Modelos de fração de cura com fatores latentes competitivos e fragilidade

Renato de Azevedo Silva

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. Dr. Antonio Carlos Pedroso de Lima

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

São Paulo, Junho de 2011

Modelos de fração de cura com fatores latentes competitivos e fragilidade

Esta dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa realizada por Renato de Azevedo Silva em 06/05/2011. O original encontra-se disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Antonio Carlos Pedroso de Lima (orientador) - IME-USP
- Profa. Dra. Gisela Tunes da Silva - IME-USP
- Prof. Dr. Enrico Antônio Colosimo - UFMG

Agradecimentos

Aos meus pais, que me apoiaram em todos os momentos.

Ao Professor Antonio Carlos, pelas muitas as horas dedicadas ao meu aprendizado, pelas conversas, pela compreensão, pela paciência, por ter sido muito mais que um orientador, mas um amigo, um conselheiro, um "pai acadêmico";

Ao Enrico e à Gisela, pelas críticas e sugestões que contribuíram para o aperfeiçoamento do texto definitivo;

À Gabriela, pelo companheirismo e cumplicidade ao longos de todos estes anos. Por ter sido minha inspiração e minha motivação.

Ao Gregório e à Sueli, professores do departamento de Estatística da Universidade Federal de Minas Gerais, por terem despertado em mim o interesse pela pesquisa e pela educação. Ao Hamilton Prado Bueno e ao Marcelo Domingos Marchesin, do departamento de Matemática da UFMG, pelo empenho no despertar do meu senso crítico e por hoje, serem para mim excelentes exemplos de relação professor-aluno.

Aos amigos do Colégio Técnico da UFMG, João Vítor e Chicão, que me incentivaram a tomar a decisão que me levou à escolha da Estatística como área de estudo e pesquisa. Ao amigo Jorge Torrejón (Jorgito), pela atenção, pelas boas horas de papo e por ter me ajudado desde o meu primeiro dia no mestrado. Aos amigos do IBGE pelo incessante apoio no período mais complicado da elaboração deste texto.

À Magaly Marçula e ao Alfredo José Mansur, pesquisadores do Instituto do Coração do Hospital das Clínicas da Faculdade de Medicina da USP, por terem gentilmente cedido os dados utilizados neste trabalho.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior pelo auxílio financeiro durante o período do curso.

Resumo

Os modelos de riscos proporcionais são muito utilizados na análise do tempo de sobrevivência, porém, assumem implicitamente que, observado um conjunto de variáveis explicativas, a população em estudo seja homogênea e que os indivíduos permaneçam sob risco durante todo o período de observação ou até que apresentem o evento de interesse.

Tais suposições não são adequadas quando os indivíduos da população em estudo possuem diferentes pré-disposições ao surgimento de uma doença ou quando estão sujeitos à cura após o período de tratamento.

Esta dissertação discute o modelo de sobrevivência com fração de cura quando o evento de interesse é caracterizado por fatores latentes competitivos, enquanto a heterogeneidade não observada entre os riscos dos pacientes é modelada através de um fator aleatório denominado termo de fragilidade.

Palavras-chave: fração de cura, fragilidade, fatores latentes competitivos.

Abstract

Proportional hazards models are widely used in the analysis of survival time, however, it is implictly assumed that given a set of explanatory variables, the population under study is homogeneous and individuals remain at risk throughout the observation period time or until they have the event of interest.

However, such assumptions are not reasonable when individuals from the population under study have different pre-dispositions to the emergence of a disease or are subject to the cure after treatment.

This work discusses the cure fraction model when the event of interest is characterized by latent competitive factors, with patient risk modeled by a random factor called frailty.

Keywords: cure rate, frailty, competitive latent factors.

Sumário

Lista de Abreviaturas **vi**

Lista de Símbolos **vii**

Lista de Figuras **viii**

Lista de Tabelas **ix**

1 Introdução **1**

1.1 Revisão Bibliográfica 4

1.2 Estrutura deste Trabalho 7

2 Motivação do Estudo **8**

3 Modelos de Sobrevivência **10**

3.1 Conceitos Fundamentais na Modelagem de Dados de Sobrevivência 10

3.2 Modelos Clássicos de Riscos Proporcionais 13

3.3 Modelos de Sobrevivência com Fração de Cura 14

3.3.1 Modelo de mistura 15

3.3.2 Modelo de cura com fatores latentes competitivos 16

3.3.3 Curados e o Tempo de Acompanhamento 19

3.3.4 Identificabilidade 20

3.3.5 Inferência 23

4 Modelos de Fração de Cura e Fragilidade **24**

4.1 Descrição do modelo 24

4.2 Identificabilidade 27

4.3 Inferência 28

5 Simulações **35**

5.1 Geração de amostras 35

5.2 Resultados 40

5.3 Discussão 53

6 Aplicações **56**

6.1 Dados do InCor 56

6.1.1 Descrição do conjunto de dados 57

7 Considerações Finais **62**

7.1 Trabalhos Futuros 63

A Proposições **64**

A.1 Função de sobrevivência com fatores latentes competitivos e fragilidade 64

A.2 Equivalência com modelo de fatores latentes com distribuição Binomial Negativa . 65

B Seleção de covariáveis **67**

Referências Bibliográficas **74**

Lista de Abreviaturas

AE	Átrio esquerdo.
AIC	Critério de Informação de Akaike.
BIC	Critério de Informação de Bayes.
EQM	Erro quadrático médio.
ICC	Insuficiência cardíaca congestiva.
iid	Independentes e identicamente distribuídas.
InCor	Instituto do Coração da Faculdade de Medicina da Universidade de São Paulo.
RC	Razão de chances.
RR	Razão de riscos.
Var	Variância.
VE	Ventrículo esquerdo.

Lista de Símbolos

ϕ	Fração de cura, possivelmente dependente de covariáveis.
$\mathscr{H}$	Classe dos modelos de fração de cura de Yakovlev e Tsodikov.
$\mathscr{X}$	Espaço amostral de $\boldsymbol{X}$.
$\mathscr{F}$	Classe das funções de distribuição com suporte em $[0, \infty)$.
Θ	Classe das funções $\theta(\boldsymbol{x}) : 0 < \theta(\boldsymbol{x}) < \infty,\ \forall \boldsymbol{x} \in \mathscr{X}$.

Lista de Figuras

3.1 Funções de sobrevivência própria e imprópria 16

5.1 Estimadores dos coeficientes - γ conhecido 40

5.2 Estimadores dos parâmetros da distribuição dos tempos latentes - γ conhecido . . 41

5.3 Estimadores das frações de cura -γ conhecido 41

5.4 Estimadores dos coeficientes - γ estimado 44

5.5 Estimadores dos parâmetros da distribuição dos tempos latentes - γ estimado . . . 44

5.6 Função de probabilidade do "estimador" de γ 45

5.7 Estimadores das frações de cura -γ estimado 46

5.8 Estimadores dos coeficientes - γ estimado 49

5.9 Estimadores dos parâmetros da distribuição dos tempos latentes - γ estimado . . . 49

5.10 Função de probabilidade do "estimador" de γ 50

5.11 Estimadores das frações de cura -γ estimado 51

6.1 Função de sobrevivência de pacientes diagnosticados com Insuficiência Cardíaca Congestiva. 58

6.2 Funções de sobrevivência das covariáveis individualmente significativas. 61

Lista de Tabelas

5.1 Resultados das simulações para γ conhecido. 41

5.2 Resultados das simulações para γ estimado e fração de censura de 25%. 46

5.3 Resultados das simulações para γ estimado e fração de censura de 50%. 51

6.1 Valor da função de log-verossimilhança para diversos valores de γ. 59

6.2 Modelo de fatores latentes competitivos com termo de fragilidade $\gamma = 0,60$ 59

B.1 Procedimento de seleção de covariáveis - ajustes univariados. 67

B.2 Significância das covariáveis no modelo múltiplo. 71

B.3 Modelo de fatores latentes competitivos com $\gamma = 0,60$. 72

B.4 Modelo de fatores latentes competitivos de riscos proporcionais. 73

Capítulo 1

Introdução

A análise estatística do tempo de sobrevivência, ou simplesmente *análise de sobrevivência*, é o ramo da Estatística que trata da modelagem do tempo até a ocorrência de um dado evento de interesse, como por exemplo, o tempo decorrido desde o diagnóstico de câncer até o falecimento do paciente por esta causa.

O que diferencia a análise de sobrevivência de outras técnicas estatísticas é a variável de interesse, que é uma característica dinâmica, ou seja, que muda com o tempo, e que geralmente é composta por uma mistura de varíáveis aleatórias (inclusive de diferentes tipos: uma discreta a outra contínua), o que a atribui um caráter híbrido. Essas características por si só já requerem um conjunto de técnicas que considere essa particularidade.

Nos estudos em que a principal característica investigada é o tempo até a ocorrência de um evento, denominado *tempo de falha*, invariavelmente não é possível, ou não é de interesse do pesquisador, obter o momento exato das falhas de todas as unidades amostrais, uma vez que o experimento ou estudo observacional necessita ser encerrado para que os dados obtidos possam ser analisados ou mesmo, que a frequência de observação não permita determinar precisamente o momento exato da falha. Para estas unidades experimentais, muitas vezes, só o que se sabe é que ao final do experimento, a unidade ainda não havia apresentado o evento de interesse. Nos casos em que dispõe-se de informações parciais a respeito do tempo de acompanhamento, dizemos que ocorreu uma *censura* e o tempo observado é denominado *tempo de censura*. A situação descrita anteriormente representa apenas um dos vários tipos de censura, sendo importante destacar que tais informações parciais são a principal característica dos conjuntos de dados de sobrevivência.

Além disso, a maior parte das medições é feita condicionalmente ao que é conhecido no instante

de sua obtenção, e sabe-se que algumas delas podem sofrer alterações ao longo do tempo. De forma geral, a população em estudo muda constantemente e o que se busca modelar é a taxa de falha dos indivíduos que continuam vivos, o que exclui a possibilidade do uso da maioria das técnicas estatísticas conhecidas.

A *função de sobrevivência* é a forma mais popular de descrever o comportamento de dados como os que consideraremos neste trabalho. Em geral, é estimada pelo estimador proposto por Kaplan e Meier (1958) que surgiu como uma proposta para representar os dados observados sem a necessidade de fazer suposições sobre a distribuição do tempo de sobrevivência e ainda hoje é muito utilizado em algumas fases da modelagem, como por exemplo, na validação de um modelo proposto.

A função de risco, também denominda *função de taxa de falha*, é uma transformação biunívoca da função de sobrevivência que permite a comparação entre os riscos em diferentes situações mensuradas nas covariáveis. As funções de sobrevivência e de risco serão discutidas em maior profundidade no Capítulo 3.

Modelos baseados na função de risco têm obtido grande destaque na análise de sobrevivência desde a construção do modelo de riscos proporcionais de Cox (1972). Uma das razões desse modelo ter se tornado tão popular é o seu poder interpretativo e a facilidade com que as dificuldades técnicas oriundas das censuras foram contornadas. Naturalmente, a construção do modelo permite que as covariáveis sejam relacionadas com o risco de falha, de modo que influenciam no nível do risco para diferentes subgrupos.

Com o intuito de fixar alguns conceitos importantes para o entendimento do presente texto, consideremos um estudo de acompanhamento de pacientes diagnosticados com um determinado tipo de câncer. O tempo de observação se inicia no momento do diagnóstico, sendo o paciente então submetido ao tratamento específico mais adequado para o tipo e o estado de desenvolvimento do tumor. Características do paciente podem estar relacionadas com a progressão da doença e com a resposta ao tratamento. Idade, hábitos alimentares e consumo de álcool podem ser considerados pelo pesquisador como importantes fontes de informação para o prognóstico do tratamento, sendo usados como covariáveis no modelo de sobrevivência. O interesse principal nesses estudos é o tempo que o paciente sobreviverá desde o diagnóstico da doença.

Durante o período de acompanhamento, um paciente pode ir à óbito por causas alheias à doença específica considerada no estudo, o que leva ao pesquisador a informação de que, até a data do óbito, tal paciente havia sobrevivido ao câncer, mas sem ser possível determinar o tempo exato desde o diagnóstico até a morte causada pelo câncer. Este caso é tratado como um tempo

censurado, ou simplesmente uma *censura*.

Com o avanço tecnológico ocorrido nos últimos anos levando à diagnósticos mais precisos e precoces, além do aumento da eficiência dos tratamentos, a chance de cura em alguns casos é considerada elevada. Consequentemente, além do objetivo de estimar o tempo de sobrevivência das pessoas acometidas pela doença, outro interesse é justamente estimar a probabilidade do indivíduo deixar de apresentar sintomas ou sinais da doença, o que também é conhecido como *fração de cura*.

Métodos comumente utilizados em análise de sobrevivência assumem implicitamente que, dado um conjunto de covariáveis, as populações são homogêneas, o que significa que indivíduos com os mesmos valores das covariáveis possuem a mesma probabilidade de morte e que permanecem sob risco até o momento em que falham ou são censurados. Conforme mencionado anteriormente, em muitos casos é importante considerar as populações heterogêneas, isto é, uma mistura de indivíduos que, apesar de terem os mesmos valores para as covariáveis, apresentam diferentes riscos, como sugere Wienke (2007).

A suscetibilidade à morte, também conhecida como heterogeneidade oculta ou *fragilidade*, é um dos principais objetos de estudo na análise de sobrevivência demográfica, onde a influência de variações entre indivíduos na probabilidade de sobrevivência não pode ser ignorada. Os estudos envolvendo indivíduos relacionados, em que variação genética, herdabilidade e outras propriedades da suscetibilidade à ocorrência de um dado evento de interesse devem ser levadas em consideração, geralmente podem ser analisados usando modelos de fragilidade. Tais modelos consideram a existência de fatores internos ou externos (por exemplo, similaridade genética e fatores ambientais), que influenciam diretamente na probabilidade de sobrevivência dos indivíduos (Yashin *et al.*, 2001).

A noção de fragilidade é uma forma conveniente de considerar o efeito da heterogeneidade e de associações não observadas nos modelos de sobrevivência. Na forma mais simples, trata-se de um fator não observado de proporcionalidade que modifica a função de risco de um indivíduo ou de indivíduos relacionados. Aplicações em problemas de análise de sobrevivência com dados multivariados podem ser encontrados em Clayton (1978).

Embora o presente texto se utilize de uma situação de desenvolvimento de um tumor para destacar os fundamentos da elaboração do modelo discutido, as possibilidades de aplicação não estão restritas à esses casos. O próprio exemplo apresentado no Capítulo 2 e melhor discutido no Capítulo 6 é um caso em que, embora intuitivamente a situação não satisfaça todas as suposições feitas na elaboração do modelo, foi considerado com a finalidade de ilustrar a aplicação do modelo

discutido no texto.

1.1 Revisão Bibliográfica

Modelos de sobrevivência com fração de cura têm sido amplamente usados na análise de dados de câncer. Um modelo para a estimação da fração de cura foi apresentado por Boag (1949), posteriormente formalizado por Berkson e Gage (1952), o qual é denominado modelo de mistura ou modelo padrão.

Os modelos de sobrevivência com fração de cura têm como principais objetivos estimar a proporção de pacientes curados em um estudo clínico e, para aqueles não curados, a probabilidade de sobrevivência em um dado instante de tempo. O modelo de mistura é assim conhecido porque considera que a população é composta de uma mistura de indivíduos curados e suscetíveis. No trabalho de Boag (1949), o tempo de sobrevivência dos indivíduos suscetíveis à doença foi considerado como tendo uma distribuição log-normal; porém, outras distribuições podem ser consideradas ou mesmo, não especificarmos uma forma paramétrica para a distribuição do tempo.

Maller e Zhou (1996) fazem uma vasta discussão dos modelos de sobrevivência com fração de cura, discutindo desde a interpretação intuitiva do modelo até técnicas utilizadas para contornar possíveis dificuldades que possam surgir em algumas aplicações.

Embora intuitivamente atraente, o modelo de mistura possui alguns inconvenientes como destacado por Chen *et al.* (1999). O fato do modelo não apresentar a característica de riscos proporcionais quando as covariáveis são relacionadas ao termo que determina a fração de cura, a dificuldade de estimar os parâmetros usando métodos iterativos de maximização da função de verossimilhança devido às instabilidades computacionais e a falta de uma extensão multivariada do modelo são alguns desses inconvenientes.

Yakovlev e Tsodikov (1996) desenvolveram um modelo alternativo, denominado modelo de risco acumulado limitado, o qual pode ser chamado também de modelo de riscos proporcionais com fração de cura. Nessa proposta, considera-se que um paciente possui um determinado número de células com potencial para desenvolver um tumor, porém essa quantidade não é observável. Neste caso, o evento de interesse ocorrerá quando um tumor for detectado. Obviamente, um paciente pode não possuir tais células ou ainda, não apresentá-las após ser submetido a um tratamento específico. Dizemos que este paciente está curado, pois, mesmo que seja acompanhado por um longo período de tempo, não apresentará o desenvolvimento do tumor.

Diversos autores trabalharam com esse modelo, considerando que o número de células com o

potencial de desenvolver um tumor - cada uma dessas células será denominada um *fator latente* - é uma variável aleatória não observável, com distribuição de Poisson.

Sumathi e Aruna Rao (2008) fazem uma revisão bibliográfica dos modelos de fração de cura, incluindo o uso de cópulas nos casos multivariados ou recorrentes, enquanto Barajas e Yin (2008) discutem o modelo semiparamétrico com fração de cura usando o enfoque bayesiano, considerando também o caso em que o limiar de cura não é conhecido. Mizoi *et al.* (2007) discutem a influência local em modelos de sobrevivência com fração de cura quando uma das covariáveis apresenta erro de medição.

Kim *et al.* (2007) apresentam uma abordagem bayesiana ao modelo dinâmico de sobrevivência com fração de cura, no qual a função de risco é aproximada por uma função particionada (*piecewise constant*). No modelo discutido, as covariáveis são fixas, bem como seus efeitos, e são consideradas importantes para a determinação da fração de curados. É considerada a estrutura de riscos competitivos, sendo os fatores latentes independentes e identicamente distribuídos, com distribuição de Poisson.

No modelo tradicional, o evento de interesse é caracterizado pelo mínimo dos tempos latentes. No entanto, no modelo discutido por Cooner *et al.* (2006) considera-se a situação em que o indivíduo resiste à ocorrência de $k-1$ eventos latentes, mas se ocorrer o k-ésimo, então o evento de interesse ocorrerá. Os autores do estudo apresentam uma argumentação no contexto biológico para a possível utilização da distribuição geométrica do número de fatores latentes. Um efeito aleatório espacial é incorporado ao modelo através do preditor linear ou do parâmetro da distribuição do número de fatores latentes.

Bakar *et al.* (2008) discutem os modelos de Berkson e Gage e o de Yakovlev e Tsodikov sob a ótica da sobrevivência relativa, considerando um grupo de pacientes como curado quando o risco decai para o nível daquele estimado para a população geral. Além disso, sugerem a possibilidade da introdução de uma forma não-paramétrica na distribuição dos tempos, o que pode ser observado na proposta de Yu (2008), em que se discute o modelo de fração de cura sem especificar uma distribuição para a tempo de sobrevivência, bem como a introdução de um termo de fragilidade gama no contexto de eventos recorrentes. A estimação dos parâmetros é feita usando o algoritmo EM e os erros padrão são estimados por *bootstrap*.

De Castro *et al.* (2010a) apresentam uma abordagem bayesiana para um modelo de sobrevivência com fração de cura, também no contexto de riscos competitivos, porém considerando que os fatores latentes independentes e identicamente distribuídos possuem uma distribuição binomial negativa. A vantagem é que o modelo passa a acomodar uma possível superdispersão, enquanto

Rodrigues *et al.* (2009) discutem uma abordagem assintótica para o mesmo modelo. No entanto, não discutem os indícios de problemas de identificabilidade do modelo proposto.

Mais informações sobre as condições de identificabilidade do modelo de mistura podem ser obtidas em Peng e Zhang (2008), enquanto Li *et al.* (2001) discutem a identificabilidade do modelo de Yakovlev e Tsodikov.

Chang *et al.* (2007) apresentam resultados assintóticos para a estimação do modelo semiparamétrico de mistura, usado em dados de sobrevivência obtidos sob a suposição de riscos competitivos, enquanto Zeng *et al.* (2006) fazem uma proposta de relaxar a suposição de que os tempos latentes sejam *iid*, inserindo no modelo um termo de fragilidade com distribuição gama. As condições de identificabilidade do modelo nos casos em que o parâmetro da distribuição do termo de fragilidade não é especificado são discutidas e uma forma de estimar os parâmetros nas condições de não-identificabilidade do modelo é apresentada. A proposta é estimar os coeficientes para diversos valores do parâmetro da distribuição postulada para a fragilidade e escolher aquele que apresentar maior critério de informação de Akaike (AIC).

Abordagem semelhante pode ser encontrada em Chen *et al.* (2002), em que os tempos latentes competitivos são tratados de forma multivariada, sendo que a correlação entre tais tempos é induzida por um fator de fragilidade. Neste caso, a estimação é feita usando o enfoque bayesiano.

Yin e Ibrahim (2005) apresentam uma classe de modelos com uma transformação de Box-Cox na função de risco e considera uma distribuição exponencial particionada para a função de sobrevivência. Como utilizam uma abordagem bayesiana, discutem as escolhas mais adequadas para as distribuições *a priori*.

Yu e Peng (2008) abordam um modelo de mistura para dados multivariados, postulando um modelo marginal e tratando a estrutura de correlação como um parâmetro de perturbação. A proposta é considerar uma distribuição Weibull para os tempos e relacionar as covariáveis com a fração de cura. Os estimadores de máxima verossimilhança são obtidos da verossimilhança marginal, enquanto as correlações entre os indivíduos não são estimadas. No ajuste do modelo, utilizam o programa implementado por Peng (2003) para estimar consistentemente os coeficientes da regressão e, para a variância, usam o método *jacknife*, assintoticamente equivalente ao estimador sanduíche, comumente utilizado neste caso.

1.2 Estrutura deste Trabalho

Esta dissertação apresenta uma discussão dos modelos de fração de cura considerando a estrutura de fatores latentes competitivos. Inicialmente assume-se que o número de fatores latentes apresenta distribuição de Poisson e, posteriormente, considera um termo de fragilidade que, se tiver distribuição gama, será equivalente ao modelo apresentado por De Castro *et al.* (2010b). Também é apresentado um maior detalhamento do modelo discutido em Zeng *et al.* (2006), que pode ser considerado uma extensão do modelo de Yakovlev e Tsodikov, em que a distribuição condicional dos tempos de sobrevivência não é completamente especificada, sendo, portanto, um modelo semiparamétrico com fração de cura.

Este trabalho está organizado da seguinte forma: no Capítulo 2 está a motivação para o uso do modelo de sobrevivência com fração de cura considerado no texto, enquanto que no capítulo seguinte é feita uma apresentação dos conceitos essenciais para o entendimento dos modelos apresentados posteriormente. O Capítulo 4 contém uma discussão detalhada do modelo de sobrevivência com fração de cura e fragilidade, que deve ser visto como o objeto principal desta dissertação. No Capítulo 5 encontram-se os resultados de simulações feitas a fim de validar a implementação dos métodos de estimação e a avaliação da adequação dos modelos discutidos. Os Capítulos 6 e 7 apresentam um exemplo de utilização do modelo abordado no trabalho e as considerações finais, respectivamente. Algumas demonstrações necessárias ao entendimento de resultados utilizados foram colocadas no Anexo A com o intuito de simplificar a compreensão do texto.

