MODELOS DE SOBREVIVÊNCIA COM FRAÇÃO DE CURA E OMISSÃO NAS COVARIÁVEIS

Renata Santana Fonseca

Dissertação apresentada ao Corpo Docente do Programa de Pós-Graduação em Matemática Aplicada e Estatística - CCET - UFRN, como requisito parcial para obtenção do título de Mestre em Matemática Aplicada e Estatística.

Área de Concentração: Probabilidade e Estatística

Orientador: Profª Drª Dione Maria Valença

Natal, março de 2009

Modelos de Sobrevivência com Fração de Cura e Omissão nas Covariáveis

Este exemplar corresponde à redação final da dissertação devidamente corrigida, defendida por Renata Santana Fonseca e aprovada pela comissão julgadora.

Natal, março de 2009

BANCA EXAMINADORA:

Profª Drª Dione Maria Valença (orientadora) - DEST - UFRN

Profª Drª Jeanete Alves Moreira - DEST - UFRN

Prof. Dr. Damião Nóbrega da Silva - DEST - UFRN

Profª Drª Silvia Maria de Freitas - DEMA - UFC

Dedicatória

Dedico este trabalho a todos aqueles que choraram saudades junto comigo.

Agradecimentos

Tudo é do Pai, toda Honra e toda Glória, é dele a vitória alcançada em minha vida.

Agradeço à minha orientadora, Dione pela orientação, paciência e por ter me mostrado o prazer de se fazer um trabalho como este.

Tenho muito que agradecer aos meus pais, Antonio e Edneide pelos valores transmitidos em todo o percurso da minha vida e pelo apoio incondicional em todas as horas.

Ao meu eterno amigo, companheiro e, mais do que nunca, grande amor, Gilmar pela compreensão, enorme paciência e dedicação que me foram mostrados ao longo destes dois anos.

Meus irmãos, Liu, Nando e Bia, sobrinha, Pam, cunhados, Tom e Jú e amigas Cíntia, Nivea, Marília, Mariese e Relva, obrigada pelos momentos de descontração em Salvador, boas recargas de ânimo pra continuar a jornada.

Aos amigos de Natal. Allan, Cecílio, Manassés, Patrícia e Renatinha. Sem vocês eu não seria nada, não teria chegado até o fim. Obrigada pelo apoio nas horas em que pensava em desistir, pelas poucas, porém gostosas farras que fizemos e pela verdadeira amizade. Com vocês aprendi muito sobre convivência e respeito, mas não posso esquecer de pedir desculpas pelas grosserias.

Aos professores da UFRN, Damião, Jeanete, Pledson, Carla e Paulo pela receptividade, pelos conselhos, dicas e principalmente por acreditarem no meu potencial e me incentivarem a continuar com os estudos.

Aos colegas do PPGMAE, Hermes, pela enciclopédia que carrega e sempre tem uma resposta pra todas as perguntas, Lenilson, Neto, Daniel, Tatiana, Enai, Camila, Aparecida e Moisés pelos consolos, pelas soluções de exercícios e programas.

Ao secretário Rafael e ex-secretário Paulo pela disposição em ajudar e esclarecer dúvidas, e a todos os funcionários do CCET.

Ao professor Heleno Bolfarine pela sugestão do tema.

À Cristiane Fernandes (Cris), a primeira pessoa que conheci em Natal, que mesmo sem saber quem eu era, me orientou e ajudou nos primeiros dias.

E à CAPES pelo apoio financeiro.

Agradeço a todos que de certa forma fizeram possível a execução deste trabalho.

Resumo

Neste trabalho estudamos o modelo de sobrevivência com fração de cura proposto por Yakovlev et al. (1993) que possui uma estrutura de riscos competitivos. Covariáveis são introduzidas para modelar o número médio de riscos e permitimos que algumas destas covariáveis apresentem omissão. Consideramos apenas os casos em que as covariáveis omissas são categóricas e as estimativas dos parâmetros são obtidas através do algoritmo *EM* ponderado. Apresentamos uma série de simulações para confrontar as estimativas obtidas através deste método com as obtidas quando se exclui do banco de dados as observações que apresentam omissão, conhecida como análise de casos completos. Avaliamos também através de simulações, o impacto na estimativa dos parâmetros quando aumenta-se o percentual de curados e de censura entre indivíduos não curados. Um conjunto de dados reais referentes ao tempo até a conclusão do curso de estatística na Universidade Federal do Rio Grande do Norte é utilizado para ilustrar o método.

Palavras - chave: Análise de Sobrevivência; Fração de cura; variáveis omissas; Algoritmo *EM*.

Abstract

In this work we study the survival cure rate model proposed by Yakovlev (1993) that are considered in a competing risk setting. Covariates are introduced for modeling the cure rate and we allow some covariates to have missing values. We consider only the cases by which the missing covariates are categorical and implement the EM algorithm via the method of weights for maximum likelihood estimation. We present a Monte Carlo simulation experiment to compare the properties of the estimators based on this method with those estimators under the complete case scenario. We also evaluate, in this experiment, the impact in the parameter estimates when we increase the proportion of immune and censored individuals among the not immune one. We demonstrate the proposed methodology with a real data set involving the time until the graduation for the undergraduate course of Statistics of the Universidade Federal do Rio Grande do Norte.

Key - words: Survival analysis; Rate cure; Missing data; *EM* algorithm .

Sumário

Capítulo 1

Introdução

Neste trabalho abordamos aspectos teóricos, computacionais e aplicados de análise estatística para modelar dados de sobrevivência extraídos de uma população na qual uma parcela dos indivíduos está imune à ocorrência do evento (ou são considerados curados). Além disso, estudamos um procedimento para tratar a ocorrência de omissão em covariáveis.

Faremos neste capítulo, uma breve introdução aos conceitos básicos de análise de sobrevivência, à modelagem para dados com fração de cura e ao problema de dados omissos, seguida de uma revisão bibliográfica com algumas propostas encontradas na literatura para tratar de dados com tais características.

1.1 Conceitos básicos de análise de sobrevivência

A análise de sobrevivência é um conjunto de técnicas estatísticas que servem para analisar dados correspondentes ao tempo até ocorrência de determinado evento; como, por exemplo, o tempo até a morte ou o tempo até a cura de um paciente, ou ainda, o tempo até a falha de um equipamento eletrônico.

Estes métodos foram originalmente designados para estudos de mortalidade, explicando, portanto, o nome "sobrevivência". No entanto, a aplicabilidade destas técnicas se estende a diversas áreas do conhecimento. Na engenharia, onde recebe o nome

de "confiabilidade", busca-se estudar o tempo até a falha de um equipamento, na criminologia o interesse pode ser o tempo até um ex-detento reincidir no crime, na educação, o tempo até a conclusão de um curso. Na literatura refere-se a este tempo como *tempo de falha*.

Seja T uma variável aleatória contínua não negativa representando o tempo de vida, $f(t)$ a sua correspondente função densidade e $F(t)$ a função distribuição acumulada. Em análise de sobrevivência, existe um especial interessse na probabilidade de um indivíduo sobreviver pelo menos até o tempo t, conhecida como função de sobrevivência que é dada por

$$S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t), \qquad \text{para } t > 0. \tag{1.1}$$

Esta é uma função monótona decrescente com as seguintes propriedades:

(i) $S(0) = 1$;

(ii) $\lim_{t\to\infty} S(t) = 0$.

Funções de sobrevivência que não satisfazem à propriedade (ii) são denominadas funções de sobrevivência impróprias ou com fração de cura, ou ainda de longa duração (Rodrigues, Cancho, & Castro 2008) e são objeto de estudo desta dissertação.

Outra função de interesse na análise de sobrevivência é a função risco, que especifica a taxa de falha instantânea no tempo t, definida como

$$h(t) = \lim_{\Delta t\to 0^+} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \tag{1.2}$$

As funções de densidade, de sobrevivência e de risco são matematicamente relacionadas. Assim, conhecendo uma delas pode-se obter as outras através das seguintes relações:

$$h(t) = -\frac{d \log S(t)}{dt} = \frac{f(t)}{S(t)}, \tag{1.3}$$

$$f(t) = -\frac{dS(t)}{dt}. \tag{1.4}$$

Uma propriedade da função risco que define uma importante classe de modelos é a proporcionalidade dos riscos, ou seja, para duas unidades observacionais distintas i e j, a razão entre os riscos $h_i(t)/h_j(t) = k$, sendo k constante para todo tempo t. Por exemplo, se $k = 3$, dizemos que o risco (de falha) do indivíduo i é três vezes o risco do indivíduo j em todo o período de acompanhamento. Um modelo cuja função risco satisfaz esta condição é chamado de modelo de riscos porporcionais.

A principal característica de dados de sobrevivência é a presença de *censura*, que é uma observação parcial da resposta (tempo de falha). Isto ocorre devido a perda de acompanhamento do indivíduo, seja porque o paciente morreu de causa diferente da estudada, mudou de cidade, ou porque, ao final do estudo, o evento não foi observado. Sem a presença de censura, as técnicas estatísticas clássicas, como análise de regressão e planejamento de experimento, poderiam ser utilizadas na análise deste tipo de dados, provavelmente usando uma transformação para a resposta. No entanto, se houver censuras, tais técnicas não podem ser utilizadas (Colosimo & Giolo 2006). Neste caso, sabe-se que o tempo entre o início do estudo e a ocorrência do evento é maior do que o observado, o que caracteriza censura à direita. No entanto, o tempo durante o qual o indivíduo esteve em observação é aproveitado. Dentre os mecanismos de censura existentes podemos citar:

- a **censura tipo I** em que experimento é realizado em um período de tempo pré-fixado, de forma que o tempo de vida do indivíduo é conhecido apenas se ocorrer antes do final do estudo;

- a **censura tipo II** em que o estudo é realizado até que o evento ocorra um número pré-estabelecido de vezes e

- a **censura aletória**, que é a mais freqüente em estudos reais, caracteriza-se pela ocorrência de interrupções aleatórias no acompanhamento dos indivíduos (por exemplo a perda de acompanhamento do paciente que se mudou ou que morreu por outra causa). Neste caso os tempos de censuras são representados por

variáveis aleatórias. Quando a distribuição da censura não envolve parâmetros de interesse ao estudo, dizemos que a censura é não infomativa.

1.2 Fração de cura

Em modelos tradicionais de tempos de falha, assume-se que em dado momento o evento de interesse irá ocorrer para todos os indivíduos observados após um espaço de tempo razoável, isto é, todos os indivíduos estão em risco durante a realização do estudo.

No entanto, em determinados experimentos, alguns indivíduos podem nunca apresentar o evento pois estão curados ou são considerados *imunes* ao evento. Se, por exemplo, o evento de interesse é a recorrência de determinado tipo de câncer, após aplicação de tratamentos, uma parte dos indivíduos em estudo pode não apresentar o retorno da doença (daí a denominação "fração de cura"). Dados com estas características podem surgir em outros contextos. Por exemplo, em estudos na área de criminologia um possível evento de interesse é o tempo até um ex-detento reincidir no crime, entretanto, alguns deles estarão reabilitados e não apresentarão tal evento. Na demografia, em estudos sobre tempo até o divórcio, alguns casais podem nunca experimentar o evento.

Modelar estes dados ignorando a existência de uma parcela de curados ou imunes na população pode conduzir o pesquisador a conclusões distorcidas, ao passo que quando esta característica é incorporada no modelo, pode-se saber, por exemplo, qual tratamento resulta em uma maior proporção de curados, ou que tratamento é mais eficiente para determinado perfil de paciente.

Um grande número de observações censuradas à direita pode ser um indicativo da presença de indivíduos imunes na população. Neste caso, Maller & Zhou (1996) recomendam um tempo de acompanhamento suficientemente grande. Por exemplo, em estudos sobre a recidiva de determinado tipo de câncer, muitos pesquisadores consideram que um paciente estará curado se não houver reincidência da doença em um período de 5 a 10 anos após aplicação do tratamento, sendo este tempo determinado

pela experiência do pesquisador.

A presença de imunes na população pode ser identificada através do gráfico da função de sobrevivência empírica conhecida como Kaplan-Meier ou estimador produto-limite (Kaplan & Meier 1958). Se a cauda direita apresenta um nível aproximadamente constante e estritamente maior do que zero durante um período de tempo razoável, caracterizando uma função de sobrevivência imprópria (função que não converge para zero a medida que o tempo cresce), há indícios da presença de imunes.

A Figura 1.1 mostra a curva de sobrevivência estimada para um conjunto de dados reais referente ao tempo até a conclusão do curso de graduação em Estatística da UFRN. Observa-se que a cauda direita alcança um patamar acima de zero por um período considerável, o que retrata o comportamento de uma função de sobrevivência imprópria.

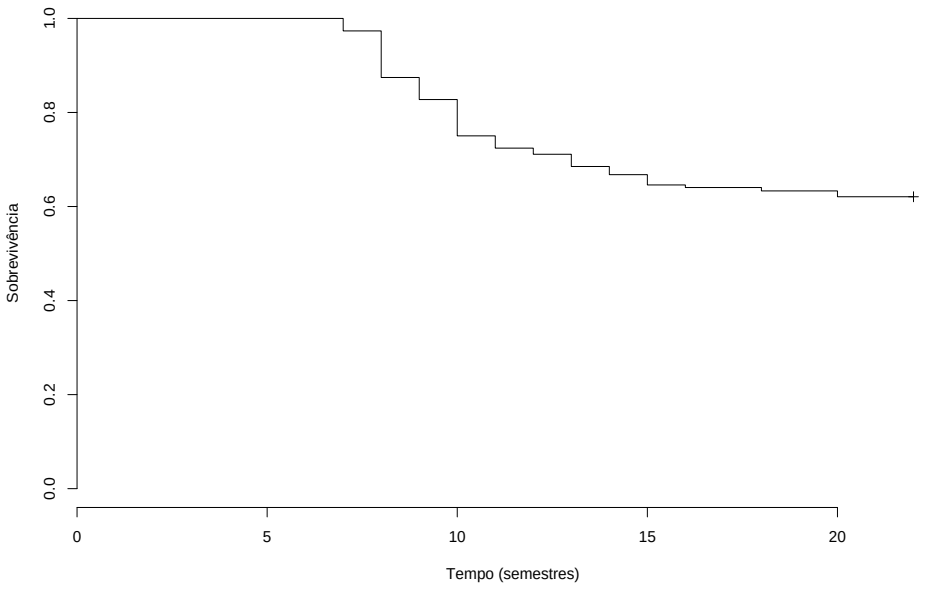

Figura 1.1: Função de Sobrevivência estimada para os dados de tempo até a conclusão do curso de graduação em Estatística da UFRN no período de 1997 a 2004. Amostra com $n = 414$ alunos.

Uma abordagem bastante popular para modelar dados com fração de cura é considerar uma mistura de distribuições, na qual uma representa o tempo de sobrevivência

da população não curada e outra é dada por uma distribuição degenerada com tempos infinitos para os imunes (Boag 1949; Berkson & Gage 1952). Neste modelo, conhecido como *modelo de mistura padrão*, é assumido que uma fração π $(0 < \pi < 1)$ da população está curada, e a restante $1 - \pi$, não está curada. Utilizando uma partição em imunes (I) e não imunes (NI), com $P(\text{I}) = \pi$ e $P(\text{NI}) = 1-\pi$ a função de sobrevivência para a população, denotada por $S_{pop}(t)$ para este modelo, é dada por

$$\begin{aligned} S_{pop}(t) &= P(T > t) \\ &= P(T > t|\text{I})P(\text{I}) + P(T > t|\text{NI})P(\text{NI}) \\ &= \pi + (1 - \pi)S^*(t), \end{aligned} \tag{1.5}$$

em que $P(T > t|\text{NI}) = S^*(t)$ denota a função de sobrevivência para a população não imune e $P(T > t|\text{I}) = 1$ para todo $t > 0$.

Algumas desvantagens deste modelo são apontadas por Chen, Ibrahim, & Sinha (1999). Na presença de covariáveis relacionadas com o parâmetro π, o modelo de mistura padrão não possui a propriedade de riscos proporcionais. Além disso, no contexto bayesiano, as distribuições a priori não informativas impróprias implicam, em geral, em distribuições a posteriori impróprias.

Neste trabalho, discutimos um modelo alternativo com estrutura de riscos competitivos[1] proposto por Yakovlev et al. (1993) e Chen, Ibrahim, & Sinha (1999), que é referido em Rodrigues, Cancho, & Castro (2008) como *modelo de tempo de promoção*. Este modelo supera tais desvantagens e será descrito com detalhes no Capítulo 2.

Modelos de sobrevivência com fração de cura têm sido extensivamente discutidos e aplicados a dados reais na literatura estatística por vários autores nos últimos anos. Técnicas bayesianas para estimação dos parâmetros do modelo de tempo de promoção são dadas em Ibrahim, Chen, & Sinha (2001). Utilizando o modelo de tempo de promoção, Zaider et al. (2001) estudam o efeito de diferentes doses de radiação sobre a fração de curados em pacientes com câncer de próstata. A proposta de Yin & Ibrahim

[1]Riscos competitivos: várias causas ou riscos podem provocar o evento de interesse.

(2005) consiste em uma classe de modelos que naturalmente liga uma família de funções de sobrevivência próprias e impróprias. Um modelo unificado, que inclui o modelo de mistura padrão e o modelo de tempo de promoção como dois casos especiais, é discutido em Rodrigues & Louzada-Neto (2006). O modelo de tempo de promoção foi estendido por Mizoi (2007) para a situação em que há erro de medição nas covariáveis. Uma abordagem semiparamétrica que permite utilizar dados correlacionados no modelo de mistura padrão é discutida em Peng et al. (2007). Um teste para avaliar a suficiência do tempo de acompanhamento em uma ampla classe de modelos de fração de cura foi proposto por Klebanov & Yakovlev (2007).

A inclusão de informações concomitantes no modelo de fração de cura é de grande importância para descrever a heterogeneidade da população. Unidades com características diferentes certamente apresentarão tempos de sobrevida e chance de cura diferentes. Contudo, algumas destas informações podem não ser observadas para todos os indivíduos, resultando em algumas lacunas na matriz de dados. Na próxima seção, descrevemos o problema e características importantes em dados com omissões nas covariáveis.

