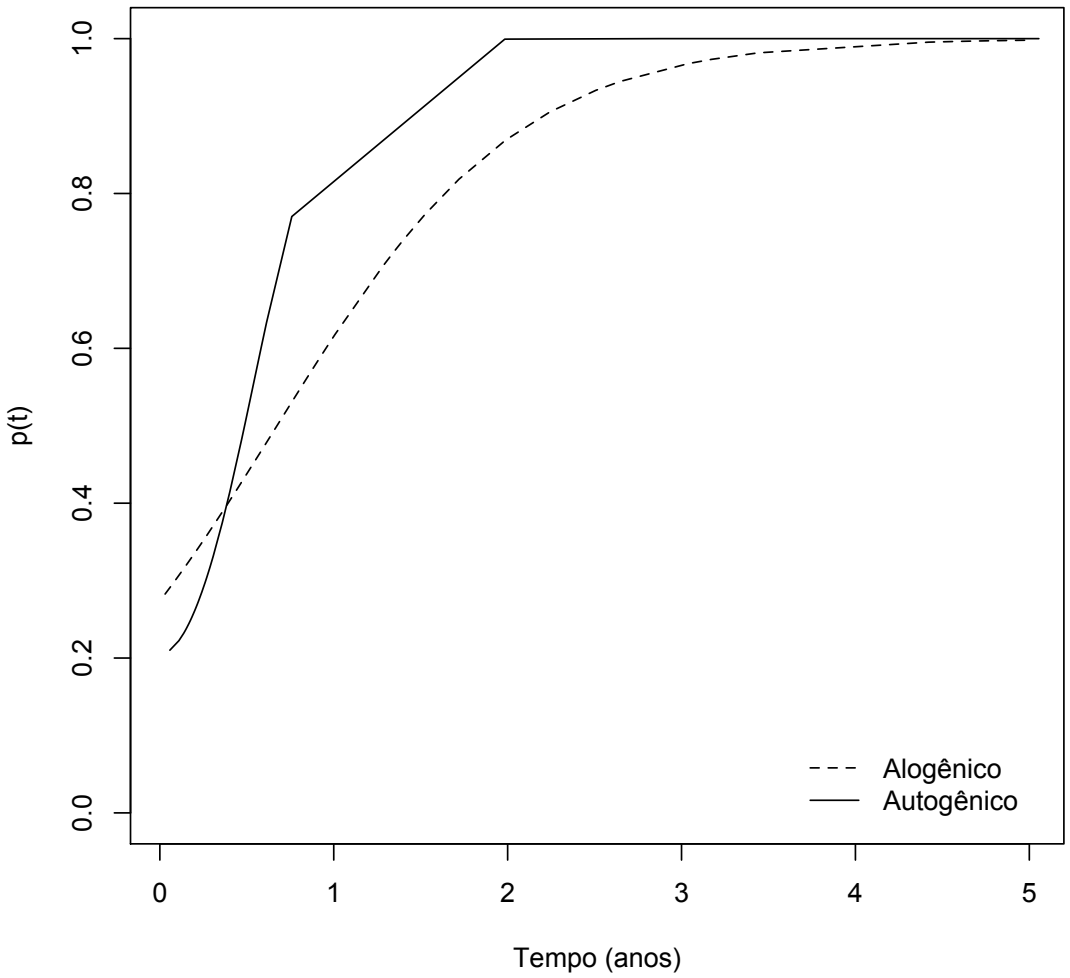

FIGURA 4.9: Função probabilidade de imunes do modelo beta exponencial para os dois grupos

Pelo gráfico, vemos que ambos os tipos de transplantes começam com uma probabilidade de cura baixa, mas a curto prazo o transplante alogênico tem um aumento maior nessa probabilidade. No entanto, a partir de um ano, o tratamento autogênico apresenta um crescimento bem acentuado e permanece com uma probabilidade de cura superior à do transplante alogênico. Desse modo é possível determinar a eficácia de cada tratamento a curto e a longo prazo, dando uma perspectiva melhor ao indivíduo que vá se sujeitar a algum deles. Observemos que essa conclusão é análoga à que foi obtida na seção 2.2.1, mas no caso anterior foi considerado o modelo exponencial, que mais tarde concluímos ser inadequado para o grupo 2. Se observarmos as curvas para os dois modelos notamos

que para o grupo 2, a curva do modelo beta exponencial está bem mais alta. Isso reflete o fato de o modelo beta exponencial conseguiu explicar melhor os dados.

