Modelo de Mistura Paramétrico com Fragilidade na Presença de Covariáveis

João Paulo Taconeli

UFSCar - São Carlos/SP

Junho/2013

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

Modelo de Mistura Paramétrico com Fragilidade na Presença de Covariáveis

João Paulo Taconeli

Orientador:

Profª. Drª. Vera Lucia D. Tomazella

Dissertação apresentada 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
Junho/2013

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

T119mm

Taconeli, João Paulo.
Modelo de mistura paramétrico com fragilidade na presença de covariáveis / João Paulo Taconeli. -- São Carlos : UFSCar, 2013.
66 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2013.

1. Análise de sobrevivência. 2. Distribuição Weibull. 3. Fragilidade. 4. Modelo de mistura. I. Título.

CDD: 519.9 (20ª)

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

FOLHA DE APROVAÇÃO

Aluno(a) : João Paulo Taconeli

DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA EM 23/04/2013
PELA COMISSÃO JULGADORA:

Presidente _____
Profa. Dra. Vera Lucia Damasceno Tomazella (DEs-UFSCar/Orientadora)

1º Examinador _____
Prof. Dr. Josemar Rodrigues (ICMC-USP)

2º Examinador _____
Profa. Dra. Juliana Cobre (ICMC-USP)

Agradecimentos

Primeiramente gostaria de agradecer à minha família, por ter me apoiado em um momento tão importante e de tantas mudanças: minha mãe Alzira, minha esposa Tânia e meus irmãos Fabinho e Guto, além de minha amiga Roseli.

Também quero manifestar minha gratidão à minha orientadora, Vera Lúcia D. Tomazella, e ao aluno de doutorado Jhon F. Gonzales, por todo conhecimento que me transmitiram, e aos professores que fizeram parte das minhas bancas: Juliana Cobre, Josemar Rodrigues e Francisco Louzada.

Não poderia deixar de agradecer também a todos os professores do Departamento de Estatística da UFSCar, especialmente aos que eu já conhecia desde a época da graduação, pela receptividade e paciência que tiveram comigo neste retorno, e também à funcionária Isabel, sempre muito atenciosa.

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

Resumo

Em análise de sobrevivência quando uma população apresenta, após um período representativo de tempo, uma quantidade expressiva de observações censuradas, podemos suspeitar que exista uma fração de indivíduos que não é susceptível ao evento de interesse. Diz-se então que esses indivíduos são "imunes", e que o conjunto de dados ao qual eles pertencem possui uma fração de cura. Os modelos de cura assumem implicitamente que todos os indivíduos que apresentaram o evento de interesse pertencem a uma população homogênea, mas no entanto podemos medir a heterogeneidade observada adicionando covariáveis ao modelo. Já a parcela da heterogeneidade que é induzida por fatores de risco não observáveis é estimada através de modelos de fragilidade. Com a finalidade de analisar dados de longa duração com heterogeneidade não observada na população, apresentamos o modelo de mistura padrão de Boag (1949) e Berkson & Gage (1952) sob um ponto de vista paramétrico, com covariáveis incidindo tanto na proporção de curados quanto na função de sobrevivência dos não curados. Peng & Zhang (2008a) realizaram uma estimação semiparamétrica deste modelo, e em nosso trabalho assumimos as distribuições de probabilidade Weibull para a parcela em risco e gama para a fragilidade. Também modelamos a proporção de curados através de modelos de regressão com diferentes funções de ligação, e testamos todos os modelos em uma base com dados reais envolvendo portadores de melanoma, realizando os ajustes tanto através da metodologia clássica quanto da bayesiana.

Abstract

Some studies involving survival data are characterized by showing a significant proportion of censored data, that is, individuals who will never experience the event of interest, even if accompanied by a long period of time. For the analysis of long-term data, we presented the standard mixture model by Berkson & Gage (1952), where we assume the Weibull distribution for the lifetime of individuals at risk and covariate. The cure rate models implicitly assume that those individuals experiencing the event of interest possess homogeneous risk. Alternatively, we consider the standard mixture model with a frailty term in order to quantify the unobservable heterogeneity among individuals. This model is characterized by the inclusion of a unobservable random variable, which represents information that can not or have not been observed. We assume frailty with a gamma distribution, obtaining the Weibull stardanrd mixture model with frailty and covariates from a point of view parametric. We realized simulation studies with the purpose of analyzing the frequentists properties of estimation procedures. Applications to real data set showed the applicability of the proposed models in which parameter estimates were determined using the maximum likelihood and bayesian approaches.

Sumário

Capítulo 1

Introdução

Experimentos em que a resposta representa o tempo até a ocorrência de um evento de interesse ocorrem com frequência em diversos ramos de conhecimento, especialmente em estudos envolvendo as áreas médica, financeira, de seguros e industrial.

Técnicas tradicionais de análise estatística, como análise de variância ou modelos de regressão, poderiam ser apropriadas para este tipo de estudo, mas nem sempre existe a garantia que todos elementos da amostra terão experimentado o evento de interesse no momento da coleta dos dados, o que torna a informação para tais indivíduos incompleta.

A classe de estudos estatísticos que lida com esta falta de informação (denominada censura) é conhecida como análise de sobrevivência (Colosimo & Giolo, 2006), e o tempo até a ocorrência do evento de interesse é comumente denominado "tempo de falha" (embora não necessariamente a medição esteja relacionada a uma falha). Os trabalhos historicamente mais importantes nessa área são o estimador não paramétrico de Kaplan-Meyer (Kaplan & Meier, 1958) e o modelo de riscos proporcionais de Cox (Cox, 1972).

Os principais objetivos quando se utiliza a análise de sobrevivência são a obtenção das estimativas da função de sobrevivência, que corresponde à probabilidade que um indivíduo sobreviva mais do que um determinado período de tempo,

e da função de risco, que mede a probabilidade de falha acontecer exatamente em um instante do tempo, dado que ela ainda não ocorreu até então.

Para atingir esses objetivos são possíveis abordagens não paramétricas, paramétricas ou semiparamétricas. Na análise de sobrevivência usual a função de sobrevivência converge para zero quando o tempo tende para infinito, ou seja, todos os indivíduos em estudo são considerados susceptíveis ao evento estudado.

A partir do momento em que avanços em áreas médicas se tornaram mais frequentes, um número maior de pacientes passou a ser considerado curado, ou imune à doença estudada. Diante disso, estimar a proporção de curados também passou a ser algo de bastante relevância. Os trabalhos apresentados por Boag (1949) e Berkson & Gage (1952), que falam sobre o modelo de mistura padrão, formaram a base do que veio a se chamar modelo de sobrevivência de longa duração (ou modelo de sobrevivência com fração de cura).

Este modelo é um dos mais utilizados dentro da literatura estatística, no entanto possui algumas restrições e desvantagens, como apontadas por Chen *et al.* (1999), que também utilizaram um modelo, que apresenta algumas vantagens em relação ao modelo de mistura padrão, denominado modelo de tempo de promoção.

Vários autores vêm discutindo a respeito de modelos envolvendo misturas de distribuições e fração de cura. Por exemplo, Farewell (1977) abordou o modelo de mistura Weibull e investigou como o fator de risco afeta o tempo de desenvolvimento de uma doença, sendo que posteriormente utilizou o modelo de riscos proporcionais de Cox (Farewell, 1982).

Goldman (1984) discutiu sobre a análise de sobrevivência quando a cura é possível. Greenhouse & Wolfe (1984) estudaram uma generalização do modelo de mistura padrão baseada na teoria de riscos competitivos. Farewell & Sprott (1986) examinaram o uso de tais modelos na inferência estatística, enquanto Kuk & Chen (1992) combinaram a formulação logística para a probabilidade de ocorrência do evento de interesse com estrutura de riscos proporcionais, propondo uma generalização semiparamétrica para o modelo de Farewell (Farewell, 1982).

Peng & Dear (2000) utilizaram a suposição de riscos proporcionais para

modelar o efeito das covariáveis sobre o tempo de falha dos pacientes não curados considerando métodos não paramétricos de estimação. Um excelente livro que aborda modelos de sobrevivência com fração de cura é o de Maller & Zhou (1996).

Quando se utiliza a abordagem paramétrica nos modelos de mistura, é necessário assumir uma distribuição de probabilidade para o tempo de falha dos pacientes não curados. As funções densidade e de sobrevivência serão derivadas desta distribuição, que por sua vez dependerá de um ou mais parâmetros (Peng & Dear, 2000). Discussões a respeito deste tipo de modelo foram feitas, dentre outros, por Berkson & Gage (1952), Farewell (1982), Farewell & Sprott (1986), Mailer & Zhou (1992) e Denham *et al.* (1996). Um problema com os modelos paramétricos se refere à verificação de suas suposições e restrições. Modelos não paramétricos são utizados quando se objetiva relaxar tais restrições.

Recentemente, modelos mais complexos de longa duração como o de Yakovlev *et al.* (1993), Chen et al. (1999), Ibrahim et al. (2001), Rodrigues et al. (2009), entre outros, vêm sendo explorados com o objetivo de explicar melhor os mecanismos biológicos envolvidos. Tsodikov (1998) definiu, em um estudo voltado à área biológica, a cura de um tumor como a probabilidade da não existência de células cancerígenas clonáveis ao final do tratamento. A proporção de pacientes sem este tipo de célula corresponderia então à fração de curados.

Uma definição alternativa do modelo de mistura padrão, que lida com células com potencial para desenvolver o tumor, foi proposta por Yakovlev *et al.* (1993) e Chen *et al.* (1999), entre outros, e denominada modelo de tempo de promoção. Yin & Ibrahim (2005) estabeleceram uma classe geral que contém como casos especiais os dois modelos de fração de cura, construída através de uma transformação de Box & Cox (Box & Cox, 1964) na função de sobrevivência populacional.

Os modelos de cura assumem implicitamente que todos os indivíduos que sofreram o evento de interesse pertencem a uma população homogênea. No entanto isso nem sempre ocorre, e uma forma de medir a heterogeneidade observada é adicionar covariáveis ao modelo. Assim, o risco da parcela que atingiu o evento de interesse deixará de ser considerado homogêneo.

Hougaard (1991) mostrou que é vantajoso considerar duas fontes de variabilidade: uma relativa aos fatores de risco observáveis (as covariáveis) e outra aos não observáveis, dado que nem sempre temos todas as informações que podem ser relevantes para um tempo de falha. Esse fato deu origem à inclusão de um novo termo nos modelos de sobrevivência, que recebeu o nome de fragilidade (Vaupel *et al.*, 1979).

Em essência, o conceito de fragilidade teve início no trabalho de Greenwood & Yule (1920) sobre tendência de acidentes e, mais tarde, Vaupel *et al.* (1979) introduziram o termo de fragilidade em modelos de sobrevivência univariados. Posteriormente, Clayton (1978) e Oakes (1982) abordaram os primeiros modelos de fragilidade para dados multivariados.

Aalen (1989), Hougaard *et al.* (1994), Longini & Halloran (1996) e Price & Manatunga (2001), dentre outros, estenderam os modelos de fragilidade considerando a fração de cura, sendo que Price & Manatunga (2001) desenvolveram um modelo considerando tanto a fração de cura quanto a fragilidade, e também assumiram uma distribuição paramétrica para a função de risco base, embora não tenham considerado o efeito de covariáveis em nenhum dos componentes do modelo. Peng & Zhang (2008a) incorporaram covariáveis e propuseram um método de estimação semiparamétrico baseado nos algoritmos EM e de múltipla imputação.

Já considerando também a abordagem bayesiana, Yin (2005) analisou dados odontológicos, sendo um indivíduo rotulado como curado quando não apresentava problema em nenhum de seus dentes, e o componente de fragilidade nesse caso foi útil para considerar a correlação causada por observações vindas do mesmo indivíduo, e utilizou o amostrador de Gibbs para obter as estimativas dos parâmetros.

Em estudos mais recentes Leng & Khalid (2010) compararam os resultados obtidos via máxima verossimilhança com os da metologia bayesiana para modelos de sobrevivência com fração de cura e fragilidade, sem a presença de covariáveis, enquanto Yu & Tiwari (2012) propuseram um enfoque bayesiano para um modelo similar, envolvendo pacientes de câncer, mas empregando uma

mistura finita de distribuições ao invés de modelar a distribuição latente através de uma distribuição paramétrica fixada, e estimaram os parâmetros através do método MCMC.

1.1 Objetivo do trabalho

A proposta deste trabalho é considerar o modelo de mistura padrão com fragilidade para analisar dados de sobrevivência censurados com fração de cura e informações observáveis entre os indivíduos, a fim de discriminar quais fatores são relevantes em cada componente do modelo, e adicionalmente quantificar a heterogeneidade não observável entre os indivíduos.

Neste trabalho o modelo de mistura com fragilidade e fração de cura é considerado incorporando covariáveis tanto na proporção de curados quanto na função de sobrevivência dos não curados, e utilizado um modelo paramétrico. Aqui o tempo de vida dos indivíduos em risco segue uma distribuição Weibull e a variável de fragilidade uma distribuição gama. Além disso, a proporção de pacientes não curados é modelada através de modelos de regressão com funções de ligação logito, probito e complementar log-log, a fim de verificar se existe alguma vantagem em se considerar alguma delas, ou se essa escolha não interfere nos resultados.

Através de um estudo de simulação checamos as propriedades assintóticas dos estimadores, e com aplicações em dados reais verificamos as estimativas dos parâmetros dos modelos propostos, inicialmente através da abordagem clássica, e na sequência utilizando a bayesiana.

1.2 Organização do trabalho

No Capítulo 2 são apresentados alguns dos principais conceitos de análise de sobrevivência, focando os que nos serão particularmente úteis. No Capítulo 3 será descrito o modelo de mistura padrão com fragilidade na presença de covariáveis. Para ilustrar a aplicabilidade do modelo, no Capítulo 4 foi utilizado

um conjunto de dados reais em que a estimação dos parâmetros foi determinada através do método de máxima verossimilhança, e também realizado um estudo de simulação com dados gerados artificialmente. No Capítulo 5 estimamos os mesmos parâmetros pela metodologia bayesiana, enquanto que no Capítulo 6 são apresentadas as conclusões e as propostas futuras.

Capítulo 2

Análise de Sobrevivência

2.1 Principais conceitos

A grande particularidade em estudos que envolvem dados de sobrevivência é o fato de que no momento da coleta uma parte da amostra ainda não terá experimentado o evento de interesse. Estas observações podem ser classificadas tanto como censuradas quanto como curadas (imunes). Desconsiderar tais observações não é apropriado, pois perderíamos informação. Por outro lado, utilizar as técnicas estatísticas tradicionais geraria resultados imprecisos, dada a diferença conceitual entre o tempo observado entre os que falharam e os censurados.

Basicamente os dados em um estudo de sobrevivência são constituídos por um conjunto de n elementos, em que observam-se seus respectivos tempos de falha (no caso do evento de interesse ter ocorrido), ou de censura (caso o evento de interesse não tenha ocorrido). Existem vários tipos de censuras, e dentre eles podemos citar (Colosimo & Giolo, 2006):

Censura tipo I: ocorre em estudos que encerram-se após um período pré-estabelecido de tempo, ou seja, sabe-se de antemão o tempo máximo até a ocorrência do evento, ou a censura.

Censura tipo II: é observada em estudos em que o encerramento se dará após uma quantidade pré-determinada de elementos atingirem o evento de interesse.

Censura tipo aleatório: acontece quando o elemento da amostra é retirado no decorrer do estudo, sem que o evento de interesse tenha sido observado, por qualquer outra razão que não diz respeito aos objetivos do experimento.

Com relação ao posicionamento do tempo de censura em relação ao tempo de ocorrência do evento, podem existir as seguintes possibilidades:

Censura à direita: é observada quando o tempo de ocorrência do evento acontece após o tempo observado, ou seja, está à direita do mesmo. É o tipo de censura que é mais comumente utilizado em análise de sobrevivência, e que será considerado neste trabalho.

Censura à esquerda: neste caso o tempo de ocorrência registrado é maior do que o tempo de falha, ou seja, o evento já teria ocorrido quando o elemento da amostra foi observado.

A censura pode ser classificada como informativa ou não-informativa. Ela será considerada informativa se sua ocorrência depender do mecanismo que gera a falha, e não-informativa caso contrário (Colosimo & Giolo, 2006).

O tempo de ocorrência pode ser pontual, que é quando sabe-se exatamente o momento em que a falha ocorreu, ou intervalar, onde os registros são feitos de tempos em tempos, e consequentemente não podemos precisar o momento da ocorrência (Wienke, 2011).

Em estudos envolvendo dados de sobrevivência as seguintes variáveis aleatórias são consideradas as mais importantes:

$T =$ tempo até a ocorrência;

$C =$ tempo até a censura;

$Y = \min(T, C)$;

$$\delta = \begin{cases} 1, \text{se} \quad T \leq C \\ 0, se \quad T > C \end{cases}$$

Outras variáveis são incluídas à medida em que sofisticamos os modelos, como será visto mais adiante.

Além das variáveis aleatórias, existem funções de extrema importância dentro dos estudos de sobrevivência:

Função de sobrevivência - é definida como a probabilidade de uma observação não falhar até o tempo t, ou seja:

$$S(t) = P(T \geq t) \tag{2.1}$$

que também pode ser expressa em termos da função de distribuição acumulada de probabilidade:

$$S(t) = 1 - F(t), \tag{2.2}$$

e possui as seguintes propriedades: $S(t) = 1$, para $t \leq 0$ e $lim_{t\to\infty} S(t) = 0$.

O gráfico abaixo mostra um exemplo de comportamento da função de sobrevivência ao longo do tempo. Neste caso específico a função de sobrevivência chega a 0 com o passar do tempo, o que significa que toda a amostra falhou.

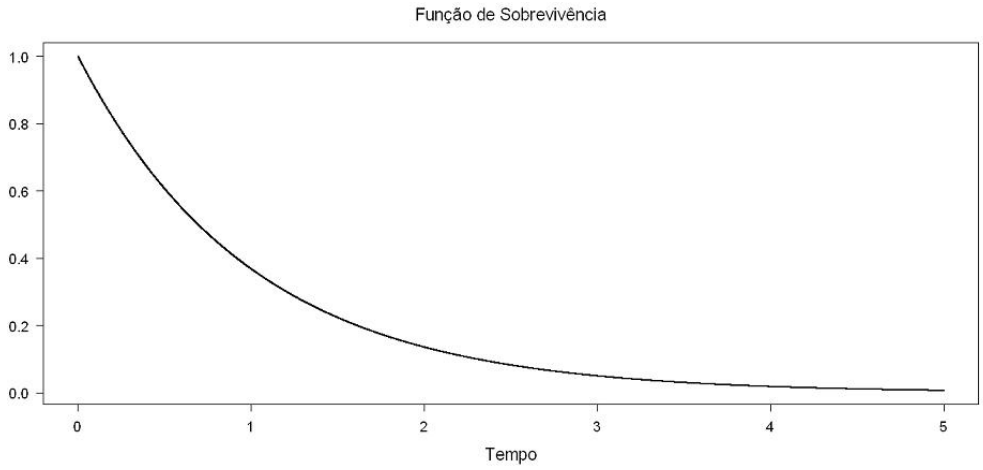

Figura 2.1: Exemplo de curva de sobrevivência.