1.3 O Problema de dados omissos

A presença de omissão em covariáveis é um problema frequente na análise estatística e pode ser causada por diversos fatores, como por exemplo, a recusa de um indivíduo a fornecer alguma informação ou a morte de um paciente antes de realizar todos os exames. Em um experimento industrial, algumas covariáveis podem não ser observadas em equipamentos quebrados devido a causas não relacionadas ao processo experimental.

A abordagem mais simples para esses casos consiste em ignorar as unidades que apresentam omissão e analisar, de maneira usual, apenas os dados completamente observados, supondo que constituem uma amostra aleatória da população de interesse. Essa prática é conhecida como *Análise de Casos Completos* (ACC) e, dependendo do

percentual de dados omissos, pode causar perda de informação e resultar em estimativas tendenciosas e ineficientes. Portanto, é importante o desenvolvimento de métodos que incorporem os dados omissos nas análises. Uma revisão sobre os métodos existentes para ajustar modelos de regressão utilizando este tipo de dado é abordada em Little (1992), que destaca maior eficiência dos métodos de imputação múltipla.

Muitas propostas vêm sendo discutidas para tratar dados omissos em diversos contextos. Diversas técnicas para análise de dados que apresentam omissão são abordadas por Roderick, Little, & Rubin (1987). Quando as covariáveis omissas são categóricas, Ibrahim (1990) propõem utilizar o algoritmo *EM ponderado* (Dempster et al. 1977) para obter estimativas dos parâmetros em modelos lineares generalizados. Esta técnica foi extendida por Ibrahim, Chen, & Lipsitz (1999) para covariáveis contínuas ou mistas (contínuas e categóricas), em que é implementada a versão Monte Carlo do algoritmo *EM* (Tanner 1996). O algoritmo *EM* ponderado é descrito em detalhes por Horton & Laird (1999) que, por meio de exemplos, ilustram sua aplicação discutindo vantagens e limitações. Na área de sobrevivência, modelos com fração de cura, efeitos aleatórios permitindo omissão não-ignorável nas covariáveis é discutido por Herring & Ibrahim (2002).

Neste trabalho seguimos a proposta dada por Chen & Ibrahim (2001) para estimação dos parâmetros do modelo de tempo de promoção quando as covariáveis apresentam omissão. No entanto, diferentemente de Chen & Ibrahim (2001), que usam a verossimilhança para dados completos, neste trabalho, seguindo a idéia de Mizoi (2007), utilizamos a verossimilhança marginal, que elimina variáveis latentes, sendo o algoritmo *EM* utilizado para imputar apenas os dados omissos. Será considerada aqui apenas a situação em que as covariáveis omissas são categóricas, e assim utilizamos o algoritmo *EM* ponderado para obter estimativas dos parâmetros. Além disso, apresentamos uma série de estudos de simulação para avaliar propriedades desse procedimento quando as covariáveis omissas são categóricas.

1.3.1 Mecanismos de omissão

Para analisar dados omissos deve-se levar em conta o processo que causa as omissões, conhecido como *mecanismos de omissão*, originalmente descritos em Rubin (1976). Em particular, é necessário considerar se a omissão nas covariáveis está relacionada com as variáveis no conjunto de dados.

Seja $\mathbf{X}$ a matrix de covariáveis, $\mathbf{T}$ o vetor de tempos até a falha e $\mathbf{C}$ o vetor de tempos de censura. Se houver omissão, pode-se particionar a matriz de covariáveis $\mathbf{X}$ em $(\mathbf{X}_{obs}, \mathbf{X}_{mis})$, em que $\mathbf{X}_{mis}$ é uma matriz de covariáveis com omissão e $\mathbf{X}_{obs}$ é uma matriz de covariáveis completamente observadas.

Considere que, para a j-ésima covariável, $j = 1, \ldots, p$, tem-se a variável aleatória indicadora de omissão $\mathbf{R}_j$, que vale 1 se $\mathbf{X}_j$ é omissa e 0 caso contrário.

O mecanismo de omissão é caracterizado pela distribuição condicional de $\mathbf{R}_j$ dado $(\mathbf{X}, \mathbf{T}, \mathbf{C})$, digamos $P(\mathbf{R}_j|\mathbf{X}, \mathbf{T}, \mathbf{C}; \boldsymbol{\tau})$, sendo $\boldsymbol{\tau}$ um vetor de parâmetros desconhecidos.

Se as omissões não dependem dos valores de $(\mathbf{X}, \mathbf{T}, \mathbf{C})$, omissos ou observados, isto é, se

$$P(\mathbf{R}_j = r_j|\mathbf{X}, \mathbf{T}, \mathbf{C}, \boldsymbol{\tau}) = P(\mathbf{R}_j = r_j|\boldsymbol{\tau}),$$

os dados são classificados como MCAR (*missing completely at random*). Em outras palavras, supor que os dados são MCAR significa supor que os indivíduos com dados omissos têm o mesmo perfil que os indivíduos completamente observados. Se, por exemplo, um indivíduo entrou tardiamente no estudo e não realiza todos os exames antes de findar o período de observação, esta suposição é adequada.

O mecanismo conhecido por omissão aleatória ou MAR (*missing at random*) supõe que as probabilidades condicionais de omissão dependem apenas do que é observado, isto é

$$P(\mathbf{R}_j = r_j|\mathbf{X}, \mathbf{T}, \mathbf{C}, \boldsymbol{\tau}) = P(\mathbf{R}_j = r_j|\mathbf{X}_{obs}, \mathbf{T}, \mathbf{C}, \boldsymbol{\tau}).$$

Se, por exemplo, as omissões ocorrem em maior quantidade para grupos com menores tempos de vida, é possível que os pacientes tenham morrido antes de realizar

todos os exames, indicando que a omissão está relacionada aos tempos de falha, então, o mecanismo de omissão aparentemente é MAR.

Se as probabilidades condicionais de omissão dependem também das quantidades não observadas e não podem ser completamente explicadas pelos dados observados, a omissão é chamada não aleatória (*missing not at random*) - MNAR ou *não ignorável*. Em termos de probabilidade, o mecanismo MNAR pode ser representado como $P(\mathbf{R_j} = r_j|\mathbf{X}, \mathbf{T}, \mathbf{C}, \boldsymbol{\tau})$ e não pode ser simplificada.

É importante salientar que ao excluir as observações omissas, pressupõe-se que a omissão é do tipo MCAR, ou seja, que os indivíduos com covariáveis completamente observadas constituem uma subamostra aleatória da amostra original, o que não é razoável na maioria das circunstâncias.

1.4 Objetivos

O objetivo deste trabalho é implementar um procedimento de estimação dos parâmetros do modelo com fração de cura em que algumas observações apresentam omissão em pelo menos uma das covariáveis. A proposta de Chen & Ibrahim (2001) consiste em implementar a versão Monte Carlo do algoritmo *EM*, permitindo que as covariáveis sejam contínuas, categóricas ou mistas. Nossa proposta difere desta pois utilizamos uma verossimilhança marginal, que elimina as variáveis latentes, simplificando o uso do algoritmo *EM*, que será empregado para imputar somente os dados omissos. Além disso, consideramos apenas os casos em que as covariáveis omissas são categóricas, assim, podemos utilizar o algoritmo *EM* ponderado proposto por Ibrahim (1990).

Por meio de um estudo de simulação, confrontaremos estimativas, precisão e probabilidades de cobertura dos parâmetros envolvidos no modelo obtidos através do algoritmo *EM*, com aqueles obtidos quando as unidades com omissão são excluídas do banco de dados (ACC). Avaliamos também, através de simulações, o impacto na estimativa dos parâmetros quando aumentamos o percentual de imunes e de censura entre

indivíduos não imunes em amostras completamente observadas e amostras com omissão, bem como o quanto o percentual de dados omissos pode interferir nos resultados das análises.

1.5 Estrutura da dissertação

Os próximos Capítulos estão organizados da seguinte maneira: no Capítulo 2 descrevemos a formulação do modelo com fração de cura proposto por Yakovlev et al. (1993), considerando que os tempos de promoção seguem uma distribuição Weibull. No Capítulo 3 apresentamos o procedimento de estimação dos parâmetros e erros padrão para o modelo com fração de cura quando covariáveis categóricas são omissas. No Capítulo 4 apresentamos estudos de simulação para comparar propriedades do estimador de máxima verossimilhança sob o modelo com fração de cura e omissão nas covariáveis pela análise de casos completos e pelo algoritmo *EM*. No Capítulo 5 um conjunto de dados reais, referentes ao tempo até a conclusão em um curso de graduação será utilizado para ilustrar o método. Por fim, serão apresentadas no Capítulo 6, as conclusões com base nos resultados obtidos e propostas para estudos futuros.

Capítulo 2

Modelo com Fração de Cura

Neste capítulo, descrevemos a formulação do modelo com fração de cura proposto por Yakovlev et al. (1993), bem como as funções de verossimilhança e verossimilhança marginal associadas, supondo que os tempos de promoção seguem uma distribuição Weibull.

2.1 Formulação do modelo

Seja M uma variável aleatória (v.a.) representando o número de causas ou riscos de ocorrência de um particular evento de interesse. É assumido que M tem distribuição $Poisson(\theta)$. Dado $M = m$, sejam Z_j, $j = 1, \ldots, m$, variáveis aleatórias contínuas não negativas, independentes e identicamente distribuídas com função distribuição acumulada $F(\cdot) = 1 - S(\cdot)$ e independentes de M, representando o tempo de ocorrência do evento devido à j-ésima causa ou risco, ou o tempo de promoção do evento. O tempo até a ocorrência do evento de interesse é definido como $Y = \min\{Z_j; 0 \leq j \leq M\}$ com $P(Z_0 = \infty) = 1$, pois se $M = 0$ não existem causas ou riscos para a ocorrência do evento de interesse. As variáveis M e Z_j são variáveis latentes, ou seja, não observáveis e Y é uma v.a. observável que pode ser censurada à direita. Para este modelo a função

de sobrevivência para a população é dada por

$$S_p(y) = P(Y > y) = P\Big[\min\{Z_0, Z_1, ..., Z_M\} > y\Big],$$

que usando a lei da probabilidade total (Magalhães, 2006; p. 29) pode ser escrita como

$$\begin{aligned} S_p(y) &= \sum_{k=0}^{\infty} P\Big[\min\{Z_0, Z_1, ..., Z_M\} > y | M = k\Big] P(M = k) \\ &= P(Z_0 > y)P(M = 0) + \sum_{k=1}^{\infty} P\Big[min\{Z_1, \ldots, Z_k\} > y\Big] \frac{\theta^k}{k!} e^{-\theta} \\ &= e^{-\theta} + \sum_{k=1}^{\infty} P\Big[Z_1 > y, ..., Z_k > y\Big] \frac{\theta^k}{k!} e^{-\theta}, \end{aligned}$$

pois $P(Z_0 > y) = 1$, para todo y. Agora, dado $M = k$, para $k = 1, \ldots, \infty$, as variáveis aleatórias Z_j, $j = 1, \ldots, M$ são independentes e identicamente distribuídas com função de sobrevivência $S(y)$, então, segue que

$$\begin{aligned} S_p(y) &= e^{-\theta} + \sum_{k=1}^{\infty} S(y)^k \frac{\theta^k}{k!} e^{-\theta} \\ &= e^{-\theta} \sum_{k=0}^{\infty} S(y)^k \frac{\theta^k}{k!} \\ &= e^{-\theta} e^{\theta S(y)} \\ &= \exp\left[-\theta(1 - S(y))\right], \end{aligned}$$

e portanto,

$$S_p(y) = \exp\left[-\theta F(y)\right], \tag{2.1}$$

em que $F(y)$, a função de distribuição acumulada de Z_j, $j = 1 \ldots, M$ é uma função de distribuição própria. Consequentemente,

$$\lim_{y \to \infty} S_p(y) = \exp(-\theta),$$

ou seja, $S_p(y)$ é uma função de sobrevivência imprópria e $\exp(-\theta) = P(M = 0)$ representa a fração de cura induzida pelo modelo.

Pode-se observar que à medida que θ cresce, a fração de cura, $\exp(-\theta)$ decresce, o que é bastante intuitivo, já que θ é o valor esperado para a variável latente M, que representa o número de causas ou riscos de ocorrência do evento. Consequentemente, o indivíduo que apresenta um elevado número de causas para ocorrência do evento terá baixa fração de cura. O raiocínio para θ decrescente é análogo.

Através das relações entre as funções de sobrevivência, risco e densidade, dadas em (1.3) e (1.4), é possível obter a função de densidade correspondente a (2.1) como sendo

$$f_p(y) = \theta f(y) \exp\{-\theta F(y)\}, \quad y > 0, \tag{2.2}$$

com $f(y) = \mathrm{d}F(y)/\mathrm{d}y$ uma função densidade própria. A função risco (taxa de falha instantânea) para a população é

$$h_p(y) = \frac{f_p(y)}{S_p(y)} = \theta f(y). \tag{2.3}$$

Quando relacionamos covariáveis ao parâmetro θ, a função risco em (2.3) possui uma estrutura de riscos proporcionais, que é uma propriedade desejável na análise de sobrevivência. Isto ocorre devido à suposição de que a variável latente M, que representa o número de riscos ou causas para a ocorrência do evento, segue uma distribuição de Poisson (Rodrigues, Cancho, & Castro 2008).

A função de sobrevivência para a população não curada é dada por

$$\begin{aligned} S^*(y) &= P(Y > y | M \geq 1) \\ &= \frac{P(Y > y, M \geq 1)}{P(M \geq 1)} \\ &= \frac{\exp\{-\theta F(y)\} - \exp(-\theta)}{1 - \exp(-\theta)}. \end{aligned} \tag{2.4}$$

Portanto, pode-se relacionar o modelo de mistura padrão e o modelo de tempo

de promoção através da relação

$$S_p(y) = \exp(-\theta) + [1 - \exp(-\theta)]S^*(y),$$

em que $S^*(y)$ é dada por (2.4) e a fração de cura é $\pi = \exp(-\theta)$. Isto mostra que o modelo (2.1) pode ser escrito como o modelo de mistura padrão com uma família específica de funções de sobrevivência $S^*(y)$ dada em (2.4).

2.2 Função de verossimilhança

Considere uma amostra aleatória de n indivíduos de uma determinada população. Define-se portanto, as seguintes variáveis:

- M_i - variável latente representando o número de causas para a ocorrência do evento no i-ésimo indivíduo, com $M_i \sim Poisson(\theta_i)$, $i = 1, \ldots, n$;
- Z_{ij} são os tempos, não observáveis, até a ocorrência do evento devido à j-ésima causa ou risco, com $j = 1, \ldots, M_i$ para o i-ésimo indivíduo, com função distribuição acumulada $F(\cdot|\boldsymbol{\lambda}) = 1 - S(\cdot|\boldsymbol{\lambda})$ que não depende de M_i, sendo $\boldsymbol{\lambda}$ um vetor de parâmetros desconhecidos;
- $y_i = \min\{T_i, C_i\}$, tempo de falha observado, com $T_i = \min\{Z_{ij}; 0 \leq j \leq M_i\}$ e C_i tempo de censura do i-ésimo indivíduo;
- δ_i - indicador de censura, com δ_i igual a 1, se $T_i \leq C_i$, e igual a zero, se $T_i > C_i$;
- $x_i' = (x_{i1}, ..., x_{ip})$ vetor p-dimensional de covariáveis associado ao i-ésimo indivíduo;

Considere os vetores n-dimensionais

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)'$$

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)'$$

$$\mathbf{M} = (M_1, M_2, \ldots, M_n)',$$

e a matriz de covariáveis de dimensão $n \times p$

$$\mathbf{X} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix},$$

denotamos o conjunto dos dados completos (que inclui os dados observados e as variáveis latentes) por $\mathcal{D}_c = (n, \mathbf{y}, \boldsymbol{\delta}, \mathbf{M}, \mathbf{X})$ e os dados observados por $\mathcal{D} = (n, \mathbf{y}, \boldsymbol{\delta}, \mathbf{X})$.

