Modelos semiparamétricos de fração de cura para dados com censura intervalar

Julio Cezar Brettas da Costa

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientador: Prof[a]. Dr[a]. Gisela Tunes da Silva

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CNPq

São Paulo, janeiro de 2016

Modelos semiparamétricos
de fração de cura para dados
com censura intervalar

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 18/02/2016. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof[a]. Dr[a]. Gisela Tunes da Silva (orientadora) - IME-USP
- Prof. Dr. Antonio Carlos Pedroso de Lima - IME-USP
- Prof. Dr. Mario de Castro Andrade Filho - ICMC-USP

À minha querida mãe, Ivani.

Agradecimentos

Agradeço especialmente à minha professora e orientadora, Gisela Tunes, pelos seus conselhos, seu constante incentivo desde a graduação e principalmente a oportunidade de trabalho e estudo conjunto que me proporcionou.

Aos meus pais, Ivani e Valdir, por todo o suporte e dedicação que me prestaram durante estes anos, sem os quais não teria sido possível alcançar esta conquista.

Ao meu estimado amigo, Thiago Akira Ferreira, pelo apoio e companheirismo.

Ao professor Antonio Carlos Pedroso de Lima, pelo grande apoio e pelos recursos essenciais disponibilizados para este trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo financiamento deste projeto.

Resumo

Modelos de fração de cura compõem uma vasta subárea da análise de sobrevivência, apresentando grande aplicabilidade em estudos médicos. O uso deste tipo de modelo é adequado em situações tais que o pesquisador reconhece a existência de uma parcela da população não suscetível ao evento de interesse, consequentemente considerando a probabilidade de que o evento não ocorra. Embora a teoria encontre-se consolidada tratando-se de censuras à direita, a literatura de modelos de fração de cura carece de estudos que contemplem a estrutura de censura intervalar, incentivando os estudos apresentados neste trabalho. Três modelos semiparamétricos de fração de cura para este tipo de censura são aqui considerados para aplicações em conjuntos de dados reais e estudados por meio de simulações.

O primeiro modelo, apresentado por Liu e Shen (2009), trata-se de um modelo de tempo de promoção com estimação baseada em uma variação do algoritmo EM e faz uso de técnicas de otimização convexa em seu processo de maximização. O modelo proposto por Lam *et al.* (2013) considera um modelo semiparamétrico de Cox, modelando a fração de cura da população através de um efeito aleatório com distribuição Poisson composta, utilizando métodos de aumento de dados em conjunto com estimadores de máxima verossimilhança. Em Xiang *et al.* (2011), um modelo de mistura padrão é proposto adotando um modelo logístico para explicar a incidência e fazendo uso da estrutura de riscos proporcionais para os efeitos sobre o tempo. Os dois últimos modelos mencionados possuem extensões para dados agrupados, utilizadas nas aplicações deste trabalho.

Uma das principais motivações desta dissertação consiste em um estudo conduzido por pesquisadores da Fundação Pró-Sangue, em São Paulo - SP, cujo interesse reside em avaliar o tempo até a ocorrência de anemia em doadores de repetição por meio de avaliações periódicas do hematócrito, medido em cada visita ao hemocentro. A existência de uma parcela de doadores não suscetíveis à doença torna conveniente o uso dos modelos estudados. O segundo conjunto de dados analisado trata-se de um conjunto de observações periódicas de cervos de cauda branca equipados com rádio-colares. Tem-se como objetivo a avaliação do comportamento migratório dos animais no inverno para determinadas condições climáticas e geográficas, contemplando a possibilidade de os cervos não migrarem.

Um estudo comparativo entre os modelos propostos é realizado por meio de simulações, a fim de avaliar a robustez ao assumir-se determinadas especificações de cenário e fração de cura. Até onde sabemos, nenhum trabalho comparando os diferentes mecanismos de cura na presença de censura intervalar foi realizado até o presente momento.

Palavras-chave: análise de sobrevivência; censura intervalar; fração de cura; anemia; migração de cervos; simulações.

Abstract

Cure rate models define an vast sub-area of the survival analysis, presenting great applicability in medical studies. The use of this type of model is suitable in situations such that the researcher recognizes the existence of an non-susceptible part of the population to the event of interest, considering then the probability that such a event does not occur. Although the theory finds itself consolidated when considering right censoring, the literature of cure rate models lacks of interval censoring studies, encouraging then the studies presented in this work. Three semiparametric cure rate models for this type of censoring are considered here for real data analysis and then studied by means of simulations.

The first model, presented by Liu e Shen (2009), refers to a promotion time model with its estimation based on an EM algorithm variation and using convex optimization techniques for the maximization process. The model proposed by Lam *et al.* (2013) considers a Cox semiparametric model, modelling then the population cure fraction by an frailty distributed as an compound Poisson, used jointly with data augmentation methods and maximum likelihood estimators. In Xiang *et al.* (2011), an standard mixture cure rate model is proposed adopting an logistic model for explaining incidence and using proportional hazards structure for the effects over the time to event. The two last mentioned models have extensions for clustered data analysis and are used on the examples of applications of this work.

One of the main motivations of this dissertation consists on a study conducted by researches of Fundação Pró-Sangue, in São Paulo - SP, whose interest resides on evaluating the time until anaemia, occurring to recurrent donors, detected through periodic evaluations of the hematocrit, measured on each visit to the blood center. The existence of a non-susceptible portion of donors turns the use of the cure rate models convenient. The second analysed dataset consists on an set of periodic observations of radio collar equipped white tail deers. The goal here is the evaluation of when these animals migrate in the winter for specific weather and geographic conditions, contemplating the possibility that deer could not migrate.

A comparative study among the proposed models is realized using simulations, in order to assess the robustness when assuming determined specifications about scenario and cure fraction. As far as we know, no work has been done comparing different cure mechanisms in the presence of interval censoring data until the present moment.

Keywords: survival analysis, cure rate, interval censoring, anaemia, deer migration, simulations.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O estudo do tempo até a ocorrência de um determinado evento é objeto de interesse comum entre pesquisadores de diversas áreas podendo encontrar, entre estes, exemplos em economia, engenharia, sociologia, e principalmente em áreas médicas e biológicas. O ramo da estatística que tem como objetivo estudar o tempo de duração de processos até a ocorrência de um determinado evento é denominado análise de sobrevivência.

A análise de sobrevivência abrange modelos de regressão que contemplam a presença de observações incompletas a respeito do tempo de um processo, levando em consideração a informação parcial que estas podem oferecer, ocasionando então a obtenção de estimativas mais precisas. Neste contexto, tais observações faltantes são denominadas censuras, sendo comum a ocorrência destas em estudos de tempos de sobrevida.

Uma observação é dita censurada quando não é possível observar a ocorrência do evento de interesse, mas sabe-se que o evento não ocorreu no período de inspeção. Existem diversos tipos de estrutura de censura, sendo os mais comuns denominados "censura à direita", "censura à esquerda" e "censura intervalar". O primeiro tipo, com teoria melhor consolidada, apresenta a maior parte das aplicações encontradas na literatura, consistindo nos casos em que o evento estudado ocorre após o período de tempo observado, sem informação precisa a respeito do tempo de ocorrência além do conhecimento do mesmo superar o tempo do indivíduo no estudo. Para exemplos de aplicações da análise de sobrevivência e tipos de censura, tem-se como referências fundamentais Colosimo e Giolo (2006), Klein e Moeschberger (2003) e Ibrahim *et al.* (2001a), este último para o contexto bayesiano.

Neste trabalho em particular, volta-se o foco ao estudo de dados com censura intervalar, recomendando-se a consulta de Sun (2006) para uma abordagem mais aprofundada deste tipo de censura. Uma observação cujo tempo de evento é desconhecido mas sabe-se que pertence a um intervalo especificado, impedindo assim precisar o tempo em que o evento realmente ocorreu, é dita censurada neste intervalo. Conforme a literatura, dados com censura deste tipo são comuns na área médica, principalmente em estudos clínicos, nos quais observa-se o indivíduo por meio de visitas periódicas e consequentemente tendo ciência apenas do intervalo de tempo em que a doença ou a aparição de um sintoma ocorreu.

Usualmente, modelos de análise de sobrevivência supõem a ocorrência incondicional do evento: se o evento não ocorreu, justifica-se que o tempo de observação não foi suficientemente extenso. Embora tal suposição aplique-se à maioria dos exemplos práticos, para alguns casos surge naturalmente a suposição de que é possível que o evento não ocorra, independentemente do período em observação. Para lidar com a situação em que existe uma parcela composta por tais tipos de indivíduos, ora não suscetíveis, ora então curados após determinado tratamento, foram propostos na área os modelos de fração de cura (também conhecidos como modelos de longa duração). Mais detalhes destes modelos surgirão com o decorrer do texto.

Uma das principais motivações deste trabalho surge com um estudo conduzido por pesquisadores da Fundação Pró-Sangue, em São Paulo - SP, cujo interesse consiste em avaliar o tempo até a ocorrência de anemia em doadores de repetição, assim como avaliar suas chances de não a

desenvolverem. Por conta da anemia ocorrer entre os instantes de observação dados pelas consultas, a censura destes dados apresenta estrutura intervalar. No contexto médico, sabe-se que certos indivíduos não são suscetíveis à anemia, independentemente de quantas doações realizem, sendo os indivíduos assim caracterizados denominados superdoadores. Pesquisadores disponibilizaram dados de todas as doações de sangue realizadas no período de janeiro de 1996 a dezembro de 2006, posteriormente analisados neste trabalho. Mais detalhes sobre o conjunto em questão podem ser encontrados em Almeida *et al.* (2013) e Almeida *et al.* (2016).

Outro conjunto de dados, oferecido pelo professor John Fieberg da Universidade de Minnesota (EUA), é aqui apresentado e analisado. O conjunto é composto por observações periódicas de cervos (cauda branca) com o uso de rádio-colares em regiões de estudo em Minnesota. O objetivo do estudo é avaliar o tempo até a migração dos animais em períodos de inverno, contemplando também a possibilidade destes não migrarem devido à existência de uma parcela não migratória. Embora os dados possuam análises na literatura envolvendo fração de cura, a abordagem deste trabalho difere por apresentar o uso de diferentes mecanismos de fração de cura e o uso da estrutura de dados agrupados. O leitor interessado pode consultar Fieberg e DelGiudice (2008) e Fieberg *et al.* (2008) para mais informações a respeito do conjunto de dados e das análises anteriormente realizadas.

Tendo como objetivo a análise destes conjuntos de dados, três diferentes modelos de fração de cura para dados com censura intervalar são aplicados e discutidos. Mantendo o foco em modelos semiparamétricos devido à flexibilidade destes, foram utilizadas as especificações de mistura simples (Xiang *et al.*, 2011), fragilidade (Lam *et al.*, 2013) e tempo de promoção (Liu e Shen, 2009) para modelar a fração de cura. Para os dois primeiros, os autores propuseram extensões para dados agrupados, aqui apresentadas e aplicadas para o conjunto de dados de migração.

Por fim, um exaustivo estudo por meio de simulações é realizado a fim de avaliar a robustez dos modelos para dados gerados a partir de diferentes especificações. As comparações são realizadas em cenários de diferentes frações de cura, taxas de censura, tamanho de amostra e mecanismo de geração de dados. Além disso, estimativas não paramétricas são obtidas através do algoritmo de Turnbull (Turnbull, 1976), devidamente apresentado neste trabalho por conta de seu uso intermediário na estimação em dois dos algoritmos estudados.

As implementações em R dos modelos de tempo de promoção e fragilidade estão disponibilizadas no CRAN (*The Comprehensive R Archive Network*) através do pacote *intercure*, desenvolvido como parte deste trabalho. Para o estimador de Turnbull, foi utilizada a função *icfit* do pacote *interval*. O modelo de mistura padrão para dados com censura intervalar apresentado em Xiang *et al.* (2011) tem sua rotina computacional disponibilizada nos materiais suplementares do artigo original.

1.1 Revisão Bibliográfica

Existem diversas pesquisas na área médica com o intuito de analisar o tempo até a ocorrência de um evento específico, podendo este ser dado pela manifestação de uma doença ou mesmo a morte do indivíduo. Porém, há em alguns casos a possibilidade de um indivíduo não ser suscetível ao evento de interesse devido a uma possível cura obtida por um tratamento ou por conta de alguma particularidade da própria observação. A proporção de indivíduos que assim se caracterizam é denominada na literatura como fração de cura. Diversas pesquisas residem em obter estimativas da fração de cura de uma determinada população ou mensurar os efeitos de certas covariáveis sobre a mesma.

Dados em que a fração de cura está presente podem facilmente ser encontrados na literatura, conforme Peng e Dear (2000) e Lam *et al.* (2005), nos quais estuda-se a recorrência de tumores de câncer de mama, Conkin *et al.* (1992) em seus trabalhos com dados de doença de descompressão obtidos de experimentos da NASA (*National Aeronautics and Space Administration*) ou em estudos de incidência de melanoma (Chen *et al.*, 1999). É importante destacar que o termo "cura" pode estar associado não somente a contextos biológicos, abrangindo também a não suscetibilidade de eventos como o primeiro casamento (Aalen, 1992), estudo do tempo de desemprego e ocorrência do primeiro divórcio, entre outros.

O primeiro modelo estatístico para fração de cura foi desenvolvido por Boag (1949) e então modificado três anos depois por Berkson e Gage (1952). Tal modelo é conhecido como "modelo de mistura padrão" e pode ser escrito como:

$$S_{pop}(t) = \pi S^*(t) + 1 - \pi, \tag{1.1}$$

em que $S_{pop}(t)$ e $S^*(t)$ representam as funções de sobrevivência da população e dos indivíduos suscetíveis, respectivamente, e π denota a probabilidade de um indivíduo ser suscetível ao evento de interesse.

O modelo de mistura padrão foi amplamente utilizado durante anos em análise de sobrevivência, apresentando muitos estudos e aplicações até atualmente, com Ma (2010) e Kim e Jhun (2008), por exemplo, apresentando usos deste para dados com censura intervalar. Uma alternativa a este modelo surge com Chen *et al.* (1999) criticando alguns pontos do mesmo e apresentando o modelo originalmente desenvolvido por Yakovlev *et al.* (1993) chamado "modelo de tempo de promoção", também conhecido como modelo *BCH* (*bounded cumulative hazard model*). Nesta abordagem em particular, consideram-se N variáveis latentes independentes e identicamente distribuídas representando as possíveis causas da ocorrência do evento de interesse. Usualmente, toma-se $N \sim Poisson(\theta)$ por conta da garantia da propriedade de riscos proporcionais (Chen *et al.*, 1999). Conforme exemplificado em Rodrigues *et al.* (2008), as causas de ocorrência do evento podem ser o número de células defeituosas que causariam o tumor (Yakovlev *et al.*, 1993) ou o número de fatores que levariam um cliente a cancelar suas operações em um banco (Hoggart e Griffin, 2001). No modelo de tempo de promoção, o tempo observado até o evento é dado como sendo o mínimo dos N tempos latentes associados às causas, considerando-se um indivíduo não suscetível ou curado quando $N = 0$. Modelos de regressão utilizando esta modelagem para a proporção de curados podem ser encontrados em Chen *et al.* (1999), Ibrahim *et al.* (2001a), Ibrahim *et al.* (2001b) e Tsodikov *et al.* (2003). Mais recentemente em Yin e Nieto-Barajas (2009) é proposta, sob o paradigma bayesiano e fazendo uso do modelo *BCH*, uma modelagem que incorpora conjuntamente covariáveis de forma multiplicativa e aditiva para os efeitos do tempo de sobrevida e cura, respetivamente.

Abordagens menos populares encontram-se disponíveis para a modelagem da fração de cura, como a classe proposta por Aalen (1992) baseada no processo de Poisson composto, que atribui efeitos aleatórios multiplicativos à função de risco para explicar a heterogeneidade dos indivíduos. Extensões e aplicações desta são revisadas e exploradas por Lam *et al.* (2013), que faz uso de um modelo semiparamétrico de riscos proporcionais de Cox (Cox, 1972) com fragilidades, cujo trabalho é melhor explorado posteriormente neste texto por conta de sua extensão para dados com censura intervalar. Em Rodrigues *et al.* (2010) é apresentado, sob uma abordagem bayesiana, um modelo considerando-se o número de lesões ou células alteradas assumindo distribuição Poisson composta ponderada.

Também pode-se encontrar na literatura alguns trabalhos que unificam a teoria de diferentes modelos de fração de cura, como em Yin e Ibrahim (2005) ou Rodrigues *et al.* (2008). Estes criam classes nas quais o modelo de mistura padrão e o modelo de tempo de promoção são casos particulares, porém, a teoria destes estudos abrange somente casos de censura à direita.

Na prática, a escolha da modelagem da fração de cura é usualmente feita de acordo com a motivação que os dados proporcionam ou a conveniência das propriedades matemáticas do modelo selecionado, justificando a diversidade de abordagens para a estimação de tal proporção nos trabalhos existentes até então.

Dados com censura intervalar surgem com naturalidade em estudos médicos e biológicos, nos quais o tempo do evento de interesse não pode ser diretamente observado, sabendo-se apenas que reside em um intervalo obtido por meio de uma sequência de consultas.

No contexto geral, não necessariamente englobando fração de cura, estimadores não paramétricos da função de sobrevivência foram propostos proporcionando um grande avanço aos recursos descritivos e inferenciais para este tipo de estrutura. Turnbull (1976) propõe um estimador não paramétrico de máxima verossimilhança utilizando um algoritmo iterativo de autoconsistência apresentado posteriormente em detalhes nesta dissertação. Posteriormente, Gentleman e Geyer

(1994) propõem condições para avaliar a unicidade da estimativa, além de condições alternativas para avaliar as estimativas como sendo ou não de máxima verossimilhança. Colosimo e Giolo (2006) apresentam uma versão modificada deste utilizando o estimador produto-limite de Kaplan-Meier. Em Sun (2006), encontra-se o algoritmo de Turnbull como uma aplicação do algoritmo EM (Dempster *et al.*, 1977), assim sugerida pelo autor original, e um conjunto de estimadores e técnicas alternativos para este tipo de conjunto de dados. Em conjunto com o método de Turnbull, uma abordagem bayesiana para obtenção não paramétrica da curva de sobrevivência é dada por Gómez *et al.* (2004) utilizando-se o amostrador de Gibbs juntamente com processos de Dirichlet, porém, sem expressão analítica, sendo necessário métodos numéricos para estimação. Neste mesmo artigo, são apresentados estimadores paramétricos e testes não paramétricos para comparação de funções de sobrevivência. Ainda no contexto bayesiano, Zhou (2004) apresenta um estimador não paramétrico com forma analítica bem definida para a função de sobrevivência, este porém demanda um grande custo computacional.

Considerando-se dados com uma proporção de curados e censuras do tipo intervalar, em Aljawadi *et al.* (2012a) são apresentados estimadores paramétricos e não paramétricos da fração de cura populacional utilizando a modelagem de tempo de promoção sem a inclusão de covariáveis. As estimativas são obtidas por meio do algoritmo EM, com uma comparação da performance da estimação da proporção de curados exibida por meio de simulações.

Um modelo paramétrico de regressão é apresentado em Aljawadi *et al.* (2013) associando covariáveis ao parâmetro de escala de uma distribuição exponencial na presença de fração de cura sob o modelo de tempo de promoção para dados com censura intervalar. Neste modelo, a fração de cura é tomada como constante.

Ma (2010) atribui efeito de covariáveis à cura e à taxa de falha por meio do uso do modelo semiparamétrico de Cox em conjunto com o modelo de mistura padrão para a cura. A estimação da taxa de falha acumulada basal é dada pelo método do maior minorante convexo e, em conjunto com esta, obtém-se estimativas dos parâmetros através do método de Newton-Raphson. Propriedades inferenciais do estimador proposto são cuidadosamente estudadas neste artigo. Em Liu e Shen (2009), algoritmo estudado neste trabalho, encontra-se um modelo de regressão semiparamétrico que faz uso do modelo *BCH* para explicar a fração de curados em dados com censura intervalar. Uma alternativa a este último é apresentada em Hu e Xiang (2013), fazendo-se o uso de *splines*.

Utilizando o modelo de Cox, Kim e Jhun (2008) impõem fragilidades para cada indivíduo com o intuito de modelar a associação entre a probabilidade de cura e o tempo até a ocorrência do evento, utilizando estimativas parciais de uma distribuição exponencial por partes para a curva de risco basal. O artigo faz uso do modelo de mistura padrão e a estimação dos parâmetros é obtida maximizando-se uma adaptação da função de verossimilhança aproximada sugerida por Goetghebeur e Ryan (2000). Ainda com o modelo semiparamétrico de Cox, o trabalho de Lam *et al.* (2013) considera o efeito de covariáveis em fragilidades para cada indivíduo, modelando simultaneamente a cura com estas utilizando a distribuição Poisson composta. Algoritmos de múltipla imputação são utilizados para o processo de estimação neste caso. Lam e Wong (2014) apresentam uma extensão deste modelo incorporando efeito de grupo para os dados. Este trabalho apresenta um estudo do modelo de fragilidade para dados com censura intervalar sem agrupamentos por meio de simulações, exibindo também aplicações em conjuntos de dados reais, contemplando o efeito de grupo para um dos conjuntos em questão.

Para dados na presença de censura intervalar com a estrutura de medidas repetidas, encontra-se no trabalho de Xiang *et al.* (2011) o uso de modelos lineares generalizados mistos para a obtenção dos efeitos das covariáveis sobre a fração de cura e a taxa de falha na presença de efeitos aleatórios para grupos, supondo o modelo de riscos proporcionais de Cox e mistura padrão para cura. Este, em conjunto com seu caso particular para dados sem efeito de grupo, será revisado e utilizado nos trabalhos aqui apresentados. Ma e Li (2010) utilizam, também para dados agrupados/medidas repetidas e cura dada por mistura padrão, modelos paramétricos de locação-escala atribuindo diferentes efeitos aleatórios para os preditores lineares associados ao parâmetro de locação e de fração de cura, respectivamente.

No contexto bayesiano, Thompson e Chhikara (2003) estendem o modelo de fração de cura para dados com censura intervalar incorporando medidas repetidas em sua modelagem. Os trabalhos encontrados em Banerjee e Carlin (2004) apresentam um modelo paramétrico incluindo fragilidades para lidar com correlação espacial fazendo uso de distribuições *a priori* com estrutura autoregressiva.

1.2 Objetivos e Organização

Este trabalho tem como objetivo principal o estudo e a comparação de diferentes modelos de regressão semiparamétricos para dados com fração de cura e censura intervalar por meio de aplicações a dados reais e simulações.

Inicialmente, apresenta-se o algoritmo de Turnbull para estimação não paramétrica da curva de sobrevivência para dados com censura intervalar, introduzindo brevemente o conceito deste tipo de censura na notação provida por Turnbull (1976).

É então apresentado um modelo de mistura padrão proposto por Xiang *et al.* (2011) para censura do tipo intervalar. O estimador apresentado aborda técnicas de modelos lineares generalizados para o processo de estimação, trabalhando com modelos mistos para a inclusão de um efeito de grupo. A rotina computacional, disponibilizada pelos autores, é utilizada nas aplicações e simulações com mudanças superficiais sobre o código original.

Em seguida, é exibida a metodologia proposta por Liu e Shen (2009), familiarizando o leitor com a notação original e a construção da função de verossimilhança dos dados. A rotina faz alternadamente, em seu processo de estimação, o uso do algoritmo de Turnbull em conjunto com algoritmos primal-dual de pontos interiores e quase-Newton para a etapa de maximização. Diferente dos outros modelos semiparamétricos abordados neste estudo, esta proposta não possui extensão para dados agrupados.

O trabalho segue com a apresentação da proposta de Lam *et al.* (2013), em que se introduz um efeito aleatório de natureza multiplicativa ao risco de um indivíduo, modelando tal efeito através de uma distribuição de Poisson composta. Em trabalhos mais recentes, os autores do artigo original adaptam o algoritmo para a inclusão de efeitos de grupo (Lam e Wong, 2014), extensão utilizada em uma das aplicações desta dissertação.

Ao familiarizar o leitor com a teoria dos algoritmos mencionados, são apresentadas aplicações destes sobre o banco de dados de anemia proveniente da Fundação Pró-Sangue. Uma cuidadosa análise descritiva é realizada apresentando evidências a respeito da existência da fração de cura, reforçando o contexto biológico do problema. Aplicações diretas dos algoritmos descritos são realizadas proporcionando estimativas pontuais e intervalares dos efeitos das covariáveis consideradas no modelo, consequentemente proporcionando as frações de cura estimadas.

A seguir, uma análise dos dados de migração de cervos é apresentada utilizando os algoritmos propostos em conjunto com extensões para efeito de grupo, controlando assim a variabilidade das observações provenientes de um mesmo animal no decorrer dos anos. Resultados a respeito do tempo até a migração e da probabilidade de não migrar são obtidos e discutidos neste trabalho, corroborando as conclusões obtidas em trabalhos anteriores provenientes dos pesquisadores (Fieberg e DelGiudice, 2008).

Após a análise dos conjuntos de dados, segue-se um estudo por meio de simulações com o intuito de avaliar o comportamento dos modelos apresentados quando suas especificações não são as mesmas do real mecanismo de fração de cura. Conclusões gerais do trabalho são apresentadas no fim deste texto.

Capítulo 2

Métodos Estatísticos

2.1 Estimador Não Paramétrico de Turnbull

Uma das maneiras mais usuais na análise de tempos de sobrevida de obter-se uma estimativa da função de sobrevivência para dados censurados à direita é por meio do estimador não paramétrico de Kaplan-Meier (Kaplan e Meier, 1958). Na presença de censura intervalar, é comum observar o uso deste mesmo estimador considerando-se a média dos extremos do intervalo de censura como tempo observado, reduzindo assim o problema para um caso de censuras à direita apenas, e então utilizando o estimador de Kaplan-Meier. Essa prática, eventualmente adotada por pesquisadores sem aprofundamento na teoria de censura intervalar, pode levar a um viés nas estimativas e espera-se que tal viés aumente conforme a amplitude dos intervalos também aumente (Dorey *et al.*, 1993; Odell *et al.*, 1992; Rücker e Messerer, 1988).

É então apresentado nesta seção, como alternativa ao estimador produto-limite, o estimador não paramétrico de máxima verossimilhança proposto por Turnbull (1976) para a função de sobrevivência de um conjunto de dados na presença de censura e truncamento intervalares. Apesar de o interesse neste trabalho consistir apenas na análise de dados com censura intervalar, será descrita a metodologia de forma mais geral, contemplando a possibilidade de truncamento.

Para isto, consideram-se $X_1, \cdots, X_N$ variáveis aleatórias independentes $(i = 1, ...N)$ provenientes da função de distribuição $F(x) = P(X \leq x)$. Considere também X_i censurada no intervalo A_i. Desta forma, o conjunto de dados pode ser representado por seus intervalos de censura por meio de N observações como $A_1, \cdots, A_N$.

Dizemos que X_i é censurada em um intervalo se A_i tem a forma $[L_i, R_i]$. Observações exatas, censuras à direita ou esquerda podem facilmente ser incorporadas com a igualdade dos extremos ou com intervalos semifechados, respectivamente. Os tempos de observação que definem os intervalos de censura são adotados como fixos ou vindos de um mecanismo aleatório independente de X_i.

Com as suposições consideradas anteriormente, e sendo L_i e R_i os limites esquerdo e direito, respectivamente, da $i-$ésima censura intervalar, a função de verossimilhança é proporcional a

$$L^*(F) = \prod_{i=1}^{N} [F(R_i+) - F(L_i-)], \tag{2.1}$$

em que L_i- e R_i+ representam os instantes imediatamente anterior a L_i e posterior a R_i, respectivamente.

Define-se então um número finito de intervalos disjuntos da forma $\{[q_j, p_j]\}_{j=1}^{m}$ construídos da seguinte maneira: $q_j \in \{L_i : i = 1, \cdots, N\}$ e $p_j \in \{R_i : i = 1, \cdots, N\}$ de forma que o intervalo aberto (q_j, p_j) não contenha nenhum elemento de $\{L_i, R_i : i = 1, \cdots, N\}$ e $q_1 \leq p_1 < q_2 \leq p_2 < \cdots < q_m \leq p_m$. Define-se também $C = \bigcup_{j=1}^{m} [q_j, p_j]$.

Além disso, toma-se $s_j = F(p_j+) - F(q_j-)$ para $1 \leq j \leq m$, com $\sum_{j=1}^{m} s_j = 1$ e $s_j \geq 0$. Se imposto que as funções de distribuição F possuem valores constantes fora do conjunto C, então o conjunto de vetores $\mathbf{s} = (s_1, \cdots, s_m)$ define classes de equivalência neste espaço de funções. Desta

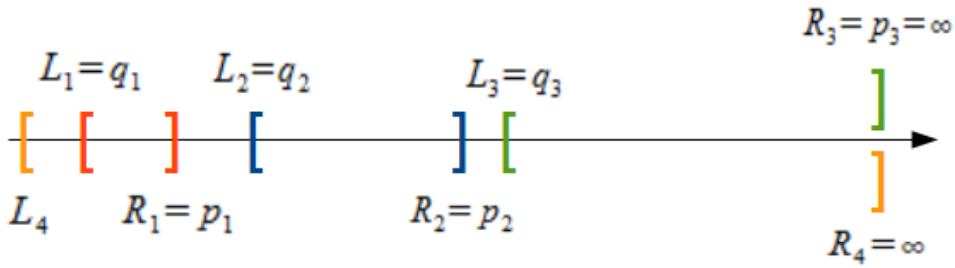

Figura 2.1: *Exemplo ilustrativo para intervalos de Turnbull.*

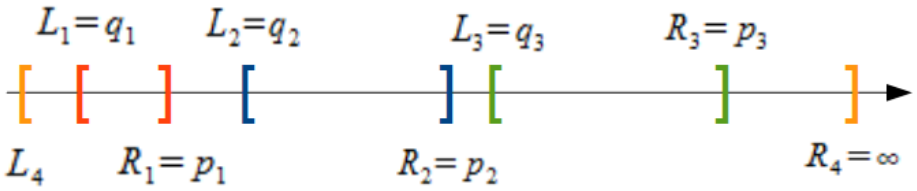

Figura 2.2: *Exemplo ilustrativo para intervalos de Turnbull sem "limiar de cura".*

forma, é dito que duas funções com o mesmo vetor $\mathbf{s}$ são equivalentes, pois possuem os mesmos valores fora de C e, de (2.1), tem-se que para qualquer comportamento da função neste conjunto, a função de verossimilhança é a mesma.

Baseado no exemplo ilustrativo apresentado em Liu e Shen (2009), a Figura 2.1 exibe a construção dos intervalos de Turnbull para um caso de quatro censuras intervalares. A figura evidencia a construção dos mesmos como sendo os intervalos fechados mais internos formados por elementos dos conjuntos $\{L_i : i = 1, \cdots, N\}$ e $\{R_i : i = 1, \cdots, N\}$ para os limites esquerdos e direitos, respectivamente.

A Figura 2.2, entretanto, apresenta o caso em que não há censura à direita tal que L_i, proveniente desta, supere o maior valor de R_j entre as censuras intervalares finitas $[L_j; R_j]$ contendo certamente a ocorrência de um evento. A situação assim construída não apresenta o denominado "limiar de cura" proposto em Zeng *et al.* (2006), utilizado em Liu e Shen (2009) e no trabalho de outros autores, implicando na impossibilidade de estimar a fração de cura para alguns estimadores propostos na literatura. Conforme exemplificado posteriormente, a ausência deste limiar implica na sobrevivência estimada pelo algoritmo de Turnbull tendendo a 0 quando $t \to \infty$. Aljawadi *et al.* (2012b) mostram que o estimador de Turnbull pode ser utilizado para estimar a proporção de curados em seu estudo comparativo com o estimador de Kaplan-Meier. No estudo aqui apresentado, utiliza-se o último dos saltos, s_m, para estimar a fração de cura associada a amostras sem "limiar de cura", exibindo desempenho razoável em aplicações e simulações.

Turnbull (1976) mostra que se pode restringir a busca por um estimador de máxima verossimilhança às classes de equivalência e que o estimador destas obtido é, com exceção de alguns casos triviais, único. Assim, reduz-se o problema de maximizar (2.1) à maximização de

$$L^*(s_1, \cdots, s_m) = \prod_{i=1}^{N} \sum_{j=1}^{m} \alpha_{ij} s_j, \tag{2.2}$$

sujeito a $\sum_{j=1}^{m} s_j = 1$ e $s_j \geq 0$, $(1 \leq j \leq m)$, em que:

$$\alpha_{ij} = \begin{cases} 1, & \text{se } [q_j, p_j] \subseteq A_i, \\ 0, & c.c. \end{cases} .$$

Pode-se notar que α_{ij} são variáveis indicadoras de que o intervalo $[q_j, p_j]$, construído a partir das classes de equivalência, está contido no intervalo de censura A_i. Para fins ilustrativos, os dados hipotéticos da Figura 2.1 apresentam $\alpha_{41} = \alpha_{42} = \alpha_{43} = 1$.

Com a notação e as devidas considerações apresentadas, pode-se então descrever o processo de obtenção do estimador de máxima verossimilhança de $\mathbf{s}$. O procedimento descrito a seguir pode ser visto como uma aplicação do algoritmo EM, entretanto, será aqui mantida a notação e metodologia do texto original.

Para $1 \leq i \leq N$, $1 \leq j \leq m$, tome $I_{ij} = 1$ se $x_i \in [q_j, p_j]$ e 0 caso contrário. Como o valor de I_{ij} pode não ser conhecido devido às censuras, fazemos uso da esperança dada por

$$E_s[I_{ij}] = \alpha_{ij} s_j / \sum_{k=1}^{m} \alpha_{ik} s_k = \mu_{ij}(\mathbf{s}). \tag{2.3}$$

Definida dessa forma, a quantidade $\mu_{ij}(\mathbf{s})$ é interpretada como a probabilidade de a i-ésima observação pertencer ao intervalo $[q_j, p_j]$ quando F pertence à classe de equivalência definida por $\mathbf{s} = (s_1, \cdots, s_m)$.

Tratando (2.3) como frequência observada ao invés de esperada, temos de imediato que a proporção de observações no intervalo $[q_j, p_j]$ é dada por

$$\sum_{i=1}^{N} \mu_{ij}(\mathbf{s}) / M(\mathbf{s}) = \pi_j(\mathbf{s}),$$

em que

$$M(\mathbf{s}) = \sum_{i=1}^{N} \sum_{j=1}^{m} \mu_{ij}(\mathbf{s}).$$

Um vetor de probabilidades $\mathbf{s}$ é denominado *autoconsistente* se

$$s_j = \pi_j(s_1, \cdots, s_m), \quad (1 \leq j \leq m). \tag{2.4}$$

A um vetor $\mathbf{s}$ que satisfaça a igualdade (2.4) dá-se o nome de *estimativa autoconsistente*. Tal igualdade fornece motivação ao seguinte processo iterativo para encontrar a solução:

1. Obtenha estimativas iniciais $s_j^0 (1 \leq j \leq m)$ de modo que $\sum_{j=1}^{m} s_j^0 = 1$ e $s_j^0 \geq 0$ para todo j.
2. Calcule $\mu_{ij}(\mathbf{s}^0)$ utilizando (2.3) para todo i, j. Com isso, calcule $M(\mathbf{s}^0)$ e $\pi_j(\mathbf{s}^0)$.
3. Atualize as estimativas fazendo:
$$s_j^1 = \pi_j(\mathbf{s}^0)$$
para todo j.
4. Repita o segundo passo utilizando as estimativas atualizadas.
5. Pare quando o critério de convergência estabelecido for satisfeito.

É possível mostrar que, satisfazendo certas condições, a solução obtida pelo processo iterativo é equivalente à solução da maximização da função de verossimilhança (2.2). Para isso, define-se

$$d_j(\mathbf{s}) = \sum_{i=1}^{N} \left\{ \frac{\alpha_{ij}}{\sum_{k=1}^{m} \alpha_{ik} s_k} - \frac{\beta_{ij}}{\sum_{k=1}^{m} \beta_{ik} s_k} \right\} \quad , 1 \leq j \leq m. \tag{2.5}$$

Pode ser mostrado que condições necessárias e suficientes para que $\mathbf{s}$, obtido através do processo iterativo, seja estimador de máxima verossimilhança são dadas por

$$d_j(\mathbf{s}) = 0 \quad ou \quad d_j(\mathbf{s}) \leq 0 \text{ com } s_j = 0, \tag{2.6}$$

para todo j.

Para a demonstração deste resultado, ou para questões referentes à construção das classes de equivalência e condições de identificabilidade, o leitor deve consultar Turnbull (1976).

Com a solução $\hat{\mathbf{s}}$ obtida, temos

$$\hat{F}(x) = \begin{cases} 0, & \text{se } x < q_1, \\ \hat{s}_1 + \hat{s}_2 + \cdots + \hat{s}_j, & \text{se } p_j < x < q_{j+1} \quad (1 \leq j \leq m-1), \\ 1, & \text{se } x > p_m. \end{cases}$$

Então, chega-se a

$$\hat{S}(x) = \begin{cases} 1, & \text{se } x < q_1, \\ 1 - \sum_{k=1}^{j} \hat{s}_k, & \text{se } p_j < x < q_{j+1} \quad (1 \leq j \leq m-1), \\ 0, & \text{se } x > p_m. \end{cases}$$

Note que a função estimada não possui valor definido em pontos pertencentes aos intervalos de equivalência $[q_j, p_j]$. Para contornar este problema, Aljawadi *et al.* (2012a) gera tempos de sobrevida aleatoriamente para cada intervalo de censura que contém um intervalo de equivalência, criando vetores de valores intermediários para a função de sobrevivência avaliada em pontos pertencentes aos intervalos de censura.

Colosimo e Giolo (2006) sugerem uma modificação do estimador de Turnbull considerando a ocorrência das observações no extremo direito do intervalo de equivalência, obtendo também com isso o número de indivíduos em risco até determinado tempo, aplicando então o estimador produto-limite de Kaplan-Meier em um processo iterativo. Entretanto, a restrição a classes de equivalência não representa grandes problemas quando o estimador é utilizado para fins descritivos ou mesmo para alguns casos inferenciais, conforme o uso em algoritmos posteriormente apresentados.

Além disso, Gentleman e Geyer (1994) apresentam condições simplificadas e de cálculo imediato para avaliação de um estimador assim obtido como sendo ou não um estimador de máxima verossimilhança, em conjunto com condições para avaliar sua unicidade. Tais resultados são contemplados pela função *icfit()* do pacote *interval*, distribuído no sistema R (R Core Team, 2015). Para detalhes destes, recomenda-se a leitura do artigo original.

2.2 Modelo de Mistura Padrão

Os modelos de mistura padrão apresentam muitos estudos na literatura considerando dados com tempos de falha exatos e censuras à direita. Entretanto, a literatura ainda apresenta-se escassa quando refere-se ao uso destes modelos sob a estrutura de censura intervalar. Neste contexto, Xiang *et al.* (2011) propõem modelos de fração de cura fazendo uso da teoria de modelos lineares generalizados mistos, incorporando efeito de grupo ao modelo por meio do uso de efeitos aleatórios nos preditores lineares relacionados aos tempos de sobrevida e proporção de cura. O efeito aleatório é adotado com o fim de controlar a variabilidade de indivíduos dentro de um mesmo grupo através de uma estrutura de correlação imposta, contemplada posteriormente na aplicação ao conjunto de dados de migração.

É apresentada a seguir a teoria proposta para dados agrupados e não agrupados encontrada em Xiang *et al.* (2011), com rotina computacional de ambos os estimadores disponibilizadas nos materiais suplementares do artigo. Conforme mencionado no artigo original, a identificabilidade para este modelo não está demonstrada e problemas na estimação podem ocorrer por conta disso. Os autores destacam a necessidade de boa evidência empírica a respeito da existência da fração

de cura, em conjunto com fatores como uma amostra de tamanho grande, um extenso tempo de observação e baixa proporção de censuras para mitigar a possibilidade de problemas no processo de estimação.

2.2.1 Modelo para Dados Não Agrupados

O conjunto de dados observados é denotado por $(A_i, \boldsymbol{x_i}, \delta_i), i = 1, \cdots, n$ tal que $A_i = [L_i, R_i]$ é o intervalo no qual a falha do indivíduo i ocorre, $\boldsymbol{x_i}$ é o vetor de covariáveis de dimensão p e $\delta_i = I(R_i < \infty)$ sendo variável indicadora de falha sob censura intervalar finita. Seja $S(t; \boldsymbol{x})$ a função de sobrevivência avaliada no tempo t para um indivíduo com vetor de covariáveis $\boldsymbol{x}$. Então,

$$L \propto \prod_{i=1}^{N} \{S(L_i, \boldsymbol{x_i}) - S(R_i, \boldsymbol{x_i})\} \tag{2.7}$$

assumindo-se $0 \leq L_i < R_i \leq \infty$, $\forall i$, implicando na ausência de instantes exatos de falha.

Define-se então a variável latente indicadora de suscetibilidade Y_i, assumindo 1 para o caso de o indivíduo ser suscetível e 0 para o caso em que este é curado. A expressão para a função de sobrevivência populacional é então dada por

$$S_{pop}(t; \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) S(t; \boldsymbol{x}_i) + 1 - \pi(\boldsymbol{x}_i), \tag{2.8}$$

em que $\pi(\boldsymbol{x}_i)$ denota $P(Y_i = 1 | \boldsymbol{x}_i)$. Modela-se tal probabilidade condicional por meio da função de ligação logito,

$$\pi_i = \pi(x_i) = \frac{\exp(\xi_i)}{1 + \exp(\xi_i)}, \tag{2.9}$$

em que $\xi_i = \boldsymbol{w}_i' \boldsymbol{b}$, $\boldsymbol{w}_i = (1, \boldsymbol{x}_i')'$ e $\boldsymbol{b}$ é o vetor de coeficientes associados à probabilidade de ser suscetível. Assume-se também a estrutura de riscos proporcionais para os indivíduos suscetíveis, implicando em

$$\lambda(t; \boldsymbol{x}_i) = \lambda_0(t) \exp(\eta_i) \quad \text{e} \quad S(t; \boldsymbol{x}_i) = S_0(t)^{\exp(\eta_i)}, \tag{2.10}$$

com $\eta_i = \boldsymbol{x}_i' \boldsymbol{\beta}$ sendo o preditor linear e $\boldsymbol{\beta}$ vetor $p-$dimensional correspondente ao efeito das covariáveis $\boldsymbol{x}_i$ sobre o componente de sobrevivência do modelo. $\lambda_0(t)$ e $S_0(t)$ denotam as funções de risco e sobrevivência basais do modelo no instante t, respectivamente. Definida a notação acima, a função de log-verossimilança do modelo de mistura padrão para dados com censura intervalar, a menos de uma constante, pode então ser expressa por

$$\log L = \sum_{i=1}^{N} \left[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) + y_i \log \left\{ S_0(L_i)^{\exp(\eta_i)} - S_0(R_i)^{\exp(\eta_i)} \right\} \right]. \tag{2.11}$$

Através da expressão (2.11), nota-se que o processo de estimação depende da função basal $S_0(t)$, aqui assumida desconhecida por tratar-se de um modelo semiparamétrico e, consequentemente, não levando à dependência de suposições a respeito da distribuição dos tempos de sobrevida. Faz-se então o uso do estimador não paramétrico de Turnbull, apresentado anteriormente, para modelar a componente de sobrevivência.

Sejam $0 = t_0 < t_1 < \cdots < t_Q = \infty$ os tempos ordenados distintos de todos os extremos dos intervalos $\{L_i, R_i;\ i = 1, \cdots, N\}$ com $\alpha_{iq} = I\left((t_{q-1}, t_q] \subset (L_i, R_i]\right)$ sendo a variável indicadora de que o evento do $i-$ésimo indivíduo que ocorreu no intervalo $(L_i, R_i]$ poderia ter ocorrido no instante t_q, em que $q = 1, \cdots, Q$. Para evitar restrições no processo de estimação, os autores modelam os saltos da função de sobrevivência por meio de $\gamma_q = \log[\log S_0(t_{q-1}) - \log S_0(t_q)]$. Com isso, pode-se

escrever a função de sobrevivência basal como

$$S_0(t_q) = \prod_{k=1}^{q} e^{-\exp(\gamma_k)} = \exp\left\{-\sum_{k=1}^{q}\exp(\gamma_k)\right\}, \quad q = 1, \cdots, Q, \tag{2.12}$$

em que $S_0(t_0) = 1$ se $\gamma_0 = -\infty$, e $S_0(t_Q) = 0$ se $\gamma_Q = \infty$. Pode-se então reescrever a função de log-verossimilhança em função dos parâmetros $\boldsymbol{\beta}, \boldsymbol{b}$, e $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{Q-1})'$:

$$\log L = \sum_{i=1}^{N} \left(y_i \log \pi_i + (1 - y_i) \log\left(1 - \pi_i\right) + P_i\right), \tag{2.13}$$

em que

$$P_i = y_i \log \sum_{q=1}^{Q} \alpha_i \left[\exp\left\{-\sum_{k=1}^{q-1}\exp\left(\gamma_k + \eta_i\right)\right\} - \exp\left\{-\sum_{k=1}^{q}\exp\left(\gamma_k + \eta_i\right)\right\}\right],$$

com $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$.

Da expressão acima, nota-se que o processo de estimação de $\boldsymbol{b}$ independe da estimação de $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, simplificando a obtenção das estimativas através da maximização de $l_C = l_{\boldsymbol{\beta}} + l_{\boldsymbol{b}}$ em que

$$l_{\boldsymbol{\beta}} = \sum_{i=1}^{N}\left(y_i \log \sum_{q=1}^{Q} \alpha_i \left[\exp\left\{-\sum_{k=1}^{q-1}\exp\left(\gamma_k + \eta_i\right)\right\} - \exp\left\{-\sum_{k=1}^{q}\exp\left(\gamma_k + \eta_i\right)\right\}\right]\right) \tag{2.14}$$

e

$$l_{\boldsymbol{b}} = \sum_{i=1}^{N}\left\{y_i \log \pi_i + (1 - y_i)\log\left(1 - \pi_i\right)\right\}. \tag{2.15}$$

A função de log-verossimilhança descrita acima depende das variáveis não observáveis y_i, fazendo-se necessário o uso de técnicas como o algoritmo EM. Para os valores y_i desconhecidos, obtém-se no passo E a esperança $y_i^{(r)} = E(y_i|\beta, b, \gamma)$ dos mesmos dada por

$$y_i^{(r)} = \delta_i + \frac{(1 - \delta_i)\,\pi_i \left(\sum_{q=1}^{Q} \alpha_{iq}\left[\exp\left\{-\sum_{k=1}^{q-1}\exp\left(\gamma_k + \eta_i\right)\right\} - \exp\left\{-\sum_{k=1}^{q}\exp\left(\gamma_k + \eta_i\right)\right\}\right]\right)}{1 - \pi_i + \pi_i \left(\sum_{q=1}^{Q} \alpha_{iq}\left[\exp\left\{-\sum_{k=1}^{q-1}\exp\left(\gamma_k + \eta_i\right)\right\} - \exp\left\{-\sum_{k=1}^{q}\exp\left(\gamma_k + \eta_i\right)\right\}\right]\right)}, \tag{2.16}$$

em que $\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\gamma}$ são obtidos da $(r-1)$-ésima iteração.

A demonstração da esperança acima, embora omitida, é obtida utilizando-se o teorema de Bayes, uma vez que tem-se disponível $P(Y_{ij} = 1|x_{ij})$ e componentes da verossimilhança (restrita à classe de equivalência) como $P(\alpha_{ijq}, \delta_{ij}|x_{ij}, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\gamma})$ para a obtenção de $P(Y_{ij}|\alpha_{ijq}, \delta_{ij}, x_{ij}, \boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\gamma})$ e, com isso, a esperança condicional desejada.

Após a obtenção dos valores $y_i^{(r)}$, o passo M do algoritmo EM resume-se à maximização de l_b e l_β, facilmente obtida com o uso da função optim() em R, concluindo assim uma iteração do processo de estimação. Procede-se então repetindo a imputação de y_i e maximizando as funções de verossimilhança dados os valores fixados das variáveis latentes até que se atinja a convergência das estimativas. Embora a convergência sugerida no artigo original se refira à convergência dos parâmetros, a rotina disponibilizada nos materiais suplementares avalia a convergência dos valores imputados y_i. As aplicações e simulações apresentadas neste trabalho consideram então a convergência segundo a rotina disponibilizada.

2.2.2 Modelo para Dados Agrupados

Conforme mencionado anteriormente, os autores estendem o algoritmo para contemplar dados agrupados, ou seja, dados que compartilham uma estrutura de correlação por pertencerem a um mesmo grupo. Para este caso, os intervalos de tempo de ocorrência dos eventos são denotados por $(A_{ij}, \boldsymbol{x}_{ij}, \delta_{ij})\,,\ j = 1, \cdots, n_i,\ i = 1, \cdots, M$, com $\sum_{i=1}^{M} n_i = N$, tal que $A_{ij} = [L_{ij}, R_{ij}]$ é o intervalo de tempo em que a falha do indivíduo j do grupo i ocorre, $\boldsymbol{x}_{ij}$ é o vetor de covariáveis de dimensão p e $\delta_{ij} = I\,(R_{ij} < \infty)$. De forma análoga ao caso sem agrupamentos, tem-se

$$L \propto \prod_{i=1}^{M} \prod_{j=1}^{n_i} \{S(L_{ij}, \boldsymbol{x}_{ij}) - S(R_{ij}, \boldsymbol{x}_{ij})\}\,, \tag{2.17}$$

com $S(t; \boldsymbol{x})$ sendo a função de sobrevivência avaliada em t para o conjunto de covariáveis $\boldsymbol{x}$. Assume-se aqui que $0 \leq L_{ij} < R_{ij} \leq \infty$ para todo $j = 1, \cdots, n_i; i = 1, \cdots, M$. Ou seja, não há observações exatas.