Função de risco - também chamada de função de taxa de falha, é definida como a probabilidade condicional de que o evento ocorra dentro de um intervalo $(t, t + \Delta(t))$, dado que não ocorreu até o início do mesmo:

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

A função de risco pode ser crescente ou decrescente. Isto ocorrerá quando a taxa de falha crescer (ou decair) conforme o tempo passar. Também pode ser constante ao longo do tempo, ou até mesmo combinações destes três tipos citados.

Função de risco acumulada - também chamada de função de taxa de falha acumulada, é definida como:

$$H(t) = \int_0^t h(u)du$$

Algumas relações bastante utilizadas entre as funções descritas são:

$$f(t) = \frac{-dS(t)}{dt} \tag{2.3}$$

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

$$H(t) = \int_0^t h(u)du = -\log S(t)$$

$$S(t) = \exp(-H(t)) = \exp\left(\int_0^t h(u)du\right)$$

Os métodos para a estimação dos parâmetros que envolvem estas funções podem ser não paramétricos, paramétricos ou semiparamétricos. Dentre os não paramétricos, o mais utilizado em estudos clínicos é o estimador de Kaplan-Meier (Kaplan & Meier, 1958), também conhecido na literatura como estimador produto-limite.

O estimador de Kaplan-Meier é uma adaptação da função de sobrevivência empírica, mas com a vantagem de trabalhar com dados censurados. Basicamente ele considera uma amostra de n elementos em que $t_1 < t_2 < \cdots t_k$, em que os k representam os tempos de falha ordenados, $d_1, d_2, \cdots, d_k$ os respectivos números de falhas em t_j e n_j o número de indivíduos sob risco em t_j. O estimador de Kaplan-Meier para a função de sobrevivência será dado por:

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

sendo que cada termo do produto acima é uma estimativa da probabilidade condicional de não falha no instante t_j dado que não existiu falha até o instante t_{j-1}. Uma desvantagem deste estimador é sua incapacidade de incorporar covariáveis. No caso, uma alternativa é estratificar a base de dados e ajustar uma curva para cada nível da covariável, o que pode-se tornar complicado em caso de muitos níveis ou de combinações de níveis de diversas covariáveis.

A Figura 2.2 ilustra a curva de sobrevivência estimada através do método de Kaplan-Meier para dados referentes a um estudo clínico em que os pacientes são observados após a remoção de um melanoma maligno. O conjunto de dados possui 205 indivíduos, sendo que 148 foram censurados no período observado. O tempo máximo observado foi de 5565 dias e a morte do paciente é o evento de interesse. Estes dados foram retirados do pacote timereg do sistema R (R Development Core Team, 2009).

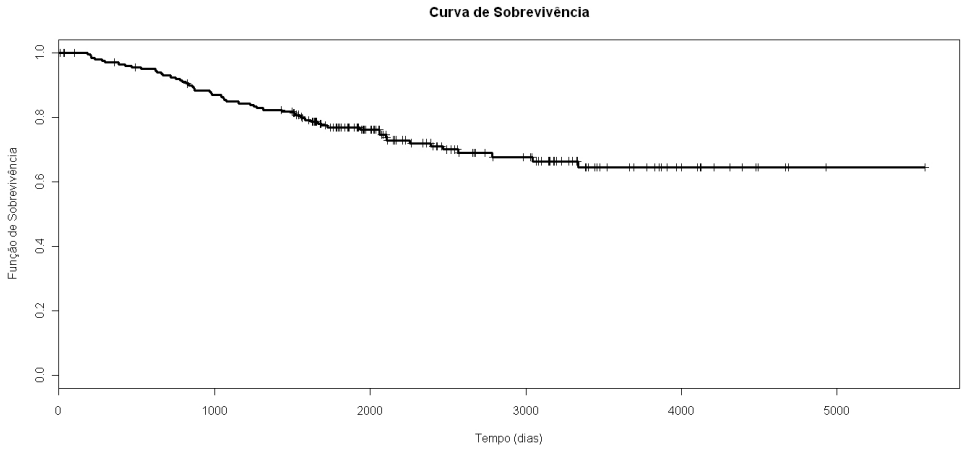

Figura 2.2: Curva de sobrevivência de Kaplan-Mayer

Observamos na Figura 2.2 que após um determinado tempo a curva se estabiliza, não havendo mais falhas. Isto sugere que ao menos parte dos indivíduos censurados no final do experimento podem ser imunes ao risco em questão ou foram curados durante o experimento. Entretanto, o estimador de Kaplan-Meier, por ser sensível a quantidades altas de censura, não é apropriado para estimar esta proporção de cura.

2.2 Modelos de probabilidade

Embora qualquer distribuição de variáveis aleatórias não negativas possa ser utilizada para representar tempos de sobrevivência, existem algumas que têm sido tradicionalmente mais empregadas e auxiliado na resolução de problemas práticos. Podemos citar como exemplo as distribuições exponencial, log-normal, Weibull e gama (Colosimo & Giolo, 2006), e outras propostas mais recentemente,

como a Weibull modificada e a Weibull modificada generalizada (Calsavara, 2011). Nesta seção apresentamos características de duas destas distribuições, que utilizaremos na formulação dos modelos que iremos testar.

2.2.1 Distribuição Weibull

A distribuição Weibull foi descrita originalmente por Waloddi Weibull em 1951 e, desde então, tem sido amplamente usada como modelo em aplicações biométricas, industriais, laboratoriais e de confiabilidade.

Esta distribuição apresenta dois parâmetros e sua função de risco é monótona (ou seja, ela acomoda riscos crescentes, decrescentes ou constantes). Sua função densidade de probabilidade é dada por:

$$f(t;\alpha,\lambda) = \alpha\lambda\,(t\lambda)^{\alpha-1}\exp\left[-\,(t\lambda)^{\alpha}\right], \quad \alpha>0,\ \lambda>0,\ t>0, \tag{2.4}$$

onde α representa o parâmetro de forma e λ o parâmetro de escala.

As funções de sobrevivência, de risco acumulado e de risco serão, respectivamente:

$$S(t) = \exp\left[-\,(t\lambda)^{\alpha}\right], \tag{2.5}$$

$$H(t) = (t\lambda)^{\alpha}, \tag{2.6}$$

$$h(t) = \alpha\lambda\,(t\lambda)^{\alpha-1}. \tag{2.7}$$

Além disso, quando $\alpha = 1$ a distribuição Weibull se reduz à distribuição exponencial.

Na figura 2.3 são mostrados os gráficos da função densidade, função de sobrevivência e função de risco da distribuição Weibull, fixando o parâmetro λ em 1 e atribuindo valores para o parâmetro α iguais a 2, 1, 0,75 e 0,5:

2.2.2 Distribuição gama

Essa distribuição tem sido uma das mais utilizadas para descrever tempos de falha, e especificamente em nosso trabalho a utilizaremos para modelar

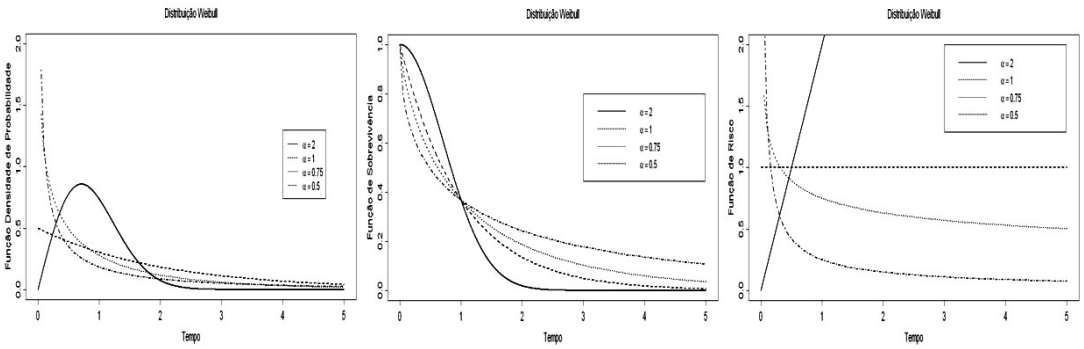

Figura 2.3: Funções de densidade, sobrevivência e de risco da distribuição Weibull.

componente de fragilidade. Esta distribuição possui diversas outras distribuições como casos especiais, e sua função densidade de probabilidade é dada por:

$$f(t|a,b) = \frac{b^a}{\Gamma(a)} t^{a-1} \exp(-tb), \quad a > 0,\ b > 0,\ t > 0,$$

onde a representa o parâmetro de forma e b o parâmetro de escala desta distribuição. Sua função de sobrevivência é calculada pela integral:

$$S(t) = \int_t^\infty \frac{b^a}{\Gamma(a)} u^{a-1} \exp(-ub)\ du.$$

Na Figura 2.4 mostramos os gráficos da função densidade, função de sobrevivência e função de risco da distribuição gama, fixando o parâmetro de forma em 1 e valores para o parâmetro de escala iguais a 2, 1, 0,5 e 0,25.

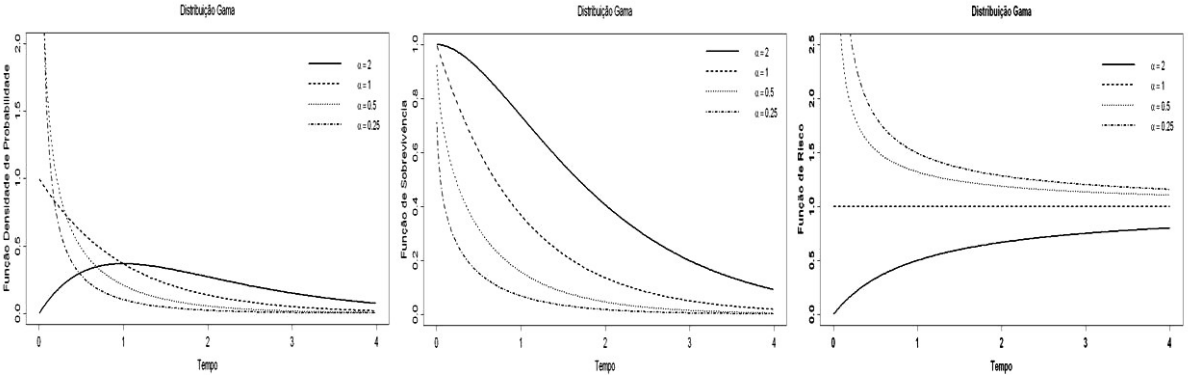

Figura 2.4: Funções de densidade, sobrevivência e de risco da distribuição gama.

Além das já citadas, outras distribuções citadas na literatura e que podem ser utilizadas em estudos de sobrevivência são a log-logística, a log-gama, a Rayleigh, a normal inversa e a distribuição de Gompertz (Wienke, 2011).

O uso de qualquer um destes modelos implica na estimação de seus respectivos parâmetros, e como o modelo é paramétrico, é utilizado o método da máxima verossimilhança (Colosimo & Giolo, 2006): em uma amostra de n elementos, em que r deles são não censurados (ou seja, possuem tempos de falha completos) e $n-r$ são censurados, e denominando $\boldsymbol{\xi}$ e $\mathbf{D}$ os vetores de parâmetros desconhecidos e dados observados, respectivamente, teremos a seguinte função de verossimilhança:

$$L(\boldsymbol{\xi}|\mathbf{D}) \propto \prod_{i=1}^{r} f(t_i|\boldsymbol{\xi}) \prod_{i=r+1}^{n} S(t_i|\boldsymbol{\xi}), \tag{2.8}$$

ou seja, a contribuição para a verossimilhança dos indivíduos que falharam será dada pela função densidade, enquanto que a contribuição dos censurados é dada pela função de sobrevivência. Dada a variável indicadora de falha relativa ao i-ésimo indivíduo, δ_i, a função de verossimilhança pode ser reescrita como:

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n} \left[f(t_i|\boldsymbol{\xi})\right]^{\delta_i} \left[S(t_i|\boldsymbol{\xi})\right]^{1-\delta_i} = \prod_{i=1}^{n} \left[h(t_i|\boldsymbol{\xi})\right]^{\delta_i} S(t_i|\boldsymbol{\xi}) \tag{2.9}$$

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\xi}$ que maximizam $L(\boldsymbol{\xi}|\mathbf{D})$, ou equivalentemente o $\log L(\boldsymbol{\xi}|\mathbf{D})$, e são encontrados resolvendo o seguinte sistema de equações:

$$U(\boldsymbol{\xi}) = \frac{\partial \log L(\boldsymbol{\xi}|\mathbf{D})}{\partial \boldsymbol{\xi}} = 0,$$

cuja solução é obtida na maioria das vezes através de algoritmos numéricos, como o método de Newton Rhapson. A obtenção de intervalos de confiança para estes parâmetros é feita através das propriedades assintóticas de tais estimadores.

2.3 Modelo de riscos proporcionais de Cox

Considerando-se um conjunto de covariáveis $\mathbf{x} = (x_1, x_2, ..., x_p)'$, a expressão geral do modelo de Cox (Cox, 1972) é dada por:

$$h(t|\mathbf{x}) = h_0(t) g(\mathbf{x}'\beta), \tag{2.10}$$

sendo g uma função não negativa com $g(0) = 1$, $h_0(t)$ a função de risco base que representa o componente não paramétrico, e $g(\mathbf{x}^{'}\beta)$ o componente paramétrico do modelo. Daí a denominação de modelo semiparamétrico.

Em linhas gerais ele trata a função de risco como uma função fatorável em duas, sendo que a primeira contempla a parcela dinâmica do risco e a segunda é associada ao efeito que as covariáveis exercem na função de risco. Normalmente utiliza-se uma função exponencial, a fim de garantir que seu resultado seja positivo.

O modelo de regressão de Cox é chamado de modelo de riscos proporcionais pois a razão entre o risco de dois indivíduos é constante no tempo, dado que a função de risco base é idêntica para todos os elementos da amostra. Assim, a diferença entre eles será explicada apenas pelas covariáveis:

$$\frac{h_i(t)}{h_j(t)} = g(\mathbf{x}_i^{'}\boldsymbol{\beta} - \mathbf{x}_j^{'}\boldsymbol{\beta}).$$

Mesmo com todas essas características que envolvem sua flexibilidade, o modelo de Cox não se ajusta necessariamente a qualquer problema envolvendo dados de sobrevivência, o que torna necessário avaliar sua adequação. Algumas extensões deste modelo envolvem a inclusão de covariáveis dependentes no tempo ou modelos que não exijam o cumprimento da suposição de proporcionalidade dos riscos.

2.4 Modelo com componente de fragilidade

Embora possam apresentar valores idênticos para todas as covariáveis, alguns indivíduos eventualmente exibem respostas diferentes no que diz respeito ao efeito de uma droga, ou de um tratamento, por exemplo. Neste contexto, existe uma variabilidade não observável em análise de sobrevivência (Hougaard, 1991).

Mesmo quando a análise se restringe a apenas um indivíduo sujeito a alguns fatores de risco (conceito bastante utilizado em análises clínicas), os tempos $t_1, ..., t_k$ relativos aos k fatores de riscos competitivos até então foram

considerados independentes - o que não parece ser muito provável dado que tais tempos pertencem à mesma unidade amostral. Desta maneira, é necessário também considerar esta dependência, o que não será o foco deste trabalho.

Assim, uma possível solução é incorporar um termo que represente esta heterogeneidade, w_i, para cada indivíduo, e considerar que w_i, $t_1, ..., t_k$ sejam independentes e identicamente distribuídos, com função de distribuição $F(.|w)$.

Em análise de sobrevivência este termo recebe o nome de fragilidade, pois teoricamente quanto maior seu valor, mais "frágil" seria o indivíduo, e por consequência, maior sua probabilidade de falha (Vaupel *et al.*, 1979).

No nosso estudo, particularmente, o termo de fragilidade atuará na função de sobrevivência dos indivíduos não imunes, como um complemento às covariáveis observadas, ajudando desta forma a estimar esta função de maneira mais precisa. Não estimaremos, desta forma, a fragilidade a nível individual.

Clayton (1978) introduziu o termo de fragilidade ao modelo de Cox (1972) de forma multiplicativa, ou seja, a variável aleatória que representa a fragilidade, W, irá agir multiplicativamente na função de risco base, da seguinte maneira:

$$h(t|W = w) = w\ h_0(t), \tag{2.11}$$

onde W é considerada uma variável aleatória latente e não negativa e $h_0(t)$ a função de risco base.

Neste contexto, de acordo com Elbers & Ridder (1982), é necessário que a distribuição do efeito aleatório tenha média 1 para o modelo ser identificável. Sua variância é interpretada como uma medida de heterogeneidade da população. Valores pequenos de $Var(W)$ deixam os valores de W todos muito próximos de 1, enquanto altos valores de $Var(W)$ apontam que os valores de W estão muito dispersos, ou seja, há uma forte heterogeneidade não observável entre os indivíduos (Calsavara, 2011).

A variável de fragilidade também pode agir de forma aditiva na função de risco, conforme proposto por diversos autores, dentre eles Rocha (1995). Neste caso o modelo tomaria a seguinte forma:

$$h(t|W = w) = w + h_0(t).$$

A obtenção das funções de risco e sobrevivência marginais para o modelo aditivo de fragilidade não serão aqui demonstradas, pois será utilizado o modelo multiplicativo (2.11).

Adicionando-se covariáveis ao modelo multiplicativo de fragilidade, este assumirá a seguinte forma:

$$h(t|\mathbf{X}{=}\mathbf{x}, W = w) = wh_0(t)\exp(\mathbf{x}^{'}\boldsymbol{\beta}), \tag{2.12}$$

em que $\mathbf{x} = (x_1, x_2, ..., x_p)'$ corresponde ao vetor de covariáveis e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$ os parâmetros a elas correspondentes, o que mostra que este modelo é uma generalização dos modelos de riscos proporcionais de Cox (Colosimo & Giolo, 2006).

Aqui teremos um parâmetro para cada covariável estudada, ao contrário do que veremos quando adicionarmos covariáveis à parcela imune, onde haverá um β_0 que corresponderá ao intercepto da regressão ajustada à fração de cura.