Covariáveis são introduzidas no parâmetro θ através da relação $\theta \equiv \theta(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, com $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ representando o vetor de coeficientes de regressão, de maneira que a fração de cura para o indivíduo i é

$$\pi(x_i) = \exp(-\theta_i) = \exp\left\{-\exp(x_i'\boldsymbol{\beta})\right\}. \tag{2.5}$$

Denotando a densidade conjunta de (y_i, δ_i, M_i) por

$$f(y_i, \delta_i, M_i) = f(y_i, \delta_i | M_i) f(M_i),$$

temos, conforme mostrado em (A.1), que

$$f(y_i, \delta_i | M_i) = S(y_i|\boldsymbol{\lambda})^{M_i - \delta_i} \left[M_i f(y_i|\boldsymbol{\lambda})\right]^{\delta_i}, \tag{2.6}$$

e portanto,

$$f(y_i, \delta_i, M_i) = S(y_i|\boldsymbol{\lambda})^{M_i - \delta_i} \left[M_i f(y_i|\boldsymbol{\lambda})\right]^{\delta_i} \frac{e^{-\theta}\theta^{M_i}}{M_i!}.$$

Sendo $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{\lambda}')'$ o vetor de parâmetros desconhecidos, a função de verossimilhança dos dados completos para $\boldsymbol{\phi}$ é então dada por

$$
\begin{aligned}
L(\boldsymbol{\phi};\mathcal{D}_c) &= \left\{\prod_{i=1}^{n} S(y_i|\boldsymbol{\lambda})^{M_i-\delta_i}[M_i f(y_i|\boldsymbol{\lambda})]^{\delta_i}\right\} \\
&\quad \times \exp\left\{\sum_{i=1}^{n}[M_i x_i'\boldsymbol{\beta} - \ln(M_i!) - \exp(x_i'\boldsymbol{\beta})]\right\}. \qquad (2.7)
\end{aligned}
$$

O logarítmo da função de verossimilhança é

$$
\begin{aligned}
\ell(\boldsymbol{\phi};\mathcal{D}_c) &= \log L(\boldsymbol{\phi};\mathcal{D}_c) \\
&= \sum_{i=1}^{n}[(M_i-\delta_i)\ln S(y_i|\boldsymbol{\lambda}) + \delta_i \ln M_i + \delta_i \ln f(y_i|\boldsymbol{\lambda})] \\
&\quad + \sum_{i=1}^{n}[M_i x_i'\boldsymbol{\beta} - \ln(M_i!) - \exp(x_i'\boldsymbol{\beta})]. \qquad (2.8)
\end{aligned}
$$

Visto que (2.8) não é observável, já que depende das variáveis latentes M_i, $i = 1,\ldots,n$, pode-se trabalhar com o logarítmo da função de verossimilhança marginal, que pode ser obtida fazendo-se o somatório nas variáveis M_i. Como mostrado no Apêndice A.2, a verossimilhança marginal é dada por

$$
L(\boldsymbol{\phi};\mathcal{D}) = \prod_{i=1}^{n}[x_i'\boldsymbol{\beta} f(y_i|\boldsymbol{\lambda})]^{\delta_i}\exp\big\{-\exp(x_i'\boldsymbol{\beta})\big[1-S(y_i|\boldsymbol{\lambda})\big]\big\}, \qquad (2.9)
$$

e o logarítmo da função de verossimilhança marginal é

$$
\ell(\boldsymbol{\phi};\mathcal{D}) = \sum_{i=1}^{n}\Big\{\delta_i\,(x_i'\boldsymbol{\beta} + \ln f(y_i|\boldsymbol{\lambda})) - \exp(x_i'\boldsymbol{\beta})\big[1-S(y_i|\boldsymbol{\lambda})\big]\Big\}. \qquad (2.10)
$$

Para obter a estimativa de máxima verossimilhança de $\boldsymbol{\phi}$, Chen & Ibrahim (2001) propõem maximizar o logarítmo da função de verossimilhança em (2.8) utilizando o algoritmo *EM*, imputando valores para as variáveis latentes M_i. Neste trabalho, será maximizado o logarítmo da função de verossimilhança em (2.10) com respeito a $\boldsymbol{\phi}$, visto que para censuras do tipo I ou aleatória, $\ell(\boldsymbol{\phi};\mathcal{D})$ tem as propriedades usuais de um logarítmo da função de verossimilhança, como citam Rodrigues, Cancho, & Castro

(2008).

Para grandes amostras, o estimador de máxima verossimilhança $\hat{\boldsymbol{\phi}}$ é tal que

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \dot{\sim} N\left(0, \mathcal{I}^{-1}(\boldsymbol{\phi})\right)$$

em que $\mathcal{I}(\boldsymbol{\phi})$ é a matriz de informação de Fisher dada por

$$\mathcal{I}(\boldsymbol{\phi}) = -E[\ddot{\ell}(\boldsymbol{\phi})]$$

com

$$\ddot{\ell}(\boldsymbol{\phi}) = \frac{\partial^2 \ell(\boldsymbol{\phi}; \mathcal{D})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}.$$

Como o cálculo de $\mathcal{I}(\boldsymbol{\phi})$ não é possível devido à presença de censuras, pode-se utilizar a matriz $-\ddot{\ell}(\boldsymbol{\phi})$ avaliada em $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$ que é uma estimativa consistente de $\mathcal{I}(\boldsymbol{\phi})$, conhecida como matriz de informação observada que será denotada aqui por $\mathcal{I}_{obs}(\hat{\boldsymbol{\phi}})$. Assim, pode-se realizar inferências, como intervalos de confiança e testes de hipóteses, para os parâmetros de interesse.

2.3 Modelo paramétrico Weibull

Suponha agora que os tempos de promoção Z_{ij} são independentes e identicamente distribuídos (*i.i.d.*) e seguem uma distribuição Weibull com parâmetros $\boldsymbol{\lambda} = (\rho, \gamma)'$, $j = 1, \ldots, M_i$, $i = 1, \ldots, n$, com densidade $f(z) = \rho z^{\rho-1} \exp(\gamma - z^\rho e^\gamma)$ e função de sobrevivência $S(z) = \exp(-z^\rho e^\gamma)$. Neste caso, o logarítmo da função de verossimilhança dos dados completos em (2.8) será dada por

$$\begin{aligned} \ell(\boldsymbol{\phi}; \mathcal{D}_c) &= \sum_{i=1}^{n} -M_i y_i^\rho e^\gamma + \delta_i \ln(M_i \rho y_i^{\rho-1} e^\gamma) \\ &\quad + \sum_{i=1}^{n} [M_i x_i' \boldsymbol{\beta} - \ln(M_i!) - \exp(x_i' \boldsymbol{\beta})]. \end{aligned} \tag{2.11}$$

O logarítmo da verossimilhança marginal em (2.10) toma a forma

$$\begin{aligned}\ell(\boldsymbol{\phi};\mathcal{D}) &= \sum_{i=1}^{n}\delta_i\left[x_i'\boldsymbol{\beta}+\gamma+\ln(\rho y_i^{\rho-1})-y_i^{\rho}e^{\gamma}\right]\\ &\quad-\sum_{i=1}^{n}\exp(x_i'\boldsymbol{\beta})\left[1-\exp(-y_i^{\rho}e^{\gamma})\right].\end{aligned}\tag{2.12}$$

Para o modelo (2.12), a matriz de informação observada, $\mathcal{I}_{obs}(\hat{\boldsymbol{\phi}})$, fica

$$\mathcal{I}_{obs}(\hat{\boldsymbol{\phi}})=\left.\begin{pmatrix}\ddot{\ell}^{\boldsymbol{\beta}\boldsymbol{\beta}} & \ddot{\ell}^{\boldsymbol{\beta}\rho} & \ddot{\ell}^{\boldsymbol{\beta}\gamma}\\ & \ddot{\ell}^{\rho\rho} & \ddot{\ell}^{\rho\gamma}\\ & & \ddot{\ell}^{\gamma\gamma}\end{pmatrix}\right|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}\tag{2.13}$$

sendo as submatrizes com elementos

$$\begin{aligned}
\ddot{\ell}^{\boldsymbol{\beta}\boldsymbol{\beta}}_{k,l} &= \frac{\partial^2\ell(\boldsymbol{\phi};\mathcal{D})}{\partial\beta_l\partial\beta_k}=-\sum_{i=1}^{n}x_{il}x_{ik}\exp(x_i'\boldsymbol{\beta})\left[1-\exp(-y_i^{\rho}e^{\gamma})\right],\quad l,k=1,\ldots,p,\\
\ddot{\ell}^{\boldsymbol{\beta}\rho}_{l} &= \frac{\partial^2\ell(\boldsymbol{\phi};\mathcal{D})}{\partial\beta_l\partial\rho}=-\sum_{i=1}^{n}x_{il}y_i^{\rho}e^{\gamma}\ln(y_i)\exp(x_i'\boldsymbol{\beta}-y_i^{\rho}e^{\gamma}),\quad l=1,\ldots,p,\\
\ddot{\ell}^{\boldsymbol{\beta}\gamma}_{l} &= \frac{\partial^2\ell(\boldsymbol{\phi};\mathcal{D})}{\partial\beta_l\partial\gamma}=-\sum_{i=1}^{n}x_{il}y_i^{\rho}e^{\gamma}\exp(x_i'\boldsymbol{\beta}-y_i^{\rho}e^{\gamma}),\quad l=1,\ldots,p,\\
\ddot{\ell}^{\rho\rho} &= \frac{\partial^2\ell(\boldsymbol{\phi};\mathcal{D})}{\partial\rho\partial\rho}=-\sum_{i=1}^{n}\left\{\frac{\delta_i}{\rho^2}+y_i^{\rho}e^{\gamma}(\ln(y_i))^2\Big[\delta_i+\exp(x_i'\boldsymbol{\beta}-y_i^{\rho}e^{\gamma})\left[1-y_i^{\rho}e^{\gamma}\right]\Big]\right\},\\
\ddot{\ell}^{\rho\gamma} &= \frac{\partial^2\ell(\boldsymbol{\phi};\mathcal{D})}{\partial\rho\partial\gamma}=-\sum_{i=1}^{n}y_i^{\rho}e^{\gamma}\ln(y_i)\Big\{\delta_i+\exp(x_i'\boldsymbol{\beta}-y_i^{\rho}e^{\gamma})\left[1-y_i^{\rho}e^{\gamma}\right]\Big\},\\
\ddot{\ell}^{\gamma\gamma} &= \frac{\partial^2\ell(\boldsymbol{\phi};\mathcal{D})}{\partial\gamma\partial\gamma}=-\sum_{i=1}^{n}y_i^{\rho}e^{\gamma}\Big\{\delta_i+\exp(x_i'\boldsymbol{\beta}-y_i^{\rho}e^{\gamma})\left[1-y_i^{\rho}e^{\gamma}\right]\Big\}.
\end{aligned}$$

Testes estatísticos podem ser realizados para o vetor de parâmetros $\boldsymbol{\beta}$, sendo $\boldsymbol{\lambda}=(\rho,\gamma)$ tratados como parâmetros de perturbação.

Considere as hipóteses $H_0:\boldsymbol{\beta}=\boldsymbol{\beta_0}$ *versus* $H_1:\boldsymbol{\beta}\neq\boldsymbol{\beta_0}$ e seja $\tilde{\boldsymbol{\lambda}}$ a estimativa de $\boldsymbol{\lambda}$ restrita à hipótese H_0, ou seja, obtida pela maximização de $\ell(\boldsymbol{\phi},\mathcal{D})$ com respeito a

$\boldsymbol{\lambda}$ para $\boldsymbol{\beta} = \boldsymbol{\beta_0}$ fixado. Então, a estatística da razão de verossimilhanças é dada por

$$\xi_{RV} = -2\Big[\ell(\boldsymbol{\beta}_0, \tilde{\boldsymbol{\lambda}}, \mathcal{D}) - \ell(\hat{\boldsymbol{\phi}}, \mathcal{D})\Big],$$

em que ξ_{RV} tem distribuição assintótica $\chi^2_{(p)}$ sob a hipótese H_0, sendo p a dimensão do vetor $\boldsymbol{\beta}$.

Capítulo 3

Modelo com Fração de Cura e Omissão nas Covariáveis

Neste capítulo descrevemos o procedimento de estimação dos parâmetros e erros padrão para o modelo com fração de cura na situação em que algumas covariáveis categóricas não são observadas para todos os indivíduos na amostra. Apresentamos a verossimilhança marginal (com respeito às variáveis latentes) sob a suposição de que o mecanismo gerador das omissões é MAR e o algoritmo *EM* ponderado será utilizado para imputar valores para as covariávieis omissas.

3.1 Estimação de máxima verossimilhança via algoritmo EM

Suponhamos que o mecanismo gerador da omissão é do tipo MAR, ou seja, a omissão não depende dos valores omissos, mas talvez dependa dos valores observados. Assumimos também que os tempos de falha e censura são sempre observados. Neste caso, é necessário considerar um modelo paramétrico para as covariáveis omissas, $p(x_i;\boldsymbol{\alpha})$, em que $\boldsymbol{\alpha}$ é um vetor de parâmetros desconhecidos, vistos aqui como parâmetros de perturbação, pois nosso principal interesse é realizar inferências sobre o

vetor de coeficientes de regressão do modelo, $\boldsymbol{\beta}$.

Considere que, para a i-ésima unidade amostral, é possível particionar o vetor x_i em $x_i = (x'_{mis,i}, x'_{obs,i})'$ com $x_{mis,i}$ um vetor de dimensão q_i de covariáveis omissas e $x_{obs,i}$ um vetor de dimensão $p - q_i$ de covariáveis completamente observadas.

Seja agora $\mathcal{D}_{obs} = (n, \mathbf{y}, \boldsymbol{\delta}, \mathbf{X}_{obs})$ o conjunto de dados observados, com $\mathbf{X}_{obs}$ a matriz de covariáveis observadas e $\boldsymbol{\psi} = (\boldsymbol{\beta}', \boldsymbol{\lambda}', \boldsymbol{\alpha}')'$, o vetor de parâmetros desconhecidos. Então o logarítmo da função de verossimilhança dos dados completos é dada por

$$\begin{aligned}
\ell(\boldsymbol{\psi}; \mathcal{D}_c) &= \sum_{i=1}^{n} \left[(M_i - \delta_i) \ln S(y_i|\boldsymbol{\lambda}) + \delta_i \ln M_i + \delta_i \ln f(y_i|\boldsymbol{\lambda}) \right] \\
&\quad + \sum_{i=1}^{n} \left[M_i x'_i \boldsymbol{\beta} - \ln(M_i!) - \exp(x'_i \boldsymbol{\beta}) \right] \\
&\quad + \sum_{i=1}^{n} \ln[p(x_i; \boldsymbol{\alpha})] \\
&= \ell(\boldsymbol{\phi}; \mathcal{D}_c) + \sum_{i=1}^{n} \ln[p(x_i; \boldsymbol{\alpha})] \qquad (3.1)
\end{aligned}$$

Para eliminar as variáveis latentes efetua-se o somatório em (3.1) sobre todos os valores possíveis de M_i, de forma análoga ao caso sem omissão, e assim obtém-se

$$\begin{aligned}
\ell(\boldsymbol{\psi}; \mathcal{D}) &= \sum_{i=1}^{n} \left[\delta_i \left(x'_i \boldsymbol{\beta} + \ln f(y_i|\boldsymbol{\lambda}) \right) - \exp(x'_i \boldsymbol{\beta}) \left(1 - S(y_i|\boldsymbol{\lambda}) \right) \right] + \sum_{i=1}^{n} \ln[p(x_i; \boldsymbol{\alpha})] \\
&= \ell(\boldsymbol{\phi}; \mathcal{D}) + \sum_{i=1}^{n} \ln[p(x_i; \boldsymbol{\alpha})]. \qquad (3.2)
\end{aligned}$$

O logarítmo da função de verossimilhança em (3.2) pode ser maximizado através do algoritmo *EM* que é um procedimento iterativo composto de dois passos a cada iteração: o passo E (*Expectation*) e o passo M (*Maximization*), por isso, o algoritmo foi chamado Algoritmo *EM* por Dempster et al. (1977) e é uma poderosa ferramenta para obter estimativas de máxima verossimilhança quando existem covariáveis não observáveis (latentes) ou omissas.

As situações onde o algoritmo *EM* pode ser aplicado incluem não somente da-

dos evidentemente incompletos (dados omissos, distribuições truncadas, observações censuradas ou truncadas), mas também uma ampla variedade de problemas onde a falta de dados não é natural ou evidente. A desvantagem do algoritmo *EM* é que sua taxa de convergência pode ser extremamente lenta quando existe uma grande fração de informação omissa. Nesta dissertação, o algoritmo *EM* é utilizado para encontrar estimativas de máxima verossimilhança quando covariáveis não são observadas para todos indivíduos.

Consideramos $\boldsymbol{\psi}^{(k)} = (\boldsymbol{\beta}'^{(k)}, \boldsymbol{\lambda}'^{(k)}, \boldsymbol{\alpha}'^{(k)})'$ as estimativas de $\boldsymbol{\psi}$ na k-ésima iteração do algoritmo. Para os dados referentes à i-ésima observação, os dados sem as variáveis latentes são representados por $\mathcal{D}_i = (y_i, \delta_i, x_i)$, e os dados observados por $\mathcal{D}_{obs,i} = (y_i, \delta_i, x_{obs,i})$. No passo E, calcula-se a esperança condicional de $\ell(\boldsymbol{\psi}|\mathcal{D}_i)$ dado a estimativa atual $\boldsymbol{\psi}^{(k)}$ e os dados observados $\mathcal{D}_{obs,i}$, para $i = 1, \ldots, n$. Denotando essa esperança por $Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$, o passo E para a i-ésima observação na k-ésima iteração pode ser escrito como

$$Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = E\left[\ell(\boldsymbol{\psi}, \mathcal{D}_i)|\mathcal{D}_{obs,i}, \boldsymbol{\psi}^{(k)}\right] \tag{3.3}$$

A seguir, descrevemos a obtenção de $Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ para covariáveis omissas categóricas utilizando o algorimo *EM* ponderado.

3.1.1 Algoritmo *EM* ponderado

No caso em que as covariáveis omissas são todas categóricas, uma técnica útil para obter etimativas de máxima verossimilhança é o algoritmo *EM* ponderado, proposto por Ibrahim (1990) que é descrito em detalhes em Horton & Laird (1999).