Capítulo 2

Motivação do Estudo

Muitas pesquisas têm sido desenvolvidas a fim de fornecer mais informações a respeito dos mecanismos capazes de provocar o surgimento e o agravamento de doenças, o que subsidia formas alternativas cada vez mais complexas e eficientes de tratamento. Em alguns casos, o tratamento disponibilizado garante ao paciente uma sobrevida muito além da imaginada há poucos anos atrás, com um eficiente controle dos sintomas causados pela doença, levando-o até mesmo à cura.

Com isso, cresce também a quantidade de estudos de avaliação e comparação de métodos de tratamento, gerando uma grande quantidade de dados com características cada vez mais peculiares. A demanda por técnicas estatísticas que permitam uma adequada análise desses dados é uma realidade cada vez mais presente nos dias atuais.

Estudos de sobrevida de pacientes diagnosticados com câncer têm sido tão frequentes quanto as propostas de novos tratamentos, muitos dos quais possibilitam inclusive uma considerável chance de cura para alguns tipos de câncer.

Outra doença que demanda estudos por sua importância em termos de saúde pública é a insuficiência cardíaca congestiva (ICC) caracterizada pela diminuição da eficiência na irrigação sanguínea, o que resulta em um fornecimento insuficiente de oxigênio e nutrientes aos tecidos periféricos.

Associados à ICC estão diversos fatores de risco, tais como acúmulo de tecido adiposo na região abdominal, sedentarismo, dieta rica em alimentos gordurosos, hábito de fumar, consumo excessivo de bebidas alcóolicas e interação com outras doenças, como por exemplo a doença de Chagas. Há também indícios da influência de pré-disposição genética e de fatores ambientais, o que evidencia a complexidade da casuística da ICC.

A fim de monitorar a incidência da ICC em pacientes que buscam acompanhamento médico no Hospital das Clínicas da Universidade de São Paulo, pesquisadores do Instituto do Coração (InCor) têm dedicado especial atenção ao registro desses casos desde janeiro de 1991, o que resultou em um banco com dados demográficos, antropométricos, biofísicos e bioquímicos de 1.352 pacientes até fevereiro de 2010, mas com a intenção de continuar acompanhando-os e de incluir registros de novos pacientes de forma contínua.

Estudos preliminares dos dados indicaram que, mesmo depois de um longo período de acompanhamento, cerca de 20% dos pacientes do estudo ainda não haviam apresentado o evento de interesse, que no caso era a morte causada por insuficiência cardíaca. Neste conjunto de dados, será considerado que todos os óbitos observados representam *falhas*, ou seja, que ocorreram devido ao agravamento da ICC, embora em alguns casos a causa do óbito seja desconhecida.

O tema deste trabalho surgiu da necessidade de um aprofundamento no conhecimento dos modelos de sobrevivência disponíveis na literatura que permita que se leve em consideração a presença de indivíduos que, embora tenham sido diagnosticados com a doença específica do estudo, após um período de tratamento passam a se comportar como se não estivessem mais sob risco de apresentar o evento de interesse, a saber, a morte pela dada doença. É importante destacar que, até o presente momento, não é aceito pela comunidade médica uma hipotética cura de pacientes diagnosticados com ICC, mas sim que a doença venha a ser controlada de tal forma que o paciente possa ter um prognóstico favorável, o que lhe garante uma sobrevida muito além do esperado relativa àqueles que não foram submetidos ao tratamento. Assim, será denominado *curado* aquele paciente que, após ter sido submetido a um tratamento, passa a ter os sintomas da ICC controlados e que não apresente aumento da suscetibilidade à doenças oportunistas que possam levá-lo à óbito. Em outras palavras, que se o paciente vier à óbito, a causa será outra que não a ICC.

A proposta deste trabalho é estudar em detalhes o modelo de fração de cura com fatores latentes competitivos e fragilidade proposto por Zeng *et al.* (2006), implementar um algoritmo para o ajuste do modelo paramétrico de fração de cura com fatores latentes e avaliá-lo quanto à influência do percentual de censuras na precisão e vício dos estimadores. Além disso, espera-se analisar a proposta de Zeng *et al.* (2006) para contornar os problemas de identificabilidade do modelo nos casos em que o parâmetro da distribuição do efeito aleatório é desconhecido. Por fim, usaremos o modelo paramétrico de fração de cura com fatores latentes competitivos para analisar os dados de tempo de sobrevivência de pacientes acometidos por insuficiência cardíaca congestiva.

Capítulo 3

Modelos de Sobrevivência

Este capítulo inicia-se com uma introdução às funções utilizadas na modelagem de dados de sobrevivência, com especial destaque para as relações funcionais e algumas de suas propriedades. Embora disponíveis na literatura, achou-se por bem exibí-las neste texto a fim de evidenciar as diferenças entre os modelos discutidos e facilitar a compreensão do capítulo seguinte. Posteriormente é feita uma breve discussão do modelo de riscos proporcionais proposto por Cox (1972) e os modelos de fração de cura de mistura de Berkson e Gage (1952) e o de Yakovlev e Tsodikov (1996).

3.1 Conceitos Fundamentais na Modelagem de Dados de Sobrevivência

Considere T uma variável aleatória contínua e não-negativa, sendo $f(t)$ a função de densidade de probabilidade avaliada no ponto t, para $t \geq 0$. A função de distribuição é definida por:

$$F(t) = \int_0^t f(u)du, \quad \forall t \geq 0. \tag{3.1}$$

Destaca-se que a função de distribuição, conforme apresentada em (3.1) possui as seguintes propriedades:

$$\begin{aligned} F(t) = 0, \quad \forall t \leq 0, \quad \text{e} \\ \lim_{t \to \infty} F(t) = 1. \end{aligned} \tag{3.2}$$

Como a função de densidade de probabilidade é, por definição, não-negativa, a função de distribuição é não-decrescente. A partir desta, define-se a função de sobrevivência $S(\cdot)$ avaliada no ponto $t \geq 0$ como:

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - \int_0^t f(u)du = 1 - P(T \leq t) = 1 - F(t). \tag{3.3}$$

Por ser definida como o complementar da função de distribuição (em relação a 1), algumas propriedades da função de sobrevivência decorrem naturalmente daquelas apresentadas em (3.2), a saber:

$$\begin{aligned} S(t) = 1, \quad \forall t \leq 0, \quad \text{e} \\ \lim_{t\to\infty} S(t) = 0. \end{aligned} \tag{3.4}$$

A primeira propriedade em (3.4) surge em decorrência da definição da função de distribuição e de sua relação com a função de sobrevivência, embora na prática não possua uma interpretação intuitiva de interesse. No entanto, a segunda pode ser entendida como a aproximação do valor limítrofe zero pela função de sobrevivência. Isto significa que, se o tempo de acompanhamento for suficientemente grande, a probabilidade do evento não ter ocorrido em um tempo anterior pode ser feita tão próxima de zero quanto se queira. Matematicamente, é equivalente a dizer que a função de sobrevivência decai assintoticamente para zero quando o tempo tende à infinito. À função de sobrevivência que apresenta esse comportamento é dada a denominação de *função de sobrevivência própria*.

Dado que a função de sobrevivência está diretamente relacionada como a função de distribuição, e esta, por sua vez, é definida como a integral da função de densidade de probabilidade, a seguinte propriedade relaciona a derivada da função de sobrevivência com a densidade de probabilidade:

$$S'(t) = \frac{d}{du}S(u)\Big|_{u=t} = \frac{d}{du}\{1 - F(u)\}\Big|_{u=t} = -\frac{d}{du}F(u)\Big|_{u=t} = -f(t). \tag{3.5}$$

Pela definição (3.3) e propriedades da função de distribuição, decorre que $S(\cdot)$ é uma função monótona não-crescente, o que fica evidente em (3.5), que mostra que a função de sobrevivência possui derivada não-positiva.

Uma outra função importante na modelagem de dados de sobrevivência é a função de risco,

também conhecida como função de taxa de falha, neste texto representada por $\lambda(\cdot)$ e definida por:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{3.6}$$

Em particular, $\lambda(t)\Delta t$ é aproximadamente a probabilidade de que ocorra uma falha no intervalo de tempo $(t, t + \Delta t]$, dado que o indivíduo sobreviveu pelo menos até tempo t.

Assim como as funções de densidade, distribuição e de sobrevivência estão relacionadas de forma biunívoca, o mesmo acontece com a função de taxa de falha, como mostrado em (3.7):

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, T > t)}{\Delta t P(T > t)} \\ &= \frac{1}{P(T > t)} \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{3.7}$$

Decorre de (3.5) que $f(t) = -S'(t)$, que substituída em (3.7) resulta na propriedade (3.8) mostrada a seguir:

$$\begin{aligned} \lambda(t) &= -\frac{S'(t)}{S(t)} \\ &= -\frac{d}{dt} \log S(t). \end{aligned} \tag{3.8}$$

A expressão em (3.8) é uma equação diferencial ordinária que pode ser resolvida por integração direta, o que resulta em:

$$\log S(t) = -\int_0^t \lambda(u) du. \tag{3.9}$$

Como a função de taxa de falha $\lambda(\cdot)$ é uma medida de risco em um dado instante de tempo, sua integral pode ser interpretada como sendo o risco de ocorrência do evento de interesse no intervalo de integração, condicionalmente ao conhecimento até o instante imediatamente anterior a cada ponto considerado no intervalo de integração. Dessa forma define-se a função acumulada de risco como sendo:

$$\Lambda(t) = \int_0^t \lambda(u) du. \tag{3.10}$$

Substituindo a expressão (3.10) em (3.9) e aplicando a transformação exponencial em ambos os lados, obtemos a forma (3.11) que relaciona a função de sobrevivência com a função de risco acumulado:

$$S(t) = \exp\{-\Lambda(t)\}. \tag{3.11}$$

Dadas as relações funcionais existentes entre $f(\cdot), F(\cdot), S(\cdot)$ e $\lambda(\cdot)$, conclui-se que todas elas apresentam a mesma informação sobre T, no sentido que uma delas é suficiente para determinar de forma única as demais. Este resultado será utilizado posteriormente neste texto quando da demonstração da identificabilidade do modelo aqui discutido.

Nos estudos de tempo de sobrevivência, a informação disponível para o *i-ésimo* paciente é geralmente apresentada na forma (t_i, δ_i), sendo t_i uma observação da variável aleatória não-negativa T_i, definida como o mínimo entre os tempos de sobrevivência (Y_i) e de censura (C_i) e δ_i uma variável indicadora de falha, isto é,

$$\delta_i = \begin{cases} 1 & \text{se} \quad Y_i \leq C_i \\ 0 & \text{caso contrário.} \end{cases} \tag{3.12}$$

A estruturação dos modelos de sobrevivência permite que aos tempos observados seja relacionado um conjunto de covariáveis consideradas importantes pelo pesquisador, chamado genericamente de $\boldsymbol{x}_i$, sendo que o índice indica que o conjunto de covariáveis refere-se ao i-ésimo indivíduo.

3.2 Modelos Clássicos de Riscos Proporcionais

Os modelos de riscos proporcionais modelam a função de taxa de falha, tratando-a como uma função fatorável em duas componentes, sendo que uma delas contempla a parcela dinâmica do risco, enquanto a outra está relacionada aos efeitos mensurados nas covariáveis. A forma geral de expressar a função de taxa de falha é:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t) \exp\{G(\boldsymbol{x}, \boldsymbol{\beta})\}, \tag{3.13}$$

em que $\lambda_0(\cdot)$ é chamada de função de risco basal e $\boldsymbol{\beta}$ é o vetor de coeficientes da regressão. O segundo termo da expressão (3.13) é convenientemente escrito na forma de exponencial para garantir que seja positivo. Este modelo implica que a razão dos riscos de dois indivíduos é constante ao longo do tempo, supondo que as covariáveis e seus efeitos sejam constantes. Embora alguns modelos acomodem a possibilidade das covariáveis produzirem efeitos aditivos no risco de falha, é conveniente assumir que o efeito das covariáveis seja multiplicativo, o que resulta na função de

risco (3.14):

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}), \tag{3.14}$$

em que o termo $\boldsymbol{x}^T\boldsymbol{\beta}$ é denominado preditor linear. Do modelo (3.14) resulta que a razão de riscos para dois indivíduos depende apenas da diferença entre os respectivos preditores lineares.

O modelo de sobrevivência de riscos proporcionais proposto por Cox (1972) é muito utilizado em estudos de sobrevivência pela versatilidade garantida pela componente semiparamétrica da expressão da função de risco, porém, por ser baseado na suposição de que todos os indivíduos do estudo permaneçam sob o risco de ocorrência do evento de interesse, torna-se inadequado na presença de indivíduos curados.

Diversos livros abordam toda a teoria envolvida na construção e aplicação do modelo de Cox, mas duas boas referências são Kalbfleisch e Prentice (2002) e Therneau e Grambsch (2000). Este último é inteiramente dedicado aos aspectos técnicos do modelo de Cox enquanto o primeiro trata da modelagem de dados de sobrevivência de forma mais abrangente, trazendo como um dos tópicos o modelo de Cox.

3.3 Modelos de Sobrevivência com Fração de Cura

No caso homogêneo considerado usualmente, supomos que toda a diferença sistemática nos tempos de sobrevivência das unidades amostrais é medida pelas covariáveis. Ou seja, excluindo os efeitos das covariáveis (valores desconhecidos, porém estimados), toda a variabilidade apresentada é puramente aleatória. Neste caso, existe uma função de sobrevivência comum à todas as unidades amostrais que é posteriomente estimada.

No modelo com fração de cura, além dos efeitos das covariáveis, ainda existe uma heterogeneidade sistemática devido a uma mistura de estratos na população, sendo um deles composto por indivíduos sob diferentes riscos de ocorrência do evento e o outro composto por aqueles que não encontram-se mais sob o risco de falha. Estes indivíduos *curados* se diferenciam das censuras no sentido que estas são observações retiradas do estudo sem que tenham apresentado o evento de interesse, enquanto os sobreviventes de longa duração são indivíduos que ainda estão no estudo, mas que supostamente não apresentarão o evento de interesse, não importa por quanto tempo sejam observados. Possivelmente, uma parcela das observações censuradas pode ser considerada como sendo composta por sobreviventes de longa duração.

3.3.1 Modelo de mistura

Em geral, um estudo clínico consiste em uma população heterogênea de pacientes, eventualmente dividida em dois grupos. Um grupo consiste naqueles pacientes que respondem favoravelmente ao tratamento, tornando-se imunes ou insuscetíveis à doença, podendo ser considerados curados. O outro grupo é constituído pelos pacientes que não respondem tão bem ao tratamento, continuando sob risco de desenvolver a doença. O interesse principal do pesquisador é determinar a fração de pacientes curados e estudar as causas da falha no tratamento para o grupo não curado. A proporção de curados é uma medida comumente utilizada na obtenção de tendências na sobrevivência de pacientes que sofrem de câncer.

O modelo de sobrevivência proposto por Berkson e Gage (1952) é representado pela função de sobrevivência populacional $S_p(\cdot)$ dada por:

$$S_p(t|\boldsymbol{X} = \boldsymbol{x}) = \phi(\boldsymbol{x}) + \big(1 - \phi(\boldsymbol{x})\big)S^*(t), \tag{3.15}$$

sendo $S^*(\cdot)$ uma função de sobrevivência própria. A fração de curados é dada por $\phi(\boldsymbol{x})$, que deve assumir um valor no intervalo $(0, 1)$. Observe que os valores extremos do intervalo não estão incluídos, uma vez que resultariam em casos particulares sem interesse prático. Por conveniência de notação, o argumento da função ϕ será omitido, a menos que seja indispensável sua indicação.

Uma forma de interpretar a equação (3.15) que define o modelo de mistura é considerar que para os curados, a "função de sobrevivência" seria uma função constante $S(t) \equiv 1, \forall t \geq 0$, enquanto que para aqueles não curados, existe uma função de sobrevivência própria $S^*(t)$ que descreve adequadamente o tempo de sobrevivência. Assim, o modelo proposto poderia ser interpretado como uma média entre estas duas funções, sendo os pesos as proporções de cada estrato na população.

É fácil verificar que a função de sobrevivência populacional $S_p(\cdot)$ tende a ϕ quando $t \to \infty$, sendo este o valor da fração de cura.

O modelo de sobrevivência com fração de cura (3.15) apresenta algumas características que devem ser destacadas. Primeiramente, a heterogeneidade existente na população mensurada pelas covariáveis será relacionada diretamente com a fração de curados. Além disso, ao contrário do que acontece nos modelos tradicionais de sobrevivência em que a função de sobrevivência decai a zero, quando há uma parcela da população imune à ocorrência do evento, a função de sobrevivência se estabiliza em um patamar, o qual denominamos fração de curados. A interpretação dessa característica decorre das suposições intrínsecas do modelo em questão. Quando consideramos um modelo de Cox, por exemplo, existe uma suposição de que todos os indivíduos permanecem sob

risco de ocorrência do evento de interesse até o momento em que ocorre a falha ou a censura.

Neste caso, se o tempo de acompanhamento for suficientemente longo, todos os indivíduos serão censurados ou falharão. Além disso, a função de taxa de falha acumulada é estritamente crescente e ilimitada, o que faz com que a função de sobrevivência se aproxime de zero. Por outro lado, quando consideramos o modelo com fração de cura, após um determinado tempo decorrido, alguns indivíduos não devem ser considerados como estando sob risco de apresentar o evento de interesse. Para essa parcela da população, não importa o quão longo seja o tempo de acompanhamento, os indivíduos não irão apresentar o evento de interesse, mesmo que venham à óbito por causas outras que não as de interesse no estudo. Isso quer dizer que, com o transcorrer do tempo, o risco de ocorrer uma falha na população tende a zero e, sendo a função de taxa de falha acumulada limitada superiormente, faz com que a função de sobrevivência seja imprópria, não decaindo para zero. A Figura (3.1) ilustra as funções de sobrevivência própria e imprópria, respectivamente.

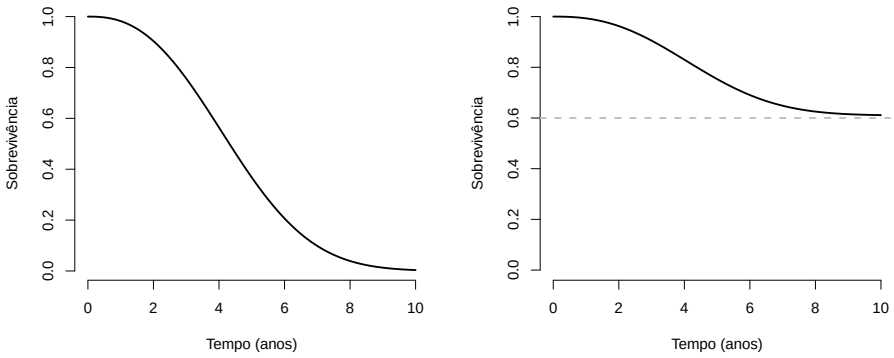

Figura 3.1: Funções de sobrevivência própria (*esq*) e imprópria (*dir*).

Uma discussão detalhada dos aspectos do modelo de mistura pode ser encontrada em Maller e Zhou (1996), enquanto a versão semiparamétrica é apresentada por Hernandez-Quintero *et al.* (2009), abordando inclusive propriedades de consistência e normalidade assintótica dos estimadores não-paramétricos de máxima verossimilhança.

3.3.2 Modelo de cura com fatores latentes competitivos

Apesar de ser intuitivamente atraente, o modelo de mistura possui alguns inconvenientes tanto do ponto de vista bayesiano quanto do frequentista, como destacado por Chen *et al.* (1999). Um modelo alternativo para dados de sobrevivência com fração de cura e propriedades desejáveis foi

proposto por Yakovlev e Tsodikov (1996) e Chen *et al.* (1999). Nesse modelo, aos pacientes curados são atribuídos tempos de sobrevivência iguais a infinito, permitindo que tempos de sobrevivência de pacientes curados e não curados possam ser representados em uma única fórmula. Para o i-ésimo indivíduo com covariáveis $\boldsymbol{x}_i$ na população, a função de sobrevivência será representada por:

$$S_p(t|\boldsymbol{X}_i = \boldsymbol{x}_i) = \exp\{-\theta(\boldsymbol{x}_i)F(t)\}, \tag{3.16}$$

em que $\theta(\cdot)$ é uma função de ligação conhecida e $F(\cdot)$ é uma função de distribuição. Sob o modelo (3.16), a fração de cura é dada por $S(\infty|\boldsymbol{x}_i) = \exp\{-\theta(\boldsymbol{x}_i)\}$ e a taxa de falha no tempo t para o i-ésimo paciente é igual a $\theta(\boldsymbol{x}_i)f(t)$. Embora consista em um abuso, a notação $S(\infty|\boldsymbol{x}_i)$ foi utilizada para representar o limite da função de sobrevivência quanto o tempo tende à infinito.

Suponha que, para um indivíduo na população em estudo, seja K o número de células potencialmente tumorosas, denominadas neste texto como *fatores latentes* e considere que K tenha distribuição de Poisson com média θ. Seja Z_i o tempo que a i-ésima célula leva até desenvolver um tumor detectável. Condicional à $K = k$, as variáveis latentes $Z_1, Z_2, \ldots, Z_k$ são consideradas independentes e identicamente distribuídas, com função de distribuição $F(\cdot) = 1 - S(\cdot)$, que não depende de k. O que se observa, possivelmente, é a variável $T = \min\{Z_i; 1 \leq i \leq k\}$ e, por convenção, considera-se que $P(Z_0 = \infty|K = 0) = 1$. Isto quer dizer que podemos observar o tempo até o aparecimento de um tumor detectável, sendo que, se uma pessoa não possui células potencialmente tumorosas, então não desenvolverá tumor algum, qualquer que seja o tempo que este indivíduo fique sob observação. A função de sobrevivência de T, ou seja, a função de sobrevivência

populacional é obtida do seguinte desenvolvimento:

$$
\begin{aligned}
S_p(t) &= P(T > t) \\
& \sum_{k\geq 0} P(T > t, K = k) \\
& P(T > t, K = 0) + \sum_{k\geq 1} P(T > t, K = k) \\
& P(T > t|K = 0)P(K = 0) + \sum_{k\geq 1} P(T > t|K = k)P(K = k) \\
& P(K = 0) + \sum_{k\geq 1} P(\min\{Z_1, \ldots, Z_k\} > t|K = k)P(K = k) \\
& P(K = 0) + \sum_{k\geq 1} P(\min\{Z_1, \ldots, Z_k\} > t)P(K = k) \\
& P(K = 0) + \sum_{k\geq 1} P(Z_1 > t) \ldots P(Z_k > t)P(K = k) \\
& P(K = 0) + \sum_{k\geq 1} S(t) \ldots S(t)P(K = k) \\
& [S(t)]^0 P(K = 0) + \sum_{k\geq 1} [S(t)]^k P(K = k) \\
& \exp(-\theta) + \sum_{k=1}^{\infty} [S(t)]^k \frac{\theta^k}{k!} \exp(-\theta) \\
& \exp(-\theta) \exp\{1 - \theta S(t)\} = \exp\{-\theta F(t)\}. \qquad (3.17)
\end{aligned}
$$

A equação (3.17) mostra que a função de sobrevivência populacional equivale à função geradora de probabilidades da distribuição postulada para o número de fatores latentes aplicada no ponto $S(t)$, que representa a função de sobrevivência de um fator latente avaliada no tempo t. Neste caso, em particular, considerou-se a distribuição de Poisson. No apêndice (A.2) é considerado o caso em que a distribuição considerada para o número de fatores latentes é a binomial negativa, o que resulta em uma função de sobrevivência equivalente àquela obtida quando considera-se uma distribuição de Poisson e um termo de fragilidade com distribuição gama.