Deriva-se, de modo similar,

$$S_{pop}(t; \boldsymbol{x}_{ij}) = \pi(\boldsymbol{x}_{ij}) S(t; \boldsymbol{x}_{ij}) + 1 - \pi(\boldsymbol{x}_{ij}), \tag{2.18}$$

em que

$$\pi_{ij} = \pi_(\boldsymbol{x}_{ij}) = \frac{\exp(\xi_{ij})}{1 + \exp\left(\xi_{ij}\right)}, \tag{2.19}$$

com $\xi_{ij} = \boldsymbol{w'_{ij}} \boldsymbol{b}$, $w_{ij} = (1, x'_{ij})'$ e $\boldsymbol{b}$ sendo o vetor de coeficientes associados à probabilidade de suscetibilidade. Têm-se então as funções de risco e sobrevivência abaixo:

$$\lambda(t; x_{ij}) = \lambda_0(t) \exp(\eta_{ij}) \quad \text{e} \quad S(t; x_{ij}) = S_0(t)^{\exp(\eta_{ij})}, \tag{2.20}$$

com $\eta_{ij} = \boldsymbol{x}'_{ij}\beta$ sendo o preditor linear, com $\boldsymbol{\beta}$ $p-$dimensional, correspondente ao efeito sobre o componente de sobrevivência do modelo.

Assumindo mais uma vez Y_{ij} como variável latente denotando a não suscetibilidade do indivíduo, a função de log-verossimilhança do modelo de mistura padrão para dados agrupados com censura intervalar, a menos de uma constante, é então expressa por

$$\log L = \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left[y_{ij} \log \pi_{ij} + (1 - y_{ij}) \log(1 - \pi_{ij}) + y_{ij} \log \left\{ S_0(L_{ij})^{\exp(\eta_{ij})} - S_0(R_{ij})^{\exp(\eta_{ij})} \right\} \right]. \tag{2.21}$$

Introduz-se os efeitos de grupo $\boldsymbol{U} = (U_1, U_2, \cdots, U_M)$ e $\boldsymbol{V} = (V_1, V_2, \cdots, V_M)$ às componentes de sobrevivência e incidência através dos preditores lineares

$$\xi_{ij} = w'_{ij} b + V_i \tag{2.22}$$

e

$$\eta_{ij} = x'_{ij}\beta + U_i, \tag{2.23}$$

em que $\boldsymbol{U}$ e $\boldsymbol{V}$ são vetores não observáveis e independentes com distribuição dada por $N(0, \theta_1 \boldsymbol{I})$ e $N(0, \theta_2 \boldsymbol{I})$, respectivamente. O processo de estimação assume que os efeitos aleatórios incorporados satisfazem $\sum_i^M U_i = 0$ e $\sum_i^M V_i = 0$, conforme proposto em McGilchrist e Aisbett (1991).

Assim como para o modelo sem efeito de grupo, contempla-se aqui o estimador não paramétrico proposto por Turnbull para a estimação de $S_0(.)$ com pequenas alterações na notação. Definem-se como $0 = t_0 < t_1 < \cdots < t_Q = \infty$ os tempos ordenados distintos de todos os extremos dos intervalos $\{L_{ij}, R_{ij}; j = 1, \cdots, n_i, i = 1, \cdots, M\}$, com $\alpha_{ijq} = I\left((t_{q-1}, t_q] \subset (L_{ij}, R_{ij}]\right)$ tal que $q = 1, \cdots, Q$. A modelagem dos saltos através da transformação $\gamma_q = \log\left[\log S_0(t_{q-1}) - \log S_0(t_q)\right]$ também aplica-se

para este caso, obtendo-se assim

$$S_0(t_q) = \prod_{k=1}^{q} e^{-\exp(\gamma_k)} = \exp\left\{-\sum_{k=1}^{q}\exp(\gamma_k)\right\}, \quad q = 1, \cdots, Q. \tag{2.24}$$

Deste modo, fixados os efeitos aleatórios $\boldsymbol{U}$ e $\boldsymbol{V}$, a função de log-verossimilhança pode ser expressa em termos de $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{\beta}$, $\boldsymbol{b}$ e $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{Q-1})'$ como

$$l_1 = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(y_{ij}\log\pi_{ij} + (1-y_{ij})\log(1-\pi_{ij}) + Q_{ij}\right), \tag{2.25}$$

com

$$Q_{ij} = y_{ij}\log\sum_{q=1}^{Q}\alpha_{ijq}\left[\exp\left\{-\sum_{k=1}^{q-1}\exp(\gamma_k+\eta_{ij})\right\} - \exp\left\{-\sum_{k=1}^{q}\exp(\gamma_k+\eta_{ij})\right\}\right],$$

em que $\eta_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + U_i$.

Conforme McGilchrist (1993), os melhores preditores lineares não viesados (BLUPs) para ambos os efeitos fixos e aleatórios $\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{U}, \boldsymbol{V}$ são obtidos por meio da maximização de $l = l_1 + l_2$, com l_1 dado em (2.25) e

$$l_2 = -\frac{1}{2}\left\{M\log(2\pi\theta_1) + \frac{1}{\theta_1}\boldsymbol{U}'\boldsymbol{U}\right\} - \frac{1}{2}\left\{M\log(2\pi\theta_2) + \frac{1}{\theta_2}\boldsymbol{V}'\boldsymbol{V}\right\}.$$

A função de verossimilhança conjunta l é diretamente obtida da verossimilhança dos dados quando fixam-se os efeitos aleatórios $\boldsymbol{U}$ e $\boldsymbol{V}$ e soma-se à verossimilhança obtida da densidade dos efeitos aleatórios demonstrada em McGilchrist (1993), dada por l_2. Deste modo l pode ser vista como uma função de log-verossimilhança penalizada quando os efeitos aleatórios são condicionalmente fixos. Por conta da função l depender também da quantidade desconhecida y_i, faz-se necessário o uso do algoritmo EM para dados incompletos. O problema pode então ser expresso em termos da maximização da função de log-verossimilhança dos dados completos $l_C = l_{\boldsymbol{\beta}} + l_b$ em que

$$\begin{aligned} l_{\boldsymbol{\beta}} = &\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(y_{ij}\log\sum_{q=1}^{Q}\alpha_{ijq}\left[\exp\left\{-\sum_{k=1}^{q-1}\exp(\gamma_k+\eta_{ij})\right\} - \exp\left\{-\sum_{k=1}^{q}\exp(\gamma_k+\eta_{ij})\right\}\right]\right) \\ &- \frac{1}{2}\left\{M\log(2\pi\theta_1) + \frac{1}{\theta_1}\boldsymbol{U}'\boldsymbol{U}\right\} \end{aligned} \tag{2.26}$$

e

$$l_{\boldsymbol{b}} = \sum_{i=1}^{M}\sum_{j=1}^{n_i}\left\{y_{ij}\log\pi_{ij} + (1-y_{ij})\log(1-\pi_{ij})\right\} - \frac{1}{2}\left\{M\log(2\pi\theta_2) + \frac{1}{\theta_2}\boldsymbol{V}'\boldsymbol{V}\right\}. \tag{2.27}$$

Para lidar com a quantidade desconhecida y_i, calcula-se primeiramente sua esperança condicional com o restante dos parâmetros fixados. Esta rotina consiste no passo E do algoritmo EM, fixando-se para cada iteração os parâmetros $\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{U}, \boldsymbol{V}, \theta_1$ e θ_2 da iteração anterior. A esperança condicional de y_{ij}, obtida de forma análoga ao caso sem agrupamento, é dada por

$$\begin{aligned} y_{ij}^r &= E(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\gamma}, \boldsymbol{U}, \boldsymbol{V}, \theta_1, \theta_2) \\ &= \delta_{ij} + \frac{(1-\delta_{ij})\,\pi_{ij}\left(\sum_{q=1}^{Q}\alpha_{ijq}\left[\exp\left\{-\sum_{k=1}^{q-1}\exp(\gamma_k+\eta_{ij})\right\} - \exp\left\{-\sum_{k=1}^{q}\exp(\gamma_k+\eta_{ij})\right\}\right]\right)}{1-\pi_{ij}+\pi_{ij}\left(\sum_{q=1}^{Q}\alpha_{ijq}\left[\exp\left\{-\sum_{k=1}^{q-1}\exp(\gamma_k+\eta_{ij})\right\} - \exp\left\{-\sum_{k=1}^{q}\exp(\gamma_k+\eta_{ij})\right\}\right]\right)}. \end{aligned} \tag{2.28}$$

Uma vez obtidos e fixados os valores y_{ij}, realiza-se o passo M do algoritmo, consistindo na maximização de l_C em relação a $(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\gamma})$ com θ_1 e θ_2 fixados. Tal otimização, conforme o processo numérico descrito no Apêndice A.1, pode ser realizada para $l_{\boldsymbol{\beta}}$ e $l_{\boldsymbol{b}}$ separadamente utilizando o método de Newton-Raphson através de rotinas como *optim*() do pacote computacional R. Com as estimativas $(\hat{\boldsymbol{\beta}}^{(r)}, \hat{\boldsymbol{U}}^{(r)}, \hat{\boldsymbol{\gamma}}^{(r)})$ e $(\hat{\boldsymbol{b}}^{(r)}, \hat{\boldsymbol{V}}^{(r)})$ da r-ésima iteração, além de y_{ij}, estima-se então θ_1 e θ_2 com o uso de estimadores de máxima verossimilhança restrita, usualmente utilizados no ajuste de modelos mistos. A obtenção de $\hat{\theta}_1$ e $\hat{\theta}_2$ e a variância assintótica dos estimadores apresentados é determinada e vista com detalhes no apêndice de Xiang *et al.* (2011), baseado em resultados obtidos em McGilchrist (1993), fielmente retratado no Apêndice A.1 deste trabalho. Como na versão sem agrupamentos, a convergência é avaliada sobre os valores y_{ij}^r esperados por iteração.

2.3 Modelo de Tempo de Promoção

Conforme anteriormente mencionado, o modelo de tempo de promoção, também denominado BCH (*bounded cumulative hazard*) *model*, apresenta-se como opção na análise inferencial de dados com fração de curados em sua população. A construção e especificação de tal mecanismo de fração de cura, desconsiderando-se o tipo de censura, segue utilizando-se metodologia e motivação fiéis às apresentadas em Chen *et al.* (1999). As seções posteriores introduzem o leitor ao algoritmo proposto em Liu e Shen (2009), possibilitando a estimação da fração de cura por meio do modelo BCH para dados com censura do tipo intervalar. A motivação original do modelo é dada a seguir.

2.3.1 Motivação

Suponha que, para cada indivíduo na população, N denote o número de células tumorosas metástase-competitivas ativas remanescentes após algum tratamento inicial para este indivíduo. Uma célula assim definida é uma célula tumorosa com potencial de realização de metástase. Além disso, assume-se que N tem distribuição *Poisson* com parâmetro θ. Toma-se $Z_i, i = 1, 2, \cdots$, como sendo o tempo aleatório para a $i-$ésima célula produzir uma doença detectável a partir da metástase. Assim sendo, Z_i pode ser visto como tempo de promoção para a i-ésima célula tumorosa. Dado N, assume-se que as variáveis aleatórias $Z_i, i = 1, 2, \cdots$, são independentes e identicamente distribuídas com distribuição $F(y) = 1 - S(y)$ não dependente de N. O tempo até a recorrência do câncer pode então ser definido através da variável aleatória $Y = \min\{Z_i, 0 \leq i \leq N\}$, com $P(Z_0 = \infty) = 1$. A função de sobrevivência para Y é dita ser a função de sobrevivência populacional, pois aplica-se a toda a população em estudo, sendo dada por

$$\begin{aligned} S_{pop}(y) &= P(\text{sem câncer metástico até o tempo } y) \\ &= P(N = 0) + P(Z_1 > y, Z_2 > y, \cdots, Z_N > y, N \geq 1). \end{aligned}$$

Após algumas operações algébricas, conforme pode ser visto em Yakovlev *et al.* (1993), tem-se

$$\begin{aligned} S_{pop}(y) &= \exp(-\theta) + \sum_{k=1}^{\infty} S(y)^k \frac{\theta^k}{k!} \exp(-\theta) \\ &= \exp(-\theta + \theta S(y)) \\ &= \exp(-\theta F(y)). \end{aligned} \tag{2.29}$$

Da equação acima, é fácil observar que a fração de cura é dada por

$$S_{pop}(\infty) \equiv P(N = 0) = \exp(-\theta).$$

Conforme mencionado em Ibrahim *et al.* (2001a), o modelo de fração de cura assim definido possui a atrativa estrutura de riscos proporcionais, ausente no modelo de mistura padrão, com o risco populacional expresso como

$$\lambda_{pop}(y) = \theta f(y),$$

em que $f(y)$ é a função densidade de probabilidade avaliada em y obtida com $f(y) = F'(y)$.

Existe uma relação matemática entre o modelo BCH e o modelo de mistura padrão simples expressa a seguir e cuidadosamente estudada em trabalhos que visam a unificação da teoria de modelos de fração de cura (Rodrigues *et al.*, 2008; Yin e Ibrahim, 2005). Conforme explicitado em Tsodikov *et al.* (2003), a função de sobrevivência da população suscetível ("não curada"), $S^*_{pop}(y)$, é dada por

$$S^*_{pop}(y) = P(Y > y | N \geq 1) = \frac{\exp(-\theta F(y)) - \exp(-\theta)}{1 - \exp(-\theta)}.$$

A relação entre os modelos é descrita então como

$$S_{pop}(y) = \exp(-\theta) + (1 - \exp(-\theta)) S^*_{pop}(y). \tag{2.30}$$

Deste modo, $S_{pop}(y)$ pode ser visto como um modelo de mistura padrão com fração de cura igual a $1 - \pi = \exp(-\theta)$ e função de sobrevivência da população suscetível dada por $S^*_{pop}(y)$. A seguir, é apresentada neste trabalho a abordagem utilizada em Liu e Shen (2009) com respeito à inclusão de covariáveis no modelo BCH, assim como a forma da função de verossimilhança para a presença de censura intervalar com base nos trabalhos de Turnbull (1976).

2.3.2 Modelagem e Função de Verossimilhança

Em seu artigo, Liu e Shen (2009) consideram a análise de regressão de um conjunto de dados na presença de censura intervalar para um modelo semiparamétrico de tempo de promoção. Na notação deste, T denota a variável não negativa representando o tempo até o evento de interesse, com $\mathbf{Z}$ representando o vetor de covariáveis associadas. Assume-se então que

$$S(t|\mathbf{Z}) = P(T \geq t|\mathbf{Z}) = \exp(-e^{\alpha + \boldsymbol{\beta}'\mathbf{Z}} F(t)), \tag{2.31}$$

em que $(\alpha, \boldsymbol{\beta})$ são coeficientes de regressão (modelando a média do número de variáveis latentes N apresentadas na seção anterior) e F é uma função de distribuição não especificada. É fácil notar que a função de sobrevivência proposta é imprópria tomando-se o limite de t tendendo ao infinito. Liu e Shen (2009) mostram que o modelo (2.31) é identificável sob a condição de que F seja limitada por 1, conforme dado a seguir.

Mantendo a notação fiel à dos autores, faz-se E_0 como sendo a esperança sob os valores verdadeiros α_0, $\boldsymbol{\beta_0}$ e F_0 que satisfazem o modelo (2.31). Para demonstrar identificabilidade, assume-se que a esperança de Z condicionada ao parâmetros de regressão satisfaça $E_0\left[\exp(\boldsymbol{\beta_0'Z})\right] < \infty$ e $|\alpha_0| < \infty$. Assim como sua condicional, a sobrevivência marginal $P_0(T \geq t) = E_0\left[P_0(T \geq t|\boldsymbol{Z})\right]$ não é própria, conforme verifica-se pela desigualdade de Jensen:

$$\lim_{t\to\infty} P_0(T \geq t) = E_0\left[\exp(-\exp(\boldsymbol{\alpha_0} + \boldsymbol{\beta_0'Z}))\right] \geq \exp\{-E_0\left[\exp(\boldsymbol{\alpha_0} + \boldsymbol{\beta_0'Z})\right]\}.$$

Além disso, para garantir a identificabilidade do modelo, Liu e Shen (2009) tomam F como sendo uma função de distribuição usual com $\lim_{t\to\infty} F(t) = 1$. Considere então $(\alpha_*, \boldsymbol{\beta_*}, F_*)$ como sendo um conjunto de parâmetros que satisfaçam a relação (2.31), ou seja, $S_*(t|\mathbf{Z}) = \exp(-e^{\alpha_* + \boldsymbol{\beta}_*'\mathbf{Z}} F_*(t))$. Fixando-se ω em um conjunto de probabilidade 1 no espaço probabilístico, então

$$S_*(t|\boldsymbol{Z(\omega)}) = S(t|\boldsymbol{Z(\omega)}) \quad \Longrightarrow \quad e^{\alpha_* + \boldsymbol{\beta}_*'\boldsymbol{Z(\omega)}} F_*(t) = e^{\alpha + \boldsymbol{\beta}'\boldsymbol{Z(\omega)}} F(t) \quad \forall t \in (0, \infty).$$

Fazendo $t \to \infty$ implica $e^{\alpha_* + \boldsymbol{\beta}_*' \boldsymbol{Z}(\boldsymbol{\omega})} = e^{\alpha + \boldsymbol{\beta}' \boldsymbol{Z}(\boldsymbol{\omega})}$ e, consequentemente, $\alpha_* + \boldsymbol{\beta}_*' \boldsymbol{Z} = \alpha + \boldsymbol{\beta}' \boldsymbol{Z}$ com probabilidade 1. Suponha então, como uma condição adicional para identificabilidade, que se $P(\boldsymbol{b}' \boldsymbol{Z} = a) = 1$ para constante real a e para algum vetor real $\boldsymbol{b}$ respectivamente, então $a = 0$ e $\boldsymbol{b} = 0$, ou seja, o evento de um preditor linear $\boldsymbol{b}' \boldsymbol{Z}$ se igualar a a só é certo quando tem-se $a = 0$ e $\boldsymbol{b} = 0$, independente do conjunto de covariáveis. Na prática, o vetor de covariáveis Z é suficientemente "comportado" para garantir a validade da suposição. Como consequência da suposição adotada, tem-se $\alpha_* = \alpha$ e $\boldsymbol{\beta_*} = \boldsymbol{\beta}$. Logo, tem-se por consequência que $F_*(t) = F(t), \forall t \in (0, \infty)$.

Para a construção da verossimilhança, considera-se que o tempo até a ocorrência do evento de interesse para cada indivíduo seja desconhecido, mas contido no intervalo fechado $[L, R]$ de tal forma que L e R denotem os instantes de observação anterior e posterior, respectivamente, à ocorrência do evento (com $R = \infty$ no caso em que nenhum evento é observado até a última observação realizada). Assim sendo, os dados são denotados como $(L_i, R_i, \boldsymbol{Z_i})$ para cada indivíduo i $(i = 1, \cdots, n)$, assumindo-se independência e igualdade de distribuição condicionada ao vetor de covariáveis com $L_i < R_i$ para qualquer i, por construção.

A contribuição para a verossimilhança de um indivíduo cujo evento ocorre entre dois instantes de observação é dada pela probabilidade de ocorrência do evento neste intervalo:

$$P_{\boldsymbol{\theta},F}(L_i \leq T_i \leq R_i | \boldsymbol{Z_i}) = \exp(-e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} F(L_i-)) - \exp(-e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} F(R_i+)),$$

em que $\boldsymbol{\theta}' = (\alpha, \boldsymbol{\beta}')$, $\tilde{\boldsymbol{Z}}_i = (1, \boldsymbol{Z_i})$ e $P_{\boldsymbol{\theta},F}$ é medida de probabilidade em função dos parâmetros $\boldsymbol{\theta}$ e da função de distribuição F, tal que $\exp(-e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} F(t))$ é função de sobrevivência conforme (2.31), sendo esta contínua à esquerda e com limites à direita.

Se $R = \infty$, ou seja, o indivíduo é curado ou o evento não ocorreu até o último instante de observação, a contribuição para a verossimilhança é dada por

$$P_{\boldsymbol{\theta},F}(L_i \leq T_i \leq \infty | \boldsymbol{Z_i}) = \exp(-e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} F(L_i-)).$$

Logo, a função de verossimilhança para as n observações é dada por

$$L_n(\boldsymbol{\theta}, F) = \prod_{i=1}^{n} \left[e^{-\exp(\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i) F(L_i-)} - e^{-\exp(\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i) F(R_i+)} \right]^{\mathbf{1}(R_i < \infty)} \left[e^{-\exp(\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i) F(L_i-)} \right]^{\mathbf{1}(R_i = \infty)}. \tag{2.32}$$

Baseando-se na função de verossimilhança acima para $(\boldsymbol{\theta}, F)$, Liu e Shen (2009) apresentam um estimador não paramétrico para F restringindo a busca com o uso de classes de equivalência, extensão direta do trabalho de Turnbull (1976) para estimação não paramétrica na presença de censuras intervalares, conforme apresentado anteriormente. Assumindo-se a existência de uma proporção de curados, Liu e Shen (2009) definem um número finito de intervalos disjuntos $\{[s_j, r_j]\}_{j=1}^{m+1}$ construídos como se segue: $s_j \in \{L_i : i = 1, \cdots, n\}$ e $r_j \in \{R_i : i = 1, \cdots, n\}$ de modo que (s_j, r_j) não contenha membro algum de $\{L_i, R_i : i = 1, \cdots, n\}$, e $s_1 \leq r_1 < s_2 \leq r_2 < \cdots s_m \leq r_m < s_{m+1} < r_{m+1} = \infty$. Na construção dos autores, s_{m+1} denota o maior dos tempos de observação. De modo similar ao trabalho de Turnbull (1976), faz-se $\boldsymbol{C} = \bigcup_{j=1}^{m} [s_j, r_j]$ e do mesmo modo nota-se que, fixado $\boldsymbol{\theta}$ e um determinado conjunto de dados, a estimativa da função de distribuição F não pode ser crescente fora do conjunto $\boldsymbol{C}$. Examinando-se (2.32) verifica-se também que, fixando-se os valores $\{F(s_j-), F(r_j+)\}_{j=1}^{m}$, a função de verossimilhança é independente do comportamento de F em cada intervalo aberto (s_j, r_j). Faz-se então $p_j = F(r_j+) - F(s_j-)$ para $j = 1, \cdots, m$. Deste modo, o vetor $\boldsymbol{p} \equiv (p_1, p_2, \cdots, p_m)$, tal que $\sum_{j=1}^{m} p_j = 1$ e $p_j \geq 0$, define uma classe de equivalência de funções de distribuição F constantes fora do conjunto $\boldsymbol{C}$. Ou seja, como em Turnbull (1976), a busca por um estimador de máxima verossimilhança pode restringir-se à classe de equivalência composta de funções escada do tipo

$$t \mapsto \sum_{j=1}^{m} p_j \mathbf{1}(t \geq r_j),$$

com a restrição $\sum_{j=1}^{m} p_j = 1$, $0 \leq p_j \leq 1$.

Sem perda de generalidade, Liu e Shen (2009) trabalham ainda com funções da classe assim definida, tomando F contínua pela direita $F(t) = F(t+)$ e assumindo valor constante igual a $F(r_j)$ em $[r_j, s_{j+1}]$ para $j = 1, \cdots, m$, com $F(0) = 0$ e $F(r_m) = F(s_{m+1}) = 1$, com saltos p_j nos instantes imediatamente anteriores a r_j (para $j = 1, \cdots, m$). Deste modo,

$$S_{\boldsymbol{Z_i}}(t) = S_{\boldsymbol{\theta},F}(t|\boldsymbol{Z_i}) = \exp\left(-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}_{\boldsymbol{i}}} F(t)\right) = \exp\left[-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}_{\boldsymbol{i}}} \sum_{j=1}^{m} p_j \mathbf{1}(t \geq r_j)\right].$$

Da expressão acima tem-se, para $j = 1, \cdots, m$, que

$$S_{\boldsymbol{Z_i}}(s_j-) = \exp\left[-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}_{\boldsymbol{i}}}(p_0 + p_1 + \cdots + p_{j-1})\right]$$

e

$$S_{\boldsymbol{Z_i}}(r_j+) = \exp\left[-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}_{\boldsymbol{i}}}(p_0 + p_1 + \cdots + p_j)\right] = S_{\boldsymbol{Z_i}}(s_{j+1}),$$

em que $p_0 \equiv 0$ por conveniência de notação.

Com isso, nota-se que a verossimilhança dos dados depende somente do vetor de parâmetros $\boldsymbol{p}$ e $\boldsymbol{\theta}$. Tomando δ_{ij} como função indicadora assumindo 1 caso $[s_j, r_j]$ pertença a $[L_i, R_i]$ e, em particular, $\delta_{i,m+1} = 1$ indicando se a i-ésima pessoa completou o acompanhamento sem ocorrência do evento (ou seja, $R_i = \infty$) a função de log-verossimilhança pode então ser escrita como

$$l_n(\boldsymbol{\theta}, \boldsymbol{F}) \equiv l_n(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m} \delta_{ij}(S_{\boldsymbol{Z_i}}(s_j-) - S_{\boldsymbol{Z_i}}(r_j+)) + \delta_{i,m+1} S_{\boldsymbol{Z_i}}(s_{m+1})\right\}. \tag{2.33}$$

Por conta da suposição da existência de uma proporção de curados, assume-se que o banco de dados possui ao menos um indivíduo não suscetível ao evento de interesse ($R_i = \infty$) cujo último tempo de acompanhamento supere r_m. Assim sendo, o termo $\delta_{i,m+1} S_{\boldsymbol{Z_i}}(s_{m+1})$ é mantido na expressão (2.33). A contribuição para a fração de cura dá-se por meio dos indivíduos cujos eventos de interesse não foram observados e com $L_i > r_m$. Essa abordagem, comum em modelos de fração de cura, denomina-se "limiar de cura" (*cure threshold*) (Zeng *et al.*, 2006). É recomendado um tempo de acompanhamento suficientemente longo para que s_{m+1} seja grande o bastante fazendo com que alguns indivíduos sejam considerados curados de acordo com o contexto científico. Para mais detalhes da modelagem aqui apresentada e visualização de um exemplo de dados com essa estrutura, o leitor deve considerar a leitura do trabalho apresentado em Liu e Shen (2009).

2.3.3 Algoritmo Computacional

Na seção anterior, foi apresentada a função de log-verossimilhança dos dados (2.33). A maximização da função de log-verossimilhança implica na obtenção dos estimadores semiparamétricos de máxima verossimilhança denotados por $(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}, \hat{F}_n)$. O possível aumento de dimensão do vetor de parâmetros $\boldsymbol{p}$ a ser estimado traz dificuldades computacionais consideráveis. Tendo em vista estes potenciais problemas, Liu e Shen (2009) consideram o uso do algoritmo EM para dados com censura intervalar como uma extensão do método utilizado na ausência de covariáveis, apresentado em Kalbeisch e Prentice (2002). A etapa de maximização é composta por duas maximizações condicionais, consistindo na obtenção de estimativas perfiladas para $\boldsymbol{p}$ e $\boldsymbol{\theta}$. Tal substituição permite o uso de técnicas de maximização diferentes para cada vetor de parâmetros através da maximização condicional da função de log-verossimilhança esperada. Esta rotina consiste essencialmente no algoritmo apresentado em Meng e Rubin (1993), denominado algoritmo ECM. Entretanto, Liu e Shen (2009) propõem o uso de técnicas de otimização convexa, demonstrando com isso estabilidade numérica e eficiência no passo de maximização do algoritmo. A seguir, são apresentadas construções e propostas encontradas em Liu e Shen (2009), mantendo a notação dos autores e omitindo algumas

passagens apresentadas pelos mesmos em seus artigos originais, indicando-se ao leitor o artigo original para mais detalhes. Além disso, recomenda-se Boyd e Vandenberghe (2004) como referência fundamental para a teoria de otimização convexa, amplamente utilizada nos métodos de estimação que se seguem.

Para simplificação de notação em (2.33), define-se para $i = 1, \cdots, n$ e $j = 1, \cdots, m$,

$$\begin{aligned}\gamma_{i,j}(\boldsymbol{\theta}, \boldsymbol{p}) &\equiv S_{\boldsymbol{Z_i}}(s_j-) - S_{\boldsymbol{Z_i}}(r_j+) = S_{\boldsymbol{Z_i}}(s_j)\left[1 - \exp\left(-p_j e^{\boldsymbol{\theta' \tilde{Z}_i}}\right)\right] > 0,\\ \gamma_{i,m+1}(\boldsymbol{\theta}, \boldsymbol{p}) &\equiv \gamma_{i,m+1}(\boldsymbol{\theta}) = S_{\boldsymbol{Z_i}}(s_{m+1}) = \exp\left(-e^{\boldsymbol{\theta' \tilde{Z}_i}}\right).\end{aligned} \tag{2.34}$$

Então, a função de log-verossimilhança (2.33) pode ser escrita como

$$l_n(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=1}^{n} \log\left\{\sum_{j=1}^{m+1} \delta_{ij}\gamma_{ij}(\boldsymbol{\theta}, \boldsymbol{p})\right\}, \tag{2.35}$$

em que $\sum_{j=1}^{m+1} \gamma_{ij}(\boldsymbol{\theta}, \boldsymbol{p}) = 1$ para cada $i = 1, \cdots, n$.

Conforme Liu e Shen (2009), a equivalência de (2.32) com (2.33) permite que cada caso de censura intervalar seja então representado por $\boldsymbol{\Delta_i} = (\delta_{i,1}, \cdots, \delta_{i,m}, \delta_{i,m+1})$ para $i = 1, \cdots, n$. O vetor $\boldsymbol{\Delta_i}$ pode ser interpretado como dados incompletos provenientes de uma distribuição multinomial com índice 1. Considere agora a variável X_{ij} assumindo valor 1 se o i-ésimo tempo t_i pertence ao intervalo $[s_j, r_j]$ e 0 caso contrário, em que $i = 1, \cdots, n$ e $j = 1, \cdots, m+1$. Com X_{ij} assim definido, o conjunto $X = \{\boldsymbol{X_i} \equiv (X_{i,1}, \cdots, X_{i,m}, X_{i,m+1}), \boldsymbol{Z_i} : i = 1, \cdots, n\}$ pode ser visto como o conjunto de dados completos. Em outras palavras, $\boldsymbol{\Delta_i}$ pode ser visto como uma observação incompleta informando que o evento tenha ocorrido em uma das classes j contidas em $[L_i, R_i]$, mas não em qual delas, com a informação precisa contida apenas no vetor completo $\boldsymbol{X_i}$. Deste modo, condicionado à variável $\boldsymbol{Z_i}$, assume-se $\boldsymbol{X_i}$ como multinomial de classe $m+1$ e índice 1. Assim, para $j = 1, \cdots, m+1$, tem-se

$$P(X_{ij} = 1 | \boldsymbol{Z_i}) = \gamma_{ij}(\theta, \boldsymbol{p}) > 0. \tag{2.36}$$

Supondo a independência dos vetores $\boldsymbol{X_i}$, o logaritmo da função de verossimilhança dos dados completos seria então escrito como

$$l_{\boldsymbol{X}}(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=1}^{n}\left\{\sum_{j=1}^{m} X_{ij}\left[\log S_{\boldsymbol{Z_i}}(s_j) + \log\left(1 - \exp(-p_j e^{\boldsymbol{\theta' \tilde{Z}_i}})\right)\right] - X_{i,m+1} e^{\boldsymbol{\theta' \tilde{Z}_i}}\right\}. \tag{2.37}$$

Evidentemente, X_{ij} pode não ser observável, entretanto, é possível utilizar a esperança desta variável como feito em Turnbull (1976) na ausência de covariáveis. Dados $\{\boldsymbol{\Delta_i}, \boldsymbol{Z_i} : i = 1, \cdots, n\}$ e os parâmetros $(\boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}})$ da atual iteração k, obtém-se

$$X_{ij}^{(k)} \equiv E[X_{ij} | \boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}}, \boldsymbol{\delta_i}, \boldsymbol{Z_i}] = \frac{\delta_{ij}\gamma_{ij}(\boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}})}{\sum_{j'=1}^{m+1} \delta_{ij'}\gamma_{ij'}(\boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}})} \geq 0,\ j = 1, \cdots, m+1. \tag{2.38}$$

Deste modo, a função de log-verossimilhança esperada em relação ao conjunto X, definindo o passo E, fica dada por

$$\begin{aligned} l^{(k)}(\boldsymbol{\theta}, \boldsymbol{p}) &\equiv E(l_X(\boldsymbol{\theta}, \boldsymbol{p}) | \boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}}, \boldsymbol{\Delta_i}, \boldsymbol{Z_i}, i = 1, \cdots, n) = \\ &= \sum_{i=1}^{n}\left\{\sum_{j=1}^{m} X_{ij}^{(k)}\left[-e^{\boldsymbol{\theta' \tilde{Z}_i}} \sum_{j'=1}^{m} p_{j'} \mathbf{1}(s_{j'} \leq s_{j-1}) + \log\left(1 - \exp(-p_j e^{\boldsymbol{\theta' \tilde{Z}_i}})\right)\right] - X_{i,m+1}^{(k)} e^{\boldsymbol{\theta' \tilde{Z}_i}}\right\}, \end{aligned} \tag{2.39}$$

em que denota-se $s_0 = \infty$. Fazendo $c_{ij}^{(k)} = \sum_{j'=1}^{m} X_{ij'}^{(k)} \mathbf{1}(s_j \leq s_{j'-1}) \geq 0$ (considerando-se independentes de $\boldsymbol{\theta}$ e $\boldsymbol{p}$, pois encarados como observações), tem-se a esperança da função de log-verossimilhança expressa por

$$l^{(k)}(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} \left[-c_{ij}^{(k)} p_j e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} + X_{ij}^{(k)} \log \left(1 - \exp(-p_j e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i}) \right) \right] - X_{i,m+1}^{(k)} e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} \right\}. \quad (2.40)$$

Liu e Shen (2009) demonstram que a função de log-verossimilhança esperada apresentada acima é uma função côncava em relação a $\boldsymbol{\theta}$ e $p_j > 0, j = 1, \cdots, m$, implicando que qualquer máximo local é também máximo global no conjunto em que $\sum_{j=1}^{m} p_j = 1$, $p_j > 0, j = 1, \cdots, m$ e com $\boldsymbol{\theta}$ pertencendo a um conjunto convexo aberto. Conforme desenvolvido no artigo original e retratado no Apêndice A.2, a alteração da restrição $p_j \geq 0, j = 1, \cdots, m$ para $p_j > 0, j = 1, \cdots, m$ segue diretamente das condições de Karush-Kuhn-Tucker (Boyd e Vandenberghe, 2004).

O passo M é definido pela maximização da função de log-verossimilhança esperada perfilada com respeito a $\boldsymbol{\theta}$ fixado $\boldsymbol{p}$ e com respeito a $\boldsymbol{p}$ fixado $\boldsymbol{\theta}$ de maneira iterativa, definindo assim o algoritmo ECM. Por conta do número de covariáveis usualmente ser muito menor do que o tamanho n da amostra, a maximização de (2.40) com respeito a $\boldsymbol{\theta}$ pode ser realizada por uma vasta gama de algoritmos. Neste trabalho, faz-se uso do algoritmo Broyden-Fletcher-Goldfarb-Shanno (BFGS) através da função *optim()* pertencente ao pacote estatístico R. Um estudo das derivadas parciais de (2.39) em Liu e Shen (2009) apresenta condições para identificar concavidade estrita.

Obtida uma estimativa para o vetor $\boldsymbol{\theta}$, a etapa seguinte do algoritmo ECM resume-se à estimação de $\boldsymbol{p}$ e consequentemente da função de distribuição F. Entretanto, a dimensão de $\boldsymbol{p}$ em conjunto com suas restrições pode trazer grandes dificuldades ao processo de estimação.

Como resultado direto da proposição de concavidade apresentada em Liu e Shen (2009), a função de log-verossimilhança esperada (2.40), fixado $\boldsymbol{\theta}$, é estritamente côncava em função de $\boldsymbol{p}$ quando há pelo menos um valor não nulo para p_j (com $p_j \in [0, 1]$) e pelo menos um valor não nulo para X_{ij}, com $j = 1, \cdots, m$, o que garante máximo global único como sendo o máximo local encontrado.

Para auxiliar a estimação de $\boldsymbol{p}$, define-se para $j = 1, \cdots, m$,

$$a_j = \sum_{i=1}^{n} \left[e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} \sum_{j'=1}^{m} X_{ij'}^{(k)} \mathbf{1}(s_j \leq s_{j'-1}) \right], \quad (2.41)$$

tal que $X_{ij}^{(k)}$ é definido em (2.38), com k sendo o número da iteração atual. Denotando $\boldsymbol{a} = (a_1, \cdots, a_m)'$, com $\boldsymbol{\theta}$ e k fixos, tem-se a seguinte função de log-verossimilhança esperada quando descartam-se de (2.40) os termos que não dependem de $\boldsymbol{p}$:

$$f(\boldsymbol{p}) \equiv -\boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{p} + \sum_{i=1}^{n} \sum_{j=1}^{m} X_{ij}^{(k)} \log \left[1 - \exp(-e^{\boldsymbol{\theta}' \tilde{\boldsymbol{Z}}_i} p_j) \right], \quad (2.42)$$

em que $\sum_{j=1}^{m} p_j = 1$ e $0 \leq p_j \leq 1$, com $j = 1, \cdots, m$. Dadas tais restrições, conforme mencionado, quando fixado $\boldsymbol{\theta}$ e expressa em função de $\boldsymbol{p}$, a função de log-verossimilhança esperada (2.40) é estritamente côncava em $\boldsymbol{p}$. Assim, o processo de estimação pode ser expresso como

$$\max f(\boldsymbol{p}) \text{ sujeita a } \sum_{j=1}^{m} p_j = 1, \text{ e } p_j \geq 0, \text{ para } j = 1, \cdots, m. \quad (2.43)$$

A partir das expressões (2.42) e (2.43), o problema pode ser abordado no contexto de otimização convexa por meio das condições de Karush-Kuhn-Tucker apresentadas em Boyd e Vandenberghe (2004). Por conta das restrições dadas pelo método utilizado (vide Apêndice A.2), a busca de $\boldsymbol{p}$ necessariamente satisfaz $p_j > 0$, para $j = 1, \cdots, m$. A resolução do sistema não linear obtido ao

avaliar-se o vetor gradiente de $f(\boldsymbol{p})$ em (2.43) trata-se de um processo computacional intensivo e não trivial. Adota-se então o método de Newton para aproximar as equações não lineares, conforme exibido no Apêndice A.2, com a próxima iteração, fixada a iteração atual $(\boldsymbol{p}, \boldsymbol{\tau}, v)$, podendo ser expressa como

$$p_j \leftarrow p_j + \psi \Delta p_j \quad \tau_j \leftarrow \tau_j + \psi \Delta \tau_j \quad v \leftarrow v + \psi \Delta v, \tag{2.44}$$

com ψ interpretado como o tamanho do passo e determinado através do mecanismo de *backtracking* (ver Apêndice A.3). Os incrementos Δp_j, $\Delta \tau_j$ e Δv são demonstrados no Apêndice A.2 por meio da resolução aproximada do sistema não linear derivado de (2.43), com suas expressões também exibidas no sumário do fim deste capítulo.

As equações (2.44) tornam o processo semelhante ao algoritmo de Turnbull no sentido de que estas são explícitas e atualizam a si próprias. Além disso, mesmo com valores grandes de dimensão para $\boldsymbol{p}$, o algoritmo exibiu soluções numéricas estáveis, conforme pode ser visualizado em Liu e Shen (2009).

2.3.4 Sumário do Algoritmo

Aqui é apresentado um sumário do algoritmo ECM com o uso da teoria apresentada até então, assim como as quantidades já definidas anteriormente, para as estimativas de máxima verossimilhança de $\boldsymbol{\theta}$ e $\boldsymbol{p}$ na presença de fração de cura com censura intervalar. Reproduz-se aqui um passo a passo de maneira fiel ao artigo de origem.

1. Inicie o processo com vetor de parâmetros iniciais $\theta^{(0)}$ e probabilidades (saltos) $\boldsymbol{p^{(0)}}$ que satisfaçam $\sum_{j=1}^{m} p_j^{(0)} = 1$ e $p_j^{(0)} > 0$.

2. Obtenha $\boldsymbol{\theta^{(k+1)}}$ através da maximização de $l^{(k)}(\boldsymbol{\theta}, \boldsymbol{p})$ com respeito a $\boldsymbol{\theta}$, em que

$$l^{(k)}(\boldsymbol{\theta}, \boldsymbol{p}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{m} \left[-c_{ij}^{(k)} p_j e^{\boldsymbol{\theta' \tilde{Z}_i}} + X_{ij}^{(k)} \log\left(1 - \exp(-p_j e^{\boldsymbol{\theta' \tilde{Z}_i}})\right) \right] - X_{i,m+1}^{(k)} e^{\boldsymbol{\theta' \tilde{Z}_i}} \right\},$$

$i = 1, \cdots, n$, $c_{ij}^{(k)} = \sum_{j'=1}^{m} X_{ij'}^{(k)} \mathbf{1}(s_j \leq s_{j'-1})$,

$$\gamma_{i,j}(\boldsymbol{\theta}, \boldsymbol{p}) = S_{\boldsymbol{Z_i}}(s_j) \left[1 - \exp\left(-p_j e^{\boldsymbol{\theta' \tilde{Z}_i}}\right)\right], \ j = 1, \cdots, m,$$
$$\gamma_{i,m+1}(\boldsymbol{\theta}, \boldsymbol{p}) = \exp\left(-e^{\boldsymbol{\theta' \tilde{Z}_i}}\right),$$
$$S_{Z_i}(s_j) = \exp\left(-e^{\boldsymbol{\theta' \tilde{Z}_i}} \sum_{j'=0}^{j-1} p_{j'}\right), \ j = 1, \cdots, m,$$

e

$$X_{ij}^{(k)} = \frac{\delta_{ij} \gamma_{ij}(\boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}})}{\sum_{j'=1}^{m+1} \delta_{ij'} \gamma_{ij'}(\boldsymbol{\theta^{(k)}}, \boldsymbol{p^{(k)}})}, \ j = 1, \cdots, m+1.$$

3. Para obter $\boldsymbol{p^{(k+1)}}$ pelo método de pontos interiores primal-dual, dado $\boldsymbol{\theta^{(k+1)}}$, defina $b_i^{(k+1)} = \exp(\boldsymbol{\tilde{Z}_i'} \boldsymbol{\theta^{(k+1)}})$ e atualize para $j = 1, \cdots, m$,

$$X_{ij}^{(k)} = \frac{\delta_{ij} \gamma_{ij}(\boldsymbol{\theta^{(k+1)}}, \boldsymbol{p^{(k)}})}{\sum_{j'=1}^{m+1} \delta_{ij'} \gamma_{ij'}(\boldsymbol{\theta^{(k+1)}}, \boldsymbol{p^{(k)}})}.$$

Comece com valores iniciais $\boldsymbol{\tilde{p}} = \boldsymbol{p^{(k)}}$; escolha $\sigma > 1$ e $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_m)$ em que $\tau_j > 0$. Atualize $\boldsymbol{\tilde{p}}$ através dos seguintes passos até a convergência para a obtenção de $\boldsymbol{p^{(k+1)}}$:

(a) Estabeleça $\eta = \sigma / \hat{\xi}$, em que $\hat{\xi} = \sum_{j=1}^{m} \tilde{p}_j \tau_j / m$;

(b) Avalie

$$\dot{f}_j(\tilde{\boldsymbol{p}}) = \sum_{i=1}^{n} b_i^{(k+1)} \left[\frac{X_{ij}^{(k+1)}}{\exp(b_i^{(k+1)}\tilde{p_j}) - 1} - \sum_{j'=1}^{m} X_{ij'}^{(k+1)} \mathbf{1}(s_j \leq s_{j'-1}) \right],$$

$$\ddot{f}_j(\tilde{\boldsymbol{p}}) = -\sum_{i=1}^{n} \frac{X_{ij} \exp(2\tilde{\boldsymbol{Z}}_{\boldsymbol{i}}{}' \boldsymbol{\theta}^{(\boldsymbol{k+1})} \exp(b_i^{(k+1)}\tilde{p_j}))}{(\exp(b_i^{(k+1)}\tilde{p_j}) - 1)^2},$$

$$v^+ = \frac{\sum_{j=1}^{m} \left[(1/\eta + \tilde{p_j}\dot{f}_j(\tilde{\boldsymbol{p}}))/(\tau_j^{(k)} - \tilde{p_j}\ddot{f}_j(\tilde{\boldsymbol{p}})) \right]}{\sum_{j=1}^{m} \left[\tilde{p_j}/(\tau_j^{(k)} - \tilde{p_j}\ddot{f}_j(\tilde{\boldsymbol{p}})) \right]},$$

$$\Delta p_j = \frac{-\tilde{p_j}v^+ + \tilde{p_j}\dot{f}_j(\tilde{\boldsymbol{p}}) + 1/\eta}{\tau_j - \tilde{p_j}\ddot{f}_j(\tilde{\boldsymbol{p}})}$$

e

$$\tau_j^+ = v^+ - \ddot{f}_j(\tilde{\boldsymbol{p}})\Delta p_j - \dot{f}_j(\tilde{\boldsymbol{p}});$$

(c) Atualize $\tilde{\boldsymbol{p}}$, $\boldsymbol{\tau}$ e v por (2.44) utilizando o mecanismo de busca linear padrão *backtracking* encontrado no Apêndice A.3.

4. Repita os passos 2 e 3 até que $||\boldsymbol{\theta}^{(\boldsymbol{k+1})} - \boldsymbol{\theta}^{(\boldsymbol{k})}|| + ||\boldsymbol{p}^{(\boldsymbol{k+1})} - \boldsymbol{p}^{(\boldsymbol{k})}|| < \epsilon_1$ e a diferença das verossimilhanças esperadas convirja para um valor pré-especificado ϵ_2.

O algoritmo apresentado tem sua rotina em C disponibilizada pelos autores originais através de contato por e-mail. Como contribuição adicional deste trabalho, a implementação em R da rotina apresentada encontra-se disponibilizada no repositório CRAN por meio do pacote *intercure*.

A convergência do algoritmo pode ser obtida através dos resultados gerais a respeito do algoritmo ECM com restrições apresentados nos trabalhos de Meng e Rubin (1993), Nettleton (1999) e Little e Rubin (2002). Liu e Shen (2009) conjecturam que a taxa de convergência de $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ seja de $n^{-1/2}$ e que para $\hat{F}_n$ a taxa seja de $n^{-1/3}$. A normalidade assintótica e eficiência de $n^{1/2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta_0})$ é obtida através da derivação da função escore de $\boldsymbol{\theta}$ tratando-se F como uma função de perturbação. A matriz de covariâncias assintótica de $\hat{\boldsymbol{\theta}}_n$ e $\hat{F}_n$ pode ser estimada por um subproduto computacional do algoritmo ECM apresentado em Van Dyk *et al.* (1995). Para as aplicações deste trabalho, a matriz de covariâncias de $\hat{\boldsymbol{\theta}}_n$ é estimada por meio da informação observada de $\boldsymbol{\theta}$ para F fixada.

A consistência forte do estimador de máxima verossimilhança foi demonstrada pelos autores sob determinadas condições de regularidade utilizando-se a distância de Hellinger, garantindo consistência global. O teorema, em conjunto com as construções necessárias para seu enunciado, é disponibilizado no Apêndice A.4, com a demonstração disponível nos materiais suplementares do artigo de origem.

2.4 Modelo de Fragilidade

Esta seção apresenta ao leitor o modelo semiparamétrico de fração de cura para dados com censura intervalar proposto em Lam *et al.* (2013), denominado simplesmente "modelo de fragilidade" em posteriores referências na literatura. A notação original dos autores foi mantida e algumas demonstrações omitidas a fim de evitar uma apresentação exaustiva remetendo, eventualmente, o leitor aos artigos originais.

2.4.1 Notação

Em Aalen (1992) é apresentado um modelo paramétrico de análise de sobrevivência cuja função de risco é dada por

$$\lambda(t|U=u) = u\lambda_0(t), \tag{2.45}$$

em que $\lambda_0(t)$ é função de risco basal, e U é tomado como efeito individual ou fragilidade, sendo modelado por uma distribuição Poisson composta. Definido dessa maneira, o modelo contempla a possibilidade de uma fração de cura, em que o indivíduo é dado como não suscetível se $U = 0$.

Lam *et al.* (2013) estendem tal modelo admitindo o efeito de covariáveis sobre a distribuição da fragilidade, consequentemente explicando incidência e heterogeneidade para um determinado indivíduo. Além disso, o modelo semiparamétrico de Cox é utilizado nesta extensão, incorporando o efeito de covariáveis sobre o tempo até a falha nos indivíduos suscetíveis. Com o intuito de apresentar a forma analítica do modelo proposto, algumas definições são apresentadas a seguir.

Considere uma amostra aleatória de tamanho n tal que T_i denote o tempo de evento e $\boldsymbol{x_i} = (x_{i1}, \cdots, x_{ip})'$ sejam as covariáveis observadas relativas ao indivíduo i, para $i = 1, ..., n$. Define-se então $\boldsymbol{x_i^{(0)}} = (1, x_{i1}^{(0)}, \cdots, x_{ip_0}^{(0)})'$ como sendo o conjunto de covariáveis associados à incidência, ou seja, influenciando a distribuição das fragilidades U_i, enquanto $\boldsymbol{x_i^{(1)}} = (x_{i1}^{(1)}, \cdots, x_{ip_1}^{(1)})'$ são as covariáveis associadas ao tempo de falha dos indivíduos suscetíveis através de efeito multiplicativo sobre o risco do indivíduo (como no modelo padrão de Cox), sendo $x_{ij}^{(k)}$ um elemento em $\boldsymbol{x_i}$, em que $p_0, p_1 \leq p$. Na terminologia de Lam *et al.* (2013), o segundo conjunto de covariáveis, diretamente associado ao risco, é dito como sendo associado à latência.

Para o caso de censura à direita, os dados são representados por $(y_i, \delta_i, \boldsymbol{x_i}; i = 1, \cdots, n)$ em que $Y_i = \min(T_i, C_i)$, C_i é tempo de censura à direita independente de T_i, e δ_i é o indicador de falha dado por $\delta_i = I(T_i \leq C_i)$.

No caso geral de censura intervalar, os dados observados são apresentados como $(l_i, r_i, \delta_i, \boldsymbol{x_i}; i = 1, \cdots, n)$, assumindo T não observável e contido no intervalo $(l_i, r_i]$ para todos os indivíduos. O indicador de falha é definido de maneira similar com $\delta_i = 1$ se $r_i < \infty$ e $\delta_i = 0$, caso contrário.