Considere que J é o número de valores possíveis para o vetor $x_{mis,i}$, por exemplo, se existem q covariáveis omissas categóricas e $c_1, \ldots, c_q$ representam o número de categorias em cada covariável, então $J = \prod_{i=1}^{q} c_i$. Seja $\mathcal{D}_i^{(j)} = (y_i, \delta_i, x_i^{(j)})$ os dados com valores imputados, em que $x_i^{(j)} = (x_{mis,i}^{(j)}, x_{obs,i})$ é o vetor de covariáveis com os valores imputados e observados, para $j = 1, \ldots, J$, e as probabilidades do vetor $x_{mis,i}$ assumir

o valor $x_{mis,i}^{(j)}$ são os pesos denotados por $w_{ij}^{(k)}$, sendo

$$w_{ij}^{(k)} = p(x_{mis,i}^{(j)}|x_{obs,i}, y_i, \delta_i, \boldsymbol{\psi}^{(k)}) = \frac{L(\boldsymbol{\psi}^{(k)}; \mathcal{D}_i^{(j)}) f(x_i^{(j)}|\alpha^{(k)})}{\sum_{j=1}^{J} L(\boldsymbol{\psi}^{(k)}; \mathcal{D}_i^{(j)}) f(x_i^{(j)}|\alpha^{(k)})}, \tag{3.4}$$

Para tornar mais clara a obtenção de $Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$, vamos reportar o cálculo de uma esperança condicional. Considerando V uma variável aleatória discreta e U uma variável aleatória qualquer, a esperança de V dado U é dada por

$$E[V|U] = \sum_j v_j p(v_j|u). \tag{3.5}$$

Agora, fazendo $V = \ell(\boldsymbol{\psi}, \mathcal{D}_i^{(j)})$ e $U = (\mathcal{D}_{obs,i}, \boldsymbol{\psi}^{(k)})$ tem-se

$$Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) = E[V|U] = \sum_j v_j p(v_j|u) = \sum_{j=1}^{J} \ell(\boldsymbol{\psi}, \mathcal{D}_i^{(j)}) w_{ij}^{(k)}.$$

Assim, para o modelo com fração de cura, e considerando o logarítmo da função de verossimilhança em (3.2), o passo E para a i-ésima observação assume a forma

$$\begin{aligned} Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) &= \sum_{j=1}^{J} w_{ij}^{(k)} \delta_i \left\{ \ln\left[f(y_i|\boldsymbol{\lambda})\right] + x_i^{(j')}\boldsymbol{\beta} \right\} \\ &\quad - \sum_{j=1}^{J} w_{ij}^{(k)} \exp(x_i^{(j')}\boldsymbol{\beta}) \left[1 - S(y_i|\boldsymbol{\lambda})\right] \\ &\quad + \sum_{j=1}^{J} w_{ij}^{(k)} \ln\left[p(x_i^{(j)}|\boldsymbol{\alpha})\right] \\ &= Q_{1i}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\psi}^{(k)}) + Q_{2i}(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(k)}) \end{aligned} \tag{3.6}$$

sendo, $Q_{1i}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\psi}^{(k)}) = \sum_{j=1}^{J} w_{ij}^{(k)} \left\{ \delta_i \left(\ln\left[f(y_i|\boldsymbol{\lambda})\right] + x_i^{(j')}\boldsymbol{\beta} \right) - \exp(x_i^{(j')}\boldsymbol{\beta}) \left[1 - S(y_i|\boldsymbol{\lambda})\right] \right\}$

e $Q_{2i}(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(k)}) = \sum_{j=1}^{J} w_{ij}^{(k)} \ln\left[p(x_i^{(j)}|\boldsymbol{\alpha})\right]$.

Note que $\sum_{j=1}^{J} w_{ij}^{(k)} = 1$, para todo k e i, e que se todas as covariáveis para o i-ésimo indivíduo são observadas, então $w_{ij}^{(k)} = 1$. A expressão (3.6) pode ser escrita como a soma entre $Q_{1i}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\psi}^{(k)})$ e $Q_{2i}(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(k)})$, sendo que a primeira não envolve o parâmetro $\boldsymbol{\alpha}$ e a segunda não envolve os parâmetros $\boldsymbol{\beta}$ e $\boldsymbol{\lambda}$. Assim, na etapa de maximização (passo M) podemos maximizá-las separadamente com respeito a $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ e $\boldsymbol{\alpha}$, respectivamente.

O passo E para todas as observações é portanto dado por

$$\begin{aligned} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) &= \sum_{i=1}^{n} Q_{1i}(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\psi}^{(k)}) + \sum_{i=1}^{n} Q_{2i}(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(k)}) \\ &= Q_1(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{\psi}^{(k)}) + Q_2(\boldsymbol{\alpha}|\boldsymbol{\psi}^{(k)}). \end{aligned} \tag{3.7}$$

O passo M consiste na maximização de (3.7) com respeito a $\boldsymbol{\psi}$, que pode ser feita através de um método iterativo, como Newton-Raphson. A função $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ é atualizada usando a estimativa atual $\boldsymbol{\psi}^{(k+1)}$, e o algoritmo é repetido até que seja satisfeito um critério de parada estabelecido. Adotamos aqui, como critério a condição

$$\left|\left|\boldsymbol{\psi}^{(k+1)} - \boldsymbol{\psi}^{(k)}\right|\right| < \varepsilon$$

com $\boldsymbol{\psi}^{(k)} = (\boldsymbol{\beta}'^{(k)}, \boldsymbol{\lambda}'^{(k)}, \boldsymbol{\alpha}'^{(k)})'$ e $\varepsilon > 0$ em que $||\mathbf{a} - \mathbf{b}||$ denota a distância Euclidiana entre os pontos $\mathbf{a}$ e $\mathbf{b}$.

Para ilustrar a implementação do algoritmo *EM* ponderado, adaptamos um exemplo dado em Horton & Laird (1999). Considere um conjunto de dados hipotético com duas covariáveis dicotômicas (x_1, x_2), a resposta (y) e a variável incadora de falha (δ).

A Tabela 3.1 mostra os dados com algumas observações omissas e a Tabela 3.2 exibe os correspondentes dados aumentados, em que $w_{ij}^{(k)}$ são os pesos estimados para o indivíduo i na k-ésima iteração. Por exemplo, $w_{31}^{(k)} = P(x_2 = 0|x_1 = \delta = 1, y = y_3, \boldsymbol{\psi}^{(k)})$ e $w_{31}^{(k)} + w_{32}^{(k)} = 1$.

Para estimar a variância assintótica de $\hat{\boldsymbol{\psi}}$ utilizamos o método de Louis (1982)

Tabela 3.1: Conjunto de dados original (hipotético).

i	y	δ	x_1	x_2
1	y_1	0	0	0
2	y_2	0	0	1
3	y_3	1	1	-
4	y_4	1	0	1
5	y_5	0	-	-

Tabela 3.2: Conjunto de dados aumentados.

i	y	δ	x_1	x_2	$w^{(k)}$
1	y_1	0	0	0	1
2	y_2	0	0	1	1
3	y_3	1	1	0	$w_{31}^{(k)}$
3	y_3	1	1	1	$w_{32}^{(k)}$
4	y_4	1	0	1	1
5	y_5	0	0	0	$w_{51}^{(k)}$
5	y_5	0	0	1	$w_{52}^{(k)}$
5	y_5	0	1	0	$w_{53}^{(k)}$
5	y_5	0	1	1	$w_{54}^{(k)}$

que é baseado no vetor gradiente (matriz de primeiras derivadas parciais) e na matriz Hessiana (matriz de segundas derivadas) de $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$.

Sejam o vetor gradiente e a matriz Hessiana de $Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$, respectivamente dados por

$$\dot{Q}_i(\hat{\boldsymbol{\psi}}) = \left[\begin{array}{ccc} \left(\frac{\partial Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\beta}}\right)' & \left(\frac{\partial Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\lambda}}\right)' & \left(\frac{\partial Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\alpha}}\right)' \end{array} \right]',$$

$$\ddot{Q}_i(\hat{\boldsymbol{\psi}}) = \left[\begin{array}{ccc} \frac{\partial^2 Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} & \frac{\partial^2 Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\lambda}} & \frac{\partial^2 Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\alpha}} \\ & \frac{\partial^2 Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\lambda}'} & \frac{\partial^2 Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\lambda}\partial \boldsymbol{\alpha}} \\ & & \frac{\partial^2 Q_i(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}'} \end{array} \right],$$

de acordo com Louis (1982), a matriz de informação observada é dada por

$$\mathcal{I}_{Louis}(\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \Big\{ -\ddot{Q}_i(\hat{\boldsymbol{\psi}}) + E\Big[\dot{\ell}(\hat{\boldsymbol{\psi}}, \mathcal{D}_i)\dot{\ell}(\hat{\boldsymbol{\psi}}, \mathcal{D}_i)'\Big] - \dot{Q}_i(\hat{\boldsymbol{\psi}})\dot{Q}_i(\hat{\boldsymbol{\psi}})' \Big\}. \qquad (3.8)$$

Se as covariáveis omissas são categóricas tem-se que

$$E\Big[\dot{\ell}(\hat{\boldsymbol{\psi}},\mathcal{D}_i)\dot{\ell}(\hat{\boldsymbol{\psi}},\mathcal{D}_i)'\Big] = \sum_{j=1}^{J} w_{ij}^{(k)}\dot{\ell}(\hat{\boldsymbol{\psi}},\mathcal{D}_i^{(j)})\dot{\ell}(\hat{\boldsymbol{\psi}},\mathcal{D}_i^{(j)})',$$

assim, a equação (3.8) assume a forma

$$\mathcal{I}_{Louis}(\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n}\left\{-\ddot{Q}_i(\hat{\boldsymbol{\psi}}) + \sum_{j=1}^{J} w_{ij}^{(k)}\dot{\ell}(\hat{\boldsymbol{\psi}},\mathcal{D}_i^{(j)})\dot{\ell}(\hat{\boldsymbol{\psi}},\mathcal{D}_i^{(j)})' - \dot{Q}_i(\hat{\boldsymbol{\psi}})\dot{Q}_i(\hat{\boldsymbol{\psi}})'\right\}. \quad (3.9)$$

Deste modo, a variância asssintótica de $\hat{\boldsymbol{\psi}}$ é dada por

$$Var(\hat{\boldsymbol{\psi}}) = \Big[\mathcal{I}_{Louis}(\hat{\boldsymbol{\psi}})\Big]^{-1}$$

então, pode-se realizar inferências sobre o vetor de parâmetros $\boldsymbol{\beta}$.

3.1.2 Modelando a distribuição das covariáveis

Uma questão crucial no tratamento de dados omissos é a especificação de um modelo de probabilidade para as covariáveis omissas. Quando uma distribuição paramétrica é especificada para as covariáveis, os parâmetros desta distribuição são tipicamente vistos como parâmetros de perturbação. A estimação dos parâmetros pode ser computacionalmente intensiva e ineficiente se existirem muitos parâmetros de perturbação e grande percentual de dados omissos. Portanto, estratégias precisam ser empregadas na especificação da distribuição das covariáveis para reduzir o número de parâmetros de perturbação. Para tal, Lipsitz & Ibrahim (1996) e Ibrahim, Chen, & Lipsitz (1999) sugerem modelar a distribuição conjunta das covariáveis como o produto de distribuições condicionais unidimensionais. Esta estratégia tem o potencial de reduzir drásticamente o número de parâmetros de perturbação que precisam ser estimados no passo M do algoritmo *EM*. A distribuição conjunta do vetor p-dimensional de covariáveis $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ pode ser escrito por meio de uma série de distribuições

condicionais unidimensionais da seguinte forma

$$p(x_{i1}, ..., x_{ip}|\boldsymbol{\alpha}) = p(x_{ip}|x_{i1}, ..., x_{i(p-1)}, \boldsymbol{\alpha}_p)...p(x_{i2}|x_{i1}, \boldsymbol{\alpha}_2)p(x_{i1}|\boldsymbol{\alpha}_1) \tag{3.10}$$

em que $\boldsymbol{\alpha}_j$ é um vetor de parâmetros para a j-ésima distribuição condicional, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2', \ldots, \boldsymbol{\alpha}_p')'$. É importante ressaltar que a equação (3.10) precisa ser especificada apenas para as covariáveis omissas.

Se as covariáveis omissas são todas categóricas dicotômicas, uma sequência de modelos logísticos (ou ligações probito e complemento log-log) podem ser modelados para cada $p(x_{ij}|x_{i1}, ..., x_{i(j-1)}, \boldsymbol{\alpha}_j)$, $j = 1, \ldots, p$. Para covariáveis categóricas com mais de dois níveis, podemos considerar um modelo logístico multinomial (Agresti 2002). Se as covariáveis consistem de contagem, é possível modelar $p(x_{ij}|x_{i1}, ..., x_{i(j-1)}, \boldsymbol{\alpha}_j)$ como um modelo de regressão Poisson.

A modelagem da distribuição das covariáveis depende da ordem de condicionamento das covariáveis. Contudo, Ibrahim, Chen, & Lipsitz (1999) e Chen & Ibrahim (2001) mostraram, através de uma análise de sensitividade que as estimativas de $\boldsymbol{\beta}$ são robustas com respeito às mudanças na ordem de condicionamento, adição de termos de interação, bem como mudanças na função de ligação.

Capítulo 4

Estudo de Simulação

Neste capítulo apresentamos estudos de simulação para comparar propriedades do estimador de máxima verossimilhança sob o modelo de fração de cura e omissão nas covariáveis pela análise de casos completos e pela imputação dos dados faltantes com o algoritmo *EM*. Para este estudo, utilizamos o ambiente `R` (versão 2.7.2).

Como a metodologia mais comumente empregada na prática é a análise de casos completos, nosso principal objetivo aqui é comparar as estimativas obtidas por este método com as estimativas de máxima verossimilhança obtidas através do algoritmo *EM*. Investigamos também as propriedades assintóticas das estimativas de máxima verossimilhança para o modelo com fração de cura quando as amostras são completamente observadas.

4.1 Obtenção dos dados simulados

Consideramos uma situação com três covariáveis associadas a cada indivíduo (x_1, x_2, x_3). Para $i = 1, \ldots, n$, assumimos para cada i que (x_{i1}, x_{i2}) são sempre observadas, sendo x_{i1} e x_{i2} independentes, com valores x_{i1} obtidos por amostragem $i.i.d.$ da distribuição normal padrão e x_{i2} obtidos por amostragem $i.i.d.$ da distribuição de Bernoulli com probabilidade de sucesso $0,6$ e que x_{i3} pode ser omissa. Consideramos que, dado x_{i1} e x_{i2}, a covariável com omissão x_{i3} tem distribuição de Bernoulli com

parâmetro α_i, então

$$p(x_{i3}|x_{i1}, x_{i2}, \alpha_i) = \alpha_i^{x_{i3}}(1-\alpha_i)^{1-x_{i3}}$$

com $x_{i1} \in \mathbb{R}$, $\quad x_{i2} = 0, 1 \quad$ e $\quad x_{i3} = 0, 1$ sendo

$$\alpha_i = \frac{\exp(\alpha_1 x_{i1} + \alpha_2 x_{i2})}{1 + \exp(\alpha_1 x_{i1} + \alpha_2 x_{i2})},$$

com $\alpha = (\alpha_1, \alpha_2) = (-0,5; 1)$ e as distribuições condicionais são independentes para cada i.

Para cada indivíduo foram gerados valores para M_i como uma amostra $i.i.d.$ da distribuição de Poisson com média $\theta_i = \exp(x_i'\boldsymbol{\beta})$, representando o número de riscos para a ocorrência do evento, com $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$. O valor fixado para o vetor $\boldsymbol{\beta}$ determina o percentual de indivíduos com $m_i = 0$, ou seja, o percentual de imunes na amostra.

Para cada indivíduo não imune, $(m_i > 0)$, foram geradas amostras de tamanho m_i para $Z_{ij} \sim Weibull(\rho, \gamma)$, com $\rho = 2$ e $\gamma = -2\log 4$. Assim, os tempos de falha são $t_i = \min\{z_{ij}; j = 1, \ldots, m_i\}$.

Geramos também censuras aleatórias a partir de uma distribuição uniforme no intervalo $(0, u)$, sendo que a mudança no valor de u afeta a proporção de censuras na amostra. Consideramos aqui a proporção de censuras calculada com respeito ao total de indivíduos sujeitos ao evento, com o intuito de avaliar separadamente o efeito do aumento da proporção de censuras entre não curados e de imunes na estimação dos parâmetros. Assim, considerando os eventos: $A \equiv$ curados; $\bar{A} \equiv$ não curados e $B \equiv$ censurados ou imunes. A proporção de censura entre não imunes utilizada nesta simulação (denotada pc_1) representa a frequência relativa de B dado $\bar{A}$, ou seja,

$$pc_1 = \frac{\text{número de indivíduos em } B \cap \bar{A}}{\text{número de indivíduos em } \bar{A}},$$

enquanto que o percentual de censurados ou imunes (denotado pc_2), que em aplicações

é interpretado simplesmente como a proporção de censuras, pode ser representado por

$$pc_2 = \frac{\text{número de indivíduos em } B}{\text{número de indivíduos na população}}.$$

Denotando por π a proporção de imunes, pode-se verificar, pelo uso de propriedades de frequencia relativa, a seguinte relação entre estas duas quantidades:

$$pc_2 = pc_1(1-\pi) + \pi. \tag{4.1}$$

Os tempos observados serão $y_i = \min\{t_i, c_i\}$ e, associado a cada tempo, tem-se $\delta_i = 1$ se $t_i \leq c_i$ e $\delta_i = 0$ se $t_i > c_i$.

Os dados omissos para x_{i3} foram gerados com um mecanismo de omissão que não depende de x_{i3}, e portanto, os dados são MAR, sendo que mecanismo de omissão dos dados pode ser ignorado na estimação dos parâmetros. Para simular omissão, adaptamos a idéia de Ibrahim, Chen, & Lipsitz (1999). A variável indicadora de omissão foi especificada como descrito na Seção 1.3.1, sendo $r_{i3} = 0$ se a variável x_{i3} é observada e $r_{i3} = 1$ se x_{i3} é omissa, $i = 1, \ldots, n$. A distribuição assumida para r_{i3} é dada por

$$P(R_{i3} = r | \mathcal{D}_{obs}, \tau) = \tau_i^r (1-\tau_i)^{1-r},$$

com

$$\tau_i = \frac{\exp(\tau_{30} + \tau_{31}x_{i1} + \tau_{32}x_{i2} + \tau_{33}y_i + \tau_{34}\delta_i)}{1 + \exp(\tau_{30} + \tau_{31}x_{i1} + \tau_{32}x_{i2} + \tau_{33}y_i + \tau_{34}\delta_i)}$$

e o valor considerado para o vetor $\boldsymbol{\tau}_3 = (\tau_{30}, \tau_{31}, \tau_{32}, \tau_{33}, \tau_{34})'$ controla o percentual de omissão na amostra.

Os comandos em `R` para geração de amostras como descrito acima, são dados no apêndice B.

4.2 Resultados para os dados simulados

Nesta seção apresentamos os resultados obtidos para amostras simuladas em diferentes situações. A Tabela 4.1 resume os casos considerados nas simulações:

Tabela 4.1: Situações consideradas para simulação.