Para a obtenção da $S(t)$ considerando o termo de fragilidade, denotamos $S(t|W = w)$ como a função de sobrevivência de um indivíduo condicional à fragilidade W, ou seja:

$$S(t|W = w) = \exp\left(-\int_0^t h(s|w)ds\right) = \exp\left(-w\int_0^t h_0(s)ds\right) = \exp\left(-wH_0(t)\right),$$

em que $H_0(t)$ representa a função de risco base acumulada no instante t.

Entretanto os dados para este modelo a nível individual não são observáveis. Consequentemente é necessário ir para o nível populacional, em que o termo de fragilidade pode ser integrado. Assim utilizando a transformada de Laplace podemos obter $S(t)$, da seguinte maneira:

$$S(t) = E[S(t|W = w)] = \int_0^\infty \exp(-wH_0(t))g(w)dw = \boldsymbol{L}_W(H_0(t)), \tag{2.13}$$

em que $g(w)$ é a função densidade da variável de fragilidade e $\boldsymbol{L}_w(H_0(t))$ é a tranformada de Laplace aplicada no ponto $H_0(t)$, o que mostra a importância desta transformação nos modelos de fragilidade (Wienke, 2011).

As derivadas da transformação de Laplace podem também ser utilizadas para obtermos outros resultados relativos à função de sobrevivência não condici-

onal:

$$f(t) = -h_0(t)\boldsymbol{L}_W{}'(H_0(t)) \quad , \quad h(t) = -h_0(t)\frac{\boldsymbol{L}_W{}'(H_0(t))}{\boldsymbol{L}_W(H_0(t))}$$
$$E(W) = -\boldsymbol{L}_W{}'(0) \quad e \quad Var(W) = \boldsymbol{L}_W{}''(0) - (\boldsymbol{L}_W{}'(0))^2,$$

onde $\boldsymbol{L}_W(.)'$ e $\boldsymbol{L}_W(.)''$ denotam as derivadas de primeira e segunda ordem da transformação de Laplace, respectivamente.

2.4.1 Modelo de fragilidade gama

Iremos considerar que W, a variável aleatória que representa a fragilidade, segue uma distribuição gama, com a seguinte função densidade de probabilidade:

$$f(w|a,b) = \frac{b^a}{\Gamma(a)} w^{a-1} \exp\{-(wb)\}, \quad w > 0, a > 0, b > 0.$$

Então sua transformada de Laplace de primeira ordem será dada por:

$$L_W(\mu) = E[e^{-\mu w}] = \left(1 + \frac{\mu}{b}\right)^{-a}. \tag{2.14}$$

Assim, sem considerar covariáveis, sua função de sobrevivência será obtida, considerando a distribuição da fragilidade $W \sim G(a,b)$ e $a = b = \sigma$, a fim de garantir que a média seja igual a 1, como indicado, e conforme (2.14):

$$S(t) = L_W(H_0(t)) = \left(1 + \frac{H_0(t)}{\sigma}\right)^{-\sigma}. \tag{2.15}$$

Consequentemente, a função densidade será dada por:

$$f(t) = \left(1 + \frac{H_0(t)}{\sigma}\right)^{-\sigma-1} h_0(t). \tag{2.16}$$

E na presença de covariáveis teremos:

$$S(t|\mathbf{x}) = L_W(H_0(t)) = \left(1 + \frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma},$$

e

$$f(t|\mathbf{x}) = \left(1 + \frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1} h_0(t)\exp(\mathbf{x}'\boldsymbol{\beta}).$$

A distribuição gama tem sido a distribuição mais comumente utilizada para representar o componente de fragilidade, mas outras distribuições também

empregadas neste sentido são a log-normal, a normal inversa e a positiva estável, por exemplo (Hougaard, 1991).

Já para a função de risco base podem ser atribuídas distribuições tradicionalmente utilizadas para representar tempos de vida, como a exponencial, log-normal, Weibull e gama, dentre outras (Wienke, 2011).

2.5 Modelo de mistura padrão

Em diversos experimentos notamos que a curva de sobrevivência estabiliza-se em um patamar estritamente maior que 0. Este fato ocorre com frequência cada vez maior na área clínica, em que novos tratamentos têm conseguido prolongar ou mesmo curar diversos tipos de tumores. Essa proporção de indivíduos que não experimentam o evento de interesse recebe o nome de curados, ou sobreviventes de longa duração. Este conceito também pode ser utilizado em outras áreas, bastando para isso que uma parcela dos elementos não venha a experimentar o evento de interesse em longos períodos de observação.

Verificar se essa proporção de curados (ou imunes, não suscetíveis - dentre outras possíveis denominações) realmente existe, e estimá-la é algo de grande relevância atualmente em análise de sobrevivência. Outro desafio é diferenciar os curados dos meramente censurados à direita.

O modelo de mistura padrão proposto por Boag (1949) e Berkson & Gage (1952) para estudos em que uma proporção substancial dos indivíduos não experimentará o evento de interesse pode ser utilizado para modelar a fração de cura, pois envolve duas subpopulações (no caso os indivíduos susceptíveis e os não susceptíveis ao evento de interesse). Sempre que o estimador da função de sobrevivência convergir para um valor maior que 0 após um extenso período de tempo, existe a possibilidade de haver uma proporção da população que seja imune, e assim jamais experimentará o evento de interesse.

O modelo de mistura padrão considera que número de riscos desconheci-

dos, N, é proveniente de uma distribuição Bernoulli com parâmetro θ, ou seja:

$$P(N|\theta) = \theta^N(1-\theta)^{1-N}, \quad 0 \leq \theta \leq 1 \tag{2.17}$$

e sua função de sobrevivência populacional, será:

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

em que $1-\theta$ representa a incidência de curados, ou imunes, e $S(t)$ a função de sobrevivência dos indivíduos em risco. É interessante notar que $S(t)$ é uma função própria mas $S_{pop}(t)$ é imprópria, pois:

$$\lim_{t\to\infty} S_{pop}(t) = 1 - \theta$$

A exemplo do que ocorre com a função $S(t)$, $S_{pop}(0) = 1$, e também é decrescente. Além disso, quando $\theta = 1$, a função de sobrevivência imprópria $S_{pop}(t)$ se reduz à função de sobrevivência própria $S(t)$, pois nesse caso não existiria uma parcela de curados.

Neste contexto, a função de densidade populacional será, de acordo com (2.3):

$$f_{pop}(t) = \theta f(t),$$

com a seguinte função de risco populacional:

$$h_{pop}(t) = \frac{f_{pop}(t)}{S_{pop}(t)} = \frac{\theta f(t)}{1 - \theta + \theta S(t)} \tag{2.19}$$

É possível incorporar covariáveis tanto na fração de cura quanto na função de sobrevivência própria deste modelo, conforme será visto no Capítulo 3.

O modelo de mistura padrão foi um dos primeiros modelos paramétricos em análise de sobrevivência de longa duração. Mas, como qualquer modelo estatístico, ele possui 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 ocorre no modelo de tempo de promoção de Yakovlev & Tsodikov (1996), em que se modela o número de causas que competem entre si e que podem causar o evento.

Uma outra característica do modelo de mistura padrão é que ele não é um modelo de riscos proporcionais. Para isto, basta observar a função de risco do modelo (2.19). Se considerarmos $f(t)$ como a função de risco base, comum a todos os indivíduos, o restante da expressão ainda dependerá de t através da função de sobrevivência $S(t)$.

Uma importante questão referente ao modelo de mistura padrão é com relação à sua identificabilidade. Laska & Meisner (1992) mostraram que se a última observação é censurada, o resultado da estimação por máxima verossimilhança não é único. Entretanto Li *et al.* (2001) mostraram que o modelo de mistura padrão é identificável quando a função de sobrevivência própria é especificada parametricamente e a fração de curados θ é um parâmetro (e que também pode ser função de covariáveis), o que é nosso caso.

2.6 Considerações finais

Neste Capítulo apresentamos uma introdução à análise de sobrevivência, destacando especialmente definições e conceitos que nos serão úteis em nosso estudo. Com o objetivo de analisar dados de sobrevivência com fração de cura e fragilidade, no Capítulo 3 apresentaremos um modelo baseado no modelo de Berkson & Gage (1952) para estimar a probabilidade de cura.

Este modelo de mistura padrão com fragilidade e na presença de covariáveis também considerará (e estimará) a heterogeneidade não obseravada através do compontente de fragilidade, a qual será atribuída uma distribuição gama. Para os indivíduos em risco consideraremos que os tempos de vida seguem distribuição Weibull.

Capítulo 3

Modelo de Mistura Padrão com Fragilidade e Covariáveis

3.1 Definições

O modelo de mistura padrão com fragilidade e covariáveis tem como objetivo analisar dados de tempo de vida considerando censura (como é feito em modelos usuais de sobrevivência), proporção de curados (como inserido através dos modelos de mistura padrão) e também a heterogeneidade não captada pelas covariáveis, que neste contexto recebe o nome de fragilidade.

Em seu trabalho, Peng & Zhang (2008a) incluíram covariáveis em ambos os componentes do modelo e assumiram uma distribuição gama para a fragilidade. Entretanto sua abordagem foi semiparamétrica, pois eles não atribuiram distribuição de probabilidade para a função de risco base. Aqui será proposta uma abordagem totalmente paramétrica, assumindo uma distribuição de probabilidade adequada para a função de risco base. Adicionalmente, Peng & Zhang (2008a) estimaram os parâmetros do modelo de forma clássica, utilizando o algoritmo EM e o método de múltipla imputação.

Calsavara (2011) estimou tanto pelo enfoque clássico quanto pelo bayesiano os parâmetros do modelo de fragilidade com fração de cura, considerando a distribuição Weibull para a função de risco base. No entanto, não incluiu

covariáveis em nenhum dos componentes do modelo.

Assim o modelo introduzido por Berkson & Gage (1952) pode ser extendido, a fim de incluir efeito de covariáveis. Seja $\mathbf{z} = (z_1, z_2, ..., z_q)'$ representando o conjunto de covariáveis que afetará a probabilidade de cura e $\mathbf{x} = (x_1, x_2, ..., z_p)'$ o conjunto de covariáveis que incidirá na função de sobrevivência dos não curados. Assim, o modelo pode ser reescrito como:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = 1 - \theta(\mathbf{z}) + \theta(\mathbf{z})S(t|\mathbf{x}), \tag{3.1}$$

em que $1-\theta(\mathbf{z})$ representa a probabilidade de um paciente ser curado dependendo das covariáveis $\mathbf{z}$ e $S(t|\mathbf{x})$ é a função de sobrevivência da distribuição do tempo de falha de pacientes não curados, dependendo de $\mathbf{x}$. Estes dois conjuntos de covariáveis podem conter covariáveis em comum, ou mesmo serem idênticos, pois uma mesma covariável pode afetar os dois subconjuntos do modelo de mistura padrão.

Algumas abordagens paramétricas foram apresentadas por Farewell (1982) e Peng *et al.* (1998), em que foram assumidas distribuições de probabilidade para a variável latente. Alternativas semiparamétricas são também bastante usuais, com a vantagem de reduzir a dependência do modelo em relação às suposições dos estudos paramétricos. A mais popular delas é o modelo semiparamétrico de mistura com taxa de cura e riscos proporcionais (Kuk & Chen, 1992), cujas funções de sobrevivência $S_{pop}(.)$ e de densidade $f_{pop}(.)$ são:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = 1 - \theta(\mathbf{z}) + \theta(\mathbf{z})\ S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta})}, \tag{3.2}$$

$$f_{pop}(t|\mathbf{x}, \mathbf{z}) = \theta(\mathbf{z}) f(t) \exp(\mathbf{x}'\boldsymbol{\beta})\ S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta})-1}, \tag{3.3}$$

em que $S_0(t)$ é uma função base de sobrevivência, arbitrária, e $\boldsymbol{\beta}$ representa o vetor de parâmetros que serão estimados para as covariáveis associadas aos indivíduos em risco. Atribuindo uma distribuição de probabilidades para t, o modelo torna-se paramétrico.

Para modelar os efeitos das covariáveis na taxa de cura, podemos utilizar diferentes funções de ligações (Peng & Zhang, 2008a). Definindo $\boldsymbol{b} = (b_0, b_1, ..., b_q)'$ como sendo o vetor de parâmetros que serão estimados para as

covariáveis associadas à fração de cura e a função de ligação logito, temos o modelo de regressão logístico:

$$\theta(\mathbf{z}) = \frac{\exp(\boldsymbol{b}'\mathbf{z})}{\mathbf{1} + \exp(\boldsymbol{b}'\mathbf{z})}$$

Outra função de ligação bastante empregada é a probito:

$$\theta(\mathbf{z}) = \mathbf{\Phi}(\boldsymbol{b}'\mathbf{z}),$$

em que Φ corresponde à função de distribuição acumulada de uma distribuição normal padrão.

Adicionalmente, também é utilizada função de ligação complemento log-log, que é dada por:

$$\theta(\mathbf{z}) = \exp(-\exp(\boldsymbol{b}'\mathbf{z}))$$

O processo de estimação empregado é similar para as três funções de ligação. Para a estimação da probabilidade de cura, basta substituir os parâmetros pelas estimativas obtidas, de acordo com as covariáveis e a função de ligação utilizada. Desta forma $1 - \theta(\mathbf{z})$ será a probabilidade de cura.

Conforme visto em (2.15), a função de sobrevivência, considerando que o componente que representa a fragilidade siga uma distribuição de probabilidade $W \sim G(\sigma, \sigma)$ e ainda sem considerar covariáveis é dada por:

$$S(t) = L_W(H_0(t)) = \left(1 + \frac{H_0(t)}{\sigma}\right)^{-\sigma},$$

com $L_W(.)$ representando a transformada de Laplace e $H_0(t)$ a função de risco base acumulada.

Na presença de covariáveis, temos:

$$S(t|\mathbf{x}) = \left(1 + \frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}. \tag{3.4}$$

Consequentemente suas respectivas funções de densidade e de risco, também com covariáveis, serão dadas por:

$$f(t|\mathbf{x}) = \left(1 + \frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1} h_0(t)\exp(\mathbf{x}'\boldsymbol{\beta}), \tag{3.5}$$

e

$$h(t|\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \left(\frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)} h_0(t),$$

em que podemos notar que o efeito das covariáveis $\mathbf{x}$ não satisfaz a suposição de riscos proporcionais.

Se σ for muito grande (o que implica que a variância da variável que representa a fragilidade é muito próxima de zero) o denominador se aproximará de 1, e a função de risco se reduzirá a:

$$h(t|\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}) h_0(t),$$

que voltará a possuir proporcionalidade nos riscos.

Substituindo (3.4) na expressão (3.1) obtemos a função de sobrevivência populacional com fração de cura e fragilidade, na presença de covariáveis:

$$S_{pop}(t|\mathbf{x}, \mathbf{z}) = 1 - \theta(\mathbf{z}) + \theta(\mathbf{z}) \left(1 + \frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}, \tag{3.6}$$

com o vetor de covariáveis $\mathbf{z}$, que afeta a probabilidade de cura podendo ter elementos em comum com o vetor de covariáveis $\mathbf{x}$, que atua na função de sobrevivência dos não curados.

Utilizando a propriedade (2.3), e aplicando-a em (3.6) chegamos à função densidade populacional:

$$f_{pop}(t|\mathbf{x}, \mathbf{z}) = \theta(\mathbf{z}) \left(1 + \frac{H_0(t)\exp(\mathbf{x}'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1} h_0(t)\exp(\mathbf{x}'\boldsymbol{\beta}). \tag{3.7}$$

3.2 Inferência

Como a abordagem proposta é totalmente paramétrica, a estimação, dentro do enfoque clássico, será feita pelo método da máxima verossimilhança. Desta maneira se torna necessária a obtenção da função de verossimilhança para os parâmetros que precisarão ser estimados. Considerando os dados observados $\mathbf{D} = (t, \delta)$, sendo t o tempo de ocorrência e δ o indicador de censura de cada elemento da população e o conjunto de parâmetros a ser estimado $\boldsymbol{\xi} = (\sigma, \mathbf{b}, \boldsymbol{\beta})$,

temos que a função de verossimilhança para o modelo de mistura padrão é dada por:

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n} \left[S_{pop}(t_i;\boldsymbol{\xi})\right]^{1-\delta_i} \left[f_{pop}(t_i;\boldsymbol{\xi})\right]^{\delta_i} \tag{3.8}$$

Portanto, aplicando (3.6) e (3.7) em (3.8), chegamos à função de verossimilhança para o modelo de mistura padrão com covariáveis e fragilidade:

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i)\left(1 + \frac{H_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\theta(\mathbf{z}_i)\left(1 + \frac{H_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1} h_0(t)\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]^{\delta_i} \tag{3.9}$$

Considerando a função de risco base seguindo uma distribuição Weibull com parâmetros α e λ, conforme (2.4), a função de verossimilhança (3.9) toma a seguinte forma, agora com $\boldsymbol{\xi} = (\sigma, \alpha, \lambda, \mathbf{b}, \boldsymbol{\beta})$:

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i)\left(1 + \frac{(t\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\theta(\mathbf{z}_i)\left(1 + \frac{(t\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1} \alpha\lambda\,(t\lambda)^{\alpha-1}\exp(\mathbf{x}_i'\boldsymbol{\beta})\right]^{\delta_i}$$

E consequentemente os seguintes parâmetros precisarão ser estimados:

- σ: parâmetro da distribuição do termo de fragilidade, gama.
- α: parâmetro de forma da distribuição da função de risco base, Weibull.
- λ: parâmetro de escala da distribuição da função de risco base, Weibull.
- $\mathbf{b} = (b_0, b_1, .., b_q)'$: parâmetros relativos às covariáveis que influenciam a probabilidade de cura, sendo b_0 o intercepto da regressão.
- $\boldsymbol{\beta} = (\beta_1, .., \beta_p)'$: parâmetros relativos às covariáveis que influenciam a função de sobrevivência.

Para se obter as estimativas de todos os parâmetros envolvidos através da metodologia clássica, será necessário maximizar (3.9). Entretanto diante de sua complexidade torna-se impossível encontrar seu ponto de máximo analiticamente. Diante disso é necessária a utilização de métodos numéricos para a obtenção de tais valores. Para tal finalidade utilizaremos a rotina *optim* do sistema R (R Development Core Team, 2009).

Com relação à identificabilidade deste modelo, Peng & Zhang (2008b) estudaram o modelo de mistura padrão com fragilidade e covariáveis em duas situações distintas: uma quando as covariáveis $\mathbf{z}$ que afetam a fração de cura são exatamente as mesmas que o conjunto $\mathbf{x}$, que representa as que interferem na distribuição latente, e também quando são conjuntos distintos. Eles mostraram que tal modelo é identificável se a fração de cura é modelada por uma função não constante e quando os dois componentes do modelo envolvem as mesmas covariáveis existe uma condição adicional que a distribuição da fragilidade seja proveniente de uma família completa. Como a distribuição gama é completa e a fração de cura não está sendo modelada por uma função constante, a identificabilidade do modelo está assegurada.