2.4.2 Modelo

Com as devidas notações, a função de risco condicional do modelo sugerido é dada por

$$\lambda(t|u_i, \boldsymbol{x_i^{(1)}}) = u_i\lambda_0(t)\exp(\boldsymbol{\beta' x_i^{(1)}}), \tag{2.46}$$

em que $\lambda_0(t)$ é uma função de risco basal arbitrária e $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{p_1})'$ é o vetor dos parâmetros de regressão associados ao tempo latente. Nesta modelagem, assume-se que a fragilidade U_i tem distribuição qui-quadrado não centralizada com zero grau de liberdade (Siegel, 1979), equivalente a uma distribuição Poisson composta e caso particular da família de distribuições estudada em Aalen (1992). A variável aleatória U_i é construída como soma de K_i variáveis aleatórias independentes $(W_{1i}, W_{2i}, \cdots, W_{k_i i})$, cada uma com distribuição χ^2_2. Deste modo, condicionando-se $K_i = k_i$, obtém-se

$$U_i = \begin{cases} 0, & \text{se } k_i = 0; \\ W_{1i} + W_{2i} + \cdots + W_{k_i i}, & \text{se } k_i > 0. \end{cases}$$

Neste modelo tem-se $K_i \sim \text{Poisson}(\eta_i/2)$, em que $\eta_i = \exp(\boldsymbol{\theta' x_i}^{(0)})$ com $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_{p_0})'$ sendo o vetor de parâmetros de regressão associados à fragilidade. Como mencionado em Lam *et al.* (2013), de modo similar aos modelos BCH, em muitos estudos de câncer associa-se a variável K como sendo o número latente de células tumorosas em metástase, sendo neste caso W_{ji} a contribuição para U_i da j-ésima célula. Neste caso em particular, as contribuições de diferentes células não se cancelam e possuem efeito aditivo.

Com isso, denotando por $g(\cdot)$ a função de densidade de probabilidade ou função de probabilidade de uma determinada variável, tomar a fragilidade U_i como Poisson composta resulta em

$$g(u|\boldsymbol{x_i}^{(0)}) = \sum_{k=0}^{\infty} g(u|k, \boldsymbol{x_i}^{(0)}) g(k).$$

Da relação da soma de variáveis aleatórias com distribuição qui-quadrado central com a distribuição gama, e por $k = 0$ implicar em $u = 0$, tem-se

$$g(u|\boldsymbol{x_i}^{(0)}) = \begin{cases} \exp(-\eta_i/2), & \text{se } u = 0; \\ \sum_{k=1}^{\infty} \frac{e^{-\eta_i/2}(\eta_i/2)^k}{k!} \times \frac{u^{k-1}e^{-u/2}}{2^k\Gamma(k)}, & \text{se } u > 0. \end{cases} \tag{2.47}$$

Assim definida, a fragilidade possui massa no ponto 0. O modelo contemplado aqui é um caso particular do modelo apresentado por Aalen (1992), diferenciando-se pelo fato de acomodar o efeito de covariáveis sobre a incidência individual. A extensão assim realizada permite que a probabilidade de cura dependa de covariáveis, tornando o processo mais natural na prática. Com as definições apresentadas, é fácil notar que $E(U_i|\boldsymbol{x_i}^{(0)}) = \eta_i$ e, além disso, dadas as covariáveis, pode-se obter a distribuição marginal de T conforme desenvolve-se abaixo:

$$\begin{aligned} S(t|\eta_i, \boldsymbol{x_i}^{(1)}) &= P(T > t|\eta_i, \boldsymbol{x_i}^{(1)}) = P(T > t, U = 0|\eta_i, \boldsymbol{x_i}^{(1)}) + P(T > t, U > 0|\eta_i, \boldsymbol{x_i}^{(1)}) \\ &= P(T > t|U = 0, \eta_i, \boldsymbol{x_i}^{(1)}) P(U = 0|\eta_i, \boldsymbol{x_i}^{(1)}) \\ &\quad + P(T > t|U > 0, \eta_i, \boldsymbol{x_i}^{(1)})[P(U > 0, k = 0|\eta_i, \boldsymbol{x_i}^{(1)}) + P(U > 0, k > 0|\eta_i, \boldsymbol{x_i}^{(1)})] \\ &= S(t|U = 0, \eta_i, \boldsymbol{x_i}^{(1)}) P(U = 0|\eta_i, \boldsymbol{x_i}^{(1)}) \\ &\quad + S(t|U > 0, \eta_i, \boldsymbol{x_i}^{(1)}) P(U > 0|k > 0, \eta_i, \boldsymbol{x_i}^{(1)}) P(k > 0|\eta_i, \boldsymbol{x_i}^{(1)}). \end{aligned} \tag{2.48}$$

De (2.46) tem-se que $S(t|u_i, x_i) = \exp(-u_i\Lambda_0(t)\exp(\boldsymbol{\beta' x_i^{(1)}}))$.
Portanto, de (2.47) e (2.48), obtém-se

$$\begin{aligned} S(t|\eta_i, \boldsymbol{x_i}^{(1)}) &= e^{-\eta_i/2} + \int_0^{\infty} \sum_{k=1}^{\infty} \frac{e^{-\eta_i/2}(\frac{\eta_i}{2})^k}{k!} \times \frac{u_i^{k-1}e^{-u_i/2}}{2^k\Gamma(k)} \exp\left\{-u_i\Lambda_0(t)\exp(\beta' x_i^{(1)})\right\} du_i \\ &= e^{-\eta_i/2} + e^{-\eta_i/2} \sum_{k=1}^{\infty} \left(\frac{\eta_i/2}{1 + 2\Lambda_0(t)\exp(\beta' x_i^{(1)})}\right)^k \frac{1}{k!} \\ &= \exp\left[-\frac{\eta_i}{2}\left\{1 - \frac{1}{1 + 2\Lambda_0(t)\exp(\boldsymbol{\beta' x_i}^{(1)})}\right\}\right], \end{aligned} \tag{2.49}$$

em que $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ é a função de risco basal acumulado. Com a função de sobrevivência construída desta maneira, a probabilidade de cura é obtida através do limite $t \to \infty$ em (2.49) levando a $\lim_{t\to\infty} S(t|\eta_i, \boldsymbol{x_i}^{(1)}) = P(K_i = 0) = \exp(-\eta_i/2)$.

De acordo com Lam *et al.* (2013), pode-se interpretar as fragilidades dependentes de covariáveis U_i como sendo índices individuais de condição ou estado de saúde que levam em consideração características dadas por covariáveis como sexo, idade e tipo de tratamento recebido, por exemplo. Com U vista desta forma, temos intuitivamente que valores pequenos da variável representam condições melhores de saúde, levando a um risco menor para o evento em questão. O artigo original menciona ainda que, para doenças tratáveis, valores relativamente pequenos de U podem ser uma indicação de cura, implicando em um comportamento mais homogêneo para os casos em que U é menor que um valor arbitrário τ suficientemente pequeno . Sem perda de generalidade e mantendo a metodologia fiel ao artigo em questão, toma-se $\tau = 0$, assumindo-se então que indivíduos com $U_i = 0$ são curados ou com risco extremamente baixo de ocorrência do evento. Assim definido, U

possui distribuição com massa em 0 e é contínua nos reais positivos.

2.4.3 Estimação

Em Lam *et al.* (2013) apresenta-se uma metodologia para dados com censura à direita em conjunto com uma extensão direta para o caso de dados na presença de censura intervalar. A ideia por trás desta extensão baseia-se na imputação de tempos de sobrevida levando-se em consideração a informação a respeito da função de sobrevivência para cada indivíduo. Deste modo, reduz-se o problema ao caso mais simples de censuras à direita, possibilitando assim o uso de técnicas já consolidadas e com boas propriedades para tal caso.

Devido à presença da fragilidade U, o processo de estimação torna-se complicado, pois a função de verossimilhança parcial, dependente das variáveis latentes K e U, não pode ser diretamente avaliada. Entretanto, a estimação torna-se simples quando supõe-se que tais variáveis são observáveis.

Suponha então um conjunto de dados contendo somente observações exatas ou censuradas à direita, com K e U observáveis. Denote $\boldsymbol{D}$ como sendo os dados completos tal que $\boldsymbol{D} = (y_i, \delta_i, \boldsymbol{x_i}, k_i, u_i; i = 1, \cdots, n)$. A estimação de $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ é obtida diretamente através da maximização da seguinte função de verossimilhança parcial L_C com dados completos dada por

$$\begin{aligned} L_C(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{D}) = & \prod_{i=1}^{n} \frac{\exp\left\{\frac{\exp(\boldsymbol{\theta'}\boldsymbol{x_i}^{(0)})}{2}\right\}\left\{\frac{\exp(\boldsymbol{\theta'}\boldsymbol{x_i}^{(0)})}{2}\right\}^{k_i}}{k_i!} \\ & \times \prod_{k_i>0} \frac{u_i^{k_i-1}\exp(\frac{-u_i}{2})}{2^{k_i}\Gamma(k_i)} \times \prod_{i=1}^{n}\left\{\frac{u_i \exp(\boldsymbol{\beta'}\boldsymbol{x_i}^{(1)})}{\sum_{m\in R(y_i)} u_m \exp(\boldsymbol{\beta'}\boldsymbol{x_m}^{(1)})}\right\}^{\delta_i}, \end{aligned} \tag{2.50}$$

em que $R(y_i)$ é o conjunto de indivíduos em risco no instante y_i.

Consequentemente, tem-se que

$$\begin{aligned} L_C(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{D}) &\propto \prod_{i=1}^{n} \exp\left\{-\frac{\exp(\boldsymbol{\theta'}\boldsymbol{x_i}^{(0)})}{2}\right\} \exp(k_i\boldsymbol{\theta'}\boldsymbol{x_i}^{(0)}) \times \prod_{i=1}^{n}\left\{\frac{\exp(\boldsymbol{\beta'}\boldsymbol{x_i}^{(1)})}{\sum_{m\in R(y_i)} u_m \exp(\boldsymbol{\beta'}\boldsymbol{x_m}^{(1)})}\right\}^{\delta_i} \\ &= L_1(\boldsymbol{\theta}) \times L_2(\boldsymbol{\beta}). \end{aligned} \tag{2.51}$$

De (2.51), nota-se que as funções de verossimilhança de $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ são ortogonais entre si e, desta forma, as estimativas de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\beta}}$ são obtidas separadamente. Observando-se (2.50), tem-se que a estimação dos parâmetros $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$ pode ser obtida com regressão Poisson e regressão de Cox, respectivamente, uma vez que se tenha o conjunto de dados completos $\boldsymbol{D}$. Além disso, a função de risco basal acumulada $\Lambda_0(t)$ pode ser estimada com o estimador de Nelson-Aalen,

$$\hat{\Lambda}_0(t) = \sum_{i:t_i\leq t}\left\{\frac{\delta_i}{\sum_{m\in R(y_i)} u_m \exp(\boldsymbol{\beta'}\boldsymbol{x_m}^{(1)})}\right\}. \tag{2.52}$$

A seguir, é apresentado um método de estimação utilizando-se imputação múltipla, adotando-se a técnica de aumento de dados apresentada em Tanner e Wong (1987a) com o aumento de dados normais assintóticos (*Asymptotic Normal Data Augmentation, ANDA*) apresentado por Wei e Tanner (1991). Estudos formais quanto às propriedades de convergência do método de imputação podem ser vistos em Tanner e Wong (1987a), encontrando-se além do escopo deste trabalho.

A ideia do estimador proposto é o aumento dos dados incompletos $(y_i,\ \delta_i,\ \boldsymbol{x}_i;\ i = 1, \cdots, n)$ para $(y_i,\ \delta_i,\ \boldsymbol{x}_i,\ k_i,\ u_i;\ i = 1, \cdots, n)$ em M processos de geração independentes por meio das distribuições preditivas $(K_i - \delta_i)|(y_i, \boldsymbol{x}_i, \delta_i)$ e $U_i|(y_i, \delta_i, \boldsymbol{x}_i)$ para um vetor de parâmetros inicial $\boldsymbol{\alpha}^{(0)} = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\beta}^{(0)})$, desenvolvidas no Apêndice A.5. Geradas as variáveis latentes, estima-se $\hat{\boldsymbol{\alpha}}$ pelo método da máxima verossimilhança para cada um dos M conjuntos de dados aumentados, consis-

tindo em uma iteração do algoritmo proposto. Adotando a média destas estimativas como $\hat{\boldsymbol{\alpha}}^{(\boldsymbol{k})}$, processo é repetido imputando-se o vetor $\boldsymbol{\alpha}_h$ $(h = 1, \cdots, M)$, proveniente de $N(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{k})}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}^{(\boldsymbol{k})}})$, nas distribuições condicionais das variáveis latentes K_i e U_i. Conforme apresentado em Wei e Tanner (1991), $N(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{k})}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}^{(\boldsymbol{k})}})$ é uma aproximação para a distribuição condicionada a dados incompletos $p(\boldsymbol{\alpha}|y_i, \delta_i, \boldsymbol{x}_i)$, justificada em situações de amostras de tamanho grande.

Embora a construção do estimador utilize argumentos bayesianos propostos em Tanner e Wong (1987a), este pode ser essencialmente tratado como uma implementação do algoritmo de Monte Carlo para o passo E do algoritmo EM (Rubin, 1987). Para incorporar a fração de cura à função de verossimilhança dos dados, considera-se o uso da restrição de cauda nula sugerida em Taylor (1995), fazendo-se U_j como 0 se o j-ésimo indivíduo é censurado além do maior tempo de falha observado t^*. Além disso, para dados com censura intervalar, os autores do artigo original fazem uso de imputações dos tempos de sobrevida por meio de suas distribuições condicionais, conforme exibido adiante.

2.4.4 Algoritmo Computacional

O algoritmo para estimação dos parâmetros de regressão $\boldsymbol{\alpha} = (\boldsymbol{\theta}', \boldsymbol{\beta}')$ e obtenção da matriz de covariâncias de $\hat{\boldsymbol{\alpha}}$, dada por $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}}$, é tecnicamente descrito a seguir e pode ser visto com mais detalhes em Lam *et al.* (2013).

1. Inicie com $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(\boldsymbol{0})}$ e $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}^{(\boldsymbol{0})}_{\boldsymbol{\alpha}} = \phi \boldsymbol{I}$, sendo $\boldsymbol{I}$ matriz identidade com dimensão $(p_0 + p_1 + 1) \times (p_0 + p_1 + 1)$ e ϕ uma constante não-negativa, preferencialmente pequena, como $\phi = 0, 1$. Além disso, tome cada y_i como sendo o ponto médio do intervalo definido por l_i e r_i para casos de censura intervalar, mantendo $y_i = l_i$ para indivíduos em que $\delta_i = 0$.

2. Calcule $\hat{\Lambda}_0^{(0)}(t)$ baseado na equação (2.52) tomando $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\boldsymbol{0})}$ e $u_i = u_i^{(0)} = \delta_i$.

3. Para o j-ésimo passo $(j = 1, 2, \cdots)$:

   (a) Gere $\hat{\boldsymbol{\alpha}}_{\boldsymbol{h}} \sim N(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{j-1})}, \hat{\boldsymbol{\Sigma}}^{(\boldsymbol{j-1})}_{\boldsymbol{\alpha}})$ para $h = 1, 2, \cdots, M$;

   (b) Gere $\boldsymbol{y_h} = (y_{h1}, \cdots, y_{hn})$ fazendo-se $y_{hi} = l_i$ se $\delta_i = 0$ para $h = 1, \cdots, M$. Para indivíduos com $\delta_i = 1$, tome $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\boldsymbol{h}}$ e $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$ para $h = 1, \cdots, M$, e gere y_{hi} a partir da distribuição condicional

$$P(Y > y | l_i, r_i, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{x_i}) = \frac{S(y|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{x_i}) - S(r_i|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{x_i})}{S(l_i|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{x_i}) - S(r_i|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{x_i})},$$

   em que a função de sobrevivência $S(y|\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \boldsymbol{x_i})$ é dada por (2.49).

   (c) Tomando-se $\boldsymbol{y} = \boldsymbol{y_h}$, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\boldsymbol{h}}$ e $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$, gere $\boldsymbol{k_h} = (k_{h1}, \cdots, k_{hn})$ $(h = 1, \cdots, M)$ a partir da distribuição *a posteriori* de K_i dada por

$$(K_i - \delta_i)|(y_i, \boldsymbol{x_i}, \delta_i) \sim \text{Poisson}\left(\frac{\exp(\boldsymbol{\theta}' \boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{0})})}{2 + 4\Lambda_0(y_i)\exp(\boldsymbol{\beta}' \boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{1})})}\right);$$

   (d) Fazendo $\boldsymbol{y} = \boldsymbol{y_h}$, $\boldsymbol{k} = \boldsymbol{k_h}$, $\boldsymbol{\alpha} = \boldsymbol{\alpha_h}$, e $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$, gere $\boldsymbol{u_h} = (u_{h1}, \cdots, u_{hn})$ $(h = 1, \cdots, M)$ a partir da *posteriori* condicional U_i dada por

$$U_i|(y_i, \delta_i, \boldsymbol{x_i}, k_i) \begin{cases} = 0, & \text{se } k_i = 0; \\ \sim Gama(k_i + \delta_i, \{0,5 + \Lambda_0(y_i)\exp(\boldsymbol{\beta}' \boldsymbol{x}_{\boldsymbol{i}}^{(\boldsymbol{1})})\}^{-1}), & \text{se } k_i > 0. \end{cases}$$

   Além disso, faça $u_{hi} = 0$ se $y_{hi} > r^*$, em que r^* é definido por $r^* = \max_j\{r_j \delta_j\}$;

(e) Para $h = 1, \cdots, M$, maximize a função de log-verossimilhança dada por $l(\boldsymbol{\alpha}|\boldsymbol{D_h}) = \log L_C(\boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{D_h})$ com $\boldsymbol{D_h} = (y_{hi}, \delta_i, \boldsymbol{x_i}, k_{hi}, u_{hi}; i = 1, \cdots, n)$ para obter as estimativas $\hat{\boldsymbol{\theta}}^{(\boldsymbol{j},\boldsymbol{h})}$, $\hat{\boldsymbol{\beta}}^{(\boldsymbol{j},\boldsymbol{h})}$, fazendo então $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\boldsymbol{j},\boldsymbol{h})}$ para obter $\hat{\Lambda}_0^{(j,h)}(t)$ baseado na equação (2.52).

(f) Atualize as estimativas tomando

$$\hat{\boldsymbol{\theta}}^{(\boldsymbol{j})} = \frac{1}{M}\sum_{h=1}^{M}\hat{\boldsymbol{\theta}}^{(\boldsymbol{j},\boldsymbol{h})}, \quad \hat{\boldsymbol{\beta}}^{(\boldsymbol{j})} = \frac{1}{M}\sum_{h=1}^{M}\hat{\boldsymbol{\beta}}^{(\boldsymbol{j},\boldsymbol{h})}, \quad \hat{\Lambda}_0^{(j)}(t) = \frac{1}{M}\sum_{h=1}^{M}\hat{\Lambda}_0^{(j,h)}(t);$$

(g) Atualize a matriz de covariâncias estimada por

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(\boldsymbol{j})} = \frac{1}{M}\sum_{h=1}^{M}\left[-\frac{\partial^2}{\partial\boldsymbol{\alpha}'\partial\boldsymbol{\alpha}}\{l_1(\boldsymbol{\theta}) + l_2(\boldsymbol{\beta})\}\right]^{-1}_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}^{(\boldsymbol{j},\boldsymbol{h})},\boldsymbol{D}=\boldsymbol{D}^{(j,h)}} \\ + \left(1 + \frac{1}{M}\right)\sum_{h=1}^{M}\frac{(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{j},\boldsymbol{h})} - \hat{\boldsymbol{\alpha}}^{(\boldsymbol{j})})(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{j},\boldsymbol{h})} - \hat{\boldsymbol{\alpha}}^{(\boldsymbol{j})})'}{M-1}.\end{aligned} \tag{2.53}$$

4. Repita o passo 3 até que seja alcançada a convergência para as estimativas de $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\beta}}$ e $\hat{\Lambda_0}(t)$.

Conforme Rubin (1987); Schenker e Gelfand (1988); Tanner e Wong (1987b), o segundo termo de (2.53) acomoda a variação entre as imputações. O fator de inflação $(1+\frac{1}{M})$ sugere que a redução na variabilidade obtida pelo aumento de M é afetada principalmente pelo termo $1/M$. Discussões a respeito da escolha de M podem ser vistas com mais detalhes em Lam *et al.* (2013).

Como destacado no mesmo artigo, a função de sobrevivência estimada $\hat{S}(y|\boldsymbol{x_i})$ é uma função escada devido à natureza da função de risco acumulado basal estimada $\hat{\Lambda}_0(y)$. Devido a isso, o processo de geração de tempos observado em (3b) pode ocasionar um grande acúmulo de empates, tornando a maximização através da verossimilhança parcial de Cox inconveniente. Uma correção é sugerida por Lam *et al.* (2013) utilizando-se uma versão suavizada do estimador de risco acumulado. Os autores sugerem então a imputação de dados através da função de sobrevivência condicional do passo (b) do algoritmo apresentado utilizando a seguinte modificação da função de risco acumulado:

$$\grave{\Lambda}_0(t) = \frac{\hat{\Lambda}_0(l_i)(r_i - t) + \hat{\Lambda}_0(r_i)(t - l_i)}{r_i - l_i} \tag{2.54}$$

para $l_i \leq t < r_i$.

Em conjunto com a adaptação proposta, o método da imputação múltipla permite que os pontos de salto na função de risco sejam atualizados a cada iteração, não se restringindo a um conjunto finito de valores, tornando improvável a presença de empates para os tempos gerados. A abordagem apresentada contorna o problema de fixar-se um conjunto predefinido de pontos de saltos, escolha dificilmente justificável.

Embora uma opção adicional seja apresentada em Lam *et al.* (2013) para a estimação da curva de risco acumulado, este trabalho restringe-se ao uso somente da abordagem acima devido à menor aglomeração dos tempos gerados pela mesma, além de evitar com esta possíveis problemas numéricos na geração dos dados. A abordagem passa a ser a única adotada pelos autores em publicações posteriores incluindo o efeito de grupo (Lam e Wong, 2014).

Métricas quanto ao desempenho do algoritmo são apresentadas através de simulações e aplicações a dados com censura à direita ou intervalar em seu artigo original. Os estudos de simulações posteriormente apresentados neste trabalho são focados na estimação da fração de curados e nas consequências de especificações errôneas. Embora a identificabilidade deste modelo não se encontre demonstrada, as simulações exibem estabilidade e precisão nas estimativas dos parâmetros. As propriedades deste estimador, assim como para sua extensão para dados agrupados apresentada a seguir, apresentam estudos somente por meio de simulações nos artigos originais.

2.4.5 Extensão para Dados Agrupados

Na seção anterior foi apresentado o algoritmo proposto por Lam *et al.* (2013), cuja construção considera fragilidades com massa no ponto 0 para acomodar a fração de curados na população. Uma extensão deste para dados agrupados é obtida em Lam e Wong (2014) e apresentada a seguir neste trabalho. Assim como no modelo de mistura, atribui-se um efeito aleatório para indivíduos que compartilham um mesmo grupo, contemplando uma possível estrutura de correlação presente no conjunto.

Considera-se uma amostra aleatória de m grupos com n_i indivíduos no i-ésimo grupo ($i = 1, \cdots, m$). Assim como apresentado anteriormente para o cenário sem agrupamento, desenvolve-se primeiramente o caso para dados censurados à direita para então posteriormente contemplar a estrutura de censura intervalar. Faz-se T_{ij} sendo o tempo exato até a ocorrência do evento a partir da origem de um determinado processo, com $\boldsymbol{x_{ij}} = (x_{ij1}, \cdots, x_{ijp})'$ denotando o conjunto de covariáveis observadas para o membro j do grupo i, em que $j = 1, \cdots, n_i$.

Definem-se as partições $\boldsymbol{x_{ij}^{(0)}} = (1, x_{ij1}^{(0)}, \cdots, x_{ijp_0}^{(0)})'$ e $\boldsymbol{x_{ij}^{(1)}} = (x_{ij1}^{(1)}, \cdots, x_{ijp_1}^{(1)})'$ como sendo os conjuntos de covariáveis associados à fragilidade e ao tempo até o evento, respectivamente, em que $p_0, p_1 \leq p$ e $x_{ijk}^{(l)} \in \boldsymbol{x_{ij}}$ para $l = 0, 1$.

Como os autores expõem, diferente de outros modelos de fração de cura para dados agrupados apresentados na literatura (como Xiang *et al.*, 2011, visto anteriormente) em que se consideram efeitos aleatórios independentes entre si associados à cura e ao tempo de sobrevida, Lam e Wong (2014) sugerem o uso de um efeito único de fragilidade U_{ij} para cada indivíduo, considerando já ambas as contribuições. Para o grupo i, os valores U_{ij} derivam-se de uma variável latente Ξ_i, assumindo valor constante em um mesmo grupo, com contribuição direta sobre a distribuição da variável latente K_ij, conforme será visto posteriormente.

Para dados com censura intervalar, o conjunto de dados é representado por $(l_{ij}, r_{ij}, \delta_{ij}, \boldsymbol{x_{ij}}; j = 1, \cdots, n_i; i = 1, \cdots, m)$ tal que, assim como na modelagem anterior, o instante T em que o evento ocorre não é diretamente observável, mas sabe-se que pertence ao intervalo $(l_{ij}, r_{ij}]$, definindo-se também $\delta_{ij} = I(r_{ij} < \infty)$. Representa-se então o conjunto de dados completos por $\boldsymbol{D} = (l_{ij}, r_{ij}, y_{ij}, \delta_{ij}, \boldsymbol{x_{ij}}, \xi_i, k_{ij}, u_{ij}; j = 1, \cdots, n_i; i = 1, \cdots, m)$.

Fixada a variável latente de grupo $\Xi_i = \xi_i$, tem-se que os efeitos U_{ij} são independentes, assim como os tempos T_{ij}. O risco condicional do modelo de fragilidade para dados agrupados é então dado por

$$\lambda(t|u_{ij}, \boldsymbol{x_{ij}^{(1)}}) = u_{ij}\lambda_0(t)\mu_{ij},$$

com $\lambda_0(t)$ sendo a função de risco basal arbitrária e $\mu_{ij} = \exp(\boldsymbol{\beta' x_{ij}^{(1)}})$ em que $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{p_1})'$ é o vetor de parâmetros associados à regressão de Cox. Tomando-se $\Xi_i = \xi_i$, o efeito aleatório U_{ij} assume uma distribuição Poisson composta, conforme abaixo:

$$U_{ij} = \begin{cases} 0, & \text{se } k_{ij} = 0, \\ W_{1ij} + W_{2ij} + \cdots + W_{k_{ij}ij}, & \text{se } k_{ij} > 0, \end{cases}$$

com $W_{hij} \sim \chi^2_2$ para $h = 1, \cdots, k_{ij}$. A variável U_{ij} assim definida constitui uma qui-quadrado não central com zero graus de liberdade, conforme Siegel (1979).

Fixado $\Xi_i = \xi_i$, o número K_{ij} de termos na soma é dado por uma variável aleatória Poisson com média $\eta_{ij}\xi_i/2$ em que $\eta_{ij} = \exp(\theta' x_{ij}^{(0)})$ com $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_{p_0})$ denotando o vetor de parâmetros de regressão associados ao efeito sobre a fração de curados.

Assim como no modelo proposto sem agrupamento, a variável latente K determina se o indivíduo está curado ou não, além de afetar o tempo de falha do evento no caso em que o indivíduo é suscetível. Deste modo, um grupo com valor grande para Ξ tende a apresentar valores grandes para K, proporcionando baixa probabilidade de cura e menores tempos de falha para o grupo inteiro. A

função densidade de probabilidade de U_{ij} condicionada à variável latente Ξ_i é expressa por

$$g(u_{ij}|\Xi_i = \xi_i, \boldsymbol{x_{ij}}) = \begin{cases} \exp(-\eta_{ij}\xi_i/2), & \text{se } u = 0, \\ \sum_{k=1}^{\infty} \frac{e^{-\eta_{ij}\xi_i/2}(\eta_{ij}/2\xi_i)^k}{k!} \times \frac{u^{k-1}e^{-u/2}}{2^k\Gamma(k)}, & \text{se } u > 0. \end{cases} \tag{2.55}$$

Construído dessa maneira, pode-se verificar que a média da fragilidade é dada por

$$\mathrm{E}(U_{ij}|\Xi_i = \xi_i, \boldsymbol{x_{ij}}) = \eta_{ij}\xi_i.$$

Através de algumas operações algébricas, pode-se expressar a função de sobrevivência condicionada a Ξ_i como

$$S(t|\Xi_i = \xi_i, \boldsymbol{x_{ij}}) = \exp\left[-\frac{\eta_{ij}\xi_i}{2}\left\{1 - \frac{1}{1 + 2\Lambda_0(t)\mu_{ij}}\right\}\right], \tag{2.56}$$

em que $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ é a função de risco basal acumulada. Através desta, deriva-se facilmente a proporção de curados condicional por $\lim_{t\to\infty} S(t|\Xi_i = \xi_i, \boldsymbol{x_{ij}}) = P(K_{ij} = 0|\Xi_i = \xi_i, \boldsymbol{x_{ij}}) = \exp(-\eta_{ij}\xi_i/2)$.

Conforme ressaltado pelos autores, para a construção da função de sobrevivência marginal, assume-se $\Xi_i \sim \text{Gama}(\omega, \omega)$ com média 1 e variância ω^{-1} para $\omega > 0$. Valores pequenos de ω estão associados a uma população mais heterogênea enquanto valores altos implicam em variância de Ξ_i menor, ocasionando menor grau de associação entre observações de um mesmo grupo. O caso de independência entre observações de um mesmo grupo é contemplado quando faz-se $\omega \to \infty$. A função de sobrevivência conjunta para o grupo i é dada por

$$S(t_{i1}, \cdots, t_{i,n_i}|\boldsymbol{x_i}) = \int S(t_{i1}, \cdots, t_{i,n_i}|\xi_i, \boldsymbol{x_i}) f(\xi_i) d\xi_i \tag{2.57}$$

$$= \mathrm{LP}_{\Xi}\left[\sum_{j=1}^{n_i} \frac{\eta_{ij}}{2}\left\{1 - \frac{1}{2\mu_{ij}\Lambda_0(t_{ij}) + 1}\right\}\right] \tag{2.58}$$

$$= \left[\frac{\omega}{\omega + \sum_{j=1}^{n_i} \frac{\eta_{ij}}{2}\left\{1 - \frac{1}{2\mu_{ij}\Lambda_0(t_{ij})+1}\right\}}\right]^{\omega}, \tag{2.59}$$

em que LP denota a transformada de Laplace e $\boldsymbol{x_i} = (\boldsymbol{x_{i1}}', \cdots, \boldsymbol{x_{in_i}}')'$. Assim como para o caso sem efeito de grupo, o modelo aqui apresentado não assume uma distribuição associada à função de risco basal, proporcionando um modelo menos restritivo ao custo de uma maior complexidade.

A interpretação no contexto biológico de U como indicadora da condição de saúde do indivíduo (sendo menor para aqueles com menor risco) ainda pode ser adotada para este caso, contemplando também o efeito do grupo: grupos com Ξ pequeno tendem a apresentar membros com baixo risco e alta proporção de indivíduos não suscetíveis ao evento de interesse. Os autores do artigo original ressaltam possíveis interpretações biológicas como efeito do ambiente, condições de vida entre indivíduos próximos ou mesmo fatores genéticos.

De modo similar à metodologia sem efeito de grupo, é apresentada a função de verossimilhança para dados censurados à direita, para então apresentar uma extensão para o caso com censura intervalar. Definem-se os dados observados como $(y_{ij}, \delta_{ij}, \boldsymbol{x_{ij}};\ j = 1, \cdots, n_i;\ i = 1, \cdots, m)$, com $\boldsymbol{x_{ij}}$ sendo vetor de covariáveis, y_{ij} sendo valor observado de $Y_{ij} = \min\{T_{ij}, C_{ij}\}$, tal que T_{ij} e C_{ij} denotam os tempos de falha e censura, respectivamente. A variável $\delta_{ij} = I(T_{ij} \leq C_{ij})$ indica falha associada à $j-$ésima ocorrência do i-ésimo grupo. A função de verossimilhança completa é então dada por

$$
\begin{aligned}
L_C(\omega,\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{D}) &= \prod_{i=1}^{m}\left(\frac{\omega^{\omega}\exp(-\omega\Xi_i)\Xi_i^{\omega-1}}{\Gamma(\omega)}\prod_{j=1}^{n_i}\frac{\exp(-\eta_{ij}\xi_i/2)(\eta_{ij}\xi_i/2)^{k_{ij}}}{k_{ij}!}\right)\\
&\times\prod_{k_{ij}>0}\left(\frac{u_{ij}^{k_{ij}-1}\exp(-u_{ij}/2)}{\Gamma(k_{ij})2^{k_{ij}}}\left(\frac{u_{ij}\mu_{ij}}{\sum_{(a,b)\in R(y_{ij})}u_{ab}\mu_{ab}}\right)^{\delta_{ij}}\right)\\
&\propto\left(\prod_{i=1}^{m}\frac{\omega^{\omega}\exp(-\omega\xi_i)\xi_i^{\omega-1}}{\Gamma(\omega)}\right)\left(\prod_{i=1}^{m}\prod_{j=1}^{n_i}\exp(-\eta_{ij}\xi_i/2)\eta_{ij}^{k_{ij}}\right)\\
&\times\prod_{k_{ij}>0}\left(\frac{\mu_{ij}}{\sum_{(a,b)\in R(y_{ij})}u_{ab}\mu_{ab}}\right)^{\delta_{ij}}\\
&= L_1(\omega)\times L_2(\boldsymbol{\theta})\times L_3(\boldsymbol{\beta}),
\end{aligned}
$$

em que $R(y_{ij})$ é o conjunto de indivíduos em risco no instante imeadiatamente anterior a y_{ij}. Assim como para o modelo sem efeito aleatório, a função de verossimilhança pode ser decomposta em funções de verossimilhança ortogonais umas às outras, facilitando o processo de estimação com processos de maximização independentes. Em linguagem R, as funções *optim*, *glm* e *coxph* possibilitam a obtenção de estimativas de máxima verossimilhança para ω, θ e β, respectivamente.

Por tratar-se de um modelo semiparamétrico, estima-se a função de risco basal acumulada através de

$$
\hat{\Lambda}_0(t) = \sum_{i,j:y_{ij}\le t}\left\{\frac{\delta_{ij}}{\sum_{(a,b)\in R(y_{ij})}u_{ab}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_{\boldsymbol{ab}}^{(\mathbf{1})})}\right\}. \tag{2.60}
$$

Assim como na modelagem sem grupos, é aqui utilizado o algoritmo de aumento de dados ANDA (*Asymptotic Normal Data Augmentation*) proposto em Tanner e Wong (1987a), em que geram-se M conjuntos (U_{ij}, K_{ij}, Ξ_i) a partir da distribuição preditiva conjunta das variáveis latentes, completando assim os dados observados. Procedimentos padrões de estimação são conduzidos para cada um dos M conjuntos aumentados, combinando-os para a obtenção de uma estimativa única para o vetor de parâmetros $\boldsymbol{\alpha} = (\gamma, \boldsymbol{\theta}', \boldsymbol{\beta}')'$, com $\gamma = \log(\omega)$. A seguir é descrito o algoritmo para a obtenção de $\hat{\boldsymbol{\alpha}}$, assim como sua respectiva matriz de covariâncias $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}}$.

Apresenta-se aqui os passos do algoritmo de imputação múltipla para dados agrupados. Por este diferir do algoritmo anterior em poucos aspectos, a extensão para grupos torna-se razoavelmente simples. Adotando-se $\gamma = \log(\omega)$, o processo de estimação, já contemplando o caso de censura intervalar, é exibido abaixo:

1. Inicie com $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(\mathbf{0})}$ e $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(\mathbf{0})} = \phi\boldsymbol{I}$, sendo $\boldsymbol{I}$ a matriz identidade com dimensão $(p_0+p_1+2)\times(p_0+p_1+2)$ e ϕ uma constante não negativa, preferencialmente pequena, como $\phi = 0, 1$. Além disso, obtenha réplicas iniciais $\boldsymbol{y}_{\boldsymbol{h}}^{(\mathbf{0})}$ $(h = 1, \cdots, M)$, fazendo $y_{hij}^{(0)} = (l_{ij} + r_{ij})/2$ se $\delta_{ij} = 1$, para $j = 1, \cdots, n_i$; $i = 1, \cdots, m$; $h = 1, \cdots, M$.

2. Calcule $\hat{\Lambda}_0^{(0)}(t)$ baseado na expressão (2.52) tomando $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\mathbf{0})}$ e $u_{ij} = u_{ij}^{(0)} = \delta_{ij}$.

3. Para o q-ésimo passo $(q = 1, 2, \cdots)$:

   (a) Gere $\hat{\boldsymbol{\alpha}}_{\boldsymbol{h}}$ de $N(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{q-1})}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(\boldsymbol{q-1})})$ para $h = 1, 2, \cdots, M$;

   (b) Gere $\boldsymbol{y_h}^{(q)} = (y_{h11}, \cdots, y_{hmn_i})$ tomando $y_{hij} = l_{ij}$ se $\delta_{ij} = 0$ para $h = 1, \cdots, M$. Para indivíduos com $\delta_{ij} = 1$, tome $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\boldsymbol{h}}$ e $\Lambda_0(t) = \hat{\Lambda}_0^{(q-1)}(t)$ para $h = 1, \cdots, M$, e gere $y_{hij}^{(q)}$ sequencialmente a partir da distribuição condicional preditiva demonstrada em Lam *et al.*

(2010), exibida a seguir:

$$P(Y_{ij} > y | l_{ij}, r_{ij}, y_{i,j-1}, \cdots, y_{i,1}, \delta_{i,j-1}, \cdots, \delta_{i1}, \boldsymbol{x_i}) = \frac{S_{j|j-1,\cdots,1}(y|\Upsilon_{ij}) - S_{j|j-1,\cdots,1}(r_{ij}|\Upsilon_{ij})}{S_{j|j-1,\cdots,1}(l_{ij}|\Upsilon_{ij}) - S_{j|j-1,\cdots,1}(r_{ij}|\Upsilon_{ij})},$$

para $l_{ij} < y \leq r_{ij}$, em que $\Upsilon_{ij} = (y_{i,j-1}, \cdots, y_{i,1}, \delta_{i,j-1}, \cdots, \delta_{i1}, \boldsymbol{x_i})$ e

$$S_{j|j-1,\cdots,1}(t|y_{i,j-1}, \cdots, y_{i1}, \delta_{i,j-1}, \cdots, \delta_{i1}, \boldsymbol{x_i}) = \left[1 + \frac{\frac{\eta_{ij}}{2}\left\{1 - \frac{1}{2\mu_{ij}\Lambda_0(t)+1}\right\}}{\omega + \sum_{k=1}^{j-1}\frac{\delta_{ik}n_{ik}}{2}\left\{1 - \frac{1}{2\mu_{ik}\Lambda_0(y_{ik})+1}\right\}}\right]^{-\omega-d_{ij}}, \tag{2.61}$$

em que $d_{ij} = \sum_{k=1}^{j-1}\delta_{ik}$, tal que particularmente, para $j = 1$, tem-se

$$S_1(y|\boldsymbol{x_i}) = \left[1 + \frac{\eta_{i1}}{2\omega}\left\{1 - \frac{1}{2\mu_{i1}\Lambda_0(y)+1}\right\}\right]^{-\omega}.$$

Conforme proposto pelos autores, como y_{ij} varia para cada iteração, utiliza-se uma restrição de cauda nula modificada, atribuindo $U_{ij} = 0$ para o caso em que a $j-$ ésima observação do $i-$ ésimo grupo é censurada além do maior limite direito de todos os intervalos finitos.

(c) Tomando-se $\boldsymbol{y} = \boldsymbol{y_h}$, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\boldsymbol{h}}$ e $\Lambda_0(t) = \hat{\Lambda}_0^{(j-1)}(t)$, gera-se $\boldsymbol{\xi_h} = (\xi_{h1}, \cdots, \xi_{hm})$ $(h = 1, \cdots, M)$ a partir da distribuição *a posteriori* de Ξ dada por

$$\Xi_i | \{(y_{ij}, \boldsymbol{x_{ij}}, \delta_{ij}),\ j = 1, \cdots, n_i\} \sim \text{Gama}(A_{ij}, B_{ij}),$$

em que

$$A_{ij} = \exp(\gamma) + \sum_{j=1}^{n_i}\delta_{ij} \qquad \text{e}$$

$$B_{ij} = \exp(\gamma) + \sum_{j=1}^{n_i}\frac{\exp(\boldsymbol{\theta'x_{ij}^{(0)}})}{2}\left\{1 - \frac{1}{2\exp(\boldsymbol{\beta'x_{ij}^{(1)}})\Lambda_0(y_{ij})+1}\right\};$$

(d)Fixando $\boldsymbol{\xi} = \boldsymbol{\xi_h}$, $\boldsymbol{y} = \boldsymbol{y_h}$, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\boldsymbol{h}}$ e $\Lambda_0(t) = \hat{\Lambda}_0^{(q-1)}(t)$, gere $\boldsymbol{k_h} = (k_{h,1,1}, \cdots, k_{h,m,n_m})$ $(h = 1, \cdots, M)$ a partir da distribuição condicional de K_i dada por

$$(K_i - \delta_i)|(y_i, \boldsymbol{x_i}, \delta_i) \sim \text{Poisson}\left(\frac{\exp(\boldsymbol{\theta'x_i^{(0)}})}{2 + 4\Lambda_0(y_i)\exp(\boldsymbol{\beta'x_i^{(1)}})}\right);$$

(e) Considerando $\boldsymbol{y} = \boldsymbol{y_h}$, $\boldsymbol{k} = \boldsymbol{k_h}$, $\boldsymbol{\alpha} = \boldsymbol{\alpha_h}$, e $\Lambda_0(t) = \hat{\Lambda}_0^{(q-1)}(t)$, gere $\boldsymbol{u_h} = (u_{h,1,1}, \cdots, u_{h,m,n_m})$ $(h = 1, \cdots, M)$ a partir da distribuição condicional de U dada por

$$U_i|(y_{ij}, \delta_{ij}, \boldsymbol{x_{ij}}, k_{ij}) \begin{cases} = 0, & \text{se } k_{ij} = 0; \\ \sim \text{Gama}(C_{ij}, D_{ij}), & \text{se } k_{ij} > 0. \end{cases}$$

em que $C_{ij} = k_{ij} + \delta_{ij}$ e $D_{ij} = 0,5 + \Lambda_0(y_{ij})\exp(\boldsymbol{\beta'x_{ij}^{(1)}})$. Além disso, faça $u_{hij} = 0$ se $y_{hij} > t^*$, em que t^* é definido por $t^* = \max_j\{r_j\delta_j\}$;

(f) Para $h = 1, \cdots, M$, maximize a função de log-verossimilhança dada por $l(\boldsymbol{\alpha}|\boldsymbol{D_h}) = \log L_C(\exp(\gamma), \boldsymbol{\theta}, \boldsymbol{\beta}|\boldsymbol{D_h})$ com $\boldsymbol{D_h} = (y_{hij}, \delta_{ij}, \boldsymbol{x_{ij}}, k_{hij}, u_{hij}; i = 1, \cdots, m; j = 1, \cdots, n_i)$ para obter as estimativas $\hat{\gamma}^{(q,h)}$, $\hat{\boldsymbol{\theta}}^{(\boldsymbol{q},\boldsymbol{h})}$, $\hat{\boldsymbol{\beta}}^{(\boldsymbol{q},\boldsymbol{h})}$, fazendo então $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(\boldsymbol{q},\boldsymbol{h})}$ para obter $\hat{\Lambda}_0^{(q,h)}(t)$ baseado na equação (2.52).

(g) Atualize as estimativas tomando

$$\hat{\gamma}^{(q)} = \frac{1}{M}\sum_{h=1}^{M}\hat{\gamma}^{(q,h)}, \quad \hat{\boldsymbol{\theta}}^{(\boldsymbol{q})} = \frac{1}{M}\sum_{h=1}^{M}\hat{\boldsymbol{\theta}}^{(\boldsymbol{q},\boldsymbol{h})},$$
$$\hat{\boldsymbol{\beta}}^{(\boldsymbol{q})} = \frac{1}{M}\sum_{h=1}^{M}\hat{\boldsymbol{\beta}}^{(\boldsymbol{q},\boldsymbol{h})} \quad \text{e} \quad \hat{\Lambda}_0^{(q)}(t) = \frac{1}{M}\sum_{h=1}^{M}\hat{\Lambda}_0^{(q,h)}(t);$$

(h) Atualize a matriz de covariâncias estimada por

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{(\boldsymbol{q})} =& \frac{1}{M}\sum_{h=1}^{M}\left[-\frac{\partial^2}{\partial\boldsymbol{\alpha}'\partial\boldsymbol{\alpha}}\left\{l_1(\exp(\gamma)) + l_2(\boldsymbol{\theta}) + l_3(\boldsymbol{\beta})\right\}\right]^{-1} \\ &+ \left(1+\frac{1}{M}\right)\sum_{h=1}^{M}\frac{(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{q},\boldsymbol{h})} - \hat{\boldsymbol{\alpha}}^{(\boldsymbol{q})})(\hat{\boldsymbol{\alpha}}^{(\boldsymbol{q},\boldsymbol{h})} - \hat{\boldsymbol{\alpha}}^{(\boldsymbol{q})})'}{M-1},\end{aligned} \tag{2.62}$$

em que as funções de log-verossimilhança l_1, l_2 e l_3 são avaliadas em $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}^{(\boldsymbol{q},\boldsymbol{h})}$ e $\boldsymbol{D} = \boldsymbol{D_h}$.

4. Repita o passo 3 até que seja alcançada a convergência para as estimativas de $\hat{\gamma}$, $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\beta}}$.

O segundo termo da expressão (2.62) contempla a variância entre imputações, conforme proposto por Rubin (1987) e utilizado no caso sem grupos. Similar ao modelo que não contempla agrupamento, o fator $(1+1/M)$ trata-se de uma correção devido ao número de imputações geradas ser finito.

Conforme a modelagem mais simples, os vetores $\boldsymbol{y_h}$ podem apresentar muito empates devido ao fato de a função de risco acumulado estimada ser uma função escada. Para contornar este problema, adota-se para a imputação de Y_{ij} a modificação anteriormente utilizada,

$$\tilde{\Lambda}_0(t) = \frac{\hat{\Lambda}_0(l_{ij})(r_{ij}-t) + \hat{\Lambda}_0(r_{ij})(t-l_{ij})}{r_{ij}-l_{ij}}, \tag{2.63}$$

em que $l_{ij} < t \leq r_{ij}$.

Conforme afirmam os autores, a modificação sugerida não induz grande viés às estimativas. O artigo de Lam e Wong (2014) apresenta um estudo do desempenho do estimador por meio de simulações em larga escala, sugerindo o uso de $M = 30$ a 50 para a obtenção de estimativas razoáveis, e valores maiores para análise de dados reais, apesar do custo computacional do algoritmo.

O estimador proposto para dados não agrupados tem seu viés, em conjunto com outras propriedades, avaliado em diferentes cenários em seções posteriores deste trabalho com o uso de simulações. Também para a extensão aqui apresentada, há um destaque na metodologia por conta dos saltos da função de risco não estarem restritos a um conjunto finito de valores, diferente de grande parte das metodologias apresentadas até então.

Assim como para o modelo anterior sem a incorporação de dados agrupados, esta extensão não apresenta uma demonstração formal da identificabilidade dos parâmetros, com as propriedades gerais do estimador sendo investigadas somente por meio de simulações em Lam e Wong (2014).

Capítulo 3

Aplicações

Neste capítulo, são exibidas três aplicações dos modelos de fração de cura previamente apresentados a conjuntos de dados reais. O primeiro dos conjuntos consiste em dados provenientes de um estudo retrospectivo com o intuito de comparar o tempo até a diminuição das mamas em mulheres que se submeteram a tratamentos envolvendo somente radioterapia e radioterapia em conjunto com quimioterapia (Finkelstein e Wolfe, 1985). Em seguida, o leitor é apresentado à análise de dados obtidos do hemocentro de São Paulo (Fundação Pró-Sangue), cujo objetivo consiste em estudar a relação entre as doações frequentes de sangue e a ocorrência de anemia, em conjunto com fatores como idade e hematócrito do doador na última doação. Por fim, apresenta-se uma aplicação associada a um conjunto de dados de migrações de cervos, destinado a responder questões a respeito da probabilidade de migração para determinados fatores climáticos e geográficos. Neste último conjunto, em especial, algoritmos para dados agrupados são aplicados a fim de controlar a variabilidade das medidas observadas a partir de um mesmo cervo.

As análises deste capítulo foram realizadas por meio do sistema estatístico R utilizando a implementação do algoritmo de Turnbull presente no pacote *interval* do repositório CRAN; as rotinas disponibilizadas por Xiang *et al.* (2011) para o modelo de mistura; o pacote *intercure*, recentemente disponibilizado no CRAN como parte deste trabalho, com rotinas baseadas nos algoritmos propostos em Liu e Shen (2009), Lam *et al.* (2013) e Lam e Wong (2014).

3.1 Dados de Câncer de Mama

Como exemplo introdutório, foram aplicados os algoritmos anteriormente apresentados ao conjunto de dados de câncer de mama encontrado em Finkelstein e Wolfe (1985), disponibilizado no pacote *interval* (Fay e Shaw, 2010) em R.

O trabalho original compara efeitos estéticos do tratamento com somente radioterapia contra radioterapia em conjunto com quimioterapia em realizados em mulheres com câncer de mama precoce. A comparação é realizada por meio de um estudo retrospectivo contemplando 46 mulheres tratadas com somente radioterapia e outras 48 tratadas com radioterapia e quimioterapia, sendo o evento de interesse definido pelo encolhimento moderado ou severo das mamas. Entretanto, o tempo exato em que a diminuição ocorre é desconhecido, limitando o experimento ao estudo dos intervalos de tempo que contêm o evento de interesse. Além disso, a diminuição pode não vir a acontecer, o que implica que existe uma parcela da população que não apresentará o evento de interesse, constituindo uma fração de cura (em que a "cura" modelada não se trata da cura biológica). O banco de dados do artigo original encontra-se reproduzido na Tabela 3.1. Medidas resumo envolvendo contagens de casos e censuras podem ser visualizadas na Tabela 3.2.

Pode-se notar, por meio da Tabela 3.2, que embora a amostra contenha quantidades próximas de observações para os diferentes tratamentos, o uso exclusivo da radioterapia apresenta um número menor de casos de redução das mamas.

