Estimação e diagnóstico na distribuição exponencial por partes em análise de sobrevivência com fração de cura

Alessandra Cristiane Sibim

Estimação e diagnóstico na distribuição exponencial por partes em análise de sobrevivência com fração de cura

Alessandra Cristiane Sibim

Orientador: ***Prof. Dr. Vicente Garibay Cancho***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA.*

USP – São Carlos
Maio/2011

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

S563e

Sibim, Alessandra Cristiane
Estimação e diagnóstico na distribuição exponencial por partes em análise de sobrevivência com fração de cura / Alessandra Cristiane Sibim; orientador Vicente Garibay Cancho -- São Carlos, 2011.
63 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

1. Análise de sobrevivência. 2. Inferência bayesiana. 3. Divergência de Kullback-Leibler. 4. Métodos MCMC. I. Cancho, Vicente Garibay, orient. II. Título.

Aos meus pais e irmãos.

Agradecimentos

Primeiramente agradeço a Deus, que me dá saúde e força para superar os obstáculos todos os dias.

Aos meus pais que sempre me ajudam, apoiam e me incentivam na árdua tarefa de estudar e a quem sempre estarei em dívida.

Ao meu orientador, Vicente Garibay Cancho, quero agradecer por ter acreditado em mim e por estar presente em cada etapa deste trabalho. Agradeço ainda pela paciência, incentivo na elaboração e condução do trabalho, portanto tem minha eterna gratidão e admiração.

Aos professores Francisco Louzada Neto e Mário de Castro, membros da banca do exame de qualificação, pelas sugestões feitas.

Aos professores e funcionários da biblioteca Prof. Achille Bassi e da pós-graduação do ICMC pelo excelente convívio.

Aos professores Carlos Aparecido dos Santos e Jacinta Ludovico Zamboti pelos primeiros ensinamentos estatísticos e pelo incentivo que me levaram a fazer a pós-graduação.

Em especial ao Wesley pelo apoio e incentivo na realização deste sonho, a Maria Inês pela grande ajuda que viabilizou a obtenção de resultados e a elaboração deste texto, a Conceição pelas doces palavras em um momento difícil.

A todos os meus amigos do curso de pós graduação Adriana, Aline, Gilberto, Marcia e William pelos momentos compartilhados no decorrer destes dois anos.

Aos demais amigos que, independentemente da distância, acompanharam-me nesta jornada. Não citarei nomes, pois a lista seria muito extensa para ser colocada aqui. Além disso, não correrei o risco de acabar esquecendo de alguém.

A todas as pessoas que não foram nominalmente mencionadas, mas que de alguma forma contribuíram para viabilizar este trabalho.

Finalmente, agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo suporte financeiro concedido para a realização deste trabalho.

"Feliz aquele que transfere o que sabe e aprende o que ensina".

"O saber se aprende com os mestres. A sabedoria, só com o corriqueiro da vida".

Cora Coralina

Resumo

O principal objetivo deste trabalho é desenvolver procedimentos inferências em uma perspectiva bayesiana para modelos de sobrevivência com (ou sem) fração de cura baseada na distribuição exponencial por partes. A metodologia bayesiana é baseada em métodos de Monte Carlo via Cadeias de Markov (MCMC). Para detectar observações influentes nos modelos considerados foi usado o método bayesiano de análise de influência caso a caso (Cho *et al.*, 2009), baseados na divergência de Kullback-Leibler. Além disso, propomos o modelo destrutivo binomial negativo com fração de cura. O modelo proposto é mais geral que os modelos de sobrevivência com fração de cura, já que permitem estimar a probabilidade do número de causas que não foram eliminadas por um tratamento inicial.
Palavras-chave: Inferência Bayesiana, Medidas de Diagnóstico Bayesiano, Divergência de Kullback-Leibler, Métodos MCMC, Análise de Sobrevivência.

Abstract

The main objective is to develop procedures inferences in a bayesian perspective for survival models with (or without) the cure rate based on piecewise exponential distribution. The methodology is based on bayesian methods for Markov Chain Monte Carlo (MCMC). To detect influential observations in the models considering bayesian case deletion influence diagnostics based on the Kullback-Leibler divergence (Cho et al., 2009). Furthermore, we propose the negative binomial model destructive cure rate. The proposed model is more general than the survival models with cure rate, since the probability to estimate the number of cases which were not eliminated by an initial treatment.

Keywords: Bayesian Inference, Measures of Diagnostic Bayesian, Kullback-Leibler Divergence, MCMC Methods, Survival Analysis

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Análise de Sobrevivência é uma das áreas da Estatística que mais cresceu nos últimos 20 anos. Este crescimento se justifica pelo desenvolvimento e aprimoramento de técnicas estatísticas combinados com avanços computacionais (Colosimo & Giolo, 2006).

Análise de sobrevivência é o termo utilizado para designar a análise estatística de dados quando a variável em estudo representa o tempo a partir um instante inicial bem definido até à ocorrência de determinado evento de interesse denominado de falha.

É comum em análise de sobrevivência a presença de observações parciais ou incompletas, ou seja, observações que não apresentam o evento de interesse. Essas observações são denominadas observações censuradas e podem ocorrer por uma variedade de razões:

(i) A perda do acompanhamento do paciente no decorrer do estudo;

(ii) A não ocorrência do evento de interesse até o término do experimento;

(iii) O término do experimento antes da ocorrência do evento de interesse.

Mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser usados na análise estatística, pois sua omissão pode acarretar conclusões viciadas e mesmo sendo incompletas, as observações censuradas fornecem informação sobre o tempo de vida dos indivíduos em estudo. As censuras são classificadas em censura do tipo I,

censura do tipo II e censura do tipo aleatório. Mais detalhes podem ser encontrados em Colosimo & Giolo (2006).

Ao se analisar dados de sobrevivência percebemos que muitas vezes há grande número de indivíduos censurados à direita. Empiricamente, esta característica é observada na estimativa da função de sobrevivência de Kaplan-Meier, que apresenta uma cauda direita em um nível aproximadamente constante e estritamente maior que 0 por um período considerável. Fato este que pode ser uma indicação de que em parte dos indivíduos em estudo o evento de interesse nunca ocorra, ou seja, uma fração da população apresenta imunidade. Isso ocorre, por exemplo em muitos estudos clínicos, especialmente em estudos de câncer, nos quais há uma fração significativa de pacientes que respondem favoravelmente ao tratamento e apresentam uma proporção de indivíduos sobreviventes ou curados, mesmo após um longo período de acompanhamento. Para tais dados de sobrevivência, uma proporção de indivíduos da população é suscetível ao evento de interesse e outros não o são. Nesta situação os modelos de sobrevivência usuais, que assumem que a função de sobrevivência converge para 0 quando o tempo de sobrevivência é suficientemente grande (função de sobrevivência própria), podem não ser adequados.

Uma abordagem bastante popular para modelar dados de sobrevivência com fração de cura (também chamados de modelos de sobrevivência de longa duração) é o modelo de mistura proposto por Boag (1949), posteriormente desenvolvido por Berkson & Gage (1952), e que foi amplamente estudado por vários autores (vide, Maller & Zhou, 1996).

A literatura sobre modelos de sobrevivência com fração de cura é extensa e está em rápido desenvolvimento. Podemos destacar como referências fundamentais os livros de Maller & Zhou (1996) e Ibrahim *et al.* (2001), como também o artigo de Tsodikov *et al.* (2003) e o artigo de Cooner *et al.* (2007). Fora do contexto das aplicações a dados de estudos clínicos, Yamaguchi (1992) considera o modelo de mistura em estudos de desemprego, ao passo que Hoggart & Griffin (2001) desenvolveram um modelo para o tempo até que um indivíduo deixe de ser cliente de um banco. Zaider *et al.* (2001) estudam modelos de longa duração adotando como base conceitos de processos estocásticos. Chen & Ibrahim

(2001) discutem estimação por máxima verossimilhança em um modelo semiparamétrico. O modelo bayesiano de Chi & Ibrahim (2007) estabelece uma flexibilidade na fração de cura, podendo esta ser nula ou positiva. Kim *et al.* (2007) propuseram um modelo semiparamétrico dinâmico bayesiano. Mizoi & Bolfarine (2007) (vide ainda Mizoi, 2004) estenderam o modelo de tempo de promoção considerando situações em que há erros de medição nas covariáveis. Tournoud & Ecochard (2007) tratam de um modelo com fração de cura em situações em que a exposição a um fator de risco ocorre em diversas ocasiões. Paes (2007) e Fonseca (2009) estudam modelos de sobrevivência com fração de cura com omissão nas covariáveis. Sen & Tan (2008) discutem estimação não-paramétrica da fração de cura na presença de censura intervalar. Lopes (2008) apresenta dois modelos com fração de cura considerando a introdução de efeitos aleatórios, um baseado no modelo de mistura e outro no modelo de tempo de promoção. Rodrigues *et al.* (2009a) propõem um modelo de sobrevivência com fração de cura unificado. Rodrigues *et al.* (2009b) discutem o uso da distribuição Conway–Maxwell Poisson na estimação da proporção de curados. Ortega *et al.* (2009) apresentam um modelo de regressão log-gama generalizado reparametrizado para incluir indivíduos que apresentam fração de cura, estimando assim, os efeitos das covariáveis sobre os tempos e realizando uma análise de sensibilidade e de resíduos sob uma abordagem clássica. Cancho *et al.* (2009) realizam um estudo de diagnóstico sob a perspectiva clássica para um modelo de regressão log–Weibull exponenciada considerando a presença de indivíduos curados. de Castro *et al.* (2010) desenvolveram um aplicativo para estimação dos parâmetros de modelos de sobrevivência com fração de cura por meio de pacote gamlss em *R* (R Development Core Team, 2010). Cancho *et al.* (2011) consideram um modelo de sobrevivência com fração de cura, baseada na distribuição binomial negativa, abrangendo como um caso especial o modelo tempo de promoção.

Rodrigues *et al.* (2010b) propuseram modelo destrutivo com taxa de cura, considerando a distribuição Poisson ponderada, para modelar o número inicial de causas ou riscos relacionadas à ocorrência de um particular evento de interesse. Em Rodrigues *et al.* (2010a) é considerada uma abordagem bayesiana para o modelo destrutivo Poisson ponderado.

Muitas famílias de distribuição foram consideradas para os tempos de vida. Neste trabalho consideramos a distribuição exponencial por partes por ser um modelo bastante flexível.

O modelo exponencial por partes (MEP) é amplamente utilizado em análise de sobrevivência, para modelar dados onde o interesse é o tempo até a ocorrência de um determinado evento, podendo ser o tempo até a morte do paciente, até a cura ou ainda reincidência de uma doença, o que comumente ocorre em estudos clínicos sobre leucemia (Breslow, 1974), transplante de coração (Aitkin *et al.*, 1983), epidemia de AIDS (Brookmeyer & Goedert, 1989), mortalidade hospitalar (Clark & Ryan, 2002), entre outros. São vários os trabalhos existentes na literatura que discutem a aplicação do MEP, Friedman (1982) considerou-o para modelar a função de risco basal do modelo de Cox, apresentando condições para a existência e distribuição assintótica dos estimadores de máxima verossimilhança (EMV) para os coeficientes da regressão e taxas de falha. Kim & Proschan (1991) fazem um estudo comparativo entre um estimador para o MEP e o estimador de Kaplan–Meier discutindo vantagens e desvantagens de ambos considerando dados sem covariáveis. Gamerman (1994) em uma abordagem bayesiana dinâmica utiliza o MEP não paramétrico considerando uma relação estocástica entre os sucessivos intervalos do tempo, fazendo uma comparação entre as estimativas bayesianas e clássicas. Barbosa *et al.* (1996) descrevem o uso do modelo exponencial por partes em dados de tempos de vida acelerados utilizando a abordagem de modelos lineares generalizados. Reineke *et al.* (1999) comparam o desempenho de um estimador exponencial por partes com o estimador de Kaplan–Meier considerando dados censurados. Ibrahim *et al.* (2001) utilizam o método de máxima verossimilhança para o MEP considerando modelos de sobrevivência com fração de cura com dados faltantes, Chen *et al.* (2002) considerando modelos de sobrevivência com fração de cura para o estudo de câncer com dados clínicos, empregam o MEP, Yin (2005) propôs um modelo de sobrevivência com fração de cura baseado na transformação de Box–Cox considerando o modelo exponencial por partes, unificando o modelo de mistura padrão (Berkson & Gage, 1952) e o modelo de tempo de promoção (Yakovlev & Tsodikov, 1996). Schmidli *et al.* (2007) consideram o modelo em uma abordagem bayesiana com

priori informativas em um estudo de tratamento clínico. Chen *et al.* (2008) utilizam um modelo de regressão logística para o estudo de caso-controle de reincidência de câncer de mama e desenvolvem uma abordagem geral baseado na pseudo-verossimilhança para acomodar os dados com covariável faltante considerando o MEP.

1.1 Critérios de comparação de modelos

Uma questão importante consiste na avaliação e escolha do modelo que melhor represente a situação em estudo. Neste trabalho utilizamos dois critérios de seleção de modelos, os quais especificamente, são usados na metodologia bayesiana em que as amostras das distribuições a *posteriori* para os parâmetros do modelo são obtidos usando métodos MCMC. A seguir descrevemos cada um deles.

A densidade preditiva condicional ordenada (CPO) é uma ferramenta de avaliação do modelo muito útil e extensamente usada na literatura estatística sob vários contextos (Ibrahim *et al.*, 2001). Seja $\boldsymbol{\mathcal{D}}$ os dados completos e $\boldsymbol{\mathcal{D}}_{(-i)}$ os dados com a i-ésima observação excluída. Denotamos a densidade a *posteriori* de $\boldsymbol{\gamma}$ dado $\boldsymbol{\mathcal{D}}_{(-i)}$ por $\pi(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}}_{(-i)})$, $i = 1, \ldots, n$ e $\boldsymbol{\gamma}$ é um vetor dos parâmetros. Assim podemos escrever a CPO_i para a i-ésima observação como

$$CPO_i = \int_{\Theta} g(t_i|\boldsymbol{\gamma})\pi(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}}_{(-i)})d\boldsymbol{\gamma} = \left\{ \int_{\Theta} \frac{\pi(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}})}{g(t_i|\boldsymbol{\gamma})} d\boldsymbol{\gamma} \right\}^{-1}, \tag{1.1}$$

em que $g(t_i|\boldsymbol{\gamma})$ é a função densidade de probabilidade. Para valores altos de CPO_i temos um melhor ajuste do modelo. Uma estimativa de Monte Carlo para CPO_i considerando uma amostra de tamanho Q da distribuição a *posteriori* $\pi(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}})$,(Chen *et al.*, 2000) é dada por

$$\widehat{CPO_i} = \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{g(t_i|\boldsymbol{\gamma}_q)} \right\}^{-1}. \tag{1.2}$$

Como em Cancho *et al.* (2010) utilizamos a estatística $B = \sum_{i=1}^{n} \log(\widehat{CPO_i})$ na seleção dos modelos, maiores valores de B indicam o melhor modelo.

O outro critério utilizado neste trabalho é o "Deviance Information Criterion" (DIC) (Spiegelhalter *et al.*, 2002) o qual é baseado na média a *posteriori* da deviance, que pode ser aproximada por

$$\bar{D} = \frac{1}{Q}\sum_{q=1}^{Q} D(\boldsymbol{\gamma}_q) \qquad \text{com} \qquad D(\boldsymbol{\gamma}) = -2\sum_{i=1}^{n} \log\left[g(t_i|\boldsymbol{\gamma}_q)\right].$$

O critério DIC também pode ser aproximado considerando amostras MCMC, por $\widehat{DIC} = 2\bar{D} - \hat{D}$, no qual

$$\hat{D} = D\left(\frac{1}{Q}\sum_{q=1}^{Q} \boldsymbol{\gamma}_q\right).$$

Menores valores de DIC determinam os melhores modelos.

1.2 Diagnóstico

Uma maneira de avaliar as suposições feitas sobre o modelo e de detectar pontos influentes pode ser efetuada pela análise de diagnóstico, a qual iniciou-se com a análise de resíduos para detectar a presença de pontos extremos e avaliar a adequação da distribuição proposta para a variável resposta. Uma referência importante neste assunto é o artigo de Cox & Snell (1968) que apresenta uma forma bastante geral de definir resíduos, usada até os dias atuais (Paula, 2004). Uma das propostas mais inovadoras nesta área foi apresentada por Cook (1986), que propôs avaliar a influência conjunta das observações sob pequenas perturbações no modelo, ao invés da avaliação pela retirada individual ou conjunta de pontos. Se essas perturbações causam efeitos desproporcionais, pode ser indício de que o modelo está mal ajustado ou que podem existir afastamentos sérios das suposições feitas para o mesmo (Carrasco, 2007). Embora a metodologia proposta por Cook (1986) venha sendo aplicada com sucesso em diferentes áreas da estatística aplicada, observamos que dependendo da complexidade do modelo a aplicação da metodologia envolve extensas manipulações algébricas e em alguns casos um duro trabalho computacional. Vários autores têm aplicado as técnicas de influência local, Ortega *et al.* (2003) apresentam uma aplicação

a modelos de regressão log-gama generalizadas com observações censuradas. Labra *et al.* (2005) realizam um estudo de influência local em um modelo com erros de medição t de Student com intercepto nulo. Ortega *et al.* (2006) aplicaram um estudo de influência local para o modelo de regressão logística.