A forma mais conveniente de introduzir no modelo a heterogeneidade mensurada pelas covariáveis é relacionando-as com o parâmetro θ através de uma ligação log-linear, ficando o modelo descrito por:

$$S_p(t|\boldsymbol{X} = \boldsymbol{x}) = \exp\{-\theta(\boldsymbol{x})F(t)\}, \qquad (3.18)$$

sendo $\theta(\boldsymbol{x}) = \exp\{\boldsymbol{\beta}^T \boldsymbol{x}\}$.

Sob o modelo (3.18), a fração de curados é obtida tomando-se o limite da função de sobrevivência quando o argumento tende à infinito.

Embora tenham sido formulados baseando-se em situações distintas, os modelos (3.16) e (3.15) podem ser relacionados, sendo possível escrever aquele no formato deste, como propõem Sumathi e Aruna Rao (2008). Fazendo algumas manipulações algébricas na expressão (3.16), podemos escrevê-la como um modelo de mistura, conforme apresentado em (3.19):

$$S_p(t|\boldsymbol{X} = \boldsymbol{x}) = e^{-\theta(\boldsymbol{x})} + (1 - e^{-\theta(\boldsymbol{x})})\frac{\exp\{-\theta(\boldsymbol{x})F(t)\} - e^{-\theta(\boldsymbol{x})}}{1 - e^{-\theta(\boldsymbol{x})}}. \tag{3.19}$$

Neste caso, a função de sobrevivência condicional é o último termo em (3.19). É importante destacar que no modelo de mistura (3.15) a função de sobrevivência da parcela não curada depende apenas do tempo, enquanto em (3.19) dependeria também das covariáveis.

3.3.3 Curados e o Tempo de Acompanhamento

Uma dificuldade existente na análise de dados de sobrevivência usando os modelos com fração de cura surge justamente na definição dos curados. É indicado fazer uma análise exploratória de dados anteriormente à etapa de modelagem, pois possibilita a identificação de indícios da presença de indivíduos imunes ou curados. Uma característica importante a ser notada é a ocorrência de um grande número de tempos censurados depois de um longo período de acompanhamento. A frequência cada vez menor de falhas na cauda da curva de sobrevivência estimada pelo Kaplan-Meier fornece uma valiosa informação a esse respeito. Entretanto, não é possível identificar com precisão aqueles indivíduos curados, principalmente entre os tempos precocemente censurados.

Sugere-se verificar se as evidências de uma proporção positiva de curados são suficientes para o ajuste de um modelo de fração de cura, uma vez que esta é uma suposição essencial para a aplicação do método de estimação.

Maller e Zhou (1996) discutem a necessidade de testar se o tempo de acompanhamento é suficientemente longo para considerar que os indivíduos censurados ao final do experimento podem ser considerados como imunes à ocorrência do evento de interesse. Diante da demanda por métodos objetivos que permitissem avaliar se as evidências amostrais corroboravam com a hipótese da presença de curados, Zhou e Maller (1995) propuseram um teste da razão de verossimilhanças para a presença de indivíduos imunes em uma amostra de dados censurados. Posteriormente, Klebanov e Yakovlev (2007) demonstraram algumas limitações do teste proposto por Maller e Zhou e propuseram um novo teste para avaliar se o tempo de acompanhamento é suficientemente

longo para detectar a presença de indivíduos curados na amostra.

De fato, após um tempo de acompanhamento suficientemente longo, os pacientes não-curados terão sido censurados ou apresentado o evento de interesse, enquanto aqueles curados terão sido censurados ou permanecerão vivos. Isto é, depois de um tempo limiar de acompanhamento, os pacientes restantes no estudo não estarão mais sob risco de falha. Barajas e Yin (2008) discutem uma forma de usar um modelo semiparamétrico de cura quando este tempo limiar é desconhecido, considerando o contexto bayesiano.

3.3.4 Identificabilidade

Embora a fração de cura seja definida como o valor da assíntota horizontal da cauda da curva de sobrevivência, estudos mostram que a proporção de sobreviventes observada na cauda direita de uma curva de sobrevivência estimada ao final do período de acompanhamento não é capaz de estimar adequadamente a fração de curados (Li *et al.*, 2001). Farewell (1982) mostrou que é difícil distinguir um indivíduo censurado no grupo suscetível de um não suscetível, evidenciando, pois, um efeito de confusão entre as censuras, o que leva à dificuldades na distinção entre o caso em que tem elevada incidência de suscetíveis com distribuição latente de cauda pesada daquela situação em que há baixa incidência de suscetíveis e curtas caudas da distribuição latente. Como um resultado, assegurar a identificabilidade de modelos de cura assume grande importância a fim de que se obtenha estimativas únicas para os parâmetros do modelo postulado quando da disposição dos dados.

Uma referência sobre as condições de identificabilidade dos modelos de mistura é o trabalho de Peng e Zhang (2008). Esta seção apresenta os casos em que é possivel garantir a identificabilidade do modelo de Yakovlev e Tsodikov conforme indicado em Li *et al.* (2001).

Considere o modelo de sobrevivência com fração de cura desenvolvido por Yakovlev e Tsodikov (1996), dado por:

$$S_p(t|\boldsymbol{x}) = \exp[-\theta(\boldsymbol{x})F(t)],\ t < \tau, \tag{3.20}$$

em que $F(t) = 1 - S(t)$ é a função de distribuição, $\theta(\boldsymbol{x}) > 0$ e $\tau > 0$, em que τ é o limiar de cura. Para demonstrar as condições de identificabilidade do modelo, sem perda de generalidade, $\boldsymbol{x}$ será considerado univariado e serão desconsiderados os casos em que $\theta(x) = 0$ e $\theta(x) = \infty$. Denotemos por $\mathscr{F} = \{F(t) : F(t)$ é distribuição própria em $[0, \infty)\}$ a classe de funções de distribuições com suporte em $[0, \infty)$; $\Theta = \{\theta(x) : 0 < \theta(x) < \infty,$ para todo $x \in \mathscr{X}\}$ o espaço dos valores possíveis de $\theta(X)$, e $\mathscr{H} = \{S_p(t|x) = \exp[-\theta(x)F(t)], t < \tau, x \in \mathscr{X}, \theta(\cdot) \in \Theta, F(\cdot) \in \mathscr{F}\}$ a classe de modelos de sobrevivência com fração de cura com riscos proporcionais. A identificabilidade do modelo é

descrita da seguinte forma:

Definição 3.3.1 *A classe de modelos de cura com riscos proporcionais $\mathscr{H}$ é identificável se, para cada $S_p(t|x)$ e $S_p^*(t|x)$ em $\mathscr{H}$ tivermos que $S_p(t|x) \equiv S_p^*(t|x)$ se, e somente se $\theta(x) \equiv \theta^*(x)$, para $x \in \mathscr{X}$ e $F(t) \equiv F^*(t)$ para quase todo $t \in (0, \tau)$.*

Considere que a função de distribuição $F(t)$ não seja especificada, ou seja, que $F(t)$ não possa ser completamente definida por um número finito de parâmetros.

Teorema 3.3.1 *(a) Se $F(t)$ e $\theta(x)$ forem não-especificados, o modelo $S_p(t|x) = \exp[-\theta(x)F(t)]$ não é identificável em $x \in \mathscr{X}, t < \tau$.*

(b) Se $F(t)$ for não especificada e $\theta(x) = \exp(a + bx)$, o modelo $S_p(t|x) = \exp[-\theta(x)F(t)]$ não será identificável, a menos que $a = 0$.

Prova 1 *(a) A demonstração será feita exibindo dois conjuntos diferentes, $\{\theta(x); F(t)\}$ e $\{\theta^*(x); F^*(t)\}$ que determinam a mesma função de sobrevivência populacional e, portanto, a mesma função de verossimilhança.*

Escolha uma constante c tal que $F(\tau) < c < 1$ e seja $\theta^*(x) = c\theta(x), \forall x \in \mathscr{X}$ e $F^*(t) = F(t)/c, \forall t < \tau$. Observe que $S_p^*(t|x) \in \mathscr{H}$.

Então temos que:

$$\begin{aligned} S_p^*(t|x) &= \exp\{-\theta^*(x)F^*(t)\} \\ &= \exp\{-c\theta(x)F(t)/c\} \\ &= \exp\{-\theta(x)F(t)\} \\ &= S_p(t|x). \end{aligned} \tag{3.21}$$

(b) Suponha que $\theta^*(x) = c\theta(x), \forall x \in \mathscr{X}$ e $F^*(t) = F(t)/c, \forall t < \tau$ para algum c tal que $F(\tau) < c \neq 1$. Então $S_p^*(t) \in \mathscr{H}$ e

$$\begin{aligned} S_p^*(t|x) &= \exp\{\exp(a + bx)cF^*(t)\} \\ &= \exp\{\exp(a + bx)cF(t)/c\} \\ &= \exp\{\exp(a + bx)F(t)\} \\ &= S_p(t|x). \end{aligned} \tag{3.22}$$

o que mostra que o modelo não é identificável. Todavia, se $a = 0$, então $S_p^*(t|x) = S_p(t|x)$ implica

em

$$\begin{aligned} S_p^*(t|x) &= S_p(t|x) \\ \Leftrightarrow \theta^*(x)F^*(t) &= \theta(x)F(t)\} \\ \Leftrightarrow \frac{\theta^*(x)}{\theta(x)} &= \frac{F(t)}{F^*(t)}. \end{aligned} \tag{3.23}$$

Observe que o lado esquerdo da equação (3.23) não depende de t, enquanto o lado direito não depende de x. Assim, podemos afirmar que essa razão é um escalar positivo, por hipótese c. Dados $\theta(x) = \exp(bx)$ e $F(t)$, então $\theta^*(x)$ e $F^*(t)$ devem satisfazer:

$$\begin{aligned} F^*(t) &= F(t)/c \\ \theta^*(x) &= c\theta(x) = \exp(\log c + bx). \end{aligned} \tag{3.24}$$

Note que não é possível reescrever $\theta^*(x)$ como sendo $\exp(b^*x)$, a menos que $c = 1$. Isto mostra que se $a = 0$, o modelo será identificável.

Um outro caso de interesse ocorre nos casos em que o modelo de fração de cura de Yakovlev e Tsodikov possui $F(t)$ especificada como uma forma paramétrica determinada por um número finito de parâmetros, como por exemplo a distribuição Weibull, gama ou a exponencial particionada. Neste caso, o modelo fica representado por:

$$S_p(t|\boldsymbol{\psi}, x) = \exp[-\theta(x)F(t;\boldsymbol{\psi})],\ t < \tau, \tag{3.25}$$

em que $\boldsymbol{\psi}$ é o vetor de parâmetros da distribuição latente do tempo.

Teorema 3.3.2 *O modelo apresentado em (3.24) será identificável em $x \in \mathscr{X}, t < \tau$ sendo $\theta(x)$ não especificado ou igual a $\exp(a + bx)$.*

Prova 2 *Suponha que existam um conjunto de parâmetros $\boldsymbol{\psi}^*$ e uma função $\theta^*(x)$ de tal forma que $S_p^*(t|\boldsymbol{\psi}^*, x) = S_p(t|\boldsymbol{\psi}, x)$. Então, a razão dada em (3.26)*

$$\frac{\theta(x)}{\theta^*(x)} = \frac{F^*(t|\boldsymbol{\psi}^*)}{F(t|\boldsymbol{\psi})} \tag{3.26}$$

será uma constante positiva c independente de x e de t para $x \in \mathscr{X}$ e $t < \tau$. Observe que $F^*(t|\boldsymbol{\psi}^*) = cF(t|\boldsymbol{\psi})$ não pertence à mesma família de distribuições que $F(t|\boldsymbol{\psi})$ se $c \neq 1$, a menos que $\boldsymbol{\psi}^* = \boldsymbol{\psi}$.

■

3.3.5 Inferência

O procedimento usualmente adotado para a estimação dos parâmetros do modelo de Cox é baseado na verossimilhança parcial, que consiste na fatoração da função de verossimilhança em dois termos, sendo que apenas um deles depende dos parâmetros interesse. O processo inferencial do modelo de Cox é amplamente discutido na literatura, sendo que uma apresentação detalhada pode ser encontrada em Therneau e Grambsch (2000).

Ao se considerar que uma parcela dos pacientes no estudo encontra-se passível de cura, um modelo que leva em conta esta possibilidade é mais adequado para a modelagem dos dados. O modelo de mistura é geralmente estimado considerando um procedimento do tipo EM, algoritmo utilizado nos casos em que uma variável latente é importante no procedimento inferencial. Neste caso, os dados observados não permitem a identificação dos indivíduos curados, sendo esta a variável latente que justifica o uso do algoritmo EM. A utilização deste possibilita que os parâmetros do modelo sejam estimados considerando uma alocação ótima dos indivíduos nos grupos dos curados e dos suscetíveis, no sentido a proporcionar a maximização da função de verossimilhança dentre todas as alocações possíveis.

Peng (2003) implementou o algoritmo para o ajuste do modelo de mistura e disponibilizou a programação em seu *site*, cujo endereço encontra-se no artigo. A descrição detalhada do procedimento iterativo de estimação é apresentado no artigo, enquanto aspectos teóricos podem ser encontrados em Maller e Zhou (1996).

O modelo de fração de cura com fatores latentes competitivos é um caso particular do modelo discutido no Capítulo 4 e os detalhes do procedimento de estimação serão apresentados na Seção 4.3.

Capítulo 4

Modelos de Fração de Cura e Fragilidade

Foram discutidos no Capítulo 3 os principais tópicos relacionados aos modelos de fração de cura propostos por Berkson e Gage (1952) e o de Yakovlev e Tsodikov (1996). No presente capítulo, será apresentada uma discussão mais detalhada do modelo proposto por Zeng *et al.* (2006), no qual introduzem um termo de fragilidade no modelo de Yakovlev e Tsodikov, a fim de considerar a heterogeneidade entre as observações. Este termo de fragilidade é uma variável não observada e é geralmente considerado como tendo distribuição gama, uma vez que esta escolha permite um adequado tratamento analítico da função de verossimilhança na etapa de inferência.

4.1 Descrição do modelo

No desenvolvimento do modelo (3.16), uma suposição crítica é que, condicional ao número de fatores latentes $K_i = k$, os tempos $(Z_1, \ldots, Z_k)$ são considerados mutuamente independentes. Tal suposição não é muito realística, uma vez que tais tempos latentes são variáveis aleatórias não observáveis pertencentes à mesma unidade amostral.

Uma alternativa a essa restrição é a introdução de um termo de fragilidade específico para cada paciente ξ_i e considerar que, condicional a $K_i = k$ e ξ_i, os tempos latentes $(Z_1, \ldots, Z_k)$ sejam mutuamente independentes e com função de distribuição $F(\cdot)$. Além disso, assume-se que, dados $\boldsymbol{x}_i$ e ξ_i, o número de fatores latentes K_i tenha distribuição de Poisson com parâmetro $\xi_i\theta(\boldsymbol{x}_i)$. Desta forma, ξ_i representa a heterogeneidade entre as quantidades de fatores latentes nos pacientes

do estudo. Procedendo de forma análoga ao desenvolvimento do modelo (3.16), define-se a função de sobrevivência como:

$$S_p(t|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}_\xi\big[\exp\{-\theta(\boldsymbol{x})F(t)\xi\}\big], \tag{4.1}$$

em que $\mathbb{E}_\xi$ denota a esperança com respeito a ξ.

A escolha geralmente feita para a distribuição do termo de fragilidade é uma gama com média unitária, uma vez que acredita-se representar de forma adequada a heterogeneidade entre as unidades amostrais, além de ser conveniente do ponto de vista matemático, pois possibilita o tratamento analítico da função de verossimilhança. A notação adotada no presente texto para a representação da densidade gama de média unitária e variância γ é:

$$g(\xi|\gamma) = \{\gamma^{1/\gamma}\Gamma(1/\gamma)\}^{-1}\xi^{1/\gamma-1}\exp(-\xi/\gamma), \tag{4.2}$$

com $\gamma > 0$. Destaca-se que, no caso limite em que $\gamma \to 0$, a densidade em (4.2) tende a uma distribuição degenerada em 1, que recai no modelo (3.16) em que os tempos de sobrevivência dos pacientes são considerados homogêneos, após excluídos os efeitos das covariáveis.

Ao resolver a equação (4.1) usando a transformação da variável de intergração $u = \xi\big[1 + \theta(x)F(t)\gamma\big]$, resulta na função de sobrevivência dada em (4.3):

$$S_p(t|\boldsymbol{X} = \boldsymbol{x}) = \{1 + \gamma\theta(\boldsymbol{x})F(t)\}^{-1/\gamma}. \tag{4.3}$$

Esta função de sobrevivência populacional pertence a um conjunto mais geral de funções de sobrevivência, dado por:

$$S_p(t|\boldsymbol{X} = \boldsymbol{x}) = G_\gamma\{\theta(\boldsymbol{x})F(t)\}, \tag{4.4}$$

em que a transformação $G_\gamma(\cdot)$ é dada por:

$$G_\gamma(x) = \begin{cases} (1+\gamma x)^{-1/\gamma} & \text{se } \gamma > 0 \\ e^{-x} & \text{se } \gamma = 0. \end{cases} \tag{4.5}$$

Note que o caso $\gamma = 0$ corresponde ao limite da função (4.3) quando $\gamma \to 0$.

As fórmulas (4.4) e (4.5) definem uma classe geral de modelos de sobrevivência, que tem como caso particular o modelo de cura com riscos proporcionais (3.16), obtido quando $\gamma = 0$. Outro caso especial de interesse particular ocorre quando $\gamma = 1$, que resulta no modelo de cura com chances proporcionais, semelhante ao modelo com função de sobrevivência própria considerado por Pettitt (1982) e Bennett (1983). Além disso, a forma geral da classe apresentada em (4.4)

não apenas tem uma importante motivação biológica, mas também pode ser reduzida aos modelos de transformações lineares estudados por Cheng *et al.* (1995) com a escolha adequada de $\theta(\boldsymbol{x})$. De fato, se escolhermos $\theta(\boldsymbol{x}) = \exp(\beta_0 + \boldsymbol{\beta}_c^T \boldsymbol{x_c})$ com $\boldsymbol{x} = (1, \boldsymbol{x}_c^T)^T$, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_c^T)^T$ e β_0 sendo o intercepto da regressão, então o modelo em (4.4) torna-se equivalente a $S(t|\boldsymbol{X} = \boldsymbol{x}) = G_\gamma \left\{\exp(\boldsymbol{\beta}_c^T \boldsymbol{x}_c)\Lambda_0(t)\right\}$, em que $\Lambda_0(t) = F(t)\exp(\beta_0)$ é a função de risco acumulado basal. Entretanto, se $\theta(\boldsymbol{x})$ tiver uma outra forma que não $\theta(\boldsymbol{x}) = \exp(\boldsymbol{\beta}^T \boldsymbol{x})$, o modelo (4.4) ficará ligeiramente diferente daquele definido pela transformação (4.5). O índice c em $\boldsymbol{x}_c$ foi usado com o intuito de representar apenas as p covariáveis, enquanto $\boldsymbol{x}$ é um vetor de dimensão $p+1$ com a constante 1 na primeira posição, uma vez que o preditor linear inclui o intercepto. Para não sobrecarregar a notação, este índice será omitido sempre que possível.

Considere primeiramente o caso em que não há heterogeneidade entre os pacientes do estudo, exceto possivelmente aquela mensurada pelas covariáveis, ou seja, $\gamma = 0$. A função de taxa de falha pode ser calculada a partir da função de sobrevivência proposta usando a relação (3.8):

$$\begin{aligned} \lambda(t|\boldsymbol{x}) &= -\frac{\partial}{\partial t}\log S_p(t|\boldsymbol{x}) \\ &= \frac{\partial}{\partial t}\theta(\boldsymbol{x})F(t) \\ &= \exp(\boldsymbol{\beta}^T\boldsymbol{x})f(t). \end{aligned} \tag{4.6}$$

Considerando a função de taxa de falha obtida em (4.6) e duas observações com covariáveis $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$, obtemos a razão de riscos (RR) dado por:

$$RR = \frac{\lambda(t|\boldsymbol{x}_1)}{\lambda(t|\boldsymbol{x}_2)} = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_1)f(t)}{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_2)f(t)} = \exp\{\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)\}. \tag{4.7}$$

Observe que a razão de riscos não depende do tempo, o que significa que o modelo possui a propriedade de riscos proporcionais, sendo que a interpretação dos coeficientes é feita como no caso do modelo de Cox.

Um outro caso particular acontece quando o $\gamma = 1$, passando o modelo de sobrevivência a apresentar a propriedade de razão de chances proporcionais. Para mostrar a propriedade, primeiramente será calculada a chance de ocorrência de falha:

$$\frac{F(t|\boldsymbol{x})}{1 - F(t|\boldsymbol{x})} = \frac{F(t|\boldsymbol{x})}{S(t|\boldsymbol{x})} = \frac{\theta(\boldsymbol{x})F(t)\{1 + \theta(\boldsymbol{x})F(t)\}^{-1}}{\{1 + \theta(\boldsymbol{x})F(t)\}^{-1}} = \theta(\boldsymbol{x})F(t). \tag{4.8}$$

Assim, a razão de chances entre dois indivíduos com covariáveis $\boldsymbol{x}_1$ e $\boldsymbol{x}_2$ será dada por:

$$RC = \frac{\theta(\boldsymbol{x}_1)F(t)}{\theta(\boldsymbol{x}_2)F(t)} = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_1)F(t)}{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_2)F(t)} = \exp\{\boldsymbol{\beta}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2)\}. \tag{4.9}$$

Conclui-se que a razão de chances neste caso particular do modelo não depende do instante de tempo em que avaliamos as funções de distribuição e de sobrevivência, podendo ser interpretada como a razão de chances de falha para os indivíduos suscetíveis, em qualquer instante de tempo.

4.2 Identificabilidade

De acordo com Zeng *et al.* (2006), quando o parâmetro γ que especifica a transformação em (4.5) é tratado como desconhecido, os parâmetros do modelo podem não ser identificáveis. Por exemplo, suponha que $\theta(\boldsymbol{x}) = \exp(\beta_0)$. Então, para qualquer $\tilde{\gamma}$ diferente de γ, pode-se determinar um $\tilde{\beta}_0$, diferente de β_0 tal que:

$$\{1 + \gamma e^{\beta_0}\}^{-1/\gamma} = \{1 + \tilde{\gamma} e^{\tilde{\beta}_0}\}^{-1/\tilde{\gamma}}. \tag{4.10}$$

Neste caso, para qualquer função de distribuição $F(t)$, será possível definir $\tilde{F}(t)$ de tal forma que

$$\{1 + \gamma e^{\beta_0} F(t)\}^{-1/\gamma} = \{1 + \tilde{\gamma} e^{\tilde{\beta}_0} \tilde{F}(t)\}^{-1/\tilde{\gamma}}. \tag{4.11}$$

Assim sendo, os dois conjuntos distintos de parâmetros (γ, β_0, F) e $(\tilde{\gamma}, \tilde{\beta}_0, \tilde{F})$ resultam na mesma função de sobrevivência, o que implica que não são identificáveis após a observação dos dados.

Além disso, na maioria das aplicações práticas, existe pouca informação nos dados de forma que o parâmetro γ possa ser estimado com um adequado grau de precisão em amostras de tamanhos pequenos ou até mesmo moderados. Nessas situações, a função de verossimilhança apresenta uma curvatura pouco acentuada na direção do parâmetro γ . De acordo com Zeng *et al.* (2006), a experiência mostra que γ pode ser bem estimado quando o tamanho da amostra for muito grande, como por exemplo $n = 1500$ ou maior, mas ainda assim, os algoritmos de estimação podem ficar instáveis. Por causa dessas limitações, considera-se o caso em que γ é fixado para o desenvolvimento do modelo. De qualquer forma, os autores discutem a estimação de γ quando há identificabilidade e sugerem uma estratégia de seleção para a escolha de γ quando este for fixado.