3.3 Métodos de seleção de modelos

A fim de comparar modelos, algumas técnicas estatísticas são comumente utilizadas. Uma delas é a do teste da razão de verossimilhança, que é apropriada para testar dois modelos encaixados, ou seja, o conjunto de covariáveis de um deles deve ser um subconjunto das covariáveis do outro. Este teste utiliza o valor do logaritmo da função de verossimilhança de cada modelo.

É possível mostrar, sob certas condições de regularidade e sob a hipótese $H_0 : \boldsymbol{\xi} = \boldsymbol{\xi}_0$, que a estatística da razão de verossimilhança $\Lambda = 2\left[l(\hat{\boldsymbol{\xi}}) - l(\hat{\boldsymbol{\xi}_0})\right]$ possui assintoticamente distribuição χ^2_p, em que p é a diferença entre o número de parâmetro dos dois modelos (Lawless, 2002).

Outra técnica tradicional utilizada para a seleção de modelos é o AIC *(Akaike Information Criterion)*, proposto por Akaike (1974). A estatística AIC

é dada por:

$$AIC = -2l(\boldsymbol{\xi}) + 2d,$$

em que $l(\boldsymbol{\xi})$ denota o logaritmo da função de verossimilhança e d é o número de parâmetros do modelo.

Schwarz (1978) propôs uma pequena alteração ao AIC, o BIC *(Bayesian Information Criterion)*, que é definido como:

$$BIC = -2l(\boldsymbol{\xi}) + d\log(n),$$

em que n representa o tamanho da amostra.

Tanto para o AIC quanto para o BIC, menores valores correspondem aos melhores modelos. É interessante notar que estes critérios comparam modelos que não são encaixados ou mesmo com números diferentes de parâmetros, pois consideram o número de parâmetros e penalizam a verossimilhança de modelos com muitos parâmetros.

3.4 Considerações finais

O modelo de mistura padrão de Berkson & Gage (1952) assume implicitamente que todos os indivíduos que falharam pertencem a uma população homogênea. Entretanto, sabemos que os indivíduos são diferentes e consequentemente há variações biológicas entre eles que não são mensuráveis.

Com o objetivo de quantificar a heterogeneidade não observável, adicionamos um termo de fragilidade ao modelo de mistura padrão de Berkson & Gage (1952) e incorporamos covariáveis na taxa de cura e na função de sobrevivência dos não curados, adotando um método de estimação paramétrico, assumindo que os tempos de vida dos indivíduos em risco seguem uma distribuição Weibull.

Capítulo 4

Aplicações

4.1 Estudos de Simulação

Com a finalidade de verificar se as propriedades assintóticas dos estimadores de máxima verossimilhança são válidas para nosso modelo, realizamos algumas simulações com dados gerados artificialmente. Inicialmente consideramos o modelo com uma covariável tanto na fração de cura quanto na parcela não imune, e na sequência incluimos o componente de fragilidade a esse modelo. Isso também nos ajudará a dimensionar se, e quanto, a inclusão do parâmetro de fragilidade afetará tais propriedades e a velocidade de convergência. Aqui ainda consideraremos a função de ligação logito.

4.1.1 Geração dos dados e resultados

Os tamanhos de amostra determinados fora 30, 100, 300 e 500, e para cada um deles foram realizadas quantidades de simulações suficientes de modo a garantir que um número em torno de 1000 convergisse para a sumarização dos resultados.

Para cada amostra gerada foram determinados os estimadores de máxima verossimilhança através de um processo de otimização que utiliza a rotina *optim* do sistema R (R Development Core Team, 2009). Também foram calculados os

valores da probabilidade de cobertura, erro quadrático médio e desvio-padrão.

Para realizar estes procedimentos, iremos gerar valores para os tempos de ocorrência, de censura, da variável indicadora de censura e da covariável incluída nos modelos, que aqui será tratada como um binária, e será utilizada tanto para incidir na fração de cura quando no tempo de sobrevivência da parcela dos não imunes.

A geração do tempo de ocorrência foi realizada através do método da inversa, e neste contexto, para o modelo ainda sem fragilidade, utilizamos a $S_{pop}(.)$ dada em (3.2) e já consideramos a distribuição Weibull para o tempo de ocorrência. Também atribuímos os seguintes valores para a geração dos dados: $\alpha = 1$, $\lambda = 1,2$, $b_0 = 0,5$, $b_1 = 3$ e $\beta_1 = 0,5$ e assumimos uma distribuição exponencial com média 10 para o tempo de censura.

Desta maneira, o algoritmo de geração dos dados terá os seguintes passos:

1. Gerar a covariável X de uma distribuição binomial com $n = 30$, 100, 300 e 500 e $p =$ 0,5;

2. Calcular a probabilidade de cura para cada elemento da amostra, $1 - \theta_i$, utilizando a covariável gerada, os valores atribuídos a b_0 e b_1 e a função de ligação logito;

3. Gerar u_i de uma distribuição uniforme $(0, 1)$;

4. Se $u_i < 1 - \theta_i$ então $y_i = \infty$, caso contrário:

$$y_i = \left[\frac{-\log\left(\frac{u_i - 1 + \pi(z_i)}{\pi(z_i)}\right)}{\lambda \exp(x_i'\beta)} \right]^{1/\alpha};$$

5. Gerar o tempo de censura, c_i, de uma distribuição exponencial com média 10;

6. Fazer $t_i = \min\{y_i, c_i\}$;

7. Se $y_i < c_i$ então $\delta_i = 1$, caso contrário $\delta_i = 0$.

As Tabelas 4.1 e 4.2 apresentam os valores das estatísticas de interesse obtidos para cada um dos tamanhos amostrais.

Tabela 4.1: Resultados das simulações para $n = 30$ e $n = 100$.

		n = 30				n = 100			
Parâmetro	Verd. Valor	Média	EQM	DP	PC	Média	EQM	DP	PC
α	1,0	$1,112$	$0,058$	$0,213$	$0,940$	$1,026$	$0,012$	$0,108$	$0,928$
λ	1,2	$1,290$	$0,435$	$0,654$	$0,863$	$1,232$	$0,089$	$0,297$	$0,933$
β_1	0,5	$0,527$	$0,389$	$0,624$	$0,912$	$0,501$	$0,079$	$0,281$	$0,943$
b_0	0,5	$0,795$	$0,840$	$0,868$	$0,986$	$0,510$	$0,146$	$0,379$	$0,948$
b_1	3,0	$3,459$	$4,485$	$2,068$	$0,948$	$3,212$	$2,002$	$1,335$	$0,970$

Tabela 4.2: Resultados das simulações para $n = 300$ e $n = 500$.

		n = 300				n = 500			
Parâmetro	Verd. Valor	Média	EQM	DP	PC	Média	EQM	DP	PC
α	1,0	$1,006$	$0,003$	$0,059$	$0,931$	$1,002$	$0,002$	$0,045$	$0,947$
λ	1,2	$1,206$	$0,026$	$0,162$	$0,918$	$1,207$	$0,016$	$0,126$	$0,926$
β_1	0,5	$0,501$	$0,027$	$0,165$	$0,925$	$0,504$	$0,017$	$0,130$	$0,924$
b_0	0,5	$0,510$	$0,042$	$0,204$	$0,926$	$0,511$	$0,026$	$0,161$	$0,931$
b_1	3,0	$3,212$	$0,790$	$0,863$	$0,946$	$3,089$	$0,301$	$0,541$	$0,954$

Notamos que embora as estimativas obtidas se mostrem satisfatórias mesmo para tamanhos de amostras menores, o aumento do tamanho amostral faz com que os valores médios dos estimadores se aproximem cada vez mais dos verdadeiros valores utilizados para a geração, da mesma forma que as estatísticas de variabilidade diminuem com o aumento da amostra.

As Figuras 4.1 e 4.2 mostram a evolução do erro quadrático médio das estimativas conforme aumentamos o tamanho amostral, enquanto as Figuras 4.3 e 4.4 mostram como se comportou a probabilidade de cobertura, para cada parâmetro.

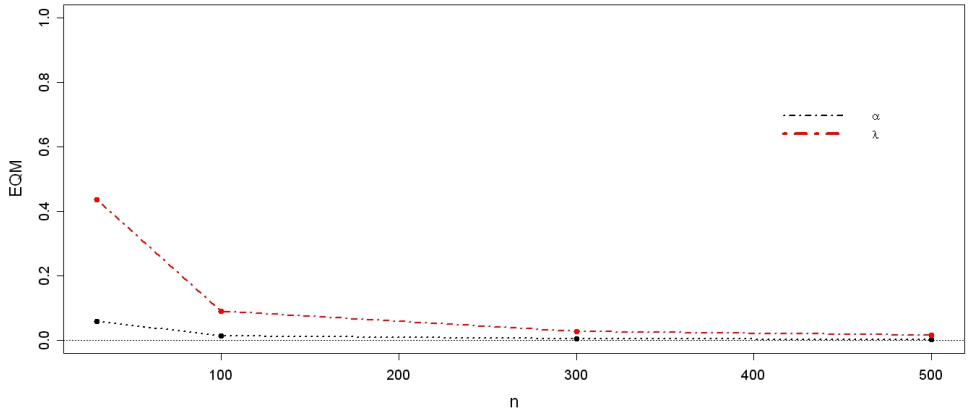

Figura 4.1: Erro quadrático médio para α e λ

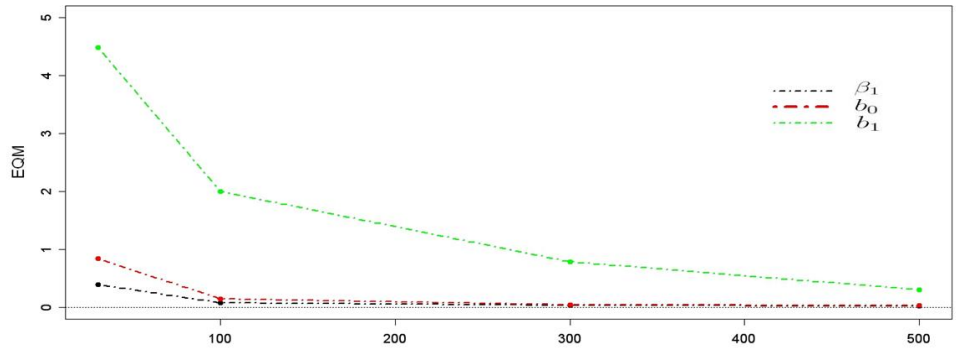

Figura 4.2: Erro quadrático médio para β_1, b_0 e b_1

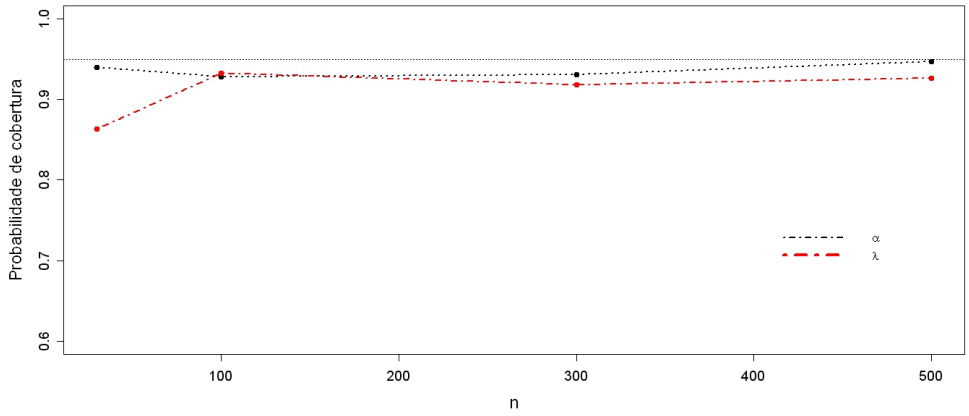

Figura 4.3: Probabilidades de cobertura para α e λ

Notamos que o erro quadrático médio diminui significativamente, e a probabilidade de cobertura se aproxima do valor nominal de 0,95 com o aumento do tamanho amostral, para todos os parâmetros.

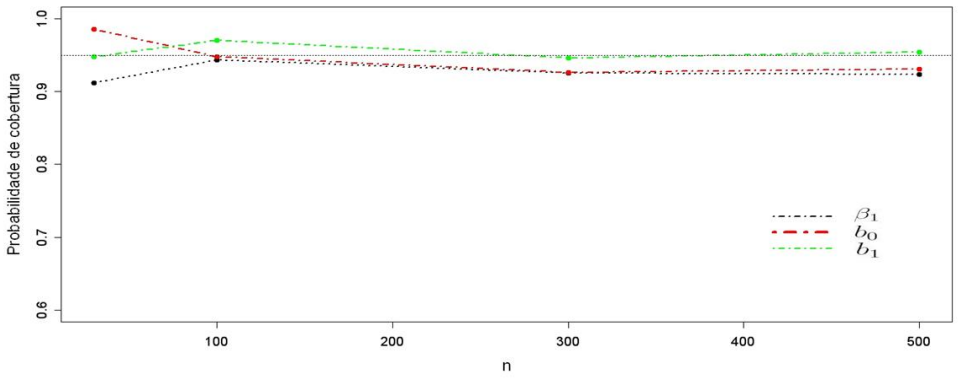

Figura 4.4: Probabilidades de cobertura para β_1, b_0 e b_1

Para a geração dos dados artificiais do modelo que incluiu o componente de fragilidade, cuja $S_{pop}(.)$ foi demonstrada em (3.6), adotamos a mesma estratégia. As diferenças principais são que nesta situação temos seis parâmetros, dada a inclusão da variável de fragilidade, com seu parâmetro σ. Atribuímos os seguintes valores para a geração dos dados: $\alpha = 1$, $\lambda = 1,2$, $\sigma = 1$, $b_0 = 0,5$, $b_1 = 3$ e $\beta_1 = 0,5$. Desta forma, o algoritmo de geração dos dados será o seguinte:

1. Gerar a covariável X de uma distribuição binomial com $n = 30$, 100, 300 e 500 e $p = 0{,}5$;

2. Calcular a probabilidade de cura para cada elemento da amostra, $1 - \theta_i$, utilizando a covariável gerada, os valores iniciais de b_0 e b_1 e a função de ligação logito;

3. Gerar u_i de uma distribuição uniforme $(0, 1)$;

4. Se $u_i < 1 - \theta_i$ então $y_i = \infty$, caso contrário:

$$y_i = \left[\frac{-\lambda^{-\alpha}\sigma + \lambda^{-\alpha}\sigma \left(\frac{u_i - 1 + \pi(z_i)}{\pi(z_i)} \right)^{\frac{-1}{\sigma}}}{\exp(x_i'\beta)} \right]^{1/\alpha} ;$$

5. Gerar o tempo de censura, c_i, de uma distribuição exponencial com média 10;

6. Fazer $t_i = \min\{y_i, c_i\}$;

7. Se $y_i < c_i$ então $\delta_i = 1$, caso contrário $\delta_i = 0$.

As Tabelas 4.3 e 4.4 apresentam os valores obtidos para cada um dos

tamanhos amostrais:

Tabela 4.3: Resultados das simulações para $n = 30$ e $n = 100$.

		n = 30				n = 100			
Parâmetro	Verd. Valor	Média	EQM	DP	PC	Média	EQM	DP	PC
α	1,0	1,422	0,811	0,796	0,985	1,125	0,095	0,282	0,970
λ	1,2	1,042	0,872	0,761	0,752	1,075	0,293	0,527	0,817
σ	1,0	1,846	4,423	2,099	0,826	1,434	1,296	1,053	0,875
β_1	0,5	0,674	1,709	1,297	0,979	0,669	0,432	0,635	0,946
b_0	0,5	0,903	2,081	1,386	0,972	0,592	0,224	0,465	0,960
b_1	3,0	3,391	4,742	2,144	0,949	3,463	3,294	1,756	0,932

Tabela 4.4: Resultados das simulações para n = 300 e n = 500.

		n = 300				n = 500			
Parâmetro	Verd. Valor	Média	EQM	DP	PC	Média	EQM	DP	PC
α	1,0	1,022	0,012	0,109	0,977	1,012	0,006	0,080	0,974
λ	1,2	1,089	0,075	0,251	0,939	1,083	0,044	0,175	0,944
σ	1,0	1,286	0,283	0,449	0,963	1,234	0,159	0,322	0,977
β_1	0,5	0,598	0,093	0,289	0,955	0,588	0,057	0,222	0,944
b_0	0,5	0,562	0,091	0,296	0,956	0,554	0,048	0,212	0,954
b_1	3,0	3,314	1,565	1,211	0,929	3,180	0,814	0,884	0,938

As Figuras 4.5 e 4.6 mostram como se comportou o erro quadrático médio dos parâmetros simulados, enquanto as Figuras 4.7 e 4.8 mostram como a probabilidade de cobertura, por tamanho de amostra, para cada parâmetro.