Por meio do estimador não paramétrico de Turnbull, obtém-se para cada um dos grupos as estimativas das funções de sobrevivência apresentadas na Figura 3.1, possibilitando melhor repre-

Tabela 3.1: *Tempos observados em meses para dados de câncer de mama*

Radioterapia	(0,7]; (0,8]; (0,5]; (4,11]; (5,12]; (5,11]; (6,10]; (7,16]; (7,14]; (11,15]; (11, 18]; ≥ 15; ≥ 17; (17, 25]; (17, 25]; ≥ 18; (19, 35]; (18, 26]; ≥ 22; ≥ 24; ≥ 24; (25, 37]; (26, 40]; (27; 34]; ≥ 32; ≥ 33; ≥ 34; (36, 44]; (36, 48]; ≥ 36; ≥ 36; (37, 44]; ≥ 37; ≥ 37, ≥ 37; ≥ 38, ≥ 40; ≥ 45; ≥ 46; ≥ 46; ≥ 46; ≥ 46; ≥ 46; ≥ 46; ≥ 46
Radioterapia + Quimioterapia	(0,22]; (0,5]; (4,9]; (4,8]; (5,8]; (8,12]; (8,21]; (10,35]; (10,17]; (11,13]; ≥ 11; (11, 17]; ≥ 11; (11, 20]; (12, 20]; ≥ 13; (13, 39]; ≥ 13; ≥ 13; (14, 17]; (14,19]; (15,22]; (16,24]; (16,20]; (16,24]; (16,60]; (17,27]; (17,23]; (17, 26]; (18, 25]; (18, 24]; (19, 32]; ≥ 21; (22, 32]; ≥ 23; (24, 31]; (24, 30]; (30, 34]; (30, 36]; ≥ 31; ≥ 32; (32, 40]; ≥ 34; ≥ 34; ≥ 35; (35, 39]; (44, 48]; ≥ 48

Tabela 3.2: *Frequências de falhas e censuras para dados de câncer de mama*

Tratamento	N	Encolhimentos	Censuras
Radioterapia	46	21	25
Radioterapia + Quimioterapia	48	35	13

sentação dos tempos de sobrevida e incidência do evento de interesse. Para efeito ilustrativo, a Figura 3.2 apresenta também o estimador de Kaplan-Meier no mesmo gráfico, utilizando o ponto médio dos intervalos em que o evento ocorreu para o cálculo. As regiões sombreadas nos gráficos se tratam dos intervalos construídos a partir das censuras intervalares, conforme apresentado na seção 2.1.

Conforme pode ser visualizado, o tratamento com radioterapia proporciona, em geral, a retração em um tempo maior em relação ao tratamento com radioterapia e quimioterapia. Nota-se também por meio destes gráficos e dos dados originais que este conjunto não possui, para nenhum dos tratamentos, eventos censurados à direita após o maior dos tempos de observação das censuras intervalares finitas, o que impossibilita o uso do estimador de tempo de promoção, dependente desta suposição, assim como implica nas curvas estimadas pelo algoritmo de Turnbull tendendo a 0 quando toma-se $t \to \infty$, conforme Figuras 3.1 e 3.2. Caso a condição seja satisfeita, tem-se como existente um limiar de cura para a amostra e, com isso, o salto do último intervalo obtido pelo algoritmo não paramétrico pode ser visto como estimativa para a proporção de curados. As probabilidades de a retração não ocorrer, neste contexto definidas como fração de cura, são estimadas com os modelos semiparamétricos de mistura padrão e fragilidade anteriormente exibidos, em conjunto com o último "salto" da estimativa não paramétrica de sobrevivência, podendo ser visualizadas na Tabela 3.3 acompanhadas de seus respectivos erros padrão obtidos por bootstrap para o método de Turnbull e através do método delta para os demais estimadores.

Tabela 3.3: *Frações de cura estimadas para os dados de câncer de mama*

Método	Radioterapia	Radioterapia + Quimioterapia
Turnbull	0,466 (0,0883)	0,055 (0,0682)
Mistura	0,368 (0,0711)	0,051 (0,0318)
Fragilidade	0,442 (0,0796)	0,031 (0,0431)

Conforme anteriormente destacado, o algoritmo proposto por Liu e Shen (2009) apresenta uma limitação para a estimação da fração de cura, assim como diversos outros propostos na literatura: deve ocorrer ao menos um evento censurado à direita com tempo de observação superior ao da maior falha da amostra em questão. O mesmo problema foi evidenciado para o algoritmo de Turnbull com as curvas estimadas tendendo a zero. Entretanto, o uso do último "salto" da sobrevivência estimada

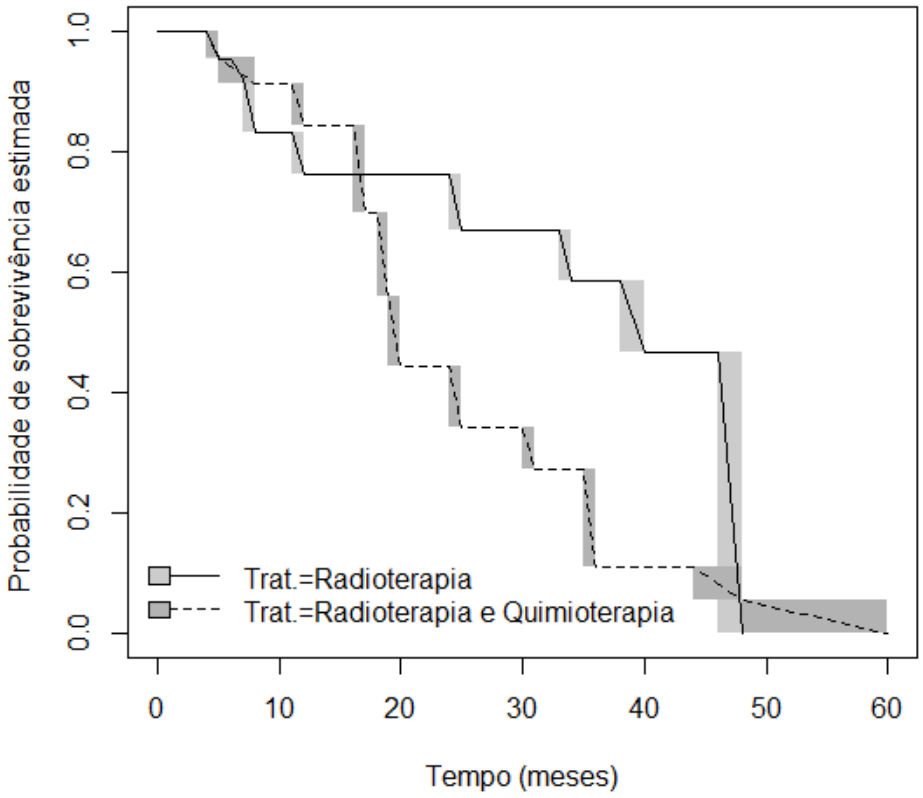

Figura 3.1: *Estimador de Turnbull para os dados de câncer de mama.*

por este algoritmo proporcionou uma estimativa razoável da fração de cura em comparação com os demais estimadores. Pode-se, desta maneira, interpretar os tempos contidos no último intervalo da curva de Turnbull como *outliers*. Naturalmente, recomenda-se uma forte motivação para o uso destes métodos, pois a suposição do limiar de cura pode ser facilmente violada por pontos aberrantes.

Para fins comparativos, foram excluídos da amostra os casos de falha cuja última inspeção ocorre em tempo superior a todos os casos de censura à direita, resultando no conjunto de estimativas da Tabela 3.4. Em particular, as estimativas para o erro padrão da proporção de cura estimada com o método de Turnbull são obtidas via bootstrap considerando $B = 300$ réplicas. Estimativas não paramétricas da curva de sobrevivência após a remoção dos indivíduos podem ser visualizadas na Figura 3.3. Para os ajustes utilizando o modelo de fragilidade adotou-se $M = 50$ e, após estabilizada a série de estimativas, foi então utilizado $M = 1000$.

Tabela 3.4: *Frações de cura estimadas para os dados de câncer de mama (removendo 3 indivíduos da amostra)*

Método	Radioterapia	Radioterapia + Quimioterapia
Turnbull	0,457 (0,0887)	0,064 (0,0829)
Mistura	0,477 (0,0745)	0,064 (0,0360)
Promoção	0,426 (0,0727)	0,117 (0,0391)
Fragilidade	0,494 (0,0789)	0,055 (0,0552)

As curvas de sobrevivência estimadas utilizando os diferentes modelos de fração de cura, além do estimador de Turnbull, são apresentadas nas Figuras 3.4 e 3.5. A suavidade das curvas obtidas pelo modelo de fragilidade deriva-se do uso da média de curvas estimadas com o uso do estimador

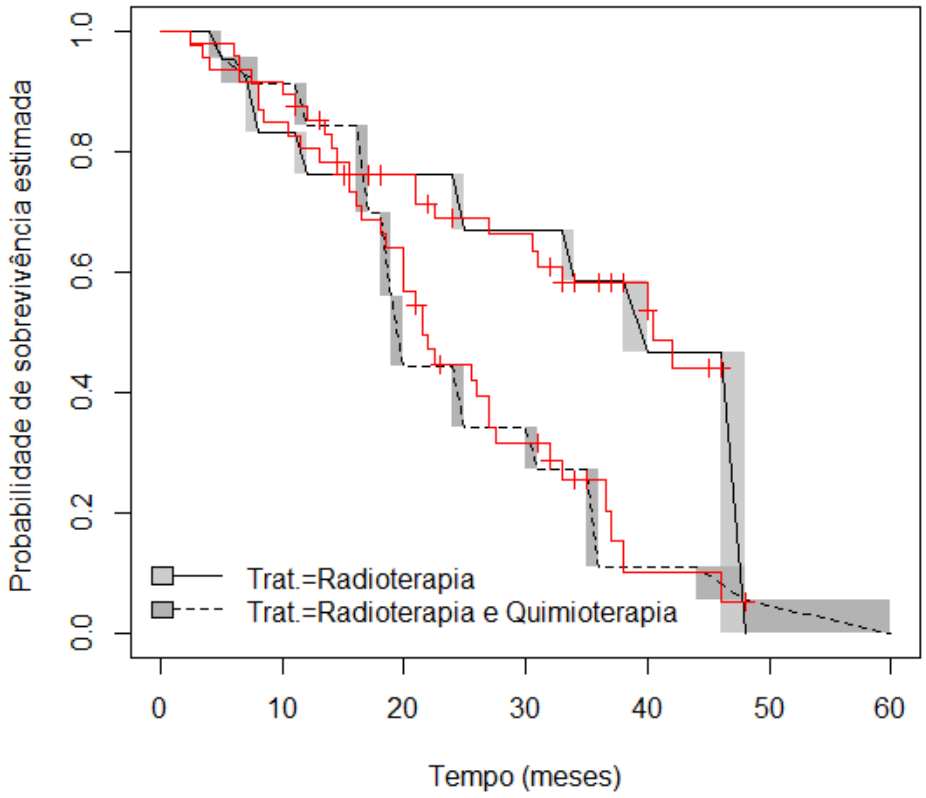

Figura 3.2: *Estimador de Turnbull (linha preta) e Kaplan-Meier com ponto médio para os dados de câncer de mama (linha vermelha).*

de Nelson-Aalen do risco acumulado e tempos inseridos para os intervalos baseados em distribuições condicionais. Pode-se notar que as curvas de sobrevivência estimadas pelo método de Turnbull para cada tratamento apresentam cruzamento entre si (vide Figuras 3.1 e 3.3). Embora as curvas de sobrevivência estimadas dos diferentes tratamentos não tenham se cruzado para os ajustes usando os modelos semiparamétricos, a construção dos modelos de mistura e fragilidade permite este comportamento em suas estimativas.

Para uma mesma amostra, diferenças podem ser visualizadas entre as estimativas obtidas com o uso de diferentes modelos. A natureza biológica do problema pode apresentar justificativas razoáveis para a escolha de um mecanismo particular de estimação da fração de curados. Entretanto, em muitas situações, tal escolha não é trivial, reduzindo as opções em função das conveniências algébricas ou computacionais de um estimador.

Os resultados dos modelos estudados neste trabalho podem levar a diferentes conclusões quanto à fração de cura, pois supõem diferentes mecanismos quanto à probabilidade de curados no conjunto de dados. Em muitas situações reais, a fração de cura é assumida como existente sem entretanto haver justificativas no contexto dos dados para uma escolha adequada do modelo. A ausência de uma especificação natural do modelo de fração de cura acarreta em dúvidas a respeito de qual metodologia usar, principalmente ao deparar-se com resultados divergentes. Tendo em vista este problema, um estudo exaustivo por meio de simulações é realizado para a avaliação da robustez dos modelos, sendo devidamente apresentado no capítulo seguinte.

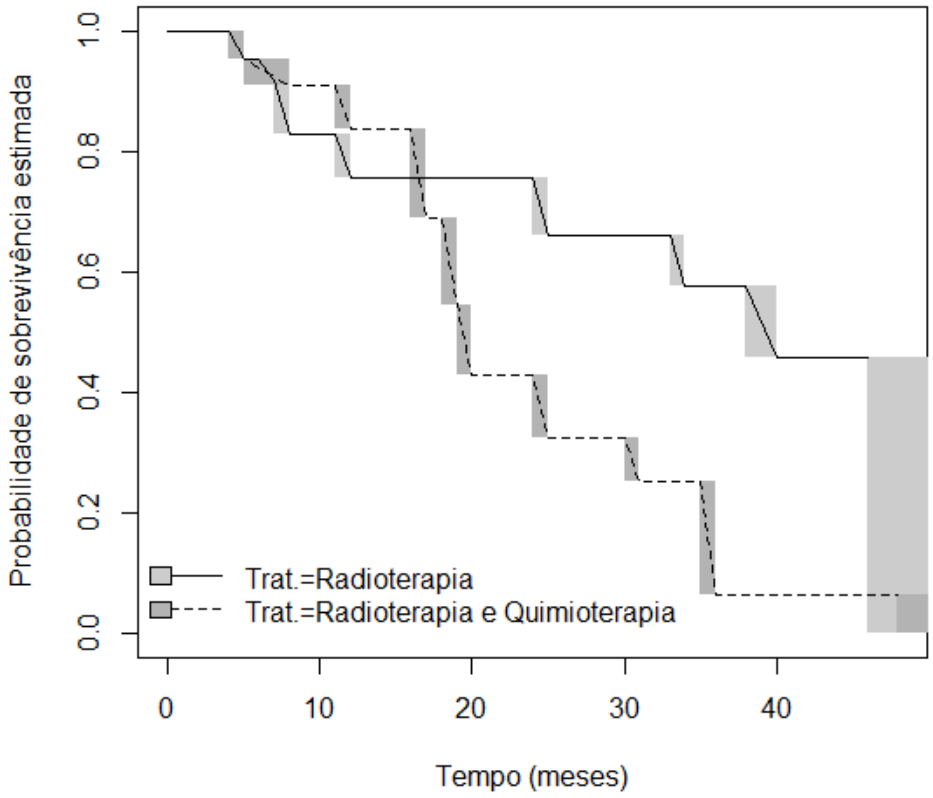

Figura 3.3: *Estimador de Turnbull após remoção de três indivíduos da amostra.*

3.2 Dados de Anemia

Um dos principais focos deste trabalho consiste na aplicação dos modelos de fração de cura anteriormente apresentados a uma base composta por dados de doações de sangue. Neste contexto, foi obtido, através de estudos conduzidos na Fundação Pró-Sangue, um conjunto de dados composto por registros de 1.600.232 doações de sangue (ou tentativas) provenientes de janeiro de 1996 a dezembro de 2006. O interesse dos pesquisadores consiste em obter o tempo até a ocorrência de anemia na população de doadores de repetição (indivíduos que doam sangue periodicamente), assim como quantificar os fatores que melhor descrevem a não suscetibilidade de determinados indivíduos, os superdoadores, à doença em estudo.

De acordo com a literatura, doações de sangue (450ml) acarretam na perda de 200 a 250mg de ferro. Deste modo, a anemia devido à deficiência de ferro no organismo é uma das complicações mais comuns entre doadores de sangue. No Brasil, permite-se que doadores do sexo masculino realizem até quatro doações por ano, com intervalo mínimo de 60 dias entre doações consecutivas. Para as mulheres, o limite permitido é de três doações por ano, com 90 dias de diferença entre doações. Na prática, estas restrições são muitas vezes desconsideradas, permitindo a ocorrência de múltiplas doações em um curto período de tempo, como o banco de dados disponibilizado apresenta.

Além disso, priorizando a saúde dos doadores e tendo em mente a minimização da ocorrência de anemia, um nível mínimo de hematócrito ou de hemoglobinas no sangue é pré-estabelecido para o processo de doação, exigindo que os doadores apresentem níveis sanguíneos acima do valor de corte definido. Para os homens, o valor de corte é de 39% de hematócrito (ou 13,0g/dL de hemoglobina), enquanto que para mulheres é estabelecido 38% (ou 12,5g/dL de hemoglobina). Indivíduos que apresentam estas condições têm suas doações sanguíneas consideradas seguras e,

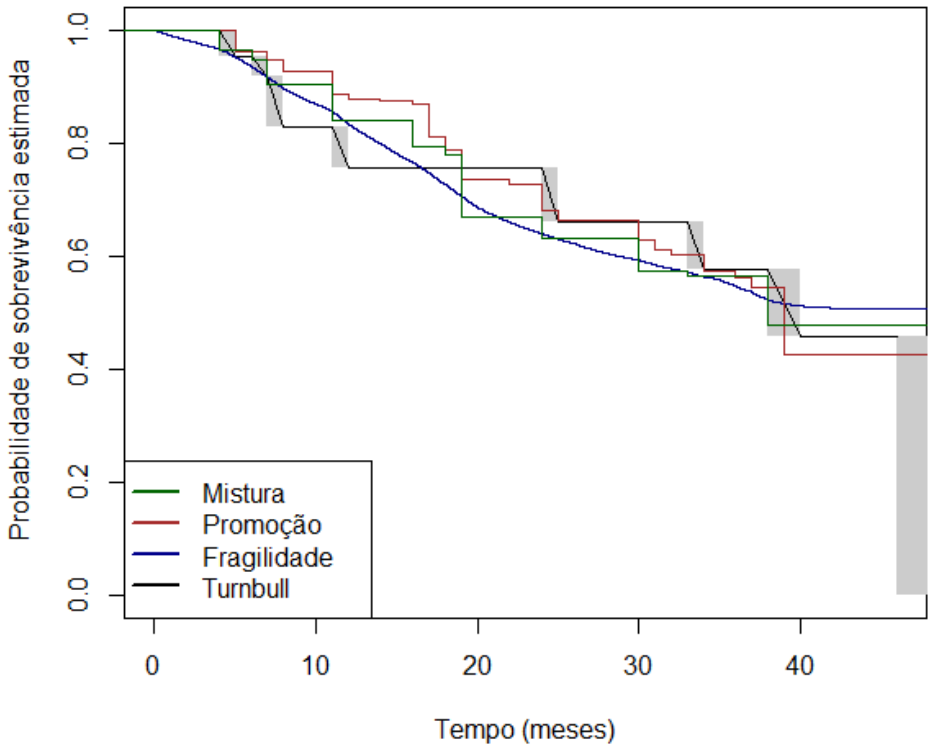

Figura 3.4: *Curvas de sobrevivência estimadas para pacientes com tratamento de somente radioterapia*

portanto, permitidas.

O problema pode ser interpretado por meio de um modelo de fração de cura com censura intervalar, considerando-se a anemia como falha, os instantes de doação como os extremos dos intervalos de censura, e os indivíduos não suscetíveis representando a fração de cura da população. Trabalha-se então na reestruturação do banco de dados: antes apresentando doações como observações, trata-se então o mesmo para a apresentação somente de um intervalo de censura ou censura à direita para cada doador como observação. Para cada doação, mede-se antes o hematócrito do indivíduo e se esta medida encontra-se abaixo de um nível crítico, definido pelos mesmos pontos de corte anteriormente mencionados, o indivíduo é considerado anêmico e consequentemente é impossibilitado de doar; caso seu hematrócito esteja acima do valor de corte definido, o procedimento de doação de sangue ocorre normalmente (Almeida *et al.*, 2013). Sob o ponto de vista estatístico, tem-se uma censura intervalar para o primeiro caso, pois sabe-se que o evento de interesse, a anemia, ocorreu entre a última doação e a atual tentativa. No segundo caso, observam-se as doações consecutivas até que ocorra a anemia ou até o fim do estudo.

Neste estudo, supõe-se que a primeira doação registrada de um mesmo indivíduo no banco de dados trata-se de sua primeira doação de sangue, definindo a origem do processo, e que este mantém suas doações realizadas no mesmo banco de sangue. Trabalha-se aqui também com a suposição de que os instantes de observação são não informativos, entretanto, é esperado que a doação de sangue em si altere a distribuição do tempo de ocorrência.

A análise dos dados foi realizada para as doações realizadas a partir de 2003, ano em que se consolida o critério de anemia utilizando o hematócrito, definindo a primeira doação a partir deste ano como início do processo. Para este conjunto, utiliza-se a idade (em anos), o número de doações nos dois anos anteriores e o hematócrito do indivíduo, variáveis coletadas anteriormente à primeira doação realizada a partir do período considerado. Embora tais quantidades tenham sido observadas

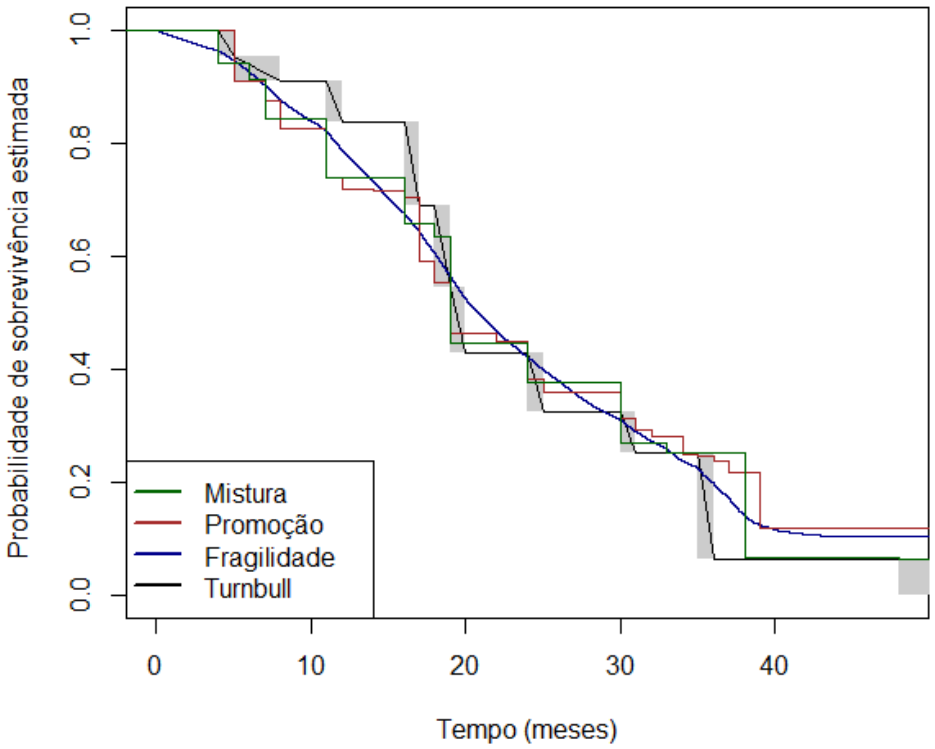

Figura 3.5: *Curvas de sobrevivência estimadas para pacientes com tratamento de radioterapia e quimioterapia*

para cada doação do conjunto, o uso de um modelo que contemple covariáveis dependentes do tempo encontra-se fora do escopo deste trabalho.

Conforme apresentado nos resultados a seguir, mesmo violando as suposições mencionadas, os ajustes derivam conclusões condizentes com o padrão esperado na área biológica.

3.2.1 Análise Descritiva

Este estudo faz uso das seguintes covariáveis para explicar o tempo até a ocorrência da anemia e sua probabilidade de cura: o número de doações realizado nos dois últimos anos associado ao indivíduo no início do estudo (basal); a idade no instante da primeira doação após 2003; e o hematócrito no instante da primeira doação após 2003. Os dados de doações iniciando-se em 2003 têm algumas medidas resumo como a média das covariáveis com seus respectivos desvios padrão apresentadas na Tabela 3.5. As informações apresentadas são descritas por:

- Sexo: Gênero do doador (M para Masculino e F para Feminino);
- N: Tamanho da amostra;
- Doações: Contagem de doações do indivíduo nos últimos dois anos observada na primeira doação a partir de 2003;
- Hct basal: Hematócrito observado anteriormente à primeira doação do indivíduo a partir de 2003;
- Idade basal: Idade, em anos, do doador na primeira doação após 2003;

- Prop. Censuras: Proporção de eventos censurados na amostra.

Tabela 3.5: *Medidas Resumo para Dados de Doadores de Sangue*

Sexo	N	Doações	Hct basal (%)	Idade basal (anos)	Prop. Censuras
F	32503	1,11 (1,151)	41,53 (2,555)	32,96 (10,587)	0,89
M	57228	1,25 (1,570)	45,30 (2,744)	33,50 (9,678)	0,99

Conforme pode ser observado pela Tabela 3.5, o evento de anemia induzida por doação pode ser interpretado como evento raro para a população masculina. Para as mulheres, a proporção observada é de 11%, corroborando o esperado: mulheres são mais suscetíveis à anemia por doações sanguíneas do que os homens.

Ambos os gêneros apresentam, para esta amostra, hematócritos e idades próximos, assim como a contagem de doações nos últimos dois anos. Devido à literatura médica apresentar evidências quanto à grande diferença do comportamento do organismo de homens e mulheres, a análise inferencial é segmentada para os diferentes sexos, implicando em um ajuste diferente dos modelos apresentados para cada gênero.

Na Figura 3.6, é apresentado o gráfico das funções de sobrevivência estimadas pelo estimador de Turnbull para este conjunto de dados. Observa-se através deste uma sobrevida menor para mulheres em relação aos homens. O gráfico evidencia, assim como visto nas tabelas, uma maior proporção de indivíduos não suscetíveis na população masculina.

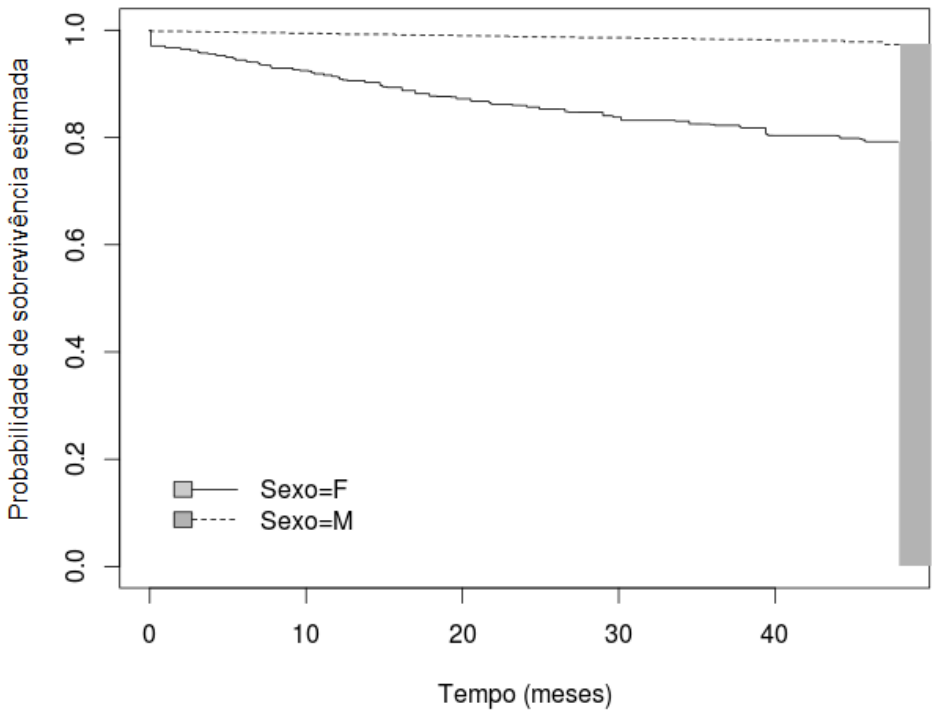

Figura 3.6: *Curvas de Turnbull para os dados de anemia*

A fim de estudar a fundo as relações de suscetibilidade e risco de desenvolvimento de anemia com as demais covariáveis do estudo, são realizadas a seguir análises inferenciais utilizando modelos semiparamétricos apresentados em seções anteriores deste trabalho.

3.2.2 Análise Inferencial

Nesta seção são apresentados resultados decorrentes dos algoritmos propostos em Liu e Shen (2009) e Lam *et al.* (2013). Para os dados estudados nesta aplicação, o estimador de mistura padrão proposto em Xiang *et al.* (2011) apresenta dificuldades computacionais para a estimação da função de sobrevivência por conta do número de parâmetros a serem estimados, implicando em tempos muito altos de processamento por iteração e inviabilizando então a aplicação deste modelo aos dados de doações sanguíneas.

Todas as análises aqui apresentadas fazem uso das covariáveis idade basal categorizada (variável indicadora assumindo 1 para idade superior ou igual a 50 anos), hematócrito basal e quantidade de doações nos dois últimos anos. As análises são particionadas por gênero por conta das diferenças de comportamento do organismo para os homens e mulheres, conforme o conhecimento de pesquisadores da área.

Utilizando-se o modelo de tempo de promoção para dados com censura intervalar proposto em Liu e Shen (2009), obtém-se as estimativas, em conjunto com seus erros padrão, apresentadas na Tabela 3.6. Para a obtenção das mesmas, foi adotada a diferença máxima de $0,0001$ entre os efeitos estimados de cada iteração para critério de convergência. Em virtude do alto tempo computacional por iteração e do volume de dados analisado, a convergência da log-verossimilhança esperada é desconsiderada nestes ajustes, adotando-se a diferença máxima de $0,001$ para o vetor de probabilidades para cada iteração. Para este problema em particular, devido às estimativas da fração de cura utilizarem apenas as estimativas associadas aos efeitos de interesse, a convergência considerada do vetor de efeitos estimados é assumida razoável.

Tabela 3.6: *Estimativas obtidas pelo estimador tempo de promoção*

Sexo	Intercepto	$\beta_{Contagem}$	β_{Hct}	β_{Idade}
Feminino	8,550 (0,2350)	-0,007 (0,0109)	-0,243 (0,0058)	-0,234 (0,0489)
Masculino	14,324 (0,4672)	0,048 (0,0151)	-0,414 (0,0110)	0,545 (0,0781)

A interpretação dos parâmetros no modelo de tempo de promoção pode ser feita por meio do risco relativo, pois a estrutura adotada nesse estimador tem a vantagem de apresentar a propriedade da proporcionalidade dos riscos. Dessa forma, com base nas estimativas e a precisão apresentada na Tabela 3.6, para o conjunto de mulheres doadoras e fixadas as demais covariáveis, o aumento de uma unidade no hematócrito basal implica na diminuição do risco de anemia em 21,57%; idades iguais ou superiores a 50 anos proporcionam diminuição de 20,86% em relação à classe de idades inferiores a 50 anos; o aumento de uma unidade no número de doações ocorridas nos dois anos anteriores em relação ao início do estudo implicam na diminuição de 0,70% do risco. Tais resultados são diretamente observáveis exponenciando-se a estimativa obtida, derivados do risco obtido a partir da expressão 2.31, com a interpretação análoga das demais estimativas.

Uma informação interessante a ser extraída da mesma tabela consiste no sinal obtido para o efeito da idade para mulheres mais velhas: a idade superior a 50 anos apresenta maior tempo até a anemia para mulheres. Nos homens, o efeito tem natureza contrária, apresentando menor sobrevivência para idosos. Este efeito para mulheres pode ser explicado no contexto biológico pela menopausa e é esperado pelos pesquisadores.

Como consequência dos resultados apresentados com o uso do estimador de tempo de promoção de Liu e Shen (2009), pode-se derivar estimativas das frações de cura utilizando-se o limite da expressão (2.31), apresentadas na Tabela 3.7, fixando-se, por exemplo, os valores médios para cada covariável associada a cada sexo (obtidas das tabelas apresentadas na análise descritiva). Para fins ilustrativos, tem-se para as mulheres:

$$1 - \pi_i = 1 - \pi(\boldsymbol{x}_i) = \exp(-e^{8,550 - 0,007 \times 1,11 - 0,243 \times 41,53}) = 0,809.$$

Ou seja, mulheres com idade inferior a 50 anos, hematócrito de 41,53 unidades e $1,11$ doações realizadas nos últimos dois anos apresenta probabilidade de $0,809$ de não ser suscetível a anemia por

doações sanguíneas. Analogamente obtém-se, para ambos os sexos e faixas etárias, as estimativas exibidas na Tabela 3.7 para os valores médios de hematócrito e doações nos dois últimos anos de cada gênero.

Em particular, para o modelo apresentado em Liu e Shen (2009) cuja fração de cura é expressa por $\exp(-e^{\boldsymbol{\beta}'\boldsymbol{x}_i})$, pode-se interpretar e^{β} como o aumento relativo do logaritmo da probabilidade de cura. Entretanto, por conta da interpretação obtida ser confusa e não imediata, opta-se apenas pela comparação direta das frações de cura obtidas, conforme a Tabela 3.7. Os erros padrão associados às estimativas das proporções de curados, assim como para as demais aplicações deste capítulo, foram estimados utilizando o método delta.

Tabela 3.7: *Frações de cura estimadas pelo modelo tempo de promoção*

Faixa Etária	Sexo	Fração de Cura
< 50	Feminino	0,812 (0,0024)
< 50	Masculino	0,987 (0,0004)
≥ 50	Feminino	0,848 (0,0067)
≥ 50	Masculino	0,978 (0,0016)

Nota-se que, para doadores do sexo masculino com idade inferior a 50 anos, hematócrito médio e frequência de doação média, a proporção de anêmicos estimada através do modelo de tempo de proporção é de $1,3\%$, evidenciando a raridade da anemia induzida por doações para os homens.

Para as estimativas obtidas pelo modelo de fragilidade de Lam *et al.* (2013), duas tabelas são utilizadas para o conjunto de estimativas: uma para aquelas referentes à fração de cura (incidência), apresentadas na Tabela 3.8, e uma segunda (Tabela 3.9) para estimativas associadas ao preditor da regressão de Cox e diretamente relacionadas ao risco, denominadas estimativas de latência na terminologia do autor.

Para os dados associados ao sexo feminino, foi adotado $M = 200$ para as primeiras 130 iterações, com aumento para $M = 400$ para as próximas 100 iterações, adotando como estimativa final a última destas. Embora o processo tenha se mostrado estável, a convergência das estimativas não foi obtida para grande rigor, com diferenças máximas inferiores a $0,01$ para os parâmetros associados à fração de curados.

Em relação aos dados de indivíduos do sexo masculino, adotou-se, assim como para o sexo feminino, $M = 200$ para as 130 primeiras iterações e $M = 400$ para as próximas 100. Adicionalmente, para resultados conservadores utilizou-se $M = 500$ para as 20 iterações seguintes, observando diferença máxima inferior a $0,01$ para os efeitos associados à proporção de curados. É importante ressaltar que, mesmo aumentando consideravelmente o número de imputações por iteração e como consequência o custo computacional, a convergência não foi observada para critérios mais rigorosos. Entretanto, mesmo sem tal rigor, o estimador do modelo de fragilidade apresenta desempenho satisfatório para muitos cenários, como apresentado posteriormente nos estudos de simulação.

Tabela 3.8: *Estimativas dos efeitos relacionados à fração de cura usando o modelo de fragilidade*

Sexo	$\theta_{Intercepto}$	$\theta_{Contagem}$	θ_{Hct}	θ_{Idade}
Feminino	7,874 (0,3098)	0,041 (0,0187)	-0,219 (0,0080)	-0,204 (0,0752)
Masculino	13,082 (0,7159)	0,114 (0,0255)	-0,387 (0,0172)	0,565 (0,1267)

As estimativas obtidas por meio do modelo de fragilidade associadas à incidência (fração de cura) apresentam efeitos com o mesmo sinal com exceção da idade: para os homens, a faixa etária dos 50 anos ou mais proporciona o aumento da ocorrência de anemia. Para as mulheres, a mesma faixa etária apresenta menor probabilidade de anemia.

Tratando-se das estimativas associadas ao risco, existem evidências de que um maior número de doações nos últimos dois anos proporciona risco menor para as mulheres, com este efeito sendo não significativo para os homens. O aumento do hematócrito, conforme esperado, apresenta maior sobrevivência para ambos os gêneros.

Tabela 3.9: *Estimativas dos efeitos relacionados ao risco usando o modelo de fragilidade*

Sexo	$\beta_{Contagem}$	β_{Hct}	β_{Idade}
Feminino	-0,087 (0,0419)	-0,077 (0,0112)	-0,174 (0,1441)
Masculino	-0,001 (0,0546)	-0,106 (0,0253)	-0,128 (0,2744)

Dado um mesmo indivíduo (fixando u_i), o modelo de fragilidade também contempla a estrutura de proporcionalidade entre os riscos, fornecendo o risco relativo através de somente $e^{\hat{\beta}}$. Deste modo, para um mesmo indivíduo i do sexo feminino e suscetível a anemia, o incremento do hematócrito em uma unidade implica na diminuição de 7,41% do risco; e o aumento de uma unidade na contagem de doações de sangue realizadas nos dois últimos anos implica em risco 8,33% menor. Da expressão (2.49), pode-se notar que os parâmetros associados à fração de cura exercem influência sobre o risco populacional, entretanto, os efeitos estimados não apresentam interpretação direta para este caso.

Conforme visto anteriormente, utilizando o modelo de fragilidade, pode-se estimar a fração de cura com $\lim_{t\to\infty} S(t|\eta_i, \boldsymbol{x}_i^{(1)}) = \exp(-\eta_i/2)$, em que $\eta_i = \exp(\boldsymbol{\theta'}\boldsymbol{x}_i^{(0)})$. Desta expressão, deriva-se que o incremento de uma unidade para uma determinada covariável com efeito sobre a fração de cura implica no aumento (ou diminuição) relativo de $1 - e^{\hat{\theta}}$ no logaritmo da probabilidade de cura. Devido à interpretação confusa, assim como para o estimador de tempo de promoção, opta-se por apenas comparar os sinais dos parâmetros estimados em conjunto com a avaliação das estimativas obtidas para a fração de cura da Tabela 3.10.

Para fins comparativos, foram estimadas, utilizando o modelo de fragilidade, as frações de cura para homens e mulheres considerando-se fixados, para cada sexo, os valores médios de hematócrito e contagem de doações nos dois anos anteriores à primeira doação. Os resultados são exibidos na Tabela 3.10.

Tabela 3.10: *Frações de cura estimadas pelo modelo de fragilidade*

Faixa Etária	Sexo	Fração de Cura
< 50	Feminino	0,855 (0,0118)
< 50	Masculino	0,993 (0,0010)
≥ 50	Feminino	0,880 (0,0129)
≥ 50	Masculino	0,988 (0,0023)

Como pode ser visto pelos resultados apresentados, o modelo de Lam *et al.* (2013) estima em $85,5\%$ indivíduos do sexo feminino com idade inferior a 50 anos não suscetíveis à anemia por repetidas doações. A estimativa para a mesma base utilizando o estimador de tempo de promoção proposto em Liu e Shen (2009) é de $81,2\%$ mulheres não suscetíveis com idade menor do que 50 anos. Para os homens com idade inferior a 50 anos, o estimador do modelo de tempo de promoção estima a proporção de curados em $98,7\%$, contra a estimativa de $99,3\%$ de curados proveniente do modelo de fragilidade.

Em todo caso, conforme visto anteriormente, o uso de diferentes especificações quanto à fração de cura resultou em estimativas consideravelmente diferentes para a proporção de curados. Em situações como esta, em que não se sabe o real mecanismo de cura dos dados, é natural o surgimento de questões quanto à robustez de cada modelo. Esse tipo de problema motiva um estudo aprofundado quanto às propriedades inferenciais da fração de cura de cada modelo, avaliadas neste trabalho por meio de simulações.

3.3 Dados de Migração

A observação periódica de animais constitui uma prática comum nas áreas de recursos naturais e biologia, definindo os estudos de telemetria. Neste contexto, muitas pesquisas utilizam rádio-colares com o intuito de monitorar periodicamente a atividade de um conjunto de animais com o objetivo

de extrair informações destes para determinado habitat ou condição climática.

Uma contribuição nesta área foi dada pelos autores Fieberg e DelGiudice (2008), que realizaram um estudo envolvendo modelos de fração de cura para avaliar o comportamento de um conjunto de cervos de cauda branca (*Odocoileus virginianus*) quanto à migração dadas determinadas condições climáticas em uma área de estudo de 1865 km^2 contida na Floresta Nacional de Chippewa, Estados Unidos. Os autores do trabalho original disponibilizaram o conjunto de dados utilizado em seus estudos, possibilitando as análises apresentadas posteriormente. A contribuição deste trabalho diferencia-se da já apresentada pelos autores por utilizar diferentes abordagens para modelar a fração de cura e, para dois algoritmos em especial, considerar variáveis latentes para modelar o efeito de cervo sobre os dados descritos a seguir.

Um total de 357 observações provenientes de 168 cervos diferentes, observados no período compreendido entre os anos de 1991 a 2006, foi analisado. Tais cervos foram capturados em áreas invernais de estudo localizadas no centro-norte de Minnesota (DelGiudice, 1998) durante os meses de janeiro a março dos anos estudados por meio de armadilhas e armas para enredar os animais, sendo posteriormente monitorados por meio de rádio-colares (DelGiudice *et al.*, 2005). O equipamento do rádio-colar pode variar como sendo de frequência muito alta, demandando observações das localizações por meio de aeronaves uma a três vezes por semana, ou como contendo sistema de posicionamento global (*global positioning system, GPS*), permitindo monitoramentos diários com determinações das localizações obtidas em períodos de uma a quatro horas. Os cervos são acompanhados até a morte ou a falha do equipamento, com novos cervos capturados anualmente para a substituição destes.

Movimentos são definidos pelos pesquisadores como migratórios quando o deslocamento é igual ou superior a dois quilômetros e realizados a partir de uma área de primavera, verão ou outono (resumidamente referidas como área de verão no artigo original) para uma área de inverno distintamente separadas e sem sobreposição, definindo assim o evento de interesse. Tais eventos de migrações são justificados por questões nutricionais ou benefícios antipredatoriais obtidos durante o inverno ao realizar a migração, conforme ressaltam os autores. Em seu trabalho original, estes supõem que os cervos que migram para uma área invernal sempre retornarão para as áreas de verão, suposição também incorporada neste trabalho. Mais detalhes do experimento como a descrição do relevo, metodologia e equipamentos para captura e histórico de predadores da região são descritos em Fieberg e DelGiudice (2008), DelGiudice *et al.* (2005) e Fieberg *et al.* (2008).

Para este estudo, considerou-se cada ano dos dados separadamente de modo que um mesmo cervo pode não apresentar migração em um certo ano e então ter migrando no ano seguinte, constituindo duas observações diferentes. Em virtude de tratar o tempo observado como um problema de análise de sobrevivência, e para isso obter uma escala temporal bem definida, a origem do processo é adotada para cada cervo como sendo o dia 20 de outubro do ano em questão, empiricamente determinada por constituir a data que antecede a migração da área de verão de qualquer um dos cervos do conjunto para qualquer ano em todo o período de 15 anos do estudo.

Além dos tempos de observação dos cervos, o conjunto de dados tem como variáveis o ano da captura, a área de estudo, e o índice de severidade do inverno associado ao ano de migração (wsi), construído como se segue: durante o período de 1 de novembro a 31 de abril, acumula-se um ponto para cada dia com temperatura ambiente inferior a -17,7 graus Celsius, e um ponto adicional para cada dia que apresentou neve com profundidade superior a 38 centímetros, implicando índices de maior valor associados a invernos mais severos.

Devido ao fato de os animais com rádio-colares não serem acompanhados continuamente, sabe-se somente que os tempos de migração estão contidos em um intervalo de tempo $[L_i, R_i]$, em que L_i representa o último instante de observação anterior à migração do cervo i e com R_i representando o primeiro instante de observação após o evento, portanto, o uso de técnicas que contemplam a estrutura de censura intervalar faz-se conveniente para este problema. Alguns cervos, entretanto, podem não migrar, com tal caso representado como uma censura à direita no conjunto de dados em questão. Neste caso, a probabilidade de um cervo não migrar constitui a fração de cura do problema em questão.

O viés de seleção dos cervos capturados é avaliado em Fieberg e Conn (2014) com o uso de cadeias de Markov ocultas em conjunto com um modelo logístico, entretanto, tais questões são tomadas como fora do escopo deste trabalho. Além disso, Fieberg *et al.* (2008) lidam com três possíveis tipos de cervo: não migratórios, condicionalmente migratórios e migratórios. Estas classificações são desconsideradas neste trabalho, assumindo que qualquer cervo do estudo apresenta um potencial de migração (migrador condicional), conforme simplificado em Fieberg e DelGiudice (2008).

3.3.1 Análise Descritiva

O conjunto de dados de migração aqui utilizado apresenta as seguintes variáveis: L, última observação anterior ao evento de interesse, considerando 20 de outubro como origem; o instante, R, da primeira observação após o evento de interesse; o índice de severidade do inverno no ano da observação; o indicador de censura; a área de estudo, correspondendo a uma das quatro diferentes regiões ("D", "I", "S", ou "W") contidas na Floresta Nacional de Chippewa; e por fim, o ano de captura.

Na Tabela 3.11 é apresentado um resumo das observações contidas no conjunto de dados disponibilizado. Através desta pode-se obter o tamanho da amostra, a proporção de censuras, o número de cervos diferentes e o índice médio de intensidade de inverno com seu respectivo desvio padrão.

Tabela 3.11: *Medidas Resumo de Dados de Migração*

N	Proporção de censuras	Cervos distintos	Índice de severidade invernal (wsi) médio
357	0,61	168	86,06 (49,07)

A variável de área de estudo encontra-se resumida na Tabela 3.12. Pode-se observar uma predominância de observações provenientes da região "W".

Tabela 3.12: *Frequência absoluta por área de estudo*

Área de estudo	N
D	31
I	95
S	111
W	120

Além disso, o conjunto de dados apresenta a informação do ano de monitoramento, que varia de 1991 a 2005, conforme pode ser visto em Tabela 3.13.

É apresentada na Figura 3.7 a função de sobrevivência estimada pelo algoritmo de Turnbull para tal conjunto de dados. Na Figura 3.8 exibem-se estimativas das curvas de sobrevivência estratificando-se os dados por área de estudo.

Pela curva obtida por meio do estimador não paramétrico de Turnbull em conjunto com os conhecimentos *a priori* do pesquisador, pode-se observar a necessidade de contemplar nos modelos de sobrevivência a possibilidade de os cervos não migrarem ou, neste contexto, existir uma proporção de curados. Na subseção a seguir são apresentadas estimativas de efeitos associados ao risco e à fração de cura, juntamente de suas interpretações, utilizando os modelos apresentados neste trabalho.

3.3.2 Análise Inferencial

Inicialmente foi ajustado um modelo de fração de cura para dados com censura intervalar utilizando o mecanismo de mistura padrão, conforme apresentado em Xiang *et al.* (2011), obtendo então os resultados da Tabela 3.14 para os efeitos relativos à fração de cura com seus respectivos erros padrão, com os efeitos relacionados ao tempo de sobrevida apresentados na Tabela 3.15, também acompanhados de seus erros padrão.

Tabela 3.13: *Frequência absoluta por ano de captura dos cervos em estudo*

Ano de Captura	N
1991	25
1992	26
1993	33
1994	32
1995	8
1996	44
1997	37
1998	5
1999	37
2000	1
2001	59
2002	24
2003	3
2004	10
2005	13

Tabela 3.14: *Parâmetros relacionados à fração de cura estimados utilizando o modelo de mistura padrão*

Parâmetro	Intercepto	b_{wsi}	b_{AreaI}	b_{AreaS}	b_{AreaW}
Estimativa	-2,071 (0,4974)	0,023 (0,0034)	0,416 (0,4810)	0,840 (0,4734)	1,055 (0,4707)

Tabela 3.15: *Parâmetros relacionados ao risco estimados utilizando o modelo de mistura padrão*

Parâmetro	β_{wsi}	β_{AreaI}	β_{AreaS}	β_{AreaW}
Estimativa	0,004 (0,0012)	0,453 (0,1966)	0,295 (0,1694)	0,281 (0,1594)

Segundo a Tabela 3.15, pode-se observar um aumento relativo de 0,4% no risco de migração para o aumento de uma unidade do índice de severidade. A área de estudo "W" apresenta risco de migração 32,45% maior em relação à área "D", aqui tomada como referência. Conforme visualizado em Figura 3.8, as áreas "I", "S" e "W" apresentaram uma menor proporção de cervos não migrantes, com a área "W" sendo aquela com maior índice de migração, efeito melhor evidenciado pelas estimativas da Tabela 3.14. Os ajustes foram obtidos adotando-se como convergência a diferença absoluta máxima de 0,0001 entre as esperanças das quantidades latentes do processo iterativo.

Na Tabela 3.14 estão apresentadas as estimativas dos parâmetros associados ao modelo para a probabilidade de um cervo, em um dado ano, migrar. Como utilizou-se uma ligação logito (vide expressão 2.9), então a interpretação pode ser feita por meio da razão de chances: a chance de um cervo não migrar é 65,18% menor na área de estudo "W" em relação à área de estudo "D"; o incremento de uma unidade no índice de intensidade do inverno (wsi) implica em uma chance 2,29% menor de não migrar. As razões de chances obtidas são exibidas na Tabela 3.16 em conjunto com seus respectivos intervalos de confiança (I.C.) de 95%.

Tabela 3.16: *Razões de chances associadas aos efeitos estimados*

Covariável	e^{-b}	I.C. de 95%
wsi	0,977	[0,971 ; 0,984]
Área I	0,660	[0,257 ; 1,693]
Área S	0,432	[0,171 ; 1,092]
Área W	0,348	[0,248 ; 1,571]

Para fins ilutrativos, fixa-se a área "W" e índice de severidade médio observado na migração (86,06), obtendo-se a fração de cura abaixo:

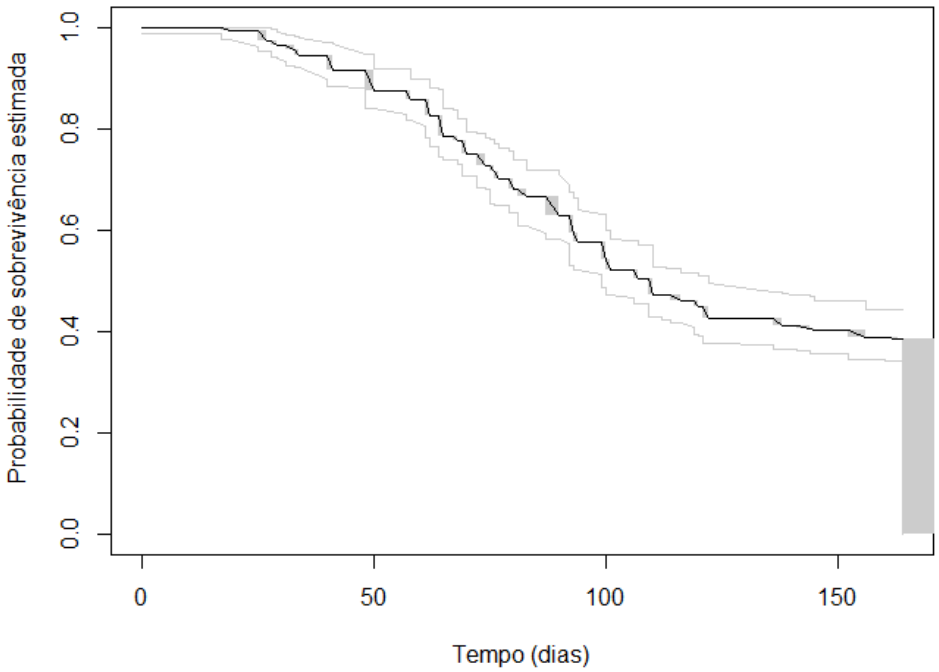

Figura 3.7: *Estimador de Turnbull para os dados de migração*

$$1-\pi_i = 1-\pi(x_i) = 1-\frac{\exp(-2,071+0,023\times 86,06+1,055)}{1+\exp(-2,071+0,023\times 86,06+1,055)} = 0,276.$$

Ou seja, 27,6% dos cervos não migram da área "W" quando o índice do inverno (wsi) encontra-se em 86,06 unidades. Em resumo, as probabilidades de não migração utilizando o estimador de Xiang *et al.* (2011), fixado índice invernal médio, são dadas na Tabela 3.17. A diferença entre a estimativa da tabela e a apresentada no exemplo ilustrativo é dada por conta do uso de mais casas decimais no cálculo de estimativas da fração de cura da Tabela 3.17. Os erros padrão exibidos em conjunto com as estimativas pontuais foram estimados por meio do método delta.

Tabela 3.17: *Frações de cura estimadas pelo modelo de mistura padrão*

Área	Fração de Cura
D	0,519 (0,1048)
I	0,416 (0,0572)
S	0,318 (0,0491)
W	0,273 (0,0438)

Mantendo-se a mesma amostra e o mesmo conjunto de covariáveis, foi realizado um ajuste utilizando o modelo de tempo de promoção proposto em Liu e Shen (2009), com resultados apresentados na Tabela 3.18. Considerando uma tolerância de $0,1$ para a diferença entre os logaritmos das verossimilhanças esperadas, obteve-se diferença máxima inferior a 10^{-15} para as estimativas dos efeitos e inferior a 10^{-4} para o vetor de probabilidades estimado. Por meio de tal ajuste, pode-se observar um aumento relativo médio de 1 % (dado por $e^{0,010}-1$) do risco de migração ao aumentar-se o índice de severidade do inverno em uma unidade. Do mesmo modo, as áreas de estudo "I", "S" e "W" apresentam risco 54,65%, 60,96% e 78,07% maior, respectivamente, em relação à área de estudo "D".

Foi então obtida, para fins ilustrativos, a fração de cura para o índice de severidade médio da

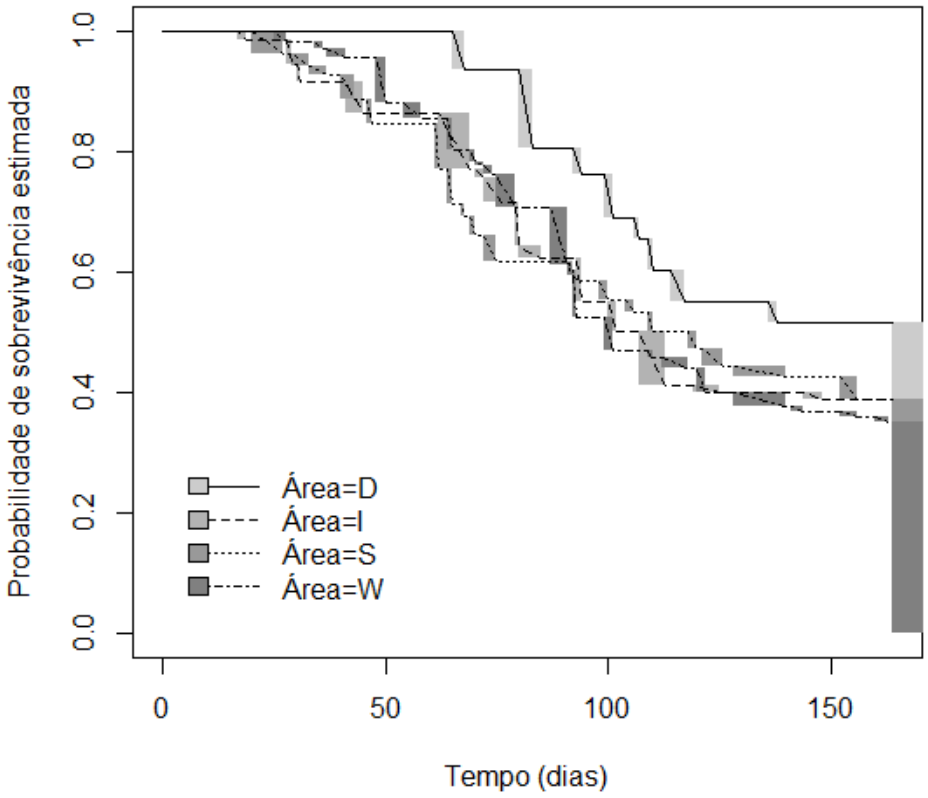

Figura 3.8: *Estimador de Turnbull para os dados de migração com estratificação por área de estudo*

Tabela 3.18: *Parâmetros estimados utilizando o modelo tempo de promoção*

Parâmetro	Intercepto	β_{wsi}	β_{AreaI}	β_{AreaS}	β_{AreaW}
Estimativa	-1,362 (0,2881)	0,010 (0,0012)	0,436 (0,2880)	0,476 (0,2834)	0,577 (0,2805)

amostra sob a área de estudo "W" utilizando o modelo de tempo de promoção:

$$1 - \pi_i = 1 - \pi(\boldsymbol{x}_i) = \exp(-e^{-1{,}362+0{,}010\times 86{,}06+0{,}577}) = 0,340.$$

O resultado acima implica que, para a área "W" com índice de severidade médio, 34% dos cervos não migram. Considerando-se precisamente os parâmetros estimados, obtém-se frações de cura para as outras áreas do estudo fixando-se o índice de severidade como a média deste. As estimativas, em conjunto com seus erros padrão estimados utilizando o método delta, são apresentadas na Tabela 3.19, diferenciando-se do exemplo acima por considerar mais algarismos significativos para os efeitos estimados no cálculo da fração de cura.

Tabela 3.19: *Frações de cura estimadas pelo modelo de tempo de promoção*

Área	Fração de Cura
D	0,537 (0,0861)
I	0,382 (0,0488)
S	0,368 (0,0450)
W	0,331 (0,0415)

Diferenças razoáveis podem ser notadas entre as estimativas de fração de cura obtidas por meio

do algoritmo de mistura padrão e tempo de promoção, conforme visualizado nas Tabelas 3.17 e 3.19, evidenciando a necessidade de uma motivação natural na escolha do mecanismo de fração de cura.

Utilizando o modelo de fragilidade de Lam *et al.* (2013), obtém-se as estimativas apresentadas nas Tabelas 3.20 e 3.21, referentes à incidência e latência do evento de migração, respectivamente. Em virtude da amostra de tamanho menor, além das 30 estimativas para "aquecimento" do processo, foram utilizadas 100 iterações com $M = 200$, seguidas por 30 iterações com $M = 1000$, 5 iterações com $M = 2000$ e finalmente 5 iterações com $M = 3000$. Mesmo com valores altos para M, implicando em custo computacional altamente intensivo, o conjunto de efeitos estimados não apresentou convergência com o rigor estabelecido de $0,001$ de diferença máxima entre as estimativas das diferentes iterações. Entretanto, o processo mostrou-se estável e com convergência de $0,01$ para estimativas associadas à fração de cura.

Tabela 3.20: *Parâmetros relacionados à fração de cura estimados utilizando o modelo de fragilidade*

Parâmetro	Intercepto	θ_{wsi}	θ_{AreaI}	θ_{AreaS}	θ_{AreaW}
Estimativa	-0,596 (0,3093)	0,010 (0,0022)	0,236 (0,3225)	0,279 (0,3052)	0,455 (0,3144)

Tabela 3.21: *Parâmetros relacionados ao risco estimados utilizando o modelo de fragilidade*

Parâmetro	β_{wsi}	β_{AreaI}	β_{AreaS}	β_{AreaW}
Estimativa	-0,003 (0,0038)	0,995 (0,6681)	1,100 (0,7217)	0,576 (0,6239)

Ao fixar-se o efeito de fragilidade dado por u_i no modelo em questão, tem-se que o risco de migração na área de estudo "W" é 77,82% maior em relação à área "D". O aumento de uma unidade do índice de severidade invernal implica em risco de migração 0,27% menor, contradizendo o aumento do risco observado utilizando-se as modelagens anteriores. Entretanto, tais efeitos apresentam-se como não significativos quando consideram-se os erros padrão estimados.

Tabela 3.22: *Frações de cura estimadas pelo modelo de fragilidade*

Área	Fração de Cura
D	0,504 (0,0941)
I	0,420 (0,0770)
S	0,404 (0,0630)
W	0,339 (0,0732)

Novamente, para fins ilustrativos, obtém-se a proporção de curados associada ao índice invernal de 86,06 unidades para cervos da área "W" para as estimativas arredondadas apresentadas na Tabela 3.20:

$$1 - \pi_i = \exp(-\eta_i/2) = \exp(-e^{-0,596+0,010\times 86,06+0,455}/2) = 0,358.$$

As estimativas das frações de cura considerando maior precisão, exibidas na Tabela 3.22, apresentaram menor diferença em relação àquelas obtidas pelo modelo de tempo de promoção, na Tabela 3.19. Assim como para as demais aplicações, o erro padrão foi estimado por meio do método delta. O sentido dos efeitos estimados de incidência apresentados em Tabela 3.20 coincidem com aqueles obtidos pelo modelo de Xiang *et al.* (2011), ou seja, a relação de áreas com maior e menor efeito de incidência permanece na mesma ordem. Embora o modelo de tempo de promoção não utilize conjuntos separados de estimativas associados à incidência e ao risco, os sinais das estimativas obtidas mostraram-se iguais aos efeitos de incidência obtidos por meio dos outros algoritmos, mantendo também a mesma relação comparativa entre as áreas.

Uma vez que um mesmo cervo pode ser observado repetidas vezes com o decorrer dos anos, pode-se considerar um efeito do animal em particular para o conjunto de dados em questão. Assim,

utilizam-se efeitos aleatórios para modelar a estrutura de correlação de medidas de um mesmo cervo, também controlando a variabilidade dos dados. Tais questões motivam o uso dos modelos mistos apresentados e discutidos neste trabalho. O modelo de Xiang *et al.* (2011) para dados agrupados proporciona as estimativas de incidência e relacionadas ao risco encontradas na Tabela 3.23 e Tabela 3.24, respectivamente, com variâncias estimadas dos efeitos aleatórios em conjunto com seus erros padrão exibidas na Tabela 3.25.

Tabela 3.23: *Parâmetros relacionados à fração de cura estimados utilizando o modelo de mistura simples para dados agrupados*

Parâmetro	Intercepto	b_{wsi}	b_{AreaI}	b_{AreaS}	b_{AreaW}
Estimativa	-2,089 (0,5668)	0,024 (0,0034)	0,371 (0,5771)	0,814 (0,5655)	1,094 (0,5656)

Tabela 3.24: *Parâmetros relacionados ao risco estimados utilizando o modelo de mistura simples para dados agrupados*

Parâmetro	β_{wsi}	β_{AreaI}	β_{AreaS}	β_{AreaW}
Estimativa	0,004 (0,0018)	0,976 (0,3297)	0,931 (0,2811)	0,667 (0,2754)

Tabela 3.25: *Variância estimada dos efeitos aleatórios associados ao modelo de mistura simples*

Parâmetro	θ_1	θ_2
Estimativa	1,180 (0,2454)	0,920 (0,3874)

Nota-se que, embora existam evidências significativas a respeito da existência do efeito de cervo, as estimativas pontuais de efeito de incidência obtidas para este caso encontram-se razoavelmente próximas às estimativas encontradas quando não se utiliza a estrutura de grupo. Destaca-se aqui o fato da amostra ser pequena, explicando os valores altos para os erros padrão obtidos (ver Tabela 3.15 e Tabela 3.24).

Ao fixar-se, como feito anteriormente, área de estudo "W" e $wsi = 86,06$, foi obtida a seguinte fração de cervos não suscetíveis à migração esperada:

$$1 - \pi_{ij} = 1 - \pi(\boldsymbol{x}_{ij}) = 1 - \frac{\exp(-2,089 + 0,024 \times 86,06 + 1,094)}{1 + \exp(-2,089 + 0,024 \times 86,06 + 1,094)} = 0,255.$$

A fração de curados estimada por meio deste exemplo pouco difere daquela obtida quando desconsidera-se o efeito de *cluster*, avaliada como 27,3% de curados. As probabilidades de cura estimadas por região de estudo, considerando a precisão computacional obtida nos efeitos estimados, são apresentadas na Tabela 3.26. As frações de cura estimadas mostram-se próximas das mesmas obtidas quanto não utiliza-se a estrutura de grupos.

Tabela 3.26: *Frações de cura estimadas pelo modelo de mistura simples considerando-se grupos*

Área	Fração de Cura
D	0,515 (0,1244)
I	0,423 (0,0711)
S	0,320 (0,0596)
W	0,262 (0,0531)

Nas Tabelas 3.27 e 3.28 são apresentadas estimativas provenientes dos efeitos associados à incidência e ao risco, respectivamente, utilizando o modelo proposto em Lam e Wong (2014), extensão natural de Lam *et al.* (2013) para a estrutura de grupos. Similarmente à aplicação a dados de doença de descompressão exibida em Lam e Wong (2014), foram utilizadas $M = 400$ imputações por iteração. Após as 100 primeiras iterações e aquecimento com 30 iterações, observando certa estacionariedade para as estimativas, utilizou-se para resultados mais conservadores $M = 1000$ para

outras 10 iterações, proporcionando maior estabilidade no processo e convergência de $0,01$ para as estimativas associadas à incidência. Como para o ajuste sem efeito de grupo, um aumento sensível de M não proporcionou convergência com grande rigor para as estimativas. Entretanto, o estudo por meio de simulação sugere que a estimativa após a centésima iteração do processo pode se mostrar razoável em termos de viés, como reforçado em Lam *et al.* (2013) para os efeitos estimados com o uso de simulação.

A estimativa do logaritmo do parâmetro ω associado ao efeito aleatório é apresentada com seu erro padrão na Tabela 3.29. Vale ressaltar que, por construção, $\omega \to \infty$ implica em Ξ degenerada em 1 e, consequentemente, independência entre indivíduos de um mesmo grupo. Logo, o baixo valor da estimativa obtida sugere alta associação entre os membros do mesmo grupo.

Tabela 3.27: *Parâmetros relacionados à fração de cura estimados utilizando modelo de fragilidade para dados agrupados*

Parâmetro	Intercepto	θ_{wsi}	θ_{AreaI}	θ_{AreaS}	θ_{AreaW}
Estimativa	-0,855 (0,4207)	0,017 (0,0039)	0,194 (0,4595)	0,784 (0,4764)	0,939 (0,4876)

Tabela 3.28: *Parâmetros relacionados ao risco estimados utilizando modelo de fragilidade para dados agrupados*

Parâmetro	β_{wsi}	β_{AreaI}	β_{AreaS}	β_{AreaW}
Estimativa	-0,007 (0,0043)	0,932 (0,7187)	0,725 (0,7419)	0,178 (0,6652)

Tabela 3.29: *Estimativa e erro padrão de* $\log(\omega)$ *utilizando o modelo de fragilidade*

Parâmetro	$\log(\omega)$
Estimativa	0,242 (0,4292)

Da expressão (2.57), pode-se obter a sobrevivência marginal de t_i fixando-se 0 para os demais tempos associados ao mesmo grupo. Deste modo, tomando-se $t \to \infty$, a fração de cura associada aos cervos da área "W" com índice de severidade invernal médio (86,06) é dada por

$$\begin{aligned}\lim_{t\to\infty} S(t|\text{Area} = \text{W}, wsi = 86,06) &= \left[\frac{\omega}{\omega + \exp(\boldsymbol{\theta}'\boldsymbol{x}_{ij}^{(0)})/2}\right]^{\omega} \\ &= \left[\frac{0,242}{0,242 + (-e^{(-0,855+0,017\times 86,06+0,939)}/2)}\right]^{0,242} \\ &= 0,264.\end{aligned}$$

Analogamente, obtém-se as frações de cura estimadas apresentadas na Tabela 3.30 utilizando-se a precisão original das estimativas obtidas. Conforme pode ser observado, o modelo de fragilidade para dados agrupados apresentou consideráveis diferenças para as estimativas pontuais em relação ao modelo de mistura padrão para dados agrupados. Entretanto, erros padrão altos foram observados associados às estimativas provenientes de ambos os modelos.

Tabela 3.30: *Frações de cura estimadas pelo modelo de fragilidade considerando-se grupos*

Área	Fração de Cura
D	0,511 (0,1026)
I	0,459 (0,0822)
S	0,308 (0,0952)
W	0,272 (0,0901)

Em resumo, as estimativas pontuais das probabilidades de cura utilizando cada um dos modelos propostos neste trabalho são apresentadas na Tabela 3.31, com representação gráfica dada pela Figura 3.9. Conforme evidenciado, o uso da estrutura de agrupamento para o modelo de mistura padrão parece não afetar as estimativas das proporções de não migrantes, entretanto, grandes diferenças podem ser notadas para o modelo de fragilidade utilizando grupo em relação a quando não se considera tal estrutura. Ressalta-se que a amostra é pequena e, como os resultados das propriedades dos estimadores são assintóticos, pode ser que a aproximação para a normalidade não esteja boa. É importante que estudos sejam feitos para avaliar o desempenho do estimador para amostras pequenas.

Tabela 3.31: *Frações de cura estimadas para dados de migração*

	Mistura	Mistura com Grupos	Promoção	Fragilidade	Fragilidade com Grupos
Área D	0,519	0,515	0,537	0,504	0,511
Área I	0,416	0,423	0,382	0,420	0,459
Área S	0,318	0,320	0,368	0,404	0,308
Área W	0,273	0,262	0,331	0,339	0,272

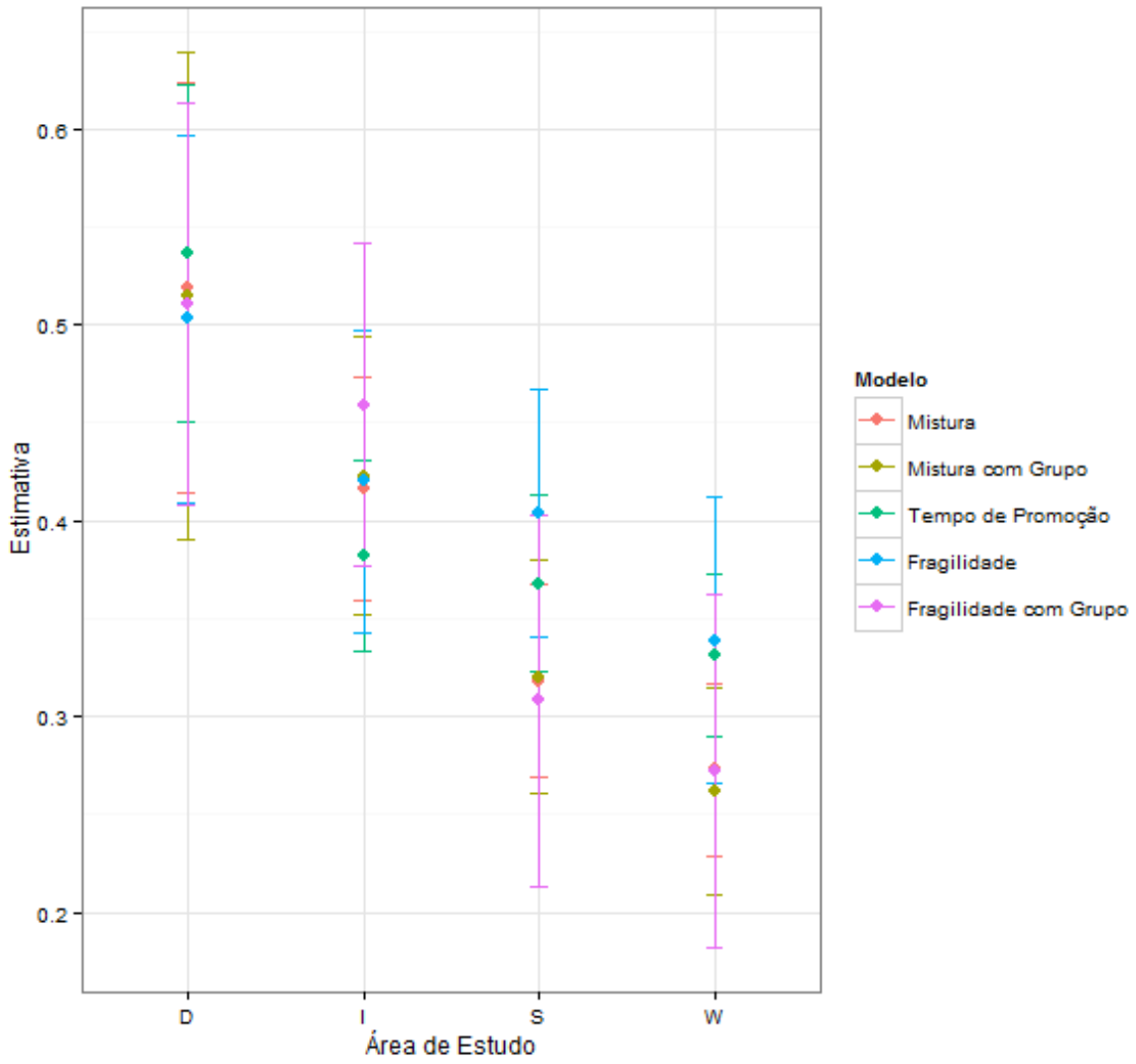

Figura 3.9: *Estimativas pontuais de fração de cura para dados de migração*

Devido às notáveis diferenças entre as estimativas, especialmente para os dados de doações, e a

complexidade quanto à escolha ideal para o mecanismo de fração de cura nos problemas apresentados, um estudo com respeito ao viés proporcionado por cada estimador em diferentes situações pode ser útil, motivando um critério de decisão adicional. O capítulo seguinte apresenta ao leitor estudos realizados por meio de simulações utilizando processos computacionalmente intensivos para melhor investigar propriedades de interesse dos estimadores aqui apresentados.

Capítulo 4

Simulações

Com o intuito de estudar a robustez e as possíveis propriedades presentes nos estimadores discutidos, potencialmente justificando a escolha de um modelo para determinadas situações práticas, foi realizado um estudo exaustivo das modelagens propostas por meio de simulações. Para isto, foram estudadas métricas de desempenho provenientes dos estimadores aplicados a conjuntos gerados a partir dos diferentes mecanismos de fração de cura anteriormente apresentados (mistura padrão, tempo de promoção e modelo de fragilidade). Diferentes cenários são estabelecidos a partir do percentual de curados, percentual de censurados, tamanho de amostra e, conforme mencionado, diferentes especificações para os modelos de fração de cura.

Foi considerada a estimação do efeito de dois tratamentos (representados por uma única variável binária) e uma covariável contínua de média zero, aqui adotada como normal padrão, cujo efeito sobre a proporção de curados é pré-estabelecido como nulo (como em Lam *et al.*, 2013, por exemplo). Tratando-se da fração de cura em si, duas possibilidades foram estudadas em conjunto com as possíveis combinações dos demais fatores: tratamentos T0 e T1 com 40% e 10% de curados, respectivamente; tratamentos T0 e T1 com 30% e 20% de proporção de curados, respectivamente.

As frações de cura são facilmente determinadas de forma analítica a partir de expressões anteriormente apresentadas, com as estimativas obtidas variando-se a covariável binária (tratamento T0 e T1) e mantendo fixada a covariável contínua como nula, permitindo melhor comparação entre as probabilidades de cura se comparado com o caso em que se fixa o valor médio amostral da covariável.

Para a proporção de censuras, utilizam-se valores de 35% a 40% e 60% a 65%, obtidas a partir de uma variável aleatória mista definida por $C = \min(c_1, c_2 \times A)$ tal que c_1 e c_2 são constantes escolhidas de modo que obtenha-se valores de censura dentro do intervalo desejado, com $A \sim \text{Exp}(1)$. Tais faixas de censura consideram o total de censurados da amostra toda, sem desconsiderar indivíduos curados. Os tamanhos de amostra n delineados para a simulação são tomados por 200, 400 e 800 unidades amostrais. As diferentes combinações de modelos, frações de cura, proporção de censura e tamanho de amostra definem 36 cenários diferentes. Para todo e qualquer cenário, foram gerados $B = 1000$ conjuntos de dados, ou seja, cada cenário contém 1000 amostras simuladas das quais são obtidas estimativas para cada uma delas.

Os diferentes estimadores apresentam diferentes taxas de convergência e critérios de parada, dificultando a uniformização do processo de simulação. Por conta disso, os estudos apresentados são flexíveis quanto a tais critérios, visando a escala do experimento e tempos computacionais viáveis. O estimador do modelo de mistura tem como critério de parada a diferença máxima inferior a $0,0001$ para as esperanças das variáveis latentes indicadoras de suscetibilidade. Para o estimador do modelo de tempo de promoção, adota-se a diferença máxima de $0,001$ para as estimativas dos efeitos e $0,005$ para os saltos da função distribuição. Em virtude do alto tempo de processamento para a convergência da esperança da log-verossimilhança, esta é desconsiderada. Para o estimador proveniente do modelo de fragilidade, utilizando $M = 50$ imputações por iteração, considera-se a diferença máxima de $0,001$ para as estimativas dos efeitos. Para todos estes, a centésima iteração é também adotada como critério de parada, como proposto de modo conservador para o modelo de

fragilidade nas simulações realizadas em Lam *et al.* (2013). A adoção da centésima iteração como critério de parada para os estimadores semiparamétricos é justificada por, em geral, estimativas associadas ao modelo de mistura e tempo de promoção convergirem com um número inferior de iterações, com o modelo de fragilidade apresentando bom desempenho para diversos cenários quando considera-se tal critério de parada, conforme apresentado posteriormente.

Adicionalmente, devido à melhor performance computacional e por utilizar seus ajustes essencialmente como controle, para o estimador não paramétrico de Turnbull implementado na função *icfit* do pacote "interval", foram mantidas as configurações padrão de número máximo de iterações igual a 10000 e tolerância de $0,000001$ entre diferenças.

A seguir detalha-se o procedimento para a geração de dados a partir de cada mecanismo. Quantidades como média estimada, em conjunto com viés, erro quadrático médio, erro padrão (calculado pelo método delta), probabilidade de cobertura e número de conjuntos considerados foram obtidas para cada cenário utilizando-se os estimadores propostos a fim de avaliar seus desempenhos quando assume-se uma especificação errônea e comparadas com as estimativas do modelo que gera os dados do cenário. Estimativas não paramétricas utilizando o algoritmo de Turnbull também são exibidas em cada cenário para fins comparativos. Para este algoritmo, em especial, considerou-se como estimativa da proporção de curados o último "salto" estimado iterativamente para cada estratificação por tratamento: por meio de simulações, as propriedades de tal estimativa sob o estrutura de censura intervalar são investigadas neste capítulo.

Em particular, o modelo de tempo de promoção não permite a estimação da fração de cura para conjuntos de dados que não apresentam o limiar de cura definido por Zeng *et al.* (2006). Por conta disso, o estimador de tempo de promoção é aplicado somente a conjuntos gerados contendo esta característica. Para o algoritmo de Turnbull, a sobrevivência estimada tende a 0 quando os dados não apresentam este limiar. Entretanto, utilizando-se o último "salto" da curva estimada, pode-se obter estimativas razoáveis da fração de cura, conforme exemplificado para o conjunto de dados de câncer de mama e reforçado no estudo que se segue.

As probabilidades de cobertura, obtidas para cada estimador com exceção do estimador de Turnbull, tratam-se da proporção de intervalos de confiança (95% de confiança) contendo o valor real da proporção de curados, construídos para cada fração de cura estimada utilizando o erro padrão obtido via método delta. Caso as propriedades de um estimador de máxima verossimilhança sejam satisfeitas, as probabilidades de cobertura estimadas devem aproximar-se de 95% com o aumento do tamanho da amostra.

Além das estimativas apresentadas nesta seção, para cada cenário da simulação, são estudadas as estimativas dos parâmetros obtidas quando utiliza-se o modelo que gera os dados do cenário em questão. Estes resultados são apresentados no Apêndice B.2.

4.1 Geração de dados baseada no Modelo de Mistura Padrão

A metodologia apresentada a seguir baseia-se nos estudos com uso de simulações realizados em Xiang *et al.* (2011) sem a presença de efeito aleatório para grupo, com pequenas alterações adotadas com o intuito de padronizar os diferentes métodos de geração quanto aos instantes de observação e censura. Geram-se, através do algoritmo a seguir, os dados com censura intervalar (L_i, R_i, δ_i) com a presença de uma fração de curados:

1. Gera-se uma variável latente binária Y_i utilizando-se o modelo logístico $P(Y_i = 1) = \frac{1}{1+e^{-(b_0+b_1x_{i1}+b_2x_{i2})}}$ para o i-ésimo indivíduo, com $x_{i1} \sim Bernoulli(0,5)$, $x_{i2} \sim N(0;1)$, $b_2 = 0$, e com b_0 e b_1 escolhidos de modo que as frações de cura para cada tratamento coincidam com seus valores desejados.

2. Gera-se o tempo de censura $C = \min(c_1, c_2 \times A)$, com $A \sim \text{Exp}(1)$. As constantes c_1 e c_2 são determinadas de acordo com a proporção de censuras desejada.

3. Caso $Y_i = 0$, toma-se $T_i = C_i$ e o indicador de censura $\delta_i = 0$.

4. Caso $Y_i = 1$, o tempo de evento é gerado a partir da distribuição com risco condicional $\lambda_i(t) = \exp(\boldsymbol{\beta}\boldsymbol{x}_i)$, com $\beta_1 = 0$ e $\beta_2 = 0,5$. Faz-se $\delta_i = 1$ se $T_i \leq C_i$ e 0 caso contrário.

5. Se $\delta_i = 0$, toma-se $L_i = C_i < R_i = \infty$.

6. Se $\delta_i = 1$, geram-se m instantes de observação a partir da soma de tempos com distribuição uniforme $Q_j \sim U(0,1; 0,5)$, de modo que $\sum_{j=0}^{m-1} Q_j \leq T_i < \sum_{j=0}^{m} Q_j$, em que $Q_0 = 0$ é o primeiro instante de observação. Satisfeita a inequação, faz-se então $L_i = \sum_{j=0}^{m-1} Q_j$ e $R_i = \sum_{j=0}^{m} Q_j$.

Com base no processo definido acima, foram simulados dados provenientes de um mecanismo de mistura padrão. A partir dos estimadores apresentados em capítulos anteriores, em conjunto com o estimador de mistura padrão de Xiang *et al.* (2011), calcula-se para cada conjunto de dados, primeiramente, as estimativas dos efeitos associados à incidência. A partir dos efeitos estimados, aplica-se a expressão para obtenção da proporção de curados proveniente de cada estimador. Tendo, para cada um dos conjuntos de dados gerados, estimativas das frações de cura para ambos os tratamentos com seus respectivos erros padrão, são então obtidos e exibidos resultados como valor médio estimado, viés médio, desvio padrão das estimativas obtidas para cada amostra, erro padrão médio, probabilidade de cobertura (para valor teórico de 95%) e, por fim, o número estimativas obtidas para cada cenário, podendo ser inferior a $B = 1000$ por conta de instabilidades numéricas ou limitações quanto ao limiar de cura.

Na Tabela 4.1, são apresentados os resultados das simulações provenientes do cenário com taxa de censura média (35 a 40%), com probabilidades de cura relacionadas aos tratamento T0 e T1 dadas por 40% e 10%, respectivamente. O viés para este cenário é substancialmente inferior utilizando-se um estimador apropriado para mistura padrão. Entretanto, as probabilidades de cobertura dos intervalos apresentam-se inferiores a 95%. Em particular, devido ao erro padrão de cada salto estimado pelo método de Turnbull ser obtido através de *bootstrap*, consequentemente tornando esse resultado inviável na simulação, o erro padrão médio associado às estimativas é desconsiderado deste estudo para este estimador. Devido a um dos conjuntos não apresentar o denominado limiar de cura, o modelo de tempo de promoção apresentou uma estimativa a menos em relação aos demais.

Tabela 4.1: *Resultados para dados gerados por modelo de mistura padrão com* $n = 200$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,401	0,001	0,004	0,062	0,049	0,883	1000
Mistura	T1	0,099	-0,001	0,002	0,040	0,029	0,836	1000
Promoção	T0	0,369	-0,031	0,005	0,063	0,047	0,799	999
Promoção	T1	0,122	0,022	0,002	0,037	0,027	0,817	999
Fragilidade	T0	0,387	-0,013	0,005	0,066	0,053	0,865	1000
Fragilidade	T1	0,082	-0,018	0,002	0,044	0,039	0,808	1000
Turnbull	T0	0,401	0,001	0,004	0,062			1000
Turnbull	T1	0,103	0,003	0,002	0,041			1000

Para o cenário de tamanho de amostra $n = 200$ com taxa de censura alta (60% a 65%) e com frações de cura de 40% e 10% para T0 e T1, respectivamente, foram obtidos os resultados apresentados em Tabela 4.2. Para tais taxas de censura, o estimador de Lam apresentou-se superior em termos de vício e erro quadrático médio para T1, mesmo não sendo o modelo associado à geração dos dados.

Ainda avaliando amostras de tamanho $n = 200$ geradas a partir da simulação de modelos de mistura padrão, para tratamentos com fração de cura 30% e 20% e taxa de censura média entre 35% e 40%, foi obtida a Tabela 4.3. As probabilidades de cobertura encontram-se mais altas para este cenário, porém, ainda bastante inferiores ao valor teórico de 95%.

Tabela 4.2: *Resultados para dados gerados por modelo de mistura padrão com* $n = 200$*, fração de cura entre* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,410	0,010	0,015	0,121	0,049	0,590	1000
Mistura	T1	0,141	0,041	0,010	0,094	0,031	0,448	1000
Promoção	T0	0,381	-0,019	0,011	0,101	0,046	0,622	949
Promoção	T1	0,174	0,074	0,013	0,085	0,032	0,410	949
Fragilidade	T0	0,386	-0,014	0,016	0,127	0,067	0,688	1000
Fragilidade	T1	0,120	0,020	0,009	0,094	0,074	0,766	1000
Turnbull	T0	0,422	0,022	0,012	0,107			1000
Turnbull	T1	0,191	0,091	0,014	0,077			1000

Tabela 4.3: *Resultados para dados gerados por modelo de mistura padrão com* $n = 200$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,297	-0,003	0,003	0,058	0,046	0,863	1000
Mistura	T1	0,196	-0,004	0,002	0,048	0,040	0,879	1000
Promoção	T0	0,285	-0,015	0,003	0,057	0,042	0,815	1000
Promoção	T1	0,203	0,003	0,002	0,047	0,036	0,865	1000
Fragilidade	T0	0,282	-0,018	0,004	0,063	0,047	0,823	1000
Fragilidade	T1	0,182	-0,018	0,003	0,054	0,055	0,895	1000
Turnbull	T0	0,300	-0,000	0,004	0,060			1000
Turnbull	T1	0,198	-0,002	0,003	0,051			1000

Na Tabela 4.4 são exibidas as métricas referentes ao cenário de tamanho de amostra $n = 200$, fração de cura de 30% e 20% e taxa de censura entre 60% e 65%. A probabilidade de cobertura associada à fração de cura para ambos os tratamentos é superior utilizando-se o algoritmo de Lam, entretanto, o mesmo estimador apresentou os maiores erros padrão para este cenário. Para o modelo de tempo de promoção, 64 estimativas não foram obtidas por conta da inexistência do limiar de cura para as respectivas amostras.

Tabela 4.4: *Resultados para dados gerados por modelo de mistura padrão com* $n = 200$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,319	0,019	0,015	0,120	0,046	0,538	1000
Mistura	T1	0,226	0,026	0,012	0,108	0,040	0,540	1000
Promoção	T0	0,307	0,007	0,010	0,102	0,043	0,579	936
Promoção	T1	0,241	0,041	0,011	0,094	0,038	0,524	936
Fragilidade	T0	0,277	-0,023	0,016	0,126	0,059	0,625	1000
Fragilidade	T1	0,191	-0,009	0,013	0,115	0,095	0,801	1000
Turnbull	T0	0,338	0,038	0,013	0,106			1000
Turnbull	T1	0,257	0,057	0,012	0,093			1000

Aumentando-se o tamanho das amostras geradas para $n = 400$, o cenário com frações de cura $0,40$ e $0,10$ com níveis de censura intermediários levou aos resultados da Tabela 4.5. Os erros quadráticos médios mostram-se próximos para este cenário. Ao aumentar os níveis de censura, obtém-se maior discrepância entre os desvios padrões empíricos e os erros padrões médios, conforme a Tabela 4.6. O estimador de fragilidade apresenta maior viés médio associado a T1 ao estimar-se

a fração de cura no cenário de proporções $0,3$ e $0,2$ com nível de censura médio, como mostra a Tabela 4.7, entretanto, seu erro quadrático médio mostrou-se próximo dos pertencentes aos demais estimadores. Com maiores taxas de censura sobre o cenário anterior, os estimadores mostraram-se todos com maior viés e probabilidades de cobertura inferior, com o modelo de tempo de promoção sobressaindo-se com o erro quadrático médio inferior (ver Tabela 4.8). Algumas estimativas não foram obtidas para o estimador de mistura devido a erros relacionados à matriz modelo não ser inversível para certos conjuntos gerados.

Tabela 4.5: *Resultados para dados gerados por modelo de mistura padrão com* $n = 400$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,398	-0,002	0,002	0,043	0,035	0,887	1000
Mistura	T1	0,101	0,001	0,001	0,029	0,021	0,842	1000
Promoção	T0	0,367	-0,033	0,003	0,044	0,033	0,738	1000
Promoção	T1	0,124	0,024	0,001	0,027	0,019	0,724	1000
Fragilidade	T0	0,388	-0,012	0,002	0,045	0,037	0,879	1000
Fragilidade	T1	0,090	-0,010	0,001	0,032	0,029	0,865	1000
Turnbull	T0	0,398	-0,002	0,002	0,043			1000
Turnbull	T1	0,102	0,002	0,001	0,031			1000

Tabela 4.6: *Resultados para dados gerados por modelo de mistura padrão com* $n = 400$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,419	0,019	0,008	0,087	0,035	0,570	1000
Mistura	T1	0,136	0,036	0,006	0,072	0,023	0,424	1000
Promoção	T0	0,385	-0,015	0,007	0,080	0,033	0,559	962
Promoção	T1	0,173	0,073	0,009	0,059	0,023	0,277	962
Fragilidade	T0	0,401	0,001	0,008	0,090	0,049	0,710	1000
Fragilidade	T1	0,116	0,016	0,006	0,073	0,060	0,807	1000
Turnbull	T0	0,419	0,019	0,009	0,090			1000
Turnbull	T1	0,163	0,063	0,008	0,064			1000

Tabela 4.7: *Resultados para dados gerados por modelo de mistura padrão com* $n = 400$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,297	-0,003	0,002	0,040	0,032	0,868	984
Mistura	T1	0,199	-0,001	0,001	0,036	0,028	0,870	984
Promoção	T0	0,285	-0,015	0,002	0,040	0,030	0,809	998
Promoção	T1	0,205	0,005	0,001	0,034	0,026	0,856	998
Fragilidade	T0	0,287	-0,013	0,002	0,042	0,034	0,839	1000
Fragilidade	T1	0,189	-0,011	0,002	0,038	0,040	0,928	1000
Turnbull	T0	0,297	-0,003	0,002	0,042			1000
Turnbull	T1	0,200	0,000	0,001	0,038			1000

Aumentando-se o tamanho da amostra para $n = 800$ e para censura média e frações de cura de 40% e 10%, com geração baseada no mecanismo de mistura, obtém-se os resultados apresentados na Tabela 4.9. Dentre os modelos semiparamétricos, o estimador de fragilidade apresentou para

Tabela 4.8: *Resultados para dados gerados por modelo de mistura padrão com* $n = 400$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,311	0,011	0,008	0,086	0,033	0,538	916
Mistura	T1	0,217	0,017	0,007	0,083	0,028	0,509	916
Promoção	T0	0,308	0,008	0,006	0,076	0,031	0,576	955
Promoção	T1	0,237	0,037	0,006	0,070	0,027	0,485	955
Fragilidade	T0	0,287	-0,013	0,009	0,093	0,043	0,629	1000
Fragilidade	T1	0,195	-0,005	0,009	0,093	0,075	0,825	1000
Turnbull	T0	0,319	0,019	0,008	0,087			1000
Turnbull	T1	0,237	0,037	0,007	0,077			1000

este cenário os melhores resultados quanto a viés e erro quadrático médio, com o estimador não paramétrico de Turnbull apresentando-se ainda mais preciso quanto a estas métricas. Para todos os cenários com tamanho de amostra $n = 800$, o estimador de Xiang *et al.* (2011) apresentou dificuldades técnicas no processo de estimação (tempos inviáveis para cada iteração do algoritmo), impedindo a avaliação de seu desempenho para cenários com este tamanho amostral e, portanto, com resultados associados não disponibilizados.

Tabela 4.9: *Resultados para dados gerados por modelo de mistura padrão com* $n = 800$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,367	-0,033	0,002	0,030	0,023	0,650	1000
Promoção	T1	0,123	0,023	0,001	0,019	0,014	0,599	1000
Fragilidade	T0	0,392	-0,008	0,001	0,030	0,026	0,915	1000
Fragilidade	T1	0,091	-0,009	0,001	0,022	0,022	0,897	1000
Turnbull	T0	0,399	-0,001	0,001	0,029			1000
Turnbull	T1	0,099	-0,001	0,000	0,021			1000

Ao aumentar o nível de censura, observa-se menor valor para o erro quadrático médio associado à fração de cura de T0 utilizando o modelo de tempo de promoção, entretanto, o uso do mesmo modelo apresentou probabilidade de cobertura bastante baixa associada à estimativa da fração de cura em T1 (vide Tabela 4.10), sugerindo que a aproximação para a normalidade adotada ao usar o método delta para estimar os erros padrão pode não ser boa. Também neste cenário, o modelo de fragilidade apresenta melhor desempenho associado às estimativas relacionadas ao tratamento T1.

Tabela 4.10: *Resultados para dados gerados por modelo de mistura padrão com* $n = 800$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,382	-0,018	0,003	0,056	0,023	0,600	936
Promoção	T1	0,171	0,071	0,007	0,043	0,017	0,131	936
Fragilidade	T0	0,403	0,003	0,004	0,063	0,034	0,718	1000
Fragilidade	T1	0,113	0,013	0,003	0,052	0,047	0,865	1000
Turnbull	T0	0,413	0,013	0,005	0,067			1000
Turnbull	T1	0,141	0,041	0,004	0,049			1000

Maiores probabilidades de cobertura podem ser observadas no cenário com frações de cura 30% e 20% para T0 e T1, respectivamente, com nível de censura médio. O estimador não paramétrico

apresentou menor viés médio para este cenário (vide Tabela 4.11).

Tabela 4.11: *Resultados para dados gerados por modelo de mistura padrão com* $n = 800$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,284	-0,016	0,001	0,027	0,021	0,808	1000
Promoção	T1	0,204	0,004	0,001	0,024	0,018	0,862	1000
Fragilidade	T0	0,291	-0,009	0,001	0,029	0,024	0,870	1000
Fragilidade	T1	0,191	-0,009	0,001	0,027	0,028	0,932	1000
Turnbull	T0	0,298	-0,002	0,001	0,029			1000
Turnbull	T1	0,197	-0,003	0,001	0,027			1000

Os resultados obtidos para dados gerados pelo modelo de mistura com $n = 800$, proporção de curados de $0,30$ e $0,20$, e alto nível de censura, são apresentados na Tabela 4.12. O erro quadrático médio das estimativas mostrou-se próximo. O estimador de fragilidade apresentou menor viés para a proporção de curados de T1.

Tabela 4.12: *Resultados para dados gerados por modelo de mistura padrão com* $n = 800$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,305	0,005	0,004	0,059	0,022	0,557	968
Promoção	T1	0,236	0,036	0,004	0,054	0,019	0,389	968
Fragilidade	T0	0,294	-0,006	0,005	0,067	0,031	0,639	1000
Fragilidade	T1	0,199	-0,001	0,005	0,068	0,058	0,859	1000
Turnbull	T0	0,309	0,009	0,005	0,071			1000
Turnbull	T1	0,224	0,024	0,005	0,066			1000

Para facilitar a visualização dos resultados das simulações com dados gerados pelo modelo de mistura e frações de cura de 40% e 10%, apresenta-se o viés médio e o erro quadrático médio nas Figuras B.1 e B.2, com variação de tamanho de amostra, nível de censura e estimador utilizado. Analogamente, obtém-se nas Figuras B.3 e B.4 os gráficos respectivos considerando-se o cenário com frações de cura de 30% e 20% associadas a T0 e T1, respectivamente. Por meio das figuras apresentadas, é notável a melhora das estimativas com o aumento do tamanho da amostra para todos os cenários com dados simulados pelo mecanismo de mistura, conforme evidenciado pela diminuição do erro quadrático médio. O estimador de mistura padrão, embora tenha a mesma especificação utilizada na geração dos dados, não apresenta necessariamente o melhor desempenho, como pode ser visto nos cenários de alta taxa de censura.

O Apêndice B.2 apresenta os efeitos estimados utilizando o modelo de mistura do qual derivam-se as estimativas das frações de cura apresentadas nesta seção quando utiliza-se o modelo de Xiang *et al.* (2011). Conforme esperado, maiores taxas de censura implicam em maior viés associado às estimativas e baixo desempenho na estimação intervalar, como exibem as probabilidades de cobertura das tabelas. O leitor pode consultar o mesmo apêndice a fim de avaliar as estimativas dos efeitos obtidos pelo modelo de tempo de promoção e fragilidade quando especificados corretamente, com o intuito de complementar as análises apresentadas nas próximas seções.

4.2 Geração de dados baseada no Modelo de Tempo de Promoção

Para a geração de dados a partir de um modelo de tempo de promoção, obtém-se a média de uma variável aleatória de distribuição Poisson $N_i \sim P(\lambda_i)$ por $\lambda_i = \boldsymbol{\beta}' x_i$, com esta representando a quantidade de células tumorosas remanescentes após um determinado tratamento. Obtidas as

contagens, geram-se então N_i variáveis exponenciais de média 1. O tempo de evento, assim como na motivação biológica mencionada na introdução deste trabalho, é tomado como o mínimo dos tempos exponenciais de cada célula tumorosa. O procedimento de geração é detalhado pelo algoritmo a seguir:

1. Geram-se, para cada unidade amostral, variáveis aleatórias N_i com distribuição Poisson de média $\exp(\alpha + \beta_0 x_{i1} + \beta_1 x_{i2})$ para o i-ésimo indivíduo, com $x_{i1} \sim Bernoulli(0,5)$, $x_{i2} \sim N(0;1)$, $\beta_1 = 0$ e com α e β_0 escolhidos de modo que as frações de cura para cada tratamento coincidam com seus valores desejados.

2. Geram-se, para cada unidade amostral, N_i tempos com distribuição exponencial de média 1. Adota-se o mínimo destes tempos como sendo o tempo do evento de interesse, T_i. Caso $N_i = 0$, toma-se $T_i = \infty$, constituindo um caso de cura.

3. Geram-se tempos de censura $C_i \sim \min(c_1, c_2 \times A)$, com $A \sim \text{Exp}(1)$. As constantes c_1 e c_2 são determinadas de acordo com a proporção de censuras desejada. Define-se então a variável indicadora de falhas $\delta_i = I(T_i \leq C_i)$.

4. Se $\delta_i = 0$, toma-se $L_i = C_i < R_i = \infty$.

5. Se $\delta_i = 1$, geram-se m instantes de observação a partir da soma de tempos com distribuição uniforme $Q_j \sim U(0,1;0,5)$, de modo que $\sum_{j=0}^{m-1} Q_j \leq T_i < \sum_{j=0}^{m} Q_j$, em que $Q_0 = 0$ é o primeiro instante de observação. Satisfeita a inequação, faz-se então $L_i = \sum_{j=0}^{m-1} Q_j$ e $R_i = \sum_{j=0}^{m} Q_j$.

De forma similar às estimativas para os dados gerados pelo modelo de mistura, as estimativas das frações de cura provém dos efeitos estimados utilizando-se cada estimador apresentado e suas expressões de proporção de curados associadas.

Assim como observado para os dados gerados através do modelo de mistura padrão, os resultados referentes à probabilidade de cobertura obtidos utilizando-se os estimadores propostos para dados gerados a partir do modelo de tempo de promoção ainda mostram-se inferiores à proporção teórica de 95% estabelecida (Tabela 4.13). Conforme esperado, o viés e o erro quadrático médio associado às estimativas do modelo de tempo de promoção são inferiores aos obtidos pelos demais modelos.

Tabela 4.13: *Resultados para dados gerados por modelo de tempo de promoção com* $n = 200$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,388	-0,012	0,005	0,068	0,049	0,822	1000
Mistura	T1	0,099	-0,001	0,002	0,042	0,029	0,837	1000
Promoção	T0	0,397	-0,003	0,004	0,060	0,047	0,850	1000
Promoção	T1	0,100	-0,000	0,001	0,034	0,024	0,823	1000
Fragilidade	T0	0,385	-0,015	0,005	0,070	0,053	0,843	1000
Fragilidade	T1	0,088	-0,012	0,002	0,044	0,044	0,870	1000
Turnbull	T0	0,397	-0,003	0,004	0,065			1000
Turnbull	T1	0,102	0,002	0,002	0,039			1000

Para maiores taxas de censura, mesmo com dados gerados por meio do modelo de tempo de promoção, o estimador proveniente do mesmo modelo apresentou um número menor de estimativas por conta de alguns dos conjuntos de dados gerados não apresentarem o limiar de cura necessário. As estimativas provenientes do modelo dos dados do cenário apresentaram os menores erros quadráticos médios, com o estimador de fragilidade apresentando maior erro quadrático estimando a proporção de curados de T0. Tais resultados, estimados sob um cenário de fração de cura 40% e 10% para os tratamentos e altas taxas de censura, podem ser observados na Tabela 4.14.

Tabela 4.14: *Resultados para dados gerados por modelo de tempo de promoção com $n = 200$, frações de cura iguais a 10% e 40% e taxa de censura entre 60% e 65%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,414	0,014	0,019	0,138	0,048	0,505	999
Mistura	T1	0,129	0,029	0,009	0,088	0,030	0,462	999
Promoção	T0	0,424	0,024	0,012	0,107	0,047	0,576	917
Promoção	T1	0,122	0,022	0,005	0,069	0,026	0,527	917
Fragilidade	T0	0,395	-0,005	0,023	0,150	0,069	0,606	1000
Fragilidade	T1	0,123	0,023	0,009	0,090	0,083	0,831	1000
Turnbull	T0	0,429	0,029	0,016	0,122			1000
Turnbull	T1	0,165	0,065	0,009	0,071			1000

Para o cenário de dados gerados pelo modelo de tempo de promoção apresentando frações de cura de 30% e 20% para T0 e T1, respectivamente, e taxa de censura entre 35% e 40%, o modelo proposto em Lam *et al.* (2013) apresentou maior viés, porém, maior probabilidade de cobertura para as estimativas referentes a T1. O estimador não paramétrico de Turnbull mostra-se como uma boa opção em termos de viés e erro quadrático médio. Os desvios padrão empíricos encontram-se consideravelmente próximos neste cenário. Tais resultados podem ser observados na Tabela 4.15. Ao aumentar-se as taxas de censura, o modelo de tempo de promoção apresenta o menor erro quadrático médio, com o estimador de fragilidade apresentando o menor viés, embora tenha apresentado também o maior erro quadrático médio, conforme pode ser visto na Tabela 4.16.

Tabela 4.15: *Resultados para dados gerados por modelo de tempo de promoção com $n = 200$, frações de cura iguais a 20% e 30% e taxa de censura entre 35% e 40%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,296	-0,004	0,003	0,057	0,046	0,872	1000
Mistura	T1	0,194	-0,006	0,003	0,051	0,039	0,859	1000
Promoção	T0	0,298	-0,002	0,003	0,051	0,043	0,880	1000
Promoção	T1	0,198	-0,002	0,002	0,045	0,036	0,863	1000
Fragilidade	T0	0,289	-0,011	0,004	0,058	0,047	0,857	1000
Fragilidade	T1	0,185	-0,015	0,003	0,054	0,055	0,937	1000
Turnbull	T0	0,300	0,000	0,003	0,056			1000
Turnbull	T1	0,196	-0,004	0,002	0,049			1000

Tabela 4.16: *Resultados para dados gerados por modelo de tempo de promoção com $n = 200$, frações de cura iguais a 20% e 30% e taxa de censura entre 60% e 65%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,345	0,045	0,014	0,108	0,047	0,546	1000
Mistura	T1	0,246	0,046	0,013	0,103	0,042	0,479	1000
Promoção	T0	0,326	0,026	0,010	0,095	0,044	0,588	872
Promoção	T1	0,226	0,026	0,008	0,087	0,037	0,536	872
Fragilidade	T0	0,281	-0,019	0,019	0,137	0,059	0,581	1000
Fragilidade	T1	0,197	-0,003	0,014	0,120	0,104	0,797	1000
Turnbull	T0	0,336	0,036	0,012	0,106			1000
Turnbull	T1	0,249	0,049	0,011	0,094			1000

Aumentando-se o volume de dados gerados para $n = 400$, os estimadores não paramétrico e de tempo de promoção apresentam os menores vícios médios para as proporções de cura geradas pelo modelo de tempo de promoção. Embora altas, nenhuma probabilidade de cobertura atingiu 95% para este cenário (fração de cura de $0,40$ e $0,10$ com níveis médios de censura), conforme a Tabela 4.17.

Tabela 4.17: *Resultados para dados gerados por modelo de tempo de promoção com* $n = 400$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,390	-0,010	0,002	0,045	0,035	0,860	1000
Mistura	T1	0,099	-0,001	0,001	0,029	0,021	0,838	1000
Promoção	T0	0,399	-0,001	0,002	0,041	0,033	0,890	1000
Promoção	T1	0,099	-0,001	0,001	0,023	0,017	0,840	1000
Fragilidade	T0	0,392	-0,008	0,002	0,045	0,038	0,881	1000
Fragilidade	T1	0,093	-0,007	0,001	0,030	0,033	0,924	1000
Turnbull	T0	0,399	-0,001	0,002	0,044			1000
Turnbull	T1	0,099	-0,001	0,001	0,028			1000

Para dados gerados pelo modelo de tempo de promoção com maior nível de censura, o estimador com a mesma especificação apresentou, com diferenças substanciais em relação aos demais, o menor erro quadrático médio, conforme a Tabela 4.18 mostra.

Tabela 4.18: *Resultados para dados gerados por modelo de tempo de promoção com* $n = 400$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,399	-0,001	0,012	0,110	0,034	0,460	1000
Mistura	T1	0,123	0,023	0,005	0,070	0,021	0,414	1000
Promoção	T0	0,418	0,018	0,007	0,082	0,033	0,572	940
Promoção	T1	0,118	0,018	0,003	0,051	0,018	0,482	940
Fragilidade	T0	0,399	-0,001	0,012	0,111	0,050	0,619	1000
Fragilidade	T1	0,115	0,015	0,005	0,069	0,067	0,844	1000
Turnbull	T0	0,410	0,010	0,011	0,105			1000
Turnbull	T1	0,139	0,039	0,005	0,057			1000

Para o cenário de nível médio de censura cujas frações de cura para T0 e T1 são dadas por $0,30$ e $0,20$, respectivamente, o erro quadrático médio dos estimadores mostrou-se próximo, com o viés do estimador não paramétrico novamente igualando-se ao viés obtido pelo estimador de tempo de promoção (Tabela 4.19). Para maiores níveis de censura, o estimador de Turnbull continua sendo a melhor opção referente ao erro quadrático médio quando descarta-se o estimador de tempo de promoção, conforme a Tabela 4.20. Neste mesmo cenário, uma estimativa não foi obtida para o modelo de fragilidade devido a problemas relacionados à implementação por conta da precisão nos dígitos significativos.

Para um maior tamanho de amostra ($n = 800$), os estimadores propostos (com exceção do modelo de mistura, por conta do tempo de processamento por iteração inviável para a escala do estudo) apresentaram erros quadráticos médios iguais sob a precisão considerada quando avaliados sob o cenário com T0 e T1 apresentando 40% e 10% de indivíduos curados com nível de censura entre 35% e 40% (Tabela 4.21). Mantendo as mesmas proporções de curados e aumentando-se os níveis de censura, foram obtidos os resultados apresentados em Tabela 4.22.

Para os dados gerados pelo modelo de promoção tais que as frações de cura são dadas por $0,3$ e $0,2$, com níveis médios de censura, os estimadores semiparamétricos avaliados apresentaram altas

Tabela 4.19: *Resultados para dados gerados por modelo de tempo de promoção com $n = 400$, frações de cura iguais a 20% e 30% e taxa de censura entre 35% e 40%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,291	-0,009	0,002	0,043	0,032	0,846	1000
Mistura	T1	0,198	-0,002	0,001	0,036	0,028	0,868	1000
Promoção	T0	0,295	-0,005	0,001	0,038	0,030	0,878	1000
Promoção	T1	0,199	-0,001	0,001	0,032	0,025	0,877	1000
Fragilidade	T0	0,289	-0,011	0,002	0,043	0,033	0,852	1000
Fragilidade	T1	0,193	-0,007	0,001	0,038	0,039	0,939	1000
Turnbull	T0	0,295	-0,005	0,002	0,042			1000
Turnbull	T1	0,199	-0,001	0,001	0,036			1000

Tabela 4.20: *Resultados para dados gerados por modelo de tempo de promoção com $n = 400$, frações de cura iguais a 20% e 30% e taxa de censura entre 60% e 65%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,311	0,011	0,011	0,105	0,032	0,425	1000
Mistura	T1	0,214	0,014	0,008	0,091	0,028	0,413	1000
Promoção	T0	0,315	0,015	0,007	0,079	0,031	0,517	921
Promoção	T1	0,216	0,016	0,005	0,069	0,026	0,511	921
Fragilidade	T0	0,292	-0,008	0,011	0,107	0,043	0,559	999
Fragilidade	T1	0,194	-0,006	0,009	0,094	0,082	0,825	999
Turnbull	T0	0,320	0,020	0,010	0,096			1000
Turnbull	T1	0,223	0,023	0,006	0,077			1000

Tabela 4.21: *Resultados para dados gerados por modelo de tempo de promoção com $n = 800$, frações de cura iguais a 10% e 40% e taxa de censura entre 35% e 40%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,398	-0,002	0,001	0,028	0,024	0,904	1000
Promoção	T1	0,099	-0,001	0,000	0,017	0,012	0,815	1000
Fragilidade	T0	0,394	-0,006	0,001	0,031	0,027	0,909	1000
Fragilidade	T1	0,096	-0,004	0,000	0,021	0,024	0,946	1000
Turnbull	T0	0,397	-0,003	0,001	0,031			1000
Turnbull	T1	0,098	-0,002	0,000	0,021			1000

Tabela 4.22: *Resultados para dados gerados por modelo de tempo de promoção com $n = 800$, frações de cura iguais a 10% e 40% e taxa de censura entre 60% e 65%*

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,416	0,016	0,004	0,065	0,024	0,506	940
Promoção	T1	0,113	0,013	0,002	0,040	0,013	0,436	940
Fragilidade	T0	0,409	0,009	0,006	0,078	0,036	0,635	1000
Fragilidade	T1	0,111	0,011	0,003	0,054	0,053	0,887	1000
Turnbull	T0	0,399	-0,001	0,008	0,091			1000
Turnbull	T1	0,121	0,021	0,003	0,046			1000

probabilidades de cobertura associadas às frações de cura estimadas, com o estimador de fragilidade atingindo a probabilidade teórica pré-estabelecida para T1, conforme Tabela 4.23.

Tabela 4.23: *Resultados para dados gerados por modelo de tempo de promoção com* $n = 800$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,298	-0,002	0,001	0,025	0,022	0,903	1000
Promoção	T1	0,199	-0,001	0,000	0,022	0,018	0,888	1000
Fragilidade	T0	0,295	-0,005	0,001	0,028	0,024	0,900	1000
Fragilidade	T1	0,197	-0,003	0,001	0,025	0,028	0,954	1000
Turnbull	T0	0,298	-0,002	0,001	0,027			1000
Turnbull	T1	0,199	-0,001	0,001	0,025			1000

Aumentando-se o nível de censuras, o estimador de fragilidade passa a apresentar grande diferença nas probabilidades de cobertura em relação ao estimador de tempo de promoção, compreendendo maior número de acertos nos intervalos definidos. O mesmo estimador apresentou, entretanto, os maiores erros padrão estimados (Tabela 4.24).

Tabela 4.24: *Resultados para dados gerados por modelo de tempo de promoção com* $n = 800$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,313	0,013	0,003	0,058	0,022	0,530	946
Promoção	T1	0,214	0,014	0,003	0,051	0,018	0,508	946
Fragilidade	T0	0,300	0,000	0,005	0,073	0,032	0,610	1000
Fragilidade	T1	0,202	0,002	0,005	0,071	0,064	0,873	1000
Turnbull	T0	0,303	0,003	0,006	0,078			1000
Turnbull	T1	0,210	0,010	0,004	0,064			1000

Nas Figuras B.5 e B.6, encontram-se os gráficos de viés e erro quadrático médio associados aos cenários gerados pelo modelo de tempo de promoção com frações de cura de 0,40 e 0,10 referentes aos tratamentos T0 e T1, respectivamente. Observa-se ao aumentar o tamanho da amostra, assim como para os dados gerados pelo modelo de mistura, uma diminuição substancial dos erros quadráticos médios associados às estimativas de todos os estimadores aqui apresentados. A mesma propriedade pode ser observada nas Figuras B.7 e B.8, quando considera-se os tratamentos com fração de cura 0,30 e 0,20, respectivamente. O estimador com a mesma especificação dos dados do cenário apresentou, sob a precisão considerada, erro quadrático médio estritamente inferior em relação a todos os outros estimadores nos cenários de alta taxa de censura.

As estimativas dos efeitos especificados no modelo de tempo de promoção são exibidas nas Tabelas B.3, B.4 e B.5 do Apêndice B.2 e, conforme esperado, com erros quadráticos médios tornando-se menores ao aumentar-se o tamanho da amostra.

4.3 Geração de dados baseada no Modelo de Fragilidade

Através do modelo de Lam *et al.* (2013), geram-se conjuntos de dados para estudo deste e dos demais estimadores. A fração de cura, neste caso, é determinada quando a variável latente K_i assume valor 0 e, consequentemente, $U_i = 0$. A seguir, detalha-se o processo para geração dos conjuntos de dados:

1. São geradas as variáveis latentes K_i por meio de distribuições Poisson com média $\exp(\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2})/2$ para o i-ésimo indivíduo, com $x_{i1} \sim Bernoulli(0,5)$, $x_{i2} \sim N(0;1)$, $\theta_2 = 0$, e com θ_0 e θ_1 escolhidos de modo que as frações de cura para cada tratamento coincidam com seus valores desejados.

2. Gera-se, para cada unidade amostral, K_i variáveis independentes entre si com distribuição qui-quadrado central com 2 graus de liberdade. Adota-se a soma destas variáveis como sendo a variável latente U_i.

3. O risco dos tempos de sobrevida T_i condicionados a u_i e $\boldsymbol{x}_i$ é dado por $\lambda(t|u_i, \boldsymbol{x}_i) = \lambda_0(t) u_i \exp(\boldsymbol{\beta' x}_i)$, com $\beta_1 = 0$ e $\beta_2 = 0,5$. Assume-se risco basal $\lambda_0(t) = 1/2$, implicando que $T_i|(u_i, \boldsymbol{x}_i)$ apresenta distribuição exponencial com média $u_i \exp(\boldsymbol{\beta' x_i})/2$. Os tempos $T_i|(u_i, \boldsymbol{x}_i)$ são então facilmente obtidos pelo método da transformação inversa utilizando as quantidades U_i geradas no passo anterior. Denota-se os tempos condicionais gerados como apenas T_i.

4. Geram-se tempos de censura $C_i \sim \min(c_1, c_2 \times A)$, com $A \sim \text{Exp}(1)$. As constantes c_1 e c_2 são determinadas de acordo com a proporção de censuras desejada. Define-se então a variável indicadora de falhas $\delta_i = I(T_i \leq C_i)$.

5. Se $\delta_i = 0$, toma-se $L_i = C_i < R_i = \infty$.

6. Se $\delta_i = 1$, geram-se m instantes de observação a partir da soma de tempos com distribuição uniforme $Q_j \sim U(0,1; 0,5)$, de modo que $\sum_{j=0}^{m-1} Q_j \leq T_i < \sum_{j=0}^{m} Q_j$, em que $Q_0 = 0$

é o primeiro instante de observação. Satisfeita a inequação, faz-se então $L_i = \sum_{j=0}^{m-1} Q_j$ e $R_i = \sum_{j=0}^{m} Q_j$.

Com os dados gerados, estima-se os efeitos associados às covariáveis para cada um dos estimadores estudados e aplica-se a expressão para a obtenção da fração de cura específica de cada um variando-se o tratamento e tomando-se a segunda covariável como nula, como feito anteriormente.

Mantida a taxa de censura entre 35% e 40% e fixando-se frações de cura em 40% e 10% para os dois tratamentos por meio do modelo de fragilidade, foi obtido menor viés para as estimativas obtidas pelo modelo com a mesma especificação do cenário, mesmo com a ausência de convergência das estimativas na precisão estabelecida. As probabilidades de cobertura se mostram inferiores aos 95% pré-estabelecidos. Os resultados podem ser visualizados na Tabela 4.25.

Tabela 4.25: *Resultados para dados gerados por modelo de fragilidade com* $n = 200$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,414	0,014	0,005	0,068	0,049	0,830	1000
Mistura	T1	0,116	0,016	0,002	0,045	0,031	0,812	1000
Promoção	T0	0,422	0,022	0,004	0,063	0,048	0,837	993
Promoção	T1	0,113	0,013	0,002	0,039	0,026	0,796	993
Fragilidade	T0	0,404	0,004	0,005	0,073	0,053	0,845	1000
Fragilidade	T1	0,100	0,000	0,002	0,048	0,046	0,899	1000
Turnbull	T0	0,423	0,023	0,005	0,068			1000
Turnbull	T1	0,120	0,020	0,002	0,044			1000

Ao aumentar a taxa de censura do mesmo cenário, observam-se os resultados fornecidos na Tabela 4.26. O estimador do modelo de fragilidade novamente se sobressai em relação aos demais quanto ao viés médio. O modelo de tempo de promoção apresentou menor erro quadrático médio para as estimativas obtidas neste cenário e para aquele apresentado em Tabela 4.25.

Tabela 4.26: *Resultados para dados gerados por modelo de fragilidade com* $n = 200$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,489	0,089	0,027	0,137	0,049	0,383	1000
Mistura	T1	0,183	0,083	0,017	0,100	0,036	0,333	1000
Promoção	T0	0,504	0,104	0,022	0,107	0,048	0,385	938
Promoção	T1	0,182	0,082	0,014	0,084	0,033	0,359	938
Fragilidade	T0	0,475	0,075	0,027	0,147	0,069	0,513	1000
Fragilidade	T1	0,168	0,068	0,015	0,103	0,090	0,785	1000
Turnbull	T0	0,503	0,103	0,026	0,124			1000
Turnbull	T1	0,215	0,115	0,020	0,081			1000

Para T0 e T1 apresentando 30% e 20% de probabilidade de cura e mantida taxa de censura entre 35% e 40%, derivam-se os resultados apresentados na Tabela 4.27, em que os erros quadráticos médios encontram-se próximos para todos os algoritmos. Foi novamente observado um menor viés médio para as estimativas do estimador de fragilidade, embora estes apresentem menor probabilidade de cobertura para T0 quando comparados ao estimador de tempo de promoção.

Mantendo tais proporções de curados e aumentando-se o nível de censuras, observa-se, com exceção do erro quadrático médio inferior ao obtido pelo modelo de tempo de promoção, melhores resultados associados ao estimador de fragilidade, estimador associado aos dados do cenário considerado (ver Tabela 4.28).

Tabela 4.27: *Resultados para dados gerados por modelo de fragilidade com* $n = 200$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,314	0,014	0,004	0,060	0,046	0,855	1000
Mistura	T1	0,220	0,020	0,003	0,052	0,041	0,866	1000
Promoção	T0	0,320	0,020	0,003	0,054	0,044	0,873	993
Promoção	T1	0,220	0,020	0,003	0,048	0,038	0,856	993
Fragilidade	T0	0,305	0,005	0,004	0,062	0,048	0,863	1000
Fragilidade	T1	0,208	0,008	0,003	0,057	0,056	0,929	1000
Turnbull	T0	0,322	0,022	0,004	0,059			1000
Turnbull	T1	0,223	0,023	0,003	0,052			1000

Tabela 4.28: *Resultados para dados gerados por modelo de fragilidade com* $n = 200$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,379	0,079	0,023	0,128	0,047	0,399	1000
Mistura	T1	0,287	0,087	0,021	0,117	0,044	0,360	1000
Promoção	T0	0,395	0,095	0,019	0,101	0,046	0,418	962
Promoção	T1	0,293	0,093	0,017	0,093	0,042	0,399	962
Fragilidade	T0	0,358	0,058	0,021	0,134	0,062	0,541	1000
Fragilidade	T1	0,263	0,063	0,020	0,128	0,100	0,750	1000
Turnbull	T0	0,402	0,102	0,022	0,108			1000
Turnbull	T1	0,309	0,109	0,021	0,097			1000

Para dados gerados pelo modelo de fragilidade com $n = 400$, frações de cura $0,40$ e $0,10$ para T0 e T1, e nível intermediário de censura, o estimador de fragilidade apresentou o menor viés médio em conjunto com as maiores probabilidades de cobertura, embora não apresente convergência em nenhum caso sob os critérios estabelecidos (Tabela 4.29). Os resultados para maiores níveis de censura são exibidos em Tabela 4.30.

Tabela 4.29: *Resultados para dados gerados por modelo de fragilidade com* $n = 400$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,412	0,012	0,002	0,048	0,035	0,834	1000
Mistura	T1	0,114	0,014	0,001	0,030	0,022	0,849	1000
Promoção	T0	0,424	0,024	0,003	0,044	0,034	0,831	987
Promoção	T1	0,112	0,012	0,001	0,026	0,018	0,815	987
Fragilidade	T0	0,411	0,011	0,003	0,050	0,038	0,863	999
Fragilidade	T1	0,103	0,003	0,001	0,033	0,034	0,931	999
Turnbull	T0	0,424	0,024	0,003	0,047			1000
Turnbull	T1	0,115	0,015	0,001	0,032			1000

Para o cenário com frações de cura de $0,30$ e $0,20$ associadas aos tratamentos T0 e T1, respectivamente, obteve-se os resultados da Tabela 4.31. Sob este cenário, todos os estimadores estudados apresentaram erro quadrático médio similar. Embora os erros quadráticos médios encontrem-se próximos ao se aumentar o nível de censuras, o estimador de fragilidade apresenta menor viés e maior probabilidade de cobertura associados (ver Tabela 4.32).

Tabela 4.30: *Resultados para dados gerados por modelo de fragilidade com* $n = 400$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,484	0,084	0,019	0,108	0,035	0,316	934
Mistura	T1	0,168	0,068	0,010	0,076	0,025	0,304	934
Promoção	T0	0,499	0,099	0,017	0,084	0,034	0,268	957
Promoção	T1	0,176	0,076	0,010	0,065	0,023	0,264	957
Fragilidade	T0	0,482	0,082	0,018	0,106	0,050	0,454	1000
Fragilidade	T1	0,161	0,061	0,010	0,080	0,072	0,785	1000
Turnbull	T0	0,491	0,091	0,020	0,107			1000
Turnbull	T1	0,196	0,096	0,014	0,066			1000

Tabela 4.31: *Resultados para dados gerados por modelo de fragilidade com* $n = 400$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,317	0,017	0,002	0,043	0,033	0,841	1000
Mistura	T1	0,217	0,017	0,002	0,038	0,029	0,838	1000
Promoção	T0	0,323	0,023	0,002	0,040	0,031	0,832	980
Promoção	T1	0,219	0,019	0,002	0,034	0,027	0,835	980
Fragilidade	T0	0,314	0,014	0,002	0,045	0,034	0,857	1000
Fragilidade	T1	0,212	0,012	0,002	0,039	0,040	0,945	1000
Turnbull	T0	0,325	0,025	0,003	0,044			1000
Turnbull	T1	0,221	0,021	0,002	0,037			1000

Tabela 4.32: *Resultados para dados gerados por modelo de fragilidade com* $n = 400$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Mistura	T0	0,377	0,077	0,014	0,092	0,034	0,328	951
Mistura	T1	0,278	0,078	0,013	0,085	0,031	0,321	951
Promoção	T0	0,394	0,094	0,014	0,075	0,033	0,298	960
Promoção	T1	0,288	0,088	0,013	0,071	0,030	0,277	960
Fragilidade	T0	0,368	0,068	0,014	0,097	0,045	0,475	1000
Fragilidade	T1	0,265	0,065	0,013	0,095	0,080	0,774	1000
Turnbull	T0	0,390	0,090	0,016	0,090			1000
Turnbull	T1	0,293	0,093	0,015	0,080			1000

Quando aumentado o tamanho da amostra para $n = 800$, sob o cenário de fração de cura $0,40$ e $0,10$ para T0 e T1, com níveis médios de censura, o modelo de fragilidade apresentou estimativas com menor viés e erro quadrático médio associados (Tabela 4.33).

Para a fração de curados associada a T0 em dados com maiores proporções de censuras, o desempenho obtido pelo estimador não paramétrico de Turnbull mostrou-se similar ao obtido pelo estimador associado ao modelo gerador dos dados do cenário, o modelo de fragilidade, como pode ser visualizado em Tabela 4.34.

Ao considerar o cenário dos dados gerados pelo modelo de fragilidade com frações de cura $0,30$ e $0,20$ para T0 e T1, obtém-se, sob nível médio de censuras, os resultados da Tabela 4.35. Embora os erros quadráticos médios encontrem-se próximos para os diferentes modelos, o estimador com especificação de fragilidade apresentou menor viés e maior probabilidade de cobertura associados

Tabela 4.33: *Resultados para dados gerados por modelo de fragilidade com* $n = 800$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,425	0,025	0,002	0,030	0,024	0,757	968
Promoção	T1	0,114	0,014	0,001	0,019	0,013	0,743	968
Fragilidade	T0	0,417	0,017	0,001	0,034	0,027	0,830	1000
Fragilidade	T1	0,109	0,009	0,001	0,023	0,025	0,941	1000
Turnbull	T0	0,424	0,024	0,002	0,034			1000
Turnbull	T1	0,116	0,016	0,001	0,024			1000

Tabela 4.34: *Resultados para dados gerados por modelo de fragilidade com* $n = 800$*, frações de cura iguais a* 10% *e* 40% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,500	0,100	0,014	0,061	0,024	0,179	962
Promoção	T1	0,173	0,073	0,008	0,050	0,017	0,173	962
Fragilidade	T0	0,486	0,086	0,013	0,075	0,035	0,379	1000
Fragilidade	T1	0,161	0,061	0,008	0,062	0,058	0,766	1000
Turnbull	T0	0,484	0,084	0,015	0,092			1000
Turnbull	T1	0,175	0,075	0,009	0,057			1000

às suas estimativas. Aumentando a proporção de censurados em relação ao cenário anterior, obtém-se os resultados da Tabela 4.36. As probabilidades de cobertura mostraram-se baixas para este conjunto de parâmetros.

Tabela 4.35: *Resultados para dados gerados por modelo de fragilidade com* $n = 800$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 35% *e* 40%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,325	0,025	0,001	0,028	0,022	0,747	962
Promoção	T1	0,220	0,020	0,001	0,024	0,019	0,785	962
Fragilidade	T0	0,321	0,021	0,001	0,031	0,024	0,808	1000
Fragilidade	T1	0,215	0,015	0,001	0,028	0,028	0,916	1000
Turnbull	T0	0,327	0,027	0,002	0,031			1000
Turnbull	T1	0,220	0,020	0,001	0,027			1000

Tabela 4.36: *Resultados para dados gerados por modelo de fragilidade com* $n = 800$*, frações de cura iguais a* 20% *e* 30% *e taxa de censura entre* 60% *e* 65%

Método de estimação	Tratamento	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.	B^*
Promoção	T0	0,393	0,093	0,012	0,056	0,024	0,168	972
Promoção	T1	0,285	0,085	0,010	0,054	0,021	0,184	972
Fragilidade	T0	0,377	0,077	0,010	0,067	0,032	0,391	1000
Fragilidade	T1	0,269	0,069	0,009	0,068	0,061	0,692	1000
Turnbull	T0	0,382	0,082	0,013	0,077			1000
Turnbull	T1	0,281	0,081	0,011	0,068			1000

O viés e o erro quadrático médio das estimativas dos dados gerados pelo modelo de fragilidade são exibidos nas Figuras B.9 e B.10, para fração de cura de 40% e 10%. Os mesmos resultados para os cenários com 30% e 20% de curados associados a T0 e T1 são apresentados em B.11 e B.12. Em termos de viés, o estimador de Lam *et al.* (2013) apresentou valores, em geral, inferiores aos demais. Considerando o erro quadrático médio, o estimador de fragilidade mostrou melhor desempenho no cenário de alta taxa de censura cujas frações de cura para T0 e T1 são dadas por 30% e 20%. Na maioria dos casos, o estimador apresentou-se mais viesado para T0, consequência direta das estimativas menos precisas associadas ao intercepto dos efeitos de incidência, como pode ser visto nas Tabelas B.6, B.7 e B.8. Ainda considerando o erro quadrático médio, o estimador baseado no modelo de tempo de promoção mostra-se, em geral, como uma boa opção sob cenários cuja especificação da cura é dada pelo modelo de fragilidade.

4.4 Discussão

Na maior parte dos cenários, os diferentes estimadores apresentaram, em média, estimativas razoáveis da fração de cura, mesmo aplicados a dados com especificações diferentes do modelo utilizado para a estimação. O estimador não paramétrico de Turnbull, consolidado na literatura para análises descritivas sobre dados de censura intervalar, apresentou boas estimativas da proporção de curados para ambos os tratamentos, desempenhando melhor ou igual, em termos de viés e erro quadrático médio, do que alguns estimadores semiparamétricos para determinados cenários. Os resultados obtidos para o limite da curva estimada por este algoritmo incentivam seu uso para a estimação de frações de cura ao trabalhar-se com variáveis categóricas por meio de estratificações, como pode ser feito para o estimador de Kaplan-Meier. Além disso, mesmo para conjuntos de dados sem o limiar de cura, o estimador apresentou desempenho similar aos demais quando se considera o último salto estimado como estimativa da fração de curados, conforme indicam os resultados da Tabela 4.16.

Sob os critérios considerados, as análises por meio de simulação mostram também que o estimador associado à especificação correta dos dados gerados de um cenário não necessariamente proporciona melhores estimativas, como evidencia o cenário de dados gerados pelo modelo de mistura com taxa alta de censuras e frações de cura de 40% e 10%: em média, as estimativas do modelo de fragilidade apresentam menor viés e valores próximos associados ao erro quadrático médio, se comparadas com as estimativas obtidas pelo estimador de mistura padrão. Entretanto, esta relação não se mantém para o cenário que se diferencia apenas por proporções de curados de 30% e 20%.

Em termos de viés, o modelo de fragilidade aparenta ser o candidato mais adequado para cenários com alto nível de censura, independente do mecanismo de fração de cura adotado para a geração dos dados. Para baixos níveis de censura, os estimadores com especificações iguais aos dados gerados mostraram-se menos viesados. Tratando-se do erro quadrático médio, sob taxas altas de censura e tamanhos de amostra $n = 200$ ou $n = 400$, o estimador de tempo de promoção apresenta na maior parte dos casos valores inferiores ou iguais aos demais, mesmo sem adotar a convergência da verossimilhança esperada como critério de parada. Tais propriedades empiricamente obtidas podem auxiliar a escolha do pesquisador em situações com especificações indefinidas associadas à fração de cura dos dados.

As simulações realizadas neste trabalho apresentam como limitação as avaliações restritas às proporções de curados, entretanto, um estudo das significâncias dos efeitos associados às covariáveis pode ser realizado para cada especificação de fração de cura.

Alguns padrões observados nas simulações auxiliam a escolha do estimador para as aplicações apresentadas: o menor viés para as estimativas do modelo de fragilidade associadas a cenários de maior nível de censura e amostras de tamanho $n = 800$ justifica a escolha para os dados de anemia, em que o evento de interesse possui baixa frequência; entre os modelos sem a estrutura de grupos, o modelo de tempo de promoção apresenta os menores erros quadráticos médios associados aos cenários de alta censura, tamanho $n = 200$ ou $n = 400$ e fração de cura 40% (características similares às do conjunto de cervos) para qualquer especificação de modelo de cura, incentivando o

uso para os dados de migração, conjunto que apresenta limiar de cura.

Resultados assintóticos foram utilizados para a obtenção dos erros padrão associados às estimativas das proporções de curados, o que pode não proporcionar boas aproximações para as quantidades em questão em vigor dos tamanhos de amostra adotados, como evidenciam as diferenças entre os erros padrão médios e os desvios padrão empíricos. Estudos mais aprofundados para a avaliação dos estimadores de fração de cura para amostras pequenas são necessários e incentivam trabalhos posteriores.

Os diferentes custos computacionais associados aos diferentes estimadores proporcionaram dificuldades na uniformização dos critérios de convergência, demandando as particularidades apresentadas para realizar o experimento em tempo viável. Entretanto, mesmo não observando a convergência da verossimilhança esperada para o modelo de tempo de promoção, o estimador deste modelo apresentou resultados satisfatórios ou mesmo superiores em alguns aspectos para grande parte dos cenários. Embora o algoritmo proposto em Lam *et al.* (2013) seja altamente intensivo, o processo é facilmente paralelizável, com implementações contemplando tal estrutura disponíveis no CRAN por meio do pacote *intercure*. Devido ao fato de as implementações dos diferentes modelos não se apresentarem otimizadas, discussões a respeito da performance computacional dos mesmos encontram-se fora do escopo deste trabalho.

Capítulo 5

Conclusão

Uma das principais contribuições deste trabalho foi dada pela análise de conjuntos de dados reais motivados pelo contexto de cada um deles quanto à fração de cura. Conforme apresentado nas aplicações, a escolha de diferentes especificações da proporção de curados pode acarretar grandes diferenças nas estimativas.

Como exemplo, foi visto que o modelo de tempo de promoção estima a proporção de mulheres não suscetíveis à anemia com idade inferior a 50 anos, fixado hematócrito médio e frequências históricas de doação média, em 81,2%. Sob as mesmas condições, o modelo de fragilidade estimou a fração de cura em 85,5%. Em particular, para o problema de doadores de repetição, o evento de anemia é dado como raro, proporcionando taxas altas de censura, o que pode justificar a escolha do modelo de fragilidade, fortificada pelo desempenho do estimador em cenários simulados com grande proporção de censuras. A incorporação de efeitos específicos para indivíduos fornece motivação adicional ao uso do modelo por conta da qualidade de vida específica e variabilidade genética dos diferentes doadores.

Para o conjunto de dados de migração, determinadas áreas de estudo ("D" e "I") apresentaram frações de cura próximas estimadas pelos diferentes algoritmos. Para este conjunto de dados, foi também incorporado o efeito de grupo às estimativas associadas ao modelo de mistura e de fragilidade, com as proporções de curados obtidas pelo modelo de mistura para dados agrupados próximas às obtidas pelo mesmo modelo desconsiderando-se tal estrutura. Valores mais altos de erro padrão podem ser observados para as estimativas envolvendo efeito de grupo para ambos os estimadores de fragilidade e mistura, potencialmente justificados pelo pequeno tamanho de amostra. Por conta do uso do método delta para a estimação dos erros padrão associados às frações de cura, a aproximação para a normalidade pode não ser boa ao se considerar o tamanho da amostra. Assim como para os dados de doadores de sangue, todos os estimadores apresentaram estimativas associadas à incidência condizentes com o padrão esperado para este conjunto de dados.

Conforme visto nas simulações, as estimativas obtidas pelo algoritmo de Turnbull apresentaram, para diferentes cenários provenientes de diferentes especificações da fração de cura, viés e erro quadrático médio próximos aos estimadores semiparamétricos aqui estudados. Pouco encontra-se na literatura quanto ao uso do estimador não paramétrico da função de sobrevivência para a estimação da fração de cura, com os trabalhos de Aljawadi *et al.* (2012b) comparando o viés de tais estimativas com as obtidas pelo estimador de Kaplan-Meier com a imputação de pontos médios para o tempo de sobrevida, estes gerados por meio de modelos de tempo de promoção. O trabalho apresentado visa o estudo da performance do estimador para diferentes mecanismos de geração de dados com proporção de curados, além de considerar o uso do estimador em cenários que não apresentam o limiar de cura, que ainda não havia sido abordado. Justifica-se seu uso desta forma por meio de conhecimentos *a priori* da existência da proporção de curados e por conta das boas estimativas obtidas nas simulações e na aplicação a dados de câncer.

Em geral, as estimativas semiparamétricas apresentaram vícios próximos para simulações de baixo nível de censura. Para níveis altos de censura, o estimador obtido a partir do modelo de fragilidade apresentou menor viés em relação aos demais nas simulações realizadas. O estimador

de mistura padrão apresentou tempos computacionais inviáveis por iteração em seu processo de estimação para amostras geradas de tamanho $n = 800$, assim como em sua aplicação aos conjuntos de dados de doadores de sangue. O aumento do número de parâmetros a serem estimados com o aumento de intervalos de observação distintos, em conjunto com problemas de identificabilidade do estimador, podem explicar as dificuldades computacionais do processo de estimação.

Também foi observado um bom desempenho associado ao modelo de tempo de promoção em relação aos demais, mesmo sem adotar a convergência da log-verossimilhança esperada como critério de parada, quanto ao erro quadrático médio em cenários de alta censura, amostras de tamanho 200 ou 400 e fração de cura próxima de 40%, incentivando sua aplicação aos dados de migração que apresentam características similares. Embora as simulações estudadas sugiram cenários que privilegiam determinados estimadores, estas informações devem ser vistas como complementares, com a escolha do modelo baseando-se principalmente na motivação biológica do problema.

As rotinas computacionais implementadas no decorrer deste trabalho encontram-se no repositório público CRAN (*The Comprehensive R Archive Network*) por meio do pacote *intercure*, tendo como próximos trabalhos a manutenção e disponibilização de novas funcionalidades no pacote. É importante ressaltar que os estudos apresentados neste trabalho possuem foco na estimação da fração de cura, entretanto, probabilidades de o evento de interesse ocorrer após um determinado tempo podem ser estudadas por meio dos modelos apresentados. Alternativas dos mesmos estimadores sob o paradigma bayesiano não foram avaliadas nestes estudos. Além disso, há também interesse em obter uma correção de probabilidades estimadas para eventos raros utilizando-se modelos de fração de cura, assim como extensões para o suporte a covariáveis dependentes do tempo.

Apêndice A

Materiais Suplementares dos Estimadores

A.1 Estimadores de Máxima Verossimilhança Restrita para o Modelo de Mistura

A obtenção do estimador de máxima verossimilhança restrita para θ_1 e θ_2, do modelo proposto por Xiang *et al.* (2011), dá-se conforme a seguir. Sejam os vetores $\eta = (\eta_{11}, \cdots, \eta_{Mn_M})'$ e $\xi = (\xi_{11}, \cdots, \xi_{Mn_M})'$ os preditores lineares introduzidos em (2.22) e (2.23). Expressa-se então em forma matricial: $\eta = X\beta + ZU$ e $\xi = Wb + ZV$ em que $X = (x_{11}, \cdots, x_{Mn_M})'$ e $W = (w_{11}, \cdots, w_{Mn_M})'$ são as matrizes modelo de dimensão $N \times p$ e $N \times (p+1)$ dos respectivos efeitos β e b; $Z = (z_1, \cdots, z_M)$ é matrix de dimensão $N \times M$ cujos vetores coluna z_i possuem elementos iguais a 1 para indivíduos do grupo i, com 0 caso contrário. Com isso, fixados os valores latentes $y^{(r)}$, a maximização da log-verossimilhança l_C dá-se através do seguinte processo iterativo:

$$\begin{bmatrix} \hat{\beta} \\ \hat{U} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ U_0 \\ \gamma_0 \end{bmatrix} + B_\beta^{-1} \left\{ \begin{bmatrix} X' & 0 \\ Z' & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \partial l_\beta/\partial\eta \\ \partial l_\beta/\partial\gamma \end{bmatrix} - \begin{bmatrix} 0 \\ U_0/\theta_1 \\ 0 \end{bmatrix} \right\} \text{ e} \tag{A.1}$$

$$\begin{bmatrix} \hat{b} \\ \hat{V} \end{bmatrix} = \begin{bmatrix} b_0 \\ V_0 \end{bmatrix} + B_b^{-1} \left\{ \begin{matrix} W' \\ Z' \end{matrix} \begin{bmatrix} \partial l_b/\partial\xi \end{bmatrix} - \begin{bmatrix} 0 \\ V_0/\theta_2 \end{bmatrix} \right\}, \tag{A.2}$$

em que β_0, b_0, U_0, V_0 e γ_0 são valores fixos e atualizados a cada iteração para dados valores θ_1 e θ_2. B_β e B_b denotam as matrizes de derivadas segundas de l_β e l_b em relação a (β, U, γ) e (b, V), respectivamente, podendo ser expressas por

$$B_\beta = \begin{bmatrix} X' & 0 \\ Z' & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} -\frac{\partial^2 l_\beta}{\partial\eta\partial\eta'} & -\frac{\partial^2 l_\beta}{\partial\eta\partial\gamma'} \\ -\frac{\partial^2 l_\beta}{\partial\gamma\partial\eta'} & -\frac{\partial^2 l_\beta}{\partial\gamma\partial\gamma'} \end{bmatrix} \begin{bmatrix} X & Z & 0 \\ 0 & 0 & I \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & U_0/\theta_1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ e}$$

$$B_b = \begin{bmatrix} W' \\ Z' \end{bmatrix} \begin{bmatrix} -\frac{\partial^2 l_\beta}{\partial\eta\partial\eta'} \end{bmatrix} \begin{bmatrix} W \\ Z \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & V_0/\theta_2 \end{bmatrix} 0 \Big].$$

Baseados em Sun (2006), os autores denotam:

$$S(t_q; x_{ij}) = \exp\left\{ -\sum_{k=1}^{q} \phi_{ijk} \right\} \quad \text{com} \quad \phi_{ijk} = \exp(\gamma_k + \eta_{ij});$$

$$g_{ij} = \sum_{q=1}^{Q} \alpha_{ijq} \left\{ S(t_{q-1}; x_{ij}) - S(t_q; x_{ij}) \right\},$$

$$c_{ijq} = \sum_{l=q}^{Q} (\alpha_{ijl} - \alpha_{ijl+1}) S(t_l; x_{ij}),$$

$$d_{ijq} = \sum_{l=q}^{Q} (\alpha_{ijl} - \alpha_{ijl+1}) S(t_l; x_{ij}) \log S(t_l; x_{ij}),$$

$$f_{ijq} = S(t_q; x_{ij}) \log S(t_q; x_{ij}), \quad f_{ij0} = f_{ijQ} = 0,$$

$$h_{ijq} = f_{ijq} \log S(t_q; x_{ij}) + f_{ijq}, \quad h_{ij0} = h_{ijQ} = 0.$$

Obtém-se os componentes das derivadas $\partial l_\beta/\partial\eta$, $\partial l_b/\partial\xi$, $\partial l_\beta/\partial\gamma$, $\partial^2 l_\beta/\partial\eta\partial\eta'$, $\partial^2 l_b/\partial\xi\partial\xi'$, $\partial^2 l_\beta/\partial\gamma\partial\gamma'$, e $\partial^2 l_\beta/\partial\eta\partial\gamma'$, conforme abaixo:

$$\frac{\partial l_\beta}{\partial \eta_{ij}} = y_{ij}/g_{ij} \sum_{q=1}^{Q} \alpha_{ijq} (f_{ijq-1} - f_{ijq}),$$

$$\frac{\partial l_b}{\partial \xi_{ij}} = y_{ij} - \pi_{ij},$$

$$\frac{\partial l_\beta}{\partial \gamma_q} = \sum_{i,j} y_{ij}\phi_{ijq}c_{ijq}/g_{ij},$$

$$-\frac{\partial^2 l_\beta}{\partial \eta_{ij}\partial \eta_{ij}} = y_{ij}\left[\left\{\frac{\sum_{q=1}^{Q}\alpha_{ijq}(f_{ijq-1}-f_{ijq})}{g_{ij}}\right\}^2 - \frac{\sum_{q=1}^{Q}\alpha_{ijq}(h_{ijq-1}-h_{ijq})}{g_{ij}}\right],$$

$$-\frac{\partial^2 l_b}{\partial \xi_{ij}\partial \xi_{ij}} = \pi_{ij}(1-\pi_{ij}),$$

$$-\frac{\partial^2 l_\beta}{\partial \gamma_q^2} = \sum_{i=1}^{M}\sum_{j=1}^{n_i} y_{ij}\phi_{ijq}c_{ijq}/g_{ij}\left\{\frac{\phi_{ijq}c_{ijq}}{g_{ij}} - (1-\phi_{ijq})\right\},$$

$$-\frac{\partial^2 l_\beta}{\partial \gamma_q \gamma_k} = \sum_{i=1}^{M}\sum_{j=1}^{n_i} y_{ij}\left(\frac{\phi_{ijq}\phi_{ijk}c_{ijq}c_{ijk}}{g_{ij}^2} + \frac{\phi_{ijq}\phi_{ijk}c_{ijk}}{g_{ij}}\right), \quad \text{para} \quad q<k,$$

$$-\frac{\partial^2 l_\beta}{\partial \gamma_q \eta_{ij}} = y_{ij}\phi_{ijq}\left\{\frac{c_{ijq}}{g_{ij}^2}\sum_{l=1}^{Q}\alpha_{ijl}(f_{ijl-1}-f_{ijl}) - \frac{c_{ijq}+d_{ijq}}{g_{ij}}\right\},$$

$$-\partial^2 l_\beta/\partial\eta_{ij}\partial\eta_{kl} = 0, -\partial^2 l_b/\partial\xi_{ij}\partial\xi_{kl} = 0, \quad \text{para} \quad (i,j)\neq(k,l).$$

Suponha $\boldsymbol{B}_\beta$ particionado de acordo com $\boldsymbol{\beta}|U|\gamma$ como

$$\boldsymbol{B}_\beta^{-1} = \begin{bmatrix} A_{11}A_{12}A_{13} \\ A_{21}A_{22}A_{23} \\ A_{31}A_{32}A_{33} \end{bmatrix} \tag{A.3}$$

e $\boldsymbol{B_b}$ uma partição de acordo com $\boldsymbol{b}|V$ tal que

$$\boldsymbol{B}_\beta^{-1} = \begin{bmatrix} A_{44}A_{45} \\ A_{54}A_{55} \end{bmatrix}. \tag{A.4}$$

Aplicando os resultados de McGilchrist (1993), os autores derivam então os estimadores de maxima verossimilhança restrita para θ_1 e θ_2:

$$\hat{\theta_1} = M^{-1}\left\{tr(A_{22}) + \hat{U}'\hat{U}\right\},$$
$$\hat{\theta_2} = M^{-1}\left\{tr(A_{55}) + \hat{V}'\hat{V}\right\},$$

com as variâncias assintóticas de $(\hat{\beta}, \hat{b}, \hat{\theta_1}, \hat{\theta_2})$:

$$\begin{aligned}
\text{var}(\hat{\beta}) &= A_{11} \\
\text{var}(\hat{b}) &= A_{44} \\
\text{var}(\hat{\theta_1}) &= 2\theta_1^2\left\{M - 2\theta_1^{-1}tr(A_{22}) + \theta_1^{-2}tr(A_{22}^2)\right\}^{-1} \\
\text{var}(\hat{\theta_2}) &= 2\theta_2^2\left\{M - 2\theta_2^{-1}tr(A_{55}) + \theta_2^{-2}tr(A_{55}^2)\right\}^{-1}
\end{aligned}$$

em que $tr(A)$ representa o traço da matriz A.

A.2 Estimação de vetor de probabilidades para modelo de tempo de promoção

Esta seção visa apresentar a metodologia proposta em Liu e Shen (2009) para a resolução de (2.43), ou seja, a partir das quantidades definidas na subseção 2.3.3, deseja-se maximizar a expressão

$$f(\boldsymbol{p}) \equiv -\boldsymbol{a^T p} + \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^{(k)} \log\left[1 - \exp(-e^{\boldsymbol{\theta' \tilde{Z}_i}} p_j)\right],$$

com $\sum_{j=1}^{m} p_j = 1$ e $0 \leq p_j \leq 1$, com $j = 1, \cdots, m$.

Conforme exibido em Liu e Shen (2009), tem-se as derivadas parciais de primeira e segunda ordem em relação a p_j dadas por:

$$\dot{f}_j(\boldsymbol{p}) = \frac{\delta}{\delta p_j} f(\boldsymbol{p}) = -a_j + \sum_{i=1}^{n} \frac{X_{ij}^{(k)} e^{\boldsymbol{\theta' \tilde{Z}_i}}}{\exp(e^{\boldsymbol{\theta' \tilde{Z}_i}} p_j) - 1} \text{ para } j = 1, \cdots, m, \tag{A.5}$$

$$\ddot{f}_j(\boldsymbol{p}) = \frac{\delta^2}{\delta p_j^2} f(\boldsymbol{p}) = -\sum_{i=1}^{n} \frac{X_{ij}^{(k)} e^{2\boldsymbol{\theta' \tilde{Z}_i}} \exp(e^{\boldsymbol{\theta' \tilde{Z}_i}} p_j)}{(\exp(e^{\boldsymbol{\theta' \tilde{Z}_i}} p_j) - 1)^2} \text{ para } j = 1, \cdots, m. \tag{A.6}$$

Para $j \neq j'$, com $j, j' = 1, \cdots, m$, tem-se $\frac{\delta^2}{\delta p_j p_j'} f(\boldsymbol{p}) = 0$, ou seja, a matriz hessiana é diagonal. Tal fato permitiu a Liu e Shen (2009) o desenvolvimento de um simples e eficiente algoritmo.