A proposta de Zeng *et al.* (2006) é realizar o ajuste do modelo para diversos valores de γ

e fazer a escolha do melhor modelo com base em um critério de qualidade de ajuste, como por exemplo o AIC (*critério de informação de Akaike*), BIC (*critério de informação bayesiano*), ou na verossimilhança baseada na validação cruzada.

4.3 Inferência

Para descrever o processo de estimação pelo método da máxima verossimilhança do modelo discutido no texto, é importante destacar algumas das suposições do modelo.

Sejam $T_i = Y_i \wedge C_i; \boldsymbol{x}_i, \delta_i = \mathbb{I}\{Y_i \leq C\}$ para $i = 1, 2, \ldots, n$ o tempo observado, o vetor de covariáveis e a variável indicadora de falha do i-ésimo indivíduo, respectivamente. Suponha que o tempo de acompanhamento seja infinito e que, com probabilidade 1, haja uma parcela positiva de curados, isto é, ($Y_i = \infty$ e $C_i = \infty$) para um conjunto não vazio de índices. Além disso, suponha que os tempos de falha e de censura sejam condicionalmente independentes, ou seja, $(C_i \perp Y_i)|\boldsymbol{x}_i$ e que razão de chances seja finita em quase todo ponto.

A função $\theta(\cdot)$ relaciona as covariáveis com o parâmetro da distribuição dos fatores latentes e pode ser escrita como:

$$\theta(\boldsymbol{x}) = \eta(\boldsymbol{\beta}^T \boldsymbol{x}), \tag{4.12}$$

sendo $\eta(\cdot)$ uma função de ligação estritamente positiva e $\beta = (\beta_0, \ldots, \beta_p)^T$. Usando as relações funcionais apresentadas na seção 3.1, é possível obter a função de densidade a partir da sobrevivência (4.4) e, então, a função de verossimilhança será dada por:

$$\begin{aligned} L(\boldsymbol{\beta}, F|\boldsymbol{x}, t, \delta) \quad = \quad & \prod_{i=1}^{n} \Bigg\{ \Bigg[\left\{ -G'\left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F(t_i)\right) \eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) f(t_i) \right\}^{\delta_i} \\ & \times \left\{ G\left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F(t_i)\right) \right\}^{1-\delta_i} \Bigg]^{\mathbb{I}(t_i < \infty)} \left[G\left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right) \right]^{\mathbb{I}(t_i = \infty)} \Bigg\}. \end{aligned} \tag{4.13}$$

Observe que a função apresentada em (4.13) é um produtório de uma função contendo três termos. O primeiro deles é a densidade de probabilidade para os indivíduos que apresentaram a falha; o segundo é a função de sobrevivência para aqueles censurados; e o terceiro é a assíntota da função de sobrevivência condicional ao vetor de covariáveis $\boldsymbol{x}$, que é a contribuição dos indivíduos curados para a função de verossimilhança.

Quando a função de verossimilhança é construída especificando-se uma classe de funções paramétricas de distribuição para o tempo de sobrevivência, a estimação dos parâmetros pode ser

feita maximizando numericamente a expressão em (4.13), porém, se for do interesse do pesquisador que as estimativas sejam feitas sem especificar uma classe paramétrica, torna-se necessário fazer uma modificação, substituindo o termo $f(\cdot)$ pelo estimador não paramétrico $F\{\cdot\}$ definido como

$$F\{t_i\} = F(t_i) - \lim_{t \uparrow t_i} F(t), \tag{4.14}$$

que representa o salto da função de distribuição empírica no ponto de interesse. No caso dessa modificação, a função $F\{\cdot\}$ possui massa distribuída nos pontos em que aconteceram as falhas, sendo que a massa total deve ser unitária. É importante destacar que os curados não contribuem para a estimação da função de distribuição $F(\cdot)$, enquanto as observações censuradas o fazem. Devido à isso, surge a necessidade de estabelecer um prazo de forma a diferenciar os indivíduos curados, denominado limiar de cura, necessário para que o modelo seja identificável. As suposições apresentadas anteriormente especificam um limiar de cura igual a infinito, embora em situações práticas como estudos de sobrevida de pacientes diagnosticados com câncer, pode ser considerado que, se um paciente não apresenta sinais de que a doença persista no período de 5 anos após o tratamento, então este pode ser considerado curado.

Considerando o caso semiparamétrico proposto por Zeng *et al.* (2006), a função a ser maximizada fica descrita por:

$$\begin{aligned} L(\boldsymbol{\beta}, F|\boldsymbol{x}, t, \delta) \quad = \quad & \prod_{i=1}^{n} \Bigg\{ \Bigg[\left\{ -G'\left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F(t_i)\right) \eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F\{t_i\} \right\}^{\delta_i} \\ & \times \left\{ G\left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F(t_i)\right) \right\}^{1-\delta_i} \Bigg]^{\mathbb{I}(t_i < \infty)} \left[G\left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right) \right]^{\mathbb{I}(t_i = \infty)} \Bigg\}. \end{aligned} \tag{4.15}$$

Para obter os estimadores que maximizam a expressão em (4.15), considera-se fixo o vetor $\boldsymbol{\beta}$ e aplicando o logaritmo na função de verossimilhanca, a função de log-verossimilhança será

proporcional a:

$$\begin{aligned} l(F|\boldsymbol{x},t,\delta) &\propto \sum_{i=1}^{n} \mathbb{I}(t_i<\infty)\Bigg[\delta_i \log F\{t_i\} + \delta_i \log\left\{-G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F(t_i)\right)\right\} \\ &\quad +(1-\delta_i)\log G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F(t_i)\right)\Bigg] \\ &\propto \sum_{i=1}^{n} \mathbb{I}(t_i<\infty)\Bigg[\delta_i \log p_i + \delta_i \log\left\{-G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F_i\right)\right\} \\ &\quad +(1-\delta_i)\log G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F_i\right)\Bigg], \end{aligned} \tag{4.16}$$

em que $p_i = F\{t_i\}$ são os saltos do estimador não-paramétrico da função de distribuição e $F_i = \sum_{j:t_j\leq t_i;\delta_j=1} p_j$.

Um ponto a ser destacado neste procedimento é que, embora a expressão (4.16) seja tratada como a função de log-verossimilhança, na realidade trata-se de uma log-verossimilhança condicional, pois é obtida considerando que os coeficientes da regressão estejam fixados.

Para estimar a função de distribuição $F(\cdot)$ usando a expressão (4.16), é necessário fazer uma restrição a fim de assegurar que a função estimada seja, de fato, uma distribuição própria, o que resulta na expressão:

$$\sum_j \delta_j \mathbb{I}(t_j<\infty)p_j = 1. \tag{4.17}$$

Em termos de notação, é conveniente considerar uma ordenação crescente dos tempos distintos de falha $t_{(1)} < t_{(2)}, \ldots, t_{(m)}$; $m = \sum_j \delta_j \mathbb{I}(t_j<\infty)$. Seja $p_{(i)}$ o tamanho do salto de $\hat{F}(\cdot)$ em $t_{(i)} < \infty$. Por definição, tem-se que $p_{(i)} > 0 \Rightarrow \delta_{(i)} = 1$ e $t_{(i)} < \infty$. Assim, a derivada da função (4.16) em relação a $p_{(i)}$ será equivalente à expressão (4.18):

$$\begin{aligned} \frac{\partial l}{\partial p_{(i)}} &= \frac{1}{p_{(i)}} + \sum_{j=1}^{n}\Bigg\{\delta_j \frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathbb{I}(t_{(i)}\leq t_j<\infty) \\ &\quad +(1-\delta_j)\frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathbb{I}(t_{(i)}\leq t_j<\infty)\Bigg\} - \kappa. \end{aligned} \tag{4.18}$$

sendo o último termo o multiplicador de Lagrange devido à restrição apresentada em (4.17).

A estimação é feita nos $m-1$ maiores saltos da função de distribuição $\hat{F}(\cdot)$, para $\boldsymbol{\beta}$ fixado, ficando as equações de estimação em função de $\boldsymbol{\beta}$. Resolvendo a equação escore apresentada em

(4.18), ou seja, igualando-a a zero e desenvolvendo o somatório, obtém-se:

$$\begin{aligned}\frac{1}{p_{(i)}} &+ \sum_{j=1}^{n}\left\{\delta_j \frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathbb{I}(t_{(i)} \leq t_j < t_{(i+1)})\right.\\ &+ \left.(1-\delta_j)\frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathbb{I}(t_{(i)} \leq t_j < t_{(i+1)})\right\}\\ &+ \sum_{j=1}^{n}\left\{\delta_j \frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathbb{I}(t_{(i+1)} \leq t_j < \infty)\right.\\ &+ \left.(1-\delta_j)\frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\mathbb{I}(t_{(i+1)} \leq t_j < \infty)\right\} - \kappa = 0. \qquad (4.19)\end{aligned}$$

Nota-se que o segundo somatório na expressão (4.19) é equivalente a $-\dfrac{1}{p_{(i+1)}}$ e que, da forma como foi definido $t_{(i)}$, tem-se que:

$$t_{(i)} \leq t_j < t_{(i+1)} \Rightarrow \begin{cases} t_j = t_{(i)} \text{ e } \delta_j = 1 \\ \text{ou } \delta_j = 0. \end{cases} \qquad (4.20)$$

A relação em (4.20) significa que, em um intervalo definido por dois tempos de falha consecutivos, os tempos observados são censuras ou são iguais ao limite inferior do intervalo. Como o tempo de sobrevivência é supostamente contínuo, é razoável considerar que os tempos de falha sejam não coincidentes, ou seja, que não haja "empates".

Substituindo a expressão (4.20) em (4.19), obtém-se uma relação entre $p_{(i)}$ e $p_{(i+1)}$, o que sugere a possibilidade de estabelecer uma forma recursiva entre essas estimativas, o que resulta em:

$$\frac{1}{p_{(i)}} + \sum_{t_{(i)} \leq t_j \leq t_{(i+1)}}\left\{\delta_j \frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j) + (1-\delta_j)\frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_j\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\right\} - \frac{1}{p_{(i+1)}} = 0. \qquad (4.21)$$

A fim de se obter uma simplificação da expressão em (4.21), considere a seguinte notação:

$$F_j = \sum_{t_k \leq t_j; \delta_k = 1} p_k = \sum_{t_k \leq t_{(i)}} p_k + \sum_{t_{(i)} < t_k < t_j < t_{(i+1)}} p_k = F_{(i)}. \qquad (4.22)$$

O segundo somatório é nulo, pois a função estimada é do tipo degrau, com saltos apenas nos

tempos distintos de falha, o que faz com que a expressão em (4.21) possa ser reescrita como:

$$\frac{1}{p_{(i+1)}} = \frac{1}{p_{(i)}} + \frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})F_{(i)}\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})F_{(i)}\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)}) + \sum_{t_{(i)}<t_j<t_{(i+1)}} \frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_{(i)}\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_{(i)}\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j). \quad (4.23)$$

Usando o fato de que $\sum_{i=1}^m p_{(i)} = 1 \Rightarrow F_{(i)} = p_{(1)} + p_{(2)} + \cdots + p_{(i)} = 1 - \left[p_{(i+1)} + \cdots + p(m)\right] = 1 - S_{(i+1)}$, a expressão em (4.23) pode ser reescrita como:

$$\frac{1}{p_{(i)}} = \frac{1}{p_{(i+1)}} - \frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})\left[1-S_{(i+1)}\right]\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})\left[1-S_{(i+1)}\right]\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)}) - \sum_{t_{(i)}<t_j<t_{(i+1)}} \frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\left[1-S_{(i+1)}\right]\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\left[1-S_{(i+1)}\right]\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j), \quad (4.24)$$

para $i = 1, 2, \ldots, m-1$.

De acordo com Zeng *et al.* (2006), a expressão (4.24) é preferível por assegurar que $0 < p_{(i)} < p_{(i+1)}; \forall i = 1, 2, \ldots, m-1$. A estimação é feita considerando $\boldsymbol{\beta}; \alpha := p_{(m)} > 0$ e κ como parâmetros independentes e que $p_{(1)}, p_{(2)}, \ldots, p_{(m-1)}$ sejam funções de $\boldsymbol{\beta}$ e α.

Denotemos por $l_r(\cdot)$ a função de log-verossimilhança restrita:

$$\begin{aligned} l_r(\alpha, \boldsymbol{\beta}, \kappa|\boldsymbol{x}, \boldsymbol{t}, \boldsymbol{\delta}) &= \sum_{i=1}^n \mathbb{I}(t_i < \infty)\Bigg\{\delta_i \log p_i + \delta_i \log\left[G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F_i\right)\right] + \delta_i \log \eta(\boldsymbol{\beta}^T\boldsymbol{x}_i) + \\ &\quad (1-\delta_i)\log\left[G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F_i\right)\right]\Bigg\} + \sum_{i=1}^n \mathbb{I}(t_i = \infty)\log G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)\right) - \\ &\quad \kappa\left[\sum_{i=1}^n \delta_i \mathbb{I}(t_i < \infty)p_i - 1\right]. \end{aligned} \quad (4.25)$$

A função de log-verossimilhança restrita em (4.25) é usada pra obter as equações de estimação de $\boldsymbol{\beta}$, de $\alpha = p_{(m)}$ que é o maior salto na função degrau $\hat{F}(\cdot)$ e κ. Este último é necessário no processo de estimação, mas não é de interesse final. Como p_i depende de $\boldsymbol{\beta}$ (e consequentemente F_i também depende de $\boldsymbol{\beta}$), será usada a propriedade da derivada implícita nas parcelas envolvendo os termos p_i e F_i. Destaca-se novamente que $\{p_{(i)}\}$ é apenas uma ordenação conveniente do conjunto $\{p_i\}$.

As equações de estimação são obtidas calculando-se as derivadas da função de log-verossimilhança restrita (4.25) em relação aos parâmetros do modelo, inclusive o parâmetro perturbador κ usado

na restrição do estimador da função de distribuição.

$$
\begin{aligned}
\frac{\partial l_r}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \mathbb{I}(t_i < \infty) \left\{ \delta_i \log p_i + \delta_i \log \left[G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F_i \right) \right] \right\} + \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \mathbb{I}(t_i < \infty) \delta_i \log \eta(\boldsymbol{\beta}^t \boldsymbol{x}_i) \\
&+ \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \mathbb{I}(t_i < \infty)(1-\delta_i) \log \left[G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F_i \right) \right] + \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \mathbb{I}(t_i = \infty) \log G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \right) \\
&- \frac{\partial}{\partial \boldsymbol{\beta}} \kappa \left[\sum_{i=1}^{n} \delta_i \mathbb{I}(t_i < \infty) p_i - 1 \right] \\
&= \frac{\partial}{\partial p_{(i)}} \frac{\partial p_{(i)}}{\partial \boldsymbol{\beta}} \sum_{i=1}^{m} \left\{ \log p_{(i)} + \log \left[G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) F_{(i)} \right) \right] \right\} + \sum_{i=1}^{n} \delta_i \frac{\partial}{\partial \boldsymbol{\beta}} \log \eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \\
&+ \sum_{i=1}^{n} \mathbb{I}(t_i < \infty)(1-\delta_i) \frac{\partial}{\partial \beta} \log \left[G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) F_i \right) \right] + \sum_{i=1}^{n} \mathbb{I}(t_i = \infty) \frac{\partial}{\partial \boldsymbol{\beta}} \log G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \right) \\
&- \kappa \left[\sum_{i=1}^{m} \delta_{(i)} \frac{\partial}{\partial \boldsymbol{\beta}} p_{(i)} \right] \\
&= \sum_{i=1}^{m} \frac{1}{p_{(i)}} \frac{\partial p_{(i)}}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{m} \frac{G'' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) F_{(i)} \right)}{G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) F_{(i)} \right)} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) x_{(i)} F_{(i)} + \eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) \frac{\partial F_{(i)}}{\partial \boldsymbol{\beta}} \right] \\
&+ \sum_{i=1}^{n} \delta_i \frac{\eta'(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i)} x_i + \sum_{i=1}^{n-1} \sum_{t_{(i)} < t_j < t_{(i+1)}} \frac{G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) F_{(i)} \right)}{G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) F_{(i)} \right)} \left[\eta'(\boldsymbol{\beta}^T \boldsymbol{x}_j) x_j F_{(i)} + \eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) \frac{\partial F_{(i)}}{\partial \boldsymbol{\beta}} \right] \\
&+ \sum_{t_{(m)} < t_j < \infty} \frac{\partial}{\partial \boldsymbol{\beta}} \log G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) F_{(m)} \right) + \sum_{i=1}^{n} \mathbb{I}(t_i = \infty) \frac{G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \right)}{G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \right)} \eta'(\boldsymbol{\beta}^T \boldsymbol{x}_i) x_i - \kappa \sum_{i=m} \frac{\partial p_{(i)}}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{m} \frac{1}{p_{(i)}} \frac{\partial p_{(i)}}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{m} \frac{G'' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) F_{(i)} \right)}{G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) F_{(i)} \right)} \left[\eta'(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) x_{(i)} F_{(i)} + \eta(\boldsymbol{\beta}^T \boldsymbol{x}_{(i)}) \frac{\partial F_{(i)}}{\partial \boldsymbol{\beta}} \right] + \sum_{i=1}^{n} \delta_i x_i \\
&+ \sum_{i=1}^{m} \sum_{t_{(i)} < t_j < t_{(i+1)}} \frac{G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) F_{(i)} \right)}{G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) F_{(i)} \right)} \left[\eta'(\boldsymbol{\beta}^T \boldsymbol{x}_j) x_j F_{(i)} + \eta(\boldsymbol{\beta}^T \boldsymbol{x}_j) \frac{\partial F_{(i)}}{\partial \beta} \right] \\
&+ \sum_{i=1}^{n} \mathbb{I}(t_i = \infty) \frac{G' \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \right)}{G \left(\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) \right)} \eta'(\boldsymbol{\beta}^T \boldsymbol{x}_i) x_i - \kappa \sum_{i=m} \frac{\partial p_{(i)}}{\partial \boldsymbol{\beta}}. \qquad (4.26)
\end{aligned}
$$

A derivada da função escore com relação a α é apresentada na expressão (4.27):

$$\begin{aligned}
\frac{\partial l_r}{\partial \alpha} &= \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}\mathbb{I}(t_i<\infty)\left\{\delta_i \log p_i + \delta_i \log G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F_i\right)\right\} + \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}\mathbb{I}(t_i<\infty)\delta_i \log \eta(\boldsymbol{\beta}^T\boldsymbol{x}_i) \\
&+ \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}\mathbb{I}(t_i<\infty)(1-\delta_i)\log G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_i)F_i\right) + \frac{\partial}{\partial \alpha}\sum_{i=i}^{n}\mathbb{I}(t_i=\infty)\log G\left(\boldsymbol{\beta}^T\boldsymbol{x}_i\right) \\
&- \frac{\partial}{\partial \alpha}\kappa\left[\sum_{i=1}^{n}\delta_i\mathbb{I}(t_i<\infty)p_i - 1\right] \\
&= \sum_{i=1}^{m}\frac{1}{p_{(i)}}\frac{\partial p_{(i)}}{\partial \alpha} + \sum_{i=1}^{m}\frac{G''\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})F_{(i)}\right)}{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})F_{(i)}\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_{(i)})\frac{\partial F_{(i)}}{\partial \alpha} \\
&+ \sum_{i=1}^{m}\sum_{t_{(i)}<t_j<t_{(i+1)}}\frac{G'\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_{(i)}\right)}{G\left(\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)F_{(i)}\right)}\eta(\boldsymbol{\beta}^T\boldsymbol{x}_j)\frac{\partial F_{(i)}}{\partial \alpha} - \kappa\sum_{i=m}\frac{\partial p_{(i)}}{\partial \alpha}.
\end{aligned} \tag{4.27}$$

A derivada com relação a κ é dada por:

$$\frac{\partial l_r}{\partial \kappa} = -\left[\sum_{i=1}^{n}\delta_i\mathbb{I}(t_i<\infty)p_i - 1\right] = -\left[\sum_{i=1}^{m}p_{(i)} - 1\right]. \tag{4.28}$$

No trabalho de Zeng *et al.* (2006), a resolução do sistema de equações de estimação é feita usando um algoritmo de Newton-Raphson. No capítulo seguinte apresentamos algumas simulações em que um modelo paramétrico foi ajustado aos conjuntos de dados simulados, e para estimar as variâncias assintóticas foi considerada a matriz de informação observada, estimada no processo de otimização da função de log-verossimilhança no ponto de máximo. A inversa dessa matriz é uma medida de curvatura da função otimizada na região da estimativa.

Capítulo 5

Simulações

Uma das propostas do presente trabalho é avaliar o modelo paramétrico de fração de cura com fatores latentes competitivos, levando-se em consideração ou não um efeito aleatório de fragilidade. Para isso, foram feitas simulações em que amostras foram geradas, com posterior ajuste do modelo discutido. As estimativas dos parâmetros foram usadas para estimar medidas como o erro quadrático médio (EQM), a variância e o vício dos estimadores.

Este capítulo apresenta alguns aspectos técnicos dos processos de geração de amostras e de estimação, além de exibir em tabelas o resumo dos resultados da simulação e em gráficos as densidades estimadas dos estimadores do modelo.

Para estimar os parâmetros do modelo proposto, foi desenvolvido um programa computacional na linguagem R (R Development Core Team, 2004), bem como uma rotina para a geração de amostras utilizadas para avaliar o desempenho do processo de estimação.

5.1 Geração de amostras

O tempo T decorrido desde o início do acompanhamento até a ocorrência do evento de interesse é determinado no instante de ocorrência do primeiro evento causado pelos fatores latentes competitivos. Supondo que cada fator latente esteja associado a uma função de sobrevivência $S(\cdot)$ e que sejam independentes e identicamente distribuídos, condicional ao número de fatores latentes

$K = k$, a distribuição do tempo de falha pode ser obtida conforme mostrado na equação (5.1).

$$\begin{aligned} F_T(t) = P(T \leq t) &= P(\min\{Z_1, \ldots, Z_k\} \leq t) \\ &= 1 - P(\min\{Z_1, \ldots, Z_k\} > t) \\ &= 1 - P(Z_1 > t, \ldots, Z_k > t) \\ &= 1 - P(Z_1 > t) \ldots P(Z_k > t) \\ &= 1 - [P(Z > t)]^k \\ &= 1 - [S(t)]^k . \end{aligned} \tag{5.1}$$

A função de densidade de probabilidade de T pode ser obtida aplicando a derivada na equação (5.1), o que resulta em:

$$f_T(t) = dF_T(t) = k\,[1 - F(t)]^{(k-1)}\, f(t). \tag{5.2}$$

Para as amostras usadas na simulação apresentada no presente capítulo, adotou-se a distribuição exponencial de parâmetro ω como geradora dos tempos associados aos fatores latentes. Sendo assim, a função de densidade de probabilidade do tempo de falha será dada por:

$$\begin{aligned} f_T(t) &= k\left[e^{-\omega t}\right]^{(k-1)} \omega e^{-\omega t} \\ &= (\omega k) e^{-(\omega k)t}, \end{aligned} \tag{5.3}$$

de onde se conclui que $T \sim exp(\omega k)$. Isto significa que o mínimo entre k variáveis aleatórias independentes com distribuição exponencial de parâmetro ω é uma variável aleatória com distribuição exponencial de parâmetro ωk. Destaca-se que neste caso k é um inteiro positivo fixado.

Está sendo considerado que o i-ésimo indivíduo possui k_i fatores latentes que podem gerar o evento de interesse. O tempo até a ocorrência do evento de interesse segundo o modelo proposto é o mínimo dos tempos latentes, assumidos aqui como realizações de uma distribuição exponencial de parâmetro ω. Assim, o tempo que pode ser observado é uma realização de uma distribuição exponencial de parâmetro $k_i\omega$, desde que $k_i > 0$.