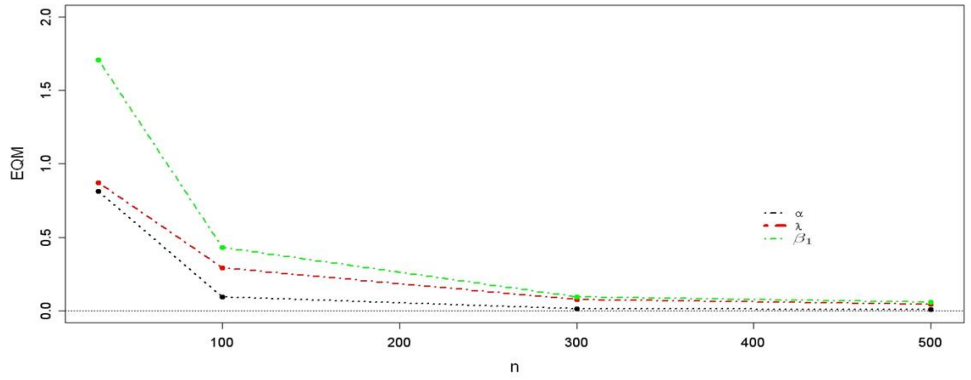

Figura 4.5: Erro quadrático médio para α, λ e β_1

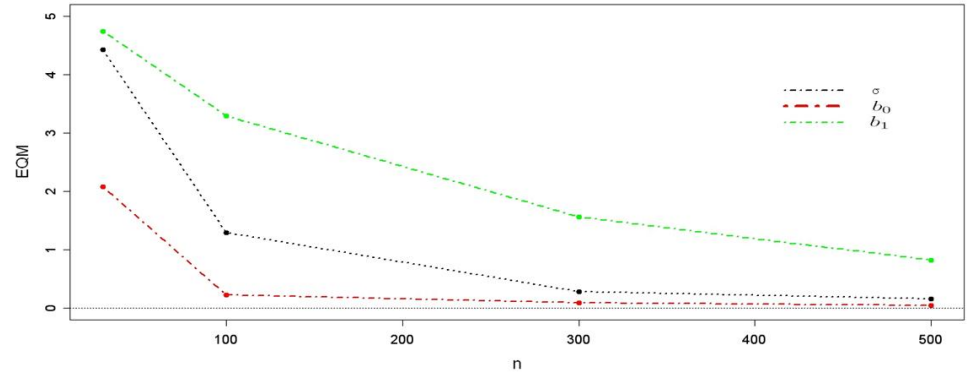

Figura 4.6: Erro quadrático médio para σ, b_0 e b_1

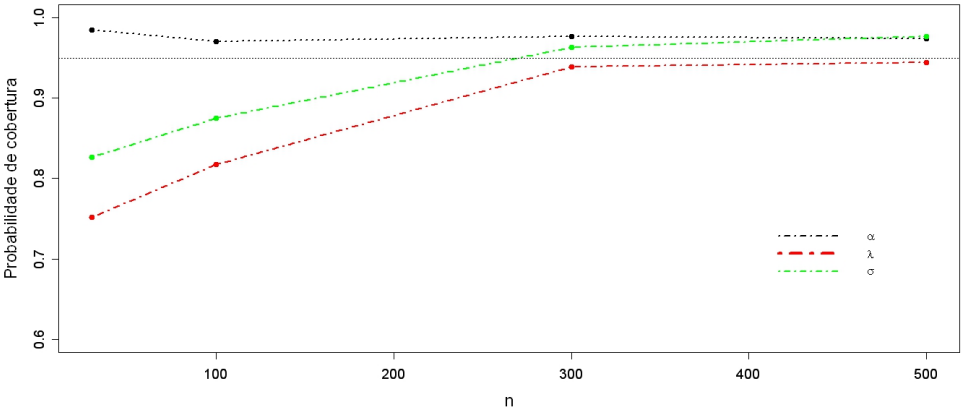

Figura 4.7: Probabilidades de cobertura para α, λ e σ

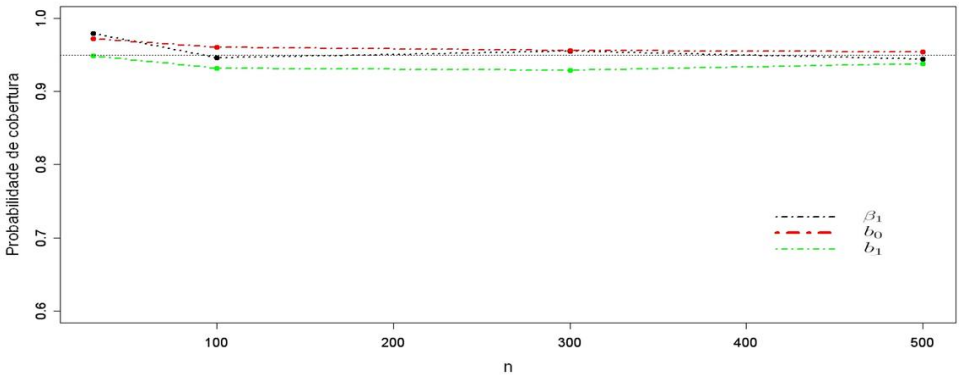

Figura 4.8: Probabilidades de cobertura para β_1, b_0 e b_1

Nestas simulações também verificamos que o aumento do tamanho amostral faz com que os valores médios dos estimadores se aproximem cada vez mais dos valores utilizados para a geração dos dados, embora de maneira mais lenta do que verificamos no modelo sem fragilidade, pois enquanto que no modelo sem fragilidade os valores para $n = 500$ praticamente são iguais aos usados na geração dos dados, no modelo com fragilidade eles ficaram um pouco mais distantes, embora se aproximando gradativamente.

Da mesma forma, as estatísticas de variabilidade diminuem e a probabilidade de cobertura se aproxima do valor nominal de 0,95 à medida que aumentamos o tamanho da amostra.

4.2 Aplicação com dados de melanoma

Nesta seção, aplicamos a teoria descrita no Capítulo 3 através de um exemplo com dados reais estudados por Kirkwood *et al.* (2000), em que iremos considerar (caso comprovemos sua adequabilidade) o modelo Weibull para os tempos de ocorrência do evento de interesse. Toda a programação utilizada neste trabalho foi desenvolvida em linguagem de programação R (R Development Core Team, 2009)

O conjunto de dados para esta aplicação provém de um estudo de melanoma, que foi realizado com o objetivo de avaliar a eficácia da aplicação de uma dosagem alta de interferon alfa-2b como forma de prevenir a recorrência do

câncer de pele. Os pacientes foram incluídos no estudo entre 1991 e 1995, tendo sido acompanhados até 1998.

A variável resposta Y representa o tempo até a morte de paciente ou tempo de censura. Nesta amostra temos $n = 417$ pacientes, com 56% de observações censuradas. As variáveis incluem y: tempo (em anos); x_1: tipo de tratamento (0: sem tratamento; 1: interferon) ; x_2: idade do paciente; x_3: categoria do nódulo $(1, 2, 3, 4)$; x_4: sexo (0: masculino; 1: feminino); x_5: capacidade funcional (0: ativo; 1: outras) e x_6: espessura do tumor(em mm).

A Figura 4.1 mostra a estimativa de Kaplan-Meier para a função de sobrevivência deste conjunto de dados. Observamos que com o passar do tempo a curva se estabiliza em um patamar acima de 0, o que é um indício de que uma parcela dos pacientes esteja curada. Assim torna-se interessante ajustar um modelo que considere essa possibilidade.

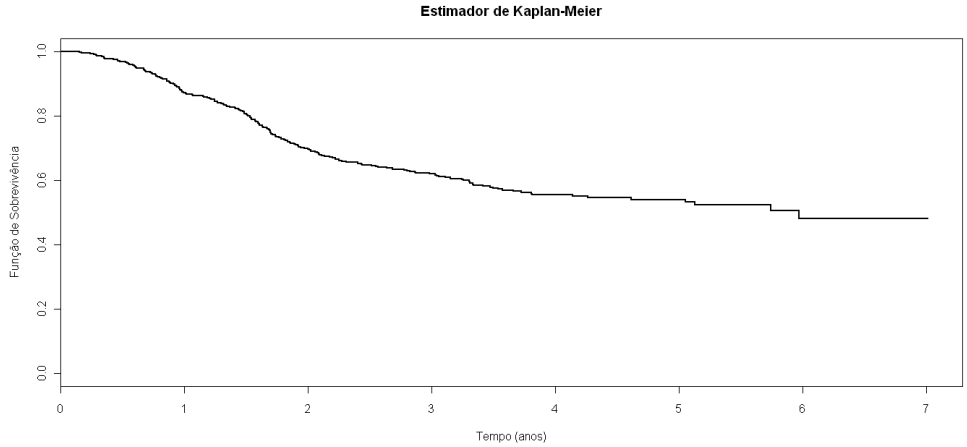

Figura 4.9: Curva de Kaplan-Meier

Além da variável resposta e do indicador de censura, existem cinco covariáveis que poderão ser relevantes tanto na probabilidade de cura quanto na função de sobrevivência dos pacientes não curados. Outros interesses serão determinar se existe e qual a proporção de curados, e se o efeito da fragilidade, que representa a heterogeneidade não explicada pelas covariáveis, também é significativo.

Antes de avaliar quais modelos poderão ser ajustados com os dados disponíveis, deve-se verificar o comportamento da função de risco dos tempos observados. Aarset (1985) propôs um método gráfico denominado TTT-Plot

(Total Time Test), em que sua versão empírica é dada por:

$$G(r/n) = \frac{\sum_{i=1}^{r} Y_{i:n} - (n-r)Y_{r:n}}{\sum_{i=1}^{r} Y_{i:n}}$$

Na fórmula acima, $r = 1, 2, ..., n$ e $Y_{i:n}$ representam as estatísticas de ordem da amostra. Caso o gráfico se aproxime de uma linha diagonal pode-se aceitar que a função de risco seja constante. Se apresentar um formato côncavo, o risco deve crescer ao longo do tempo, e em caso de um formato convexo, o risco é decrescente.

Por outro lado, caso o formato da curvatura se altere de côncavo para convexo, o risco tem forma unimodal, enquanto que se passar de convexo para côncavo a função de risco teria forma de banheira (Calsavara, 2011).

A seguir, mostramos na Figura 4.2 o TTT Plot para o conjunto de dados a ser analisado:

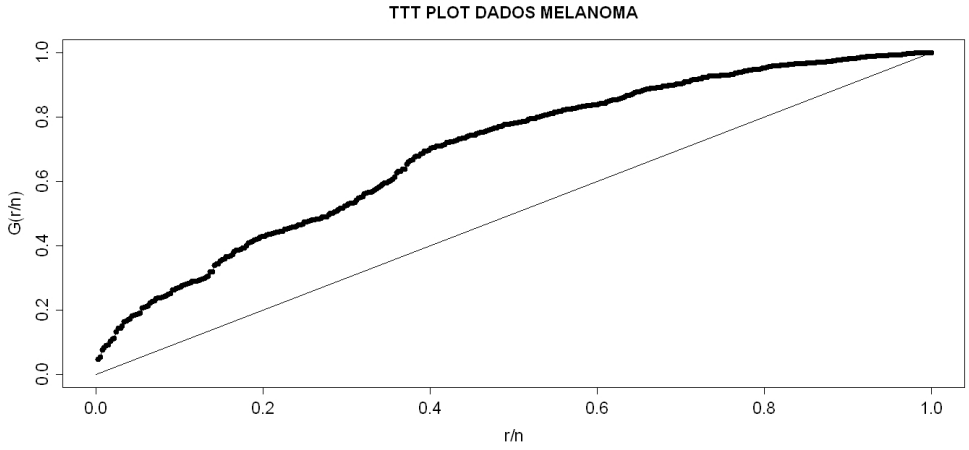

Figura 4.10: TTT Plot Dados Melanoma.

Como a curva apresentou formato côncavo, assumimos que a função de risco cresce com o passar do tempo, o que inviabilizaria, por exemplo, a utilização da distribuição exponencial para a função de risco base. Assim, está justificada a utilização da distribuição Weibull, que acomoda esse tipo de comportamento, monótono e crescente, da função de risco.

4.2.1 Modelos e resultados

Analisamos este conjunto de dados reais considerando três modelos: o Modelo 1, que considerou covariáveis apenas na fração de cura, e apresenta as seguintes funções:

$$S_{pop}(t|\mathbf{z}) = 1 - \theta(\mathbf{z}) + \theta(\mathbf{z})S(t), \tag{4.1}$$

e

$$f_{pop}(t|\mathbf{z}) = \theta(\mathbf{z})f(t); \tag{4.2}$$

o Modelo 2, que considerou covariáveis tanto na fração de cura quanto na função de sobrevivência dos não curados, conforme visto em (3.2) e (3.3); e o Modelo 3, que adicionou o componente de fragilidade ao Modelo 2, e que foi demonstrado em (3.6) e (3.7).

Os três modelos serão ajustados e comparados, todos considerando como função de risco base a distribuição Weibull (cuja escolha já foi justificada) e testando a mesma covariável tanto na fração de cura quanto na função de sobrevivência dos não curados, que é a covariável x_3, que identifica a categoria do nódulo (possuindo os níveis 1, 2, 3 e 4).

Esta escolha foi baseada em outros trabalhos que ajustaram modelos com essa mesma base de dados, mostrando a categoria do nódulo é a mais (quando não única) covariável significativa (Rodrigues *et al.*, 2008). A exemplo do que também foi feito nestes estudos, também a trataremos como uma variável quantitativa (Ortega *et al.*, 2008), o que gerará apenas um parâmetro a ser estimado, ao invés dos quatro que seriam necessários se optássemos criar uma variável binária para cada um dos níveis.

Primeiramente exibiremos as estimativas de máxima verossimilhança dos parâmetros de cada modelo e seus respectivos desvios-padrão, para cada uma das três funções de ligação, bem como as estatísticas de comparação de modelos e testes de significância entre os modelos. Na sequência exibiremos análises gráficas e as proporções estimadas de curados para cada modelo e função de ligação.

Desta forma, no Modelo 1, para cada um dos links, quatro parâmetros serão estimados: os dois referentes à distribuição de probabilidade da função de

sobrevivência dos não curados, Weibull (α e λ); b_0, que representará o intercepto, e b_1, que é o parâmetro associado à covariável utilizada na regressão relativa à probabilidade de cura (o nível do tumor). Os resultados das estimativas de máxima verossimilhança dos parâmetros do modelo e seus desvios-padrão são apresentados na Tabela 4.5.

Tabela 4.5: Resultados para o modelo 1.

	Link Logito		Link Probito		Link Compl. Log-Log	
Parâmetros	EMV	DP	EMV	DP	EMV	DP
α	$1,6168$	$0,1057$	$1,6165$	$0,1057$	$1,6171$	$0,1057$
λ	$0,4519$	$0,0275$	$0,4518$	$0,0275$	$0,4516$	$0,0275$
b_0	$-1,1647$	$0,2658$	$-0,7240$	$0,1629$	$0,4252$	$0,1716$
b_1	$0,4675$	$0,1033$	$0,2910$	$0,0634$	$-0,3254$	$0,0730$
Log Veross.	$-517,591$		$-517,587$		$-517,853$	
AIC	$1043,182$		$1043,175$		$1043,708$	
BIC	$1059,314$		$1059,308$		$1059,840$	

Observamos na Tabela 4.5 que os parâmetros da distribuição Weibull tiveram valores muito próximos para os três links, e, como era de se esperar, b_0 e b_1, que dependem da função de ligação escolhida, tiveram valores distintos de acordo com o link. Também notamos que todas as estimativas apresentaram variabilidade muito baixa, o que atesta a significância de todos os parâmetros. Já com relação aos valores das estatísticas AIC e BIC e do logaritmo da função de verossimilhança, podemos verificar que todos foram praticamente iguais para as três funções de ligação utilizadas.

Para Modelo 2, cuja $S_{pop}(.)$ e $f_{pop}(.)$ foram mostradas em (3.2) e (3.3), o parâmetro β_1 foi acrescentado, e será relativo à covariável adicionada à função de sobrevivência dos não curados. Desta maneira, para cada função de ligação, cinco parâmetros serão ajustados: os quatro do modelo anterior e o novo parâmetro β_1. A Tabela 4.6 sumariza os resultados obtidos.

Tabela 4.6: Resultados para o Modelo 2.

	Link Logito		Link Probito		Link Compl. Log-Log	
Parâmetros	EMV	DP	EMV	DP	EMV	DP
α	$1,6084$	$0,1045$	$1,6081$	$0,1045$	$1,6069$	$0,1046$
λ	$0,3174$	$0,0559$	$0,3172$	$0,0559$	$0,3156$	$0,0560$
β_1	$0,2136$	$0,0883$	$0,2137$	$0,0883$	$0,2160$	$0,0885$
b_0	$-0,9736$	$0,2962$	$-0,6064$	$0,1827$	$0,3005$	$0,1952$
b_1	$0,3989$	$0,1080$	$0,2487$	$0,0666$	$-0,2795$	$0,0760$
Log Veross.	$-514,4713$		$-514,467$		$-514,6631$	
AIC	$1038,943$		$1038,934$		$1039,326$	
BIC	$1059,108$		$1059,100$		$1059,492$	

Novamente as estimativas apresentaram variabilidade muito baixa para todos os parâmetros do modelo e os três links não mostraram diferenças relevantes em relação às estatísticas de comparação AIC, BIC e o logaritmo da verossimilhança.

Já em relação ao Modelo 1, os valores de AIC e BIC mostraram uma redução, o que indica que a adição do novo parâmetro pode ter sido significativa. Para determinar se o Modelo 2 trouxe ganhos significativos em relação ao Modelo 1, foi realizado o teste da razão de verossimilhança, visto na Seção 2.5. A estatística deste teste apresentou um valor de $6,2$, que originou um valor-p de $0,012$. Deste modo concluímos que existe diferença entre esses dois modelos, ou seja, a adição de mais um parâmetro se mostrou relevante na explicação da resposta.

Passando para o Modelo 3, conforme $S_{pop}(.)$ e $f_{pop}(.)$ definidas em (3.6) e (3.7), adicionamos o termo de fragilidade, e o seu parâmetro a ser estimado, σ, é relativo à função de distribuição escolhida para este novo componente - que no caso foi a gama com parâmetros (σ, σ). Desta forma, $1/\sigma$ será o valor estimado da heterogenidade não observada. A Tabela 4.7 resume os valores obtidos no ajuste do Modelo 3.

Tabela 4.7: Resultados para o Modelo 3.

	Link Logito		Link Probito		Link Compl. Log-Log	
Parâmetros	EMV	DP	EMV	DP	EMV	DP
α	$2,4285$	$0,3015$	$2,4301$	$0,3021$	$2,4301$	$0,3022$
λ	$0,3312$	$0,0586$	$0,3311$	$0,0587$	$0,3292$	$0,0587$
σ	$0,6645$	$0,3487$	$0,6603$	$0,3478$	$0,6555$	$0,3453$
β_1	$0,5402$	$0,1675$	$0,5408$	$0,1678$	$0,5456$	$0,1686$
b_0	$-0,6777$	$0,3886$	$-0,4214$	$0,2427$	$0,1148$	$0,2670$
b_1	$0,3789$	$0,1230$	$0,2364$	$0,0760$	$-0,2794$	$0,0926$
Log Veross.	$-506,0406$		$-506,0191$		$-506,1804$	
AIC	$1024,081$		$1024,038$		$1024,361$	
BIC	$1048,280$		$1048,237$		$1048,559$	

Com base nos resultados da Tabela 4.7 observamos que alguns desvios-padrão cresceram proporcionalmente em relação ao valor de seus respectivos parâmetros (mais precisamente σ e b_0, que agora já não é mais significativo). Também notamos que novamente houve uma redução dos valores das estatísticas de comparação AIC e BIC em relação aos Modelos 1 e 2, o que é um indício de que o modelo com fragilidade se ajustou melhor aos dados.

Para comprovar que o Modelo 3 também trouxe ganhos significativos em relação ao Modelo 2, novamente foi realizado o teste da razão de verossimilhança, mostrado na na Seção 2.5. Desta vez a estatística do teste apresentou um valor de $16,9$, e consequentemente um valor-p $< 0,001$. Assim, podemos afirmar que a inclusão do termo de fragilidade foi algo significativo para o estudo.