Peng & Dey (1995) apresentam duas distintas abordagens bayesianas para detectar observações influentes no ajuste de modelos de regressão, uma é baseada na distribuição a *posteriori* e a segunda baseada na distribuição preditiva. Quatro medidas específicas são propostas, entre elas, destacamos a divergência Kullback-Leibler (K-L). Recentemente, Cho *et al.* (2009), propuseram um método bayesiano de análise de influência caso a caso para dados de sobrevivência, baseado na divergência Kullback-Leibler, no qual desenvolveram medidas de diagnóstico para avaliar a influência de um caso nas distribuições a *posteriori* conjuntas e marginais fundamentadas na divergência Kullback-Leibler.

1.2.1 Análise de influência caso a caso

Uma maneira comum de avaliar a influência de uma observação no ajuste de um modelo é por meio da deleção de casos (Cook & Weisberg, 1982).

Suponha que $K(P, P_{(-i)})$ denota a divergência K-L entre P e $P_{(-i)}$, em que P denota a distribuição a *posteriori* de $\boldsymbol{\gamma}$ para os dados completos e $P_{(-i)}$ é a distribuição a *posteriori* de $\boldsymbol{\gamma}$ sem o i-ésimo caso. Especificamente,

$$K(P, P_{(-i)}) = \int \pi(\boldsymbol{\gamma}|\mathbf{D}) \log \left\{ \frac{\pi(\boldsymbol{\gamma}|\mathbf{D})}{\pi(\boldsymbol{\gamma}|\mathbf{D}_{(-i)})} \right\} d\boldsymbol{\gamma}, \tag{1.3}$$

$K(P, P_{(-i)})$ mede o efeito de omitir o i-ésimo caso dos dados completos na distribuição a *posteriori* de $\boldsymbol{\gamma}$. Note que $K(P, P_{(-i)}) \neq K(P_{(-i)}, P)$ em geral. Após alguma álgebra, podemos demonstrar uma expressão simplificada para $K(P, P_{(-i)})$ dada por,

$$K(P, P_{(-i)}) = \log E_{\boldsymbol{\gamma}} \left[\frac{L(\boldsymbol{\gamma}|\mathbf{D}_{(-i)})}{L(\boldsymbol{\gamma}|\mathbf{D})} \Big| \mathbf{D} \right] + E_{\boldsymbol{\gamma}} \left[\log \left\{ \frac{L(\boldsymbol{\gamma}|\mathbf{D})}{L(\boldsymbol{\gamma}|\mathbf{D}_{(-i)})} \right\} \Big| \mathbf{D} \right], \tag{1.4}$$

em que, $E[\cdot|\mathbf{D}]$ representa a média a *posteriori* de $\boldsymbol{\gamma}$, $L(\boldsymbol{\gamma}|\mathbf{D}) = \prod_{k=1}^{n} f(t_k|\boldsymbol{\gamma})$ a função

verossimilhança para os dados completos e $L(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}}_{(-i)}) = \prod_{k=1,k\neq i}^{n} f(t_k|\boldsymbol{\gamma})$ a função verossimilhança sem a i-ésima observação. A equação (1.4) pode ser reescrita como

$$\begin{aligned} K(P, P_{(-i)}) &= \log E_{\boldsymbol{\gamma}}[\{g(t_i|\boldsymbol{\gamma})\}^{-1}|\boldsymbol{\mathcal{D}}] + E_{\boldsymbol{\gamma}}[\log\{g(t_i|\boldsymbol{\gamma})\}|\boldsymbol{\mathcal{D}}], \\ &= -\log(CPO_i) + E_{\boldsymbol{\gamma}}[\log\{g(t_i|\boldsymbol{\gamma})\}|\boldsymbol{\mathcal{D}}]. \end{aligned} \tag{1.5}$$

Da Equação (1.5) uma estimativa de Monte Carlo para $K(P, P_{(-i)})$ considerando uma amostra de tamanho Q da distribuição a *posteriori* de $p(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}})$ é dada por

$$\widehat{K(P, P_{(-i)})} = -\log(\widehat{CPO_i}) + \frac{1}{Q}\sum_{q=1}^{Q}\log[g(t_i|\boldsymbol{\gamma}_q)], \tag{1.6}$$

sendo $(\widehat{CPO_i})$ como descrito na Equação (1.2).

Segundo McCulloch (1989) e Cho *et al.* (2009), a calibração de $K(P, P_{(-i)})$ pode ser efetuada resolvendo em p_i a equação

$$K(P, P_{(-i)}) = K(B(0,5), B(p_i)) = -\frac{\log\{4p_i(1-p_i)\}}{2}, \tag{1.7}$$

em que $B(p)$ denota uma distribuição de Bernoulli com probabilidade de sucesso p. Isto implica que descrever resultados usando $\pi(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}}_{(-i)})$ ao invés de $p(\boldsymbol{\gamma}|\boldsymbol{\mathcal{D}})$ é equivalente com a descrição de um evento não observado com probabilidade p_i quando a probabilidade correta é 0,5. Após o cálculo de $K(P, P_{(-i)})$ em (1.5) a solução em p_i da Equação (1.7) é calculada por $p_i = 0,5[1 + \sqrt{1 - \exp\{-2K(P, P_{(-i)})\}}]$. Isto implica que $0,5 \leq p_i \leq 1$. Além disso, se $p_i \gg 0,5$ implica que o i-ésimo caso é influente.

Neste trabalho analisamos modelos de sobrevivência sem e com fração de cura e ainda modelos destrutivos com fração de cura com o intuito de analisar qual proporciona um melhor ajuste. São propostas análise de influência caso a caso para investigar possíveis problemas com o modelo ajustado. O estudo é feito com enfoque bayesiano. São apresentados estudos de simulação e estudos com dados reais. A proposta deste trabalho é interessante uma vez que não encontramos na literatura análises de diagnósticos para os

modelos aqui discutidos.

1.3 Organização dos capítulos

No Capítulo 2 desenvolvemos métodos bayesianos via Monte Carlo em cadeias de Markov (MCMC) para um modelo de regressão exponencial por partes e realizamos um estudo de diagnóstico. No Capítulo 3 abordamos modelos de sobrevivência com fração de cura, fazemos uma análise de sensibilidade na estimativa dos parâmetros a *posteriori* para um conjunto de dados reais propiciando a escolha de um modelo final com melhor ajuste aos dados. No Capítulo 4 tratamos de um modelo destrutivo com fração de cura baseado na distribuição binomial negativa, neste modelo temos a inserção de um parâmetro (p) que permite estimar a probabilidade do número de causas não destruídas, também realizamos um estudo de diagnóstico considerando um conjunto de dados de melanoma. Finalmente, no Capítulo 5 apresentamos as conclusões com base nos resultados obtidos e as propostas futuras.

Capítulo 2

Modelo Exponencial por Partes

Segundo Ibrahim *et al.* (2001), o MEP é um dos modelos mais populares utilizados em análise de sobrevivência, a popularidade e importância se deve ao fato deste modelo ser capaz de acomodar funções de taxa de falha com diversas formas, não havendo a necessidade de impormos restrições quanto a forma da função de risco para obtermos um ajuste apropriado do modelo aos dados, como acontecem com alguns modelos como por exemplo, com o modelo exponencial e Weibull. Esta característica torna o modelo bastante flexível. Uma das dificuldades em se trabalhar com o modelo exponencial por partes está em definir a partição do eixo dos tempos a ser utilizada. Em geral, tal partição é escolhida arbitrariamente como discutido por Gamerman (1994), Barbosa *et al.* (1996), (Ibrahim *et al.*, 2001), através de algum critério de seleção de modelo (Yin, 2005) ou aleatoriamente (Demarqui *et al.*, 2008).

2.1 Formulação do modelo

Seja T uma variável aleatória não-negativa representando o tempo até a ocorrência de um evento de interesse, denominado tempo de falha. Considere uma partição finita e arbitrária do eixo dos tempos, tal que, $s_0 < s_1 < \ldots < s_J < \infty$, com $s_0 = 0$ e $s_J > t$, para algum t observado, com $t > 0$, admita que tal partição divida o eixo do tempo em J

intervalos disjuntos, denotados por $I_1 = (s_0, s_1], I_2 = (s_1, s_2], \ldots, I_J = (s_{J-1}, s_J]$ (Ibrahim *et al.*, 2001).

O MEP, algumas vezes chamado modelo semiparamétrico é caracterizado pela aproximação da função taxa de falha, $h(t)$ por segmentos de retas definidos pelos intervalos determinados pela partição $\{s_0, ..., s_J\}$, isto é, assume-se que em cada intervalo $I_j = (s_{j-1}, s_j]$, $j = 1, ..., J$ a função taxa de falha é constante e denotada por $h(t) = \lambda_j$, $\lambda_j > 0$, $\forall t \in I_j$. Consequentemente, a função taxa acumulada, $H(t)$, associada ao j-ésimo intervalo, $I_j = (s_{j-1}, s_j]$, é dada pela soma das áreas dos retângulos cujas bases são determinadas pelos intervalos definidos pela partição $\{s_0, ..., s_J\}$, e com alturas dada pela função taxa de falha, $h(t)$, ou seja, $H(t) = \sum_{k=1}^{j-1} \lambda_k(s_k - s_{k-1}) + \lambda_j(t - s_{j-1})$, para $t \in I_j$, $j = 1, ..., J$.

Temos que a função densidade de probabilidade e a função de sobrevivência do MEP são expressas respectivamentes por

$$f(t|\boldsymbol{\lambda}) = \begin{cases} \lambda_1 \exp\{-\lambda_1 t\}, & \text{se } t \in I_1; \\ \lambda_j \exp\left\{-\left[\sum_{k=1}^{j-1} \lambda_k(s_k - s_{k-1}) + \lambda_j(t - s_{j-1})\right]\right\}, & \text{se } t \in I_j, j > 1, \end{cases} \tag{2.1}$$

e

$$S(t|\boldsymbol{\lambda}) = \begin{cases} \exp\{-\lambda_1 t\}, & \text{se } t \in I_1; \\ \exp\left\{-\left[\sum_{k=1}^{j-1} \lambda_k(s_k - s_{k-1}) + \lambda_j(t - s_{j-1})\right]\right\}, & \text{se } t \in I_j, j > 1, \end{cases} \tag{2.2}$$

com $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)$, $\lambda_j > 0$, $\forall j = 1, \ldots, J$.

Apresentamos na Figura 2.1 os gráficos da função de sobrevivência e da função taxa de falha para o MEP. Os mesmos foram gerados considerando quatro partições no eixo dos tempos ($J = 4$) e com função de taxa de falha assumindo os seguintes valores $\lambda_1 = 0,017$; $\lambda_2 = 0,043$; $\lambda_3 = 0,120$ e $\lambda_4 = 0,760$.

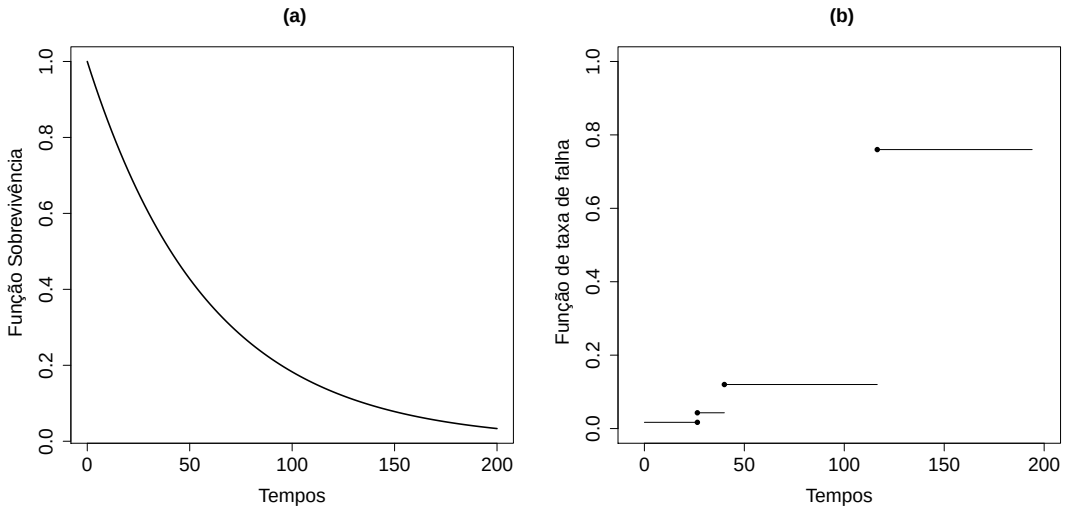


FIGURA 2.1: (a) Função de sobrevivência e (b) função de taxa de falha para o MEP.

2.2 Modelo de regressão semiparamétrico (MRSP)

Em muitas situações práticas temos que o tempo de vida (T) é influenciado por uma ou mais covariáveis. Por exemplo, o tempo de sobrevivência de um paciente pode depender de sua idade, da quantidade de cigarros que fuma por dia e de uma série de outros fatores. Uma maneira de determinar o relacionamento entre o tempo de vida e o conjunto de covariáveis é por meio de um modelo de regressão. Podemos citar duas classes importantes de modelos de regressão: os modelos de riscos proporcionais para T e modelos de locação-escala ou modelos de testes acelerados para $\log(T)$. Abordaremos neste trabalho apenas a primeira classe. Uma descrição detalhada sobre modelos de testes acelerados pode ser obtida em Lawless & Lawless (1982) e Nelson (1990). A família de modelos de riscos proporcionais é largamente utilizada na maioria das vezes que tratamos de dados sobrevivência e pode ser definida pela função de risco como

$$h(t;\boldsymbol{x}) = h_0(t)\exp\{\boldsymbol{x}^\top\boldsymbol{\beta}\} \tag{2.3}$$

em que, $\boldsymbol{\beta}$ é um vetor $p \times 1$ de coeficientes de regressão associados ao vetor de covariáveis $\boldsymbol{x}^\top = (x_1, x_2, \ldots, x_p)$ e h_0 é a função de risco para um indivíduo, também chamada de

função basal.

A principal suposição considerada para este modelo é a proporcionalidade entre os riscos. Se não assumirmos uma particular forma para $h_0(t)$, obtemos o modelo de riscos proporcionais de Cox (Cox, 1972), que tem ampla literatura de aplicações e extensões (Lawless & Lawless, 1982). Uma outra abordagem é obtida se considerarmos uma representação paramétrica para $h_0(t)$. Desta forma obtemos uma família paramétrica de riscos proporcionais. Considerando a formulação do modelo como apresentada na Seção (2.1) e assumindo a função de risco $h_j(t) = \lambda_j \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$, podemos expressar a função densidade de probabilidade para o MRSP por

$$f(t;\boldsymbol{\gamma}) = \begin{cases} \lambda_1 \exp(\boldsymbol{x}^\top \boldsymbol{\beta}) \exp\{-\lambda_1 t \ \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\}, \text{se } t \in I_1; \\ \lambda_j \exp(\boldsymbol{x}^\top \boldsymbol{\beta}) \exp\left\{-\left[\sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) + \lambda_j (t - s_{j-1})\right] \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}\right\}, \\ \text{se } t \in I_j, j > 1 \end{cases} \quad (2.4)$$

A partir da equação (2.4) podemos mostrar que a função de sobrevivência do modelo é dada por

$$S(t;\boldsymbol{\gamma}) = \begin{cases} \exp\{-\lambda_1 t \ \exp(\boldsymbol{x}^\top \boldsymbol{\beta})\}, \text{se } t \in I_1; \\ \exp\left\{-\left[\sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) + \lambda_j (t - s_{j-1})\right] \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}\}\right\}, \text{se } t \in I_j, j > 1 \end{cases} \quad (2.5)$$

com $\boldsymbol{\gamma} = (\boldsymbol{\lambda}, \boldsymbol{\beta})$, $\lambda_j > 0$, $\forall j = 1, \ldots, J$.

2.3 Inferência bayesiana

Denotamos por $\boldsymbol{D} = (n, \mathbf{t}, \boldsymbol{X}, \delta)$ os dados observados, com $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)^\top$, com $\delta_i = 1$ se o i-ésimo indivíduo for falha e 0 caso contrário. $\boldsymbol{X}$ é uma matriz $n \times p$ de covariáveis. Assumindo $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_j)^\top$, podemos escrever a

função de verossimilhança de $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\lambda})$ para n indivíduos por

$$\begin{aligned} L(\boldsymbol{\gamma}; \boldsymbol{D}) &= \prod_{i=1}^{n}\prod_{j=1}^{J}(\lambda_j \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}))^{\nu_{ij}\delta_i} \exp\Bigg\{ -\nu_{ij}\Bigg[\lambda_j(t_i - s_{j-1}) \\ &\quad + \sum_{k=1}^{j-1}\lambda_k(s_k - s_{k-1})\Bigg] \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\Bigg\}. \end{aligned} \tag{2.6}$$

sendo que $\nu_{ij} = 1$ indica se o i-ésimo indivíduo é falha ou censura no j-ésimo intervalo e 0 caso contrário, $\boldsymbol{x}_i^\top$ denota o vetor $p \times 1$ de covariáveis para o i-ésimo indivíduo e $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ é o vetor dos coeficientes de regressão correspondentes. O indicador ν_{ij} é necessário para definir a função verossimilhança sobre cada um dos J intervalos. Realizamos a partição do eixos dos tempos como sugerido por Yin (2005) que afirma que uma forma razoável de partição seria equilibrar o número de falhas nos intervalos, além de garantir que pelo menos uma falha aconteça em cada intervalo de tempo.

Assumindo independência entres os parâmetros do MRSP a densidade a *priori* conjunta é expressa por $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})$ em que os parâmetros $\boldsymbol{\beta_k}$ seguem uma distribuição Normal, com média μ_{β_k} e variância $\sigma^2_{\beta_k}$, em notação $\boldsymbol{\beta_k} \sim N(\mu_{\beta_k}, \sigma^2_{\beta_k})$, $k = 1, \ldots, p$ e $\lambda_j \sim \pi(\lambda_j)$, $j = 1, \ldots, J$. Combinando essa densidade a *priori* com a verossimilhança (2.6), a densidade a *posteriori* conjunta é dada por

$$\begin{aligned} \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{D}) &= L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D})\pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda}) \\ &\propto \prod_{i=1}^{n}\prod_{j=1}^{J}(\lambda_j \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}))^{\nu_{ij}\delta_i} \exp\Bigg\{ -\nu_{ij}\Bigg[\lambda_j(t_i - s_{j-1}) \\ &+ \sum_{k=1}^{j-1}\lambda_k(s_k - s_{k-1})\Bigg] \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})\Bigg\} \exp\left\{\frac{\boldsymbol{\beta}(2\mu - \boldsymbol{\beta})}{2\sigma^2}\right\}\pi(\boldsymbol{\lambda}), \end{aligned} \tag{2.7}$$

em que $\pi(\boldsymbol{\lambda})$ é uma distribuição a *priori* de λ.

Como a densidade a *posteriori* conjunta não é uma densidade padrão usamos métodos de Monte Carlo via cadeias de Markov (MCMC), tais como o amostrador de Gibbs e Metropolis-Hasting. Podemos mostrar que as densidades a *posteriori* condicionais para o amostrador de Gibbs são expressas por $\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}, \boldsymbol{D}) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D})\pi(\boldsymbol{\beta})$ e $\pi(\boldsymbol{\lambda}|\boldsymbol{\beta}, \boldsymbol{D}) \propto L(\boldsymbol{\beta}, \boldsymbol{\lambda}; \boldsymbol{D})\pi(\boldsymbol{\lambda})$.

Como as densidades a *posteriori* condicionais não possuem forma fechada usaremos o algoritmo de Metropolis-Hastings dentro do ciclo do algoritmo de Gibbs (Gilks *et al.*, 1996) para gerar amostras de $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$.

2.4 Aplicação

Nesta Seção ilustraremos a metodologia descrita anteriormente com um conjunto de dados simulados e um conjunto de dados reais.

2.4.1 Dados simulados

Para examinar o desempenho das medidas de diagnósticos propostas neste trabalho, nós consideramos um conjunto de dados simulados. Consideramos as covariáveis x_{i1} e x_{i2}, $i = 1, \ldots n$, sendo x_{i1} gerado de uma distribuição Bernoulli com parâmetro $p = 0,5$ e x_{i2} de uma distribuição normal $(0, 2)$. Os tempos de falha T_i foram gerados de uma distribuição exponencial(λ_i) com função de risco $h(t_i) = \lambda_1 \exp(\beta_1 x_{i1} + \beta_2 x_{i2})$ em que, $\beta_1 = 0,8$, $\beta_2 = 1,2$ e $\lambda_1 = 0,6$; os tempos de censura C_i também foram gerados de uma distribuição exponencial com parâmetro λ_c, em que o mesmo é escolhido de forma a monitorar a quantidade de dados censurados, neste caso temos 16% de dados censurados. Assumimos ainda, que T_i e C_i são independentes. Os tempos de sobrevivência t_i, $i = 1, \ldots 50$, foram tomados como $t_i = \min(T_i, C_i)$, δ_i é o indicador de censuras com $\delta_i = 1$, se $T_i \leq C_i$, e 0, se $T_i > C_i$. Selecionamos os casos 13, 28 e 39 para perturbar afim de criarmos observações influentes no conjunto de dados. Os casos perturbados 13 e 39 representam tempos de falha e o caso 28 tempo de censura. Nós escolhemos um, dois ou três casos e perturbamos a variável resposta da seguinte maneira $\tilde{t}_i = t_i + 5\tau_t$, em que τ_t são os desvios padrão dos tempos t_i. Na Tabela 2.1 o conjunto de dados (a) denota os dados originais simulados sem nenhuma perturbação e os conjuntos de dados (b)–(f) denota os dados com casos perturbados.

Consideramos uma análise bayesiana com densidades a *priori* independentes para

os parâmetros do modelo, com $\beta_k \sim N(0,100)$, k=1,2, e $\lambda_j \sim G(1;0,0001)$, com j=1 em que, $N(0,100)$ denota uma distribuição normal com média 0 e variância 100 e $G(1;0,0001)$ uma distribuição gama com média $1/0,0001$ e variância $1/(0,0001)^2$. A distribuição a *posteriori* para os parâmetros obtida não é tratável analiticamente, portanto métodos MCMC podem ser considerados. Como as condicionais não tem forma fechada geramos amostras das distribuição a *posteriori* dos parâmetros usando o algoritmos de Gibbs com Metropolis-Hasting da seguinte forma: geramos duas cadeias paralelas cada uma com 35.000 iterações para cada parâmetro, descartamos as primeiras 5.000 iterações afim de eliminar os efeitos dos valores iniciais e evitar problemas de correlação, consideramos espaçamento de tamanho 10, resultando uma amostra de Gibbs de tamanho 6.000. Para monitorar a convergência do amostrador de Gibbs utilizamos a aproximação desenvolvida por Gelman e Rubin(1992). Afim de avaliar a robustez do modelo relacionado às escolhas dos hiperparâmetros das *priori*, um estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas. Na Tabela 2.1 apresentamos as estimativas a *posteriori* para os dados simulados com e sem perturbações em alguns casos e os valores para os critérios de seleção B e DIC (discutidos na Seção 1.1). Para o cálculo da CPO (Equação 1.1), temos que, se $\delta_i = 1$ então $g(t_i|\boldsymbol{\gamma}) = f(t;\boldsymbol{\gamma})$ (Equação 2.4) e se $\delta_i = 0$ temos $g(t_i|\boldsymbol{\gamma}) = S(t;\boldsymbol{\gamma})$ (Equação 2.5). Os critérios B e DIC demonstram que o modelo sem casos perturbados (a) apresenta o melhor ajuste para os dados.

TABELA 2.1: Médias e desvios padrão a *posteriori* para os parâmetros do modelo dos dados simulados.

Dados	Casos	β_1		β_2		λ_1		B	DIC
	perturbados	Média	DP.	Média	DP.	Média	DP.		
a	Nenhum	0,747	0,327	1,002	0,079	0,516	0,136	-66,410	132,463
b	13	-0,009	0,468	0,360	0,109	0,179	0,066	-135,895	248,703
c	28	-0,487	0,453	0,616	0,126	0,308	0,110	-130,559	236,797
d	39	2,384	0,336	0,810	0,105	0,091	0,023	-98,637	180,818
e	{13,28}	-0,353	0,513	0,305	0,129	0,160	0,065	-146,771	276,849
f	{13,28,39}	0,792	0,351	0,062	0,089	0,055	0,015	-152,808	291,991

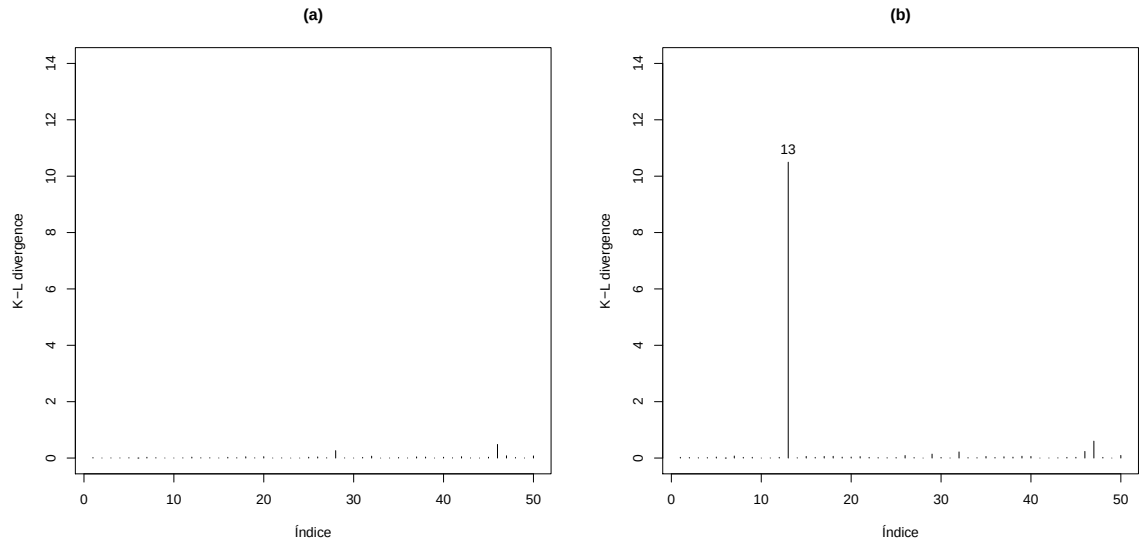
Considerando as amostras de Gibbs, estimamos as medidas de divergência de K–L (apresentado na Seção 1.2.1) para cada um dos casos (a,b,c,d,e e f), esses resultados são

graficados na Figura 2.2. Ainda estimamos as medidas de calibração (discutidos na Seção 1.2.1) referentes a cada caso. Os resultados apresentados na Tabela 2.2 indicam que sem perturbação nos dados (a), os casos selecionados não são influentes, pois obtivemos valores pequenos para $K(P, P_{(-i)})$ em cada caso. Entretanto, após perturbação (b–f) obtivemos medidas $K(P, P_{(-i)})$ grandes e calibração próximas ou iguais ao valor 1 indicando que estes casos são influentes.

TABELA 2.2: Medidas de diagnósticos para os dados simulados.

Nomes dos Dados	Identificação dos casos	$K(P, P_{(-i)})$	Calibração
a	13	0,011	0,574
	28	0,262	0,819
	39	0,001	0,530
b	13	10,496	1,000
c	28	11,267	1,000
d	39	8,174	1,000
e	13	3,813	0,999
	28	2,932	0,999
f	13	2,129	0,996
	28	2,563	0,998
	39	1,975	0,995

Na Figura 2.2 plotamos $K(P, P_{(-i)})$ para o modelo proposto, podemos notar que K-L identifica os casos influentes apresentando valores maiores de $K(P, P_{(-i)})$ para os casos perturbados se comparados aos outros casos.



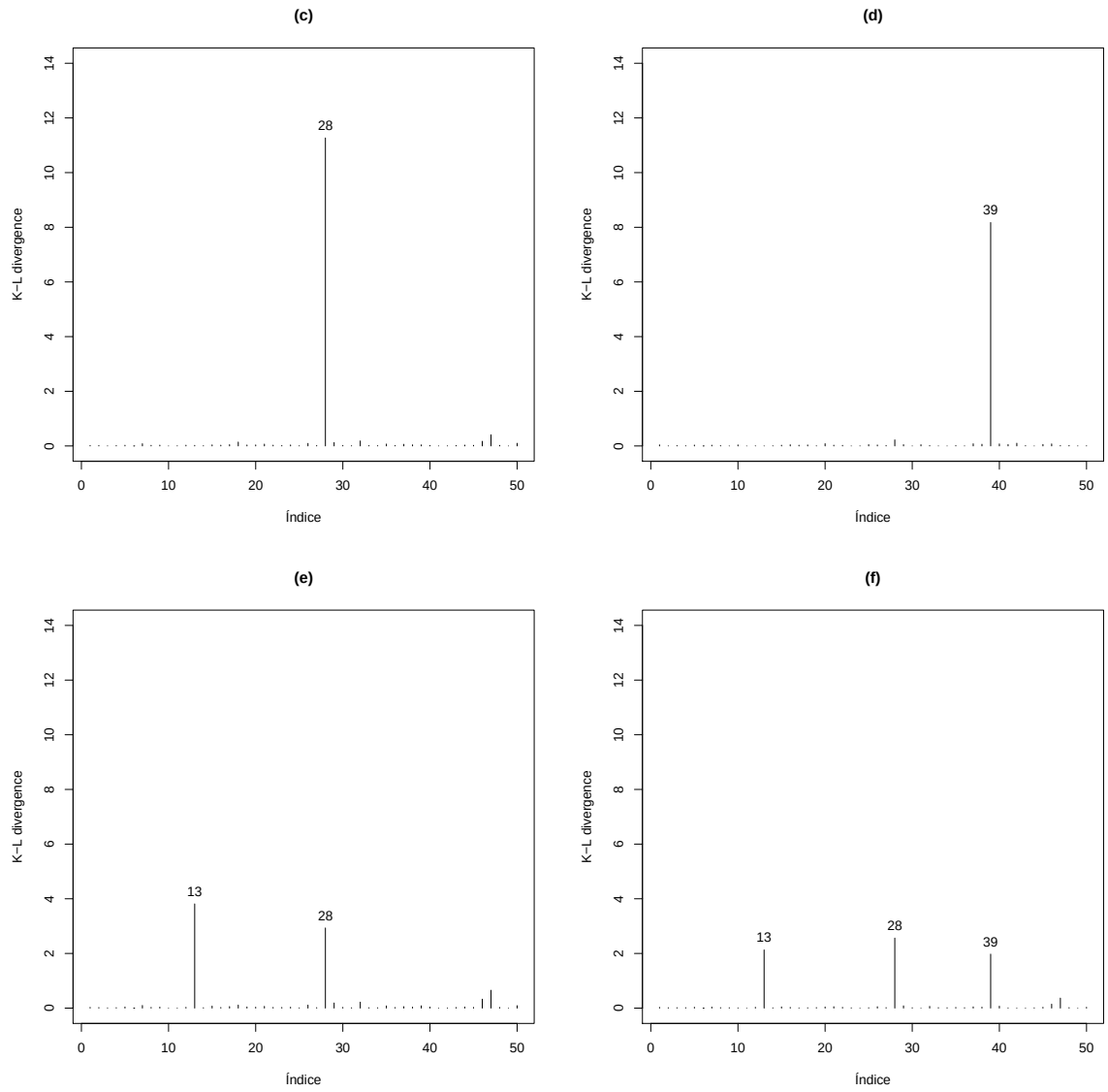


FIGURA 2.2: Gráfico de índices de $K(P, P_{(-i)})$ para os dados simulados.

2.4.2 Dados de infecção renal

Os dados considerados nesta seção referem-se ao estudo descrito em McGilchrist & Aisbett (1991), os quais relacionam os tempos de reincidência de infecção renal em 38 pacientes. Consideramos a presença de cinco covariáveis: x_1, idade do paciente (com média de 43,7 anos e desvio padrão de 14,8 anos); x_2, gênero do paciente (0-masculino, 1-feminino); x_3, x_4 e x_5 indicam o tipo da doença apresentada pelo paciente, sendo GN (glomerulo nefrite), AN (nefrite aguda) e PKD (rim policístico) respectivamente. Em nosso estudo consideramos somente o primeiro tempo (em dias) observado. Há no conjunto de

dados 28,9 % de censuras.

Pelo método gráfico baseado no tempo total em teste (TTT) transformado, descrito por Barlow & Campo (1975) identificamos a forma da função de risco com o objetivo de avaliar se o modelo é adequado aos tempos de vida. Em nosso caso, temos indício de que os dados apresentam a função de risco não monótona, como observado na Figura 2.3, assim ajustamos MRSP (descrito na Seção 2.2), considerando $J = 1, 2, 3, 4$ e 5 com as cinco covariáveis descritas anteriormente.

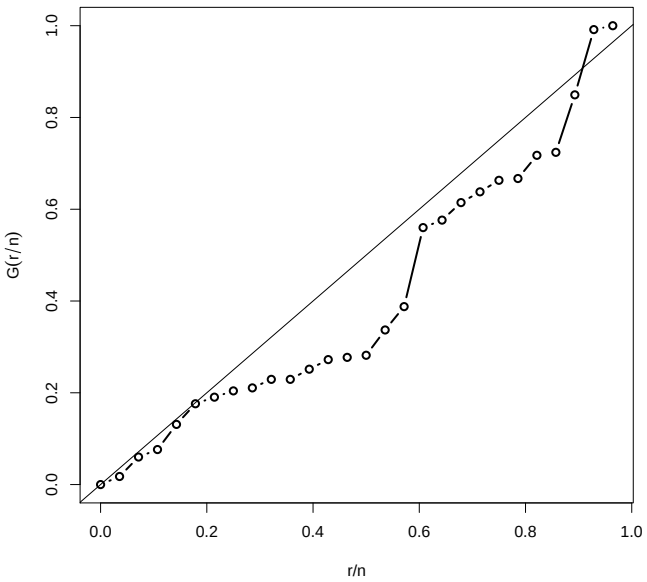


FIGURA 2.3: Gráfico TTT plot para dados de infecção renal.

Para nossa análise bayesiana consideramos densidades a *priori* independentes para os parâmetros β_k, ou seja, $\beta_k \sim N(0; 100)$, k=1,...5 e escolhemos para λ densidades a *priori* dependentes, por existir uma relação de dependência entre os parâmetros em cada partição, com a seguinte distribuição, $\phi_1 \sim G(0,1; 0,01)$, $\xi \sim G(0,1; 0,01)$ e $\phi_j \sim N(\phi_{j-1}; \xi)$, j=2,...J. Como os parâmetros λ_j são não negativos, adotamos a seguinte parametrização, $\lambda_j = \log(\phi_j)$, com j=1,...J. Considerando essas densidades a *priori* geramos amostras de Gibbs através do algoritmo de Gibbs com Metropolis-Hasting da seguinte forma: geramos duas cadeias paralelas cada uma com 35.000 iterações com um *burn in* de 5.000 e saltos de tamanho 10, resultando uma amostra de Gibbs de tamanho 6.000. Para monitorar a

convergência do amostrador de Gibbs utilizamos a aproximação desenvolvida por Gelman e Rubin(1992). Na Tabela 2.3 apresentamos os valores para os critérios de seleção B e DIC para os modelos ajustados. Ambos critérios indicam um MRSP com $J = 4$ intervalos disjuntos na partição do eixo dos tempos. Para avaliar a robustez do modelo relacionado às escolhas dos hiperparâmetros das *priori*, um estudo de sensibilidade foi realizado, sendo que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas e não alteraram os resultados da Tabela 2.3 e Tabela 2.4.

TABELA 2.3: Critérios de seleção de modelos considerando os dados de infecção renal.

Modelo	Critério	
J	B	DIC
1	-160,677	318,500
2	-163,058	319,500
3	-162,515	320,500
4	-160,619	317,000
5	-164,677	323,200

Na Tabela 2.4 apresentamos os resumos da distribuição a *posteriori* do MRSP para o melhor modelo. Para todos os parâmetros observamos valores de $\hat{R}$ próximos de 1, indicando que as iterações foram suficientes para se chegar a convergência. A "highest posterior density" (HPD)(Chen & Shao, 1999) (maiores detalhes consulte Bernardo & Smith, 2000) com 95% de credibilidade indica que somente as covariáveis sexo e PKD são significativas.

TABELA 2.4: Resumos a *posteriori* para os tempos.

Parâmetro	Média	Desvio Padrão	$HPD(95\%)$	$\hat{R}$
$\beta_{1(\text{idade})}$	-0,001	0,016	(-0,034; 0,028)	1,001
$\beta_{2(\text{sexo})}$	-1,175	0,484	(-2,100;-0,195)	1,003
$\beta_{3(\text{GN})}$	0,251	0,596	(-0,944; 1,391)	1,001
$\beta_{4(\text{AN})}$	0,782	0,628	(-0,373; 2,077)	1,001
$\beta_{5(\text{PKD})}$	-2,028	1,016	(-4,135;-0,169)	1,006
λ_1	0,020	0,017	(0,001; 0,053)	1,001
λ_2	0,051	0,044	(0,003; 0,135)	1,002
λ_3	0,014	0,015	(0,000; 0,041)	1,001
λ_4	0,030	0,027	(0,001; 0,076)	1,002

Com as amostras de Gibbs estimamos as medidas de divergência de K–L para cada uma das observações, esses resultados são graficados na Figura 2.4, em que as observações 10, 15 e 21 apresentam maiores valores quando comparados com as demais observações.

Para verificarmos se essas observações são influentes estimamos a calibração da medida de divergência K–L, essas estimativas são apresentados na Tabela 2.5 conjuntamente com as respectivas estimativas da divergência K–L, indicando que as três observações são influentes.

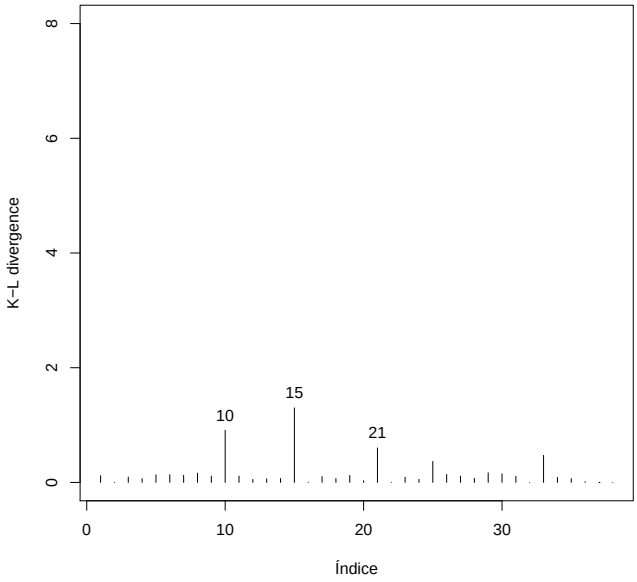


FIGURA 2.4: Gráfico de índices de $K(P, P_{-i})$ para os dados de infecção renal.

TABELA 2.5: Identificação dos casos influentes para dados de infecção renal.

Identificação do caso			Influência caso a caso	
Paciente	Tempo	Idade	$K(P, P_{(-i)})$	Calibração
10	154	51,500	0,910	0,957
15	536	17,000	1,300	0,981
21	562	46,500	0,601	0,918

Na Tabela 2.6 apresentamos as novas estimativas dos parâmetros após a exclusão de cada caso influente e também as variações relativas (VR) nas estimativas após termos excluídos um dos três pontos influentes ou descartados todos de uma só vez (representados pelo conjunto $I = \{10, 15, 21\}$). As VR (em porcentagem) para cada parâmetro estimado são obtidas por $VR_{\gamma_d} = |(\hat{\gamma}_d - \hat{\gamma}_{d(I)})/\hat{\gamma}_d| \times 100\%$, em que $\hat{\gamma}_{d(I)}$ denota a média a *posteriori* de γ_d, com $d = 1, \ldots, 9$ após o conjunto I de observações ter sido removido. A densidade a *posteriori* máxima (HPD) com 95% de credibilidade para cada nova estimativa é

apresentada entre parênteses na Tabela 2.6.

Notamos valores altos de VR para o parâmetro $\widehat{\beta_1}$ quando há exclusão de qualquer um dos pontos influentes e também quando excluímos o conjunto I, observamos que a estimação dos parâmetros do modelo apresenta sensibilidade em relação aos pontos influentes, observando-se a ocorrência de mudança de sinal no valor do parâmetro como ocorre em $\widehat{\beta_1}$ quando excluímos o caso 10. Em adição nós notamos que ocorreram algumas mudanças nas estimativas de alguns coeficientes, particularmente $\widehat{\beta_5}$ deixa de ser significativo após retirada do caso 21 e do conjunto I de observações influentes.

TABELA 2.6: Estimativas, VR (em %) e a correspondente HPD (95%) ajustados para o conjunto de dados de infecção renal.

Parâmetros	Observações descartadas			
	10	15	21	I
$\beta_{1(\text{idade})}$	0,006 464 (-0,023;0,039)	-0,018 970 (-0,048;0,015)	-0,004 135 (-0,037;0,029)	-0,008 370 (-0,037;0,025)
$\beta_{2(\text{sexo})}$	-1,938 64 (-3,114;-0,815)	-1,137 03 (-2,007;-0,178)	-1,400 19 (-2,318;-0,380)	-2,188 86 (-3,271 ;-1,030)
$\beta_{3(\text{GN})}$	0,593 136 (-0,483;1,716)	0,086 65 (-1,029;1,117)	0,264 05 (-0,800;1,461)	0,431 71 (-0,648; 1,424)
$\beta_{4(\text{AN})}$	0,575 26 (-0,599;1,868)	0,941 20 (-0,199;2,193)	0,972 24 (-0,242; 2,214)	0,711 09 (-0,465 ;1,947)
$\beta_{5(\text{PKD})}$	-2,801 38 (-5,020;-0,770)	-2,367 16 (-4,347; -0,551)	-0,188 90 (-2,937; 2,393)	-0,555 72 (-3,143; 1,825)
λ_1	0,025 23	0,039 91	0,024 15	0,042 105
λ_2	0,061 20	0,082 61	0,051 01	0,119 132
λ_3	0,027 85	0,035 137	0,019 28	0,075 408
λ_4	0,047 56	0,110 264	0,045 50	0,182 500

Baseados nestas análises concluímos que o modelo de regressão final é dado por

$$S(t_i;\boldsymbol{\lambda},\boldsymbol{\beta}) = \exp\left\{-\left[\sum_{k=1}^{j-1}\lambda_k(s_k - s_{k-1}) + \lambda_j(t - s_{j-1})\right]\exp\{x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4 + x_{i5}\beta_5\}\right\},$$

$i = 1, \ldots, 38$ e $j = 1, \ldots, 4$.

As médias a *posteriori*, os desvios padrão e a densidade a *posteriori* máxima (HPD) com 95% de credibilidade para λ_j, $j = 1, \ldots, 4$ e para os β_j são apresentados na Tabela 2.7.

TABELA 2.7: Resumo a *posteriori* para o modelo final.

Parâmetro	Média	Desvio Padrão	HPD (95%)
$\beta_{2(\text{sexo})}$	-1,154	0,493	(-2,124;-0,186)
$\beta_{3(\text{GN})}$	0,241	0,534	(-0,743;1,343)
$\beta_{4(\text{AN})}$	0,736	0,529	(-0,283;1,798)
$\beta_{5(\text{PKD})}$	-2,081	0,956	(-4,070;-0,339)
λ_1	0,017	0,010	(0,001;0,038)
λ_2	0,043	0,028	(0,004;0,098)
λ_3	0,012	0,008	(0,001;0,029)
λ_4	0,026	0,018	(0,002;0,060)

Os valores para os critérios de seleção B e DIC são -159,321 e 315,000 respectivamente. Comparados com os valores da Tabela 2.3 percebemos que temos um melhor ajuste para o modelo. Ainda na Tabela 2.7 podemos notar que as covariáveis sexo e PKD são significativas, ou seja, a infecção renal está relacionada com o sexo do indivíduo em estudo, havendo diferenças entre as taxas de infecção no sexo feminino e masculino e também no tipo de doença, na qual a doença PKD se destaca das outras duas, como podemos observar na Figura 2.5.

Na Figura 2.5 apresentamos as estimativas da função de sobrevivência estratificada por sexo dos indivíduos em estudo de acordo com cada tipo de doença apresentada, podemos notar que a função de sobrevivência de pacientes com doença PKD é superior aos dos pacientes com doenças AN e GN. Também podemos observar que o tempo de sobrevivência é maior em pacientes do sexo feminino.

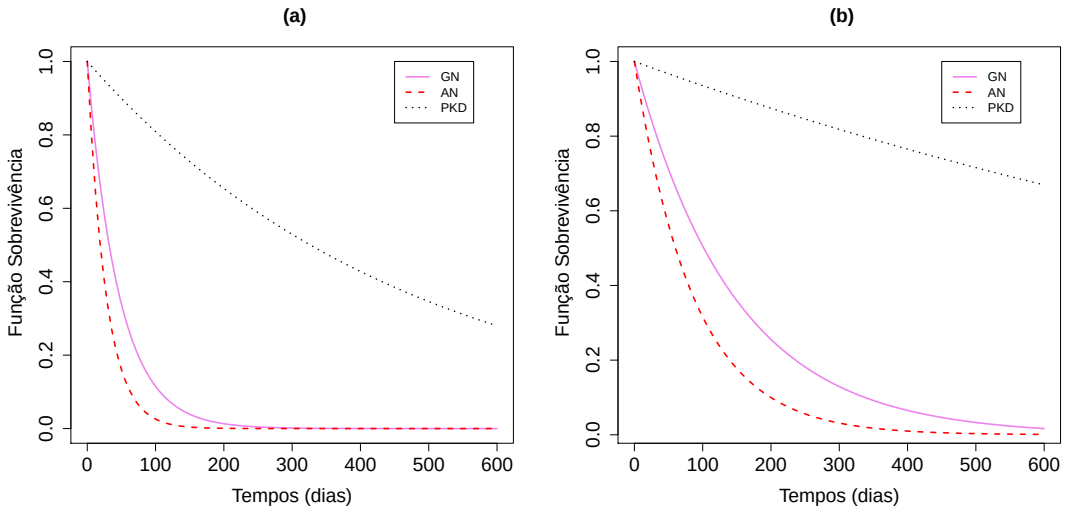


FIGURA 2.5: Estimativas da função de sobrevivência por sexo: (a) sexo masculino e (b) sexo feminino.

2.5 Conclusões

Podemos observar por meio dos dados simulados que a divergência de Kullback-Leibler é um método eficaz para detectar observações influentes. Pelo conjunto de dados reais identificamos quais pontos influenciaram o ajuste do modelo, pela retirada de tais pontos obtivemos um modelo final e estimamos seus parâmetros. Ainda podemos observar pelos critérios de seleção de modelos que o modelo final obtido tem indícios de ser o ideal para o conjunto de dados considerado.

Capítulo 3

Modelo de sobrevivência com fração de cura

Modelos de sobrevivência incorporando fração de cura estão se tornando cada vez mais populares, principalmente na análise dos dados clínicos de câncer, sendo utilizado para modelar dados de tempo do evento para vários tipos de câncer em que uma proporção significativa dos pacientes são "curados". Modelos com fração de cura são baseados em mecanismos de ativação de fatores latentes que envolvem tempos de falha em dois diferentes níveis: um tempo de falha observado (T) que é o tempo observado para ocorrer o evento de interesse e o tempo de um evento latente (R_i), $i = 1, 2, \ldots M$ que é o tempo de ativação dos M fatores latentes que geram o tempo de falha observado (T) (Cooner *et al.*, 2007).

3.1 Formulação do modelo

Suponha que, para um indivíduo da população, seja denotado por M o número de causas que competem para produzir um evento de interesse. Assumimos que M segue uma determinada distribuição de probabilidade. Além disso, seja denotado por R_i o tempo (aleatório) para a i-ésima causa produzir o evento de interesse. Isto é, R_i pode ser encarada como um tempo de incubação para a i-ésima causa. As variáveis R_i, $i = 1, 2, \ldots$, são

independentes e identicamente distribuídas (i.i.d.) com função de distribuição comum, $F(\cdot) = 1 - S(\cdot)$ e são independentes de M. Assim, o tempo para o evento de interesse ser observado pode ser definido pela variável aleatória

$$T = \min\{R_1, \ldots, R_M\},$$

com $P(T = \infty | M = 0) = 1$.

Tsodikov *et al.* (2003), Rodrigues *et al.* (2008) entre outros demonstram que $S_{\text{pop}}(t) = \phi(S(t))$, em que $\phi(\cdot)$ é a função geradora de probabilidade para o número de causas competitivas M. Desta forma temos que

$$\begin{aligned}
S_{\text{pop}}(t) &= P(\text{não ocorrer o evento de interesse até o tempo } \ t) \qquad (3.1)\\
&= P(R_0 > t, M = 0) + P(\min\{R_1, \ldots, R_M\} > t, M \geq 1)\\
&= P(R_0 > t, M = 0) + P(R_1 > t, R_2 > t, \ldots, R_M > t, M \geq 1)\\
&= P(R_0 > t, M = 0) + P(R_1 > t, M = 1) + P(R_1 > t, R_2 > t, M = 2) +\\
&\quad \ldots + P(R_1 > t, \ldots, R_M > t, M = m) + \ldots\\
&= P(R_0 > t; M = 0)P(M = 0) + P(R_1 > t; M = 1)P(M = 1) +\\
&\quad P(R_1 > t, R_2 > t, \ldots, R_M > t; M = m)P(M = m) + \ldots\\
&= p_0 + S(t)p_1 + (S(t))^2 p_2 + \ldots + (S(t))^m p_m + \ldots\\
&= p_0 + \sum_{m=1}^{\infty} p_m (S(t))^m\\
&= \phi(S(t))
\end{aligned}$$

Podemos observar que $S_{\text{pop}}(t)$ não é uma função de sobrevivência própria, isto é $\lim_{t\to\infty} S_{\text{pop}}(t) > 0$.

Como em de Castro *et al.* (2009), consideramos que M segue uma distribuição binomial negativa com parâmetros θ e α (Piegorsch, 1990), com a seguinte função de

probabilidade

$$P(M = m; \theta, \alpha) = \frac{\Gamma(\alpha^{-1} + m)}{\Gamma(\alpha^{-1})\, m!} \left(\frac{\theta\alpha}{1 + \theta\alpha}\right)^m \left(\frac{1}{1 + \alpha\theta}\right)^{1/\alpha}, \tag{3.2}$$

$m = 0, 1, 2, \ldots$, $\theta > 0$ e $\alpha > -1/\theta$, de modo que

$$E[M] = \theta, \qquad Var(M) = \theta + \alpha\theta^2 \tag{3.3}$$

Como observado por Rodrigues *et al.* (2009a) a variância de M (número de causas ou riscos) em (3.3) proporciona ao mesmo tempo dois cenários importantes, o de sobredispersão que ocorre quando $\alpha > 0$ e pode ser interpretado no caso de pesquisas biológicas na área de câncer como um agrupamento de células cancerígenas ou tumorais e a subdispersão que ocorre no sentido inverso quando α assume valores negativos. A distribuição de probabilidade em (3.2) é muito flexível no sentido de fornecer ligações entre as distribuições geométrica, Poisson e binomial negativa, pois se α =1 temos o modelo geométrico e se $\alpha \to 0$ temos o modelo de Poisson.

A função geradora de probabilidade para o modelo binomial negativo é dada por

$$\phi(s) = \left[\frac{1}{1 + \alpha\theta(1 - s)}\right]^{1/\alpha}, \quad 0 \leq s \leq 1. \tag{3.4}$$

De (3.1) e (3.4) a função de sobrevivência populacional é expressa por

$$S_{\text{pop}}(t) = [1 + \alpha\theta F(t)]^{-1/\alpha}, \tag{3.5}$$

A proporção de indivíduos curados é determinada por $p_0 = \lim_{t\to\infty} S_{\text{pop}}(t)$, neste caso temos $p_0 = [1 + \alpha\theta]^{\frac{-1}{\alpha}}$. Da equação (3.5) podemos obter a função densidade populacional

$$\begin{aligned} f_{\text{pop}}(t) &= -\frac{d}{dt}\Big(S_{\text{pop}}(t)\Big) \\ &= \theta f(t)\Big[1 + \alpha\theta F(t)\Big]^{-(1+1/\alpha)}, \end{aligned} \tag{3.6}$$

em que $f(t) = \frac{d}{dt}(F(t))$.

3.2 Função de Verossimilhança

Supomos m indivíduos e M_i a variável que representa o número de causas competindo que podem produzir um câncer detectável para o i-ésimo indivíduo. Assumimos que os M_i's são variáveis aleatórias com distribuição binomiais negativas i.i.d. com média θ_i, $i = 1, \ldots, m$. Nós enfatizamos aqui que os M_i's não são observados e podem ser vistos como variáveis latentes. Entretanto, supomos que $R_{i1}, \ldots, R_{iM_i}$ são os tempos de incubação i.i.d. para cada célula cancerígena M_i do i-ésimo indivíduo, $i = 1, \ldots, m$ e todos tem função de distribuição acumulada $F(\cdot|\boldsymbol{\lambda})$. Seja Y_i o tempo observado tal que, $Y_i = \min(T_i, C_i)$, com $T_i = \min\{R_{i1}, \ldots, R_{iM_i}\}$, C_i o tempo de censura e δ_i o indicador de censura igual a 1 se $Y_i = T_i$ e 0 caso contrário. Nós incluímos as covariáveis no modelo por meio do valor esperado do número de causas que competem pra produzir um câncer, ou seja, $E(M_i) = \theta_i = \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta})$, $i = 1, \ldots, m$, em que $\boldsymbol{\beta}$ é um vetor $k \times 1$ de coeficientes de regressão. Os dados observados são $\boldsymbol{\mathcal{D}} = (n, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{X})$, com $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)^\top$ e $\boldsymbol{X}$ uma matrix $m \times k$ que contém as covariáveis.

Baseados na distribuição binomial negativa em (3.2) com $\alpha > 0$ e, considerando o modelo exponencial por partes para os tempos de promoção R_i como apresentado na Seção 2.1, com

$$f_j(y_i; \boldsymbol{\lambda}) = \lambda_j \exp\left\{ -\left[\sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) + \lambda_j (y_i - s_{j-1}) \right] \right\} \tag{3.7}$$

e

$$F_j(y_i; \boldsymbol{\lambda}) = 1 - \exp\left\{ -\left[\sum_{k=1}^{j-1} \lambda_k (s_k - s_{k-1}) + \lambda_j (y_i - s_{j-1}) \right] \right\}, \tag{3.8}$$

podemos expressar a função de verossimilhança da seguinte forma

$$L(\boldsymbol{\gamma}; \boldsymbol{\mathcal{D}}) \propto \prod_{i=1}^{m} \prod_{j=1}^{J} \left\{ \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) f_j(y_i; \boldsymbol{\lambda}) \right\}^{\delta_i \nu_{ij}} \left\{ 1 + \alpha \exp(\boldsymbol{x_i}^\top \boldsymbol{\beta}) F_j(y_i; \boldsymbol{\lambda}) \right\}^{-\nu_{ij}(\delta_i + 1/\alpha)}, \tag{3.9}$$

em que, $\boldsymbol{\gamma} = (\boldsymbol{\lambda}, \alpha, \boldsymbol{\beta})$ é o vetor de parâmetros, $\boldsymbol{\mathcal{D}} = (n, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{X}, \boldsymbol{\nu})$, $\boldsymbol{\nu} = (\nu_{11}, \ldots, \nu_{mJ})$

com $\nu_{ij} = 1$ se $s_{j-1} < y_i \leq s_j$ e $\nu_{ij} = 0$ caso contrário, $j = 1, \ldots, J$ e $i = 1, \ldots, m$.

3.3 Inferência bayesiana

Assumindo independência entre os parâmetros do modelo a densidade a *priori* conjunta é expressa por

$$\pi(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda}) = \pi(\alpha)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda})$$

em que o parâmetro α tem distribuição $G(a, b)$, com $G(a, b)$ denotando uma distribuição gama com média a/b e variância a/b^2. $\boldsymbol{\beta_l} \sim \pi(\boldsymbol{\beta}_l)$, $l = 1, \ldots, k$ e $\boldsymbol{\lambda}_j \sim \pi(\boldsymbol{\lambda}_j)$, $j = 1, \ldots, J$. Combinando essa densidade a *priori* com a verossimilhança (3.9), a densidade a *posteriori* conjunta é dada por

$$\pi(\alpha, \boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\mathcal{D}}) = L(\boldsymbol{\gamma}; \boldsymbol{\mathcal{D}})\pi(\alpha)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda}).$$

Como a densidade a *posteriori* conjunta não é uma densidade padrão usamos métodos de Monte Carlo via cadeias de Markov (MCMC), tais como o amostrador de Gibbs e Metropolis-Hasting. Podemos mostrar que as densidades a *posteriori* condicionais para o amostrador de Gibbs são expressas por

$$\pi(\boldsymbol{\beta}|\alpha, \boldsymbol{\lambda}, \boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\gamma}; \boldsymbol{\mathcal{D}})\pi(\boldsymbol{\beta});$$

$$\pi(\boldsymbol{\lambda}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\gamma}; \boldsymbol{\mathcal{D}})\pi(\boldsymbol{\lambda});$$

$$\pi(\alpha|\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\mathcal{D}}) \propto L(\boldsymbol{\gamma}; \boldsymbol{\mathcal{D}})\pi(\alpha).$$

Como as densidades a *posteriori* condicionais não possuem forma fechada usaremos o algoritmo de Metropolis-Hastings dentro do ciclo do algoritmo de Gibbs (Gilks *et al.*, 1996) para gerar amostras dos parâmetros envolvidos no modelo.

3.4 Aplicação

Nesta seção ilustraremos a metodologia descrita anteriormente com um conjunto de dados simulados e um conjunto de dados reais.

3.4.1 Dados simulados

Para examinar o desempenho das medidas de diagnósticos propostas neste trabalho, nós consideramos um conjunto de dados simulados, em que a covariável x_i, $i = 1, \ldots n$ foi gerada de uma distribuição normal (1; 0,5). Os tempos de falha T_i para o modelo (3.1) foram simulados de uma distribuição exponencial com parâmetro $\lambda = \exp(-3)$, consideramos a proporção de curados $p_{0i} = [1 + \alpha\theta_i)]^{-1/\alpha}$ com $\alpha = 2$, $\theta_i = \exp(\beta_0 + \beta_1 x_i)$ com $\beta_0 = -0,5$ e $\beta_1 = 0,7$ e a proporção de observações censuradas $p_{ci} = p_{0i} + 0,07$. Os tempos observados e os indicadores de censura foram gerados como segue:

1. Gerar $u_i \sim \text{uniforme}(0, 1)$;

2. Se $u_i < p_{0i}$, fazer $t_i = \infty$; caso contrário, fazer

$$t_i = F^{-1}\left(\frac{u_i^{-\alpha} - 1}{\alpha \exp(\beta_0 + \beta_1 x_i)}\right)$$

   com $F^{-1}(t) = -\log(1 - t)/e^{\lambda}$;

3. Gerar

$$c_i \sim \exp\left(\frac{e^{\alpha}(p_{ci} - p_{0i})}{1 - (p_{ci} - p_{0i})}\right);$$

4. Considerar $y_i = \min\{t_i, c_i\}$;

5. Se $y_i < c_i$, então $\delta_i = 1$, caso contrário, temos $\delta_i = 0$, $i = 1, \ldots, n$.

Para a análise bayesiana consideramos densidades a *priori* independentes para os parâmetros do modelo, com $\beta_l \sim N(0; 100)$, com $l = 0, 1$; $\alpha \sim G(1; 0, 01)$ e $\lambda_j \sim G(1; 0, 001)$, com j=1. Considerando essas densidades a *priori* geramos amostras de Gibbs

através do algoritmo de Gibbs com Metropolis-Hasting da seguinte forma: geramos duas cadeias paralelas cada uma com 35.000 iterações para cada parâmetro, descartamos as primeiras 5.000 iterações afim de eliminar os efeitos dos valores iniciais e evitar problemas de correlação, consideramos espaçamento de tamanho 10, resultando uma amostra de Gibbs de tamanho 6.000. Para monitorar a convergência do amostrador de Gibbs utilizamos a aproximação desenvolvida por Gelman e Rubin(1992). Afim de avaliar a robustez do modelo relacionado às escolhas dos hiperparâmetros das *priori*, um estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas.

Selecionamos os casos 14, 54 e 80 para perturbar afim de criarmos observações influentes no conjunto de dados. Nós escolhemos um ou dois casos e perturbamos a variável resposta da seguinte maneira $\tilde{t}_i = t_i + 6\tau_t$, em que τ_t são os desvios padrão dos tempos t_i. Na Tabela 3.1 apresentamos as estimativas a *posteriori* para os dados simulados. O conjunto de dados (a) denota os dados originais simulados sem nenhuma perturbação e os conjuntos de dados (b)–(f) denota os dados com casos perturbados, informamos ainda na Tabela 3.2 os valores para os critérios de seleção B e DIC (discutidos na seção 1.1). Esses critérios demonstram que o modelo sem casos perturbados (a) apresenta o melhor ajuste para os dados.

TABELA 3.1: Médias e desvios padrão a *posteriori* para os parâmetros do modelo dos dados simulados com fração de cura.

Dados	Casos	α		β_0		β_1		λ_1	
	perturbados	Média	DP	Média	DP	Média	DP	Média	DP
a	Nenhum	1,6764	1,2652	-0,5200	0,6972	0,5397	0,4917	0,0423	0,0127
b	14	6,3882	1,4722	3,0854	1,3681	1,0965	0,8882	0,0011	0,0009
c	54	6,7836	1,4893	3,8010	1,4194	0,6344	0,9039	0,0010	0,0008
d	80	6,5851	1,3550	4,5487	1,6104	-0,0673	0,8929	0,0010	0,0008
e	{14,54}	6,4562	1,3063	4,5235	1,5415	0,5336	0,8627	0,0004	0,0004
f	{14,54,80}	6,7219	1,3998	4,8643	1,3707	0,1222	0,8458	0,0005	0,0004

TABELA 3.2: Critérios de seleção de modelos.

Critério	Dados					
	a	b	c	d	e	f
B	-229,1443	-247,1962	-245,8269	-247,5562	-253,4850	-251,3234
DIC	458,1000	489,3000	485,2000	490,1000	505,0000	501,3000

Considerando as amostras de Gibbs, foram estimadas as medidas de divergência de K–L (apresentado na Seção 1.2.1) para cada um dos casos (a,b,c,d,e e f), esses resultados são graficados na Figura 3.1. Ainda foram estimadas as medidas de calibração (discutidos na Seção 1.2.1) referentes a cada caso. Os resultados apresentados na Tabela 3.3 indicam que sem perturbação nos dados (a), os casos selecionados não são influentes, pois obtivemos valores pequenos para $K(P, P_{(-i)})$ em cada caso. Entretanto, após perturbação (b–f) obtivemos medidas maiores de $K(P, P_{(-i)})$ e medidas de calibração maiores que 0,5 e próximas ao valor 1 indicando que estes casos são influentes.

TABELA 3.3: Medidas de diagnósticos para os dados simulados com fração de cura .

Nomes dos Dados	Identificação dos casos	$K(P, P_{(-i)})$	Calibração
a	14	0,0817	0,6942
	54	0,0265	0,6136
	80	0,1159	0,7274
b	14	2,1781	0,9967
c	54	2,7368	0,9989
d	80	1,9271	0,9946
e	14	0,3117	0,8405
	80	0,2130	0,7945
	54	0,3280	0,8468
f	80	0,2108	0,7932

Na Figura 3.1 plotamos $K(P, P_{(-i)})$ para o modelo proposto, podemos notar que K-L identifica os casos influentes apresentando valores maiores de $K(P, P_{(-i)})$ para os casos perturbados se comparados aos outros casos.

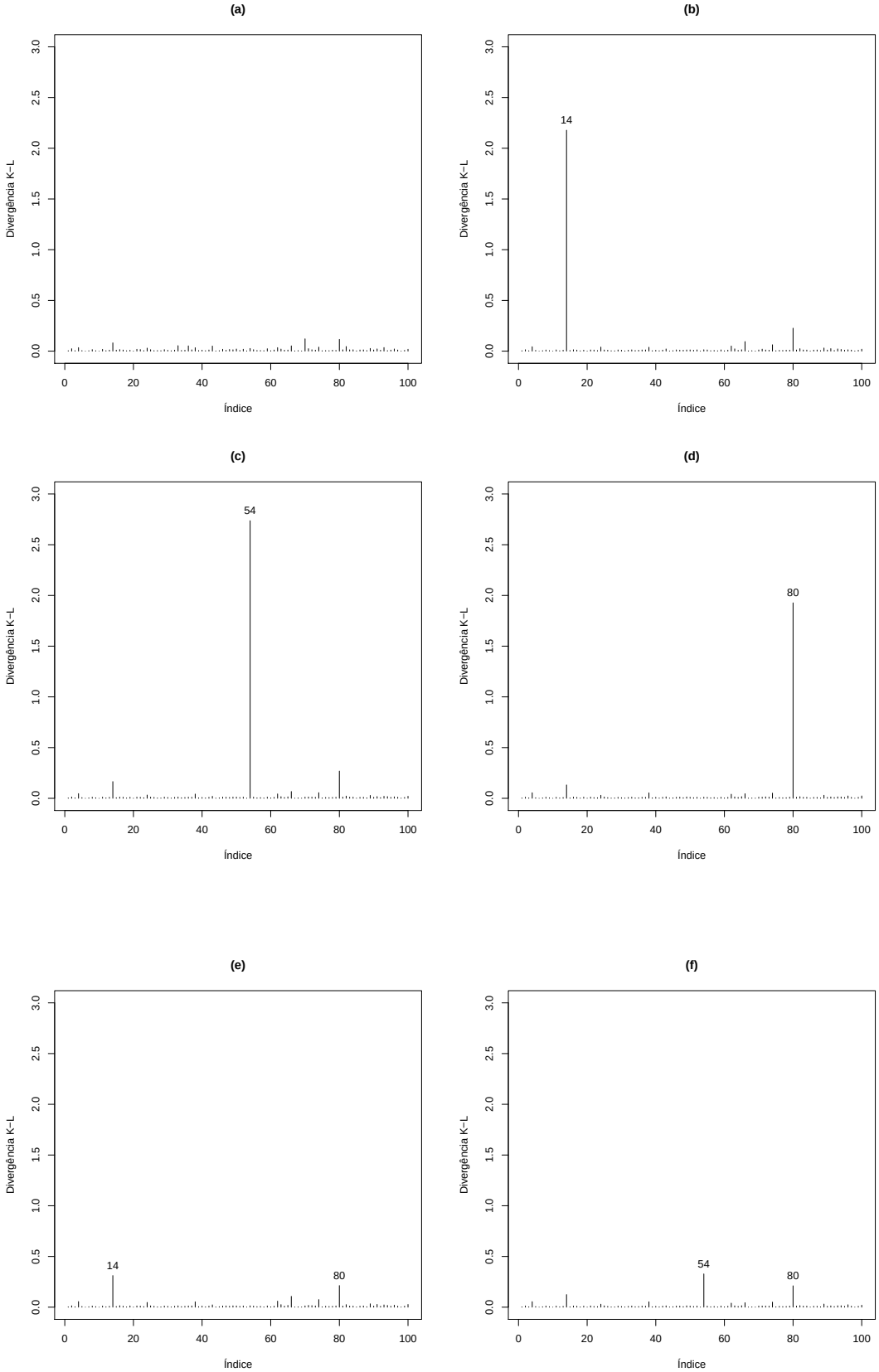


FIGURA 3.1: Gráfico de índices de $K(P, P_{(-i)})$ para os dados simulados com fração de cura.

3.4.2 Dados de melanoma

Os dados considerados nesta seção provém de um estudo de melanoma com o objetivo de avaliar a eficácia da aplicação de uma dosagem alta de interferon alfa-2b como forma de prevenir a recorrência de câncer. Os dados foram obtidos entre 1991 e 1995, mas houve acompanhamento dos pacientes até 1998. A variável resposta (y) representa o tempo até a morte dos paciente ou tempo de censura. Maiores detalhes deste conjunto de dados pode ser encontrado em Ibrahim *et al.* (2001). Da amostra original foram desconsiderados 10 pacientes devido a presença de dados faltantes, resultando em $n = 417$ pacientes, com 56% de observações censuradas. As variáveis consideradas neste estudo incluem y : tempo (em anos); x_1: categoria do nódulo (1,2,3,4) ; x_2: idade (anos); x_3: espessura do tumor(em mm).

Na Figura 3.2 apresentamos a estimativa de Kaplan–Meier da função de sobrevivência por categoria do nódulo para os dados de melanoma, na qual observamos a existência de uma apreciável fração de indivíduos "curados", pelo menos no que diz respeito ao intervalo de tempo abrangido pelo estudo.

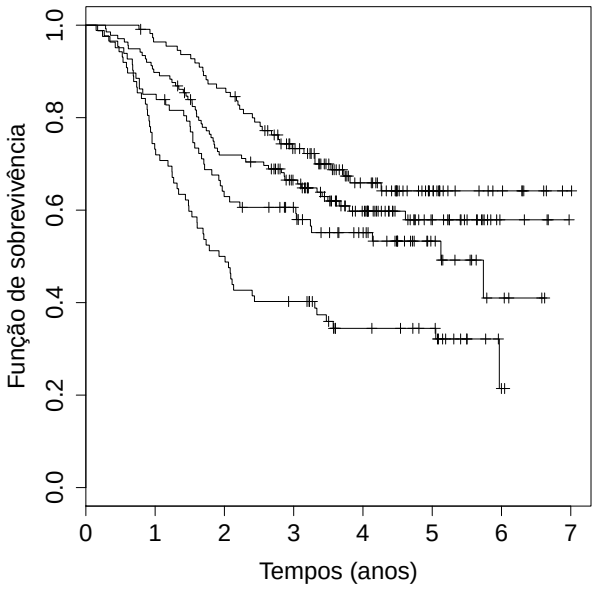


FIGURA 3.2: Estimativa de Kaplan-Meier da função de sobrevivência por categoria nódulo (1 a 4, de cima para baixo).

Uma análise bayesiana foi considerada e para cada componente de $\boldsymbol{\beta}$, consideramos densidades a *priori* independentes com uma distribuição normal com média 0 e variância 1000. Para $\lambda_1, \ldots, \lambda_J$, consideramos uma distribuição gama , ou seja, $\lambda_j \sim G(1; 0.01)$, j=1,... J. Para o parâmetro α consideramos uma distribuição a *priori* informativa, $\alpha \sim G(1; 1)$ com a finalidade de garantir que a distribuição a *posteriori* seja própria (Chen *et al.*, 1999). Assim como no Capítulo 2 temos que a distribuição a *posteriori* para os parâmetros obtida não é tratável analiticamente, portanto métodos MCMC podem ser considerados. Como as condicionais não tem forma fechada geramos amostras das distribuição a *posteriori* dos parâmetros usando o algoritmos de Gibbs com Metropolis-Hasting da seguinte forma: geramos duas cadeias paralelas cada uma com 35.000 iterações com um *burn in* de 5.000 e saltos de tamanho 10, resultando uma amostra de Gibbs de tamanho 6.000. Para monitorar a convergência do amostrador de Gibbs utilizamos a aproximação desenvolvida por Gelman e Rubin(1992). Afim de avaliar a robustez do modelo relacionado às escolhas dos hiperparâmetros das *priori*, um estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas. A Tabela 3.4 apresenta os valores para os critérios de seleção B e DIC para os modelos ajustados. Ambos critérios indicam $J = 5$ como melhor partição do eixo dos tempos.

TABELA 3.4: Critérios de seleção de modelos de acordo com cada partição.

Modelo	Critério	
J	B	DIC
1	-536,5428	1068
2	-533,1140	1066
3	-525,3784	1051
4	-524,8778	1050
5	-517,3055	1035
6	-521,7236	1043

Na Tabela 3.5 apresentamos o resumo a *posteriori* dos parâmetros do modelo com o melhor ajuste dentre os considerados. Para todos os parâmetros observamos valores de $\hat{R}$ próximos a 1, indicando que as iterações foram suficientes para se chegar a convergência. A " highest posterior density" (HPD)(Chen & Shao, 1999) com 90% de credibilidade indica que somente a covariável, que representa a espessura do tumor é não significativa.

TABELA 3.5: Resumo a *posteriori* dos parâmetros do modelo.

Parâmetro	Média	Desvio Padrão	$HPD(90\%)$	$\hat{R}$
α	1,1628	0,6455	(0,0305; 2,0301)	1,0181
β_0(intercepto)	-1,9439	0,5011	(-2,7961; -1,1754)	1,0044
β_1(cat. nódulo)	0,5500	0,1320	(0,3326; 0,7601)	1,0091
β_2(idade)	0,0143	0,0076	(0,0021; 0,0271)	1,0068
β_3(espessura)	0,0326	0,0308	(-0,0208; 0,0799)	1,0022
λ_1	0,0869	0,0314	(0,0350; 0,1350)	1,0238
λ_2	0,2166	0,0663	(0,1038; 0,3183)	1,0045
λ_3	0,5312	0,1513	(0,2837; 0,7777)	1,0218
λ_4	0,4280	0,1280	(0,2257; 0,6366)	1,0044
λ_5	0,4255	0,1630	(0,1631; 0,6811)	1,0030

Com as amostras de Gibbs foram estimadas as medidas de divergência de K–L para cada uma das observações, esses resultados podem ser observados na Figura 3.3, em que destacamos as observações 23, 47, 176 e 330 que apresentaram maiores valores quando comparados com as demais observações. Para verificar se essas observações são influentes estimamos a calibração da medida de divergência K–L, essas estimativas são apresentados na Tabela 3.6 conjuntamente com as respectivas estimativas da divergência K–L ($K(P, P_{-i})$), podemos observar pelos valores da calibração que nenhuma das observações são influentes, pois deveríamos ter resultados similares aos apresentados na Tabela 3.3.

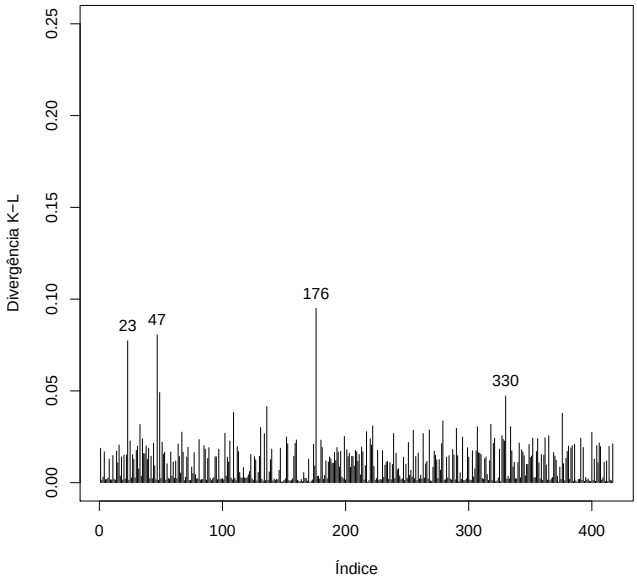


FIGURA 3.3: Gráfico de índices de $K(P, P_{-i})$ para dados de melanoma.

A título de ilustração realizamos uma análise de sensibilidade na estimativa dos parâmetros a *posteriori*, considerando as três observações com maiores valores de calibração.

TABELA 3.6: Identificação dos casos influentes para dados de melanoma.

Identificação do caso	Influência caso a caso	
Paciente	$K(P, P_{-i})$	Calibração
23	0,0773	0,6892
47	0,0804	0,6927
176	0,0950	0,7079
330	0,0472	0,6501

Na Tabela 3.7 apresentamos as novas estimativas dos parâmetros após a exclusão das observações "influentes", uma a uma ou excluindo todas de uma vez, representado por $I = \{23; 47; 176\}$, também calculamos as variações relativas (VR) (em porcentagem), as quais são obtidas por $VR_{\gamma_d} = |(\hat{\gamma}_d - \hat{\gamma}_{d(I)})/\hat{\gamma}_d| \times 100\%$, em que $\hat{\gamma}_{d(I)}$ denota a média a *posteriori* de γ_d, com $d = 1, \ldots, 10$ após o conjunto I de observações ter sido removido. A densidade a *posteriori* máxima (HPD) com 90% de credibilidade para cada nova estimativa é apresentada entre parênteses na Tabela 3.7.

TABELA 3.7: Estimativas, VR (em %) e a correspondente HPD (90%) ajustados para o conjunto de dados de melanoma.

Parâmetros	Observações descartadas			
	23	47	176	I
α	1,1257 3,1 (0,0482;2,0205)	1,1321 2,6 (0,0709;2,0923)	1,1943 2,7 (0,1309;2,1762)	0,9802 15,7 (0,0002;1,8020)
β_0(intercepto)	-1,9694 1,3 (-2,7262;-1,1558)	-1,9649 1,0 (-2,8593;-1,0755)	-1,9100 1,7 (-2,9123;-0,9971)	-1,8940 2,5 (-2,7013;-1,0281)
β_1(cat. nódulo)	0,5434 1,2 (0,3373;0,7368)	0,5567 1,2 (0,3342;0,7571)	0,5436 1,1 (0,3490;0,7381)	0,5094 7,3 (0,3102;0,7115)
β_2(idade)	0,0137 4,1 (0,0015;0,0255)	0,0139 2,7 (-0,0004;0,0264)	0,0133 6,9 (-0,0011;0,0262)	0,0127 11,1 (-0,0002;0,0253)
β_3(espessura)	0,0306 6,1 (-0,0179;0,0792)	0,0413 26,6 (-0,0128;0,0868)	0,0323 0,9 (-0,0205;0,0826)	0,0307 5,8 (-0,0139;0,0811)
λ_1	0,0992 14,1	0,0869 0,0	0,0939 8,0	0,1053 21,1
λ_2	0,2264 4,5	0,2184 0,8	0,2180 0,6	0,2436 12,4
λ_3	0,5560 4,6	0,5212 1,8	0,5307 0,0	0,5856 10,2
λ_4	0,4737 10,6	0,4108 4,0	0,4526 5,7	0,4839 13,0
λ_5	0,4954 16,4	0,4059 4,6	0,4695 10,3	0,5042 18,4
Critério B	-516,6873	-516,7518	-516,3218	-516,1163

Notamos maiores valores de VR para o parâmetro $\widehat{\beta}_3$ quando há exclusão da observação 47, o mesmo ocorre com o parâmetro $\widehat{\alpha}$ quando excluímos as observações do conjunto I, de maneira geral temos as maiores variações relativas (VR) nos parâmetros $\widehat{\lambda}$ quando excluímos o conjunto I. O parâmetro $\widehat{\beta}_2$ apresenta sensibilidade em relação as observações "influentes", pois quando excluímos as observações 47 ou 176 há alteração na significância do parâmetro, resumidamente, $\widehat{\beta}_2$ passa a ser não significativo no modelo. Observamos que a exclusão de qualquer observação praticamente não altera o valor do critério de selção B.

De acordo com o exposto, sugerimos como alternativa a um modelo final excluir apenas a covariável x_3 que é não significativa. Um resumo a *posteriori* dos parâmetros do modelo final é apresentado na Tabela 3.8.

TABELA 3.8: Resumo dos parâmetros a *posteriori* para o modelo final.

Parâmetro	Média	Desvio Padrão	HPD (90%)
α	1,1680	0,6801	(0,0107; 2,0978)
$\beta_{0(\text{intercepto})}$	-1,7930	0,4672	(-2,5175; -0,9859)
$\beta_{1(\text{cat. nódulo})}$	0,5242	0,1212	(0,3320; 0,7245)
$\beta_{2(\text{idade})}$	0,0150	0,0077	(0,0023; 0,0275)
λ_1	0,0878	0,0324	(0,0339; 0,1382)
λ_2	0,2176	0,0655	(0,1060; 0,3185)
λ_3	0,5267	0,1494	(0,2887; 0,7707)
λ_4	0,4268	0,1247	(0,2231; 0,6190)
λ_5	0,4291	0,1585	(0,1875; 0,6896)

Foram obtidos os valores dos critérios de seleção, B= -516,8008 e DIC= 1033, que comparados aos valores da Tabela 3.4 nos leva a atestar que este modelo é o modelo com melhor ajuste aos dados. Comparando os valores da Tabela 3.8 aos da Tabela 3.5 percebemos que maiores alterações nos resumos a *posteriori* dos parâmetros do modelo estão relacionados aos coeficientes $\boldsymbol{\beta}$.

Na Tabela 3.9 apresentamos as estimativas da proporção de curados por categoria de nódulo para o modelo final, as quais revelam que quanto maior a categoria do nódulo menor a proporção de curados.

Na Tabela 3.10 avaliamos a proporção de indivíduos curados em cada categoria de nódulo de acordo com a variação de suas idades. Consideramos os quantis 5%, 50% e 95% das idades que correspondem a 29, 47 e 70 anos respectivamente, notamos que com o aumento da idade a proporção de indivíduos curados diminui em cada categoria de nódulo. Podemos concluir que as covariáveis idade e categoria do nódulo influenciam na proporção de indivíduos curados.

TABELA 3.9: Resumo a *posteriori* para a fração de cura estratificada por categoria do nódulo.

Fração Cura			Percentil	
	Média	Desvio padrão	2,5%	97,5%
p_{0_1}	0,6459	0,0490	0,5367	0,7273
p_{0_2}	0,5267	0,0448	0,4249	0,5989
p_{0_3}	0,4046	0,0449	0,3039	0,4793
p_{0_4}	0,2940	0,0499	0,1860	0,3861

TABELA 3.10: Resumo a *posteriori* para a fração de cura (p_0) estratificada por categoria do nódulo e por idade.

				Percentil	
Idade	Fração de Cura	Média	Desvio Padrão	2,50%	97,50%
29	p_{0_1}	0,7016	0,0543	0,5822	0,7935
	p_{0_2}	0,5892	0,0549	0,4724	0,6849
	p_{0_3}	0,4672	0,0563	0,3507	0,5664
	p_{0_4}	0,3498	0,0595	0,2256	0,4632
47	p_{0_1}	0,6459	0,0490	0,5367	0,7273
	p_{0_2}	0,5267	0,0448	0,4249	0,5989
	p_{0_3}	0,4046	0,0449	0,3039	0,4793
	p_{0_4}	0,2940	0,0499	0,1860	0,3861
70	p_{0_1}	0,5679	0,0623	0,4384	0,6802
	p_{0_2}	0,4459	0,0576	0,3280	0,5532
	p_{0_3}	0,3302	0,0554	0,2168	0,4338
	p_{0_4}	0,2326	0,0556	0,1191	0,3407

A Figura 3.4 ilustra os resultados mostrados na Tabela 3.10. É fácil notar que com o aumento da idade dos pacientes e quanto maior a categoria do nódulo, menor a fração de cura. Há uma diferença significativa na proporção de curados entre a primeira categoria de nódulo e a última. Já entre a terceira e a quarta categoria não há grande variação da fração de cura.

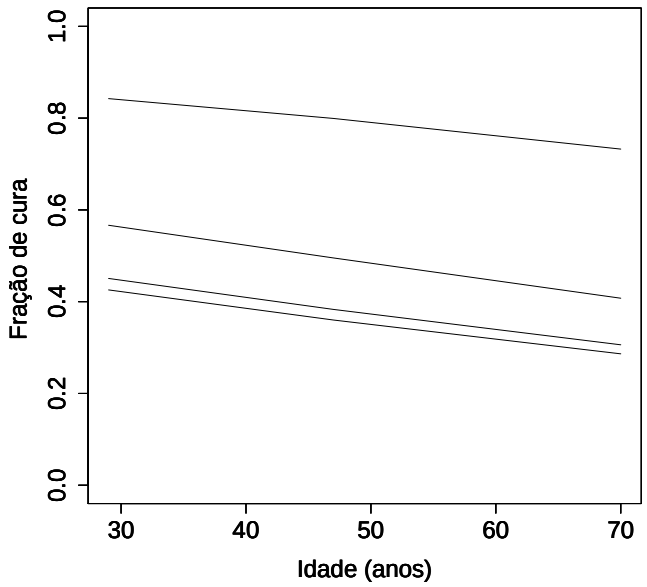


FIGURA 3.4: Proporção de curados por categoria nódulo (1 a 4, de cima para baixo).

3.5 Conclusões

Neste capítulo realizamos um estudo de simulação com o objetivo de avaliar a eficiência da divergência de Kullback-Leibler em detectar observações influentes na presença de proporção de cura. Consideramos um conjunto de dados de melanoma e realizamos uma análise de diagnóstico e de sensibilidade de possíveis observações influentes com o objetivo de avaliar se tais observações influenciam na estimativa dos parâmetros do modelo. Diante de tal análise propomos um modelo final, em que optamos por excluir a covariável x_3 a qual foi não significativa. Pela Tabela 3.7 temos que as covariáveis idade e categoria do nódulo são significativas, ou seja, há uma relação entre a idade do paciente e a classificação quanto ao tumor, percebemos que a fração de cura diminui em indíviduos mais velhos e com maior categoria de nódulo.

Capítulo 4

Um modelo destrutivo com fração de cura

Recentemente, Rodrigues *et al.* (2010b) propuseram o modelo destrutivo com fração de cura, considerando a distribuição Poisson ponderada, para modelar o número inicial de causas ou riscos relacionadas à ocorrência de um particular evento de interesse. Essa proposta inclui como caso particular os modelos de sobrevivência com fração de cura proposta por Yakovlev & Tsodikov (1996) (veja também Chen *et al.*, 1999) e permite estimar a proporção de causas latentes que não foram removidos pelo tratamento inicial. Seguindo Rodrigues *et al.* (2010b), propomos um modelo destrutivo com fração de cura, assumindo que o número inicial de causas (ou riscos) latentes relacionadas à ocorrência de um particular evento de interesse é modelado pela distribuição binomial negativa apresentada no Capítulo 3. Vale ressaltar que nos modelos destrutivos temos a inclusão de um parâmetro (p), que permite estimar a probabilidade do número de causas que não foram eliminadas por um tratamento inicial, por exemplo em um estudo de câncer, p é responsável por avaliar o número de células que não foram eliminadas ou destruídas. Os modelos destrutivos podem ser adequados para modelar qualquer tipo de dados de sobrevivência que apresente uma fração de sobreviventes nos mais variados contextos.

4.1 Formulação do modelo

Suponha que para um indivíduo da população, seja M uma variável aleatória discreta que representa o número de causas que competem para produzir um evento de interesse e suponha que M segue uma distribuição binomial negativa, com a função de probabilidade dada em (3.2). Suponha que dado $M = m$, sejam W_i, $i = 1, 2, \ldots, m$, variáveis aleatórias independentes de M, com distribuição de Bernoulli com probabilidade de sucesso p indicando a presença da i-ésima causa competitiva. Seja a variável N, que representa o número total de causas competindo dentre as M causas competitivas iniciais que não foram destruídas, definida como

$$N = \begin{cases} W_1 + W_2 + \cdots + W_M, & \text{se } M > 0, \\ 0, & \text{se } M = 0. \end{cases} \tag{4.1}$$

Por danificação ou destruição, queremos dizer que $N \leq M$. O tempo (aleatório) para a i-ésima causa produzir o evento de interesse é denotado por R_i, $i = 1, 2, \ldots$. Assumimos que condicionado a N as variáveis R_i são i.i.d. com função distribuição acumulada $F(t)$ e função de sobrevivência $S(t) = 1 - F(t)$. Também assumimos que $R_1, R_2, \ldots$ são independentes de N. O número total de causas competindo N e o tempo R_i são não observadas e podem ser interpretadas como variáveis latentes na formulação do modelo. Então, o tempo observado para a ocorrência do evento de interesse é definida por

$$T = \min\{R_1, R_2, \ldots, R_N\}$$

com $P(T = \infty | N = 0) = 1$, o que leva a uma proporção de indivíduos "curados". As distribuições exponencial, Weibull e exponencial por partes podem ser usadas para modelar as variáveis R_i (Ibrahim *et al.*, 2001), como ao longo desta dissertação neste capítulo também consideramos o modelo exponencial por partes para modelar as variáveis R_i. Como observado no início do Capítulo 3, Tsodikov *et al.* (2003) e Rodrigues *et al.*

(2008) demonstram que a função de sobrevivência populacional é dada por

$$S_{pop}(t) = P(T \geq t) = \phi(S(t)), \tag{4.2}$$

sendo $\phi(\cdot)$ é a função geradora de probabilidade para o número de causas competitivas N. Podemos mostrar que a função geradora de probabilidades para N é dada por $\phi(s) = [1 + \alpha\theta p(1-s)]^{-1/\alpha}$, assim a função de sobrevivência imprópria é

$$S_{\text{pop}}(t) = \phi(S(t)) = \{1 + \alpha\theta p F(t)\}^{-1/\alpha}, \tag{4.3}$$

se $p = 1$, $S_{pop}(t)$ em (4.3) se reduz ao modelo (3.5).

A proporção de curados é determinada por $p_0 = \lim_{t\to\infty} S_{\text{pop}}(t) = [1 + \alpha\theta p]^{\frac{-1}{\alpha}}$. E a correspondente função densidade de (4.3) é expressa por

$$\begin{aligned} f_{\text{pop}}(t) &= -\frac{d}{dt}\Big(S_{pop}(t)\Big) \\ &= \theta p f(t)\{1 + \alpha\theta p F(t)\}^{-(1+1/\alpha)}, \end{aligned} \tag{4.4}$$

com $f(t) = \frac{d}{dt}(F(t))$. Observe que $f_{\text{pop}}(t)$ não é uma função de densidade própria, já que $S_{\text{pop}}(t)$ é uma função de sobrevivência imprópria. A correspondente função de taxa de falha é dada por

$$h_{\text{pop}}(y) = \frac{\theta p f(t)}{1 + \theta\alpha p F(t)}. \tag{4.5}$$

4.2 Inferência bayesiana

Como já dissemos anteriormente, em análise de sobrevivência é comum a existência de dados que não sejam completamente observados e estejam sujeitos a censura à direita. Seja C_i o tempo de censura do i-ésimo indivíduo. Em uma amostra de tamanho n, observamos $Y_i = \min\{T_i, C_i\}$ e $\delta_i = \mathrm{I}(T_i \leq C_i)$, onde $\delta_i = 1$ se Y_i é o tempo de vida e $\delta_i = 0$ se é tempo de censura, para $i = 1, \ldots, n$. Seja $\boldsymbol{\gamma}$ o vetor de parâmetros da distribuição dos tempos não observados, $R_{i1}, \ldots, R_{i\,N_i}$. Note que o modelo destrutivo binomial negativo

(Seção 4.1) é não identificável no sentido Li *et al.* (2001). Para contornar esse problema, propomos relacionar os parâmetros p e θ do modelo as covariáveis $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$, respectivamente, sem elementos em comum e $\boldsymbol{x}_2$ sem a coluna de *uns*. Adotamos as funções de ligações

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x}_{1i}^\top \boldsymbol{\beta}_1 \quad \text{e} \quad \log(\theta_i) = \boldsymbol{x}_{2i}^\top \boldsymbol{\beta}_2, \tag{4.6}$$

$i = 1, \ldots, n$ com $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ denotando os vetores com k_1 e k_2 coeficientes.

A função de verossimilhança utilizada para fazermos inferências acerca dos parâmetros do modelo é expressa por

$$L(\boldsymbol{\vartheta}; \boldsymbol{\mathcal{D}}) \propto \prod_{i=1}^{n}\prod_{j=1}^{J} \left\{\theta_i p_i f_j(y_i; \boldsymbol{\lambda})\right\}^{\delta_i \nu_{ij}} \left\{1 + \alpha\theta_i p_i F_j(y_i; \boldsymbol{\lambda})\right\}^{-\nu_{ij}(\delta_i + 1/\alpha)} \tag{4.7}$$

com $\boldsymbol{\vartheta} = (\alpha, \boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\lambda}^\top)^\top$, $\boldsymbol{\mathcal{D}} = (\boldsymbol{\nu}, \boldsymbol{y}, \boldsymbol{\delta}, \boldsymbol{x_1}, \boldsymbol{x_2})$, $\boldsymbol{\nu} = (\nu_{11}, \ldots, \nu_{mJ})$ com $\nu_{ij} = 1$ se $s_{j-1} < y_i \leq s_j$ e $\nu_{ij} = 0$ caso contrário, $j = 1, \ldots, J$ e $i = 1, \ldots, m$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$, $\boldsymbol{x_1} = (x_{11}, \ldots, x_{1n})^\top$, $\boldsymbol{x_2} = (x_{21}, \ldots, x_{2n})^\top$. Enquanto que $f_j(y_i; \boldsymbol{\lambda})$ foi exposto na Equação (3.7) e $F_j(y_i; \boldsymbol{\lambda})$ na Equação (3.8).

Como já visto na Seção (3.2) aqui também assumimos *priori* independentes para os parâmetros do modelo. Assim a densidade a *priori* conjunta é expressa por

$$\pi(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\lambda}) = \pi(\alpha)\pi(\boldsymbol{\beta}_1)\pi(\boldsymbol{\beta}_2)\pi(\boldsymbol{\lambda})$$

em que o parâmetro α tem distribuição $G(a, b)$, $\boldsymbol{\lambda}_i \sim \pi(\boldsymbol{\lambda}_i)$, $i = 1, \ldots, J$, $\boldsymbol{\beta_1} \sim N(0, \sigma_1^2)$ e $\boldsymbol{\beta_2} \sim N(0, \sigma_2^2)$. Combinando essa densidade a *priori* com a verossimilhança (4.7), a densidade a *posteriori* conjunta é dada por

$$\pi(\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\lambda}|\boldsymbol{\mathcal{D}}) = L(\boldsymbol{\gamma}; \boldsymbol{\mathcal{D}})\pi(\alpha)\pi(\boldsymbol{\beta}_1)\pi(\boldsymbol{\beta}_2)\pi(\boldsymbol{\lambda}).$$

Como a densidade a *posteriori* conjunta não é uma densidade padrão usamos métodos de Monte Carlo via cadeias de Markov (MCMC), tais como o amostrador de Gibbs e Metropolis-Hasting. As densidades a *posteriori* condicionais para o amostrador de

Gibbs são expressas por

$$\pi(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \alpha, \boldsymbol{\lambda}, \mathcal{D}) \propto L(\boldsymbol{\gamma}; \mathcal{D})\pi(\boldsymbol{\beta}_1);$$

$$\pi(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1, \alpha, \boldsymbol{\lambda}, \mathcal{D}) \propto L(\boldsymbol{\gamma}; \mathcal{D})\pi(\boldsymbol{\beta}_2);$$

$$\pi(\boldsymbol{\lambda}|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \alpha, \mathcal{D}) \propto L(\boldsymbol{\gamma}; \mathcal{D})\pi(\boldsymbol{\lambda});$$

$$\pi(\alpha|\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\lambda}, \mathcal{D}) \propto L(\boldsymbol{\gamma}; \mathcal{D})\pi(\alpha).$$

Como as densidades a *posteriori* condicionais não possuem forma fechada usaremos o algoritmo de Metropolis-Hastings dentro do ciclo do algoritmo de Gibbs (Gilks *et al.*, 1996) para gerar amostras dos parâmetros envolvidos no modelo.

4.3 Aplicação

Analisamos aqui o mesmo conjunto de dados descritos no Capítulo 3. Ressaltando que consideramos as covariáveis categoria do nódulo (x_1) (1, n=111; 2, n=137; 3, n=87; 4, n=82), idade (x_2) (em anos; média = 48,0 e desvio padrão = 13,1) e espessura do tumor (x_3) (em mm; média = 3,9 e desvio padrão = 3,2). A categoria do nódulo 1-4, respectivamente, é codificada a partir do número de gânglios envolvidos na doença (0, 1, 2–3, e $\geq$4). Para fins ilustrativos, ligamos os parâmetros p em (4.6) a categoria do nódulo (x_1) e espessura do tumor (x_3) enquanto θ em (4.6) a idade (x_2).

Obtemos as estimativas bayesianas considerando as seguintes distribuições a *priori*: para cada componente de $\boldsymbol{\beta}$, assumimos uma distribuição a *priori* normal com média 0 e variância 10^3, para λ_j assumimos uma distribuição a *priori* gama , ou seja, $\lambda_j \sim G(a, b)$ com hiperparâmetros $a = 1$ e $b = 0,01$, $j = 1, \ldots, J$ e para o parâmetro α consideramos uma distribuição a *priori* informativa, $\alpha \sim G(1; 1)$ com a finalidade de garantir que a distribuição a *posteriori* seja própria (Chen *et al.*, 1999). Geramos amostras de Gibbs através do algoritmo de Gibbs com Metropolis-Hasting da seguinte forma: geramos duas cadeias paralelas cada uma com 30.000 iterações com um *burn in* de 5.000 e saltos de tamanho 10,

resultando uma amostra de Gibbs de tamanho 5.000. Para monitorar a convergência do amostrador de Gibbs utilizamos a aproximação desenvolvida por Gelman e Rubin(1992). Afim de avaliar a robustez do modelo relacionado às escolhas dos hiperparâmetros das *priori*, um estudo de sensibilidade foi realizado, no qual constatamos que as estimativas dos parâmetros a *posteriori* não apresentaram diferenças significativas. Utilizamos os critérios de seleção B e DIC (discutidos na Seção 1.1) para avaliar a melhor partição do eixo dos tempos. Na Tabela 4.1 apresentamos os valores para os critérios de seleção B e DIC para os modelos ajustados. Ambos critérios indicam $J = 5$ como melhor partição do eixo dos tempos.

TABELA 4.1: Critérios de seleção para o modelo destrutivo.

Modelo	Critério	
J	B	DIC
1	-536,5428	1068
2	-530,9640	1061
3	-526,5614	1051
4	-524,7432	1049
5	-516,1186	1032
6	-521,9981	1044

Na tabela 4.2 apresentamos os resumos a *posteriori* dos parâmetros do modelo destrutivo para a melhor partição. Para todos os parâmetros observamos valores de $\hat{R}$ próximos a 1, indicando que as iterações foram suficientes para se chegar a convergência. A " highest posterior density" (HPD)(Chen & Shao, 1999) com 90% de credibilidade indica que todas as covariáveis são significativas.

TABELA 4.2: Estimativas a *posteriori* para o modelo MEP destrutivo.

Parâmetro	Média	Desvio Padrão	$HPD(90\%)$	$\hat{R}$
α	1,5381	0,5552	(0,6215; 2,4064)	1,0664
β_{1_0}(intercepto)	-4,9180	1,6616	(-7,8063; -2,3406)	1,0314
β_{1_1}(cat. nódulo)	1,9339	0,7081	(0,8302; 3,0156)	1,0464
β_{1_3}(espessura)	0,3014	0,1685	(0,0325; 0,5681)	1,0119
β_{2_2}(idade)	0,0208	0,0067	(0,0097; 0,0317)	1,0437
λ_1	0,0767	0,0237	(0,0369; 0,1129)	1,0721
λ_2	0,2001	0,0541	(0,1132; 0,2680)	1,0824
λ_3	0,4922	0,1281	(0,2749; 0,6837)	1,0575
λ_4	0,4226	0,1088	(0,2421; 0,5919)	1,0326
λ_5	0,4328	0,1401	(0,2251; 0,6764)	1,0414

De acordo com as estimativas obtidas na Tabela 4.2, podemos notar que o sinal positivo do parâmetro $\widehat{\beta}_{2_2}$ indica que o número de causas competitivas aumentam ao passo que o a fração de cura diminui. Já para os parâmetros $\widehat{\beta}_{1_1} > 0$ e $\widehat{\beta}_{1_3} > 0$ temos que maiores valores para categoria do nódulo implica em estimativas menores para a fração de cura em ambos os casos.

Com as amostras de Gibbs foram estimadas as medidas de divergência de K–L para cada uma das 417 observações do conjunto de dados, esses resultados são graficados na Figura 4.1, em que destacamos as observações 23, 109, 176, 199, 356 e 376 por apresentar maiores valores quando comparados com as demais observações. Para verificar se essas observações são influentes procedemos como nos capítulos anteriores estimando a calibração da medida de divergência K–L, essas estimativas são apresentados na Tabela 4.3 conjuntamente com as respectivas estimativas da divergência K–L, realizamos uma análise de sensibilidade para avaliar se estas observações alteram as estimativas dos parâmetros do modelo, para tanto escolhemos as três observações com maiores valores de calibração.

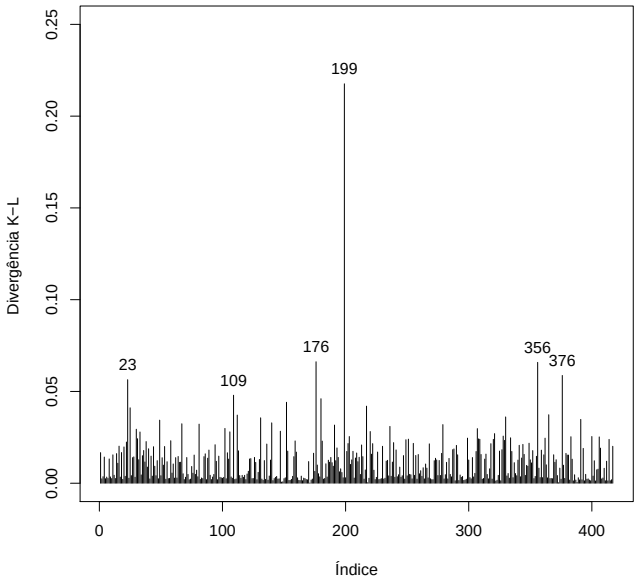


FIGURA 4.1: Gráfico de índices de $K(P, P_{-i})$ para dados de melanoma considerando o modelo destrutivo.

TABELA 4.3: Identificação dos casos influentes para dados de melanoma.

Identificação do caso	Influência caso a caso	
Paciente	$K(P, P_{-i})$	Calibração
23	0,0564	0,6633
109	0,0479	0,6511
176	0,0661	0,6760
199	0,2176	0,7970
356	0,0658	0,6756
376	0,0587	0,6664

Na Tabela 4.4 apresentamos as novas estimativas dos parâmetros após a exclusão das observações que apresentaram o maior valor de calibração, excluímos uma a uma ou todas de uma vez, representado pelo conjunto $I = \{176; 199; 356\}$, também calculamos as variações relativas (VR) (em porcentagem), as quais são obtidas por $VR_{\gamma_d} = |(\hat{\gamma}_d - \hat{\gamma}_{d(I)})/\hat{\gamma}_d| \times 100\%$, em que $\hat{\gamma}_{d(I)}$ denota a média a *posteriori* de γ_d, com $d = 1, \ldots, 10$ após o conjunto I de observações ter sido removido. A densidade a *posteriori* máxima (HPD) com 90% de credibilidade para cada nova estimativa é apresentada entre parênteses na Tabela 4.4.

TABELA 4.4: Estimativas, VR (em %) e a correspondente HPD (90%) ajustados para o conjunto de dados de melanoma.

Parâmetros	Observações descartadas			
	176	199	356	I
α	1,3205 14,1 (0,3218;2,1390)	1,3976 9,1 (0,5879;2,2163)	1,3945 9,3 (0,5447;2,2604)	1,6302 5,9 (0,4138;2,6432)
β_{1_0}(intercepto)	-4,0530 17,5 (-6,7572;-0,5416)	-5,6092 14,0 (-8,4381;-0,8271)	-5,0643 2,9 (-7,7219;-0,4346)	-6,2116 26,3 (-9,4406;-0,7053)
β_{1_1}(cat. nódulo)	1,6533 30,0 (0,3196;2,7913)	2,2807 17,9 (0,6146;3,6301)	2,0413 5,5 (0,4185;3,2709)	2,5115 29,8 (0,4506;4,0103)
β_{1_3}(espessura)	0,2372 21,3 (-0,0785;0,4990)	0,3841 27,4 (-0,0401;0,7055)	0,3234 7,2 (-0,0474;0,6140)	0,4159 37,9 (-0,0380;0,7578)
β_{2_2}(idade)	0,0170 18,2 (0,0048;0,0276)	0,0182 12,5 (0,0081;0,0279)	0,0185 11,0 (0,0071;0,0297)	0,0202 2,8 (0,0081;0,0330)
λ_1	0,0979 27,6	0,0852 11,0	0,0878 14,4	0,0787 2,6
λ_2	0,2218 10,8	0,2013 0,5	0,1995 0,2	0,2042 2,0
λ_3	0,5605 13,8	0,4906 0,3	0,5218 6,0	0,5304 7,7
λ_4	0,4728 11,8	0,4186 0,9	0,4413 4,4	0,4336 2,6
λ_5	0,5225 20,7	0,4491 3,7	0,4640 7,2	0,5024 16,0
Critério B	-516,8554	-516,1045	-516,4568	-506,8090

Observando a Tabela 4.4 percebemos maiores valores de VR para o parâmetro $\widehat{\beta}_{1_3}$ quando há exclusão da observação influente 199 e do conjunto I, o que também ocorre com o parâmetro $\widehat{\beta}_{1_1}$ quando excluímos a observação 176, com relação aos parâmetros $\widehat{\lambda}$ temos as maiores variações relativas (VR) quando excluímos a observação 176. O parâmetro $\widehat{\beta}_{1_3}$ apresenta sensibilidade em relação as observações influentes, pois quando excluímos qualquer um dos casos ou todos, há alteração na significância do parâmetro, resumidamente, $\widehat{\beta}_{1_3}$ passa a ser não significativo no modelo. Os demais parâmetros não apresentam mudanças significativas.