Observe que, quanto maior o número de fatores latentes em um indivíduo, menor será o tempo esperado até que aconteça o evento de interesse. Segundo a formulação do modelo de Yakovlev e Tsodikov (1996), um indivíduo que não possui nenhum desses fatores latentes não apresentará o evento de interesse, mesmo que o tempo de acompanhamento seja muito longo. Convencionalmente, consideramos que o tempo de falha para estes indivíduos é infinito.

Nos estudos de simulação, é importante que tenhamos sob controle as principais características das amostras. Uma dessas características importantes de se controlar é a proporção de tempos

censurados. Supondo que o tempo de censura tenha distribuição exponencial com parâmetro μ, é possível relacionar os parâmetros das distribuições dos tempos de falha e de censura com a probabilidade de observar-se uma censura, conforme apresentado em (5.4):

$$\begin{aligned} P(T > C) = \mathbb{E}\left[P(T > C|C)\right] &= \int_0^{+\infty} P(T > c|C = c)\mu e^{-\mu c} dc \\ &= \int_0^{+\infty} e^{-\omega c}\mu e^{-\mu c} dc \\ &= \frac{\mu}{\mu+\omega}\int_0^{+\infty}(\omega+\mu)e^{-(\omega+\mu)c} dc \\ &= \frac{\mu}{\mu+\omega}. \end{aligned} \tag{5.4}$$

Intuitivamente, aumentando o parâmetro μ do tempo de censura, diminuímos o tempo médio de censura, o que faz com que frequentemente se observe tempos de falha superiores aos de censura, resultando em uma maior proporção de observações censuradas.

A equação (5.4) apresenta a probabilidade da variável T ser maior do que C, ou seja, de uma observação censurada. Para gerar uma amostra com a proporção q, com $(0 < q < 1)$ de censuras, devemos escolher valores positivos de ω e μ tal que, dado $K_i = k_i$,

$$\frac{\mu}{\mu + k_i\omega} = q \Leftrightarrow k_i\omega = \frac{(1-q)}{q}\mu. \tag{5.5}$$

No lado direito da equação (5.5) podemos ver que, pré-fixando um valor para a proporção de censuras q e para o parâmetro da distribuição do tempo de censura (μ), então a variabilidade entre os números de fatores latentes será necessariamente "compensada" por uma heterogeneidade nos valores de ω, o que contradiz a suposição de que os tempos de falha devem ser *iid*.

A alternativa para esta situação foi primeiramente gerar uma covariável indicadora de grupo e determinar os parâmetros do preditor linear (β_0 e β_1) de tal forma que as proporções de curados nos dois grupos sejam controladas. Posteriormente, os valores de θ foram obtidos por (5.6).

$$\theta_x = \theta(x) = \exp(\beta_0 + \beta_1 x). \tag{5.6}$$

O número de fatores latentes de cada indivíduo é uma realização de uma distribuição de Poisson com média θ_x, com $x \in \{0, 1\}$. A fim de controlar a proporção de censuras, foi estabelecida uma relação semelhante à apresentada em (5.5), porém, substituindo k_i pelo valor esperado θ_x.

Após o processo de geração dos tempos de censura, o número de fatores latentes $k_.$ foi gerado para cada indivíduo supondo uma distribuição de Poisson com parâmetro θ_x, dada a covariável indicadora de grupo. Em seguida, definiu-se o tempo de falha como sendo o mínimo dentre os

k_i tempos exponenciais *iid* com parâmetro ω previamente escolhido. No caso descrito por um modelo com fração de cura, uma parcela da amostra não terá o tempo de sobrevivência descrito por uma distribuição exponencial, uma vez que considera-se que o tempo até a ocorrência do evento é infinito.

Nas simulações realizadas, considerou-se 1000 amostras com cada conjunto de parâmetros, sendo que os resultados dos ajustes para cada amostra simulada foram utilizados para a estimação do EQM, da variância e do vício. Optou-se por tamanhos de amostra iguais a $n = 250$ e $n = 1000$, com proporções de censuras de 25% e 50%, uma vez que valores superiores à estes causavam instabilidade no processo de estimação nos casos em que se dispunha de amostras pequenas. Em cada amostra simulada, foi gerada uma única covariável a fim de classificar as unidades amostrais em dois grupos, aqui denominados controle e tratamento, de forma balanceada.

Dada esta covaríavel, as frações de cura para os grupos controle e tratamento são obtidas em:

$$\begin{aligned}
&P(K=0|X=x) = e^{-\theta(x)} = \exp\{-e^{\beta_0+\beta_1 x}\} \\
&p_0 = P(K=0|X=0) = \exp\{-e^{\beta_0}\} \Leftrightarrow \beta_0 = \log\{-\log(p_0)\} \\
&p_1 = P(K=0|X=1) = \exp\{-e^{\beta_0+\beta_1}\} \Leftrightarrow \beta_0 + \beta_1 = \log\{-\log(p_1)\}.
\end{aligned} \tag{5.7}$$

Para os casos em que o efeito da fragilidade foi considerado na geração dos dados, obtemos as frações de cura na seguinte expressão:

$$\begin{aligned}
&p_0 = \{1+\gamma\exp(\beta_0)\}^{(-1/\gamma)} \Leftrightarrow \beta_0 = \log\left(\frac{p_0^{-\gamma}-1}{\gamma}\right) \\
&p_1 = \{1+\gamma\exp(\beta_0+\beta_1)\}^{(-1/\gamma)} \Leftrightarrow \beta_1 = \log\left(\frac{p_1^{-\gamma}-1}{\gamma}\right) - \log\left(\frac{p_0^{-\gamma}-1}{\gamma}\right) \\
&\beta_1 = \log\left(\frac{p_1^{-\gamma}-1}{p_0^{-\gamma}-1}\right).
\end{aligned} \tag{5.8}$$

As frações de cura para esses grupos foram ajustadas em $0,20$ e $0,60$, respectivamente, e os correspondentes coeficientes do preditor linear foram obtidos pela relação (5.7) nos casos sem fragilidade e por (5.8) quando γ era positivo.

Nas amostras com fragilidade, a variável latente foi gerada de uma distribuição gama com média 1 e parâmetro de escala γ. Dados os valores dos coeficientes e a covariável x, determinou-se $\theta(x)$ para cada indivíduo da amostra usando a ligação $\log(\theta) = \beta_0 + \beta_1 x_i$, valor utilizado como parâmetro da distribuição de Poisson para gerar os números de fatores latentes.

Embora o uso da distribuição exponencial com parâmetro $k_i\omega$ seja mais eficiente computacionalmente, optou-se pela geração de k_i realizações de uma exponencial($\omega = 0,025$), das quais o mínimo foi considerado como o tempo de falha da i-ésima unidade amostral.

Ao final desse processo de geração, tínhamos a disposição uma amostra de tempos de acompanhamento até a ocorrência do evento de interesse (ou até a censura), a covariável indicadora de grupo e a variável indicadora de falha. Foi desenvolvida uma função que estima o ponto de máximo da função de log-verossimilhança e a matriz hessiana, também conhecida como matriz de informação observada, necessária para a estimação dos desvios assintóticos dos estimadores.

Um caso importante de ser analisado, embora irreal na prática, ocorre quando o valor do parâmetro γ é conhecido. O caso mais realístico ocorre quando o parâmetro γ é desconhecido. Nesse caso, foi utilizado o procedimento proposto por Zeng *et al.* (2006), em que o modelo é estimado para vários valores de gama e é escolhido aquele que apresenta maior AIC.

Para gerar as amostras e determinar o melhor modelo, foram considerados os seguintes valores de γ: $\{0,00; 0,25; 0,50; 0,75; 1,00; 1,25\}$. Especificamente no caso da geração das amostras, foi desconsiderado o último desses valores, admitindo-o como valor possível no processo de estimação.

O algoritmo implementado para o ajuste do modelo levou em consideração que os tempos associados aos fatores latentes competitivos possuíam distribuição na classe Weibull, com função de densidade de probabilidade dada por:

$$f(x) = \upsilon\omega(\omega x)^{(\upsilon-1)} \exp\{-(\omega x)^{(\upsilon)}\}. \tag{5.9}$$

O caso particular usado nas simulações foi *Weibull*($\upsilon = 1; \omega = 0,025$).

Em avaliação prévia dos procedimentos de geração de amostras e de estimação mostrou que a diferença entre proporção observada de censura na parcela suscetível da população e os valores teóricos pode ser considerada desprezível.

5.2 Resultados

No primeiro cenário simulado, considerou-se o caso em que o valor do parâmetro da distribuição da fragilidade é conhecido. A Tabela 5.1 apresenta medidas descritivas dos estimadores dos coeficientes da regressão, os parâmetros da distribuição dos tempos latentes e das frações de cura para os grupos controle e tratamento.

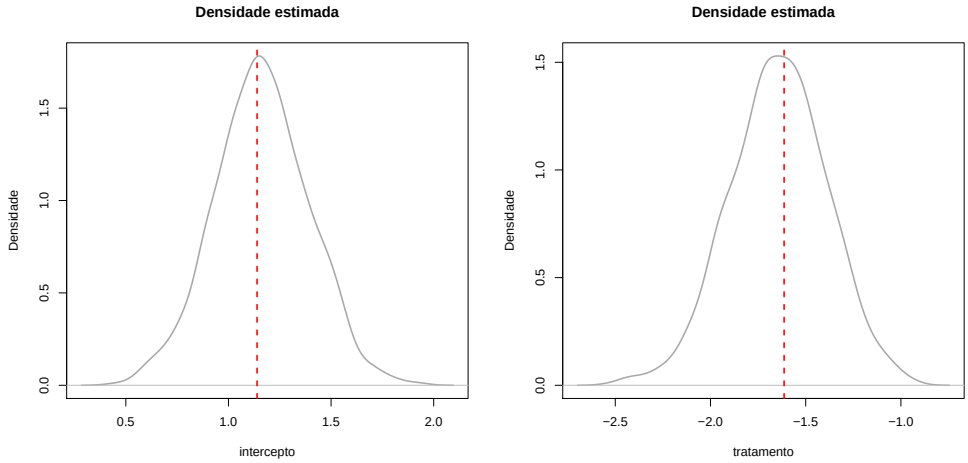

Figura 5.1: Densidade dos estimadores dos coeficientes de regressão.

A Figura (5.1) mostra um esboço das densidades dos estimadores dos coeficientes da regressão, obtidas de 1000 amostras simuladas de tamanho 250 contendo cerca de 25% de censuras, com $\gamma = 0,75$. A linha tracejada indica o valor ajustado na geração dos dados. As formas das densidades estimadas sugerem uma aproximação da distribuição normal assintótica, resultado conhecido para os estimadores de máxima verossimilhança.

As Figuras (5.2) e (5.3) apresentam esboços das densidades estimadas dos estimadores dos parâmetros da distribuição dos tempos latentes e das frações de cura, respectivamente.

Em todos os casos, observa-se que as estimativas encontram-se distribuídas em torno dos valores verdadeiros usados na geração das amostras simuladas, indicando não haver vícios consideráveis, o que é corroborado pela Tabela 5.1. Em alguns casos, os valores de EQM, variância e vício foram menores do que $0,00005$, mas como adotou-se como padrão o uso de quatro casas decimais, tais valores aparecem como $0,0000$ nas tabelas.

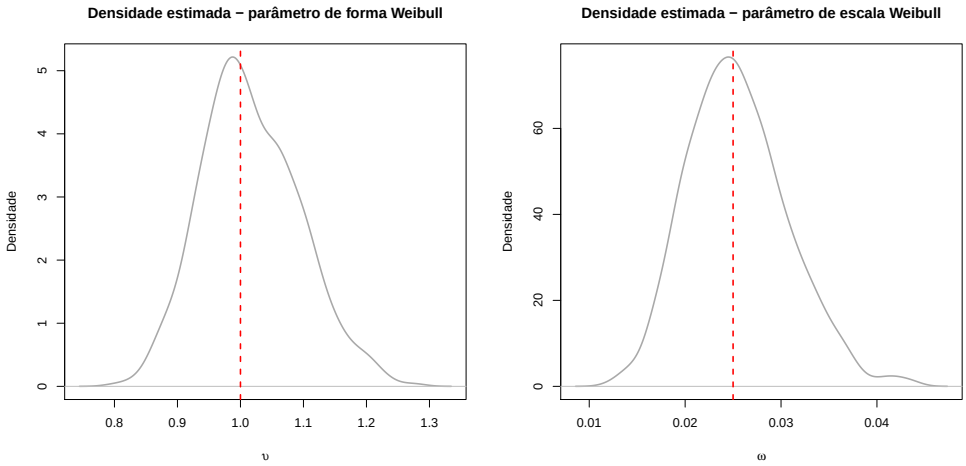

Figura 5.2: Densidade dos estimadores dos parâmetros de forma ($\hat{\upsilon}$) e de escala ($\hat{\omega}$) da distribuição weibull dos tempos latentes.

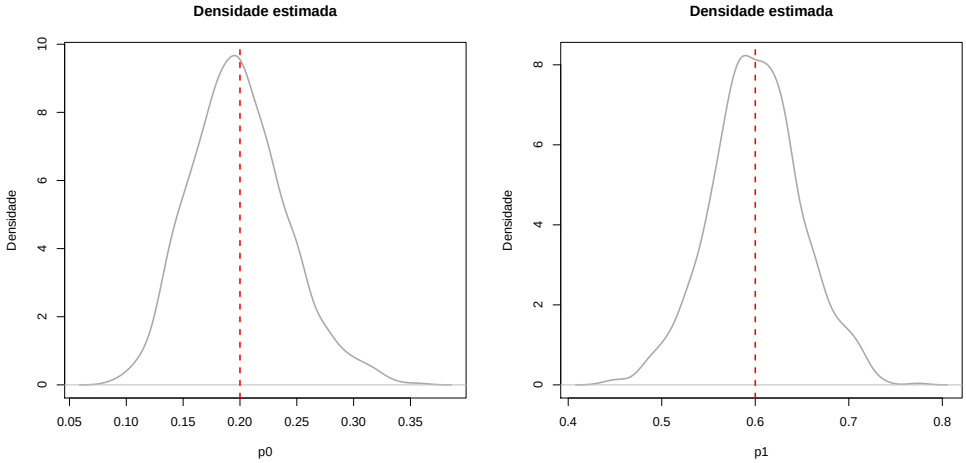

Figura 5.3: Densidade dos estimadores das frações de cura.

Tabela 5.1: Resultados das simulações para γ conhecido

Gama	**N**	**Parâmetro**	**Valor**	**Média**	**Mediana**	**EQM**	**Var**	**Vício^2**	**Vício**
0,00	250	β_0	0,4760	0,4754	0,4733	0,0207	0,0207	0,0000	-0,0006
		β_1	-1,1480	-1,1522	-1,1495	0,0372	0,0372	0,0000	-0,0042
		υ	1,0000	1,0102	1,0034	0,0059	0,0058	0,0001	0,0102

continua na próxima página

Tabela 5.1 – continuação

Gama	N	Parâmetro	Valor	Média	Mediana	EQM	Var	Vício^2	Vício
		ω	0,0250	0,0254	0,0253	0,0000	0,0000	0,0000	0,0004
		p_0	0,2000	0,2022	0,2008	0,0021	0,0021	0,0000	0,0022
		p_1	0,6000	0,5998	0,5999	0,0022	0,0022	0,0000	-0,0002
	1000	β_0	0,4760	0,4830	0,4820	0,0047	0,0047	0,0001	0,0070
		β_1	-1,1480	-1,1550	-1,1540	0,0097	0,0096	0,0001	-0,0070
		υ	1,0000	1,0050	1,0040	0,0015	0,0015	0,0000	0,0050
		ω	0,0250	0,0250	0,0250	0,0000	0,0000	0,0000	0,0000
		p_0	0,2000	0,1980	0,1980	0,0005	0,0005	0,0000	- 0,0020
		p_1	0,6000	0,6000	0,5990	0,0006	0,0006	0,0000	0,0000
0,25	250	β_0	0,6840	0,6900	0,6888	0,0276	0,0275	0,0000	0,0060
		β_1	-1,2910	-1,3119	-1,3157	0,0426	0,0422	0,0004	-0,0209
		υ	1,0000	1,0196	1,0149	0,0067	0,0063	0,0004	0,0196
		ω	0,0250	0,0258	0,0255	0,0000	0,0000	0,0000	0,0008
		p_0	0,2000	0,2007	0,1987	0,0019	0,0019	0,0000	0,0007
		p_1	0,6000	0,6026	0,6043	0,0023	0,0023	0,0000	0,0026
	1000	β_0	0,6840	0,6869	0,6828	0,0072	0,0072	0,0000	0,0029
		β_1	-1,2910	-1,3030	-1,2981	0,0112	0,0111	0,0001	-0,0120
		υ	1,0000	1,0050	1,0032	0,0014	0,0013	0,0000	0,0050
		ω	0,0250	0,0252	0,0252	0,0000	0,0000	0,0000	0,0002
		p_0	0,2000	0,1998	0,2003	0,0005	0,0005	0,0000	-0,0002
		p_1	0,6000	0,6022	0,6026	0,0006	0,0006	0,0000	0,0022
0,50	250	β_0	0,9050	0,9236	0,9196	0,0367	0,0364	0,0003	0,0186
		β_1	-1,4460	-1,4550	-1,4443	0,0562	0,0562	0,0001	-0,0090
		υ	1,0000	1,0161	1,0113	0,0065	0,0062	0,0003	0,0161
		ω	0,0250	0,0255	0,0252	0,0000	0,0000	0,0000	0,0005
		p_0	0,2000	0,1986	0,1968	0,0017	0,0017	0,0000	-0,0014
		p_1	0,6000	0,5960	0,5967	0,0022	0,0022	0,0000	-0,0040
	1000	β_0	0,9050	0,9088	0,9060	0,0100	0,0100	0,0000	0,0038
		β_1	-1,4460	-1,4483	-1,4460	0,0143	0,0143	0,0000	-0,0023
		υ	1,0000	1,0044	1,0020	0,0014	0,0014	0,0000	0,0044
		ω	0,0250	0,0251	0,0250	0,0000	0,0000	0,0000	0,0001

continua na próxima página

Tabela 5.1 – continuação

Gama	N	Parâmetro	Valor	Média	Mediana	EQM	Var	Vício^2	Vício
		p_0	0,2000	0,1999	0,2000	0,0005	0,0005	0,0000	-0,0001
		p_1	0,6000	0,5991	0,6000	0,0006	0,0006	0,0000	-0,0009
0,75	250	β_0	1,1390	1,1624	1,1578	0,0531	0,0525	0,0005	0,0234
		β_1	-1,6130	-1,6443	-1,6394	0,0665	0,0655	0,0009	-0,0313
		υ	1,0000	1,0163	1,0089	0,0063	0,0061	0,0003	0,0163
		ω	0,0250	0,0255	0,0251	0,0000	0,0000	0,0000	0,0005
		p_0	0,2000	0,1988	0,1966	0,0018	0,0018	0,0000	-0,0012
		p_1	0,6000	0,6007	0,6007	0,0024	0,0024	0,0000	0,0007
	1000	β_0	1,1390	1,1468	1,1470	0,0134	0,0133	0,0001	0,0078
		β_1	-1,6130	-1,6220	-1,6210	0,0164	0,0163	0,0001	-0,0090
		υ	1,0000	1,0045	1,0040	0,0015	0,0014	0,0000	0,0045
		ω	0,0250	0,0252	0,0250	0,0000	0,0000	0,0000	0,0002
		p_0	0,2000	0,1994	0,1990	0,0005	0,0005	0,0000	-0,0006
		p_1	0,6000	0,6000	0,6000	0,0006	0,0006	0,0000	0,0000
1,00	250	β_0	1,3860	1,4105	1,4110	0,0684	0,0678	0,0006	0,0245
		β_1	-1,7920	-1,8139	-1,8230	0,0830	0,0826	0,0005	-0,0219
		υ	1,0000	1,0182	1,0110	0,0070	0,0066	0,0003	0,0182
		ω	0,0250	0,0257	0,0250	0,0000	0,0000	0,0000	0,0007
		p_0	0,2000	0,1993	0,1960	0,0017	0,0017	0,0000	-0,0007
		p_1	0,6000	0,5984	0,5970	0,0026	0,0026	0,0000	-0,0016
	1000	β_0	1,3860	1,3833	1,3793	0,0159	0,0159	0,0000	-0,0027
		β_1	-1,7920	-1,7934	-1,7947	0,0186	0,0186	0,0000	-0,0014
		υ	1,0000	1,0026	1,0028	0,0014	0,0014	0,0000	0,0026
		ω	0,0250	0,0252	0,0251	0,0000	0,0000	0,0000	0,0002
		p_0	0,2000	0,2012	0,2011	0,0004	0,0004	0,0000	0,0012
		p_1	0,6000	0,6009	0,6007	0,0006	0,0006	0,0000	0,0009

Nos casos em que considerou-se γ desconhecido, adotou-se o procedimento sugerido por Zeng *et al.* (2006) de ajustar o modelo para diversos valores deste parâmetro e escolher aquele que resultasse no maior AIC.

Considerando cerca de 25% de censura, foi feita uma simulação com 1000 réplicas de amostras pequenas ($n = 250$) e moderadas ($n = 1000$), e valores de γ entre 0 e 1. A Tabela 5.2 apresenta medidas descritivas das estimativas dos parâmetros. Nota-se que em alguns casos o parâmetro da distribuição da fragilidade não é bem estimado, o que resulta em um aumento da variabilidade dos coeficientes da regressão, possivelmente acompanhada de vício não desprezível.

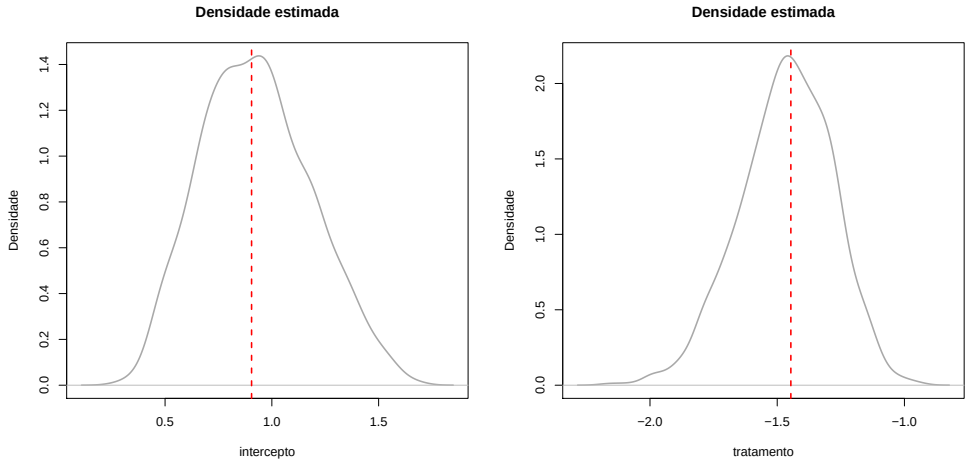

Figura 5.4: Densidade dos estimadores dos coeficientes de regressão.

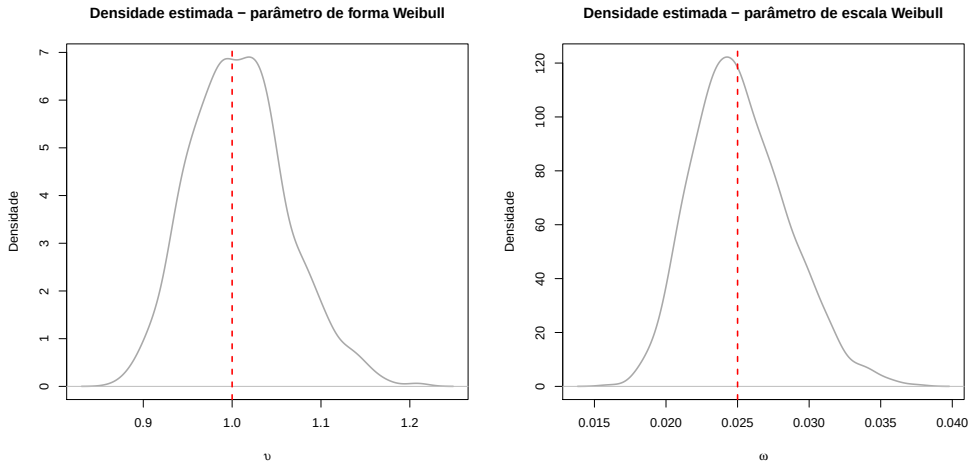

Figura 5.5: Densidade dos estimadores dos parâmetros de forma ($\hat{\upsilon}$) e de escala ($\hat{\omega}$) da distribuição weibull dos tempos latentes.

As Figuras (5.4) e (5.5) apresentam, respectivamente, a densidade dos estimadores dos coeficientes da regressão e dos parâmetros da distribuição dos tempos latentes obtidos das amostras

de tamanho $n = 1000$ simuladas considerando $gamma = 0,50$. Observe que as distribuições aparentam uma suave assimetria, o que pode ser a causa dos vícios apresentados na Tabela 5.2.

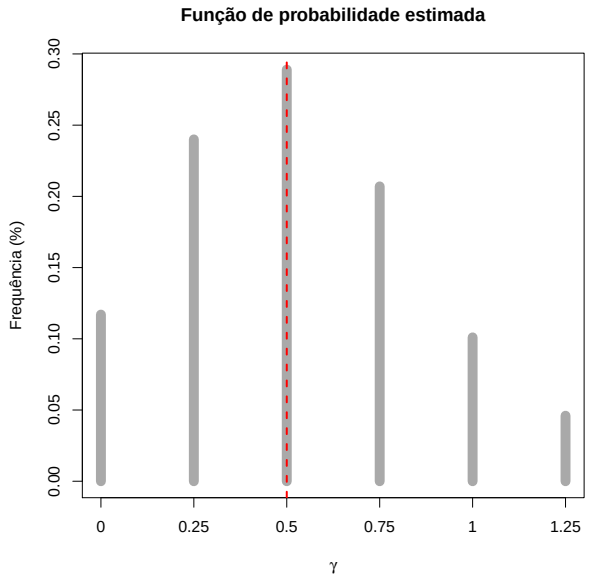

Figura 5.6: Função de probabilidade do estimador de γ.

O parâmetro da distribuição do termo de fragilidade não é bem estimado para amostras menores, porém, para amostras de tamanho $n = 1000$, os vícios induzidos nos demais parâmetros tornam-se desprezíveis. A Figura (5.6) apresenta a distribuição dos valores de γ que resultaram nos melhores modelos para os dados simulados.

É importante destacar que, mesmo com o erro associado à escolha do melhor modelo, as estimativas das frações de cura não sofrem muita influência em termos de vício, mesmo para amostras pequenas, como pode ser visto na Figura (5.7)

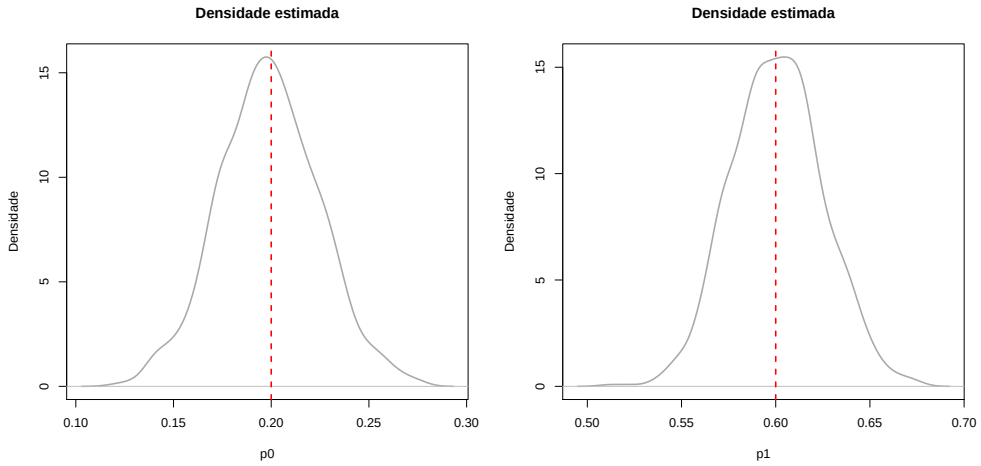

Figura 5.7: Densidade dos estimadores das frações de cura.

Tabela 5.2: Resultados das simulações para γ estimado e fração de censura de 25%

Gama	**N**	**Parâmetro**	**Valor**	**Média**	**Mediana**	**EQM**	**Var**	**Vício^2**	**Vício**
0,00	250	β_0	0,4760	0,6741	0,5986	0,1345	0,0952	0,0393	0,1981
		β_1	-1,1480	-1,2777	-1,2444	0,0898	0,0729	0,0169	-0,1297
		υ	1,0000	1,0459	1,0390	0,0115	0,0094	0,0021	0,0459
		ω	0,0250	0,0236	0,0234	0,0000	0,0000	0,0000	-0,0014
		γ	0,0000	0,2705	0,0000	0,2235	0,1503	0,0732	0,2705
		p_0	0,2000	0,2089	0,2072	0,0023	0,0022	0,0001	0,0089
		p_1	0,6000	0,5995	0,5984	0,0025	0,0025	0,0000	-0,0005
	1000	β_0	0,4760	0,5581	0,5155	0,0314	0,0247	0,0068	0,0821
		β_1	-1,1480	-1,1969	-1,1869	0,0184	0,0160	0,0024	-0,0489
		υ	1,0000	1,0187	1,0151	0,0023	0,0020	0,0003	0,0187
		ω	0,0250	0,0243	0,0245	0,0000	0,0000	0,0000	-0,0007
		γ	0,0000	0,1238	0,0000	0,0581	0,0427	0,0153	0,1238
		p_0	0,2000	0,2060	0,2056	0,0006	0,0005	0,0000	0,0060
		p_1	0,6000	0,5995	0,5992	0,0006	0,0006	0,0000	-0,0005
0,25	250	β_0	0,6840	0,7937	0,7153	0,1398	0,1277	0,0121	0,1097
		β_1	-1,2910	-1,3526	-1,3376	0,0823	0,0785	0,0038	-0,0616
		υ	1,0000	1,0357	1,0294	0,0105	0,0093	0,0013	0,0357

continua na próxima página

Tabela 5.2 – continuação

Gama	N	Parâmetro	Valor	Média	Mediana	EQM	Var	Vício^2	Vício
		ω	0,0250	0,0248	0,0246	0,0000	0,0000	0,0000	-0,0002
		γ	0,2500	0,4048	0,2500	0,2259	0,2020	0,0239	0,1548
		p_0	0,2000	0,2076	0,2068	0,0024	0,0023	0,0001	0,0076
		p_1	0,6000	0,5978	0,5968	0,0024	0,0024	0,0000	-0,0022
	1000	β_0	0,6840	0,7207	0,6928	0,0482	0,0468	0,0014	0,0367
		β_1	-1,2910	-1,3086	-1,2937	0,0246	0,0243	0,0003	-0,0176
		υ	1,0000	1,0106	1,0084	0,0029	0,0028	0,0001	0,0106
		ω	0,0250	0,0249	0,0248	0,0000	0,0000	0,0000	-0,0001
		γ	0,2500	0,3002	0,2500	0,0887	0,0862	0,0025	0,0502
		p_0	0,2000	0,2019	0,2017	0,0007	0,0007	0,0000	0,0019
		p_1	0,6000	0,5984	0,5994	0,0006	0,0006	0,0000	-0,0016
0,50	250	β_0	0,9050	0,9748	0,9512	0,1649	0,1600	0,0049	0,0698
		β_1	-1,4460	-1,4927	-1,4541	0,1021	0,0999	0,0021	-0,0467
		υ	1,0000	1,0268	1,0212	0,0108	0,0101	0,0007	0,0268
		ω	0,0250	0,0256	0,0252	0,0000	0,0000	0,0000	0,0006
		γ	0,5000	0,5718	0,5000	0,2274	0,2223	0,0051	0,0718
		p_0	0,2000	0,2013	0,2008	0,0022	0,0022	0,0000	0,0013
		p_1	0,6000	0,5987	0,5990	0,0022	0,0022	0,0000	-0,0013
	1000	β_0	0,9050	0,9281	0,9157	0,0666	0,0661	0,0005	0,0231
		β_1	-1,4460	-1,4662	-1,4578	0,0335	0,0331	0,0004	-0,0202
		υ	1,0000	1,0081	1,0058	0,0031	0,0030	0,0001	0,0081
		ω	0,0250	0,0252	0,0249	0,0000	0,0000	0,0000	0,0002
		γ	0,5000	0,5182	0,5000	0,1083	0,1080	0,0003	0,0182
		p_0	0,2000	0,1989	0,1982	0,0007	0,0007	0,0000	-0,0011
		p_1	0,6000	0,6007	0,6003	0,0006	0,0006	0,0000	0,0007
0,75	250	β_0	1,1390	1,1450	1,1710	0,1605	0,1604	0,0000	0,0060
		β_1	-1,6130	-1,6370	-1,6404	0,0997	0,0991	0,0006	-0,0240
		υ	1,0000	1,0186	1,0141	0,0094	0,0091	0,0003	0,0186
		ω	0,0250	0,0266	0,0259	0,0000	0,0000	0,0000	0,0016
		γ	0,7500	0,7183	0,7500	0,2199	0,2189	0,0010	-0,0317
		p_0	0,2000	0,1944	0,1938	0,0023	0,0022	0,0000	-0,0056
		p_1	0,6000	0,6024	0,6011	0,0023	0,0023	0,0000	0,0024

continua na próxima página

Tabela 5.2 – continuação

Gama	N	Parâmetro	Valor	Média	Mediana	EQM	Var	Vício^2	Vício
	1000	β_0	1,1390	1,1513	1,1480	0,0687	0,0685	0,0001	0,0123
		β_1	-1,6130	-1,6237	-1,6219	0,0353	0,0352	0,0001	-0,0107
		υ	1,0000	1,0063	1,0045	0,0032	0,0031	0,0000	0,0063
		ω	0,0250	0,0252	0,0249	0,0000	0,0000	0,0000	0,0002
		γ	0,7500	0,7640	0,7500	0,1070	0,1068	0,0002	0,0140
		p_0	0,2000	0,2004	0,1999	0,0007	0,0007	0,0000	0,0004
		p_1	0,6000	0,6008	0,6009	0,0006	0,0006	0,0000	0,0008
1,00	250	β_0	1,3860	1,2798	1,3193	0,1750	0,1637	0,0113	-0,1062
		β_1	-1,7920	-1,7438	-1,7423	0,1105	0,1082	0,0023	0,0482
		υ	1,0000	1,0007	0,9980	0,0091	0,0091	0,0000	0,0007
		ω	0,0250	0,0273	0,0263	0,0000	0,0000	0,0000	0,0023
		γ	1,0000	0,8388	1,0000	0,2314	0,2054	0,0260	-0,1612
		p_0	0,2000	0,1916	0,1923	0,0025	0,0024	0,0001	-0,0084
		p_1	0,6000	0,6033	0,6049	0,0025	0,0025	0,0000	0,0033
	1000	β_0	1,3860	1,3673	1,3929	0,0591	0,0588	0,0004	-0,0187
		β_1	-1,7920	-1,7807	-1,7812	0,0328	0,0327	0,0001	0,0113
		υ	1,0000	0,9997	1,0024	0,0026	0,0026	0,0000	-0,0003
		ω	0,0250	0,0253	0,0252	0,0000	0,0000	0,0000	0,0003
		γ	1,0000	0,9635	1,0000	0,0919	0,0905	0,0013	-0,0365
		p_0	0,2000	0,1969	0,1986	0,0007	0,0007	0,0000	-0,0031
		p_1	0,6000	0,5997	0,5999	0,0006	0,0006	0,0000	-0,0003

A fim de avaliar o quanto o percentual de censuras na amostra afeta as estimativas dos parâmetros do modelo, foi feita uma simulação considerando que cerca de 50% da amostra estava censurada. As distribuições dos estimadores dos coeficientes da regressão apresentaram desvios mais evidentes da normalidade, mesmo para as amostras de tamanho 1000. Tais estimadores apresentaram um vício consideravelmente maior quando comparados àqueles do caso em que se tinha 25% de censuras na amostra.

Uma parcela do vício observado nos estimadores é devida ao problema de estimação do parâmetro da distribuição da fragilidade, que conforme observado na Figura (5.10) apresentou grande vari-

abilidade em torno do verdadeiro valor $\gamma = 0,50$. Os vícios são particularmente elevados para o intercepto (β_0) quando a amostra é de tamanho $n = 250$, como pode ser observado na Tabela 5.3.

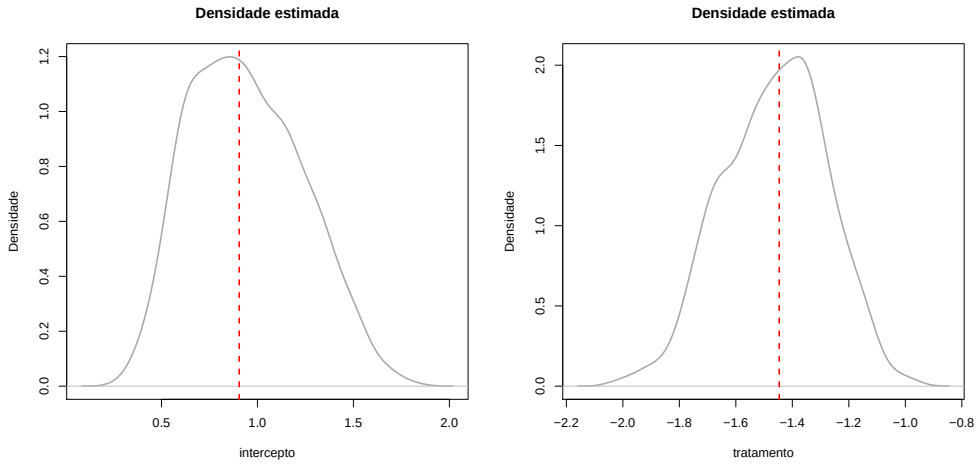

Figura 5.8: Densidade dos estimadores dos coeficientes de regressão.

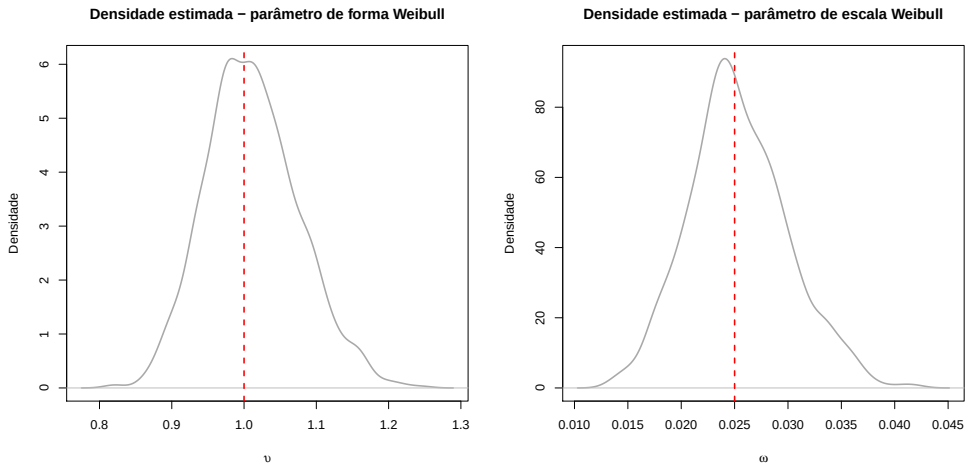

Figura 5.9: Densidade dos estimadores dos parâmetros de forma ($\hat{\upsilon}$) e de escala ($\hat{\omega}$) da distribuição weibull dos tempos latentes.

As irregularidades apresentadas na densidade estimada dos coeficientes são devidas ao parâmetro de suavização usado na estimação da densidade e não necessariamente à verdadeira forma da distribuição dos estimadores. Destaca-se que, embora os coeficientes da regressão sejam claramente

afetados pela dificuldade de estimar parâmetro γ, os estimadores dos parâmetros da distribuição dos tempos latentes são aparentemente robustos e não apresentam desvios de regularidade ou vício considerável. A Figura (5.9) mostra as densidades estimadas dos parâmetros da distribuição Weibull postulada para os tempos associados aos fatores latentes.

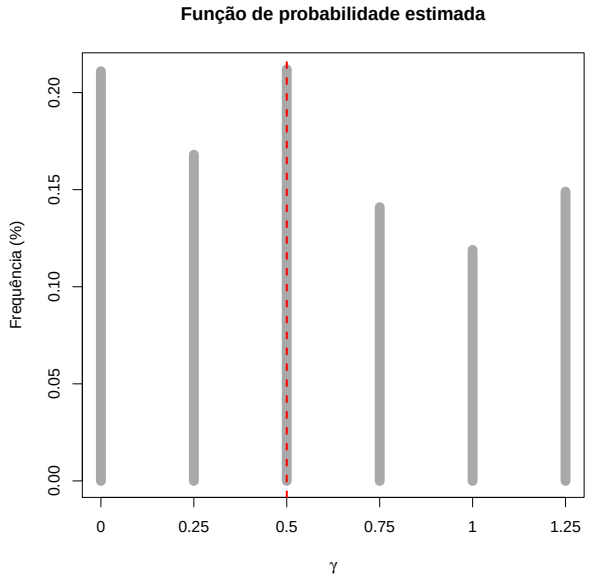

Figura 5.10: Função de probabilidade estimada do $\hat{\gamma}$.

Cabe destacar ainda que, mesmo com os problemas de estimação do parâmetro da fragilidade, as estimativas das frações de cura estão consideravelmente próximas dos valores verdadeiros usados na simulação.

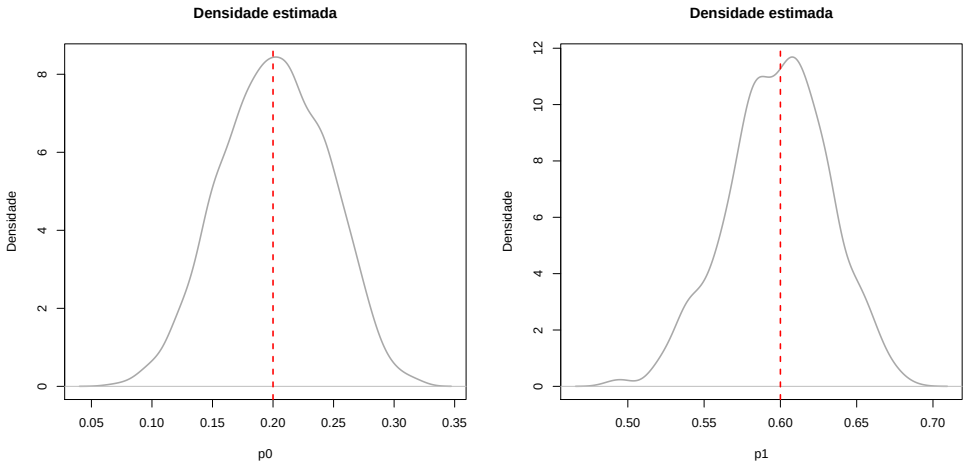

Figura 5.11: Densidade dos estimadores das frações de cura.

Tabela 5.3: Resultados das simulações para γ estimado e fração de censura de 50%

Gama	**N**	**Parâmetro**	**Valor**	**Média**	**Mediana**	**EQM**	**Var**	**Vício^2**	**Vício**
0,00	250	β_0	0,4760	0,7132	0,6520	0,2089	0,1526	0,0563	0,2372
		β_1	-1,1480	-1,2717	-1,2552	0,0986	0,0832	0,0154	-0,1237
		υ	1,0000	1,0584	1,0517	0,0160	0,0126	0,0034	0,0584
		ω	0,0250	0,0241	0,0235	0,0001	0,0001	0,0000	-0,0009
		γ	0,0000	0,3565	0,0000	0,3575	0,2304	0,1271	0,3565
		p_0	0,2000	0,2184	0,2171	0,0059	0,0055	0,0003	0,0184
		p_1	0,6000	0,5926	0,5955	0,0045	0,0045	0,0001	-0,0074
	1000	β_0	0,4760	0,5984	0,5484	0,0618	0,0468	0,0150	0,1224
		β_1	-1,1480	-1,2100	-1,2039	0,0266	0,0227	0,0039	-0,0620
		υ	1,0000	1,0248	1,0231	0,0037	0,0030	0,0006	0,0248
		ω	0,0250	0,0241	0,0239	0,0000	0,0000	0,0000	-0,0009
		γ	0,0000	0,1918	0,0000	0,1232	0,0864	0,0368	0,1918
		p_0	0,2000	0,2103	0,2088	0,0016	0,0015	0,0001	0,0103
		p_1	0,6000	0,5968	0,5956	0,0010	0,0010	0,0000	-0,0032
0,25	250	β_0	0,6840	0,8204	0,7551	0,1976	0,1789	0,0187	0,1364
		β_1	-1,2910	-1,3663	-1,3454	0,1045	0,0988	0,0057	-0,0753
		υ	1,0000	1,0428	1,0374	0,0146	0,0127	0,0018	0,0428

continua na próxima página

Tabela 5.3 – continuação

Gama	N	Parâmetro	Valor	Média	Mediana	EQM	Var	Vício^2	Vício
		ω	0,0250	0,0255	0,0248	0,0001	0,0001	0,0000	0,0005
		γ	0,2500	0,4492	0,2500	0,2986	0,2589	0,0397	0,1992
		p_0	0,2000	0,2112	0,2103	0,0056	0,0055	0,0001	0,0112
		p_1	0,6000	0,5963	0,6016	0,0045	0,0045	0,0000	-0,0037
	1000	β_0	0,6840	0,7565	0,7063	0,0717	0,0664	0,0053	0,0727
		β_1	-1,2910	-1,3246	-1,3080	0,0280	0,0268	0,0011	-0,0336
		υ	1,0000	1,0169	1,0139	0,0039	0,0036	0,0003	0,0169
		ω	0,0250	0,0248	0,0248	0,0000	0,0000	0,0000	-0,0002
		γ	0,2500	0,3660	0,2500	0,1551	0,1417	0,0135	0,1160
		p_0	0,2000	0,2060	0,2060	0,0017	0,0017	0,0000	0,0060
		p_1	0,6000	0,5978	0,6004	0,0011	0,0011	0,0000	-0,0022
0,50	250	β_0	0,9050	0,9977	0,9600	0,2003	0,1917	0,0086	0,0927
		β_1	-1,4460	-1,5107	-1,4769	0,1092	0,1051	0,0041	-0,0647
		υ	1,0000	1,0352	1,0332	0,0148	0,0136	0,0012	0,0352
		ω	0,0250	0,0261	0,0253	0,0001	0,0001	0,0000	0,0011
		γ	0,5000	0,5818	0,5000	0,2833	0,2766	0,0067	0,0818
		p_0	0,2000	0,1977	0,1990	0,0051	0,0051	0,0000	-0,0023
		p_1	0,6000	0,5971	0,5997	0,0047	0,0046	0,0000	-0,0029
	1000	β_0	0,9050	0,9445	0,9190	0,0872	0,0856	0,0016	0,0395
		β_1	-1,4460	-1,4629	-1,4530	0,0342	0,0340	0,0003	-0,0169
		υ	1,0000	1,0115	1,0090	0,0043	0,0041	0,0001	0,0115
		ω	0,0250	0,0253	0,0250	0,0000	0,0000	0,0000	0,0003
		γ	0,5000	0,5590	0,5000	0,1856	0,1821	0,0035	0,0590
		p_0	0,2000	0,2014	0,2020	0,0019	0,0019	0,0000	0,0014
		p_1	0,6000	0,5984	0,6000	0,0011	0,0011	0,0000	-0,0016
0,75	250	β_0	1,1390	1,1715	1,1479	0,2534	0,2524	0,0010	0,0325
		β_1	-1,6130	-1,6236	-1,6013	0,1095	0,1094	0,0001	-0,0106
		υ	1,0000	1,0146	1,0134	0,0121	0,0119	0,0002	0,0146
		ω	0,0250	0,0265	0,0259	0,0001	0,0001	0,0000	0,0015
		γ	0,7500	0,6940	0,7500	0,2840	0,2809	0,0031	-0,0560
		p_0	0,2000	0,1870	0,1915	0,0057	0,0055	0,0002	-0,0130
		p_1	0,6000	0,5897	0,5979	0,0061	0,0060	0,0001	-0,0103

continua na próxima página

Tabela 5.3 – continuação

Gama	N	Parâmetro	Valor	Média	Mediana	EQM	Var	Vício^2	Vício
	1000	β_0	1,1390	1,1418	1,1534	0,0917	0,0917	0,0000	0,0028
		β_1	-1,6130	-1,6137	-1,6087	0,0379	0,0379	0,0000	-0,0007
		υ	1,0000	0,9997	0,9973	0,0039	0,0039	0,0000	-0,0003
		ω	0,0250	0,0254	0,0252	0,0000	0,0000	0,0000	0,0004
		γ	0,7500	0,7262	0,7500	0,1776	0,1770	0,0006	-0,0238
		p_0	0,2000	0,1936	0,1960	0,0020	0,0020	0,0000	-0,0064
		p_1	0,6000	0,5980	0,5994	0,0013	0,0013	0,0000	-0,0020
1,00	250	β_0	1,3860	1,3210	1,3252	0,2512	0,2469	0,0043	-0,0650
		β_1	-1,7920	-1,7506	-1,7453	0,1284	0,1267	0,0017	0,0414
		υ	1,0000	1,0102	1,0062	0,0121	0,0120	0,0001	0,0102
		ω	0,0250	0,0279	0,0272	0,0001	0,0001	0,0000	0,0029
		γ	1,0000	0,8158	1,0000	0,2802	0,2462	0,0339	-0,1842
		p_0	0,2000	0,1827	0,1832	0,0054	0,0051	0,0003	-0,0173
		p_1	0,6000	0,5923	0,5962	0,0058	0,0057	0,0001	-0,0077
	1000	β_0	1,3860	1,3375	1,3670	0,0893	0,0869	0,0024	-0,0485
		β_1	-1,7920	-1,7677	-1,7675	0,0368	0,0362	0,0006	0,0243
		υ	1,0000	0,9946	0,9988	0,0037	0,0037	0,0000	-0,0054
		ω	0,0250	0,0258	0,0254	0,0000	0,0000	0,0000	0,0008
		γ	1,0000	0,8902	1,0000	0,1678	0,1558	0,0120	-0,1098
		p_0	0,2000	0,1882	0,1921	0,0020	0,0019	0,0001	-0,0118
		p_1	0,6000	0,5987	0,5999	0,0013	0,0013	0,0000	-0,0013

5.3 Discussão

Um ponto bastante criticado em dados simulados está relacionado com a distribuição usada para a geração dos tempos de sobrevivência. A distribuição exponencial é muito utilizada pela facilidade de tratar analiticamente o procedimento de geração de amostras, possibilitando maior controle de algumas características desejáveis nos dados simulados. Porém, possui o inconveniente de resultar em dados muito dispersos devido à acentuada assimetria à direita da função de densidade, além de estar relacionada à uma taxa de falha constante ao longo do tempo. Zeng *et al.* (2006)

utilizam dados simulados de uma distribuição exponencial padrão tanto para os tempos de falha quanto para os tempos de censura, o que faz com que não se tenha muito controle das frações de curados e de censuras na amostra simulada. Mesmo com os inconvenientes relacionados ao uso da distribuição exponencial na geração dos dados de sobrevivência, optou-se por considerá-la nas simulações apresentadas neste trabalho por possibilitar, ainda que de forma imprecisa, controlar as probabilidades de cura e de censura de parcela suscetível.

Um outro ponto a ser destacado é a forma de simular as censuras de forma realística, levando-se em conta as suposições consideradas na formulação do modelo. A escolha arbitrária dos parâmetros da distribuição exponencial utilizada para a geração dos tempos de censura pode resultar em uma grande proporção de censuras em tempos relativamente baixos, o que levaria à subestimação da fração de cura, ou altos, o que diminuiria a proporção de censuras daqueles suscetíveis. Nas situações práticas, quando o interesse é investigar a proporção de curados em uma amostra, é razoável supor que o tempo de acompanhamento é longo o suficiente para observar a ocorrência da falha (ou possivelmente de censura) dos indivíduos suscetíveis. Indivíduos censurados após esse tempo limiar são indicativo de cura, enquanto censuras precoces não fornecem evidência sobre a condição de cura ou suscetibilidade, o que pode levar a uma subestimação da fração de curados.

A análise dos resultados obtidos nas simulações realizadas considerando o modelo de Yakovlev e Tsodikov (1996) e o modelo de fração de cura com fatores latentes competitivos e fragilidade permite que algumas conclusões sejam tiradas. Uma delas é que, quando se considera amostras com cerca de 25% de censuras, os coeficientes da regressão são bem estimados quando o parâmetro da distribuição da fragilidade γ é igual a 0. Mesmo com tamanhos de amostrais moderados ($n \approx 1000$) e percentual de censuras em torno de 50%, os coeficientes são razoavelmente bem estimados, mas nota-se um aumento em suas variâncias. No entanto, quando o termo de fragilidade é usado para simular as amostras, frequentemente o modelo ajustado não é o mesmo considerado na geração. Aparentemente existe uma tendência de subestimar o parâmetro γ da fragilidade, o que se reflete diretamente nas estimativas dos coeficientes da regressão, uma vez que estes estão funcionalmente relacionados à γ.

Assim sendo, mesmo para amostras de tamanho elevado, o modelo pode apresentar dificuldades em termos de identificabilidade, uma vez que a função de log-verossimilhança possui uma curvatura demasiadamente suave na direção de γ. Tal fenômeno foi relatado por Zeng *et al.* (2006), embora tenham sugerido que amostras maiores que $n = 1500$ seriam suficientes para a estimação de γ. O estudo de simulação apresentado pelos autores do trabalho anteriormente referenciado mostra que, quando ocorre um erro de especificação da transformação determinada pelo parâmetro γ, os coeficientes tornam-se viciados, sobretudo o intercepto β_0. Nos casos em que o parâmetro da

distribuição da fragilidade é tratado como conhecido, observa-se que os estimadores não apresentam vício considerável, mesmo para amostras de tamanho $n = 250$ e com 50% de censura.

É importante destacar também que estes coeficientes devem ser interpretados com muita cautela, uma vez que estão funcionalmente relacionados ao γ e às frações de cura. À rigor, deveriam ser interpretados em um dado instante de tempo, condicionalmente aos valores das covariáveis. A presença de censuras em proporções em torno de 50% a 60% não causam vício aparente aos estimadores, embora sejam responsáveis pelo aumento das respectivas variabilidades.

Capítulo 6

Aplicações

6.1 Dados do InCor

A insuficiência cardíaca congestiva (ICC) é uma condição caracterizada pela perda da capacidade de suprir a demanda sanguínea do organismo, debilitando o paciente e possivelmente, levando-o à óbito.

O Instituto do Coração (InCor) do Hospital das Clínicas da Universidade de São Paulo, referência na pesquisa de causas e tratamentos de pacientes diagnosticados com doenças cardíacas, desenvolveu um estudo de acompanhamento de pacientes com prognóstico de insuficiência cardíaca por um longo período, visando levantar as principais características relacionadas com a evolução do estado de saúde dos pacientes. Embora não haja cura efetiva para a insuficiência cardíaca, uma análise preliminar dos dados apresentada por Tunes da Silva *et al.* (2010) indica que os tratamentos disponíveis atualmente podem atenuar sua gravidade, propiciando aos pacientes um prognóstico otimista.

É importante destacar que os 18 anos de coleta de dados do estudo favorecem a hipótese de que alguns pacientes se comportam como se estivessem curados, o que motiva a aplicação do modelo de sobrevivência que leva em consideração esta possibilidade. Embora possa parecer inadequado o uso do modelo discutido no presente trabalho, serão apresentados neste capítulo os resultados da análise dos dados utilizando o modelo de fração de cura com fatores latentes competitivos, diferenciando estes dos fatores observáveis medidos pelas covariáveis. Tal abordagem é justificada pela complexidade das interações existentes entre os inúmeros fatores genéticos e ambientais envolvidos na causa e evolução da ICC.

Embora os pacientes tivessem sido acompanhados até a observação do óbito por qualquer que tenha sido a causa, será considerado que o mesmo, quando ocorrido, teria tido como causa fundamental complicações decorridas do agravamento da insuficiência cardíaca, pelo menos para fins ilustrativos.

6.1.1 Descrição do conjunto de dados

Os dados utilizados neste trabalho referem-se ao tempo de sobrevivência de pacientes diagnosticados com insuficiência cardíaca congestiva, atendidos pelo InCor no período compreendido entre janeiro de 1991 e fevereiro de 2010. No total, existem 1400 registros, dos quais 48 foram excluídos da análise por apresentarem inconsistência de dados.

Variáveis demográficas, antropométricas, biofísicas e bioquímicas consideradas pelos pesquisadores como importantes na investigação dos fatores associados à ICC também encontram-se disponíveis no banco de dados. Foram utilizadas na análise as seguintes covariáveis: idade, sexo, peso, altura, índice de massa corporal (IMC), cardiomiopatia ou etiologia, classe funcional, hipertensão arterial, etilismo, tabagismo, cor ou raça, diabetes, espessura do septo interventricular, espessura da parede posterior do ventrículo esquerdo, diâmetro diastólico do ventrículo esquerdo, fração de ejeção, diâmetro do átrio esquerdo, fração de ejeção do ventrículo esquerdo, ritmo cardíaco medido no eletrocardiograma, sorologia de doença de Chagas, dosagem de sódio, hemoglobina, contagem de leucócitos, colesterol, triglicérides, HDL, LDL, glicemia, creatinina e informações do histórico cirúrgico do paciente, como cateterismo e evento cirúrgico em geral, excetuando-se os transplantes.

Devido às diferentes escalas de mensuração das variáveis numéricas consideradas no estudo, optou-se por utilizá-las na forma categorizada assim como em Tunes da Silva *et al.* (2010), sendo que os valores faltantes foram agrupados em uma única categoria. A definição das categorias foi feita usando os valores de referência conhecidos na literatura médica e, para a idade, adotou-se uma categorização conveniente, de acordo com a distribuição dos valores observados na amostra. Para cada covariável, foi adotada como referência a faixa de normalidade, sempre que havia uma disponível e os efeitos relativos dos desvios dessa faixa de referência foram estimados pelos coeficientes do modelo. Os dados apresentam cerca de 50% de censuras, sendo que 20% dos tempos observados consistem de censuras que ocorrem depois de um longo período de acompanhamento, como pode ser visto na Figura (6.1), o que sugere a necessidade de ajuste de um modelo que leve em consideração a fração de curados. Pelo fato do estudo contemplar uma grande quantidade de covariáveis, foi adotado um procedimento de seleção de variáveis, que consistiu em ajustar o modelo de sobrevivência com fração de cura, sem o termo de fragilidade, considerando uma co-

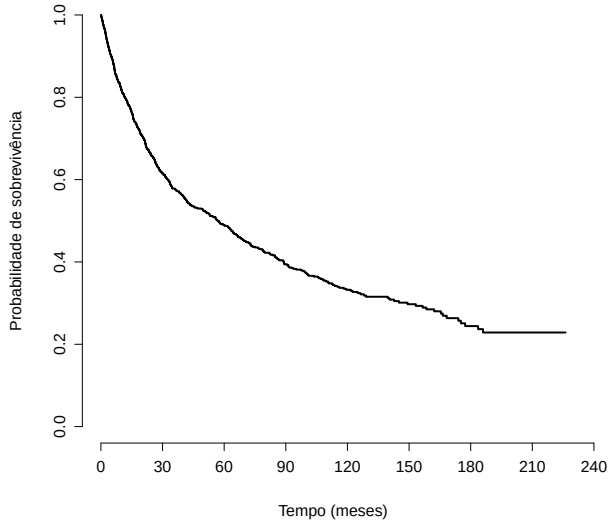

Figura 6.1: Função de sobrevivência de pacientes diagnosticados com Insuficiência Cardíaca Congestiva.

variável por vez, removendo aquelas não significativas ao nível de 15%. A Tabela B.1 no Apêndice B mostra os resultados do ajuste preliminar usado para a seleção de covariáveis do modelo e a Figura (6.2) mostra as curvas de sobrevivência de Kaplan-Meier para as covariáveis individualmente significativas na etapa de seleção anteriormente descrita. É importante destacar que o artifício de considerar os valores faltantes em uma categoria possibilita que tais registros não sejam removidos da amostra no momento da estimação do modelo. Como pode ser observado na Figura (6.2), as categorias originadas do agrupamento de registros com informações indisponíveis não encontram-se demasiadamente afastadas das demais curvas de sobrevivência. Com isso, essas categorias serão utilizadas na fase de estimação, mas não é de interesse obter coeficientes a elas associados.

Posteriormente, foi ajustado um modelo de fração de cura com as variáveis selecionadas na etapa anterior e aquelas não significativas ao nível de 10% foram retiradas. Considerou-se como significativas aquelas covariáveis em que pelo menos uma categoria apresentou diferença significativa da categoria de referência. A Tabela B.2 no Apêndice B mostra os resultados do ajuste do modelo composto pelas covariáveis individualmente significativas.

Para investigar se a inclusão de um termo de fragilidade poderia melhorar o ajuste do modelo, considerou-se o procedimento sugerido por Zeng *et al.* (2006) em que diversos valores do parâmetro de fragilidade foram utilizados, optando pelo modelo que resulta em maior AIC. Uma vez que os modelos ajustados possuíam o mesmo número de covariáveis, escolher o modelo com maior

Tabela 6.1: Valor da função de log-verossimilhança para diversos valores de γ.

γ	**log-verossimilhança**
0,0	-3599
0,2	-3599
0,4	-3598
0,6	-3597
0,8	-3598
1,0	-3601
1,2	-3606
1,4	-3604

AIC torna-se equivalente à escolher aquele que proporciona o maior valor para função de log-verossimilhança, desde que contenham a mesma quantidade de parâmetros estimados. A Tabela 6.1 mostra essas medidas para diversos valores do parâmetro da distribuição do termo de fragilidade. O critério proposto para a escolha do modelo mais adequado aos dados deve considerar a presença de um termo de fragilidade com distribuição gama com média unitária e parâmetro de escala $\gamma = 0,60$. Os resultados do ajuste desse modelo estão dispostos na Tabela B.3. Nota-se que, com a introdução do termo de fragilidade, as covariáveis *HDL* e *ECG* tornaram-se não significativas. Estas foram retiradas do modelo e os coeficientes associados às demais covariáveis foram recalculados.

Portanto, o modelo final é aquele que considera as seguintes covariáveis: idade, sexo, cardiomiopatia, classe funcional, hemoglobina e LDL, além de um termo de fragilidade. A Tabela 6.2 apresenta os coeficientes estimados e suas respectivas probabilidades de significância.

Tabela 6.2: Modelo de fatores latentes competitivos com termo de fragilidade $\gamma = 0,60$

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
Intercepto		-0,1736	0,7310	0,0564	0,8123
Idade (anos)	até 39				
	de 40 a 49	-0,0413	0,1290	0,1031	0,7482
	de 50 a 59	0,2711	0,1320	4,2288	0,0397
	60 ou mais	0,2948	0,1970	2,2408	0,1344
Sexo	Masculino				
	Feminino	-0,3390	0,1210	7,8522	0,0051
Cardiomiopatia	Idiopática				
	Isquêmica	0,0773	0,1400	0,3046	0,5810

continua na próxima página

Tabela 6.2 – continuação

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
	Alcóolica	-0,1237	0,2380	0,2707	0,6029
	Hipertensiva	-0,4444	0,1500	8,7354	0,0031
	Chagásica	0,8325	0,1320	39,883	0,0000
Classe funcional	I ou II				
	III	0,2535	0,1120	5,1351	0,0234
	IV	0,5113	0,1610	10,1231	0,0015
Hemoglobina	Normal				
	Baixa	-0,0533	0,1298	0,1677	0,6821
	Alta	-0,7285	0,4040	3,2541	0,0712
LDL	Ótimo				
	Desejável	-0,4351	0,1837	5,5881	0,0181
	Limítrofe	-0,1237	0,1920	0,4134	0,5203
	Alto	-0,2467	0,2080	1,4126	0,2346

A distribuição do tempo associado à cada fator latente deve ser considerada como sendo uma Weibull com parâmetros de forma e escala iguais a, respectivamente $\upsilon = 1,03$ e $\omega = 0,0080$, considerando a notação apresentada na expressão (5.9).

Destaca-se que deve ser ter cuidado ao interpretar os coeficientes como uma medida de relação entre a covariável e o risco de morte do paciente causada pela ICC, uma vez que esse risco varia ao longo do tempo.

Todavia, pode ser de interesse do pesquisador considerar o modelo com fração de cura sem o termo de fragilidade, contemplando as mesmas covariáveis a fim de interpretar os coeficientes à elas associados. A Tabela B.4 no Apêndice B apresenta os coeficientes do modelo de fatores latentes competitivos com riscos proporcionais.

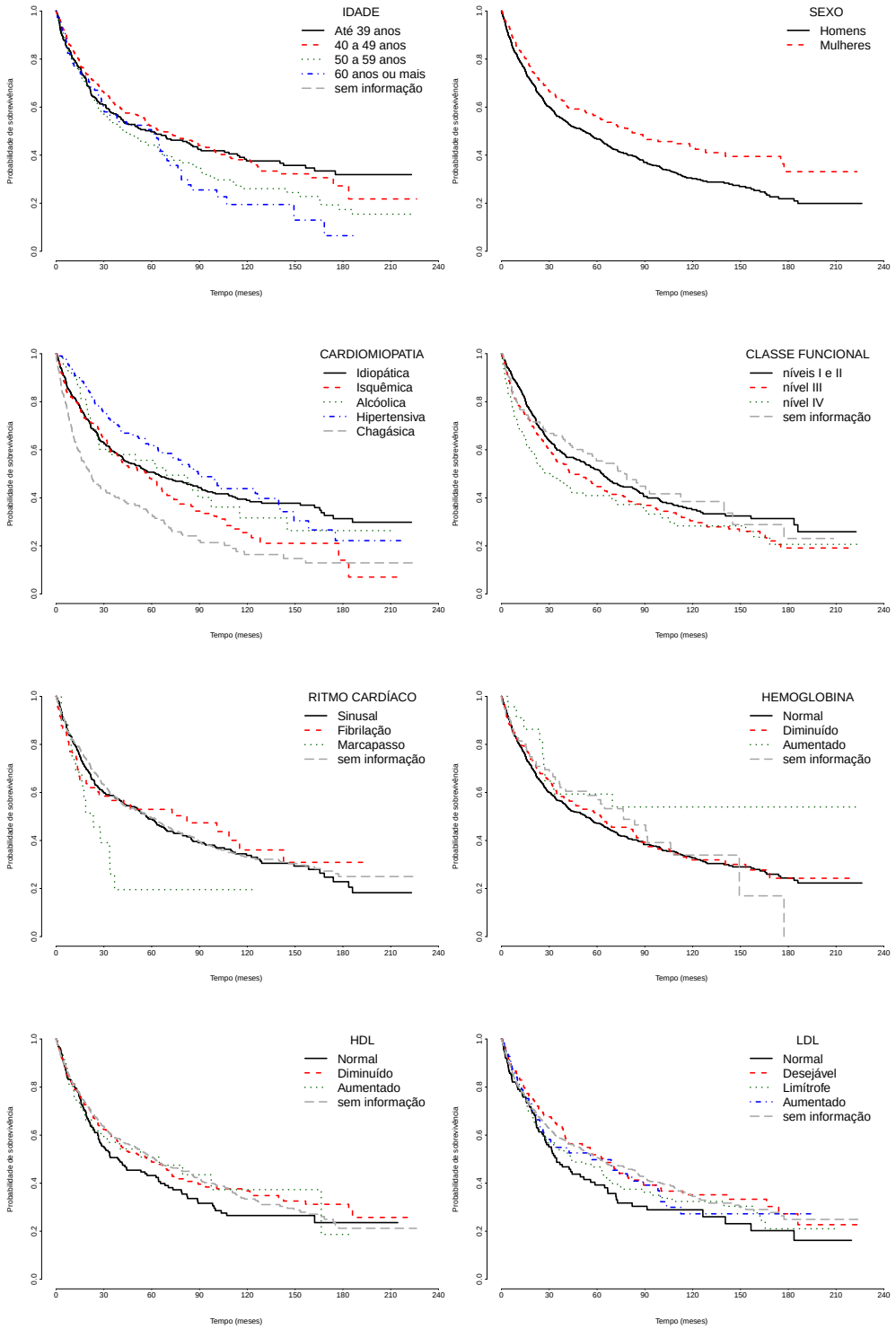

Figura 6.2: Funções de sobrevivência das covariáveis individualmente significativas.

Capítulo 7

Considerações Finais

No modelo de riscos proporcionais com fração de cura proposto por Yakovlev e Tsodikov (1996) apresentado no Capítulo 3, as covariáveis são relacionadas diretamente ao parâmetro da distribuição de Poisson, que por hipótese, representa o número médio de fatores latentes entre os pacientes de cada grupo - tratamento ou controle. Pela estrutura do modelo proposto, tais covariáveis não são relacionadas apenas à fração de cura, mas também aos tempos de sobrevivência dos pacientes não curados.

Surge naturalmente a ideia de relacionar as covariáveis tanto com a fração de curados através do parâmetro da Poisson quanto com a distribuição dos tempos de sobrevivência, o que pode causar problemas de identificabilidade e, portanto, deve ser melhor investigada.

No modelo discutido no Capítulo 4, a escolha da distribuição gama para o efeito aleatório foi feita de forma a tornar o problema tratável analiticamente, embora seja possível considerar outras distribuições para a fragilidade.

O modelo de sobrevivência com fatores latentes competitivos e fragilidade tende a proporcionar um melhor ajuste aos dados, uma vez que considera o efeito aleatório da fragilidade. Deve-se destacar, no entanto, que apenas os casos particulares em que o modelo apresenta a característica de proporcionalidade de riscos ou de chances ($\gamma = 0$ e $\gamma = 1$, respectivamente) possuem uma interpretação dos coeficientes da regressão como uma medida do efeito das covariáveis à eles associadas na sobrevida do paciente. Nos demais casos, a interpretação deve ser feita sempre condicionalmente a um instante de tempo previamente estabelecido.

Sobre o processo de estimação deste modelo, Zeng *et al.* (2006) destacam que a curvatura da função de log-verossimilhança na direção do parâmetro γ é pouco acentuada para os tamanhos

amostrais usuais, o que leva às dificuldades para estimar adequadamente o parâmetro de fragilidade. Com isso, nas ocasiões práticas, o pesquisador deve avaliar qual modelo atenderá de maneira mais adequada aos objetivos previamente estabelecidos. Naturalmente, muitas considerações devem ser feitas antes de se optar por um ou outro modelo, priorizando a precisão das estimativas de sobrevivência ou a interpretação dos coeficientes estimados.