Além disso, o valor estimado da variância da variável que representa a fragilidade, será desta forma $Var(\sigma) = 1/\hat{\sigma} = 1/0,66 = 1,51$. Dado que tal variável possui média 1, podemos justificar a inclusão do termo de fragilidade, que representará a variabilidade não explicada pela covariável.

Como uma análise complementar, nas Figuras 4.11 e 4.12, são comparadas as funções de sobrevivência geradas pelos Modelos 1, 2 e 3 com a curva de Kaplan-Meier, por nível da covariável. Neste caso, como já foi visto que as estatísticas de comparação mostraram que as três funções de ligação tiveram desempenhos praticamente idênticos, utilizamos apenas a ligação logito.

Visualmente, os três modelos mostraram ajustes satisfatórios e próximos à curva de Kaplan-Meier, para os quatro níveis da covariável testada. Entretanto observamos que em alguns trechos iniciais da curva de sobrevivência, o Modelo 3 tem uma aderência um pouco melhor do que os outros dois modelos.

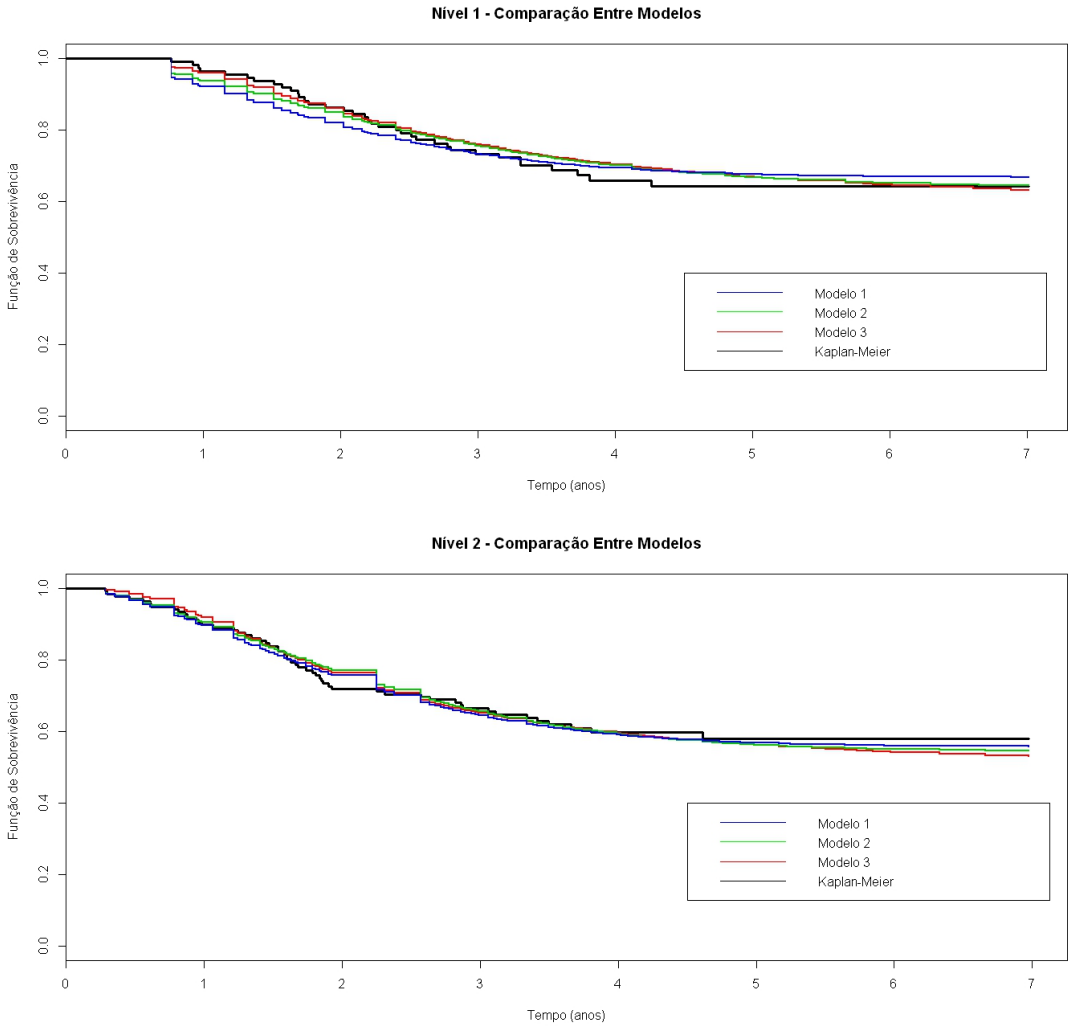

Figura 4.11: Curvas de sobrevivência estimadas e de Kaplan Mayer para os níveis 1 e 2 da covariável x_3.

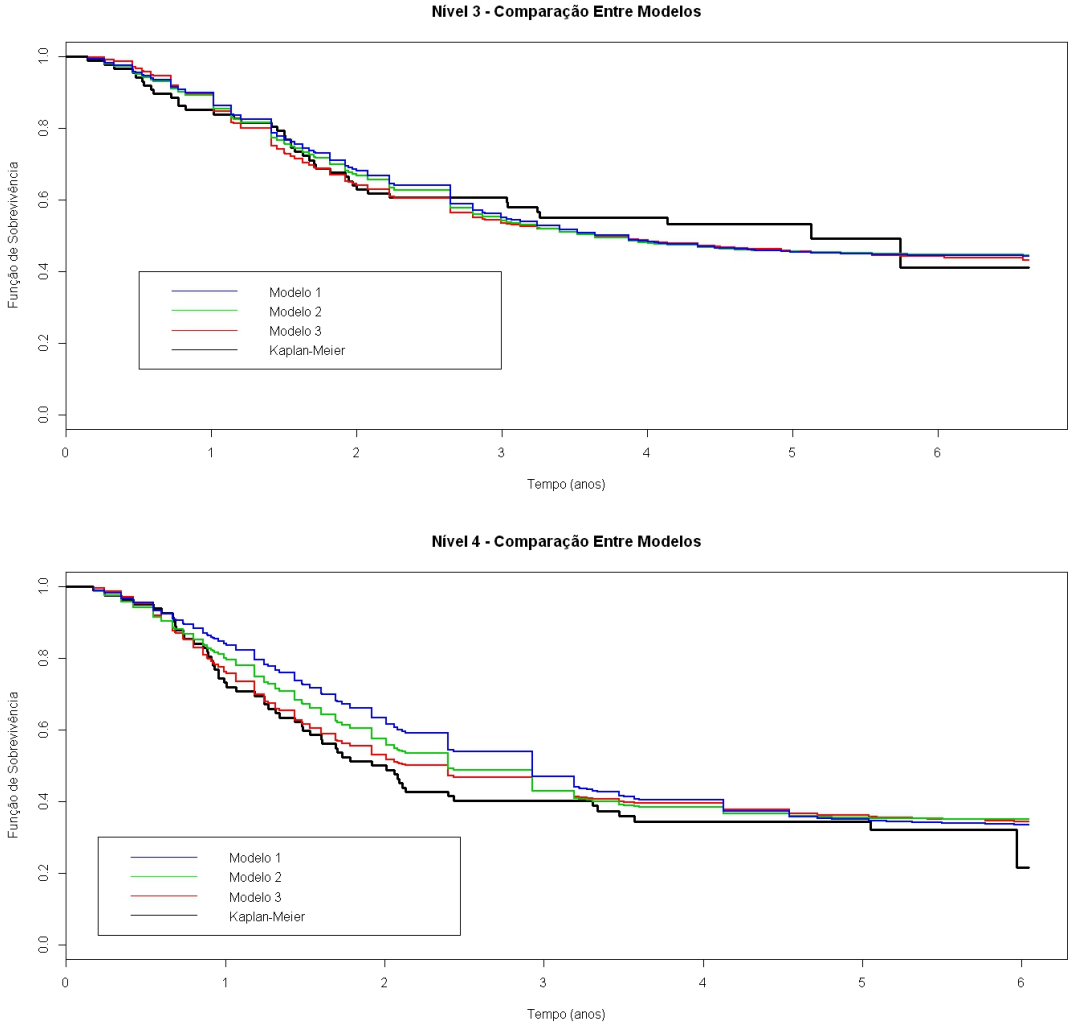

Figura 4.12: Curvas de sobrevivência estimadas e de Kaplan Mayer para os níveis 3 e 4 da covariável x_3.

Na sequência, calculamos para cada modelo a proporção de curados, através das respectivas estimativas de b_0 e b_1 aplicadas à cada função de ligação. Os resultados são exibidos na Tabela 4.8:

Tabela 4.8: Estimativas das proporções de cura.

	Modelo 1			Modelo 2			Modelo 3		
Níveis	Logito	Probito	Compl.	Logito	Probito	Compl.	Logito	Probito	Compl.
Nível 1	0,668	0,667	0,669	0,640	0,640	0,640	0,574	0,573	0,572
Nível 2	0,557	0,556	0,550	0,544	0,543	0,538	0,480	0,480	0,473
Nível 3	0,441	0,441	0,438	0,444	0,444	0,442	0,387	0,387	0,384
Nível 4	0,331	0,330	0,341	0,349	0,349	0,357	0,302	0,3001	0,307
Total	0,518	0,517	0,517	0,510	0,510	0,510	0,451	0,450	0,448

Notamos que, em relação aos níveis da covariável ocorreu o que esperávamos, ou seja, quanto maior o nível do tumor menor a proporção de cura, pois nesse caso a doença estaria num estágio mais avançado. Também percebemos que os resultados foram bastante similares dentro de cada modelo, independente da função de ligação escolhida, e que os Modelos 1 e 2 apresentaram resultados bastante próximos entre si, ao contrário do Modelo 3, que estimou uma proporção de curados menor que os outros.

É interessante notar que se tivéssemos utilizado a análise de sobrevivência tradicional, trataríamos os 56% de pacientes que não chegaram ao evento de interesse como meros censurados, ou seja, assumiriamos que todos eles acabariam falhando.

Incorporando a fração de cura e covariáveis, como nos Modelos 1 e 2, já poderíamos afirmar que em torno de 51% dos pacientes estariam curados, ou imunes, e nesse caso apenas 5% viriam a morrer por conta do melanoma no futuro.

E finalmente, ao considerarmos também o termo de fragilidade, através do Modelo 3, concluímos que a proporção de cura foi de aproximadamente 45%. Sendo assim, dos 56% originalmente censurados, 11% não estariam imunes.

Como um meio de comparação gráfica entre as funções de ligação, as Figuras 4.13 e 4.14 apresentam os gráficos do ajuste da curva de Kaplan-Meier juntamente com a sobrevivência estimada pelos Modelo 2 e Modelo 3 considerando a categoria do nódulo no nível 1 para as três diferentes funções de ligação.

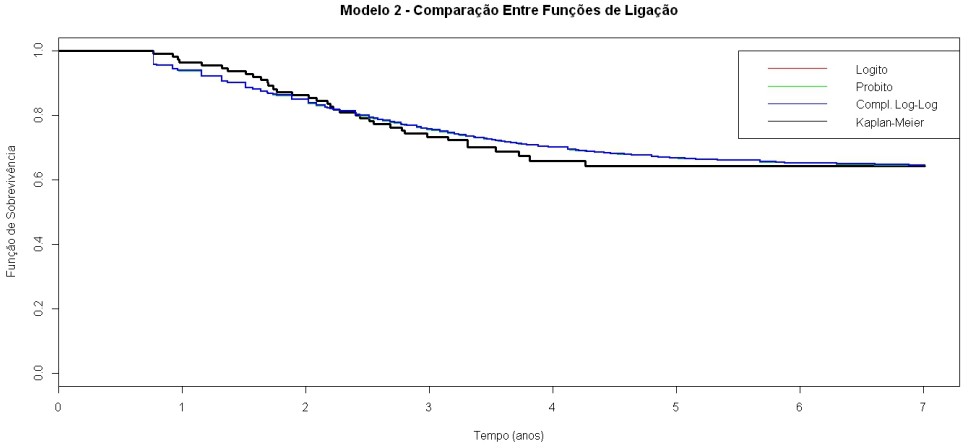

Figura 4.13: Comparação entre funções de ligação - Modelo 2.

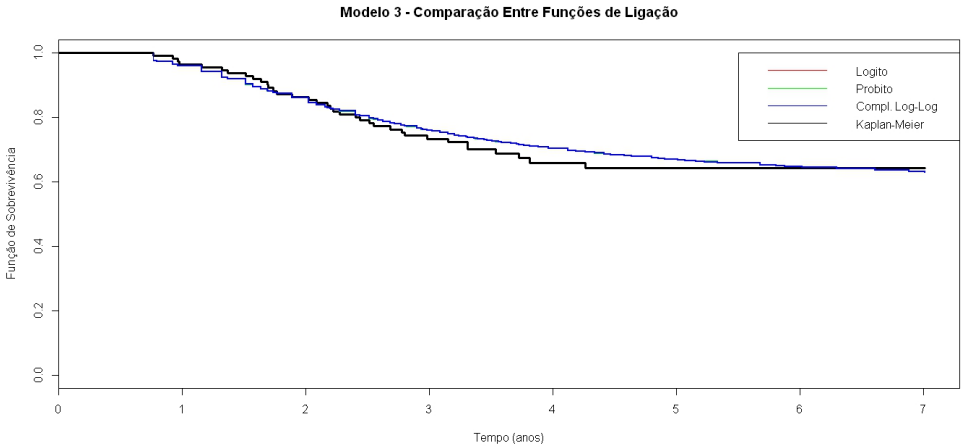

Figura 4.14: Comparação entre funções de ligação - Modelo 3.

Podemos notar que, visualmente, em ambos os modelos as três funções produziram ajustes rigorosamente iguais (o que já esperado dada a proximidade que as estatísticas de comparação já haviam mostrado), e que todos ajustes ficaram muito próximos à curva de Kaplan-Meier.

Além disso, mais uma vez é possível observar que no início da curva de sobrevivência o Modelo 3 fica um pouco mais próximo das estimativas de Kaplan-Meier do que o Modelo 2, reforçando a diferença entre ambos e a importância do componente de fragilidade para a melhoria do ajuste.

4.3 Considerações finais

Neste capítulo consideramos uma aplicação com dados reais de melanoma, em que assumimos que os tempos de ocorrência do evento dos indivíduos em risco seguem distribuição Weibull, obtendo assim o Modelo 3, denominado modelo de fragilidade com fração de cura e covariáveis.

Além disso fizemos a análise dos dados comparando-o a um modelo considerando covariáveis apenas na fração de cura (Modelo 1) e a outro com covariáveis na fração de cura e na função de sobrevivência dos não curados (Modelo 2).

Observamos que o Modelo 3 propiciou melhor ajuste aos dados, já que permitiu quantificar a heterogeneidade não observável, e ainda que as três funções de ligação utilizadas no processo de estimação do parâmetro de cura produziram resultados extremamente próximos.

Também realizamos estudos de simulação com dados artificiais, em que verificamos as propriedades assintóticas dos estimadores.

Capítulo 5

Abordagem Bayesiana dos Modelos

Neste capítulo vamos considerar o método bayesiano para estimar os parâmetros dos Modelos 1, 2 e 3 considerando os mesmos dados reais utilizados na Seção 4.2, de pacientes portadores de melanoma. Aqui, novamente a covariável utilizada será a categoria do nódulo.

Dentro do enfoque bayesiano, o desconhecimento é representado através de modelos probabilísticos para os parâmetros. Desta maneira, não existe nenhuma distinção entre quantidades observáveis e os parâmetros de um modelo estatístico. Todos são considerados quantidades aleatórias.

Dentre os componentes da inferência bayesiana, estão a distribuição a priori, $\pi(\theta)$, que expressa o conhecimento que temos a respeito do parâmetro de interesse θ antes de observarmos os dados, e a função de verossimilhança $L(\mathbf{x}|\theta)$, que utiliza os dados observados no experimento e que conecta as distribuições a priori e posteriori.

Por sua vez, a distribuição a posteriori, $\pi(\theta|\mathbf{x})$, expressa o conhecimento que temos a respeito do parâmetro depois de observarmos os dados. Assim, fazendo uma analogia com o Teorema de Bayes, podemos escrever:

$$\pi(\theta|\mathbf{x}) \propto L(\mathbf{x}|\theta)\pi(\theta).$$

Isto significa que a probabilidade $\pi(\theta)$ será revista com base nos dados observados $\mathbf{x}$, dando origem a $\pi(\theta)|\mathbf{x})$. Ou seja, os dados observados "corrigem"

a informação inicial. A partir da distribuição a posteriori do parâmetro, podemos examinar qualquer aspecto de θ (média, variância, percentis, etc).

5.1 Distribuições a priori dos modelos

Para a estimação dos parâmetros dos três modelos através da metodologia bayesiana, será necessário determinar a distribuição a priori para cada um de seus parâmetros. A seguir mostramos as distribuições escolhidas, sendo que o Modelo 1 utilizará apenas quatro delas, o Modelo 2, cinco, e o Modelo 3 utilizará todas. Para os parâmetros positivos assumiremos distribuições a priori seguindo a distribuição gama, e para os parâmetros que podem assumir valores dentro do intervalo dos números reais escolhemos como distribuição a priori a distribuição normal:

$$\pi(\alpha) = \frac{b^a}{\Gamma(a)}\ \alpha^{a-1}\exp(-b\alpha), \quad a, b > 0$$

$$\pi(\lambda) = \frac{d^c}{\Gamma(c)}\ \lambda^{c-1}\exp(-d\lambda), \quad c, d > 0$$

$$\pi(\sigma) = \frac{s^r}{\Gamma(r)}\ \sigma^{r-1}\exp(-s\sigma), \quad r, s > 0$$

$$\pi(b_0) = (2\pi h)^{-1/2} \quad \exp\left(\frac{-(b_0-g)^2}{2h}\right) \quad g \in R, \quad h > 0$$

$$\pi(b_1) = (2\pi m)^{-1/2} \quad \exp\left(\frac{-(b_1-k)^2}{2m}\right) \quad k \in R, \quad m > 0$$

$$\pi(\beta_1) = (2\pi q)^{-1/2} \quad \exp\left(\frac{-(\beta_1-p)^2}{2q}\right) \quad p \in R, \quad q > 0,$$

onde a, b, c, d, r e s são hiperparâmetros relativos às distribuições gama atribuídas às prioris dos parâmetros α, λ e σ e g, respectivamente e h, k, m, p e q são os hiperparâmetros relativos às prioris dos parâmetros b_0, b_1 e β_1, respectivamente.

Todos os três modelos serão ajustados apenas com a função de ligação logito, dado que detectamos que não existiu, através dos resultados obtidos pela metodologia clássica, vantagens de uma função de ligação em relação às demais

(ver Seção 4.2). Assim, para cada um dos modelos serão definidas as densidades a priori conjuntas, as densidades a posteriori e as condicionais completas a posteriori para cada parâmetro.

5.2 Distribuições a posteriori para o Modelo 1

Supondo que para o Modelo 1 os parâmetros de interesse são independentes, a priori conjunta será o produto de suas respectivas funções densidade. Deste modo:

$$\pi(\alpha, \lambda, b_0, b_1) = \pi(\alpha)\ \pi(\lambda)\ \pi(b_0)\ \pi(b_1) \tag{5.1}$$

Para a obtenção da densidade a posteriori, é necessário definirmos a função de verossimilhança, que é relacionada à $S_{pop}(.)$ e à $f_{pop}(.)$. Para o Modelo 1, considerando (4.1) e (4.2), o vetor de parâmetros $\boldsymbol{\xi} = (\alpha, \lambda, b_0 \text{ e } b_1)$ e o risco base proveniente de uma distribuição Weibull, a função de verossimilhança será dada por:

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i) \exp\left[-(t_i\lambda)^{\alpha}\right]\right]^{1-\delta_i} \left[\theta(\mathbf{z}_i)\alpha\lambda\,(t_i\lambda)^{\alpha-1} \exp\left[-(t_i\lambda)^{\alpha}\right]\right]^{\delta_i} \tag{5.2}$$

E, combinando (5.1) e (5.2), obtemos a densidade a posteriori:

$$\pi(\alpha, \lambda, b_0, b_1|\mathbf{D}) \propto \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i) \exp\left[-(t_i\lambda)^{\alpha}\right]\right]^{1-\delta_i} \ \times$$

$$\left[\theta(\mathbf{z}_i)\alpha\lambda\,(t_i\lambda)^{\alpha-1} \exp\left[-(t_i\lambda)^{\alpha}\right]\right]^{\delta_i} \ \alpha^{a-1} \exp(-b\alpha) \ \times$$

$$\lambda^{c-1} \exp(-d\lambda) \ \exp\left(\frac{-(b_0 - g)^2}{2h}\right) \ \exp\left(\frac{-(b_1 - k)^2}{2m}\right)$$

Desta forma, as distribuições condicionais completas a posteriori de cada parâmetro serão:

$$\pi(\alpha|\lambda, b_0, b_1, \mathbf{D}) \propto \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i) \exp\left[-(t_i\lambda)^{\alpha}\right]\right]^{1-\delta_i} \left[\alpha\lambda^{\alpha} t_i^{\alpha-1} \exp\left[-(t_i\lambda)^{\alpha}\right]\right]^{\delta_i} \times$$

$$\alpha^{a-1} \exp(-b\alpha)$$

$$\pi(\lambda|\alpha,b_0,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]\right]^{1-\delta_i}\left[\lambda^{\alpha}t_i^{\alpha-1}\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]\right]^{\delta_i}\times$$

$$\lambda^{c-1}\exp(-d\lambda)$$

$$\pi(b_0|\alpha,\lambda,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]\right]^{1-\delta_i}\left[\theta(\mathbf{z}_i)\right]^{\delta_i}\exp\left(\frac{-(b_0-g)^2}{2h}\right)$$

$$\pi(b_1|\alpha,\lambda,b_0,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]\right]^{1-\delta_i}\left[\theta(\mathbf{z}_i)\right]^{\delta_i}\exp\left(\frac{-(b_1-k)^2}{2m}\right)$$

Como estas densidades condicionais a posteriori não possuem a forma de nenhuma das distribuições conhecidas, para gerar valores para os quatro parâmetros de interesse utilizaremos o algoritmo de Metropolis-Hastings (Hastings, 1970). Com ele, conseguiremos simular amostras da distribuição conjunta utilizando as distribuições condicionais completas destes parâmetros.

De acordo com a teoria das cadeias de Markov, esperamos que as cadeias geradas convirjam para uma distribuição estacionária, que no caso também é nossa distribuição de interesse. Para verificar se tais cadeias de fato convergiram, existem alguns testes que podemos fazer, tanto numéricos quanto visuais. Neste estudo, a convergência das cadeias será avaliada pelo método numérico de Geweke (1992) e pela análise gráfica da densidade a posteriori.

O diagnóstico de Geweke considera duas partes distintas da cadeia de Markov gerada (normalmente os primeiros 10% e os últimos 50%), e compara as médias destas partes, utilizando um teste de diferença entre as médias, a fim de verificar se estas duas partes da cadeia são provenientes da mesma distribuição (hipótese nula). A estatística teste é uma estatística Z padrão, e desta forma não descartamos a convergência da cadeia para casos em que $-2 \leq Z \leq 2$.

Para a geração das cadeias do Modelo 1, consideramos distribuições a priori independentes, assumindo valores dos hiperparâmetros de forma que as prioris fiquem não informativas, ou seja, em princípio os parâmetros terão uma variabilidade muito alta: $\alpha \sim G(0,001;0,001)$, $\lambda \sim G(0,001;0,001)$, $b_0 \sim N(1;1000)$ e $b_1 \sim N(1;1000)$. Geramos, através de uma programação realizada no sistema R (R Development Core Team, 2009), uma cadeia de 100000 elementos para cada parâmetro. Descartamos as primeiras 5000 e as restantes foram tomadas de 20 em 20, o que gerou uma amostra de tamanho 4750.

A Tabela 5.1 apresenta as médias a posteriori, desvios-padrão, intervalos de 95% de credibilidade e os valores da estatística de Geweke para cada um dos quatro parâmetros do Modelo 1.

Tabela 5.1: Resultados das distribuições a posteriori para o Modelo 1.

Parâmetro	Média a posteriori	Desvio padrão	Intervalo de credibilidade	Estatística Z
α	$1,528$	$0,151$	$(1,196; 1,799)$	$-0,569$
λ	$0,421$	$0,076$	$(0,265; 0,554)$	$-0,431$
b_0	$-1,080$	$0,283$	$(-1,636; -0,521)$	$0,052$
b_1	$0,487$	$0,132$	$(0,264; 0,778)$	$-0,047$

Observamos que os valores obtidos pela metodologia bayesiana (média e desvio-padrão) são bastante similares aos do método clássico, mostrados na Seção 4, e que os valores da estatística Z de Geweke ficaram próximos de zero, e isso é um indício de convergência da cadeia gerada.

A Figura 5.1 exibe as densidades marginais a posteriori para todos parâmetros do modelo:

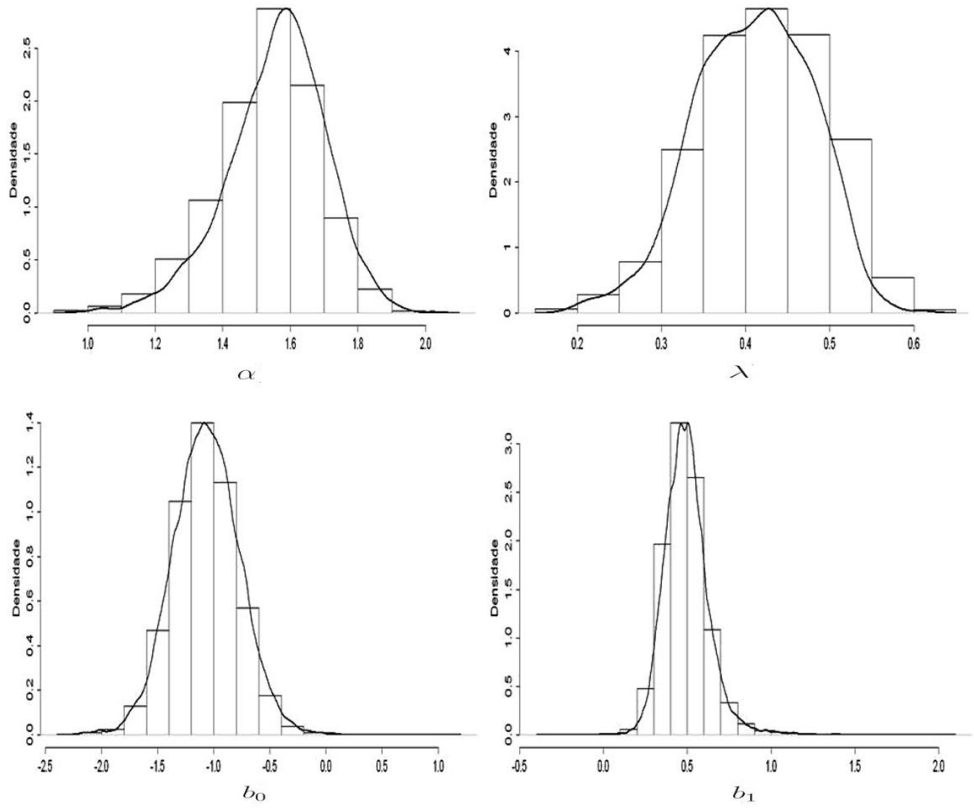

Figura 5.1: Densidades marginais a posteriori dos parâmetros - Modelo 1.

5.3 Distribuições a posteriori para o Modelo 2

Considerando o Modelo 2, mostrado em (3.2) e (3.3), sua priori conjunta, dada a suposição de independência dos parâmetros será:

$$\pi(\alpha, \lambda, \beta_1, b_0, b_1) = \pi(\alpha)\ \pi(\lambda)\ \pi(\beta_1)\ \pi(b_0)\ \pi(b_1), \tag{5.3}$$

com a seguinte função de verossimilhança, considerando a função de risco base seguindo uma distribuição Weibull e o vetor de parâmetros $\boldsymbol{\xi} = (\alpha, \lambda, b_0, b_1 \text{ e } \beta_1)$:

$$L(\boldsymbol{\xi}|\mathbf{D}) = \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i} \times$$

$$\left[\theta(\mathbf{z}_i)\alpha\lambda^{\alpha}t_i^{\alpha-1} \exp(\mathbf{x}_i'\boldsymbol{\beta}) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{\delta_i} \tag{5.4}$$

e combinando (5.3) e (5.4) temos a densidade a posteriori:

$$\pi(\alpha, \lambda, \beta_1, b_0, b_1|\mathbf{D}) \propto \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i) + \theta(\mathbf{z}_i) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i} \times$$

$$\left[\theta(\mathbf{z}_i)\alpha\lambda^{\alpha}t_i^{\alpha-1} \exp(\mathbf{x}_i'\boldsymbol{\beta}) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{\delta_i} \times$$

$$\alpha^{a-1} \exp(-b\alpha)\ \lambda^{c-1} \exp(-d\lambda)\ \ \exp\left(\frac{-(b_0-g)^2}{2h}\right)\ \ \exp\left(\frac{-(b_1-k)^2}{2m}\right)\ \ \exp\left(\frac{-(\beta_1-p)^2}{2q}\right)$$

Desta forma, as distribuições condicionais completas a posteriori de cada parâmetro são:

$$\pi(\alpha|\lambda, \beta_1, b_0, b_1, \mathbf{D}) \propto \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i)) + \theta(\mathbf{z}_i)) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i} \times$$

$$\left[\alpha\lambda^{\alpha}t_i^{\alpha-1} \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{\delta_i}\ \ \alpha^{a-1} \exp(-b\alpha)$$

$$\pi(\lambda|\alpha, \beta_1, b_0, b_1, \mathbf{D}) \propto \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i)) + \theta(\mathbf{z}_i)) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i} \times$$

$$\left[\lambda^{\alpha}t_i^{\alpha-1} \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{\delta_i}\ \ \lambda^{c-1} \exp(-d\lambda)$$

$$\pi(b_0|\alpha, \lambda, \beta_1, b_1, \mathbf{D}) \propto \prod_{i=1}^{n} \left[1 - \theta(\mathbf{z}_i)) + \theta(\mathbf{z}_i)) \exp\left[-(t_i\lambda)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i} [\theta(\mathbf{z}_i))]^{\delta_i} \times$$

$$\exp\left(\frac{-(b_0-g)^2}{2h}\right)$$

$$\pi(b_1|\alpha,\lambda,\beta_1,b_0,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i))+\theta(\mathbf{z}_i))\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i}\left[\theta(\mathbf{z}_i))\right]^{\delta_i}\times$$

$$\exp\left(\frac{-(b_0-k)^2}{2m}\right)$$

$$\pi(\beta_1|\alpha,\lambda,b_0,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i))+\theta(\mathbf{z}_i))\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]^{\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]^{1-\delta_i}\times$$

$$\left[\exp(\mathbf{x}_i'\boldsymbol{\beta})\exp\left[-\left(t_i\lambda\right)^{\alpha}\right]^{\exp(\mathbf{x}'\boldsymbol{\beta})}\right]^{\delta_i}\ \exp\left(\frac{-(\beta_1-p)^2}{2q}\right)$$

Novamente as densidades condicionais a posteriori não apresentaram nenhuma forma conhecida, assim para gerar valores para os quatro parâmetros de interesse novamente recorreremos aos algoritmo de Metropolis-Hastings.

A exemplo do que fizemos no Modelo 1, consideramos as mesmas distribuições a priori independentes, com o parâmetro adicional $\beta_1 \sim N(1;1000)$ e também geramos uma cadeia de 100000 elementos para cada parâmetro. Descartamos as primeiras 5000 e as restantes foram tomadas de 20 em 20, chegando novamente a uma amostra de tamanho 4750.

A Tabela 5.2 apresenta as médias a posteriori, desvios-padrão, intervalos de 95% de credibilidade e valores da estatística de convergência de Geweke para cada um dos cinco parâmetros do Modelo 2.

Tabela 5.2: Resultados das distribuições a posteriori para o Modelo 2.

Parâmetro	Média a posteriori	Desvio padrão	Intervalo de credibilidade	Estatística Z
α	$1,608$	$0,103$	$(1,408;1,808)$	$-0,224$
λ	$0,344$	$0,053$	$(0,246;0,452)$	$-0,097$
β_1	$0,173$	$0,082$	$(0,011;0,337)$	$0,145$
b_0	$-1,002$	$0,288$	$(-1,555;-0,420)$	$-0,571$
b_1	$0,412$	$0,107$	$(0,198;0,625)$	$0,526$

Observamos novamente que os valores obtidos pela metodologia bayesiana ficaram próximos aos do método clássico, mostrados na Seção 4, e que os valores da estatística de Geweke ficaram dentro do limite em que não se pode descartar a convergência da cadeia.

A Figura 5.2 exibe as densidades marginais a posteriori para todos parâmetros do Modelo 2. Observamos que alguns parâmetros obtiveram uma melhor

convergência do que outros, e isso se deve ao fato de termos aumentado o número de parâmetros a serem estimados no processo iterativo:

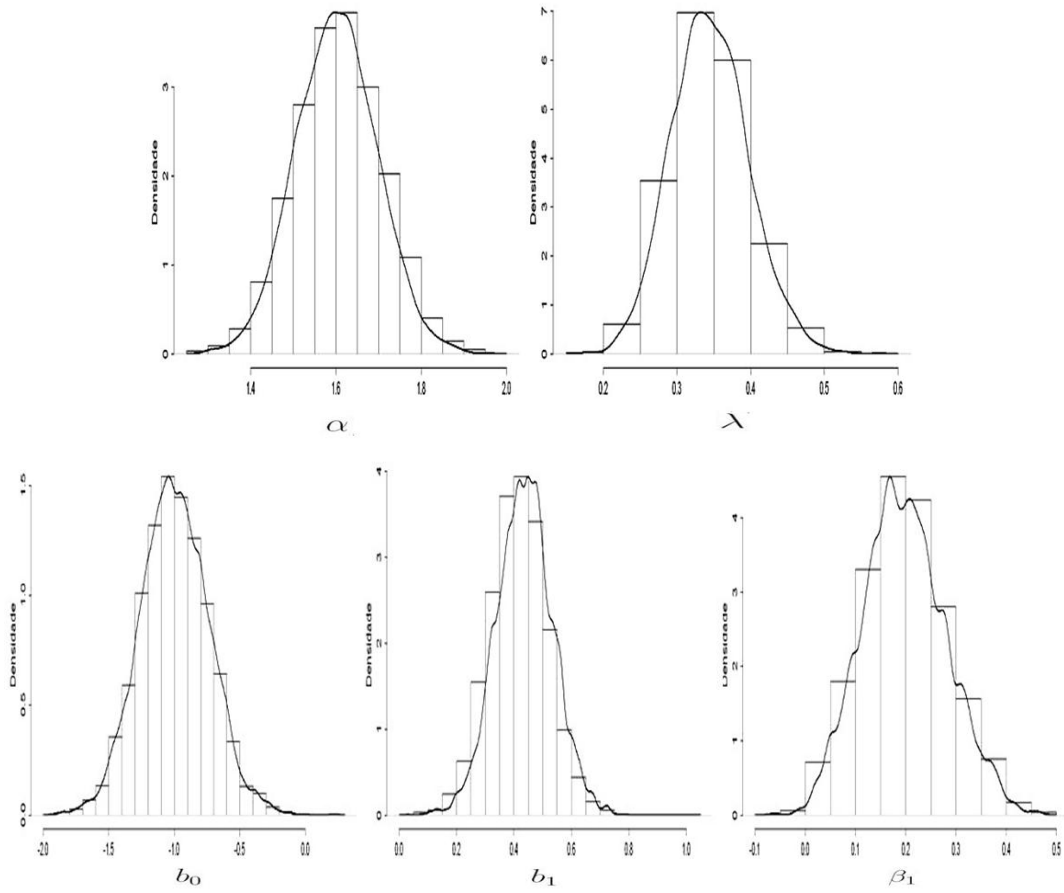

Figura 5.2: Densidades marginais a posteriori dos parâmetros - Modelo 2.

5.4 Distribuições a posteriori para o Modelo 3

Para o Modelo 3, temos a inclusão do termos de fragilidade, com seu parâmetro σ Assim, sua densidade priori conjunta, dada a suposição que os parâmetros de interesse são independentes, será:

$$\pi(\alpha, \lambda, \sigma, \beta_1, b_0, b_1) = \pi(\alpha)\ \pi(\lambda)\ \pi(\sigma)\ \pi(\beta_1)\ \pi(b_0)\ \pi(b_1), \tag{5.5}$$

Considerando a S_{pop} e a f_{pop} dadas respectivamente por (3.6) e (3.7), o vetor de parâmetros $\boldsymbol{\xi} = (\alpha, \lambda, \sigma, b_0, b_1$ e $\beta_1)$, a função de risco base Weibull, e a função de verossimilhança dada em (3.9), obtemos a seguinte posteriori:

$$\pi(\alpha,\lambda,\sigma,\beta_1,b_0,b_1|\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\exp(\mathbf{x}_i'\boldsymbol{\beta})\ \alpha\lambda^{\alpha}t_i^{\alpha-1}\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1}\right]^{\delta_i} \alpha^{a-1}\exp(-b\alpha)\ \lambda^{c-1}\exp(-d\lambda)\times$$

$$\sigma^{r-1}\exp(-s\sigma)\ \ \exp\left(\frac{-(b_0-g)^2}{2h}\right)\ \ \exp\left(\frac{-(b_1-k)^2}{2m}\right)\ \ \exp\left(\frac{-(\beta_1-p)^2}{2q}\right).$$

O conjunto das distribuições condicionais completas a posteriori para o Modelo 3 será dado por:

$$\pi(\alpha|\lambda,\sigma,\beta_1,b_0,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\alpha\lambda^{\alpha}t_i^{\alpha-1}\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1}\right]^{\delta_i} \alpha^{a-1}\exp(-b\alpha)$$

$$\pi(\lambda|\alpha,\sigma,\beta_1,b_0,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\lambda^{\alpha}t^{\alpha-1}\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1}\right]^{\delta_i} \lambda^{c-1}\exp(-d\lambda)$$

$$\pi(\sigma,|\alpha,\beta_1,b_0,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1}\right]^{\delta_i} \sigma^{r-1}\exp(-s\sigma)$$

$$\pi(b_0|\alpha,\lambda,\sigma,\beta_1,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$[\theta(\mathbf{z}_i)]^{\delta_i}\ \exp\left(\frac{-(b_0-g)^2}{2h}\right)$$

$$\pi(b_1|\alpha,\lambda,\sigma,\beta_1,b_0,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^{\alpha}\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$[\theta(\mathbf{z}_i)]^{\delta_i}\ \exp\left(\frac{-(b_0-k)^2}{2m}\right)$$

$$\pi(\beta_1|\alpha,\lambda,\sigma,b_0,b_1,\mathbf{D}) \propto \prod_{i=1}^{n}\left[1-\theta(\mathbf{z}_i)+\theta(\mathbf{z}_i)\left(1+\frac{(t_i\lambda)^\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma}\right]^{1-\delta_i} \times$$

$$\left[\exp(\mathbf{x}_i'\boldsymbol{\beta})\ \left(1+\frac{(t_i\lambda)^\alpha\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sigma}\right)^{-\sigma-1}\right]^{\delta_i}\ \exp\left(\frac{-(\beta_1-p)^2}{2q}\right)$$

E, a exemplo dos Modelos 1 e 2, as densidades condicionais a posteriori não apresentaram nenhuma forma conhecida, o que sugere a utilização do algoritmo de Metropolis-Hastings.

Como a convergência, considerando as mesmas prioris não informativas dos Modelos 1 e 2 e para o parâmetro de fragilidade $\sigma \sim G(1;0,001)$, não foi obtida, optamos por reduzir a variabilidade dos três parâmetros que não estavam convergindo, que foram b_0, b_1 e σ. Neste contexto, assumimos $\sigma \sim G(1;1)$, $b_0 \sim N(0;1)$ e $b_1 \sim N(0;1)$, mudança essa que levou à convergência da cadeia.

A Tabela 5.3 apresenta as médias a posteriori, desvios-padrão, intervalos de 95% de credibilidade e valores da estatística de convergência de Geweke para cada um dos cinco parâmetros do Modelo 3.

Tabela 5.3: Resultados das distribuições a posteriori para o Modelo 3.

Parâmetro	Média a posteriori	Desvio padrão	Intervalo de credibilidade	Estatística Z
α	$2,385$	$0,294$	$(1,868;3,030)$	$0,442$
λ	$0,312$	$0,055$	$(0,215;0,430)$	$-0,343$
σ	$0,767$	$0,428$	$(0,235;1,926)$	$-0,580$
β_1	$0,570$	$0,168$	$(0,274;0,931)$	$0,820$
b_0	$-0,468$	$0,369$	$(-1,108;0,340)$	$-0,319$
b_1	$0,352$	$0,129$	$(0,119;0,627)$	$0,232$

Observamos novamente que os valores obtidos pela metodologia bayesiana ficaram próximos aos do método clássico, mostrados na Seção 4, e que os valores da estatística de Geweke ficaram dentro do limite em que podemos não rejeitar a convergência da cadeia. Embora o intercepto da regressão logística, b_0, neste caso pareça não ser significativo, a covariável permanece sendo relevante para explicar a proporção de curados, dado que o coeficiente linear, b_1, mostrou um desvio-padrão baixo, ao contrário do que ocorreu com b_0.

Também observamos que o valor estimado da variância da variável de fragilidade, foi de $Var(\sigma) = 1/\hat{\sigma} = 1/0,77 = 1,30$. Como definimos esta variável de modo que ela tenha média 1, podemos afirmar que devemos considerar no modelo um componente que represente a variabilidade não explicada pela covariável.

A Figura 5.3 exibe as densidades marginais a posteriori para todos parâmetros do Modelo 3:

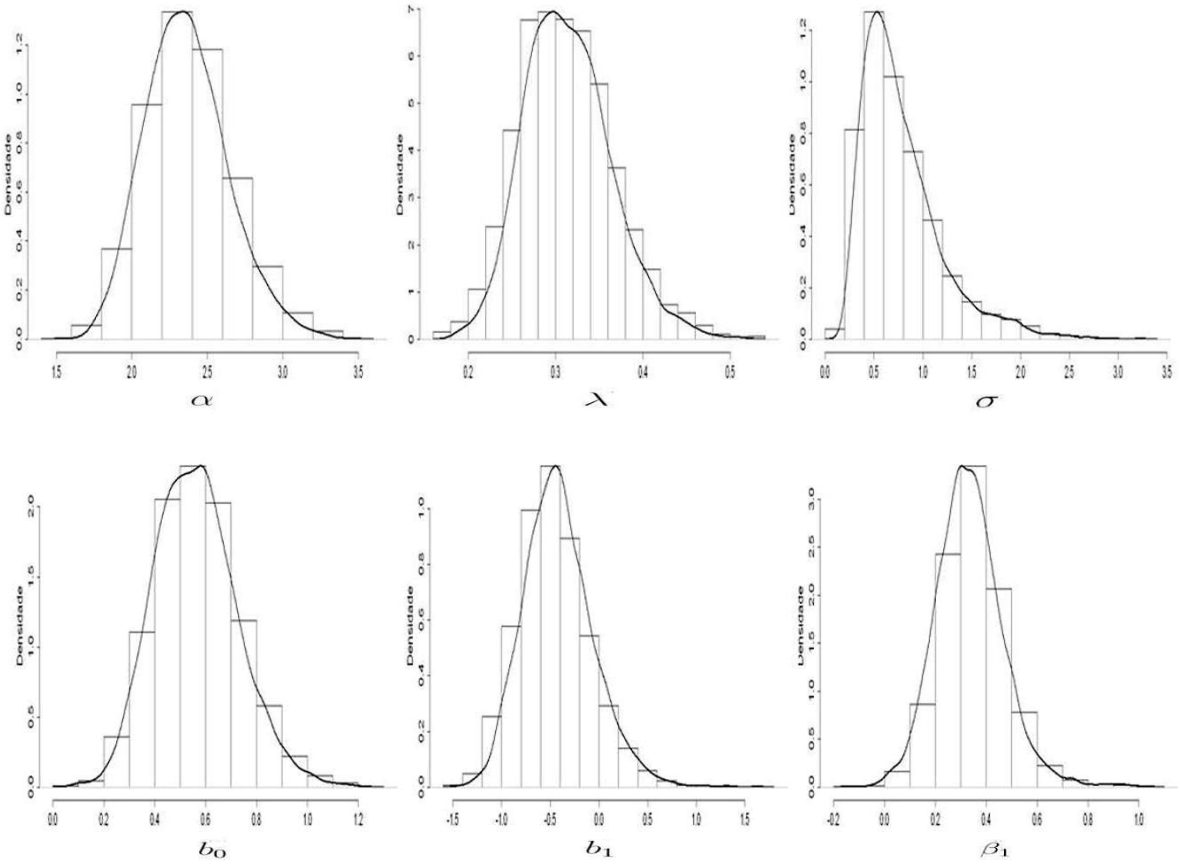

Figura 5.3: Densidades marginais a posteriori dos parâmetros - Modelo 3.

Finalmente, calculamos para cada modelo a proporção de curados, através das estimativas de b_0 e b_1 aplicadas à função de ligação logito. Os resultados são exibidos na Tabela 5.4.

Tabela 5.4: Estimativas das proporções de cura pelo método bayesiano.

	Modelo 1	Modelo 2	Modelo 3
Nível 1	0.644	0.643	0.529
Nível 2	0.526	0.544	0.441
Nível 3	0.406	0.442	0.357
Nível 4	0.296	0.344	0.281
Total	0.487	0.510	0.416

Percebemos que os resultados obtidos foram razoavelmente próximos aos da metodologia clássica, e também que ao considerar o componente de fragilidade, através do Modelo 3, a proporção de cura foi de aproximadamente 42%. Sendo assim, dos 56% originalmente censurados, 42% estariam imunes, e consequentemente não falhariam mesmo em um período maior de observação, fato que não seria contemplado se utilizássemos a análise de sobrevivência tradicional.

5.5 Considerações finais

Neste capítulo utilizamos os dados do melanoma, e estimamos novamente os parâmetros dos três modelos propostos no Capítulo 4, só que desta vez dentro de uma abordagem de estimação bayesiana. Após definirmos as distribuições a priori para cada um dos parâmetros, notamos que as distribuições condicionais a posteriori não apresentaram nenhuma forma conhecida.

Diante disso, utilizamos um método MCMC, no caso o algoritmo de Metropolis Hastings, para a geração dos dados de cada um dos parâmetros, que nos forneceram estatísticas a posteriori satisfatórias, pois os estimaram de forma bastante similar ao enfoque clássico, detectando significância da covariável em ambos os componentes do modelo de mistura, bem como do termo de fragilidade. Apenas para o Modelo 3 foi necessária uma redução da variabilidade das distribuições a priori de alguns parâmetros a fim de obtermos a convergência da cadeia

gerada. Também calculamos a proporção de cura para cada um dos três modelos, onde os resultados também foram similares aos obtidos via metodologia clássica.

Capítulo 6

Conclusões e Propostas Futuras

Neste trabalho apresentamos brevemente os principais conceitos de análise de sobrevivência bem como algumas distribuições comumente utilizadas nessa área. Vimos o que caracteriza o modelo de longa duração em análise de sobrevivência, no qual apresentamos o modelo de mistura padrão e como ele é construído a partir de uma variável binária que classifica os indivíduos de uma população em curados e não curados. Como alternativa para modelar dados de sobrevivência com fração de cura, um termo de fragilidade foi adicionado ao modelo de mistura padrão com o objetivo de quantificar a heterogeneidade não observável entre os indivíduos na população.

A principal contribuição foi apresentar o modelo de mistura padrão de Berkson & Gage (1952) com fragilidade na presença de covariáveis. Motivado pela flexibilidade da distribuição Weibull em acomodar diversas formas para a taxa de falha, consideramos que os indivíduos em risco são modelados por essa distribuição e propomos um método de estimação totalmente paramétrica. Para esta finalidade, foi empregado o método de estimação de máxima verossimilhança.

Como ilustração, aplicamos o modelo apresentado em um conjunto de dados reais de pacientes com Melanoma, e ajustamos esses dados a três modelos, o Modelo 1 considerando covariáveis apenas na fração de cura, o Modelo 2 considerando covariáveis na fração de cura e na função de sobrevivência dos não curados, e o Modelo 3, com fragilidade, fração de cura e covariáveis na fração de

cura e na função de sobrevivência dos não curados.

Observamos que o modelo de fragilidade com fração de cura na presença de covariáveis, o Modelo 3, é interessante para esse conjunto de dados. De forma geral os modelos de fragilidades são interessantes para resolver questões de heterogeneidade não observada nos estudos em análise de sobrevivência e o modelo de fragilidade com fração de cura estudado aqui pode conduzir a vários trabalhos futuros.

Na sequência estendemos este trabalho, realizando estudos de simulação, onde comprovamos as propriedades assintóticas dos estimadores dos parâmetros, e apresentamos uma metodologia bayesiana para a estimativa dos parâmetros dos modelos estudados, e a aplicamos na mesma base de dados estudada na aplicação, e notamos que tal metodologia apresentou resultados bastante próximos dos obtidos via máxima verossimilhança.

Como propostas futuras estão a escolha de outras distribuições de probabilidade para o componente de fragilidade e para o risco da parcela não imune.

Também pode ser de grande utilidade a determinação do custo de estimação da fração de cura com a inclusão da fragilidade ao modelo de mistura padrão, e também a utilização do modelo de partição produto (Denison *et al.*, 2002) como ferramenta de seleção de covariáveis.

Referências Bibliográficas

Aalen, O. (1989). A linear regression model for the analysis of lifetimes. *Statistics in Medicine*, **8**, 907–925.

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

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

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, **11**(1), 15–53.

Box, G. & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, **26**, 211–252.

Calsavara, V. (2011). *Modelos de sobrevivência com fração de cura usando um termo de fragilidade e tempo de vida Weibull modificada generalizada*. Master's thesis, Universidade Federal de São Carlos.

Chen, M.-H., Ibrahim, J. G. & Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, **94**(447), 909–919.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemological studies in familial tendency in chronical disease incidency. *Biometrik*, **65**, 141–151.

Colosimo, E. & Giolo, S. (2006). *Análise de Sobrevivência Aplicada*. Edgard Blucher, São Paulo, SP.

Cox (1972). Regression models and lifetables. *Journal Royal Statistical Society*, **34**, 187–220.

Denham, J. W., Denham, E. E., Dear, K. B. G., & Hudson, G. V. (1996). The follicular non-hodgkin's lymphomas-i: the possibility of cure. *The European Journal of Cancer*, **32**, 470–479.

Denison, D. G. T., Adams, N. M., Holmes, C. C. & Hand, D. J. (2002). Bayesian partition modelling. *Computational Statistics & Data Analysis*, **38**(4), 475–485.

Elbers, C. & Ridder, G. (1982). True and spurious duration dependence: the identifiability of the proportional hazard model. *The review of economic studies*, **49(3)**, 403–409.

Farewell, V. & Sprott, D. (1986). Mixture models in survival analy- sis: Are they worth the risk? *The Canadian Journal of Statistics*, **14**, 257–262.

Farewell, V. T. (1977). A model for binary variable with time-censored observations. *Biometrika*, **38**, 43–46.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, **38**(2), 1041–1046.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, **4(2)**, 169–193.

Goldman, A. (1984). Survivorship analysis when cure is a possibility: a monte carlo study. *Statistics in medicine*, **3**, 153–163.

Greenhouse, j. & Wolfe, R. (1984). A competing risks derivation of a mixture model for the analysis of survival data. *Communication in Statistics Theory and Methods*, **13**, 3133–3154.

Greenwood, M. & Yule, G. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal Statistical Society*, **83**, 255–279.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **59**, 97–109.

Hougaard, P. (1991). Modelling heterogeneity in survival data . *Journal of Applied Probability*, (1).

Hougaard, P., Myglegaard, P. & Borch-Johnsen, K. (1994). Heterogeneity models for disease susceptibility with application to diabetic nephorpathy. *Biometrics*, **50**, 1178–1188.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**(421), 457–481.

Kirkwood, J. M., Ibrahim, J., Sondak, V., Richards, J., Flaherty, L., Ernstoff, M., Smith, T., Rao, U., Steele, M. & Blum, H. (2000). High- and low-dose interferon alfa-2b in high-risk melanoma: First analysis of intergroup trial E1690/S9111/C9190. *Journal of Clinical Oncology*, **18**(12), 2444–2458.

Kuk, A. & Chen, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, **79**, 531–541.

Laska, E. & Meisner, M. (1992). Nonparametric estimation and testing in a cure model. *Biometrics*, **48**, 1223–1334.

Lawless, J. (2002). *Statiscal Models and Methods for Lifetime Data*. Wiley, New York,NY, second edition.

Leng, O. Y. & Khalid, Z. M. (2010). A comparative study of maximum likelihood and bayesian estimation approaches in estimating frailty mixture survival model parameters. *Proceedings of the 6th IMT-GT Conference on Mathematics, Statistics and its Applications*, **1**, ,478–492.

Li, C., Taylor, J. & Sy, J. (2001). Identifiability of cure models. *Statististics and Probability Letters*, **54**, 389–395.

Longini, I. & Halloran, M. (1996). A frailty mixture model for estimating vaccine efficacy. *Applied S*, **45**, 165–173.

Mailer, R. A. & Zhou, S. (1992). Estimating the proportion of immunes in a censored sample. *Biometrika*, **79**, 731– 739.

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

Oakes, D. (1982). A mode for association in bivariate survival data. *Journal of the Royal Statistical Society*, **44**.

Ortega, E. M., Cancho, V. G. & Lachos, V. H. (2008). Assessing influence in survival data with a cure fraction and covariates. *SORT*, **32(2)**, 115–140.

Peng, Y. & Dear, K. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, **56**(1), 237 – 243.

Peng, Y. & Zhang, J. (2008a). Estimation method of the semiparametric mixture cure gamma frailty mod. *Statistics i*, **27**, 5177–5194.

Peng, Y. & Zhang, J. (2008b). Identifiability of a mixture cure fraitly model. *Statistics and Probability Letters*, **78**, 2604–2608.

Peng, Y., Dear, K. & Denham, J. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, **17**(425), 813–830.

Price, D. & Manatunga, A. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in M*, **20**, 1515–1527.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rocha, C. (1995). *Modelos com fragilidade em análise de sobrevivência*. Master's thesis, Universidade de Lisboa.

Rodrigues, J., Cancho, V. & de Castro, M. (2008). *Teoria Unificada de Análise de Sobrevivência*. Associação Brasileira de Estatística, Sï¿½o Paulo,SP.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Tsodikov, A. (1998). Asymptotic efficiency of a proportional hazards model with cure. *Statistics & Probabillty Letters*, **39**, 237–244.

Vaupel, J., Manton, K. & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16(3)**, 439–454.

Wienke, A. (2011). *Frailty Models in Survival Analysis*. Chapman & Hall.

Yakovlev, A. & Tsodikov, A. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. World Scientific, Singapore.

Yakovlev, A., Asselain, B., Bardou, V., Fourquet, A., Hoang, T., Rochefedière, A. & Tso (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biométrie*, **12**, 66–82.

Yin, G. (2005). Bayesian cure rate frailty models with application to a root canal therapy study. *Biometrics*, **61**, 552–558.

Yin, G. & Ibrahim, J. G. (2005). Cure rate models: a Unified approach. *The Canadian Journal of Statistics*, **33**(4), 559–570.

Yu, B. & Tiwari, R. C. (2012). A bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data. *Canadian Journal of Statistics*, **40**, 40–54.