Primeiramente, apresenta-se a solução ótima com o uso dos multiplicadores de Lagrange. Seja v e τ_j, $j = 1, \cdots, m$ os multiplicadores referentes às restrições de igualdade e desigualdade, respectivamente. Com isso, faz-se $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_m)$ e $1 = (1, \cdots, 1)^T$, e assim, o problema é visto como a maximização da função $\Gamma(\boldsymbol{p}, \boldsymbol{\tau}, v) = f(\boldsymbol{p}) + \boldsymbol{\tau^T p} + v(1 - 1^T \boldsymbol{p})$. No contexto de otimização convexa, as condições necessárias e suficientes para que um ponto $\boldsymbol{p^*}$ seja ótimo estão expressas nas condições de Karush-Kuhn-Tucker (KKT), devidamente apresentadas em Boyd e Vandenberghe (2004). Especificamente, os pontos $\boldsymbol{p^*}$ em $[0,1]^m$, $\boldsymbol{\tau^*} = (\tau_1^*, \cdots, \tau_m^*)$ em $\mathbb{R}^m$ e v^* em $\mathbb{R}$ compõem a solução ótima se e somente se

$$\begin{aligned} &\sum_{j=1}^{m} p_j^* = 1, \quad p_j^* \geq 0, \ \tau_j^* \geq 0, \ \tau_j^* p_j^* = 0, \\ &\dot{f}_j(\boldsymbol{p^*}) + \tau_j^* - v^* = 0, \quad j = 1, \cdots, m. \end{aligned} \tag{A.7}$$

Devido ao fato da matriz hessiana de $f(\boldsymbol{p})$ ser diagonal, e eliminando por substituição τ_j, tem-se então

$$\sum_{j=1}^{m} p_j^* = 1, \quad p_j^* \geq 0, \ v^* \geq \dot{f}_j(\boldsymbol{p^*}), \ [v^* - \dot{f}_j(\boldsymbol{p^*})] p_j^* = 0, \quad j = 1, \cdots, m. \tag{A.8}$$

Entretanto, se $p_j^* = 0$, tem-se $\dot{f}_j(\boldsymbol{p^*}) = \infty$, contradizendo a segunda desigualdade apresentada. Assim, os pontos de máximo $\boldsymbol{p^*}$ e v^* devem satisfazer:

$$\dot{f}_j(\boldsymbol{p^*}) - v^* = 0, \ \sum_{j=1}^{m} p_j^* = 1, \text{ e } p_j^* > 0, \quad j = 1, \cdots, m. \tag{A.9}$$

Uma implicação direta das condições acima é que, para $j = 1, \cdots, m$, tem-se $p_j < 1$. Tais condições sugerem valores iniciais no "interior", ou seja, $0 < p_j < 1$ para cada j. As condições KKT contribuem no sentido de que as restrições do ponto interior se mantém a cada iteração do algoritmo

ECM.

A maximização de F através da resolução do sistema não-linear (A.7) é, na maior parte dos casos, um processo computacional intensivo e não trivial. Liu e Shen (2009) sugerem o uso do método de Newton, aproximando as equações não lineares por expansões de Taylor, reduzindo o problema à resolução das seguintes equações com passos $(\Delta \boldsymbol{p}, \Delta \boldsymbol{\tau}, \Delta v)$:

$$\begin{bmatrix} \nabla^2 f(\boldsymbol{p}) & \boldsymbol{I} & \mathbf{1} \\ \text{diag}(\boldsymbol{\tau}) & \text{diag}(\boldsymbol{p}) & \mathbf{0} \\ \boldsymbol{-1^T} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{p} \\ \Delta \boldsymbol{\tau} \\ \Delta v \end{bmatrix} = \begin{bmatrix} \nabla f(\boldsymbol{p}) + \boldsymbol{\tau} - v\mathbf{1} \\ \text{diag}(\boldsymbol{\tau})\boldsymbol{p} - (\mathbf{1}/\boldsymbol{\eta})\mathbf{1} \\ 0 \end{bmatrix}, \tag{A.10}$$

em que $\nabla^2 f(\boldsymbol{p})$ é a matriz hessiana, $\text{diag}(\boldsymbol{p})$ e $\text{diag}(\boldsymbol{\tau})$ são matrizes diagonais com elementos $p_1, \cdots, p_m$ e $\tau_1, \cdots, \tau_m$ em suas respectivas diagonais principais, $\boldsymbol{I}$ é matriz identidade e $\mathbf{1}$ é um vetor composto apenas por 1 em seus componentes, com η sendo um escalar positivo para o controle das restrições de desigualdade. A construção de tal sistema com notação similar pode ser visualizada na página 610 de Boyd e Vandenberghe (2004).

Como $\nabla^2 f(\boldsymbol{p})$ é uma matriz diagonal, o sistema (A.10) pode ser resolvido explicitamente. As seguintes derivações podem ser obtidas, denotando $\tau_j^+ = \tau_j + \Delta\tau_j$ e $v^+ = v + \Delta v$, conforme disponível para o leitor em Liu e Shen (2009):

$$\tau_j^+ = v^+ - \ddot{f}_j(\boldsymbol{p})\Delta p_j - \dot{f}_j(\boldsymbol{p}),\ j = 1, \cdots, m,$$

$$\Delta p_j = \frac{-p_j v^+ + p_j \dot{f}_j(\boldsymbol{p}) + 1/\eta}{\tau_j - p_j \ddot{f}_j(\boldsymbol{p})},$$

$$v^+ = \frac{\sum_{j=1}^m \left[(1/\eta + p_j \dot{f}_j(\boldsymbol{p}))/(\tau_j - p_j \ddot{f}_j(\boldsymbol{p}))\right]}{\sum_{j=1}^m \left[p_j/(\tau_j - p_j \ddot{f}_j(\boldsymbol{p}))\right]},$$

válidas quando não chegam a uma indeterminação por conta do denominador nulo. Com isso, tem-se o necessário para calcular o passo do algoritmo de Newton Δp_j para cada $j = 1, \cdots, m$. Após calcular τ_j^+ através de (A.2), obtém-se diretamente os passos de Newton $\Delta\tau_j = \tau_j^+ - \tau_j$ e $\Delta v = v^+ - v$. O processo iterativo do algoritmo de Newton fica então determinado pelas seguintes atualizações a cada iteração:

$$p_j \leftarrow p_j + \psi\Delta p_j \quad \tau_j \leftarrow \tau_j + \psi\Delta\tau_j \quad v \leftarrow v + \psi\Delta v$$

em que ψ é interpretado como o tamanho do passo. Para este, além de levar em conta restrições de KKT, deve-se também considerar que passos "pequenos" proporcionam maior trabalho computacional enquanto passos "grandes" não asseguram a convergência do algoritmo. Neste trabalho usaremos para o cálculo de ψ , assim como usado em Liu e Shen (2009), o mecanismo de busca linear padrão *backtracking*, o qual assegura que as restrições mantenham-se satisfeitas. A implementação deste mecanismo é descrita e apresentada no Apêndice A.3, podendo ser encontrada com maiores detalhes nas notas auxiliares de Liu e Shen (2009) disponibilizadas online, ou de forma mais geral em Boyd e Vandenberghe (2004).

A.3 Busca linear do algoritmo primal-dual de pontos interiores

Conforme mencionado neste trabalho e em Liu e Shen (2009), detalhes do algoritmo a seguir podem ser encontrados em Boyd e Vandenberghe (2004). Métodos do ponto interior forçam a busca em direção de um ponto ótimo contido na área definida pelas restrições de desigualdade. O termo "primal-dual" se refere a problemas homônimos de Lagrange que providenciam suporte teórico para as técnicas de otimização convexa. Os autores ainda mencionam o fato de que estes algoritmos tornaram-se amplamente utilizados na prática devido à sua eficiência e alta acurácia em problemas de otimização não linear convexa com restrições para um número grande de dimensões.

Para a estimação de parâmetros proposta por Liu e Shen (2009), o método primal-dual do ponto interior é um procedimento iterativo. Dadas as iterações atuais $\boldsymbol{p^{(l)}}$, $\boldsymbol{\tau^{(l)}}$, $v^{(l)}$, $l = 0, 1, \cdots$, a próxima iteração é obtida então por

$$p_j^{(l+1)} = p_j^{(l)} + \psi \Delta p_j, \quad \tau_j^{(l+1)} \leftarrow \tau_j^{(l)} + \psi \Delta \tau_j, \quad v^{(l+1)} \leftarrow v^{(l)} + \psi \Delta v,$$

em que ψ é determinado por um procedimento *backtracking* padrão de busca linear para garantir $p_j^{(l+1)} > 0$ e $\tau_j^{(l+1)} > 0$. A busca linear primeiramente calcula o comprimento do maior passo positivo para $\boldsymbol{\tau}$ que seja menor ou igual a 1, ou seja,

$$\psi_{\max}^{(l)} = \min\{1, \min\{-\tau_i^{(l)}/\Delta\tau_i | \Delta\tau_i < 0\}\}. \tag{A.11}$$

O algoritmo então procede com $\psi = 0,99\psi_{\max}^{(l)}$, multiplicando ψ por $\rho \in (0, 1)$ até que $p_j^{(l+1)} > 0$. O *backtracking* continua então multiplicando ψ por ρ até que

$$||F_\eta(\boldsymbol{p^{(l+1)}}, \boldsymbol{\tau^{(l+1)}}, v^{(l+1)})||_2 \leq (1 - \pi\psi)||F_\eta(\boldsymbol{p^{(l)}}, \boldsymbol{\tau^{(l)}}, v^{(l)})||_2, \tag{A.12}$$

em que

$$F_\eta(\boldsymbol{p}, \boldsymbol{\tau}, v) = \begin{pmatrix} \nabla f(\boldsymbol{p}) + \boldsymbol{\tau} - v\mathbf{1} \\ \text{diag}(\boldsymbol{\tau})\boldsymbol{p} - 1/\eta\mathbf{1} \\ 1 - \boldsymbol{p^T}\mathbf{1} \end{pmatrix}. \tag{A.13}$$

Conforme Boyd e Vandenberghe (2004), o valor de π é tipicamente escolhido como sendo entre $0,01$ e $0,1$, e ρ escolhido entre $0,3$ e $0,8$. O valor para η é escolhido como sendo um fator σ multiplicado por $1/\hat{\mathcal{E}}$, em que $m/\hat{\mathcal{E}} = \sum_{j=1}^m p_j^{(l)}\tau_j^{(l)}$ é dito ser o atual hiato de dualidade. Atualiza-se p_j, τ_j, v, até que

$$m\hat{\mathcal{E}} < \epsilon_2 \quad \text{e} \quad \left[\sum_{j=1}^m (\dot{f}_j(\boldsymbol{p}) + \tau_j - v)^2\right]^{1/2} < \epsilon_2,$$

para um valor ϵ_2 pequeno pré-especificado.

A.4 Consistência do estimador no modelo de tempo de promoção

Devido aos desafios técnicos encontrados por conta da censura intervalar, o estudo para a demonstração da consistência forte para o estimador de máxima verossimilhança $(\hat{\boldsymbol{\theta}_n}, \hat{F}_n)$ com a estrutura apresentada de fração de cura é feito utilizando-se a distância de Hellinger, providenciando consistência global. A distância de Hellinger é uma distância L_2 definida como

$$h(q_1, q_2) = \left(\frac{1}{2}\int(\sqrt{q_1} - \sqrt{q_2})^2 dv\right)^{1/2} = \left(1 - \int\sqrt{q_1 q_2}dv\right)^{1/2} \tag{A.14}$$

tal qual q_1 e q_2 não dependem da medida dominante v. Esta é uma verdadeira medida de distância que satisfaz $h(q_1, q_2) \leq 1$ e $h(q_1, q_2) = 0$ se e somente se $q_1 = q_2$. A distância de Hellinger foi utilizada em van der Vaart e Wellner (2000) para a demonstração da consistência forte do caso não paramétrico com censura intervalar. Liu e Shen (2009) propoem uma extensão para estimadores de máxima verossimilhança semiparamétricos com fração de cura e censura intervalar, cuja prova é amplamente baseada na teoria apresentada em van der Vaart e Wellner (1996).

Antes da apresentação do teorema a respeito da consistência do estimador, introduz-se aqui uma construção similar à utilizada em van der Vaart e Wellner (2000), necessária para lidar com a aleatoriedade dos tempos de inspeção. O uso desta construção, com a modificação para permitir uma proporção de curados, leva à mesma função de verossimilhança apresentada na expressão (2.32), para a qual deriva-se o eficiente algoritmo apresentado.

Toma-se K como um inteiro aleatório positivo que denota o número de inspeções para uma pessoa, e $\boldsymbol{Y_K} = \{Y_{K,1}, \cdots, Y_{K,K}\}$ denota os tempos de inspeção para o evento de interesse, em que $Y_{K,1} < \cdots < Y_{K,K}$. Na prática, é natural assumir $E_0(K) < \infty$. Os tempos aleatórios de observação para uma pessoa são denotados pela matriz triangular $\boldsymbol{Y} = \{Y_{k,j} : j = 1, \cdots, k, k = 1, 2, \cdots\}$, em que k é valor observado de K. Seja então $\boldsymbol{Y_k} = \{Y_{k,1}, \cdots, Y_{k,k}\}$ a $k-$ésima linha da matriz $\boldsymbol{Y}$ com realização denotada por $\boldsymbol{y_k} = \{y_{k,1}, \cdots, y_{k,k}\}$.

A informação que tem-se a respeito do tempo de evento T é que o mesmo encontra-se em um dos intervalos $[Y_{K,j-1}, Y_{K,j})$, $j = 1, \cdots, K$, ou $[Y_{K,K}, Y_{K,K+1}]$, em que $Y_{K,0} \equiv 0$ e $Y_{K,K+1} \equiv \infty$. Denota-se também $\boldsymbol{\Lambda_k} = (\Lambda_{k,1}, \cdots, \Lambda_{k,k+1})$, em que $\Lambda_{k,j} = 1_{[Y_{k,j-1}, Y_{k,j})}(T)$ para $j = 1, \cdots, k$, e $\Lambda_{k,k+1} = 1_{[Y_{k,k}, Y_{k,k+1})}(T)$ com valores observados denotados por $\lambda_k = (\lambda_{k,1}, \cdots, \lambda_{k,k+1})$.

Assume-se então que, condicionado ao vetor de covariáveis $\boldsymbol{Z}$, o par $(K, \boldsymbol{Y})$ é independente do tempo de evento T. Também é feita pelos autores a suposição de inspeção não informativa tal que a distribuição de $(K, \boldsymbol{Y_K}, \boldsymbol{Z})$ não dependa do tempo de evento T e dos parâmetros de interesse $\boldsymbol{\theta}$ e F. Dado o vetor de covariáveis $\boldsymbol{Z}$, a função de sobrevivência condicional de T segue o modelo de cura apresentado em (2.31). Assim, tem-se a seguinte distribuição condicional para $\boldsymbol{\Lambda_k}$:

$$\begin{aligned}(\boldsymbol{\Lambda_k}|K, \boldsymbol{Y_K}, \boldsymbol{Z}) &\sim \text{Multinomial}_{K+1}(1, \Delta S_K(\boldsymbol{Z})),\\ \Delta S_K(\boldsymbol{Z}) &\equiv (1 - S(Y_{K,1}|\boldsymbol{Z}), \cdots, S(Y_{K,K-1}|\boldsymbol{Z}) - S(Y_{K,K}|\boldsymbol{Z}), S(Y_{K,K}|\boldsymbol{Z})).\end{aligned}$$

Denotando-se $\boldsymbol{V} = (K, \Lambda_k, \boldsymbol{Y}_K, \boldsymbol{Z})$ com realização dada por $\boldsymbol{v} = (k, \lambda_k, \boldsymbol{y}_k, \boldsymbol{z})$ e fazendo-se $\tilde{\boldsymbol{z}}^T = (1, \boldsymbol{z}^T)$, adota-se então, conforme Liu e Shen (2009), uma versão da densidade da distribuição de $\boldsymbol{V}$, sendo esta uma função de verossimilhança para uma observação $\boldsymbol{v}$, dada por

$$p_{\theta,F}(\boldsymbol{v}) = \sum_{j=1}^{k}\lambda_{k,j}\left[\exp\left(-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}}F(y_{k,j-1})\right) - \exp\left(-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}}F(y_{k,j})\right)\right] + \lambda_{k,k+1}\left[\exp\left(-e^{\boldsymbol{\theta'}\tilde{\boldsymbol{Z}}}F(y_{k,k})\right)\right]. \tag{A.15}$$

Assim, tendo introduzidas notações e construções anteriores, prova-se a consistência de Hellinger satisfazendo-se as seguintes condições:

C1 O parâmetro $\boldsymbol{\theta}$ é restrito ao conjunto compacto $\boldsymbol{\Theta}$ em $\mathbb{R}^{d+1}$, e função de distribuição basal F pertencente ao conjunto $\mathcal{B}$ de todas as funções de sub-distribuições em $(0, \infty)$.

C2 $E_0(||\boldsymbol{Z}||_2) < \infty$ para a norma euclidiana $||\,.\,||_2$ em $\mathbb{R}^p$. Para o parâmetro verdadeiro θ_0, $E(\exp(e^{|\theta_0'\tilde{\boldsymbol{Z}}|})) < \infty$.

C3 A verdadeira função parâmetro F_0 é monotonicamente crescente e satisfaz

$$E_0\left\{\frac{1}{\min_{j=1,\cdots,K}(F_0(Y_{K,j}) - F_0(Y_{K,j-1}))^2}\right\} < \infty.$$

Com isso, tem-se o necessário para enunciar o seguinte teorema:

Teorema A.1 *Sob as condições de regularidade C1-C3, as estimativas de máxima verossimilhança* $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ *e* $\hat{F}_n$ *satisfazem*

$$h(p_{\hat{\boldsymbol{\theta}}_{\boldsymbol{n}},\hat{F}_n}, p_{\boldsymbol{\theta_0},F_0}) \to_{a.s.} 0.$$

A.5 Distribuições preditivas do modelo semiparamétrico de Lam-Wong

Fazendo uso da notação apresentada na seção deste modelo e tomando f como sendo função densidade de probabilidade ou função de probabilidade, tem-se a distribuição preditiva de k_i dada por

$$f(k_i|y_i, \boldsymbol{x_i}, \delta_i) \propto f(y_i, \delta_i|\boldsymbol{x_i}, k_i) f(k_i|\boldsymbol{x_i}). \tag{A.16}$$

Para o primeiro termo,

$$f(y_i, \delta_i|\boldsymbol{x_i}, k_i) = \int f(y_i, \delta_i, u_i|\boldsymbol{x_i}, k_i) du_i = \int f(y_i, \delta_i|u_i, \boldsymbol{x_i}, k_i) f(u_i|k_i, \boldsymbol{x_i}) du_i.$$

Porém,

$$f(u_i|k_i, \boldsymbol{x_i}) = \left[\mathbb{1}_{\{0\}}(k_i)\mathbb{1}_{\{0\}}(u_i) + \mathbb{1}_{\mathbb{N}^+}(k_i)\mathbb{1}_{u_i>0} \frac{u_i^{k_i-1} e^{-u_i/2}}{2^{k_i}\Gamma(k_i)} \right].$$

Deste modo,

$$f(y_i, \delta_i|\boldsymbol{x_i}, k_i) = \int f(y_i, \delta_i|u_i, \boldsymbol{x_i}, k_i)\mathbb{1}_{\{0\}}(k_i)\mathbb{1}_{\{0\}}(u_i) du_i + \int f(y_i, \delta_i|u_i, \boldsymbol{x_i}, k_i)\mathbb{1}_{\mathbb{N}^+}(k_i)\mathbb{1}_{u_i>0} \frac{u_i^{k_i-1} e^{-u_i/2}}{2^{k_i}\Gamma(k_i)} du_i.$$

Pelo fato do integrando da primeira parcela assumir valor nulo para todos os pontos com exceção do ponto u_i, a primeira integral é nula, implicando em

$$f(y_i, \delta_i|\boldsymbol{x_i}, k_i) = \int f(y_i, \delta_i|u_i, \boldsymbol{x_i}, k_i) \frac{u_i^{k_i-1} e^{-u_i/2}}{2^{k_i}\Gamma(k_i)} du_i \mathbb{1}_{\mathbb{N}^+}(k_i)\mathbb{1}_{u_i>0},$$

ou seja,

$$\begin{aligned}
f(y_i, \delta_i|\boldsymbol{x_i}, k_i) &= \int [u_i \lambda_0(y_i) \exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i} [\exp(-u_i \Lambda_0(y_i) \exp(\boldsymbol{\beta' x_i^{(1)}}))] \frac{u_i^{k_i-1} e^{-u_i/2}}{2^{k_i}\Gamma(k_i)} du_i \mathbb{1}_{\mathbb{N}^+}(k_i)\mathbb{1}_{u_i>0} \\
&= \frac{\Gamma(k_i+\delta_i)(\lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))^{\delta_i}}{2^{k_i}\Gamma(k_i)[0,5 + \Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i+k_i}} \\
&\times \int \frac{u_i^{\delta_i+k_i-1} e^{-u_i(0,5+\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))} [0,5 + \Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i+k_i}}{\Gamma(k_i+\delta_i)} \mathbb{1}_{\mathbb{N}^+}(k_i)\mathbb{1}_{u_i>0} \\
&= \frac{\Gamma(k_i+\delta_i)(\lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))^{\delta_i}}{2^{k_i}\Gamma(k_i)[0,5 + \Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i+k_i}} \mathbb{1}_{\mathbb{N}^+}(k_i) \\
&= \frac{(k_i+\delta_i-1)!(\lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))^{\delta_i}}{2^{k_i}(k_i-1)![0,5 + \Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i+k_i}} \mathbb{1}_{\mathbb{N}^+}(k_i).
\end{aligned}$$

Logo,

$$f(y_i, \delta_i|\boldsymbol{x_i}, k_i) = \frac{(k_i+\delta_i-1)!(\lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))^{\delta_i}}{2^{k_i}(k_i-1)![0,5 + \Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i+k_i}} \mathbb{1}_{\mathbb{N}^+}(k_i). \tag{A.17}$$

Sabe-se também, por construção, que

$$f(k_i|\boldsymbol{x_i}) = \frac{(\frac{\exp((\theta x_i^{(0)}))}{2})^{k_i} e^{-\frac{\exp((\theta x_i^{(0)}))}{2}}}{k_i!}. \tag{A.18}$$

Relacionando (A.16) com (A.17) e (A.18), é fácil notar que

$$f(k_i|y_i, \boldsymbol{x_i}, \delta_i) \propto \frac{(\frac{\exp(\theta x_i^{(0)})}{2})^{k_i}}{k_i!} \times \frac{(k_i+\delta_i-1)!}{2^{k_i}(k_i-1)![0,5+\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i+k_i}} \mathbb{1}_{\mathbb{N}^+}(k_i).$$

Desenvolvendo a expressão acima, obtém-se

$$f(k_i|y_i, \boldsymbol{x_i}, \delta_i) \propto \frac{(\frac{\exp(\theta x_i^{(0)})}{4+2\Lambda_0(y_i)\exp(\beta' x_i^{(1)})})^{k_i}}{k_i!} \times \frac{(k_i+\delta_i-1)!}{(k_i-1)!} \mathbb{1}_{\mathbb{N}^+}(k_i).$$

Fazendo $c = \frac{\exp(\theta x_i^{(0)})}{4+2\Lambda_0(y_i)\exp(\beta' x_i^{(1)})}$, com alguma manipulação algébrica, tem-se

$$f(k_i|y_i, \boldsymbol{x_i}, \delta_i) \propto \frac{c^{k_i-\delta_i}(k_i+\delta_i-1)!e^{-c}}{k_i!(k_i-1)!} \mathbb{1}_{\mathbb{N}^+}(k_i).$$

Substituindo os possíveis valores de δ_i, nota-se que a expressão acima é análoga a

$$f(k_i|y_i, \boldsymbol{x_i}, \delta_i) \propto \frac{c^{k_i-\delta_i}e^{-c}}{(k_i-\delta_i)!} \mathbb{1}_{\mathbb{N}}(k_i-\delta_i).$$

Ou seja:

$$(K_i-\delta_i)|(y_i, \boldsymbol{x_i}, \delta_i) \sim \text{Poisson}\left(\frac{\exp(\boldsymbol{\theta' x_i^{(0)}})}{2+4\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})}\right).$$

Com k_i disponível, a distribuição preditiva de u_i dá-se através de

$$f(u_i|y_i, \delta_i, \boldsymbol{x_i}, k_i) \propto f(y_i, \delta_i|u_i, \boldsymbol{x_i}, k_i) f(u_i|k_i, \boldsymbol{x_i}). \tag{A.19}$$

Sabe-se que a função conjunta de y_i e δ_i é dada por

$$f(y_i, \delta_i|u_i, \boldsymbol{x_i}, k_i) = [u_i\lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})]^{\delta_i}[\exp(-u_i\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))] \propto u_i^{\delta_i} e^{-u_i\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})}. \tag{A.20}$$

Se $k_i = 0$, tem-se necessariamente que $u_i = 0$ por construção. Para $k_i > 0$ a distribuição condicional de u_i dado k_i é dada por

$$f(u_i|k_i, \boldsymbol{x_i}) = \frac{u_i^{k_1-1}e^{-u_i/2}}{2^{k_i}\Gamma(k_i)} \propto u_i^{k_i-1}e^{-u_i/2}. \tag{A.21}$$

De (A.19), (A.20) e (A.21), tem-se

$$f(u_i|y_i, \delta_i, \boldsymbol{x_i}, k_i) \propto u_i^{k_i+\delta_i-1} e^{-u_i(0,5+\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}}))}.$$

Portanto,

$$U_i|(y_i, \delta_i, \boldsymbol{x_i}, k_i) \begin{cases} = 0, & \text{se } k_i = 0; \\ \sim \text{Gama}(k_i+\delta_i, \{0,5+\Lambda_0(y_i)\exp(\boldsymbol{\beta' x_i^{(1)}})\}^{-1}), & \text{se } k_i > 0. \end{cases}$$

Apêndice B

Gráficos e Estimativas dos Parâmetros no Estudo de Simulação

B.1 Gráficos de viés e erro quadrático médio das simulações

As Figuras B.1, B.2, B.3 e B.4 exibem viés e erro quadrático médio obtidos para diferentes cenários simulados por meio do modelo de mistura padrão.

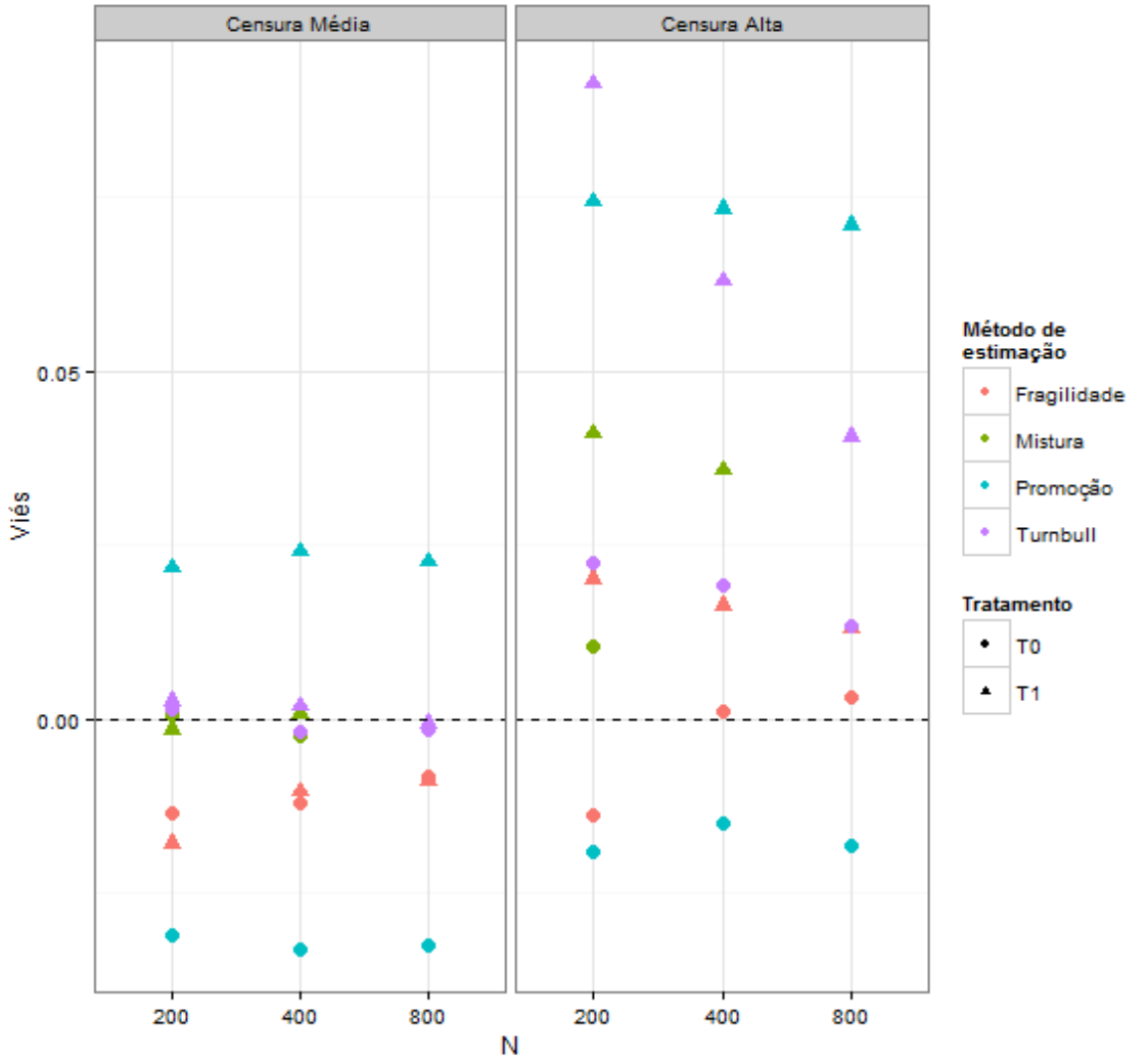

Figura B.1: *Viés para dados gerados pelo modelo de mistura padrão com frações de cura de T0 e T1 dadas por* 40% *e* 10%

Para conjuntos de dados gerados pelo modelo de tempo de promoção, tem-se o viés e o erro quadrático médio dos diferentes estimadores apresentados nas Figuras B.5, B.6, B.7 e B.8.

Os resultados associados aos dados gerados pelo modelo de fragilidade são ilustrados nas Figuras B.9, B.10, B.11 e B.12.

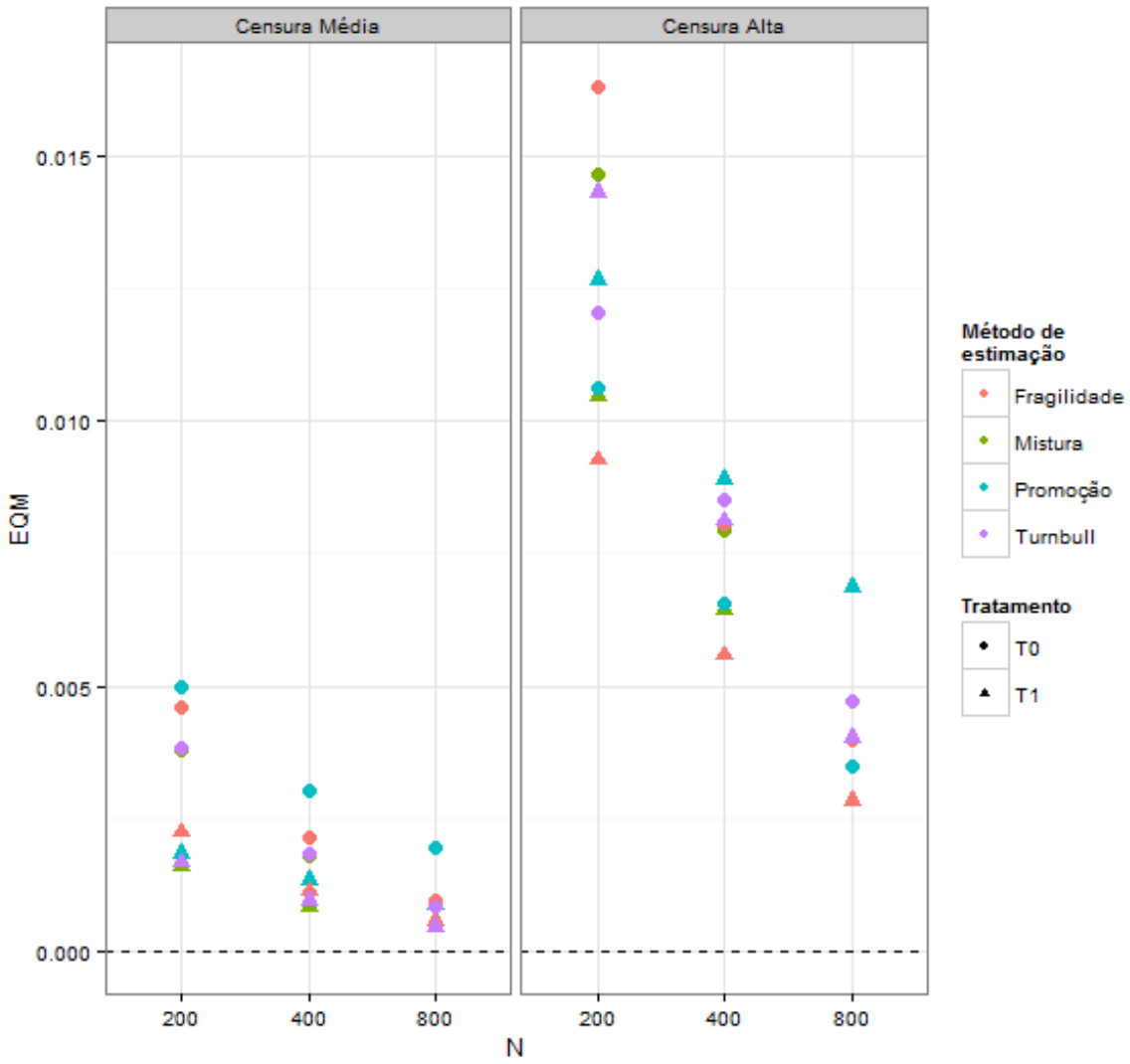

Figura B.2: *Erro quadrático médio para dados gerados pelo modelo de mistura padrão com frações de cura de T0 e T1 dadas por* 40% *e* 10%

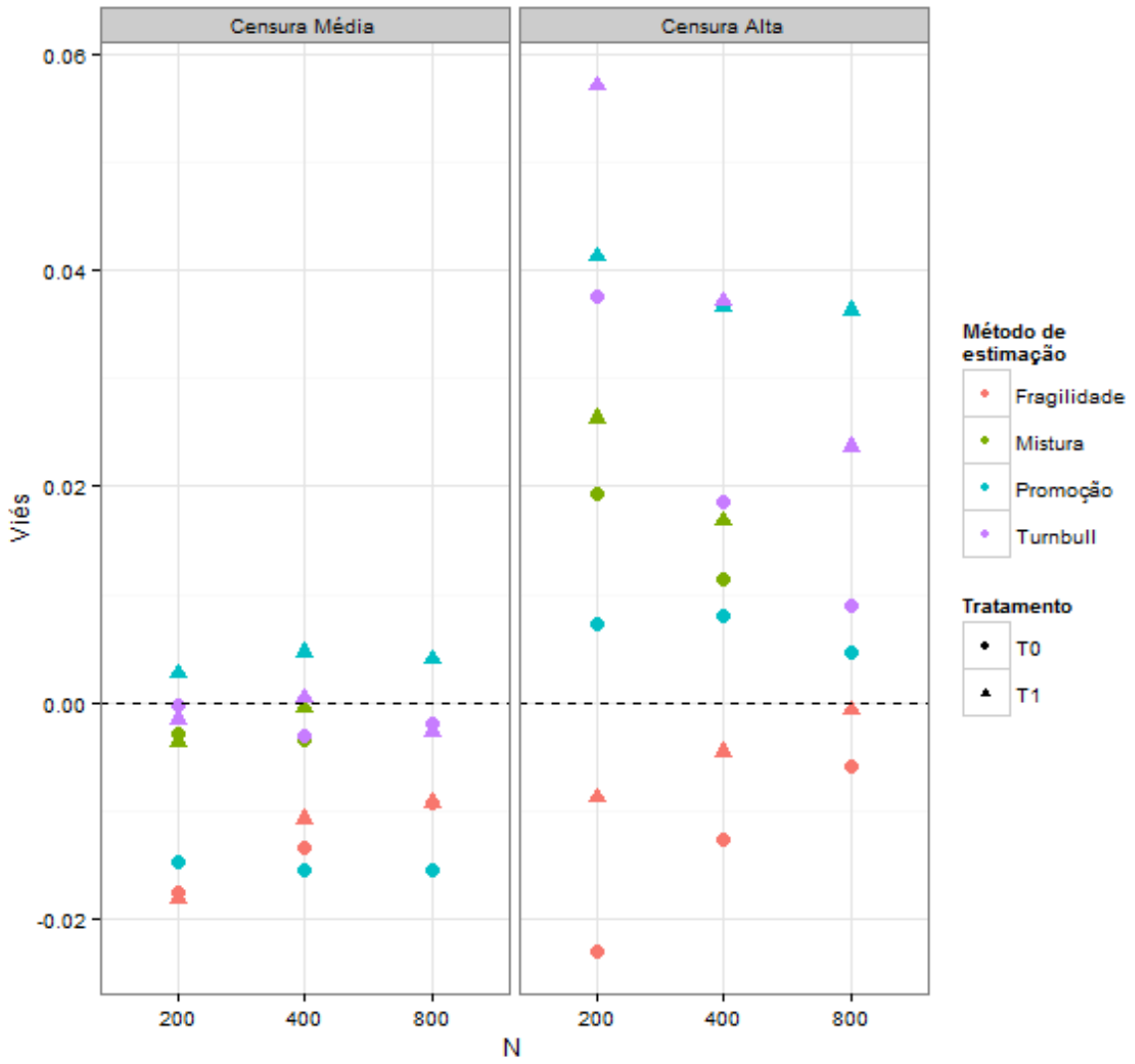

Figura B.3: *Viés para dados gerados pelo modelo de mistura padrão com frações de cura de T0 e T1 dadas por* 30% *e* 20%

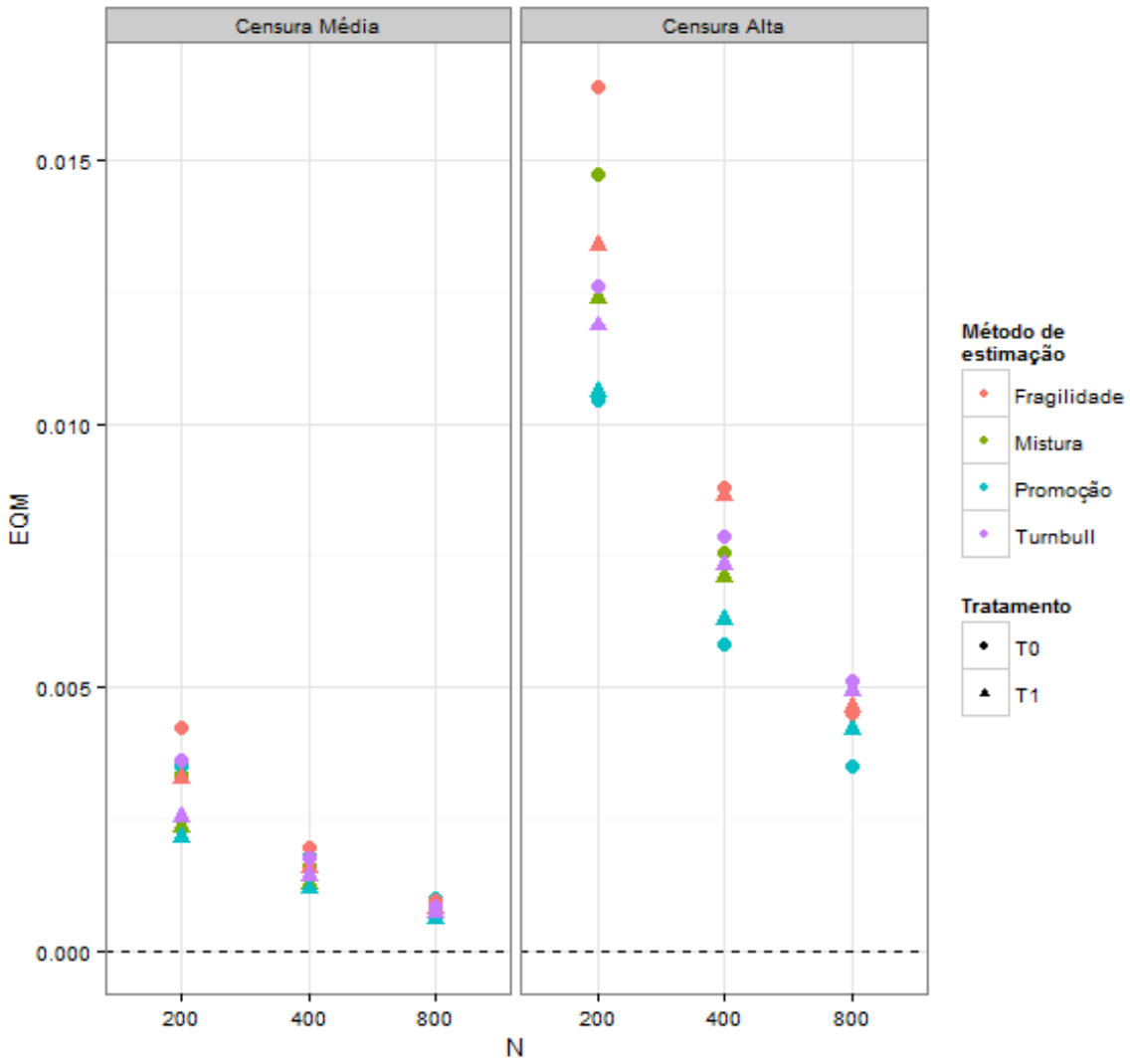

Figura B.4: *Erro quadrático médio para dados gerados pelo modelo de mistura padrão com frações de cura de T0 e T1 dadas por* 30% *e* 20%

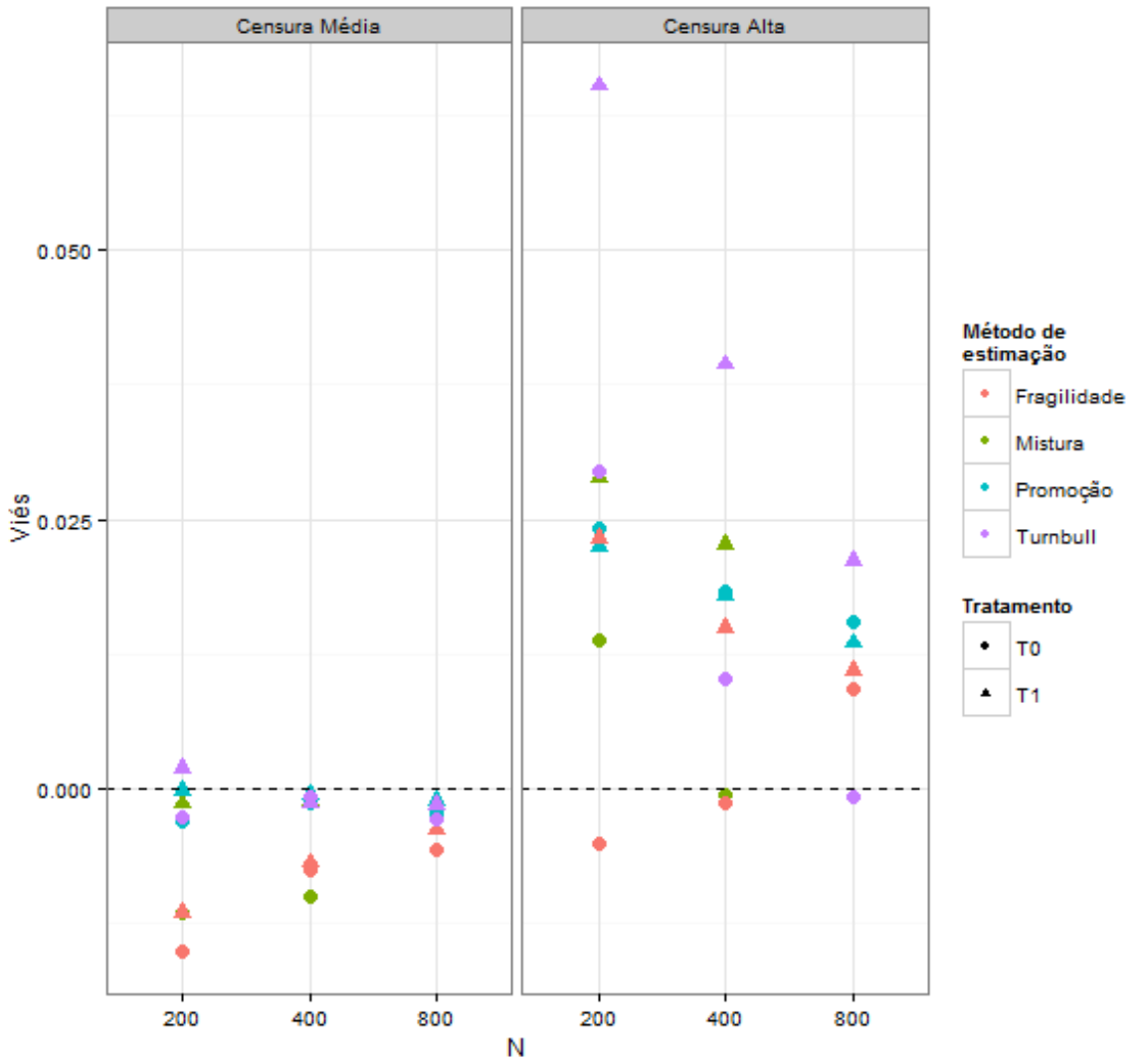

Figura B.5: *Viés para dados gerados pelo modelo de tempo de promoção com frações de cura de T0 e T1 dadas por* 40% *e* 10%

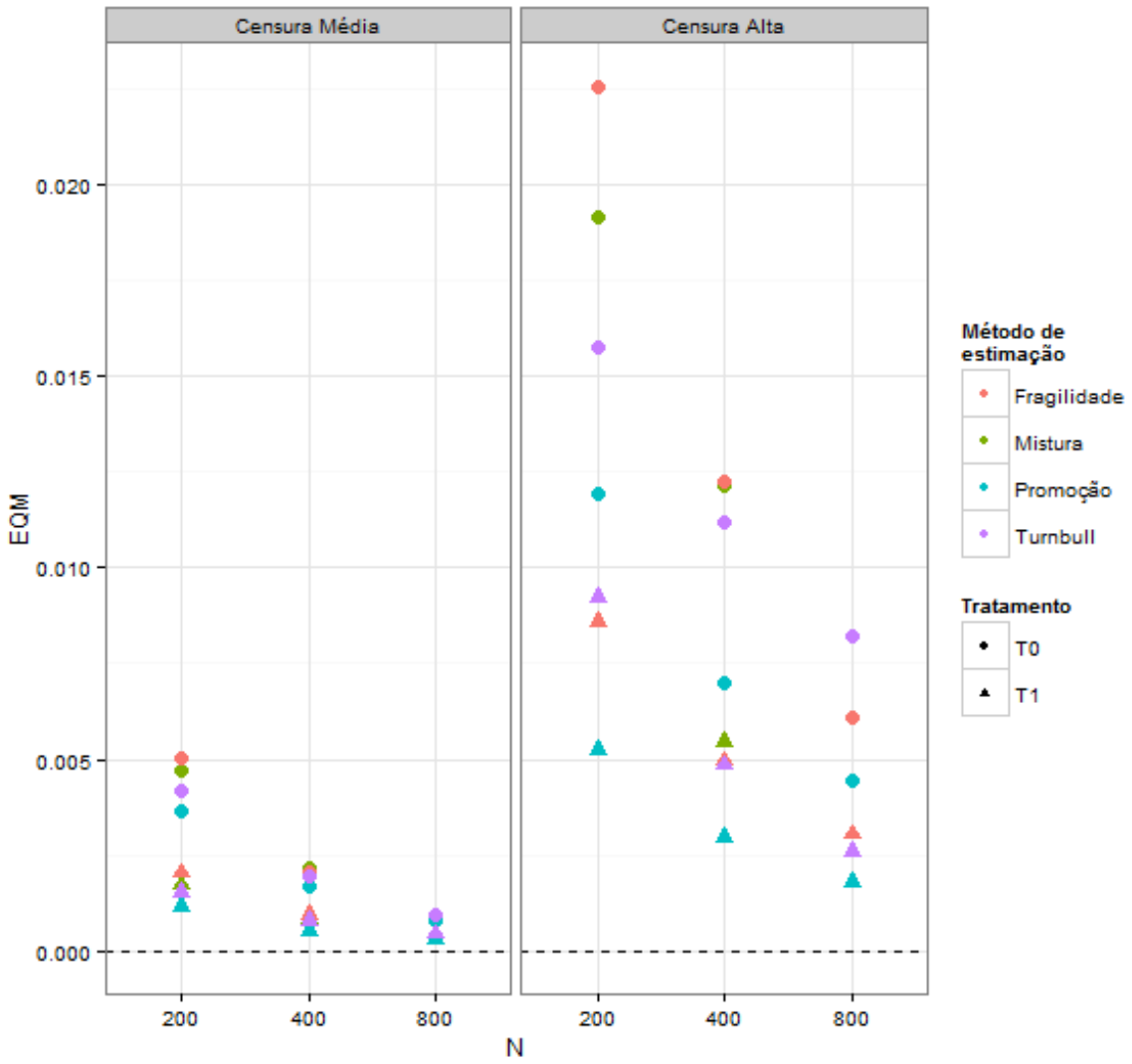

Figura B.6: *Erro quadrático médio para dados gerados pelo modelo de tempo de promoção com frações de cura de T0 e T1 dadas por* 40% *e* 10%

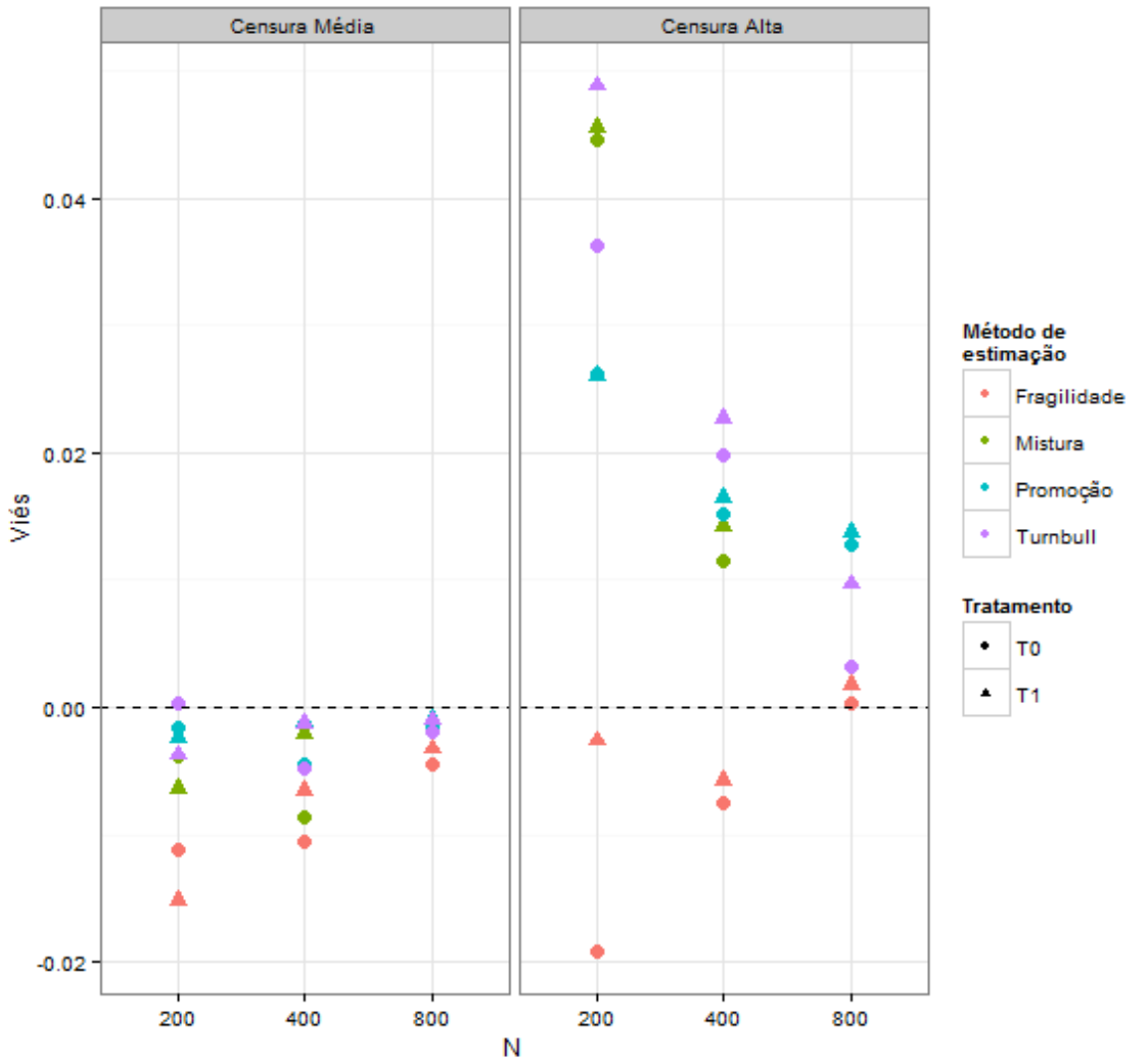

Figura B.7: *Viés para dados gerados pelo modelo de tempo de promoção com frações de cura de T0 e T1 dadas por* 30% *e* 20%

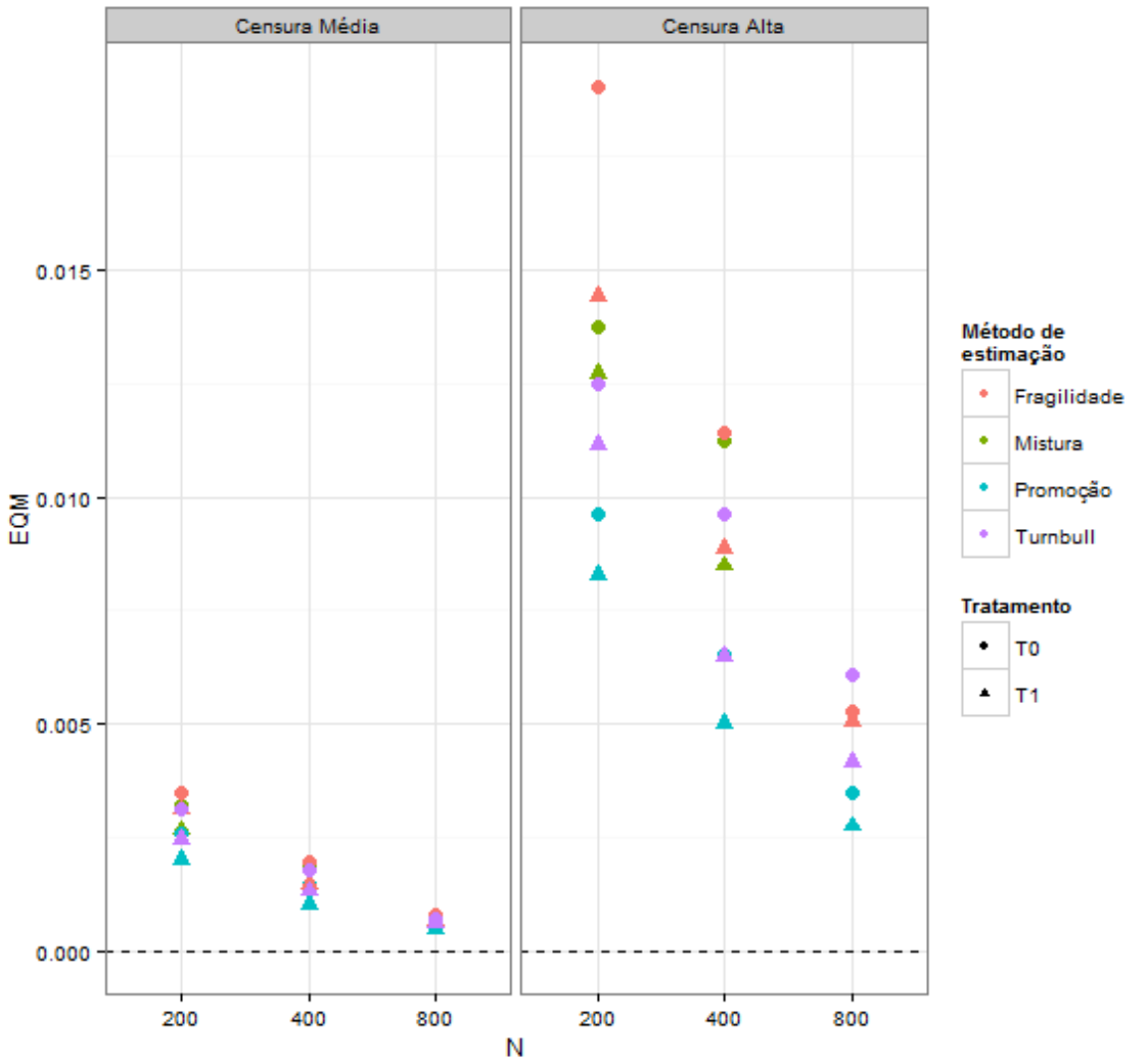

Figura B.8: *Erro quadrático médio para dados gerados pelo modelo de tempo de promoção com frações de cura de T0 e T1 dadas por* 30% *e* 20%

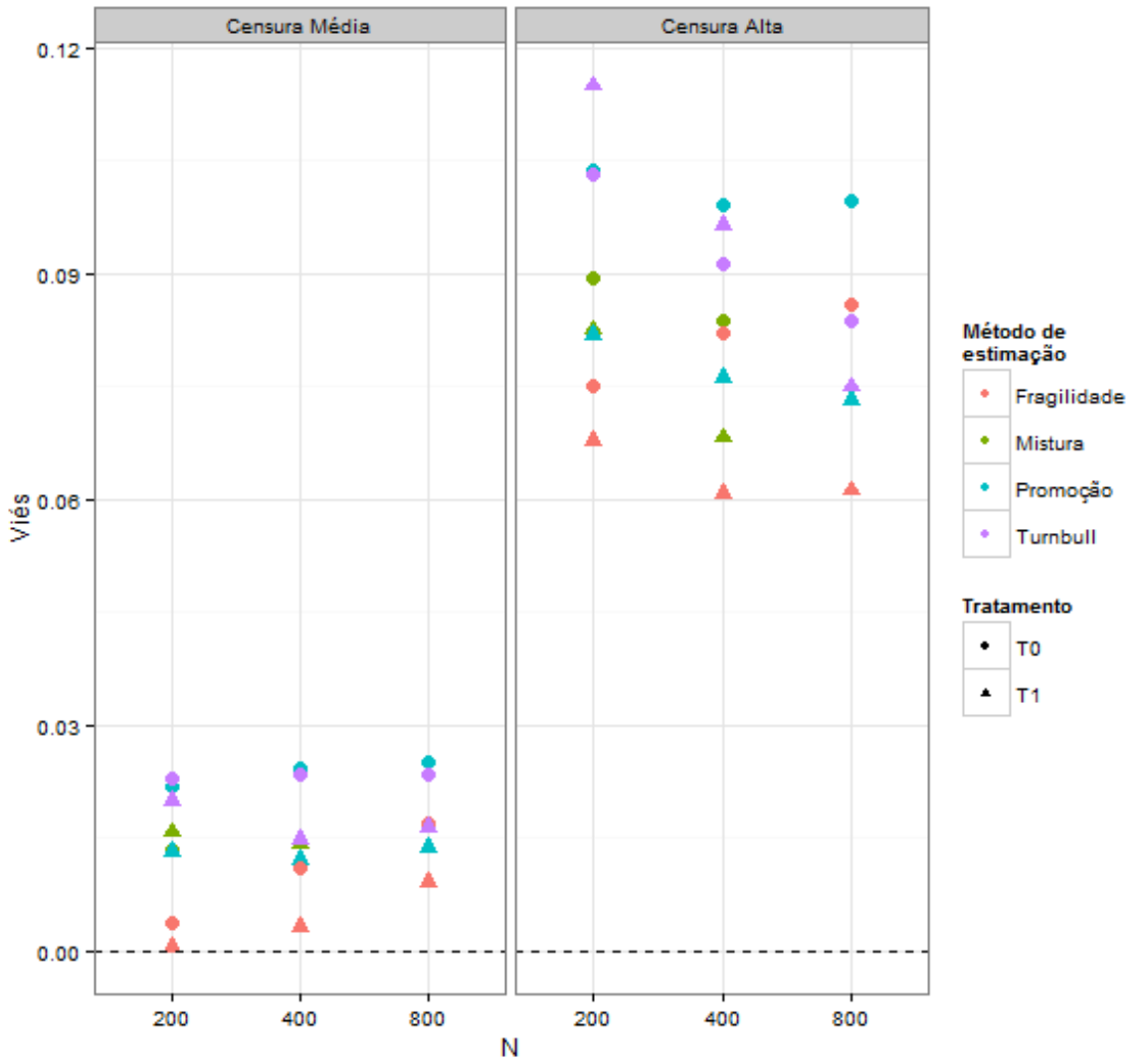

Figura B.9: *Viés para dados gerados pelo modelo de fragilidade com frações de cura de T0 e T1 dadas por* 40% *e* 10%

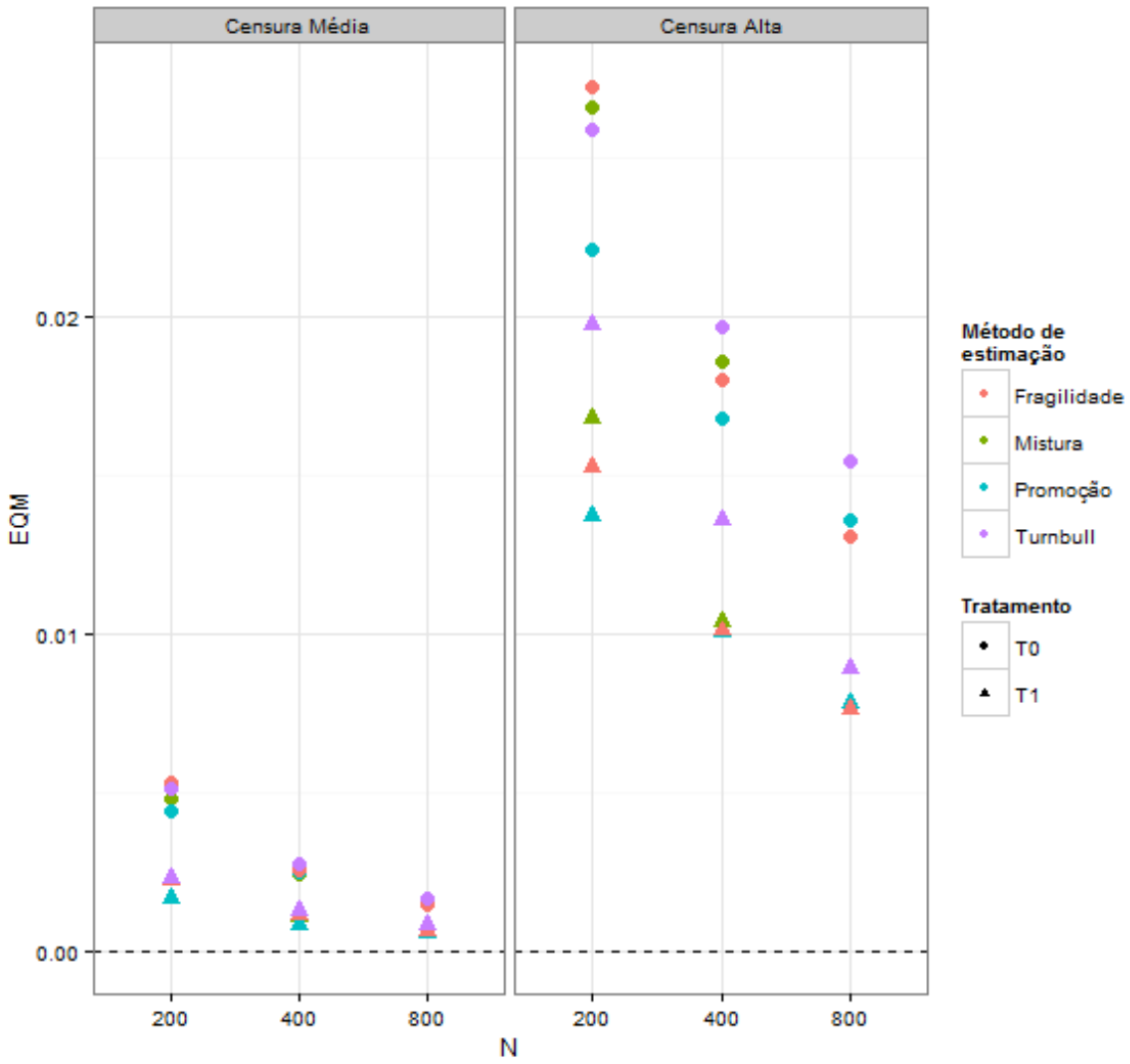

Figura B.10: *Erro quadrático médio para dados gerados pelo modelo de fragilidade com frações de cura de T0 e T1 dadas por* 40% *e* 10%

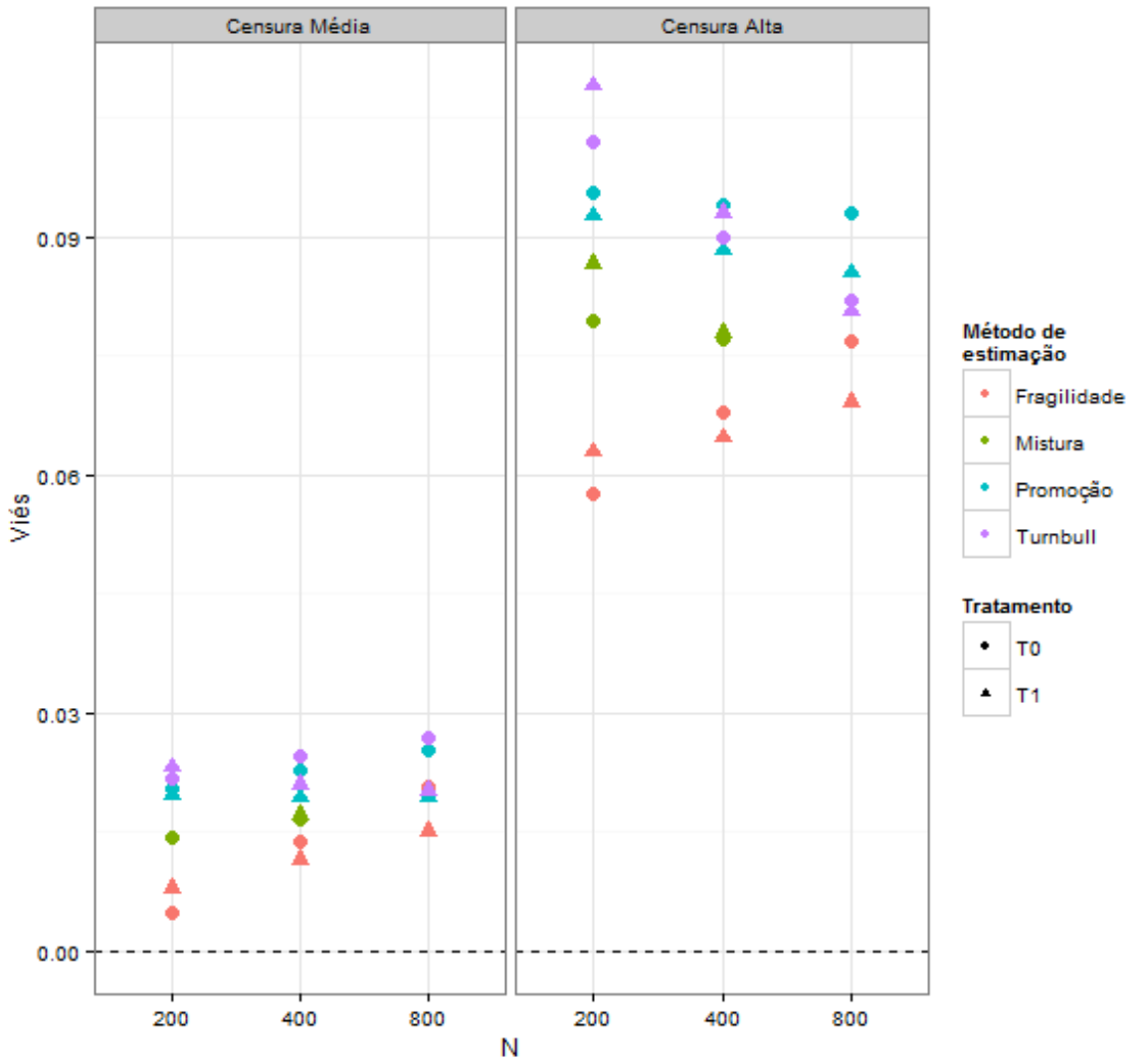

Figura B.11: *Viés para dados gerados pelo modelo de fragilidade com frações de cura de T0 e T1 dadas por* 30% *e* 20%

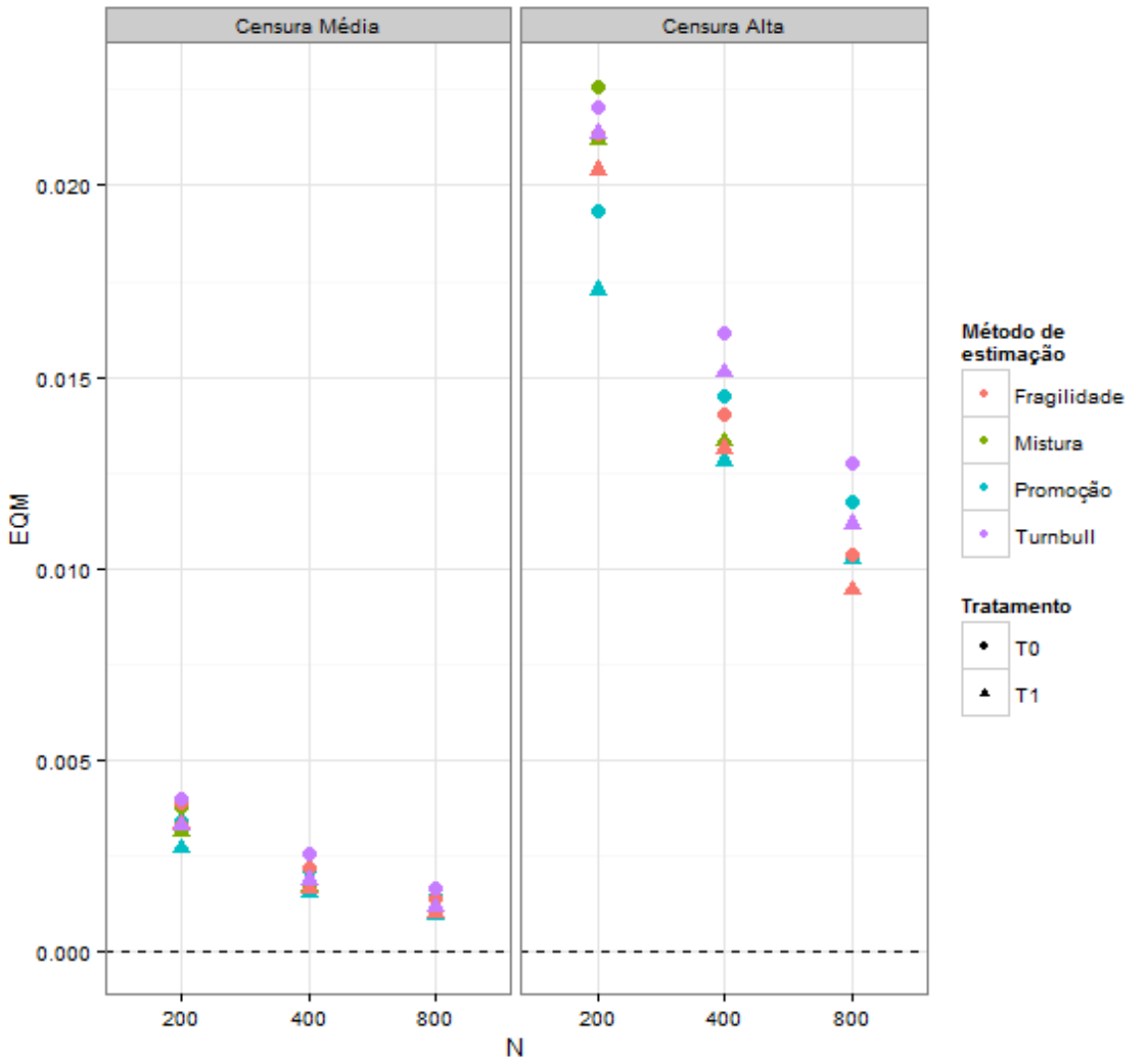

Figura B.12: *Erro quadrático médio para dados gerados pelo modelo de fragilidade com frações de cura de T0 e T1 dadas por* 30% *e* 20%

B.2 Efeitos estimados na simulação

Embora o foco dos estudos de simulações anteriormente realizados esteja na obtenção e análise de estimativas de fração de cura, esta seção apresenta, de maneira complementar ao leitor, os resultados de simulação das estimativas dos parâmetros de regressão obtidas para cada base, da qual derivam-se as frações de cura estimadas. As tabelas apresentadas nesta seção exibem apenas resultados dos parâmetros estimados quando o mecanismo de fração de cura do estimador é o mesmo que o utilizado para gerar o conjunto de dados. Por exemplo: para as bases geradas a partir do modelo de fragilidade, compara-se apenas as estimativas dos parâmetros obtidas pelo algoritmo proposto por Lam *et al.* (2013), pois para este sabemos os reais parâmetros por trás da geração dos dados.

B.2.1 Modelo de Mistura Padrão

As tabelas a seguir apresentam as estimativas obtidas para cada estimador aplicado aos dados gerados pela mesma especificação, tal que os parâmetros do preditor linear associados à fração de cura são dados por β_0 (intercepto) e β_1 (efeito associado à covariável binária). Uma segunda covariável assumindo distribuição normal de média nula tem efeito dado por $\beta_2 = 0$, com estimativas associadas a tal parâmetro também apresentadas nesta seção.

A Tabela B.1 com as métricas referentes ao modelo de mistura padrão para amostras de tamanho $n = 200$, assim como as posteriores a esta, apresenta em seu corpo: fração de cura real dos diferentes tratamentos, taxa de censura (média ou alta); parâmetro estimado; valor real do parâmetro a ser estimado; média das estimativas pontuais; viés médio das estimativas; erro quadrático médio das estimativas; desvio padrão empírico das estimativas; erro padrão médio das estimativas; e a probabilidade de cobertura estimada para intervalos construídos com 95% de confiança. A Tabela B.2 apresenta o mesmo conjunto de resultados para amostras de tamanho $n = 400$. Devido aos problemas quanto à identificabilidade com o aumento da dimensão do vetor a ser estimado, além do alto custo computacional, não foram obtidas estimativas utilizando o modelo de mistura padrão para amostras com $n = 800$.

Tabela B.1: *Métricas das estimativas de efeitos obtidas utilizando o modelo de mistura padrão para dados gerados pelo mesmo mecanismo* ($n = 200$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	0,41	0,41	0,00	0,07	0,26	0,21	0,89
(40%, 10%)	Censura Média	β_1	1,79	2,00	0,21	1,49	1,21	1,69	0,90
(40%, 10%)	Censura Média	β_2	0,00	0,00	0,00	0,06	0,24	0,18	0,87
(40%, 10%)	Censura Alta	β_0	0,41	0,47	0,06	1,72	1,31	3,37	0,60
(40%, 10%)	Censura Alta	β_1	1,79	2,23	0,44	7,52	2,71	7,98	0,61
(40%, 10%)	Censura Alta	β_2	0,00	0,24	0,24	0,67	0,78	0,20	0,54
(30%, 20%)	Censura Média	β_0	0,85	0,88	0,03	0,08	0,29	0,22	0,88
(30%, 20%)	Censura Média	β_1	0,54	0,56	0,02	0,17	0,41	0,34	0,91
(30%, 20%)	Censura Média	β_2	0,00	-0,01	-0,01	0,05	0,23	0,17	0,88
(30%, 20%)	Censura Alta	β_0	0,85	0,92	0,07	1,11	1,05	0,29	0,54
(30%, 20%)	Censura Alta	β_1	0,54	0,62	0,08	2,59	1,61	0,56	0,62
(30%, 20%)	Censura Alta	β_2	0,00	0,25	0,25	0,36	0,55	0,19	0,53

B.2.2 Modelo de Tempo de Promoção

Para o modelo de tempo de promoção, foram obtidos os resultados apresentados nas Tabelas B.3, B.4 e B.5. As mesmas quantidades de interesse apresentadas na subseção anterior são exibidas nas tabelas que se seguem.

B.2.3 Modelo de Fragilidade

Em Tabela B.6, observa-se as mesmas métricas para dados gerados pelo modelo de fragilidade com estimativas obtidas utilizando-se a mesma especificação. As Tabelas B.7 e B.8 exibem tais

Tabela B.2: *Métricas das estimativas de efeitos obtidas utilizando o modelo de mistura padrão para dados gerados pelo mesmo mecanismo* ($n = 400$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	0,41	0,42	0,01	0,03	0,18	0,15	0,89
(40%, 10%)	Censura Média	β_1	1,79	1,82	0,02	0,15	0,39	0,28	0,87
(40%, 10%)	Censura Média	β_2	0,00	0,01	0,01	0,03	0,16	0,13	0,88
(40%, 10%)	Censura Alta	β_0	0,41	0,34	-0,06	0,17	0,41	0,15	0,57
(40%, 10%)	Censura Alta	β_1	1,79	1,85	0,06	2,04	1,43	0,57	0,50
(40%, 10%)	Censura Alta	β_2	0,00	0,16	0,16	0,12	0,31	0,13	0,54
(30%, 20%)	Censura Média	β_0	0,85	0,87	0,02	0,04	0,20	0,16	0,88
(30%, 20%)	Censura Média	β_1	0,54	0,53	-0,00	0,09	0,30	0,24	0,89
(30%, 20%)	Censura Média	β_2	0,00	-0,00	-0,00	0,02	0,15	0,12	0,88
(30%, 20%)	Censura Alta	β_0	0,85	0,81	-0,04	0,22	0,47	0,16	0,53
(30%, 20%)	Censura Alta	β_1	0,54	0,57	0,03	0,67	0,82	0,26	0,56
(30%, 20%)	Censura Alta	β_2	0,00	0,15	0,15	0,16	0,37	0,12	0,47

Tabela B.3: *Métricas das estimativas de efeitos obtidas utilizando o modelo de tempo de promoção para dados gerados pelo mesmo mecanismo* ($n = 200$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	-0,09	-0,08	0,01	0,03	0,17	0,13	0,86
(40%, 10%)	Censura Média	β_1	0,92	0,93	0,01	0,03	0,18	0,17	0,93
(40%, 10%)	Censura Média	β_2	0,00	0,00	0,00	0,01	0,10	0,08	0,91
(40%, 10%)	Censura Alta	β_0	-0,09	-0,16	-0,07	0,10	0,30	0,13	0,58
(40%, 10%)	Censura Alta	β_1	0,92	0,95	0,03	0,07	0,26	0,17	0,82
(40%, 10%)	Censura Alta	β_2	0,00	0,00	0,00	0,02	0,13	0,08	0,81
(30%, 20%)	Censura Média	β_0	0,19	0,19	0,01	0,02	0,14	0,12	0,89
(30%, 20%)	Censura Média	β_1	0,29	0,30	0,01	0,03	0,17	0,16	0,93
(30%, 20%)	Censura Média	β_2	0,00	0,01	0,01	0,01	0,09	0,08	0,92
(30%, 20%)	Censura Alta	β_0	0,19	0,12	-0,07	0,08	0,27	0,12	0,59
(30%, 20%)	Censura Alta	β_1	0,29	0,30	0,01	0,06	0,25	0,17	0,82
(30%, 20%)	Censura Alta	β_2	0,00	-0,00	-0,00	0,02	0,13	0,09	0,81

resultados aumentando-se o tamanho da amostra para $n = 400$ e $n = 800$, respectivamente.

Tabela B.4: *Métricas das estimativas de efeitos obtidas utilizando o modelo de tempo de promoção para dados gerados pelo mesmo mecanismo* ($n = 400$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	-0,09	-0,08	0,00	0,01	0,11	0,09	0,89
(40%, 10%)	Censura Média	β_1	0,92	0,93	0,01	0,02	0,13	0,12	0,92
(40%, 10%)	Censura Média	β_2	0,00	-0,00	-0,00	0,00	0,07	0,06	0,91
(40%, 10%)	Censura Alta	β_0	-0,09	-0,14	-0,05	0,05	0,23	0,09	0,58
(40%, 10%)	Censura Alta	β_1	0,92	0,93	0,01	0,03	0,18	0,12	0,83
(40%, 10%)	Censura Alta	β_2	0,00	-0,00	-0,00	0,01	0,09	0,06	0,82
(30%, 20%)	Censura Média	β_0	0,19	0,20	0,01	0,01	0,11	0,08	0,88
(30%, 20%)	Censura Média	β_1	0,29	0,28	-0,01	0,02	0,13	0,12	0,92
(30%, 20%)	Censura Média	β_2	0,00	-0,00	-0,00	0,00	0,07	0,06	0,92
(30%, 20%)	Censura Alta	β_0	0,19	0,15	-0,04	0,05	0,23	0,09	0,52
(30%, 20%)	Censura Alta	β_1	0,29	0,29	0,00	0,03	0,17	0,12	0,84
(30%, 20%)	Censura Alta	β_2	0,00	0,00	0,00	0,01	0,08	0,06	0,83

Tabela B.5: *Métricas das estimativas de efeitos obtidas utilizando o modelo de tempo de promoção para dados gerados pelo mesmo mecanismo* ($n = 800$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	-0,09	-0,08	0,01	0,01	0,08	0,06	0,90
(40%, 10%)	Censura Média	β_1	0,92	0,92	0,00	0,01	0,09	0,08	0,93
(40%, 10%)	Censura Média	β_2	0,00	-0,00	-0,00	0,00	0,05	0,04	0,91
(40%, 10%)	Censura Alta	β_0	-0,09	-0,13	-0,04	0,03	0,18	0,07	0,51
(40%, 10%)	Censura Alta	β_1	0,92	0,93	0,00	0,02	0,13	0,08	0,80
(40%, 10%)	Censura Alta	β_2	0,00	0,00	0,00	0,00	0,06	0,04	0,82
(30%, 20%)	Censura Média	β_0	0,19	0,19	0,00	0,01	0,07	0,06	0,90
(30%, 20%)	Censura Média	β_1	0,29	0,29	-0,00	0,01	0,09	0,08	0,94
(30%, 20%)	Censura Média	β_2	0,00	0,00	0,00	0,00	0,04	0,04	0,93
(30%, 20%)	Censura Alta	β_0	0,19	0,15	-0,03	0,03	0,16	0,06	0,53
(30%, 20%)	Censura Alta	β_1	0,29	0,29	-0,00	0,01	0,12	0,08	0,82
(30%, 20%)	Censura Alta	β_2	0,00	-0,00	-0,00	0,00	0,06	0,04	0,85

Tabela B.6: *Métricas das estimativas de efeitos obtidas utilizando o modelo de fragilidade para dados gerados pelo mesmo mecanismo* ($n = 200$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	0,61	0,59	-0,01	0,04	0,20	0,15	0,85
(40%, 10%)	Censura Média	β_1	0,92	0,97	0,05	0,08	0,28	0,27	0,95
(40%, 10%)	Censura Média	β_2	0,00	0,01	0,01	0,02	0,14	0,13	0,93
(40%, 10%)	Censura Alta	β_0	0,61	0,38	-0,22	0,25	0,44	0,21	0,55
(40%, 10%)	Censura Alta	β_1	0,92	0,97	0,05	0,30	0,55	0,44	0,89
(40%, 10%)	Censura Alta	β_2	0,00	0,11	0,11	0,10	0,30	0,23	0,84
(30%, 20%)	Censura Média	β_0	0,88	0,87	-0,01	0,03	0,18	0,13	0,87
(30%, 20%)	Censura Média	β_1	0,29	0,29	-0,00	0,05	0,23	0,23	0,95
(30%, 20%)	Censura Média	β_2	0,00	0,02	0,02	0,02	0,12	0,12	0,93
(30%, 20%)	Censura Alta	β_0	0,88	0,73	-0,15	0,18	0,40	0,18	0,56
(30%, 20%)	Censura Alta	β_1	0,29	0,30	0,01	0,28	0,53	0,40	0,88
(30%, 20%)	Censura Alta	β_2	0,00	0,11	0,11	0,10	0,29	0,23	0,84

Tabela B.7: *Métricas das estimativas de efeitos obtidas utilizando o modelo de fragilidade para dados gerados pelo mesmo mecanismo* ($n = 400$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	0,61	0,57	-0,03	0,02	0,14	0,10	0,86
(40%, 10%)	Censura Média	β_1	0,92	0,95	0,03	0,04	0,18	0,19	0,95
(40%, 10%)	Censura Média	β_2	0,00	0,01	0,01	0,01	0,10	0,09	0,92
(40%, 10%)	Censura Alta	β_0	0,61	0,37	-0,24	0,15	0,31	0,15	0,47
(40%, 10%)	Censura Alta	β_1	0,92	0,97	0,05	0,14	0,37	0,33	0,92
(40%, 10%)	Censura Alta	β_2	0,00	0,11	0,11	0,05	0,20	0,17	0,82
(30%, 20%)	Censura Média	β_0	0,88	0,84	-0,04	0,02	0,12	0,09	0,86
(30%, 20%)	Censura Média	β_1	0,29	0,30	0,01	0,03	0,16	0,16	0,94
(30%, 20%)	Censura Média	β_2	0,00	0,02	0,02	0,01	0,08	0,08	0,93
(30%, 20%)	Censura Alta	β_0	0,88	0,69	-0,18	0,11	0,27	0,13	0,48
(30%, 20%)	Censura Alta	β_1	0,29	0,30	0,01	0,11	0,32	0,29	0,92
(30%, 20%)	Censura Alta	β_2	0,00	0,09	0,09	0,05	0,19	0,16	0,83

Tabela B.8: *Métricas das estimativas de efeitos obtidas utilizando o modelo de fragilidade para dados gerados pelo mesmo mecanismo* ($n = 800$)

Fração de cura	Censura	Parâmetro	Valor Real	Média	Viés Médio	EQM	D.P. Empírico	Erro Padrão Médio	P.C.
(40%, 10%)	Censura Média	β_0	0,61	0,56	-0,05	0,01	0,09	0,07	0,83
(40%, 10%)	Censura Média	β_1	0,92	0,94	0,01	0,02	0,12	0,13	0,96
(40%, 10%)	Censura Média	β_2	0,00	0,02	0,02	0,00	0,06	0,06	0,93
(40%, 10%)	Censura Alta	β_0	0,61	0,36	-0,25	0,11	0,22	0,10	0,38
(40%, 10%)	Censura Alta	β_1	0,92	0,96	0,04	0,07	0,25	0,25	0,94
(40%, 10%)	Censura Alta	β_2	0,00	0,11	0,11	0,03	0,14	0,12	0,77
(30%, 20%)	Censura Média	β_0	0,88	0,82	-0,06	0,01	0,08	0,07	0,81
(30%, 20%)	Censura Média	β_1	0,29	0,30	0,01	0,01	0,11	0,11	0,96
(30%, 20%)	Censura Média	β_2	0,00	0,03	0,03	0,00	0,05	0,06	0,91
(30%, 20%)	Censura Alta	β_0	0,88	0,67	-0,21	0,08	0,18	0,09	0,39
(30%, 20%)	Censura Alta	β_1	0,29	0,30	0,01	0,05	0,22	0,21	0,93
(30%, 20%)	Censura Alta	β_2	0,00	0,09	0,09	0,02	0,12	0,11	0,80

Referências Bibliográficas

Aalen (1992) O.O. Aalen. Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2:951–972. Citado na pág. 2, 3, 23, 24

Aljawadi *et al.* (2013) B. A. I. Aljawadi, M. R. A. Bakar, N. A. Ibrahim e M. Al-Omari. Parametric maximum likelihood estimation of cure fraction using interval-censored data. *Journal of Advanced Computing*, 1:43–58. Citado na pág. 4

Aljawadi *et al.* (2012a) Bader A. I. Aljawadi, Mohd Rizam A. Bakar e Noor Akma Ibrahim. Nonparametric versus parametric estimation of the cure fraction using interval censored data. *Communications in Statistics: Theory and Methods*, 41:4251–4275(25). Citado na pág. 4, 10

Aljawadi *et al.* (2012b) Bader A. I. Aljawadi, Mohd Rizam A. Bakar e Noor Akma Ibrahim. Turnbull versus Kaplan-Meier estimators of cure rate estimation using interval censored data. *Pertanika Journal of Science and Technology*, 20:243–255. Citado na pág. 8, 75

Almeida *et al.* (2013) F.N. Almeida, E.C. Sabino, G. Tunes, G.B. Schreiber, P.P.S.B. Silva, A.B.F. Carneiro-Proietti, J.E. Ferreira e A. Mendrone-Junior. Predictors of low haematocrit among repeat donors in São Paulo, Brazil: Eleven year longitudinal analysis. *Transfusion and Apheresis Science*, 49:553–559. Citado na pág. 2, 38

Almeida *et al.* (2016) F.N. Almeida, G. Tunes, J.C.B. Costa, E. Sabino, A. Mendrone-Júnior e J. Ferreira. A provenance model based on declarative specifications for intensive data analyses in hemotherapy information systems. *Future Generation Computer Systems*, 59:105–113. Citado na pág. 2

Banerjee e Carlin (2004) S. Banerjee e B. P. Carlin. Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics*, 60:268–275. Citado na pág. 5

Berkson e Gage (1952) J. Berkson e R.P. Gage. Survival curves for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515. Citado na pág. 3

Boag (1949) J.W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11:15–44. Citado na pág. 3

Boyd e Vandenberghe (2004) S.P. Boyd e L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK. Citado na pág. 19, 20, 80, 81, 82

Chen *et al.* (1999) M-H. Chen, J.G. Ibrahim e D. Sinha. A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94:909–919. Citado na pág. 2, 3, 15

Colosimo e Giolo (2006) E. Colosimo e S. Giolo. *Análise de Sobrevivência Aplicada*. Edgard Blücher, São Paulo. Citado na pág. 1, 4, 10

Conkin *et al.* (1992) J. Conkin, S. Bedahl e H. Van Liew. A computerized databank of decompression sickness incidence in altitude chambers. *Aviation, Space and Enviromental Medicine*, 63: 819–824. Citado na pág. 2

Cox (1972) D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34 (2):187–220. Citado na pág. 3

DelGiudice (1998) G. D. DelGiudice. Surplus killing of white-tailed deer by wolves in Northcentral Minnesota. *Journal of Mammalogy*, 79:227–235. Citado na pág. 44

DelGiudice *et al.* (2005) G. D. DelGiudice, B. A. Sampson, D. W. Kuehn, M. Carstensen Powell e J. Fieberg. Understanding margins of safe capture, chemical immobilization, and handling of free-ranging white-tailed deer. *Wildlife Society Bulletin*, 33:677–687. Citado na pág. 44

Dempster *et al.* (1977) A.P. Dempster, N.M. Laird e D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39 (1):1–38. Citado na pág. 4

Dorey *et al.* (1993) F.J. Dorey, R.J. Little e N. Schenker. Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, 12:1589–1603. Citado na pág. 7

Fay e Shaw (2010) Michael P. Fay e Pamela A. Shaw. Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. *Journal of Statistical Software*, 36:1–34. URL http://www.jstatsoft.org/v36/i02/. Citado na pág. 33

Fieberg e Conn (2014) J. Fieberg e P. Conn. A hidden markov model to identify and adjust for selection bias: an example involving mixed migration strategias. *Ecology and Evolution*, 4: 1903–1912. Citado na pág. 45

Fieberg e DelGiudice (2008) J. Fieberg e G. DelGiudice. Exploring migration data using interval-censored time-to-event models. *Journal of Wildlife Management*, 72:1211–1219. Citado na pág. 2, 5, 44, 45

Fieberg *et al.* (2008) J. Fieberg, G. DelGiudice e D. Kuehn. Understanding variation in autumn migration of northern white-tailed deer by long-term study. *Journal of Mammalogy*, 89:1529–1539. Citado na pág. 2, 44, 45

Finkelstein e Wolfe (1985) D.M. Finkelstein e R.A. Wolfe. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41:731–740. Citado na pág. 33

Gentleman e Geyer (1994) R. Gentleman e C.J. Geyer. Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81:618–623. Citado na pág. 3, 10

Goetghebeur e Ryan (2000) E. Goetghebeur e L. Ryan. Semiparametric regression analysis of interval-censored data. *Biometrics*, 56:1139–1144. Citado na pág. 4

Gómez *et al.* (2004) G. Gómez, M.L. Calle e R. Oller. Frequentist and Bayesian approaches for interval-censored data. *Statistical Papers*, 45:139–173. Citado na pág. 4

Hoggart e Griffin (2001) C. Hoggart e J.E. Griffin. A Bayesian partition model for customer attrition. *In: George, E. I. (ed.), Bayesian Method with Applications to Science, Policy, and Ofcial Statistics, Selected Papers from the ISBA 2000*, páginas 223–232. Citado na pág. 3

Hu e Xiang (2013) T. Hu e L. Xiang. Efficient estimation for semiparametric cure models with interval-censored data. *Journal of Multivariate Analysis*, 121:139–151. Citado na pág. 4

Ibrahim *et al.* (2001a) J.G. Ibrahim, M-H. Chen e D. Sinha. *Bayesian Survival Analysis*. Springer, New York. Citado na pág. 1, 3, 16

Ibrahim *et al.* (2001b) J.G. Ibrahim, M-H. Chen e D. Sinha. Bayesian semiparametric models for survival data with a cure fraction. *Biometrics*, 57:383–388. Citado na pág. 3

Kalbeisch e Prentice (2002) J.D. Kalbeisch e R.L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, Hoboken, NJ. Citado na pág. 18

Kaplan e Meier (1958) E.L. Kaplan e P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53 (282):457–481. Citado na pág. 7

Kim e Jhun (2008) Y. Kim e M. Jhun. Cure rate model with interval censored data. *Statistics in Medicine*, 27:3–14. Citado na pág. 3, 4

Klein e Moeschberger (2003) J. Klein e M. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, 2nd ed. Citado na pág. 1

Lam *et al.* (2010) K. F. Lam, Y. Xu e TL Cheung. A multiple imputation approach for clustered interval-censored survival data. *Statistics in Medicine*, 29:680–693. Citado na pág. 30

Lam e Wong (2014) K.F. Lam e K.Y. Wong. Semiparametric analysis of clustered interval-censored survival data with a cure fraction. *Computational Statistics and Data Analysis*, 79: 165–174. Citado na pág. 4, 5, 27, 28, 32, 33, 50

Lam *et al.* (2005) K.F. Lam, D.Y. Fong e O.Y. Tang. Estimating the proportion of cured patients in a censored sample. *Statistics in Medicine*, 24:1865–1879. Citado na pág. 2

Lam *et al.* (2013) K.F. Lam, K.Y. Wong e F. Zhou. A semiparametric cure model for interval-censored data. *Biometrical Journal*, 55:771–788. Citado na pág. iii, v, 2, 3, 4, 5, 22, 23, 24, 25, 26, 27, 28, 33, 41, 42, 43, 49, 50, 51, 55, 56, 63, 67, 72, 73, 100

Little e Rubin (2002) R.J.A. Little e D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ. Citado na pág. 22

Liu e Shen (2009) H. Liu e Y. Shen. A semiparametric regression cure model for interval-censored data. *Journal of the American Statistical Association*, 104:1168–1178. Citado na pág. iii, v, 2, 4, 5, 8, 15, 16, 17, 18, 19, 20, 21, 22, 33, 34, 41, 42, 43, 47, 80, 81, 82, 83

Ma (2010) S. Ma. Mixed case interval censored data with a cured subgroup. *Statistica Sinica*, 20: 1165–1181. Citado na pág. 3, 4

Ma e Li (2010) S. Ma e J. Li. Interval-censored data with repeated measurements and a cured subgroup. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59:693–705. Citado na pág. 4

McGilchrist (1993) C.A. McGilchrist. REML estimation for survival models with frailty. *Biometrics*, 49:221–225. Citado na pág. 14, 15, 79

McGilchrist e Aisbett (1991) C.A. McGilchrist e C.W. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47 (2):461–466. Citado na pág. 13

Meng e Rubin (1993) X. Meng e D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80:267–278. Citado na pág. 18, 22

Nettleton (1999) D. Nettleton. Convergence properties of the EM algorithm in constrained parameter spaces. *Canadian Journal of Statistics*, 27:639–648. Citado na pág. 22

Odell *et al.* (1992) P.M. Odell, K.M. Anderson e R.B. D'Agostinho. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48:951–959. Citado na pág. 7

Peng e Dear (2000) Y. Peng e K.B.G. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56:237–243. Citado na pág. 2

R Core Team (2015) R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/. Citado na pág. 10

Rücker e Messerer (1988) G. Rücker e D. Messerer. Remission duration: an example of interval-censored observation. *Statistics in Medicine*, 7:1139–1145. Citado na pág. 7

Rodrigues *et al.* (2008) J. Rodrigues, V. G. Cancho e M. de Castro. *Teoria Unificada de Análise de Sobrevivência*. Associação Brasileira de Estatística, São Paulo. Citado na pág. 3, 16

Rodrigues *et al.* (2010) J Rodrigues, V.G. Cancho, M. de Castro e N. Balakrishnan. A Bayesian destructive weighted poisson cure rate model and an application to a cutaneous melanoma data. *Statistical Methods in Medical Research*, 21:585–597. Citado na pág. 3

Rubin (1987) D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. Citado na pág. 26, 27, 32

Schenker e Gelfand (1988) N. Schenker e A.H. Gelfand. Asymptotic results for multiple imputation. *Annals of Statistics*, 16:1550–1566. Citado na pág. 27

Siegel (1979) A. Siegel. The noncentral chi-squared distribution with zero degrees of freedom and testing for uniformity. *Biometrika*, 66:381–386. Citado na pág. 23, 28

Sun (2006) J. Sun. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, New York. Citado na pág. 1, 4, 78

Tanner e Wong (1987a) M.A. Tanner e W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540. Citado na pág. 25, 26, 30

Tanner e Wong (1987b) M.A. Tanner e W.H. Wong. An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, 29:23–32. Citado na pág. 27

Taylor (1995) J.M.G. Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907. Citado na pág. 26

Thompson e Chhikara (2003) L.A. Thompson e R.S. Chhikara. A Bayesian cure rate model for repeated measurements and interval censoring, 2003. Citado na pág. 5

Tsodikov *et al.* (2003) A.D. Tsodikov, J.G. Ibrahim e A.Y. Yakovlev. Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98:1063–1078. Citado na pág. 3, 16

Turnbull (1976) B.W. Turnbull. The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38:290–295. Citado na pág. 2, 3, 5, 7, 8, 10, 16, 17, 19

van der Vaart e Wellner (2000) A.W. van der Vaart e J.A. Wellner. Preservation theorems for Glivenko-Cantelli and uniform Glivenko-Cantelli classes. *In: Gine, E.; M,D.; Wellner, J., editors. High Dimensional Probability II. Boston:Birkhäauser*, 27:113–132. Citado na pág. 83

van der Vaart e Wellner (1996) A.W. van der Vaart e J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York. Citado na pág. 83

Van Dyk *et al.* (1995) David A. Van Dyk, Xiao-Li Meng e Donald B. Rubin. Maximum likelihood estimation via the ECM algorithm: Computing the asymptotic variance. *Statistica Sinica*, 5:55–75. Citado na pág. 22

Wei e Tanner (1991) G.C.G. Wei e M.A. Tanner. Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, 47:1297–1309. Citado na pág. 25, 26

Xiang *et al.* (2011) L. Xiang, X. Ma e K.K.W. Yau. Mixture cure model with random effects for clustered interval-censored survival data. *Statistics in Medicine*, 30:995–1006. Citado na pág. iii, v, 2, 4, 5, 10, 15, 28, 33, 41, 45, 47, 49, 50, 56, 57, 60, 61, 78

Yakovlev *et al.* (1993) A.Y. Yakovlev, B. Asselain, V.J. Bardou, A. Fourquet, T. Hoang, A. Rochefediere e A.D. Tsodikov. A simple stochastic model of tumor recurrence and its applications to data on pre-menopausal breast cancer. *B. Asselain, M. Boniface, C. Duby, C. Lopez, J.P. Masson, J. Tranchefort (Eds.), Biometrie et Analyse de Donnees Spatio-Temporelles*, 12:66–82. Citado na pág. 3, 15

Yin e Ibrahim (2005) G. Yin e J.G. Ibrahim. Cure rate models: a unified approach. *The Canadian Journal of Statistics*, 33:559–570. Citado na pág. 3, 16

Yin e Nieto-Barajas (2009) G. Yin e L.E. Nieto-Barajas. Bayesian cure rate model accommodating multiplicative and additive covariates. *Statistics and Its Interface*, 2:513–521. Citado na pág. 3

Zeng *et al.* (2006) D. Zeng, G. Yin e J.G. Ibrahim. Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101:670–684. Citado na pág. 8, 18, 56

Zhou (2004) M. Zhou. Nonparametric Bayes estimator of survival functions for doubly/interval censored data. *Statistica Sinica*, 14:533–546. Citado na pág. 4