7.1 Trabalhos Futuros

Diante da flexibilidade do modelo de cura com fatores latentes competitivos e fragilidade, poderia ser interessante discutir quando supor que a distribuição dos tempos latentes é do tipo exponencial, Weibull, ou mesmo não especificá-la, usando a abordagem semiparamétrica proposta por Zeng *et al.* (2006). Além disso, um fato importante de ser investigado é se devemos ou não considerar a super-dispersão usando o modelo de fragilidade, usando possivelmente um teste de Wald para o parâmetro γ que determina a transformação (4.5). Este teste, no caso, deveria considerar que a hipótese nula encontra-se na fronteira do espaço paramétrico, uma vez que $\gamma \geq 0$ e o interesse seria testar $H_0 : \gamma = 0$ contra $H_1 : \gamma > 0$.

Tunes da Silva *et al.* (2010) optaram por utilizar o modelo de Cox na análise dos dados de sobrevivência de pacientes atendidos pelo InCor, uma vez que esta abordagem possibilita a análise da qualidade do ajuste e das suposições necessárias ao adequado uso do modelo. Tal escolha evidencia a necessidade de métodos eficientes que permitam avaliar se as suposições relacionadas aos modelos com fração de cura estão ou não atendidas.

Uma proposta que surgiu ao longo da discussão deste trabalho foi a de avaliar a intensidade com que o erro de especificação do modelo pode afetar as estimativas dos coeficientes,em particular, no caso em o modelo de fração de cura com riscos proporcionais é ajustado utilizando dados gerados considerando o efeito de fragilidade.

Apêndice A

Proposições

Algumas etapas da obtenção de fórmulas importantes foram omitidas do texto para não efetar a fluidez da leitura, mas dada a relevância para a compreensão dos resultados usados, foram incluídas neste capítulo para servirem como uma referência.

A.1 Função de sobrevivência com fatores latentes competitivos e fragilidade

O modelo proposto por Zeng *et al.* (2006) pode ser obtido daquele proposto por Yakovlev e Tsodikov (1996) com a inserção de um termo de fragilidade com distribuição gama. A equação (A.1) mostra como o modelo (4.3) pode ser obtido.

$$\begin{aligned} S(t|\boldsymbol{x}) &= \int_0^\infty e^{\theta(\boldsymbol{x})F(t)\xi} \frac{1}{\gamma^{1/\gamma}\Gamma(1/\gamma)} \xi^{1/\gamma-1} \exp(-\xi/\gamma) d\xi \\ &= \int_0^\infty \frac{1}{\gamma^{1/\gamma}\Gamma(1/\gamma)} \xi^{1/\gamma-1} \exp\{-\xi/\gamma[1+\theta(\boldsymbol{x})F(t)\gamma] d\xi. \end{aligned} \tag{A.1}$$

Fazendo a transformação de variáveis $u = \xi[1+\theta(\boldsymbol{x})F(t)\gamma]$; $du = [1+\theta(\boldsymbol{x})F(t)\gamma]d\xi$, temos que:

$$
\begin{aligned}
S(t|\boldsymbol{x}) &= \int_0^\infty \frac{1}{\gamma^{1/\gamma}\Gamma(1/\gamma)} \left[\frac{u}{1+\theta(\boldsymbol{x})F(t)\gamma}\right]^{1/\gamma-1} e^{-u/\gamma} \frac{1}{1+\theta(\boldsymbol{x})F(t)\gamma} du \\
&= [1+\theta(\boldsymbol{x})F(t)\gamma]^{-1/\gamma} \int_0^\infty \frac{1}{\gamma^{1/\gamma}\Gamma(1/\gamma)} u^{1/\gamma-1} e^{-u/\gamma} du \\
&= [1+\theta(\boldsymbol{x})F(t)\gamma]^{-1/\gamma}. \qquad \text{(A.2)}
\end{aligned}
$$

Ressaltando que no último passo da expressão acima, tem-se a integral de uma densidade gama em todo o seu suporte, o que permite a resolução de forma trivial.

A.2 Equivalência com modelo de fatores latentes com distribuição Binomial Negativa

Um fato importante a ser destacado é a equivalência dos modelos de Zeng *et al.* (2006) e aquele proposto por De Castro *et al.* (2010a), no qual a distribuição postulada para o número de fatores latentes é uma binomial negativa. A seguir apresentamos a demonstração dessa equivalência.

Sejam W uma variável aleatória com distribuição binomial negativa e u tal que $0 < u < 1$. A função geradora de probabilidade de W avaliada no ponto u, representada pela notação $A_p(u)$, é dada por:

$$
\begin{aligned}
A_p(u) &= \sum_{k\geq 0} u^k P(W=k) \\
&= \sum_{k\geq 0} u^k \frac{\Gamma(\gamma^{-1}+k)}{k!\Gamma(\gamma^{-1})}\left(\frac{\gamma\theta}{1+\gamma\theta}\right)^k (1+\gamma\theta)^{-\gamma^{-1}} \\
&= \sum_{k\geq 0} u^k \frac{\Gamma(\gamma^{-1}+k)}{k!\Gamma(\gamma^{-1})}\left(\frac{\gamma\theta}{1+\gamma\theta}\right)^k \left(1-\frac{\gamma\theta}{1+\gamma\theta}\right)^{\gamma^{-1}} \\
&= \sum_{k\geq 0} \frac{\Gamma(\gamma^{-1}+k)}{k!\Gamma(\gamma^{-1})}\left(\frac{u\gamma\theta}{1+\gamma\theta}\right)^k \left(1-\frac{\gamma\theta}{1+\gamma\theta}\right)^{\gamma^{-1}} \\
&= \sum_{k\geq 0} \frac{\Gamma(\gamma^{-1}+k)}{k!\Gamma(\gamma^{-1})}\left(\frac{u\gamma\theta}{1+\gamma\theta}\right)^k \left(1-\frac{u\gamma\theta}{1+\gamma\theta}\right)^{\gamma^{-1}} \left(1-\frac{u\gamma\theta}{1+\gamma\theta}\right)^{-\gamma^{-1}} \left(1-\frac{\gamma\theta}{1+\gamma\theta}\right)^{\gamma^{-1}} \\
&= \{1+\gamma\theta(1-u)\}^{-\gamma^{-1}} \sum_{k\geq 0} \frac{\Gamma(\gamma^{-1}+k)}{k!\Gamma(\gamma^{-1})}\left(\frac{u\gamma\theta}{1+\gamma\theta}\right)^k \left(1-\frac{u\gamma\theta}{1+\gamma\theta}\right)^{\gamma^{-1}} \\
&= \{1+\gamma\theta(1-u)\}^{-\gamma^{-1}}, \qquad \text{(A.3)}
\end{aligned}
$$

sendo que a última igualdade é válida uma vez que o somatório da função de probabilidade da binomial negativa em todos os pontos de massa vale 1.

A equação apresentada em (A.3) pode ser avaliada no ponto $u = S(t)$, para $t \geq 0$, conforme apresentado em (A.4):

$$
\begin{aligned}
A_p\big(S(t)\big) &= \big\{1+\gamma\theta\big(1-S(t)\big)\big\}^{-\gamma^{-1}} \\
&\quad \{1+\gamma\theta F(t)\}^{-\gamma^{-1}}. \qquad \text{(A.4)}
\end{aligned}
$$

Se considerarmos que o parâmetro θ é uma função de covariáveis $\boldsymbol{x}$, digamos $\theta(\boldsymbol{x}) = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}$, tem-se a equivalência dos modelos.

Apêndice B

Seleção de covariáveis

Tabela B.1: Procedimento de seleção de covariáveis - ajustes univariados.

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
Idade (anos)	até 39				
	de 40 a 49	-0,0459	0,1010	0,2070	0,6489
	de 50 a 59	0,2191	0,1010	4,6780	0,0305
	60 ou mais	0,2510	0,1490	2,8490	0,0914
Sexo	Masculino				
	Feminino	-0,3040	0,0973	9,7300	0,0018
Peso	Normal				
	Baixo	0,1618	0,1200	1,8196	0,1770
	Alto	0,0365	0,1320	0,0758	0,7830
Altura	Normal				
	Baixo	-0,1725	0,1370	1,5880	0,2076
	Alto	-0,0429	0,1140	0,1410	0,7075
IMC	Normal				
	Baixo	0,0663	0,2160	0,0941	0,7590
	Sobrepeso	-0,1737	0,1160	2,2249	0,1358
	Obeso	-0,3893	0,1710	5,1563	0,0232
Cardiomiopatia	Idiopática				
	Isquêmica	0,2015	0,1080	3,4700	0,0626

continua na próxima página

Tabela B.1 – continuação

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
	Alcóolica	0,0011	0,1920	0,0000	0,9953
	Hipertensiva	-0,2507	0,1230	4,1600	0,0414
	Chagásica	0,6443	0,1000	41,4000	0,0000
Classe funcional	I ou II				
	III	0,1865	0,0892	4,3710	0,0366
	IV	0,3533	0,1233	8,2070	0,0042
Hipertensão	Não				
	Sim	-0,2380	0,1153	4,2800	0,0386
Etilismo	Ex-etilista				
	Etilista	0,2030	0,1238	2,6900	0,1010
Tabagismo	Não fumante				
	Fumante	0,1080	0,1240	0,7570	0,3840
	Ex-fumante	-0,3420	0,3650	0,8760	0,3490
Cor ou raça	Branca				
	Preta	-0,3100	0,2680	1,3400	0,2469
	Parda	-0,2480	0,2550	0,9500	0,3299
Diabetes	Não				
	Dep. insulina	-1,2382	0,7091	3,0493	0,0808
	Não dep. insulina	-0,0451	0,1790	0,0634	0,8012
Espessura do septo	Normal				
	Aumentada	-0,0975	0,1820	0,2880	0,5917
Espessura da parede	Normal				
	Aumentada	-0,0721	0,1870	0,1480	0,7005
Diâmetro diastólico VE	Normal				
	Aumentado	0,0379	0,1770	0,0461	0,8301
Fração de ejeção	Normal				
	Muito baixa	0,0416	0,2350	0,0313	0,8596
	Baixa	0,0627	0,2180	0,0827	0,7737
Diâmetro AE	Normal				
	Aumentado	-0,0932	0,1000	0,8600	0,3538

continua na próxima página

Tabela B.1 – continuação

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
Fração de ejeção VE	Normal				
	Muito baixa	0,1760	0,2680	0,4330	0,5100
	Baixa	0,2810	0,2470	1,2890	0,2560
ECG	Sinusal				
	Fibrilação	-0,0405	0,1742	0,0541	0,8161
	Marcapasso	0,6564	0,2766	5,6324	0,0176
Cateterismo	Não				
	Sim	-0,1240	0,1110	1,2400	0,2649
Sorologia Chagas	Não				
	Sim	-0,0957	0,1147	0,6970	0,4038
Sódio	Normal				
	Baixo	0,1452	0,1050	1,9034	0,1680
	Alto	-0,0783	0,4110	0,0364	0,8490
Hemoglobina	Normal				
	Baixa	-0,0577	0,1030	0,3160	0,5737
	Alta	-0,6067	0,3360	3,2620	0,0709
Leucócitos	Normal				
	Baixo	0,4985	0,2925	2,9050	0,0883
	Alto	-0,0586	0,1425	0,1690	0,6808
Colesterol	Ótimo				
	Limítrofe	0,0629	0,1165	0,2920	0,5890
	Alto	-0,1764	0,1495	1,3940	0,2380
Triglicérides	Ótimo				
	Limítrofe	-0,1007	0,1623	0,3850	0,5350
	Alto	0,1295	0,1386	0,8730	0,3500
HDL	Ótimo				
	Baixo	-0,1840	0,1160	2,5090	0,1132
	Alto	-0,1500	0,2080	0,5190	0,4712
LDL	Ótimo				
	Desejável	-0,3040	0,1460	4,3140	0,0378
	Limítrofe	-0,1260	0,1520	0,6940	0,4049

continua na próxima página

Tabela B.1 – continuação

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
	Alto	-0,1920	0,1640	1,3630	0,2429
Glicemia	Normal				
	Inapropriada	-0,0533	0,1009	0,2790	0,5980
	Elevada	-0,0829	0,1276	0,4220	0,5160
Creatinina	Normal				
	Baixa	0,3105	0,2408	1,6620	0,1970
	Alta	-0,0585	0,0875	0,4470	0,5040
Evento cirúrgico	Não				
	Sim	0,0087	0,2040	0,0018	0,9660

Tabela B.2: Significância das covariáveis no modelo múltiplo.

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
Intercepto		-1,2098	0,8921	1,8400	0,1751
Idade (anos)	até 39				
	de 40 a 49	-0,0004	0,1031	0,0000	0,9970
	de 50 a 59	0,2828	0,1050	7,2600	0,0071
	60 ou mais	0,2338	0,1570	2,2200	1,3630
Sexo	Masculino				
	Feminino	-0,2453	0,0981	6,2500	0,0124
Cardiomiopatia	Idiopática				
	Isquêmica	0,1467	0,1115	1,7300	0,1879
	Alcóolica	-0,0725	0,1939	0,1400	0,7086
	Hipertensiva	-0,2889	0,1249	5,3500	0,0207
	Chagásica	0,6269	0,1016	38,1000	0,0000
Classe funcional	I ou II				
	III	0,2216	0,0900	6,0600	0,0138
	IV	0,3783	0,1221	9,5900	0,0020
ECG	Sinusal				
	Fibrilação	0,0563	0,1746	0,1040	0,7469
	Marcapasso	0,7112	0,2692	6,9800	0,0082
Hemoglobina	Normal				
	Baixa	-0,0148	0,1032	0,0207	0,8857
	Alta	-0,7672	0,3539	4,7000	0,0302
HDL	Ótimo				
	Baixo	-0,2484	0,1206	4,2400	0,0395
	Alto	-0,2079	0,2068	1,0100	0,3148
LDL	Ótimo				
	Desejável	-0,3726	0,1452	6,5800	0,0103
	Limítrofe	-0,1549	0,1532	1,0200	0,3120
	Alto	-0,3002	0,1691	3,1500	0,0759

Tabela B.3: Modelo de fatores latentes competitivos com termo de fragilidade $\gamma = 0,60$.

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
Intercepto		-0,3193	0,7520	0,1801	0,6713
Idade (anos)	até 39				
	de 40 a 49	-0,0497	0,1280	0,1498	0,6987
	de 50 a 59	0,2728	0,1320	4,2919	0,0383
	60 ou mais	0,2803	0,1960	2,0355	0,1537
Sexo	Masculino				
	Feminino	-0,3522	0,1210	8,5079	0,0035
Cardiomiopatia	Idiopática				
	Isquêmica	0,1002	0,1390	0,5159	0,4726
	Alcóolica	-0,1414	0,2380	0,3535	0,5521
	Hipertensiva	-0,3973	0,1500	7,0482	0,0079
	Chagásica	0,8318	0,1320	39,9605	0,0000
Classe funcional	I ou II				
	III	0,2692	0,1120	5,7564	0,0164
	IV	0,5269	0,1610	10,7763	0,0010
ECG	Sinusal				
	Fibrilação	0,0863	0,2180	0,1559	0,6929
	Marcapasso	0,3789	0,3690	1,0517	0,3051
Hemoglobina	Normal				
	Baixa	-0,0442	0,1300	0,1160	0,7334
	Alta	-0,6544	0,3990	2,6942	0,1007
HDL	Ótimo				
	Baixo	-0,2076	0,1500	1,9199	0,1659
	Alto	-0,2350	0,2600	0,8142	0,3669
LDL	Ótimo				
	Desejável	-0,4005	0,1850	4,6785	0,0305
	Limítrofe	-0,1215	0,1950	0,3869	0,5339
	Alto	-0,2033	0,2130	0,9066	0,3410

Tabela B.4: Modelo de fatores latentes competitivos de riscos proporcionais.

Covariável	**Nível**	**Estimativa**	**Erro padrão**	χ^2	**Valor-p**
Intercepto		-0,6649	0,6487	1,0505	0,3054
Idade (anos)	até 39				
	de 40 a 49	0,0113	0,1040	0,0118	0,9134
	de 50 a 59	0,2614	0,1056	6,1332	0,0133
	60 ou mais	0,2824	0,1552	3,3106	0,0688
Sexo	Masculino				
	Feminino	-0,2581	0,0992	6,7730	0,0093
Cardiomiopatia	Idiopática				
	Isquêmica	0,1200	0,1128	1,1327	0,2872
	Alcóolica	-0,0614	0,1951	0,0991	0,7529
	Hipertensiva	-0,3050	0,1255	5,9084	0,0151
	Chagásica	0,6337	0,1019	8,6410	0,0000
Classe funcional	I ou II				
	III	0,2057	0,0900	5,2157	0,0224
	IV	0,3332	0,1241	7,2105	0,0072
Hemoglobina	Normal				
	Baixa	-0,0449	0,1041	0,1861	0,6662
	Alta	-0,6822	0,3435	3,9446	0,0470
LDL	Ótimo				
	Desejável	-0,3139	0,1470	4,5575	0,0328
	Limítrofe	-0,0434	0,1527	0,0808	0,7762
	Alto	-0,1821	0,1657	1,2075	0,2718

Referências Bibliográficas

Bakar, M. R. A., Salah, K. A., Ibrahim, N. A. e Haron, K. (2008). Cure fraction, modelling and estimating in a population-based cancer survival analysis, *Malasyan Journal of Mathematical Sciences* **2**: 113–134. 5

Barajas, L. E. N. e Yin, G. (2008). Bayesian semiparametric cure rate model with an unknown threshold, *Scandinavian Journal of Statistics* **35**: 540–556. 5, 20

Bennett, S. (1983). Analysis of survival data by the proportional odds model, *Statistics in Medicine* **2**: 273–277. 25

Berkson, J. e Gage, R. (1952). Survival curve for cancer patients following treatment, *Journal of the American Statistical Association* **47**: 501–515. 4, 10, 15, 24

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society* **11**: 15–44. 4

Chang, I. S., Hsiung, C. A., Wen, C. C., Wu, Y. J. e Yang, C. C. (2007). Non-parametric maximum-likelihood estimation in a semiparametric mixture model for competing-risks data, *Scandinavian Journal of Statistics* **34**: 870–895. 6

Chen, M. H., Ibrahim, J. G. e Sinha, D. (1999). A new bayesian model for survival data with a surviving fraction, *Journal of the American Statistical Association* **94**: 909–919. 4, 16, 17

Chen, M. H., Ibrahim, J. G. e Sinha, D. (2002). Bayesian inference for multivariate survival data with a cure fraction, *Journal of Multivariate Analysis* **80**: 101–126. 6

Cheng, S. C., Wei, L. J. e Ying, Z. (1995). Analysis of transformation models with censored data, *Biometrika* **82**: 835–845. 26

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65**: 141–151. 3

Cooner, F., Banerjee, S. e McBean, A. M. (2006). Modelling geographically referenced survival data with a cure fraction, *Statistical Methods in Medical Research* **15**: 307–324. 5

Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal statistical Society, Series B* **34**: 187–220. 2, 10, 14

De Castro, M., Cancho, V. G. e Rodrigues, J. (2010a). A note on unified approach for cure rate models, *Brazilian Journal of Probability and Statistics* **24**: 100–103. 5, 65

De Castro, M., Rodrigues, J. e Cancho, V. G. (2010b). A hands-on approach for fitting long-term survival models under the gamlss framework, *Computer Methods and Programs in Biomedicine* **97**: 168–177. 7

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors, *Biometrics* **38**: 1041–1046. 20

Hernandez-Quintero, A., Dupuy, J. F. e Escarela, G. (2009). Analysis of a semiparametric mixture model for competing risks, *Annuals of the Institute of Statistical Mathematics*. 16

Kalbfleisch, J. D. e Prentice, R. L. (2002). *The Statistical Analysis of failure time data*, second edition edn, Wiley, New York. 14

Kaplan, E. L. e Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**: 457–481. 2

Kim, S., Chen, M. H., Dey, D. K. e Gamerman, D. (2007). Bayesian dynamic models for survival data with a cure fraction, *Lifetime Data Analysis* **13**: 17–35. 5

Klebanov, L. B. e Yakovlev, A. Y. (2007). A new approach to testing for sufficient follow-up in cure-rate analysis, *Journal of Statistical Planning and Inference* **137**: 3557–3567. 19

Li, C. S., Taylor, J. M. G. e Sy, J. P. (2001). Identifiability of cure models, *Statistics & Probability Letters* **54**: 389–395. 6, 20

Maller, R. A. e Zhou, S. (1996). *Survival Analysis With Long Term Survivors*, first edn, Wiley, New York. 4, 16, 19, 23

Mizoi, M. F., Pedroso de Lima, A. C. e Bolfarine, H. (2007). Cure rate model with measurement error, *Communications in Statistics - Simulation and Computation* **36**: 185–196. 5

Peng, Y. (2003). Fitting semiparametric cure models, *Computational Statistics & Data Analysis* **41**: 481–490. 6, 23

Peng, Y. e Zhang, J. (2008). Identifiability of a mixture cure frailty model, *Statistics and Probability Letters* **78**: 2604–2608. 6, 20

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks, *Journal of the Royal Statistical Society* **44**: 234–243. 25

R Development Core Team (2004). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3. 35

Rodrigues, J., Cancho, V. G., De Castro, M. e Louzada-Neto, F. (2009). On the unification of long-term survival models, *Statistics and Probability Letters* **79**: 753–759. 6

Sumathi, K. e Aruna Rao, K. (2008). Cure rate models - a partial review with an application to recurrent event or count data, *InterStat*. 5, 19

Therneau, T. e Grambsch, P. (2000). *Modeling Survival Data - Extending the Cox Model*, first edn, Springer, New York. 14, 23

Tunes da Silva, G., Bhering, F. L. e Cesar, R. S. (2010). Relatório de análise estatística sobre o projeto: Prognóstico em portadores de insuficiência cardíaca em uma coorte com até 18 anos de seguimento, *Technical Report RAE-CEA-10P08*, IME-USP, São Paulo. 56, 57, 63

Wienke, A. (2007). *Frailty Models in survival Analysis*, PhD thesis, Institut für Medizinische Epidemiologie, Biometrie und Informatik - Universitãt Halle-Wittenberg. 3

Yakovlev, A. Y. e Tsodikov, A. D. (1996). *Stochastic Models of Tumor Latency and Their Bioestatistical Applications*, first edn, World Scientific, Singapore. 4, 10, 17, 20, 24, 36, 54, 62, 64

Yashin, A. I., Iachine, I. A., Begun, A. Z. e Vaupel, J. W. (2001). Hidden frailty: Myths and reality, *Technical Report 34*, Department of Statistics and Demography, SDU - Odense University. 3

Yin, G. e Ibrahim, J. G. (2005). A general class of bayesian survival models with zero and nonzero cure fractions, *Biometrics* **61**: 403–412. 6

Yu, B. (2008). A frailty mixture cure model with application to hospital readmission data, *Biometrical Journal* **50**: 386–394. 5

Yu, B. e Peng, Y. (2008). Mixture cure models for multivariate survival data, *Computational Statistics & Data Analysis* **52**: 1524–1532. 6

Zeng, D., Yin, G. e Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction, *Journal of the American Statistical Association* **101**: 670–694. 6, 7, 9, 24, 27, 29, 32, 34, 39, 43, 53, 54, 58, 62, 63, 64, 65

Zhou, S. e Maller, R. A. (1995). The likelihood ratio test for the presence of immunes in a censored sample, *Statistics* **27**: 181–201. 19