Caso	Omissão	% Imunes	% Censura	Tamanho da Amostra	Seção
1	sem	variando	0	variando	4.2.1
2		variando	variando	fixo ($n = 300$)	
3	com	variando	0	fixo ($n = 300$)	4.2.2
4		fixo (45%)	variando	fixo ($n = 300$)	

4.2.1 Resultados para amostras completamente observadas

A partir da Tabela 4.2 observamos que, para amostras moderadas ($n = 50$) não apresentando omissão nas covariáveis nem censuras entre os não imunes (caso 1), as estimativas dos parâmetros apresentam vieses, especialmente no parâmetro γ, para diferentes proporções de imunes (10%, 45% e 65%), no entanto, aumentando-se o tamanho da amostra este problema é aparentemente sanado.

Na Tabela 4.3, que mostra os resultados para amostras em que as covariáveis não apresentam omissão mas apresentam censura entre os não imunes (caso 2), podemos ver que o percentual de censuras na amostra, afeta sensivelmente as estimativas dos parâmetros. Observamos vieses e erros quadráticos médios acentuados nas estimativas dos parâmetros quando o percentual de censuras é alto (50%). Por exemplo, na Tabela 4.2 para 65% de imunes (que, na prática, aparecem no conjunto de dados como observações censuradas) todos os parâmetros são bem estimados. No entanto, a Tabela 4.3 mostra que, em uma situação similar, quando o percentual de imunes é de 45% e o percentual de censuras entre os não imunes é de 30%, o que resulta em um percentual de censuras entre imunes e não imunes de aproximadamente 62%[1], as estimativas para os parâmetros β_2 e γ apresentam viés acentuado.

[1]Com base na relação (4.1), tem-se na Tabela 4.3 que com $pc_1 = 0,3$ e $\pi = 0,45$ obtem-se $pc_2 = 0,3 \times 0,55 + 0,45 = 0,615$.

Tabela 4.2: Estimativas (média), erros-padrão (EP) e raiz dos erros quadráticos médios (REQM) dos parâmetros em 1500 réplicas considerando amostras sem omissão e sem censura entre os não imunes, variando o tamanho da amostra e o percentual de imunes.

Parâmetro	$n = 50$			$n = 150$			$n = 300$		
	média	EP	REQM	média	EP	REQM	média	EP	REQM
10% de imunes									
$\beta_1 = 2.0$	2.154	0.3202	0.3552	2.050	0.1635	0.1709	2.021	0.1119	0.1139
$\beta_2 = 3.0$	3.219	0.5261	0.5697	3.064	0.2768	0.2840	3.033	0.1774	0.1804
$\beta_3 = 2.0$	2.145	0.4352	0.4589	2.046	0.2230	0.2276	2.019	0.1531	0.1543
$\rho = 2.0$	2.149	0.2535	0.2915	2.041	0.1311	0.1374	2.019	0.0905	0.0924
$\gamma = -2.8$	-2.962	0.5232	0.5565	-2.830	0.2735	0.2795	-2.799	0.1846	0.1865
45% de imunes									
$\beta_1 = 1.5$	1.634	0.3444	0.3691	1.547	0.1698	0.1763	1.519	0.1185	0.1200
$\beta_2 = -1.5$	-1.608	0.4807	0.4927	-1.534	0.2710	0.2731	-1.514	0.1813	0.1818
$\beta_3 = 1.0$	1.095	0.4210	0.4315	1.031	0.2224	0.2247	1.012	0.1570	0.1575
$\rho = 2.0$	2.161	0.3453	0.3812	2.048	0.1783	0.1847	2.019	0.1221	0.1235
$\gamma = -2.8$	-3.001	0.5395	0.5887	-2.854	0.2856	0.2970	-2.804	0.2004	0.2029
65% de imunes									
$\beta_1 = -0.5$	-0.511	0.3261	0.3263	-0.503	0.1719	0.1719	-0.497	0.1426	0.1426
$\beta_2 = -1.5$	-1.605	0.6953	0.7033	-1.516	0.4078	0.4081	-1.490	0.3467	0.3470
$\beta_3 = -0.2$	-0.201	0.5495	0.5495	-0.220	0.3003	0.3009	-0.213	0.2590	0.2594
$\rho = 2.0$	2.189	0.5322	0.5648	2.047	0.2911	0.2948	2.022	0.2595	0.2604
$\gamma = -2.8$	-3.054	0.8098	0.8574	-2.851	0.4355	0.4426	-2.813	0.3993	0.4014

Tabela 4.3: Estimativas (média), erros-padrão (EP) e raiz dos erros quadráticos médios (REQM) dos parâmetros em 1500 réplicas, amostras de tamanho $n = 300$ sem omissão nas covariáveis, variando o percentual de censura entre os não imunes e de imunes.

Parâmetro	10% de censura			30% de censura			50% de censura		
	média	EP	REQM	média	EP	REQM	média	EP	REQM
10% de imunes									
$\beta_1 = 2.0$	2.023	0.1203	0.1224	2.027	0.1311	0.1339	2.127	0.1660	0.2092
$\beta_2 = 3.0$	2.979	0.1937	0.1948	2.935	0.2364	0.2453	3.363	0.3584	0.5099
$\beta_3 = 2.0$	1.948	0.1667	0.1746	1.840	0.1895	0.2481	2.097	0.2656	0.2826
$\rho = 2.0$	2.028	0.0956	0.0996	2.024	0.1083	0.1108	1.944	0.1351	0.1463
$\gamma = -2.8$	-2.690	0.2065	0.2222	-2.573	0.2587	0.3269	-3.357	0.4063	0.7119
45% de imunes									
$\beta_1 = 1.5$	1.529	0.1184	0.1219	1.505	0.1370	0.1371	1.686	0.1816	0.2598
$\beta_2 = -1.5$	-1.587	0.1847	0.2041	-1.765	0.1967	0.3302	-1.779	0.2626	0.3830
$\beta_3 = 1.0$	0.939	0.1583	0.1697	0.727	0.1687	0.3211	1.025	0.3082	0.3092
$\rho = 2.0$	2.040	0.1258	0.1320	2.167	0.1516	0.2258	1.815	0.2168	0.2852
$\gamma = -2.8$	-2.729	0.1984	0.2032	-2.502	0.2238	0.3510	-2.914	0.3898	0.4144
65% de imunes									
$\beta_1 = -0.5$	-0.523	0.1088	0.1111	-0.531	0.1235	0.1274	-0.557	0.1353	0.1468
$\beta_2 = -1.5$	-1.600	0.2221	0.2434	-1.722	0.2391	0.3264	-1.973	0.2892	0.5545
$\beta_3 = -0.2$	-0.303	0.1721	0.2006	-0.446	0.1925	0.3122	-0.749	0.2063	0.5868
$\rho = 2.0$	2.024	0.1587	0.1606	2.061	0.1842	0.1941	2.240	0.2563	0.3512
$\gamma = -2.8$	-2.724	0.2470	0.2517	-2.613	0.2779	0.3205	-2.403	0.3255	0.4923

4.2.2 Efeito das omissões nas estimativas dos parâmetros

Analisando a Tabela 4.4, que mostra os resultados para o caso 3, observa-se que as estimativas dos parâmetros obtidas através do algoritmo *EM*, em geral, apresentam com desempenho, mesmo quando o percentual de omissão é alto. O mesmo não é observado para as estimativas obtidas com a análise de casos completos (ACC), que apresentam vicios crescentes de acordo o percentual de omissão. Observa-se também que as estimativas obtidas com ACC são ainda piores quando o percentual de imunes é médio ou alto (45% e 60%).

A Tabela 4.5 mostra que, na presença de censura para não imunes (caso 4), o algoritmo *EM*, em geral, não apresenta bom desempenho quanto à estimativa dos parâmetros. No entanto, as estimativas referentes à análise de casos completos apresentam resultados muito piores e grandes vieses para todos os parâmetros envolvidos no modelo, mesmo quando os percentuais de omissão e de censura são baixos.

Mesmo para o caso sem omissão, foram obtidos resultados insatisfatórios quanto à estimativa dos parâmetros na presença de censura entre os não imunes. Portanto, os vieses encontrados nas estimativas obtidas através do algoritmo *EM* na Tabela 4.5 podem ser devido à presença de censuras e não a dados omissos.

Para visualizar o efeito das omissões nas estimativas dos parâmetros, na raíz quadrada dos erros quadráticos médios e nas probabilidades de cobertura[2] a um nível nominal de 95%, variamos o percentual de dados omissos entre 5 e 50%. Os resultados são mostrados nas Figuras (4.1), (4.2), (4.3), (4.4) e (4.5). Podemos ver que, para os parâmetros β_1, ρ e γ, o uso do algoritmo *EM* resulta em vieses relativos[3] bem próximos de zero, erros quadráticos médios constantes em função do percentual de omissão e probabilidades de cobertura próximas ao valor nominal de 95%, enquanto o método ACC fornece vieses e erros quadráticos médios crescentes em função do percentual de omissão e probabilidades de cobertura insatisfatórias, principalmente

[2]Probabilidade de cobertura: Percentual de intervalos de confiança que contém o verdadeiro valor do parâmetro.

[3]Viés relativo: O resultado da diferença entre a estimativa e o verdadeiro valor do parâmetro divido pelo verdadeiro valor do parâmetro.

Tabela 4.4: Estimativas (média), erros-padrão (EP) e raiz dos erros quadráticos médios (REQM) dos parâmetros em 1500 réplicas, amostras de tamanho $n = 300$ sem censura entre não imunes, variando o percentual de omissão e de imunes.

Parâmetro	Método	15 % de omissão			30% de omissão			50% de omissão		
		média	EP	REQM	média	EP	REQM	média	EP	REQM
10% de imunes										
$\beta_1 = 2.0$	*EM*	2.020	0.1164	0.1180	2.030	0.1280	0.1314	2.026	0.1416	0.1440
	ACC	2.169	0.1269	0.2112	2.018	0.1590	0.1599	1.969	0.2082	0.2105
$\beta_2 = 3.0$	*EM*	3.033	0.1839	0.1869	3.037	0.2051	0.2083	3.027	0.2081	0.2099
	ACC	3.108	0.1970	0.2244	3.134	0.2468	0.2807	3.179	0.2940	0.3440
$\beta_3 = 2.0$	*EM*	2.020	0.1624	0.1636	2.029	0.1715	0.1739	2.038	0.1868	0.1907
	ACC	2.079	0.1689	0.1864	2.068	0.1956	0.2070	2.060	0.2404	0.2477
$\rho = 2.0$	*EM*	2.018	0.0932	0.0949	2.027	0.0964	0.0999	2.024	0.1009	0.1037
	ACC	2.082	0.1024	0.1311	2.104	0.1181	0.1575	2.173	0.1520	0.2305
$\gamma = -2.8$	*EM*	-2.800	0.1948	0.1967	-2.809	0.2082	0.2115	-2.808	0.2150	0.2179
	ACC	-2.979	0.2048	0.2907	-3.004	0.2365	0.3309	-3.191	0.2796	0.5028
$\alpha_1 = -0.5$	*EM*	-0.509	0.1406	0.1409	0.508	0.1676	0.1678	-0.506	0.1890	0.1891
	ACC	-	-	-	-	-	-	-	-	-
$\alpha_2 = 1.0$	*EM*	1.015	0.1856	0.1862	1.030	0.2092	0.2113	1.019	0.2444	0.2452
	ACC	-	-	-	-	-	-	-	-	-
45% de imunes										
$\beta_1 = 1.5$	*EM*	1.524	0.1185	0.1210	1.526	0.1206	0.1232	1.521	0.1320	0.1337
	ACC	1.535	0.1399	0.1441	1.288	0.1864	0.2820	0.977	0.4977	0.7222
$\beta_2 = -1.5$	*EM*	-1.531	0.1788	0.1815	-1.524	0.1924	0.1939	-1.535	0.2300	0.2327
	ACC	-1.635	0.1985	0.2398	-1.707	0.2567	0.3301	-2.060	1.1000	1.2344
$\beta_3 = 1.0$	*EM*	1.024	0.1575	0.1593	1.016	0.1749	0.1756	1.143	1.3709	1.3783
	ACC	1.010	0.16941	0.1697	0.855	0.2263	0.2687	0.873	1.5335	1.5387
$\rho = 2.0$	*EM*	2.031	0.1150	0.1191	2.029	0.1216	0.1250	2.020	0.1463	0.1476
	ACC	2.104	0.1355	0.1708	2.810	0.2671	0.8529	2.542	1.8625	1.9398
$\gamma = -2.8$	*EM*	-2.821	0.1896	0.1957	-2.810	0.1978	0.2014	-2.923	1.3629	1.3713
	ACC	-2.998	0.2216	0.3164	-4.355	0.4059	1.6332	-6.579	2.6387	4.6316
$\alpha_1 = -0.5$	*EM*	-0.510	0.1518	0.1521	-0.517	0.1882	0.1890	-0.514	0.2316	0.2321
	ACC	-	-	-	-	-	-	-	-	-
$\alpha_2 = 1.0$	*EM*	1.003	0.1781	0.1781	1.014	0.1988	0.1992	1.022	0.2437	0.2447
	ACC	-	-	-	-	-	-	-	-	-
65% de imunes										
$\beta_1 = -0.5$	*EM*	-0.506	0.1075	0.1077	-0.511	0.1099	0.1105	-0.513	0.1146	0.1153
	ACC	-0.881	0.1617	0.4139	-1.128	0.2727	0.6843	-0.118	0.5238	0.6486
$\beta_2 = -1.5$	*EM*	-1.526	0.2160	0.2175	-1.526	0.2187	0.2202	-1.525	0.2365	0.2378
	ACC	-1.827	0.2956	0.4411	-2.064	0.5273	0.7717	-2.247	1.2015	1.4150
$\beta_3 = -0.2$	*EM*	-0.190	0.1892	0.1895	-0.198	0.2000	0.2000	-0.233	0.7143	0.7150
	ACC	-0.621	0.2586	0.4942	-0.858	0.6418	0.9185	-1.013	1.0366	1.3173
$\rho = 2.0$	*EM*	2.027	0.1542	0.1566	2.035	0.1571	0.1609	2.035	0.1504	0.1543
	ACC	2.657	0.3441	0.7412	2.621	1.1807	1.3339	1.909	0.9150	0.9195
$\gamma = -2.8$	*EM*	-2.809	0.2511	0.2537	-2.826	0.2531	0.2587	-2.818	0.2497	0.2538
	ACC	-4.243	0.5275	1.5620	-5.446	1.1924	2.9274	-3.715	0.9534	1.3402
$\alpha_1 = -0.5$	*EM*	-0.509	0.1473	0.1476	-0.505	0.1596	0.1596	-0.514	0.2156	0.2160
	ACC	-	-	-	-	-	-	-	-	-
$\alpha_2 = 1.0$	*EM*	1.006	0.1887	0.1888	1.021	0.1922	0.1933	1.010	0.2711	0.2713
	ACC	-	-	-	-	-	-	-	-	-

Tabela 4.5: Estimativas (média), erros-padrão (EP) e raiz dos erros quadráticos médios (REQM) dos parâmetros em 1500 réplicas, amostras de tamanho $n = 300$ sem censura entre os não imunes, com 45% de imunes e variando o percentual de censura e de omissão.

10% de censura									
Parâmetro	Método	15 % de omissão			30% de omissão			50% de omissão		
		média	EP	REQM	média	EP	REQM	média	EP	REQM
$\beta_1 = 1.5$	*EM*	1.516	0.1144	0.1155	1.511	0.1181	0.1186	1.512	0.1176	0.1182
	ACC	1.669	0.1443	0.2219	1.817	0.1737	0.3613	1.778	0.2033	0.3441
$\beta_2 = -1.5$	*EM*	-1.586	0.2106	0.2274	-1.580	0.2068	0.2217	-1.579	0.2196	0.2333
	ACC	-1.740	0.2417	0.3408	-1.920	0.2771	0.5033	-1.997	0.3236	0.5927
$\beta_3 = 1.0$	*EM*	0.897	0.1501	0.1820	0.894	0.1541	0.1873	0.891	0.1711	0.2031
	ACC	0.859	0.1702	0.2212	0.685	0.1959	0.3706	0.647	0.2287	0.4205
$\rho = 2.0$	*EM*	2.045	0.1182	0.1266	2.042	0.1227	0.1298	2.044	0.1155	0.1235
	ACC	2.316	0.1596	0.3538	2.193	0.1779	0.2627	2.148	0.2049	0.2529
$\gamma = -2.8$	*EM*	-2.708	0.1902	0.2008	-2.704	0.1948	0.2064	-2.703	0.1977	0.2096
	ACC	-3.282	0.2460	0.6245	-3.224	0.2942	0.5391	-3.060	0.3489	0.4520
$\alpha_1 = -0.5$	*EM*	-0.511	0.1509	0.1514	-0.497	0.1479	0.1479	-0.499	0.1772	0.1772
	ACC	-	-	-	-	-	-	-	-	-
$\alpha_2 = 1.0$	*EM*	1.007	0.2408	0.2409	0.984	0.2478	0.2484	0.978	0.2845	0.2854
	ACC	-	-	-	-	-	-	-	-	-
30% de censura										
$\beta_1 = 1.5$	*EM*	1.503	0.1314	0.1314	1.491	0.1330	0.1333	1.456	0.1030	0.1121
	ACC	1.672	0.1533	0.2305	1.927	0.2370	0.4881	2.364	0.5565	1.0278
$\beta_2 = -1.5$	*EM*	-1.737	0.2081	0.3150	-1.698	0.2204	0.2962	-1.604	0.1832	0.2106
	ACC	-1.895	0.2334	0.4584	-2.320	0.3270	0.8823	-2.779	0.8282	1.5237
$\beta_3 = 1.0$	*EM*	0.735	0.1773	0.3186	0.693	0.2049	0.3692	0.582	0.1899	0.4594
	ACC	0.726	0.1905	0.3341	0.555	0.2741	0.5231	0.499	0.9670	1.0891
$\rho = 2.0$	*EM*	2.140	0.1474	0.2029	2.130	0.1483	0.1974	2.092	0.1111	0.1442
	ACC	2.390	0.1864	0.4326	2.573	0.2955	0.6448	2.723	1.4720	1.6400
$\gamma = -2.8$	*EM*	-2.519	0.2198	0.3354	-2.491	0.2241	0.3596	-2.419	0.1687	0.3921
	ACC	-2.997	0.2624	0.3449	-3.547	0.4467	0.8937	-5.589	1.2442	3.0785
$\alpha_1 = -0.5$	*EM*	-0.510	0.1438	0.1442	-0.523	0.1651	0.1666	-0.581	0.1624	0.1813
	ACC	-	-	-	-	-	-	-	-	-
$\alpha_2 = 1.0$	*EM*	0.972	0.1790	0.1812	0.906	0.1980	0.2192	0.792	0.1778	0.2736
	ACC	-	-	-	-	-	-	-	-	-

para os parâmetros β_1 e ρ.