Sugerimos como alternativa a um modelo final excluir o conjunto de observações I de acordo com o critério de seleção de modelos (B) (Tabela 4.4) e a covariável x_3, uma vez que ela passa a ser não significativa após a exclusão de observações influentes. Um resumo a *posteriori* dos parâmetros do modelo final é apresentado na Tabela 4.5.

TABELA 4.5: Resumo dos parâmetros a *posteriori* para o modelo final.

Parâmetro	Média	Desvio Padrão	HPD (90%)
α	1,5449	0,6448	(0,4987; 2,5224)
β_{1_0}(intercepto)	-2,6957	0,5930	(-3,6777; -1,7810)
β_{1_1}(cat. nódulo)	1,1031	0,3710	(0,5724; 1,6613)
β_{2_2}(idade)	0,0233	0,0074	(0,0111; 0,0353)
λ_1	0,0854	0,0280	(0,0391; 0,1286)
λ_2	0,2197	0,0611	(0,1224; 0,3210)
λ_3	0,5650	0,1440	(0,3235; 0,8024)
λ_4	0,4532	0,1149	(0,2631; 0,6318)
λ_5	0,5362	0,1767	(0,2624; 0,8163)

Obtivemos os valores dos critérios de seleção, B= -509,6407 e DIC= 1019, ao compará-los aos valores da Tabela 4.1 temos que ao excluir as observações I e a covariável x_3, obtemos um modelo com melhor ajuste.

Assim como no Capítulo 3 as maiores alterações nos resumos a *posteriori* dos parâmetros do modelo estão relacionados aos coeficientes $\boldsymbol{\beta}$.

Na Tabela 4.6 apresentamos as estimativas baseadas em 5000 amostras da distribuição a *posteriori* dos parâmetros para o modelo final destrutivo, as quais revelam que quanto maior a categoria do nódulo maior a probabilidade de células não destruídas. A probabilidade da presença de causas competindo estratificada por categoria do nódulo em (4.6) é dada por $p_i = \exp(\beta_{1_0} + x_{1_i}\beta_{1_1})/\{1 + \exp(\beta_{1_0} + x_{1_i}\beta_{1_1})\}$. Notamos ainda ao nível de 5% uma diferença significativa entre a primeira e a última categoria do nódulo.

TABELA 4.6: Resumo a *posteriori* para a probabilidade da presença de causas competindo (p) estratificada por categoria do nódulo.

Probabilidade			Percentil	
	Média	Desvio padrão	2,5%	97,5%
p_1	0,1776	0,0630	0,0783	0,3241
p_2	0,3861	0,1236	0,1877	0,6745
p_3	0,6269	0,1559	0,3281	0,9292
p_4	0,8016	0,1338	0,4871	0,9888

TABELA 4.7: Resumo a *posteriori* para a fração de cura (p_0) estratificada por categoria do nódulo e por idade.

				Percentil	
Idade	Fração de Cura	Média	Desvio Padrão	2,50%	97,50%
29	p_{0_1}	0,7640	0,0420	0,6771	0,8404
	p_{0_2}	0,6156	0,0427	0,5275	0,6991
	p_{0_3}	0,5085	0,0378	0,4361	0,5860
	p_{0_4}	0,4513	0,0357	0,3823	0,5203
47	p_{0_1}	0,6894	0,0439	0,6019	0,7711
	p_{0_2}	0,5280	0,0374	0,4482	0,5967
	p_{0_3}	0,4223	0,0366	0,3491	0,4906
	p_{0_4}	0,3691	0,0417	0,2867	0,4514
70	p_{0_1}	0,5814	0,0570	0,4694	0,6923
	p_{0_2}	0,4182	0,0440	0,3296	0,5033
	p_{0_3}	0,3234	0,0457	0,2350	0,4131
	p_{0_4}	0,2788	0,0522	0,1755	0,3838

A Figura 4.2 ilustra os resultados apresentados na Tabela 4.7, ou seja, a proporção de cura por categoria do nódulo de acordo com as idades de 29, 47, e 70 anos, que correspondem aos quantis como descrito no Capítulo 3. É fácil notar que com o aumento da idade e quanto maior a categoria do nódulo, menor a fração de cura. Há uma diferença significativa na proporção de curados entre a primeira categoria de nódulo e a última. Já entre a terceira e a quarta categoria não há grande variação da fração de cura.

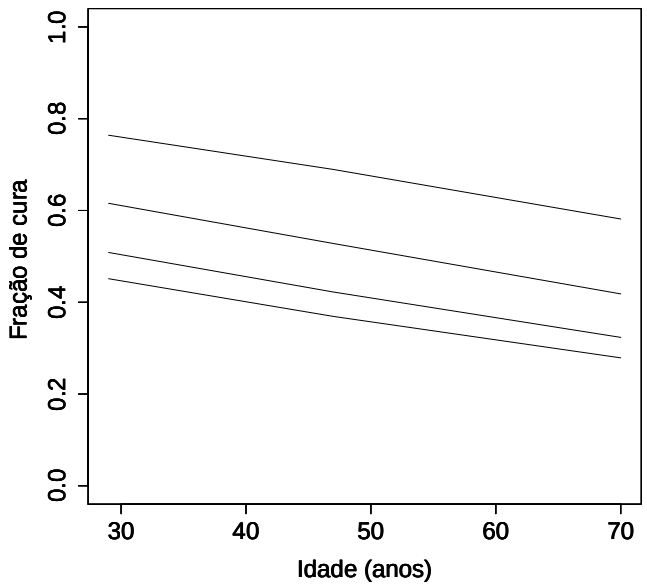


FIGURA 4.2: Proporção de curados por categoria nódulo (1 a 4, de cima para baixo).

4.4 Conclusões

Com relação ao exposto neste capítulo, de maneira geral, notamos que modelo proposto permite estimar a probabilidade da presença de causas competitivas após um tratamento inicial. Observamos que a probabilidade de sobrevivência diminui com a idade dos pacientes e que a fração de cura é mais baixa para pacientes com maior categoria de nódulo. Na aplicação do conjunto de dados referentes ao estudo de melanoma, observamos que o modelo destrutivo com fração de cura proporciona um melhor ajuste se comparado com o modelo abordado no Capítulo 3, o que podemos notar pelo critério de seleção de modelos.

Capítulo 5

Considerações Finais e Propostas Futuras

Com os estudos desenvolvidos ao longo deste texto, observamos que o uso do modelo exponencial por partes é adequado para estimar os tempos de vida, além de ser um modelo que apresenta grande flexibilidade ao acomodar funções de taxa de falha com diversas formas. Determinamos a melhor partição do eixo dos tempos de acordo com os principais critérios de seleção de modelos. Sendo esta uma questão fundamental para o processo de inferência.

Podemos observar que a divergência de Kullback-Leibler detecta os pontos influentes no modelo, o que pudemos comprovar por meio dos estudos de simulação e dos dados reais.

Com relação aos modelos destrutivos observamos que eles proporcionam um melhor ajuste (para o conjunto de dados considerado nesta dissertação), se comparados aos modelos com fração de cura, o que pudemos observar pelos critérios de seleção de modelos. Foi possível ainda, calcular a proporção de curados e estimar a probabilidade de causas competindo após um tratamento inicial, ou seja células que ainda não foram destruídas pelo tratamento.

Verificamos que o uso de métodos Bayesianos com técnicas de simulação de Monte

Carlo em Cadeias de Markov (MCMC), são apropriadas para a obtenção de sumários a *posteriori* de interesse.

A implementação computacional foi desenvolvida nos sistemas OpenBUGS (Spiegelhalter *et al.*, 2007) e R (R Development Core Team, 2010), para detalhes do programa pode ser consultado aos autores do trabalho

Como propostas futuras podemos realizar um estudo de diagnóstico embasado na abordagem clássica utilizando a técnica de influência local proposta por Cook (1986). Considerar os macanismos de ativação com base em fatores latentes encontrado em Cooner *et al.* (2007) utilizando tanto aboradagem clássica como bayesiana.

Referências

Aitkin, M., Laird, N. & Francis, B. (1983). Covariance analysis of censured survival data. *Journal of the American Statistical Association*, **78**, 264–292.

Barbosa, E., Colosimo, E. & Louzada-Neto, F. (1996). Accelerated life tests analyzed by a piecewise exponential distribution via generalized linear models. *IEEE Transactions on Reliability*, **45**, 619–623.

Barlow, R. & Campo, R. (1975). *Total Time on Test Processes and Applications to Failure Data Analysis*. California University Berkeley Operations Research Center.

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

Bernardo, J. & Smith, A. (2000). *Bayesian Theory*. John Wiley & Sons.

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

Breslow, N. (1974). Covariance analysis of censured survival data. *Biometrics*, **30**, 34–41.

Brookmeyer, R. & Goedert, J. J. (1989). Censoring in an epidemic with an application to hemophilia-associated aids. *Biometrics*, **45**, 325–335.

Cancho, V., Ortega, E. & Bolfarine, H. (2009). The Log-exponentiated-Weibull Regression Models with Cure Rate: Local Influence and Residual Analysis. *Journal of Data Science*, **7**, 433–458.

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

Cancho, V. G., Rodrugues, J. & de Castro, M. (2011). A flexive model for survival with a cure rate: A bayesian approach. *Journal of Applied Statistics*, **38**, 57–70.

Carrasco, J. M. F. (2007). *Modelo de Regressão Log-Weibull Modificado e a Nova Distribuição Weibull Modificada Generalizada*. Dissertação de mestrado, Universidade de São Paulo - Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba.

Chen, J., Ayyagari, R., Chatterjee, N., Pee, D., Schairer, C., Byrne, C., Benichou, J. & Gail, M. H. (2008). Breast Cancer Relative Hazard Estimates From Case–Control and Cohort Designs With Missing Data on Mammographic Density. *Journal of the American Statistical Association*, **103**, 976–988.

Chen, M. & Shao, Q. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, **8**, 69–92.

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

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**, 909–919.

Chen, M. H., Shao, Q. M. & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. Springer-Verlag, New York.

Chen, M. H., Harrington, D. & Ibrahim, J. (2002). Bayesian cure rate models for malignant melanoma: a case–study of Eastern Cooperative Oncology Group trial E1690. *Applay Statistics*, **51**, 135–150.

Chi, Y. Y. & Ibrahim, J. G. (2007). Bayesian approaches to joint longitudinal and survival models accommodating both zero and nonzero cure fractions. *Statistica Sinica*, **17**, 445–462.

Cho, H., Ibrahim, J. G., Sinha, D. & Shu, H. (2009). Bayesian case influence diagnostics for survival models. *Biometrics*, **65**, 116–124.

Clark, D. & Ryan, L. (2002). Concurrent prediction of hospital mortality and length of stay from risk factors on admission. *Health Services Research*, **37**, 631–645.

Colosimo, E. A. & Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Editora Edgard Blücher, Brasil.

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

Cook, R. & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall New York.

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

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Cox, D. R. & Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society*, **30**, 248–275.

de Castro, M., Cancho, V. & Rodrigues, J. (2009). A Bayesian Long-term Survival Model Parametrized in the Cured Fraction. *Biometrical Journal*, **51**, 443–455.

de Castro, M., Cancho, V. & Rodrigues, J. (2010). A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer methods and programs in biomedicine*, **97**(2), 168–177.

Demarqui, F., Loschi, R. & Colosimo, E. (2008). Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis*, **14**, 333–356.

Fonseca, R. (2009). *Modelos de sobrevivência com fração de cura e omissão nas covariáveis*. Tese de doutorado, Universidade Federal do Rio Grande do Norte.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, **10**, 101–113.

Gamerman, D. (1994). Bayes estimation of the piece-wise exponential distribution. *IEEE Transactions on Reliability*, **43**, 128–131.

Gilks, W., Gilks, W., Richardson, S. & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC.

Hoggart, C. J. & Griffin, J. E. (2001). A Bayesian partition model for customer attrition. In E. I. George, editor, *Bayesian Methods with Applications to Science, Policy, and Official Statistics (Selected Papers from ISBA 2000 – Creta, Greece)*, pages 61–70. International Society for Bayesian Analysis.

Ibrahim, J., Chen, M. & Sinha, D. (2001). *Bayesian survival analysis*. Springer Verlag.

Kim, J. & Proschan, F. (1991). Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability*, **40**, 134–139.

Kim, S., Chen, M. H., Dey, D. K. & Gamerman, D. (2007). Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Analysis*, **13**, 17–35.

Labra, F. V., Aoki, R. & Bolfarine, H. (2005). Local influence in null intercept measurement error regression under a student-t model. *Journal of Applied Statistics*, **32**, 723–740.

Lawless, J. & Lawless, J. (1982). *Statistical models and methods for lifetime data*. Wiley New York.

Li, C., Taylor, J. & Sy, J. (2001). Identifiability of cure models. *Statistics & Probability Letters*, **54**(4), 389–395.

Lopes, C. (2008). *Modelos de sobrevivência com fração de cura e efeitos aleatórios*. Tese de doutorado, Universidade de São Paulo.

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

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association*, **84**, 473–478.

McGilchrist, C. & Aisbett, C. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461–466.

Mizoi, M. & Bolfarine, H. (2007). Cure Rate Model with Measurement Error. *Communications in Statistics: Simulation and Computation*, **36**, 185–196.

Mizoi, M. F. (2004). *Influência local em modelos de sobrevivência com fração de cura*. Tese de doutorado, Instituto de Matemática e Estatística – Universidade de São Paulo, São Paulo.

Nelson, W. (1990). *Accelerated testing: statistical models, test plans and data analyses*. Wiley New York.

Ortega, E., Bolfarine, H. & Paula, G. (2003). Influence diagnostics in generalized log–gamma regression models. *Computational Statistics & Data Analysis*, **42**, 165–186.

Ortega, E., Cancho, V. & Paula, G. (2009). Generalized log-gamma regression models with cure fraction. *Lifetime Data Analysis*, **15**, 79–106.

Ortega, E. M. M., Cancho, V. G. & Bolfarine, H. (2006). Influence diagnostics in exponentiated-Weibull regression models with censored data. *SORT. Statistics and Operations Research Transactions*, **30**, 171–192.

Paes, A. (2007). *Uso de modelos com fração de cura na análise de dados de sobrevivência com omissão nas covariáveis*. Tese de doutorado, Universidade de São Paulo.

Paula, G. A. (2004). *Modelos de Regressão com Apoio Computacional*. Instituto de Matemática e Estatística - Universidade de São Paulo, Brasil.

Peng, F. & Dey, D. (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, **23**, 199–213.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**, 863–867.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reineke, D., Pohl, E. & Murdock Jr, W. (1999). Maintenance-policy cost-analysis for a series system withhighly-censored data. *IEEE Transactions on Reliability*, **48**, 413–419.

Rodrigues, J., Cancho, V. G. & de Castro, M. (2008). *Teoria Unificada de Análise de Sobrevivência*. ABE, Brasil.

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

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

Rodrigues, J., Cancho, V., Castro, M. & Balakrishnan, N. (2010a). A Bayesian destructive weighted Poisson cure rate model and an application to a cutaneous melanoma data. *Statistical Methods in Medical Research*.

Rodrigues, J., de Castro, M., Balakrishnan, N. & Cancho, V. (2010b). Destructive weighted Poisson cure rate models. *Lifetime Data Analysis*.

Schmidli, H., Bretz, F. & Racine-Poon, A. (2007). Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in medicine*, **26**, 4925–4938.

Sen, A. & Tan, F. (2008). Cure-rate estimation under Case-1 interval censoring. *Statistical Methodology*, **5**, 106–118.

Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**, 583–639.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2007). Openbugs: User manual, version 3.0. 2. *MRC Biostatistics Unit, Cambridge*.

Tournoud, M. & Ecochard, R. (2007). Application of the promotion time cure model with time-changing exposure to the study of HIV/AIDS and other infectious diseases. *Statistics in medicine*, **26**, 1008–1021.

Tsodikov, A. D., Ibrahim, J. G. & Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, **98**, 1063–1078.

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

Yamaguchi, K. (1992). Accelerated Failure-Time Regression Models with a Regression Model of Surviving Fraction: An Application to the Analysis of "Permanent Employment" in Japan. *Journal of the American Statistical Association*, **87**, 284–292.

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

Zaider, M., Zelefsky, M. J., Hanin, L. G., Tsodikov, A. D., Yakovlev, A. Y. & Leibel, S. A. (2001). A survival model for fractionated radiotherapy with an application to prostate cancer. *Physics in Medicine and Biology*, **46**, 2745–2758.