Para determinar se os grupos têm efeito sobre a fração de cura podemos incluir a variável grupo como uma covariável do modelo ligada à fração de cura. Ajustamos novamente o modelo beta exponencial e obtivemos que o p-valor do coeficiente da covariável grupo não é significativo (p-valor=0.62886). Desse modo, concluímos que não há efeito da variável grupo sobre a fração de cura.

Capítulo 5

Considerações Finais

A proposta desse trabalho foi juntar o conceito das distribuições ponderadas com o modelo de mistura padrão em análise de sobrevivência. Fomos motivados por situações em que as amostras não são realmente amostras aleatórias simples mas sim amostras ponderadas.

Neste trabalho vimos brevemente o que caracteriza um modelo de longa duração em análise de sobrevivência. Apresentamos o modelo de mistura de padrão e como ele é construído a partir de uma variável binária que classifica os indivíduos de uma população em imunes e suscetíveis.

A investigação sobre as distribuições ponderadas se restringiu à classe das distribuições exponenciais ponderadas, que inclui a distribuição exponencial *length biased* e a distribuição beta exponencial. Nessa classe pudemos verificar mais facilmente a influência da função peso na dispersão das distribuições.

Com os estudos desenvolvidos nesse texto observamos que realizar uma análise tradicional quando a amostra é proveniente de uma v.a. ponderada resulta em estimativas bastante viesadas para os parâmetros dos modelos, exceto a fração de cura. Através de simulações pudemos quantificar esse erro, quando comparamos as estimativas em cada modelo utilizando a média, o desvio padrão e erro quadrático médio.

Extensões desse trabalho seriam investigar novas distribuições exponenciais ponderadas e suas propriedades e abordar o modelo de mistura padrão exponencial ponderado sob o ponto de vista bayesiano.

Referências Bibliográficas

[1] Balka, J., Desmond, A. F., McNicholas, P. D. (2009). *Review and Implementation of Cure Models Based on First Hitting Times for Wiener Processes.* Lifetime Data Analysis 15, 147-176.

[2] Bayarri, M. J., DeGroot, M. H. (1992). *A "BAD" View of Weighted Distributions and Selection Models.* Bayesian Statistics 4, 17-33.

[3] Berkson, J., Gage, R. P. (1952). *Survival Cure of Cancer Patients Following Treatment.* Journal of the American Statistical Association, 47(259), 501-515.

[4] Boag, J. W. (1949). *Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy.* Journal of the Royal Statistical Society B, 11(1), 15-35.

[5] Cnann, A. (1985). *Survival Models with Two Phases and Length Biasing Sampling.* Communications in Statistics 14(4), 861-886.

[6] Chen, M.-H., Ibrahim, J. G., Sinha, D. (1999). *A New Bayesian Model for Survival Data with a Surviving Fraction.* Journal of American Statistical Association, 94(447), 909-919.

[7] Colosimo, E. A., Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada.* Edgar Blücher, São Paulo, SP.

[8] de Castro, M., Cancho, V. G., Rodrigues, J. (2009). *A Hands-on Approach for Long-Term Survival Models Under the GAMLSS Framework.* Computer Methods and Programs in Biomedicine 2009.

[9] Gupta, R. C., Kirmani, S. N. U. A. (1990). *The Role of Weighted Distributions in Stocastic Modeling.* Communications in Statistics, Theory and Methods 19(9), 3147-3162.

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

[11] Kaplan, E.L., Meier, P. (1958). *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association 53, 457–481.