O parâmetro β_2 é bem estimado pelo algoritmo *EM* para amostras com pequeno percentual de imunes (10%), mas quando o percentual de imunes é maior (45%), observa-se um pequeno viés que permanece constante em função do percentual de omissão. Quando o método utilizado é a ACC, observa-se que as estimativas, bem como probabilidade de cobertura pioram com o aumento do percentual de omissão.

Curiosamente, os resultados obtidos com os dois métodos para o parâmetro β_3, que é o coeficiente associado à covariável omissa, são equivalentes para o caso em que há 10% de imunes na amostra. No entanto, apesar do viés relativo para a estimativa ACC estar sempre mais próximo de zero, do que as estimativas *EM*, pode-se notar que esta quantidade apresenta uma tendência crescente com o aumento do percentual de omissão. Além disso, os vieses observados com a ACC é negativo quando o percentual de omissão é de no máximo 15% e positivo para percentuais de omissão maior do que 15%, enquanto o algoritmo *EM* apresenta vieses sempre negativos. Quando o percentual de imunes é de 45%, as estimativas *EM* apresentam viés acentuado, porém constante, enquanto as estimativas ACC apresentam vieses muito maiores e crescentes segundo o percentual de omissão.

Podemos observar que, quando o percentual de imunes é de 45%, a probabilidade de cobertura para o método ACC decresce e depois cresce em função do percentual de omissão. Isto provavelmente deve-se ao fato de que com o aumento do percentual de omissão, os erros padrão tornam-se elevados e consequentemente, os intervalos de confiança resultantes são bastante amplos e cobrem o verdadeiro valor do parâmetro, mas perdem o sentido prático por serem imprecisos.

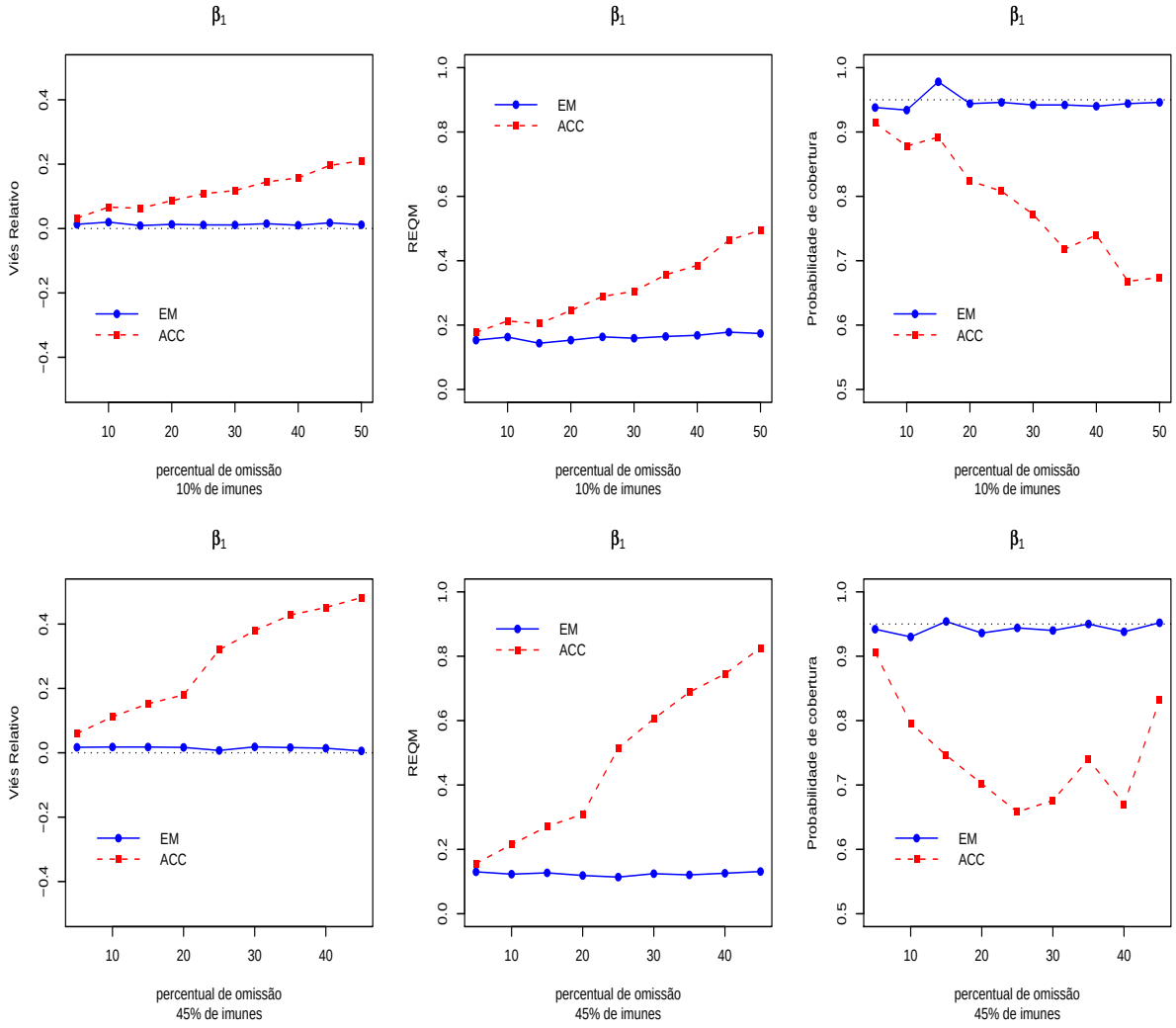

Figura 4.1: Viés relativo, raíz do erro quadrático médio e probabilidades de cobertura para estimativas do parâmetro β_1 através dos métodos *EM* e ACC para 10% e 45% de imunes - 500 simulações de amostras de tamanho $n = 200$ e 15% de censura entre não imunes.

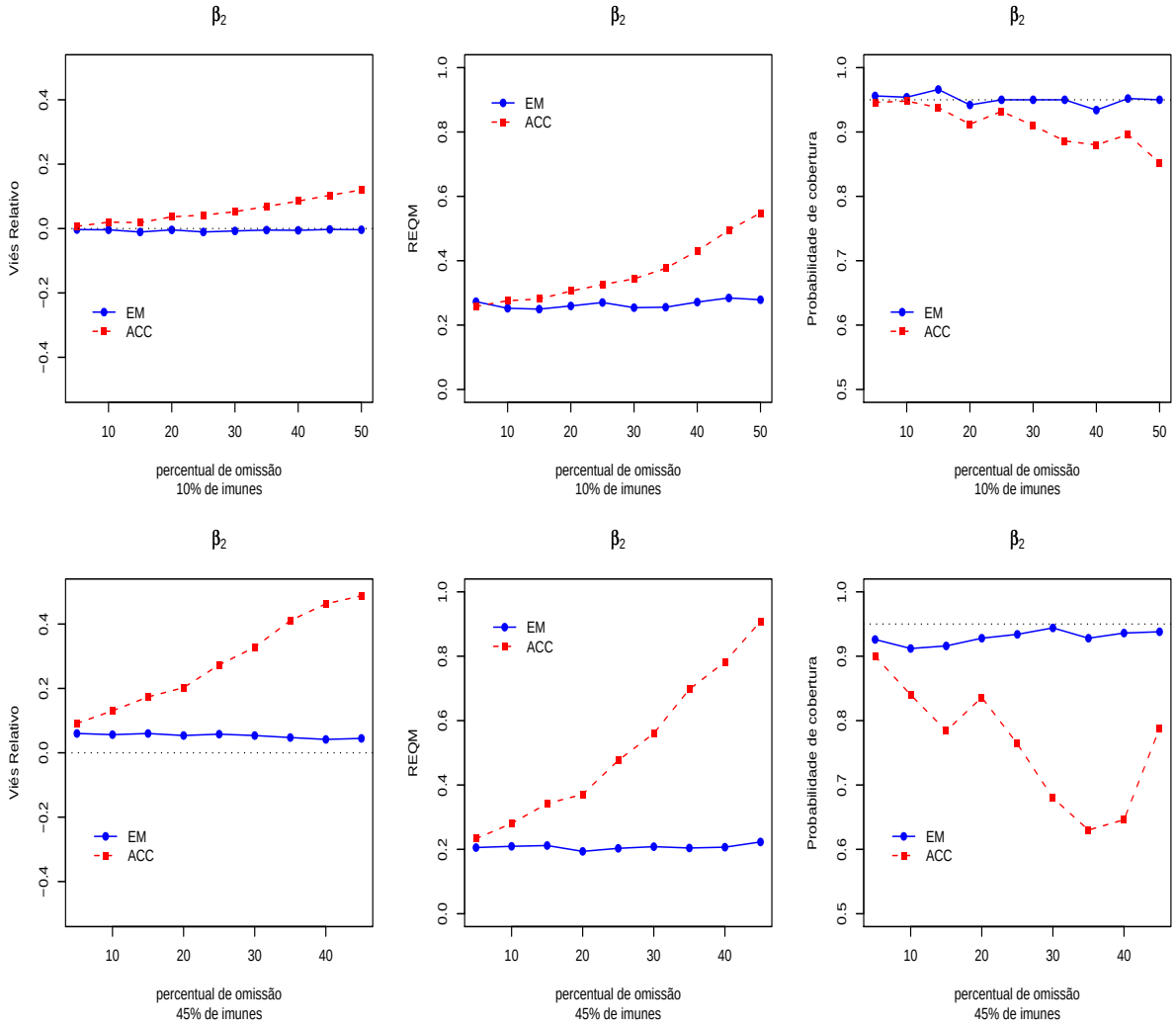

Figura 4.2: Viés relativo, raíz do erro quadrático médio e probabilidades de cobertura para estimativas do parâmetro β_2 através dos métodos *EM* e ACC para 10% e 45% de imunes - 500 simulações de amostras de tamanho $n = 200$ e 15% de censura entre não imunes.

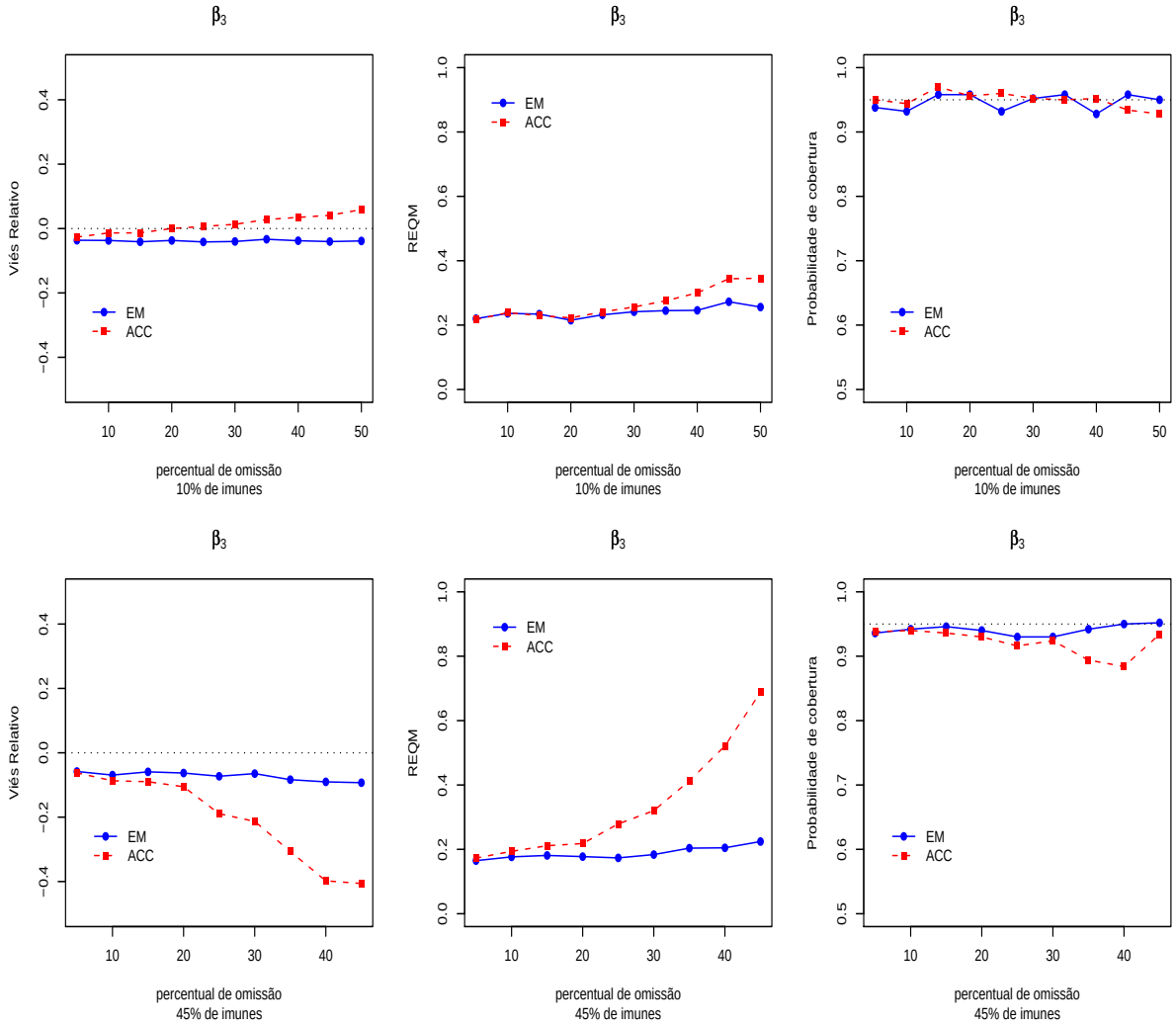

Figura 4.3: Viés relativo, raíz do erro quadrático médio e probabilidades de cobertura para estimativas do parâmetro β_3 através dos métodos *EM* e ACC para 10% e 45% de imunes - 500 simulações de amostras de tamanho $n = 200$ e 15% de censura entre não imunes.

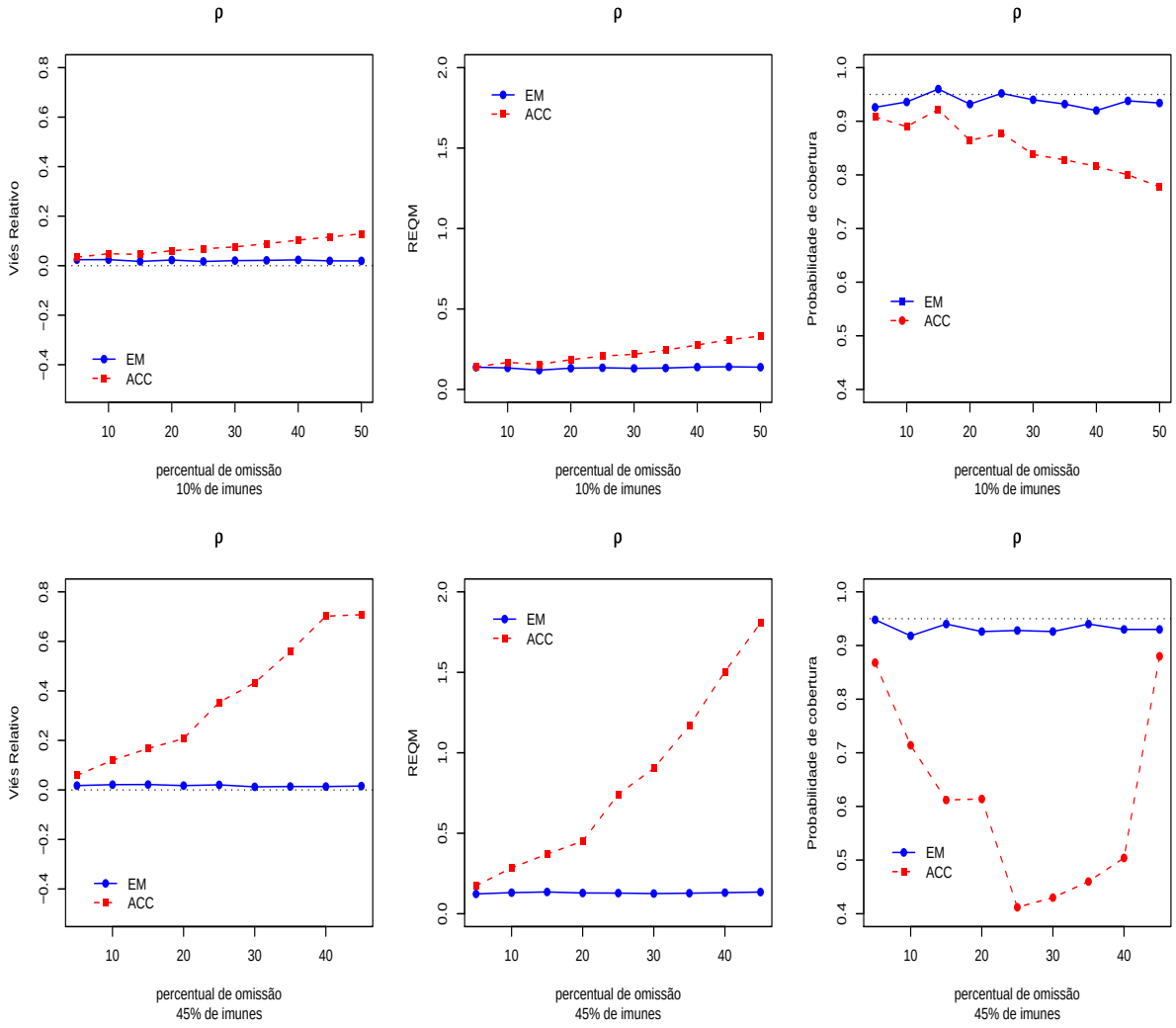

Figura 4.4: Viés relativo, raíz do erro quadrático médio e probabilidades de cobertura para estimativas do parâmetro ρ através dos métodos *EM* e ACC para 10% e 45% de imunes - 500 simulações de amostras de tamanho $n = 200$ e 15% de censura entre não imunes.