[12] Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, T. W., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B., Ramsay, N. K. C. (1987). *Comparsion of Autologous and Allogenic Bone Marrow Transplantation for Treatment of High-risk Refractory Acute Lymphoblastic Leukemia.* New England Journal of Medicine, 317(8), 461-467.

[13] Kokonendiji, C. C., Mizère, D., Balakrishnan, N. (2008). *Connections of the Poisson Weigh Function to Overdispersion and Underdispersion.* Journal of Statistical Planning and Inference 138, 1287-1296.

[14] Larose, D. T., Key, D. K. (1996). *Weighted Distributions Viewed in the Context of Model Selection: A Bayesian Perspective* The Journal of the Spanish Statistical Society, Vol. 5, 1, 227-246.

[15] Li, C., Taylor, J. M. G., Sy, J. (2001). *Identifiability of Cure Models.* Statistics & Probability Letters 54, 389-395.

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

[17] Nadarajah, S., Kotz, S. (2006). *The Beta Exponencial Distribution* Reliability Engineering and System Safety 91, 689–697.

[18] Navarro, J., Ruiz, J. M., Del Aguila, Y. (2001). *Parametric Estimation from Weighted Samples.* Biometrical Journal 43(3), 297-311.

[19] Pakes, A. G., Navarro, J., Ruiz, M. J., Del Aguila, Y. (2003). *Characterizations using weighted distributions.* Journal of Statistical Planning and Inference 116, 389-420.

[20] Patil, G. P. (2002). *Weighted Distributions.* Encyclopedia of Environmetrics vol 4, 2369-2377.

[21] Rao, C. R. (1985). *Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent?.* A Celebration of Statistics, New York: Springe, 543-569.

[22] Rigby, R. A., Stasinopoulos, D.M. (2005) *Generalized Additive Models for Location Scale and Shape.* Applied Statistics, 54(3) 507-554.

[23] Rigby, R. A., Stasinopoulos, D.M. (2007) *Generalized Additive Models for Location Scale and Shape (GAMLSS) in R.* Journal of Statistical Software, 23(7), 1-46.

[24] Rigby, R. A., Stasinopoulos, D.M., Akantziliotou, C. (2008) *Instruction on How to Use the GAMLSS Package in R.* www.gamlss.com.

[25] Rodrigues, J. (2008). *Uma Visão Unificada do Mecanismo de Seleção Populacional.* GIB-UFSCar 21/11/2008

[27] Rodrigues, J., Cancho, V. G., de Castro, M.(2008). *Teoria Unificada de Análise de Sobrevivência.* SINAPE 2008.

[27] Rodrigues, J., Cancho, V. G., de Castro, M., Louzada-Neto, F. (2009). *On the Unification of the Long-Term Models* Statistics & Probability Letters, 79(6), 753-759.

[28] Shao, Q., Zhou, X. (2004). *A New Parametric Model for Survival Data with Long-term Survivors.* Statistics in Medicine 23, 3525-3543.

[29] Yakovlev, A. Y., Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and their Biostatistical Applications.* World Scientific, Singapore.

Apêndice A

Pacote GAMLSS para R

O GAMLSS é uma estrutura geral para o ajuste de modelos de regressão em que a distribuição da variável resposta não está restrita a nenhuma família de distribuições. A sigla vem do inglês, Generalized Additive Models for Location, Scale and Shape, que pode ser traduzido como Modelos Aditivos Generalizados para Locação, Escala e Forma.

Modelos GAMLSS são modelos de regressão semi-paramétricos. Eles são paramétricos no sentido em que admitimos uma distribuição de probabilidades para a variável resposta e o termo "semi"vem do fato de que na estimação dos parâmetros da distribuição da resposta, como função das variáveis explicativas, podem ser empregadas técnicas não-paramétricas de suavização de funções.

A proposta do GAMLSS foi feita por Rigby & Stasinopoulos (2005) com o intuito de superar limitações dos modelos lineares generalizados (GLM) e modelos aditivos generalizados (GAM).

Como nos modelos GAMLSS a distribuição da variável resposta não está restrita a nenhuma família de distribuições, podemos trabalhar com distribuições altamente assimétricas, discretas ou contínuas. A parte sistemática do modelo é expandida de forma a permitir a modelagem dos parâmetros da distribuição da resposta como função linear e/ou não linear, paramétrica e/ou não paramétrica aditiva das variáveis explicativas e/ou efeitos aleatórios.

Sejam $y^T = (y_1, y_2, ..., y_n)$ o vetor da variável resposta Y e $g_k(.)$ a função de ligação dos parâmetros da distribuição de Y com as variáveis explicativas. O modelo

estatístico geral do GAMLSS para os k parâmetros da distribuição de Y é dado por

$$g_k(\theta_k) = X_k\beta_k + \sum_{j=1}^{J_k} Z_{jk}\gamma_{jk}, \tag{A.1}$$

em que $\beta_k^T = (\beta_{1k}, \beta_{2k}, ..., \beta_{J_k k})$ é o vetor do parâmetros, X_k é a matriz do modelo com dimensão $(n \times J_k)$. Z_{jk} é uma matriz de delineamento conhecida e γ_{jk} é um vetor de parâmetros de efeitos aleatórios.

Como submodelo, temos a formulação *semi-paramétrica aditiva* do modelo GAMLSS dada por

$$g_k(\theta_k) = X_k\beta_k + \sum_{j=1}^{J_k} h_{jk}(x_{jk}), \tag{A.2}$$

em que x_{jk} são colunas da matriz X_k e h_{jk}, (j=1,2,...J_k) são funções suavizadoras não-paramétricas.

Quando não existem termos aditivos em nenhum dos parâmetros da distribuição de Y, temos um modelo GAMLSS *linear simples* dado por

$$g_k(\theta_k) = X_k\beta_k. \tag{A.3}$$

O modelo (A.2) pode ser estentido para o caso não linear

$$g_k(\theta_k) = h_k(X_k\beta_k) + \sum_{j=1}^{J_k} h_{jk}(x_{jk}), \tag{A.4}$$

que chamaremos de modelo GAMLSS *não-linear semiparamétrico aditivo*. E quando no modelo (A.4) não existem termos aditivos temos o modelo GAMLSS *não-linear* dado por

$$g_k(\theta_k) = h_k(X_k\beta_k). \tag{A.5}$$

A.1 Pacote GAMLSS em R

Juntamente com a formulação dos modelos GAMLSS, Rigby & Stasinopoulos (2005) publicaram um pacote com a implementação do GAMLSS no R. A atual versão do pacote suporta distribuições da variável resposta com até quatro parâmetros.

A sintaxe do GAMLSS é bem simples e foi baseada na dos pacotes GLM e GAM, fazendo com que a transição destes pacotes para o pacote GAMLSS ocorra sem problemas.

Dada essa simplicidade, a inclusão de covariáveis no modelo é bem fácil, bastando o acréscimo de um argumento na função *gamlss* para especificar quais são as covariáveis e a qual parâmetro elas são relacionadas.

Existem extensões do pacote GAMLSS que ampliam seu uso, esses pacotes adicionais são

- **gamlss.cens** para ajustar modelos com dados censurados,
- **gamlss.dist** contendo mais famílias de distribuições,
- **gamlss.mx** para ajustar distribuições de misturas finitas,
- **gamlss.nl** para ajustar modelos não lineares e
- **gamlss.tr** para distribuições truncadas.

A extensão **gamlss.dist** permite também a implementação de novas famílias de distribuições, o procedimento é simples e requer, além das funções de densidade e distribuição acumulada, todas as derivadas de primeira e segunda ordem do logaritmo da função densidade. No caso das derivadas, não necessariamente elas precisam ser analíticas podendo ser obtidas através de algum método numérico de diferenciação.

O pacote dispõe de mais de um algoritmo de maximização. O primeiro, desenvolvido por Rigby & Stasinopoulos (2005), não depende de valores iniciais muito precisos para garantir a convergência. O segundo é uma generalização do algoritmo de Cole & Green que utiliza as derivadas da função de densidade e tem um desempenho melhor quando as estimativas dos parâmetros do modelo são altamente autocorrelacionadas. Existe ainda um terceiro algoritmo que é uma mistura dos dois algoritmos citados anteriormente.

Uma das vantagens desse pacote é a análise de resíduos e a verificação de ajuste do modelo aos dados. Facilmente obtém-se os gráficos da densidade estimada dos resíduos, gráficos QQ dos resíduos e o *wormplot* que nos dá indícios do ajuste do modelo, assim como as medidas AIC, BIC e desvio global.

A.1.1 Exemplo

Vamos ilustrar a funcionalidade do programa através de um exemplo, por simplicidade os dados foram simulados. Consideremos tempos de vida censurados gerados de uma distribuição $Gama(2,\theta)$ com p da fração de cura igual a $0,75$. Implementamos o modelo de mistura padrão com distribuição exponencial *length biased* para o tempo de vida. O modelo é dado pelas seguintes funções de densidade e sobrevivência

$$f_{ldp}(x) = px\theta^2 e^{-\theta x} \quad \text{(A.6)}$$

$$S_{ldp}(x) = 1 - p + pe^{-\theta x}(1+\theta x). \quad \text{(A.7)}$$

É importante lembrar que a estrutura do pacote é fixa quanto à nomenclatura dos parâmetros, são eles μ, σ, ν, τ. Nesse caso, p é o parâmetro μ e θ é o parâmetro σ.

Foi necessária a implementação dessa família exponencial ponderada de longa duração, para isso a extensão **gamlss.dist** foi utilizada, assim como a extensão **gamlss.cens** pois estaremos modelando a versão para dados censurados da família implementada.

Falta ainda uma série de especificações como as funções de ligação que serão disponibilizadas, informar se a distribuição é contínua ou discreta, o número de parâmetros envolvidos, os valores que esses parâmetros podem assumir, valores inicias para a maximização e as derivadas do logaritmo da função densidade. Com a nova família implementada, vamos ajustar o modelo aos dados. O comando utilizado foi o seguinte:

```
m.pond=gamlss(Surv(tempo,censura)~1,family=cens(EP),n.cyc=1000)
```

Nesse comando especificamos que estamos utilizando a versão censurada da família implementada, que foi chamada de EP e estipulamos como 1000 o número máximo de iterações para garantir a convergência do algoritmo.

O programa foi bem veloz e a convergência do algoritmo é atingida em 2 ou 3 iterações. Utilizamos a função *summary* para obter sumário das informações contidas no objeto GAMLSS recém criado.

```
> summary(m.pond)
******************************************************************
Family:  c("EPrc", "right censored Lenght Biased Weighted Exponencial Mixture Model" )
```

```
Call:  gamlss(formula = Surv(tempo, censura) ~ 1, family = cens(EP),      n.cyc = 1000)

Fitting method: RS()

------------------------------------------------------------------
Mu link function:  logit
Mu Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
 1.174e+00   2.521e-01   4.658e+00   1.006e-05

------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
 6.964e-01   8.691e-02   8.013e+00   2.380e-12

------------------------------------------------------------------
No. of observations in the fit:  100
Degrees of Freedom for the fit:  2
      Residual Deg. of Freedom:  98
                      at cycle:  2

Global Deviance:     220.3789
            AIC:     224.3789
            SBC:     229.5892
******************************************************************
```

Nesse sumário encontramos qual família foi utilizada para o ajuste, qual algoritmo foi escolhido. Vemos também as funções de ligação de cada parâmetros assim como as estimativas e a significância deles. Observemos que os valores das estimativas apresentados no sumário precisam ser transformados pelas suas respectivas funções de ligação, o que pode ser feito com os comandos

```
> m.pond$mu.fv[1]
[1] 0.7613868
> m.pond$sigma.fv[1]
[1] 2.006536
```

Por último temos o número de observações, os graus de liberdade, o número de iterações e os valores do desvio global, AIC e BIC.

Juntamente com essas medidas de ajuste o programa do GAMLSS também oferece uma análise de resíduos, a função *plot* mostra os gráficos de resíduos contra os preditos,

resíduos contra um índice ou uma variável especificada, a densidade estimada dos resíduos e um QQ-Plot.

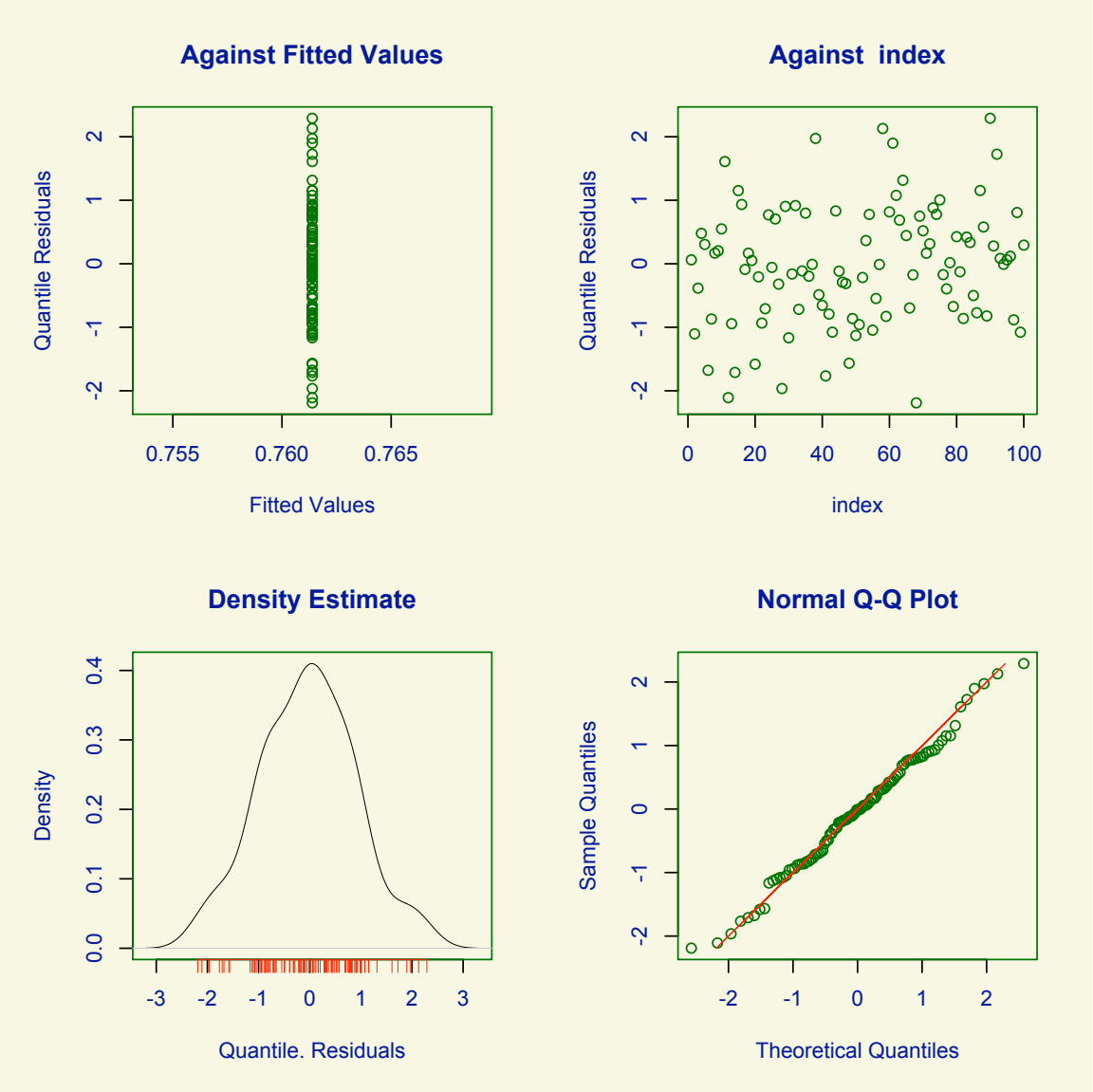

FIGURA A.1: Gráficos dos resíduos, densidade estimada e QQ-Plot.

A função *plot* também fornece informações sobre a média, variância, coeficiente de assimetria e curtose dos resíduos.

```
**********************************************************************
      Summary of the Quantile Residuals
                           mean   =  -0.03951061
                       variance   =  0.8833122
                 coef. of skewness  =  0.04980854
                 coef. of kurtosis  =  2.820389
```

```
Filliben correlation coefficient  =  0.9961603
******************************************************************
```

Mais uma ferramenta de diagnóstico oferecida pelo pacote GAMLSS é o gráfico *wormplot*, que é uma versão sem tendência do QQ-Plot. No wormplot, pontos na região entre as curvas indicam que o modelo está bem ajustado aos dados, enquanto que a presença de pontos dentro dessas regiões dão evidências de que o ajuste é ruim.

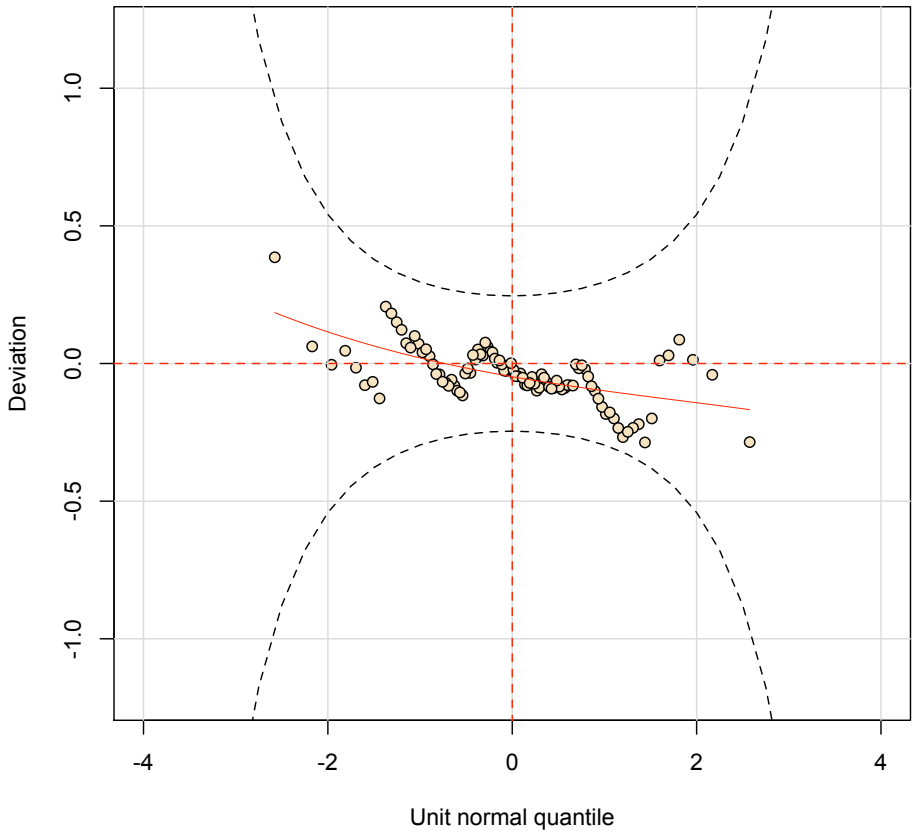

FIGURA A.2: Gráfico *wormplot*.

Pelo gráfico acima, temos todos os pontos entre as duas curvas, isso sugere que o modelo está bem ajustado ao conjunto de dados. O pacote também dispõe de um comando que dá os gráficos do deviance perfilado e intervalos de confiança perfilados para cada um dos parâmetros.

O comando é o *prof.dev*, essa rotina calcula o deviance perfilado para vários valores do parâmetro para calcular o intervalo. Devemos informar o objeto gamlss que

estamos utilizando, qual parâmetro estamos interessados, os valores mínimo e máximo que deve ser calculado o deviance, o tamanho do incremento que será adicionado a cada passo e a confiança desejada.

```
prof.dev(m.pond,"mu",min=0.5,max=0.9,step=0.01)
prof.dev(m.pond,"sigma",min=1.5,max=2.5,step=0.01)
```

O intervalo de confiança perfilado obtido para o parâmetro p foi:

$$IC_p = (0,6665343; 0,843524)$$

e intervalo de confiança perfilado obtido para o parâmetro θ foi:

$$IC_\theta = (1,683624; 2,367436)$$

Os comandos mostrados nesse pequeno tutorial são apenas aqueles que foram utilizados no decorrer deste trabalho, o pacote GAMLSS oferece mais recursos e os detalhes podem ser encontrado no manual do pacote e num artigo publicado pelos autores do pacote.

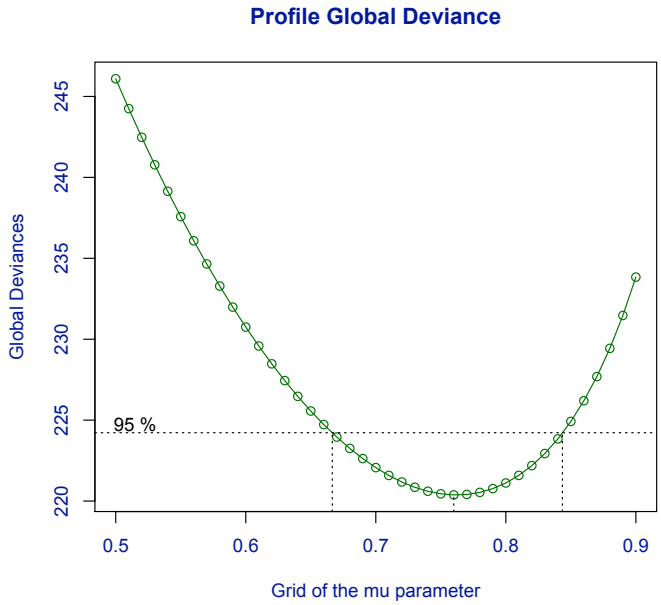

FIGURA A.3: Deviance e IC para p

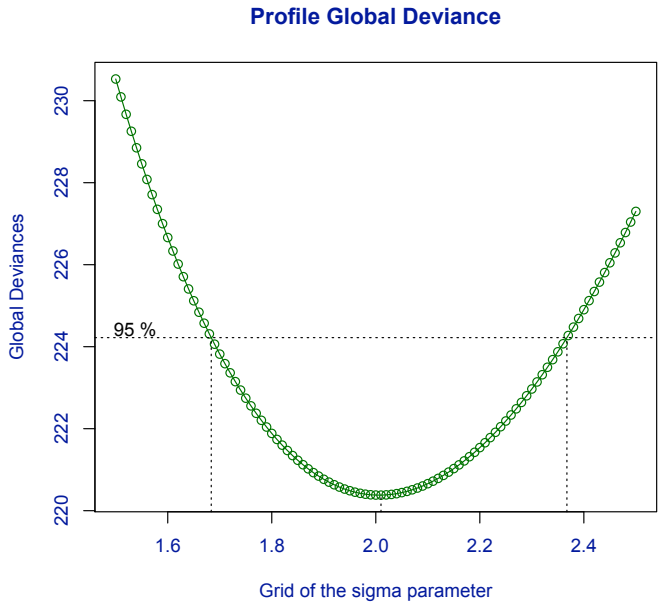

FIGURA A.4: Deviance e IC para θ