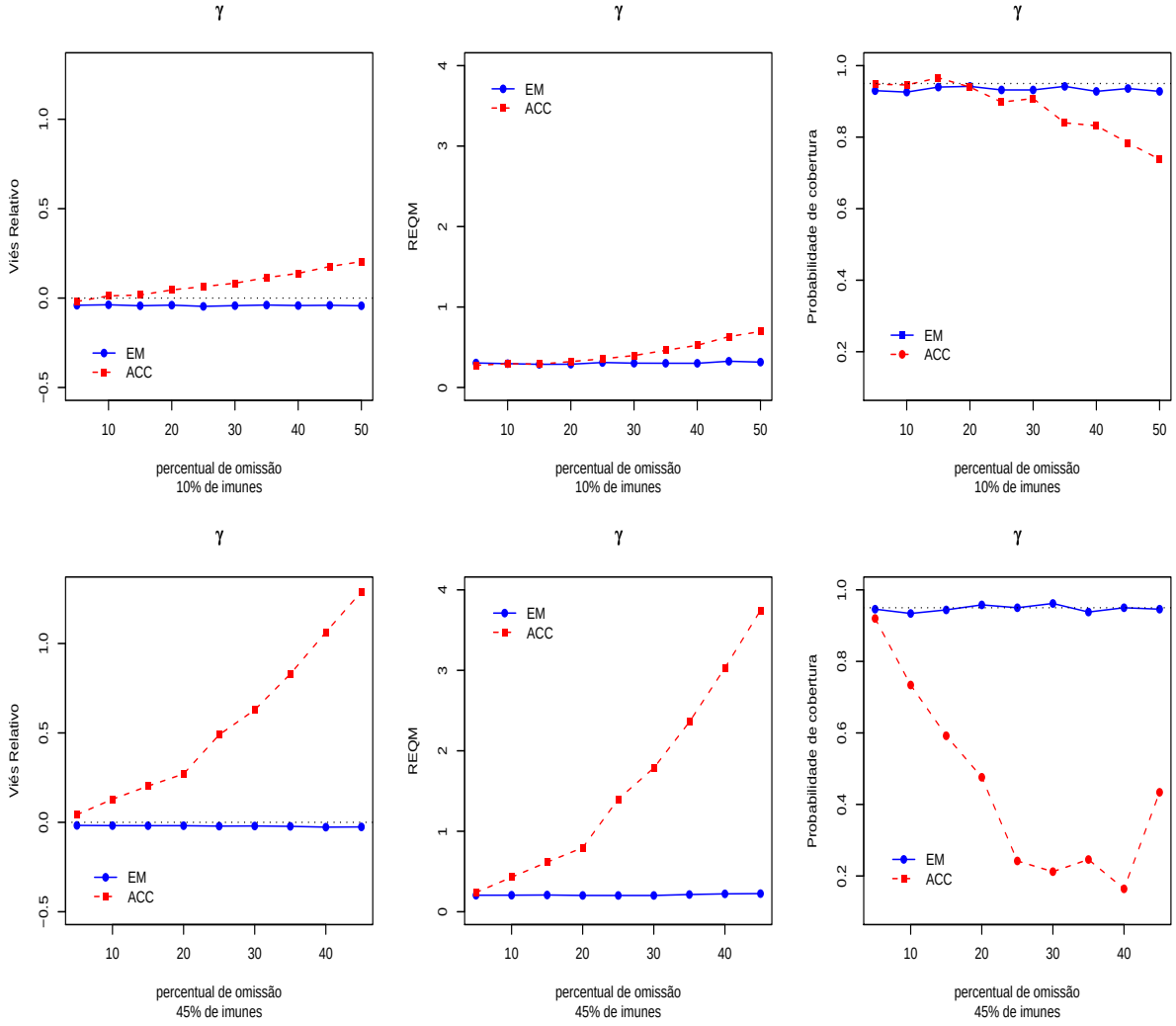

Figura 4.5: Viés relativo, raíz do erro quadrático médio e probabilidades de cobertura para estimativas do parâmetro γ através dos métodos *EM* e ACC para 10% e 45% de imunes - 500 simulações de amostras de tamanho $n = 200$ e 15% de censura entre não imunes.

Capítulo 5

Aplicação

Para elucidar o método de estimação de máxima verossimilhança via algoritmo *EM*, consideramos um conjunto de dados reais, referentes ao tempo até a conclusão do curso de graduação em Estatística da Universidade Federal do Rio Grande do Norte de 414 alunos ingressos entre os anos de 1997 e 2004 no referido curso, que no momento da inscrição para o vestibular responderam a um questionário sócio-econômico contendo 25 questões, das quais foram escolhidas cinco para modelar a fração de cura. Os dados foram obtidos de Freire & Valença (2006).

5.1 Ajuste do modelo aos dados de evasão escolar

A Figura 1.1 exibida no Capítulo 1 mostra a curva de sobrevivência empírica (Kaplan-Meier) para estes dados. Observamos que a cauda direita alcança um patamar acima de zero por um período considerável, o que caracteriza o comportamento de uma função de sobrevivência imprópria.

Para este exemplo, o indivíduo imune ao evento é aquele aluno que nunca concluirá o curso de estatística, seja porque mudou de cidade, foi aprovado em outro vestibular, etc. Portanto o termo "fração de cura" pode ser substituído aqui por "taxa de evasão".

Consideramos no modelo apenas as variáveis:

- Gênero: $x_1 = \begin{cases} 1 & \text{feminino} \\ 0 & \text{masculino} \end{cases}$
- Idade: $x_2 = \begin{cases} 1 & \text{menos de 24 anos} \\ 0 & \text{24 anos ou mais} \end{cases}$
- Transporte: $x_3 = \begin{cases} 1 & \text{transporte coletivo} \\ 0 & \text{carro ou outro meio de transporte} \end{cases}$
- Grau de instrução da mãe: $x_4 = \begin{cases} 1 & \text{até nível médio completo} \\ 0 & \text{superior ou mais} \end{cases}$
- Moradia: $x_5 = \begin{cases} 1 & \text{mora com os pais} \\ 0 & \text{mora só, com amigos ou com parentes} \end{cases}$

Dentre as cinco covariáveis analisadas, apenas duas, sexo (x_1) e grau de instrução da mãe (x_4), foram significativas segundo o teste de Wald. A variável sexo foi observada para todos os indivíduos, enquanto a variável grau de instrução da mãe foi omitida em $12,1\%$ dos casos.

Os resultados para estes dados utilizando o modelo com fração de cura e os dois métodos abordados nesta dissertação para tratar dados com omissão, a análise de casos completos e estimação de máxima verossimilhança via algoritmo *EM* são mostrados a seguir. Consideramos que os tempos de promoção, Z_{ij} têm distribuição $Weibull(\rho, \gamma)$, $i = 1, \ldots, n$ e $j = 1, \ldots, M_i$.

Assumimos que o mecanismo que gera as omissões é do tipo MAR, ou seja, o mecanismo de omissão é ignorável. Como apenas a variável x_2 tem valores omissos, consideramos a distribuição conjunta somente para esta covariável dado x_1. Um possível modelo para a distribuição condicional unidimensional é o modelo logístico (bem como outras funções de ligação). Portanto, o modelo considerado para a distribuição conjunta das covariáveis é dado por

$$p(x_{i4}|x_{i1}, \boldsymbol{\alpha}_i) = \frac{\exp(\alpha_0 + \alpha_1 x_{i1})}{1 + \exp(\alpha_0 + \alpha_1 x_{i1})}. \tag{5.1}$$

Como o percentual de dados omissos na amostra é pequeno, espera-se que a perda de informação provocada pelo uso da análise de casos completos não seja desastrosa. No entanto, o percentual de censuras é de $69,6\%$, o que pode provocar viés nas estimativas dos parâmetros, como mostrado atavés de simulações (ver Capítulo 4).

Excluindo-se as unidades com observações omissas, restam 365 alunos na amostra. As estimativas dos parâmetros para o modelo com fração de cura e erros-padrão são mostrados na Tabela 5.1.

Observamos que as estimativas para os dois métodos são similares. Entretanto, os erros padrão dos parâmetros ρ e γ são cerca de duas vezes maiores quando o método utilizado é a ACC.

Tabela 5.1: Estimativas para os dados de evasão no curso de estatística da UFRN no período de 1997 a 2004.

Variável	Método	Estimativa	Erro Padrão	Valor p
sexo	ACC	0,387	0,172	0,012
	EM	0,488	0,164	0,001
instrução	ACC	-0,822	0,142	$<$ 0,001
materna	*EM*	-0,909	0,136	$<$ 0,001
ρ	ACC	4,208	0,283	-
	EM	4,329	0,142	-
γ	ACC	-10,535	0,701	-
	EM	-10,833	0,368	-

A partir dos resultados mostrados na Tabela 5.1, pode-se calcular a taxa de evasão para o indivíduo i que é dada por

$$\pi_{ACC}(x_i) = \exp\big[-\exp(0,387x_{i1} - 0,822x_{i2})\big] \tag{5.2}$$

para a análise de casos completos, e

$$\pi_{EM}(x_i) = \exp\big[-\exp(0,488x_{i1} - 0,909x_{i2})\big] \tag{5.3}$$

para o algoritmo *EM*.

Podemos interpretar os resultados do modelo ajustado para esta amostra através das taxas de evasão estimadas que são dadas na Tabela 5.2.

Tabela 5.2: Estimativas para taxa de evasão.

Sexo	Instrução materna	Taxa de evasão (%)	
		ACC	*EM*
feminino	superior ou mais	$24,4$	$19,6$
masculino	superior ou mais	$36,8$	$36,8$
feminino	até médio	$52,4$	$51,9$
masculino	até médio	$64,0$	$66,8$

Um aluno que teve sua taxa de evasão estimada em $24,4\%$ e $19,6\%$ pela ACC e pelo algoritmo *EM*, respectivamente, as menores taxas estimadas para esta amostra, tem as seguintes características: é do sexo feminino e sua mãe tem nível superior ou mais.

Suponha agora um aluno do sexo masculino cuja mãe tem nível superior ou mais, a taxa de evasão estimada para este aluno é de $36,8\%$ e $36,8\%$ pela ACC e pelo algoritmo *EM*, respectivamente, ou seja, as mulheres têm mais chances de concluir o curso de estatística do que os homens.

O indivíduo que apresentou as maiores taxas de evasão estimadas na amostra, $64,0\%$ e $66,8\%$ pela ACC e pelo algoritmo *EM*, respectivamente, é do sexo masculino, e sua mãe tem grau de instrução de, no máximo, nível médio completo.

Através da Figura 5.1, que mostra a estimativa para a sobrevivência média pelos métodos ACC e *EM*, podemos ver que estas estimativas são quase idênticas, no entanto, se afastam da curva de Kaplan-Meier na cauda direita, o que mostra que os dois métodos de estimação utilizados subestimaram a taxa de evasão. Pode-se portanto, suspeitar da validade da suposição que o número de causas M_i tenha uma distribuição de Poisson.

Neste caso, observamos que, devido ao baixo percentual de dados omissos, as quantidades estimadas através dos dois métodos podem levar a estimativas muito próximas para a taxa de evasão, embora a ACC tenha super-estimado a menor taxa e subestimado o pior cenário.

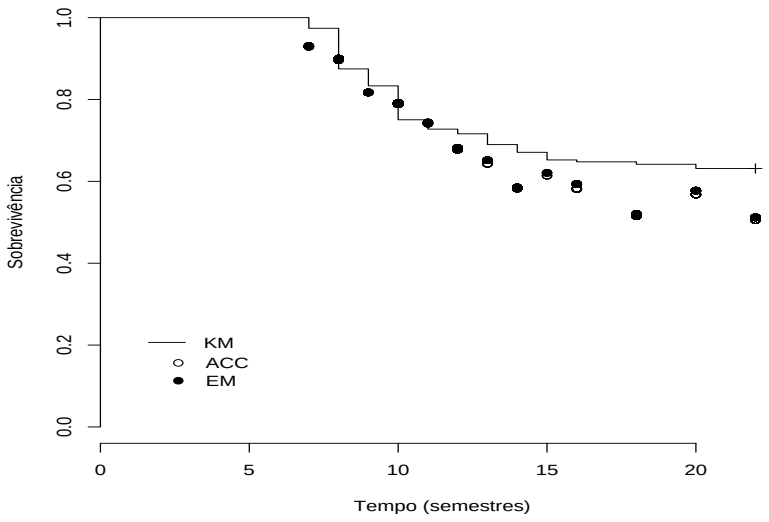

Figura 5.1: Estimativa para a função de sobrevivência pelos métodos Kaplan-Meier, ACC e *EM*.

Capítulo 6

Considerações Finais

6.1 Conclusão

Abordamos nesta dissertação modelos de sobrevivência para o caso em que existe uma proporção significativa de curados ou imunes ao evento de interesse na população. Além disso, discutimos um método de estimação dos parâmetros do modelo com fração de cura quando algumas observações apresentam omissões nas covariáveis.

Através de um estudo de simulação de Monte Carlo, avaliamos o desempenho das estimativas obtidas com o uso do algoritmo *EM* e da ACC, variando-se o percentual de censuras, de omissão e o tamanho da amostra.

Com os resultados das simulações, vimos que, com o incremento do percentual de censuras entre indivíduos não imunes, as estimativas dos parâmetros tanto na análise de casos completos quanto no algoritmo *EM* apresentam vícios crescentes, sendo que o algoritmo *EM* fornece vícios bem menores.

Analisando amotras em que apenas os indivíduos imunes são censurados, observamos que as estimativas obtidas através do algoritmo *EM*, em geral, apresentam bom desempenho, enquanto aquelas obtidas com a análise de casos completos apresentam vícios que tornam-se maiores com o aumento do percentual de imunes e de dados omissos.

Em geral, o algoritmo *EM* apresentou melhor desempenho do que a análise de casos completos, o que mostra que a exclusão de observações omissas pode prejudicar às análises de dados e dependendo do percentual de dados omissos, as perdas podem ser ainda maiores.

A análise do conjunto de dados referentes ao tempo até a conclusão do curso de estatística da UFRN mostrou a adequabilidade do modelo a contextos não biológicos, conforme era usualmente descrito na literatura. Neste caso a fração de cura é interpretada como taxa de evasão e os resultados do modelo mostram-se coerentes com as taxas observadas de evasão no curso considerado.

6.2 Pesquisas futuras

Nas simulações realizadas neste trabalho foi considerado que apenas as covariáveis categóricas apresentam dados omissos e que o mecanismo que gera as omissões é do tipo MAR.

Em pesquisas futuras pode-se permitir que tanto covariáveis categóricas quanto contínuas apresentem omisão e assim, verificar quais vantagens obtém-se com o uso do algoritmo *EM* sobre a análise de casos completos.

Este trabalho pode ser extendido também para o caso em que o mecanismo de omissão é MNAR, ou seja, não ignorável.

Como foi visto através do estudo de simulação, censuras entre indivíduos não imunes gera vieses e imprecisão nas estimativas dos parâmetros pelos dois métodos de estimação utilizados, com isso, vê-se a necessidade de desenvolver um modelo que discrimine indivíduos imunes e censurados.

Na prática, modelar covariáveis binárias utilizando o modelo logístico (ou qualquer outro paramétrico) pode não ser adequado. Neste caso, mesmo as estimativas sob o algoritmo *EM* podem ser tendenciosas. Pode-se em pesquisas futuras, realizar um estudo de simulação em que covariáveis omissas sejam geradas através de um modelo e, no momento da estimação, seja considerado um outro modelo (paramétrico ou não

paramétrico). Assim, pode-se saber se a mudança na função de ligação na modelagem das covariáveis afeta as estimativas dos parâmetros e o quanto as análises podem ser prejudicadas.

Considerar omissões no modelo unificado de fração de cura (Rodrigues & Louzada-Neto 2006) que permite o uso de outras distribuições de probabilidade para o número de causas ou riscos.

Um outro tema de estudo é considerar que as covariáveis, além de apresentarem omissão, estão sujeitas a erro de medição.

Apêndice A

Obtenção da Função de Verossimilhança

A.1 Verossimilhança

Mostraremos nesta seção, os cálculos para a obtenção da função de verossimilhança (2.7).

Vimos que $y_i = min\{T_i, C_i\}$, com $T_i = min\{Z_{i0}, Z_{i1}, \ldots, Z_{i,M_i}\}$, sendo $M_i \sim Poisson(\theta)$.

Considere f_T e S_T as funções densidade e de sobrevivência de T_i, respectivamente, e g e G as funções densidade e de sobrevivência de C_i, respectivamente, para $i = 1, \ldots, n$.

Então, a função de sobrevivência condicional dos tempos de falha, T_i, é

$$\begin{aligned} S_T(t|m_i) = P(T_i > t|M_i = m_i) &= P(min\{Z_{i0}, Z_{i1}, \ldots, Z_{i,m_i}\} > t) \\ &= P(Z_{i0} > t, Z_{i1} > t, \ldots, Z_{i,m_i} > t) \end{aligned}$$

Sabemos que, dado M_i, as variáveis aleatórias Z_i são *iid* com função de sobrevivência $S(t|\boldsymbol{\lambda})$, além disso, $P(Z_{i0} > t) = 1$

Assim,

$$\begin{aligned} S_T(t|m_i) &= P(Z_{i0} > t)P(Z_{i1} > t)\dots P(Z_{i,m_i} > t) \\ &= 1S(t|\boldsymbol{\lambda})\dots S(t|\boldsymbol{\lambda}) \\ &= S(t|\boldsymbol{\lambda})^{m_i}. \end{aligned}$$

Sabemos ainda que

$$f_T(t|m_i) = -\frac{d}{dt}S_T(t|m_i) = m_i f(t|\boldsymbol{\lambda})S(t|\boldsymbol{\lambda})^{m_i-1},$$

e a distribuição condicional de $(y_i, \delta_i|m_i)$ é obtida como segue

$$\begin{aligned} P(y_i = t, \delta_i = 0|M_i = m_i) &= P(C_i = t, T_i > C_i|M_i = m_i) \\ &= P(T_i > C_i|C_i = t, M_i = m_i)P(C_i = t) \\ &= S_T(t|m_i)g(t) \\ &= S(t|\boldsymbol{\lambda})^{m_i}g(t) \end{aligned}$$

e

$$\begin{aligned} P(y_i = t, \delta_i = 1|M_i = m_i) &= P(T_i = t, T_i \leq C_i|M_i = m_i) \\ &= P(T_i \leq C_i|T_i = t, M_i = m_i)P(T_i = t|M_i = m_i) \\ &= f_T(t|m_i)G(t) \\ &= m_i f(t|\boldsymbol{\lambda})S(t|\boldsymbol{\lambda})^{m_i-1}G(t). \end{aligned}$$

Assim, a distribuição de (y_i, δ_i) dado $M_i - m_i$, supondo que a censura é não informativa, é dada por

$$f(y_i, \delta_i|m_i) = S(t|\boldsymbol{\lambda})^{m_i-\delta_i}\left[m_i f(t|\boldsymbol{\lambda})\right]^{\delta_i} \tag{A.1}$$

Dessa forma, a densidade conjunta de (y_i, δ_i, m_i) pode ser escrita como

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\delta}, \mathbf{m}) &= \prod_{i=1}^{n} f(y_i, \delta_i, m_i) \\ &= \prod_{i=1}^{n} f(y_i, \delta_i | m_i) f(m_i) \\ &= \prod_{i=1}^{n} S(t|\boldsymbol{\lambda})^{m_i - \delta_i} \left[n_i f(t|\boldsymbol{\lambda})\right]^{\delta_i} \frac{\theta^{m_i} e^{-\theta}}{m_i!} \\ &= \prod_{i=1}^{n} S(t|\boldsymbol{\lambda})^{m_i - \delta_i} \left[m_i f(t|\boldsymbol{\lambda})\right]^{\delta_i} exp\left\{\sum_{i=1}^{n} \left[m_i ln\theta - ln(m_i!) - \theta\right]\right\} \end{aligned}$$

Então, a função de verossimilhança dos dados completos é dada por

$$L(\theta, \boldsymbol{\lambda}; \mathcal{D}_c) = \prod_{i=1}^{n} S(t|\boldsymbol{\lambda})^{M_i - \delta_i} \left[M_i f(t|\boldsymbol{\lambda})\right]^{\delta_i} exp\left\{\sum_{i=1}^{n} \left[M_i ln\theta - ln(M_i!) - \theta\right]\right\}$$

Incorporando covariáveis ao modelo através de θ pela função de ligação $\theta_i \equiv \theta(x_i'\beta) = exp(x_i'\beta)$, tem-se

$$L(\boldsymbol{\phi}; \mathcal{D}_c) = \prod_{i=1}^{n} S(t|\boldsymbol{\lambda})^{M_i - \delta_i} \left[M_i f(t|\boldsymbol{\lambda})\right]^{\delta_i} exp\left\{\sum_{i=1}^{n} \left[M_i x_i'\beta - ln(M_i!) - exp(x_i'\boldsymbol{\beta})\right]\right\}$$

A.2 Verossimilhança marginal

Para obter a verossimilhança marginal, fazemos o somatório da distribuição conjunta de (y_i, δ_i, m_i) nas variáveis não observadas m_i.

$$\begin{aligned} f(y_i, \delta_i) &= \sum_{m_i=0}^{\infty} f(y_i, \delta_i, m_i) \\ &= \sum_{m_i=0}^{\infty} f(y_i, \delta_i | m_i) f(m_i) \end{aligned}$$

De (A.1) vem

$$\begin{aligned} f(y_i,\delta_i) &= \sum_{m_i=0}^{\infty} S(y_i|\boldsymbol{\lambda})^{m_i-\delta_i}[m_i f(y_i|\boldsymbol{\lambda})]^{\delta_i}\frac{\theta^{m_i}e^{-\theta}}{m_i!} \\ &= e^{-\theta}\left[\frac{f(y_i|\boldsymbol{\lambda})}{S(y_i|\boldsymbol{\lambda})}\right]^{\delta_i}\sum_{m_i=0}^{\infty}\frac{m_i^{\delta_i}\,[\theta S(y_i|\boldsymbol{\lambda})]^{m_i}}{m_i!} \\ &= e^{-\theta}\left[\frac{f(y_i|\boldsymbol{\lambda})}{S(y_i|\boldsymbol{\lambda})}\right]^{\delta_i}\sum_{m_i=0}^{\infty}m_i^{\delta_i}\,[\theta S(y_i|\boldsymbol{\lambda})]^{m_i}\,\frac{e^{\theta S(y_i|\boldsymbol{\lambda})}e^{-\theta S(y_i|\boldsymbol{\lambda})}}{m_i!} \\ &= e^{-\theta[1-S(y_i|\boldsymbol{\lambda})]}\left[\frac{f(y_i|\boldsymbol{\lambda})}{S(y_i|\boldsymbol{\lambda})}\right]^{\delta_i}\underbrace{\sum_{m_i=0}^{\infty}\frac{m_i^{\delta_i}\,[\theta S(y_i|\boldsymbol{\lambda})]^{m_i}\,e^{-\theta S(y_i|\boldsymbol{\lambda})}}{m_i!}}_{E[V_i^{\delta_i}]} \end{aligned}$$

Sendo V_i uma v.a. com distribuição Poisson de parâmetro $\theta S(y_i|\boldsymbol{\lambda})$ com

$$E[V_i^{\delta_i}] = \begin{cases} E[V_i] & \text{se} \quad \delta_i = 1 \\ 1 & \text{se} \quad \delta_i = 0 \end{cases}$$

então

$$E[V_i^{\delta_i}] = E[V_i]^{\delta_i} = \theta^{\delta_i} S(y_i|\boldsymbol{\lambda})^{\delta_i}.$$

Daí, segue que

$$f(y_i,\delta_i) = e^{-\theta[1-S(y_i|\boldsymbol{\lambda})]} f(y_i|\boldsymbol{\lambda})^{\delta_i}\theta^{\delta_i},$$

portanto, a verossimilhança marginal é dada por

$$L(\theta,\boldsymbol{\lambda};\mathcal{D}) = \prod_{i=1}^{n} exp\{-\theta[1-S(y_i|\boldsymbol{\lambda})]\}[f(y_i|\boldsymbol{\lambda})\theta]^{\delta_i}$$

incluindo covariáveis, temos

$$L(\theta,\boldsymbol{\lambda};\mathcal{D}) = \prod_{i=1}^{n} exp\big[-exp(x_i'\boldsymbol{\beta})[1-S(y_i|\boldsymbol{\lambda})]\big]\big[f(y_i|\boldsymbol{\lambda})exp(x_i'\boldsymbol{\beta})\big]^{\delta_i}.$$

Apêndice B

Aspectos Computacionais

Apresentamos aqui os programas em R para obtenção dos dados simulados e estimativas dos parâmetros para o modelo com fração de cura.

```
# algotitmo EM

mv.EM=function(b0,l0,a0,d, y, x.au, erro = 10^(-3), r, col.mis){
# col.mis é o numero da coluna q apresenta omissao
p = ncol(x.au)
n = nrow(x.au)
w = numeric()
pw = numeric()
pr = numeric()
dw = numeric()
psi = c(b0, l0, a0)
psi1 = psi*1.01
bet = b0
lambda1 = l0[1]
lambda2 = l0[2]
```

```
alpha = a0
m = 0
  while (max(abs(psi1-psi))>erro){
  m = m + 1
  theta = exp(x.au%*%bet)   # parametro de cura pi = e^-theta
  # densidade weibull
  fy = lambda1*y^(lambda1-1)*exp(lambda2-y^lambda1*exp(lambda2))
  Sy = exp(-y^lambda1*exp(lambda2)) # sobrevivencia weibull
  xobs = x.au[,-col.mis]
     for(i in 1:nz){
        if(r[i]==0){w[i]=1}
        else{
        # p(xmis|xobs) probabilidade de sucesso
          pr[i] =  exp(xobs[i,]%*%alpha)/(1+exp(xobs[i,]%*%alpha))
          pw[i] = xobs[i,]%*%bet[-col.mis]
          # denominador de wi
          dw[i] = (exp(pw[i] + bet[col.mis])*fy[i])^d[i]*exp(-exp(pw[i] +
           bet[col.mis])*(1-Sy[i]))*pr[i] + (exp(pw[i])*fy[i])^d[i]*exp(-
            exp(pw[i])*(1-Sy[i]))*(1-pr[i])
          w[i] = ((exp(pw[i] + x.au[i,col.mis]*bet[col.mis])*fy[i])^d[i]*exp(-
          exp(pw[i] + x.au[i,col.mis]*bet[col.mis])*(1-Sy[i]))*pr[i]^x.au[i,
          col.mis]*(1-pr[i])^(1-x.au[i,col.mis]))/dw[i]
          }
        }

  Q1 = function(p2){
    b = matrix(c(p2[1:p]),ncol=1)
    lam1 = p2[p+1]
    lam2 = p2[p+2]
```

```
    q1 = sum(w*(d*(x.au%*%b + lam2 + log(lam1) + (lam1-1)*log(y) -
    y^lam1*exp(lam2)) - exp(x.au%*%b)*(1-exp(-y^lam1*exp(lam2)))))
    return(-q1)
  }
  Q2 = function(a){
    q2 = numeric()
    q2 = w*(x.au[,col.mis]*(xobs%*%a-log(1+exp(xobs%*%a)))+
     (1-x.au[,col.mis])*log(1-(exp(xobs%*%a)/(1+exp(xobs%*%a)))))
    q2s = sum(q2)
    return(-q2s)
  }

bl.max = optim(c(bet, lambda1, lambda2) , Q1, method = "BFGS", hessian = T)
a.max = optim(alpha, Q2, method = "BFGS", hessian = T)
bet = bl.max$par[1:p]
lambda1 = bl.max$par[p+1]
lambda2 = bl.max$par[p+2]
alpha = a.max$par
psi = psi1
psi1 = c(bet, lambda1, lambda2, alpha)
}
return(psi1)
#print(cat("int = " , m, "\n",
#"coeficientes = " , bet, "\n",
#"rho = ", lambda1, "\n",
#"gama = ", lambda2, "\n",
#"alpha = ", alpha, "\n"))
}
```

```
# .....gerando covariáveis.....
for(j in 1:m){
  x1 = rnorm(n)
  x2 = rbinom(n, 1, .6)
  alfa3 = exp(-0.5*x1 + x2)/(1 + exp(-0.5*x1 + x2)) # prob de sucesso
  x3 = rbinom(n, 1, alfa3)
  x = matrix(c(x1, x2, x3), nrow = n)
  p = ncol(x)
# .....fixando parâmetros.....
  b = matrix(c(2,3,2), ncol=1)  # coeficientes da regressão
  th = exp(x%*%b)  # theta - numero medio de celulas doentes
  fc = exp(-th)    #fração de cura
  N = rpois(n,th)         # gerando variaveis latentes
  pimune[j] = mean(N==0) # PERCENTUAL DE IMUNES
# gerando tempo de promoção de cada Ni e censuras
    for(i in 1:n){
        if (N[i]==0) te[i]=13
        else te[i]= min(rweibull(N[i], 2, 4))
      ce[i] = runif(1,0,8)
      if (N[i]==0) y[i] = te[i]
      else y[i] = min(te[i], ce[i])
      d[i] = ifelse(te[i] < ce[i], 1, 0)
      if (N[i]==0) d[i] = 0
    }

# percentual de censura para não imunes
pcensura[j] = sum(y==ce & N>0)/sum(N>0)
```

```
# prob de omissao - x3
po3 = exp(1.65-1.5*y - 1*x1)/(1 + exp(1.65- 1.5*y - 1*x1))

#  indicadores de omissão - x3
r = rbinom(n, 1, po3)

# percentual de omissão
pmis[j] = mean(r)



dados = matrix(c(y, d, x, r), ncol = 6) # dados completos
dobs = dados[r==0,] # dados observados

#ajuste sem omissão
  l = function(phi){ #verossimilhança marginal - weibull
  xcov = dobs[,3:5]
  yc = dobs[,1]
  dc = dobs[,2]
  b1 = phi[1]
  b2 = phi[2]
  b3 = phi[3]
  b = matrix(c(b1,b2, b3), ncol = 1)
  lam1 = phi[4]
  lam2 = phi[5]
  li = sum(dc*(xcov%*%b + lam2 + log(lam1*yc^(lam1-1))-
   yc^lam1*exp(lam2))-exp(xcov%*%b)*(1-exp(-yc^lam1*exp(lam2))))
  return(-li)
  }
  b0c = c(b[1], b[2], b[3])
```

```
  l0c = c(2,-2*log(4))
  phi0c = c(b0c, l0c) # chute inicial
  phi.max = optim(phi0c, l, hessian = T)

  beta1cc[j] = (phi.max$par)[1]
  beta2cc[j] = (phi.max$par)[2]
  beta3cc[j] = (phi.max$par)[3]
  l1cc[j] = (phi.max$par)[4]
  l2cc[j] = (phi.max$par)[5]

# aumentando os dados

pd = ncol(dados)
nz = n + sum(r==1)
Z = matrix(0, nrow=nz, ncol=pd)
col.mis = 5
k = 1
    for(i in 1:n){
        if(r[i]==0){Z[k,]=dados[i,]
                    k = k + 1}
        else{Z[k,]=dados[i,]
              Z[k,col.mis] = 0
              Z[k+1,]=dados[i,]
              Z[k+1,col.mis] = 1
              k=k+2}
            }

  x.au = Z[,3:5]
  y.au = Z[,1]
```

```
  d.au = Z[,2]
  r.au = Z[,6]
  fit1 = glm(Z[,5] ~ Z[,3] + Z[,4] - 1, family = binomial(link = logit))
  a0 = coef(fit1)      #chute para alpha

ajust = mv.EM(c(beta1cc[j], beta2cc[j], beta3cc[j]), c(l1cc[j],l2cc[j]),
a0, d.au, y.au, x.au=x.au, erro = 10^(-3), r.au, col.mis=3)

  beta1[j] = ajust[1]
  beta2[j] = ajust[2]
  beta3[j] = ajust[3]
  rho[j] = ajust[4]
  gama[j] = ajust[5]
  alpha1[j] = ajust[6]
  alpha2[j] = ajust[7]
 }
```

Referências

Agresti, A. (2002). *Categorical data analysis 2nd*. New York: Wiley.

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

Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* **11**(1), 15–44.

Chen, M. & J. Ibrahim (2001). Maximum Likelihood Methods for Cure Rate Models with Missing Covariates. *Biometrics* **57**(1), 43–52.

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

Colosimo, E. & S. Giolo (2006). *Análise de sobrevivência Aplicada*. Edgard Blücher.

Dempster, A., N. Laird & D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**(1), 1–38.

Freire, M. P. d. S. & D. M. Valença (2006). Tópicos de Análise de Sobrevivência e a Aplicação a Dados Referentes ao Tempo para Conclusão de um Curso de Graduação. *Monografia - UFRN.*

Herring, A. & J. Ibrahim (2002). Maximum likelihood estimation in random effects cure rate models with nonignorable missing covariates. *Biostatistics* **3**(3), 387.

Horton, N. & N. Laird (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* **8**(1),

37–50.

Ibrahim, J. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**(411), 765–769.

Ibrahim, J., M. Chen, & S. Lipsitz (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models. *Biometrics* **55**(2), 591–596.

Ibrahim, J., M. Chen, & D. Sinha (2001). *Bayesian Survival Analysis*. Springer.

Kaplan, E. & P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53**(2), 457–481.

Klebanov, L. & A. Yakovlev (2007). A new approach to testing for sufficient follow-up in cure-rate analysis. *Journal of Statistical Planning and Inference* **137**(11), 3557–3569.

Lipsitz, S. & J. Ibrahim (1996). A conditional model for incomplete covariates in parametric regression models.

Little, R. (1992). Regression with missing X's: a review. *Journal of the American Statistical Association* **87**(420), 1227–1237.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**(2), 226–233.

Magalhães, M. (2006). *Probabilidade e Variáveis Aleatórias*. EdUSP.

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

Mizoi, M. (2007). Cure Rate Model with Measurement Error. *Communications in Statistics-Simulation and Computation* **36**(1), 185–196.

Peng, Y., J. Taylor, & B. Yu (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Analysis* **13**(3), 351–369.

Roderick, J., A. Little, & D. Rubin (1987). *Statistical analysis with missing data*. J. Wiley.

Rodrigues, J., V. Cancho, & M. Castro (2008). *Teoria unificada da análise de sobrevivência*. 18º SINAPE.

Rodrigues, J. & F. Louzada-Neto (2006). On the unification of the long term survival models. Manuscrpit submitted to*Statistcs and Probability Letters.*

Rubin, D. (1976). Inference and missing data. *Biometrika* **63**(3), 581–592.

Tanner, M. (1996). *Tools for statistical inference*. Springer-Verlag.

Yakovlev, A., B. Asselain, V. Bardou, A. Fourquet, T. Hoang, A. Rochefediere, & A. Tsodikov (1993). A simple stochastic model of tumor recurrence and its application to data on premenopausal breast cancer. *Biometrie et Analyse de Donnees Spatio-Temporelles* **12**, 67–82.

Yin, G. & J. Ibrahim (2005). A General Class of Bayesian Survival Models with Zero and Nonzero Cure Fractions. *Biometrics* **61**(2), 403–412.

Zaider, M., M. Zelefsky, L. Hanin, A. Tsodikov, A. Yakovlev, & S. Leibel (2001). A survival model for fractionated radiotherapy with an application to prostate cancer. *Phisics in Medicine and Biology* **46**(10), 2745–2758.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo