MODELOS DE SOBREVIVÊNCIA COM FRAÇÃO DE CURA E ERRO DE MEDIDA NAS COVARIÁVEIS

Autor: Rafael Ribeiro de Lima

DISSERTAÇÃO PARA OBTENÇÃO DO TÍTULO DE
MESTRE EM MATEMÁTICA APLICADA E ESTATÍSTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM
MATEMÁTICA APLICADA E ESTATÍSTICA DA
UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

Área de concentração: Métodos Estatísticos
Orientador(a): Prof. Dra. Dione Maria Valença

Natal, agosto de 2008.

Agradecimentos

Agradeço a todos que me ajudaram direta e indiretamente.

À Dione, pela amizade, por ter sempre acreditado em mim e por todo o apoio incondicional nesses anos de convivência.

Ao professor Heleno, pela co-orientação e pela sugestão dos temas.

A todos os professores e funcionários do PPGMAE. Em especial, aos professores: Pledson, pela amizade e pelo entusiasmo que nos transmite sempre; Paulo, por me ensinar tudo sobre simulação; Damião, por sempre nos mostrar a admiração que a Estatística merece; Elias, que nos impressiona pela incrível união entre simplicidade e conhecimento; André, que me incentivou a fazer mestrado quando eu ainda era um graduando de matemática; Marcelo, pelo exemplo de didática em sala de aula.

Aos meus colegas de mestrado, em especial a Nonato, João, Helenice, Hermes, os Márcios, as Renatas, Patrícia, Lenílson, Cecílio e Allan.

Aos meus pais, Natália e Luiz, e irmãos, George, Giselle e Júlia. Qualquer coisa boa que eu escrever sobre eles será pouco visto o que eles fazem e representam para mim.

À Gisa, minha namorada, por todo o amor, atenção e incentivo nos momentos difíceis.

Aos meus amigos da ACA, da música, da Engenharia Civil, da Matemática, da rua e de infância.

À CAPES e ao CNPq pelo suporte financeiro.

Ao meu avô (*in memorian*).

A Deus.

Resumo

Neste trabalho estudamos o modelo de sobrevivência com fração de cura proposto por Yakovlev et al. (1993), fundamentado em uma estrutura de riscos competitivos concorrendo para causar o evento de interese, e a abordagem proposta por Chen et al. (1999), na qual covariáveis são introduzidas para modelar o número de riscos. Estudamos o caso em que covariáveis são medidas com erro, e para a obtenção de estimadores consistentes consideramos a utilização do método do escore corrigido. Um estudo de simulação é realizado para avaliar o comportamento dos estimadores obtidos por este método em amostras finitas. A simulação visa identificar não apenas o impacto sobre os coeficientes de regressão das covariáveis medidas com erro (Mizoi et al. 2007), mas também sobre os coeficientes de covariáveis medidas sem erro. Verificamos também a adequação da distribuição exponencial por partes ao modelo com fração de cura e erro de medida. Ao final, são feitas aplicações do modelo envolvendo conjuntos de dados reais.

Palavras-Chave: Análise de sobrevivência, fração de cura, erro de medida.

Abstract

In this work, we study the survival cure rate model proposed by Yakovlev et al. (1993), based on a competing risks structure concurring to cause the event of interest, and the approach proposed by Chen et al. (1999), where covariates are introduced to model the risk amount. We focus the measurement error covariates topics, considering the use of corrected score method in order to obtain consistent estimators. A simulation study is done to evaluate the behavior of the estimators obtained by this method for finite samples. The simulation aims to identify not only the impact on the regression coefficients of the covariates measured with error (Mizoi et al. 2007) but also on the coefficients of covariates measured without error. We also verify the adequacy of the piecewise exponential distribution to the cure rate model with measurement error. At the end, model applications involving real data are made.

Key-words: Survival analysis, cure rate, measurement error.

Sumário

Capítulo 1

Introdução

1.1 Análise de Sobrevivência

Em modelos de sobrevivência, estamos sempre interessados no tempo até a ocorrência de um evento de interesse, comumente denominado tempo de sobrevivência ou de vida. Seja o tempo até a ocorrência da falha de uma lâmpada, até a morte de um paciente com câncer, até a recidiva de um tumor, até um cidadão quitar uma dívida bancária ou mesmo até um aluno concluir a sua graduação, todos esses fenômenos podem ser objeto de estudo da análise de sobrevivência.

No desenvolvimento do texto que se segue, consideramos que os indivíduos são as unidades de estudo (lâmpadas, pacientes, estudantes, etc.), e dados de sobrevivência são o conjunto de observações em relação ao tempo de vida dos indivíduos.

Seja T uma variável aleatória contínua não-negativa, com função densidade de probabilidade $f(\cdot)$ e função de distribuição acumulada $F(\cdot)$, representando o tempo até a ocorrência do evento de interesse para um indivíduo. Definimos então a função de sobrevivência, $S(\cdot)$, como sendo a probabilidade de T ser superior a um certo tempo t, ou seja, a probabilidade de o indivíduo sobreviver ao tempo t. Desta forma, temos

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

Note que $S(t)$ é uma função monótona decrescente, sendo $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

O risco (ou taxa de falha) de um intervalo $[t, t + \Delta t)$ é definido como sendo a probabilidade de a falha ocorrer neste intervalo, dado que que não ocorreu antes de t, dividida pelo comprimento do intervalo, Δt, ou seja,

$$\frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

Se $\Delta t \to 0$, temos a taxa de falha instantânea no tempo t condicionada à sobrevivência até o tempo t. Então, a função risco associada a T, $h(\cdot)$, é definida como

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

Temos ainda a função risco acumulado, $H(\cdot)$, dada por

$$H(t) = \int_0^t h(u)du. \tag{1.3}$$

Observe que a função risco acumulado é tal que

$$H(\infty) = \lim_{t\to\infty} H(t) = \infty,$$

ou seja, não é limitada superiormente.

As seguintes relações podem ser estabelecidas entre as definições acima:

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

$$h(t) = \frac{f(t)}{S(t)}, \tag{1.5}$$

$$S(t) = exp(-H(t)). \tag{1.6}$$

Portanto, para as funções $f(\cdot)$, $F(\cdot)$, $S(\cdot)$, $h(\cdot)$ e $H(\cdot)$, basta conhecermos uma delas para obtermos as demais. Para maiores detalhes, recomendamos ver o livro do Lawless (1982).

Um fenômeno comum em dados de sobrevivência é a ocorrência de censuras, ou seja, para alguns indivíduos em estudo não sabemos seu tempo

exato de vida, mas apenas que este excede um valor. São observações incompletas ou parciais, e isso ocorre quando o indívíduo não pôde ser acompanhado até a ocorrência do evento de interesse, seja por ele ter abandonado o experimento, por perda de informação ou pela limitação do tempo de acompanhamento do experimento, dentre outros motivos. Entretanto, tais observações mesmo censuradas fornecem informações sobre o tempo de vida do indivíduo e devem ser utilizadas na análise estatística, e a sua omissão no cálculo das estatísticas de interesse pode ocasionar conclusões viciadas.

1.2 Modelos em Análise de Sobrevivência

Vários modelos paramétricos podem ser utilizados na análise de dados de sobrevivência. Dentre os modelos univariados, alguns ocupam uma posição de destaque por sua comprovada adequação a várias situações práticas, dentre os quais podemos citar: exponencial, Weibull, gama e log-normal.

Quando as circunstâncias do problema em estudo não suportam a utilização de um modelo paramétrico, a solução pode estar na utilização de distribuições ou métodos semi-paramétricos, dentre os quais destacamos o modelo exponencial por partes (MEP).

Caracterizamos a seguir três modelos importantes no desenvolvimento deste trabalho.

1.2.1 Modelo Exponencial

Se T tem distribuição exponencial, sua função risco é constante e dada por

$$h(t) = \lambda, \quad t \geq 0,$$

com $\lambda > 0$. A partir desta, a função densidade de probabilidade e a função de sobrevivência são obtidas através das relações (1.3) à (1.6), e dadas respectivamente por

$$f(t) = \lambda \exp(-\lambda t), \ t \geq 0, \tag{1.7}$$

e

$$S(t) = \exp(-\lambda t),\ t \geq 0. \tag{1.8}$$

1.2.2 Modelo Weibull

Se T tem distribuição Weibull, sua função risco é da forma

$$h(t) = \lambda\beta(\lambda t)^{\beta-1},$$

com $\lambda > 0$ e $\beta > 0$, parâmetros do modelo. Note que a distribuição exponencial é um caso particular da distribuição Weibull, quando $\beta = 1$. A função densidade de probabilidade e a função de sobrevivência são obtidas através das relações (1.3) à (1.6), e dadas respectivamente por

$$f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}],\ t > 0, \tag{1.9}$$

e

$$S(t) = \exp[-(\lambda t)^{\beta}],\ t > 0. \tag{1.10}$$

É comum encontrarmos a distribuição Weibull parametrizada de outras formas. Uma delas, utilizada em Mizoi et al. (2007), considera um vetor de parâmetros $\lambda = (\rho, \gamma)$ tal que as funções densidade de probabilidade e de sobrevivência são dadas por

$$f(t) = e^{\gamma}\rho t^{\rho-1}\exp(-t^{\rho}e^{\gamma}),\ t > 0, \tag{1.11}$$

e

$$S(t) = \exp(-t^{\rho}e^{\gamma}),\ t > 0. \tag{1.12}$$

Relacionando as duas parametrizações, temos $e^{\gamma} = \lambda^{\beta}$ e $\rho = \beta$.

A função risco é monótona crescente se $\beta > 1$, decrescente se $\beta < 1$ e constante se $\beta = 1$ (distribuição exponencial).

O modelo Weibull tem sido bastante utilizado na prática por fornecer uma boa descrição de diversos tipos de dados de sobrevivência.

1.2.3 Modelo Exponencial por Partes

O modelo exponencial por partes ou modelo exponencial particionado (MEP) tem sido discutido extensivamente na literatura em diferentes contextos. Dentre os trabalhos que serviram de referência no seu estudo e entendimento, podemos citar Friedman (1982), Kim & Proschan (1991), Lindsey & Ryan (1993), Chen & Ibrahim (2001), Demarqui (2006), dentre outros.

Uma grande vantagem do MEP é o fato dele ser capaz de acomodar funções risco com diversas formas, sejam elas monótonas ou não, o que o torna bastante flexível. Outra vantagem reside no fato de podermos controlar o seu grau de parametricidade, podendo-se até trabalhar na sua versão não-paramétrica.

Considere uma partição arbitrária e finita de $\Re^+$, $\{s_1, \ldots, s_k\}$, tal que $0 = s_0 < s_1 < s_2 < \ldots < s_k < \infty$. Desta forma, o eixo do tempo $\Re^+$ fica dividido em k intervalos disjuntos, denotados por $I_1 = (s_0, s_1], I_2 = (s_1, s_2], \ldots, I_k = (s_{k-1}, s_k]$. O MEP se caracteriza pela aproximação da função risco por valores constantes em cada um dos intervalos definidos pela partição $\{s_1, \ldots, s_k\}$, isto é, assume-se que em cada intervalo $I_j = (s_{j-1}, s_j]$, $j = 1, \ldots, k$, a função risco seja constante e denotada por $h(t) = \lambda_j$, $\lambda_j > 0$, $\forall t \in I_j$. Desta forma, a função risco acumulado $H(t)$, associada ao j-ésimo intervalo, é dada pela soma das áreas de retângulos cujas bases são determinadas pelos intervalos definidos pela partição $\{s_1, \ldots, s_k\}$, e com alturas dadas pela função risco $h(t)$, ou seja,

$$H(t) = \begin{cases} \lambda_1 t, \text{ se } t \in I_1; \\ \sum_{r=1}^{j-1} \lambda_r (s_r - s_{r-1}) + \lambda_j (t - s_{j-1}), \text{ se } t \in I_j,\ j > 1. \end{cases} \tag{1.13}$$

Assim, utilizando as relações (1.6) e (1.5) respectivamente, a função de sobrevivência e a função densidade de probabilidade da distribuição exponencial por partes são dadas por

$$S(t|\lambda_1,\ldots,\lambda_k) = \begin{cases} \exp\left(-\lambda_1 t\right), \text{se } t \in I_1; \\ \exp\left\{-\left[\sum_{r=1}^{j-1}\lambda_r(s_r - s_{r-1}) + \lambda_j(t - s_{j-1})\right]\right\}, \text{se } t \in I_j,\ j > 1, \end{cases} \tag{1.14}$$

e

$$f(t|\lambda_1,\ldots,\lambda_k) = \begin{cases} \lambda_1 \exp\left(-\lambda_1 t\right), \text{se } t \in I_1; \\ \lambda_j \exp\left\{-\left[\sum_{r=1}^{j-1}\lambda_r(s_r - s_{r-1}) + \lambda_j(t - s_{j-1})\right]\right\}, \text{se } t \in I_j,\ j > 1, \end{cases} \tag{1.15}$$

com $\lambda_j > 0$, $\forall j = 1,\ldots,k$.

1.3 Fração de Cura

Nos modelos de sobrevivência usuais, considera-se que todos os indíviduos envolvidos no experimento atingirão o evento de interesse se forem acompanhados por um tempo suficientemente grande, mas considerar tal abordagem pode não ser adequado em algumas situações.

Existem dados de sobrevivência nos quais uma porcentagem dos indivíduos não apresentará a ocorrência deste evento, mesmo se acompanhados por um longo período de tempo. Certamente uma lâmpada cedo ou tarde falhará, porém um paciente curado de um câncer pode nunca vir a sofrer o retorno (ou recidiva) do tumor, um cidadão pode nunca vir a quitar uma dívida e um aluno pode abandonar e nunca vir a se formar em um curso. Diz-se então que esses indívíduos são imunes ao evento de interesse, e o conjunto de dados de sobrevivência aos quais eles pertencem possui uma fração de cura ou de curados.

Um fato que pode vir a indicar a presença de imunes num conjunto de dados de sobrevivência é a ocorrência de uma alta proporção de censura a direita, ou seja, ao final de um experimento, quando o seu tempo limite foi atingido, existe um grande número de indivíduos que não apresentaram o

evento de interesse. A Figura 1.1 apresenta a curva de sobrevivência estimada (estimador Kaplan-Meier) para um conjunto de dados reais referente ao tempo que alunos levam para a conclusão do curso de graduação em estatística da UFRN. Observamos que a cauda direita se acomoda num valor bem acima do zero por um período considerável, retratando desta forma o comportamento de uma função de sobrevivência imprópria, ou seja, que não tende a zero quando o tempo cresce. É nesse contexto que surgem, como proposta para tratar de dados com estas características, os modelos de sobrevivência com fração de cura.

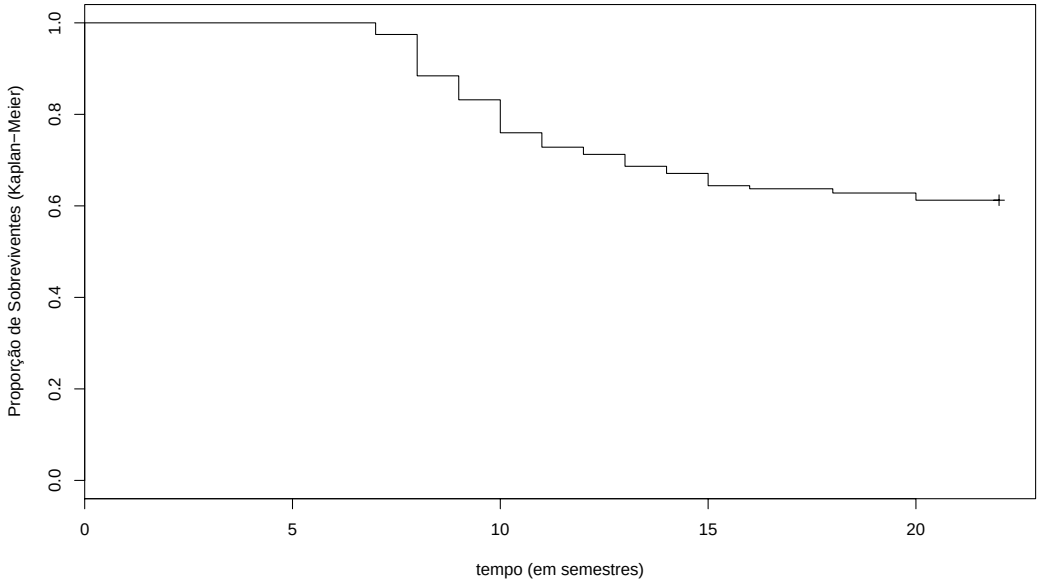

Figura 1.1: Estimativas Kaplan-Meier para os dados de tempo até a conclusão do curso de graduação em Estatística da UFRN no período de 1997 à 2004. Amostra com $n = 354$ alunos.

1.4 Análise de Sobrevivência com Fração de Cura

Uma alternativa para lidar com a presença de fração de cura em um conjunto de dados é considerarmos uma mistura de distribuições paramétricas, sendo uma função de sobrevivência imprópria considerada para a população total (curados e não curados) e uma função de sobrevivência própria para a parte da população formada pelos não curados. Boag (1949) e Berkson & Gage (1952) propõem um modelo de mistura de distribuições paramétricas, o qual considera que a população possui uma porcentagem π de indivíduos imunes ou curados. Utilizando uma partição em curados (C) e não curados (NC), a função de sobrevivência populacional é dada por

$$\begin{aligned} S_p(t) &= P(T \geq t) \\ &= P(T \geq t|C)P(C) + P(T \geq t|NC)P(NC) \\ &= 1 \times \pi + S_{NC}(t)(1 - P(C)) \\ &= \pi + S_{NC}(t)(1 - \pi). \end{aligned}$$

Assim,

$$S_p(t) = \pi + (1 - \pi)S_{NC}(t), \quad t \geq 0, \tag{1.16}$$

sendo $S_{NC}(\cdot)$ a função de sobrevivência própria associada aos indivíduos não curados.

Desde então, diversos autores vêm discutindo a respeito de modelos envolvendo mistura de distribuições e fração de cura. Fareweel (1982) reanalisa um experimento toxicológico, feito originalmente por Pierce et al. (1979) que utilizou um modelo de riscos proporcionais de Cox, usando uma mistura de distribuições paramétricas e considerando a presença de indivíduos sobreviventes. Goldman(1984) discute sobre a análise de sobrevivência quando a

cura é possível. Greenhouse & Wolfe (1984) estudam uma generalização do modelo de mistura baseada na teoria de riscos competitivos. Farewell (1986) examina o uso de tais modelos na inferência estatística. Halpern & Brown (1987) comparam o poder dos testes log-rank e de Wilcoxon generalizado em situações onde modelos de fração de cura são apropriados. Gray & Tsiatis (1989) estudam testes para comparar diferenças entre frações de cura, enquanto Sposto et al.(1992) comparam testes de diferença entre proporção de curados. Laska & Meisner (1992) trabalham com estimação não-paramétrica em modelos com fração de cura. Kuk & Chen (1992) propõem um modelo de mistura combinando regressão logística e de riscos proporcionais. Um excelente livro que aborda modelos de sobrevivência com fração de cura é o de Maller & Zhou (1996).

Alternativamente, Yakovlev et al. (1993) propõem uma nova classe de mistura envolvendo uma estrutura de riscos competitivos, onde eles impõem um limite superior à função de risco acumulado populacional de tal forma que

$$H_p(\infty) = \lim_{t\to\infty} H_p(t) = \theta, \quad \theta > 0,$$

e definem $H_p(t) = \theta F(t)$, sendo $F(\cdot)$ uma função de distribuição de uma variável não-negativa. Desta forma, utilizando a relação (1.6), temos que a função de sobrevivência populacional (curados e não curados) é dada por

$$S_p(t) = exp(-\theta F(t)), \quad t \geq 0,$$

e a fração de cura por

$$\pi = \lim_{t\to\infty} exp(-\theta F(t)) = exp(-\theta),$$

o que resulta numa função de sobrevivência própria para os indivíduos não curados, de acordo com (1.16), dada por

$$S_{NC}(t) = \frac{S_p(t) - \pi}{1 - \pi} = \frac{exp(-\theta F(t)) - exp(-\theta)}{1 - exp(-\theta)}.$$

Tsodikov (1998) observa que quando o parâmetro θ é definido como uma função de covariáveis observadas, o modelo proposto por Yakovlev et al. (1993) apresenta a estrutura de riscos proporcionais.

Chen et al. (1999) propõem uma versão bayesiana do modelo introduzido por Yakovlev et al. (1993), e desenvolvem vários resultados em trabalhos posteriores. Chen et al. (2001) propõem estimadores de máxima verossimilhança para dados com omissão, ou seja, dados incompletos. Chen et al. (2002 a, b e c) desenvolvem métodos Bayesianos para omissão em covariáveis, análises multivariadas e aplicações em dados de câncer.

1.5 Erro de Medida

É comum a ocorrência de erro de medida em procedimentos de mensuração. Seja pela imprecisão do instrumento de medida, pela falta de habilidade do técnico responsável, por intempéries presentes no ambiente de medição ou mesmo por algum entrevistado faltar com a verdade numa pesquisa de opinião, os erros podem ocorrer comumente na prática, e tais erros poderão afetar a precisão de estimadores de uma função que dependa de tais medidas. Em estudos de comparação de grupos, parte da diferença (ou mesmo toda) entre eles pode ser devido a tais erros.

Modelos com erro de medida em geral apresentam-se da seguinte forma: uma variável resposta Y deve ser obtida em termos de variáveis independentes, dentre as quais distinguimos dois tipos: X representando as variáveis que são medidas sem erro, e Z as que não podem ser observadas de maneira exata, para os indivíduos em estudo. Assim, os parâmetros do modelo relacionando Y e (X, Z) não podem ser estimados diretamente ajustando Y à (X, Z), pois Z não é observável. Em vez disso, nós podemos observar uma variável W que é relacionada à Z. O principal objetivo dos modelos com erro de medida é obter estimativas não-viciadas dos parâmetros indiretamente, ajustando o modelo para Y em termos de (X, W). Porém, substituindo Z por

W, sem que sejam feitos certos ajustes (ou correções), obtemos estimativas assintoticamente viciadas (Stefanski 1985). Em geral, ocorre que o estimador baseado na observação W, conhecido como estimador naive, é inconsistente e a estimativa é atenuada (em direção à zero) pelo erro de medida.

Um aspecto importante na utiliação de modelos com erro de medida está na suposição sobre a variável não observada (Z). Esta pode ser considerada como uma constante (modelo funcional) ou como uma variável aleatória (modelo estrutural). Dentre os modelos com erro de medida, podemos citar o método do escore corrigido, simulação e extrapolação (SIMEX) e o de regressão-calibração. Para um estudo detalhado a respeito de modelos com erro de medida, aconselhamos a leitura do livro do Carroll et al. (1995).

Mizoi et al. (2007) introduzem erro de medida nas covariáveis associadas ao modelo com fração de cura proposto por Chen et al. (1999) e estudam, através de simulação, as propriedades dos estimadores obtidos através do método do escore corrigido (Nakamura 1990, Gimenez & Bolfarine 1997).

1.6 Objetivos

Este trabalho tem como objetivo principal o aprofundamento no estudo do modelo introduzido por Yakovlev et al. (1993), utilizando a abordagem de Chen et al. (1999) e seguindo a linha de pesquisa de Mizoi et al. (2007). Especificamente desejamos:

- Realizar um estudo de simulação abrangendo aspectos complementares ao trabalho de Mizoi et al. (2007), analisando as influências da presença de erro de medida, em amostras finitas, na estimação dos parâmetros não só das covariáveis medidas com erro mas também das medidas sem erro, por meio do método do escore corrigido;

- Analisar a adequação do modelo semi-paramétrico exponencial por partes (MEP) ao modelo de sobrevivência com fração de cura e erro de

medida, utilizando o método do escore corrigido;

- Realizar aplicações do modelo de sobrevivência com fração de cura utilizando dois conjuntos de dados reais, para consolidar a teoria estudada no âmbito prático.

1.7 Conteúdo dos Capítulos

Os capítulos que seguem encontram-se organizados da seguinte forma: no Capítulo 2 é apresentada a formulação do modelo, tomando como base a abordagem de Chen et al. (1999). No Capítulo 3, apresentamos a extensão para o caso em que uma covariável é medida com erro, assim como o processo de estimação dos parâmetros utilizando o método do escore corrigido. No Capítulo 4 apresentamos os estudos de simulação. No Capítulo 5 realizamos duas aplicações do modelo com fração de cura a dados reais. A primeira a dados considerados sem erro de medida, provenientes de uma pesquisa sobre evasão escolar de um curso de graduação, e a segunda a dados com erro de medida, provenientes de um estudo sobre câncer de mama. A conclusão do trabalho é apresentada no Capítulo 6, com uma discussão sobre os resultados obtidos e a proposição de possíveis pesquisas futuras.

Capítulo 2

Modelo de Sobrevivência com Fração de Cura

2.1 Descrição do Modelo

Consideramos o modelo de sobrevivência com fração de cura introduzido por Yakovlev et al. (1993). Para tanto, imaginemos um cenário clínico, onde pacientes que foram tratados de um câncer estão sendo observados quanto ao tempo de remissão do tumor, ou seja, o tempo até a volta ou recidiva do câncer. Suponha que, para um indivíduo da amostra envolvida, N denota o número de células potencialmente capazes de se tornarem cancerígenas, e consequentemente de ocasionar o reaparecimento do tumor. Seja R_i, $0 \leq i \leq N$, o tempo para que a i-ésima célula se torne, de forma detectável, cancerígena. Dentre as N células capazes, a primeira que se tornar cancerígena ocasionará o reaparecimento do câncer. Temos então uma estrutura denominada de riscos competitivos. Considere adicionalmente as seguintes suposições:

1. $N \sim Poisson(\theta)$ de média θ, ou seja, $P(N = n) = \dfrac{e^{\theta}\theta^{n}}{n!}$, para $\theta > 0$ e $n = 0, 1, 2, \ldots$. O uso da distribuição de Poisson se deve ao fato de N representar a contagem do número de células potencialmente cancerígenas;

2. $R_1, R_2, \ldots, R_N$: variáveis aleatórias independentes e identicamente dis-

tribuídas (i.i.d.), independentes de N, com função de distribuição $F(\cdot)$ e função de sobrevivência $S(\cdot) = 1 - F(\cdot)$. Estas variáveis representam os tempos das N células potencialmente cancerígenas desenvolverem metástase detectável;

3. T: tempo até a ocorrência do evento de interesse, definido como

$$T = min\{R_0, R_1, \ldots, R_N\},$$

com $P(R_0 = \infty) = 1$, R_0 representando o tempo associado à uma célula que nunca apresentará o evento de interesse. Desta forma, se o indivíduo tiver $N = 0$, ele possuirá apenas o tempo R_0 e nunca apresentará o evento de interesse.

A proposição seguinte apresenta a forma da função de sobrevivência para indivíduos curados e não-curados.

Proposição 2.1 *Com base nas suposições* (1),(2) *e* (3) *dadas acima, temos*

$$S_p(t) = P(T \geq t) = exp(-\theta F(t)), \quad t \geq 0.$$

Prova:

$$\begin{aligned}
S_p(t) &= P(T \geq t) = P(min\{R_0, R_1, \ldots, R_N\} \geq t) \\
&= \sum_{k=0}^{\infty} P(min\{R_0, R_1, \ldots, R_k\} \geq t | N = k) P(N = k) \\
&= P(R_0 \geq t) P(N = 0) + \sum_{k=1}^{\infty} P(min\{R_0, R_1, \ldots, R_k\} \geq t) P(N = k) \\
&= \frac{exp(-\theta)\theta^0}{0!} + \sum_{k=1}^{\infty} P(R_0 \geq t, R_1 \geq t, \ldots, R_k \geq t) P(N = k) \\
&= exp(-\theta) + \sum_{k=1}^{\infty} P(R_0 \geq t) P(R_1 \geq t) \ldots P(R_k \geq t) \frac{exp(-\theta)\theta^k}{k!} \\
&= exp(-\theta) + exp(-\theta) \sum_{k=1}^{\infty} \frac{[S(t)\theta]^k}{k!} \\
&= exp(-\theta) + exp(-\theta)[exp(S(t)\theta) - 1] \\
&= exp(-\theta F(t))
\end{aligned}$$

□

Note que utilizamos a identidade matemática $\sum_{m=0}^{\infty} \frac{a^m}{m!} = e^a, \quad a > 0$, na demonstração da proposição 2.1.

Temos então que a fração de cura π induzida pelo modelo é dada por

$$\pi = \lim_{t \to \infty} S_p(t) = exp(-\theta).$$

Este resultado é análogo à probabilidade de o indivíduo não possuir nenhuma célula potencialmente cancerígena, ou seja,

$$P(N = 0) = exp(-\theta).$$

Observamos que a medida que θ cresce, a média de N aumenta e a fração de cura $exp(-\theta)$ tende a zero, diminuindo a probabilidade de cura. O resultado inverso vale, quando θ diminui o número médio de células potencialmente cancerígenas (N) diminui e a fração de cura aumenta tendendo à 1.

2.2 Incluindo Covariáveis ao Modelo

A fim de considerar populações heterogêneas em relação à presença de curados e não-curados, é de interesse incluir covariáveis ao modelo. Chen et al. (1999) desenvolvem uma versão deste modelo no qual covariáveis são introduzidas no parâmetro θ através da relação

$$\theta \equiv \theta(\mathbf{x}'\beta) = exp(\mathbf{x}'\beta), \tag{2.1}$$

sendo $\mathbf{x}= (x_1, \ldots, x_p)'$ um vetor de covariáveis e $\beta = (\beta_1, \ldots, \beta_p)'$ um vetor p-dimensional de coeficientes de regressão associados a $\mathbf{x}$ (x' denota o vetor x transposto). Caso se queira trabalhar com um intercepto, basta adicionar β_0 ao vetor de coeficientes e o valor 1 ao vetor de covariáveis, ficando $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ e $\mathbf{x}= (1, x_1, \ldots, x_p)'$. A fração de cura fica relacionada ao vetor de covariáveis através da expressão

$$\pi = exp(-exp(\mathbf{x}'\beta)). \tag{2.2}$$

2.3 Função de Verossimilhança

Suponha uma amostra com n indivíduos e considere que associado a cada indivíduo i, $i = 1, \ldots, n$, temos as seguintes variáveis:

- N_i: variáveis i.i.d., não observáveis, com $N_i \sim Poisson(\theta_i)$;
- $R_{i1}, R_{i2}, \ldots, R_{iN_i}$: variáveis i.i.d., não observáveis, com função de distribuição $F(\cdot|\lambda)$ e função de sobrevivência $S(\cdot|\lambda) = 1 - F(\cdot|\lambda)$, sendo λ o vetor de parâmetros;
- y_i: tempo observado, dado por $y_i = min\{T_i, C_i\}$, com $T_i = min\{R_{i0}, R_{i1}, \ldots, R_{iN_i}\}$ e C_i o tempo de censura do indivíduo i;
- ν_i: indicador de falha, com $\nu_i = \begin{cases} 1, & \text{se } y_i \text{ é tempo de falha;} \\ 0, & \text{se } y_i \text{ é tempo de censura.} \end{cases}$
- $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$: vetor de covariáveis associadas a cada indivíduo, introduzidas no modelo através do parâmetro θ_i segundo a relação $\theta_i = exp(x_i'\beta)$, sendo $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ o vetor dos coeficientes de regressão.

Obtemos então os vetores n-dimensionais:

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)', \quad \nu = (\nu_1, \nu_2, \ldots, \nu_n)' \quad \text{e} \quad \mathbf{N} = (N_1, N_2, \ldots, N_n)',$$

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

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

Denotamos o conjunto dos dados completos (com as variáveis não observáveis ou latentes) por D_c=(n,$\mathbf{y}$,ν,$\mathbf{N}$,$\mathbf{X}$). Sendo $\phi = (\beta', \lambda')'$ o vetor de

parâmetros, a função de verossimilhança dos dados completos é dada por (vide Apêndice A)

$$L(\phi; D_c) = \Big\{ \prod_{i=1}^{n} S(y_i|\lambda)^{N_i-\nu_i} [N_i f(y_i|\lambda)]^{\nu_i} \Big\} \times \\ exp\Big\{ \sum_{i=1}^{n} \Big[N_i x_i^{'} \beta - \ln(N_i!) - exp(x_i^{'}\beta) \Big] \Big\}. \tag{2.3}$$

Note que (2.3) não pode ser obtida, pois depende de N que é uma variável não observável. Aplicando-lhe o logaritmo, obtemos o logaritmo da função de verossimilhança dos dados completos,

$$\begin{aligned} l(\phi; D_c) &= \sum_{i=1}^{n} \Big\{ (N_i - \nu_i) \ln S(y_i|\lambda) + \nu_i \ln N_i + \nu_i \ln f(y_i|\lambda) \Big\} \\ &+ \sum_{i=1}^{n} \Big[N_i x_i^{'} \beta - \ln(N_i!) - exp(x_i^{'}\beta) \Big]. \end{aligned} \tag{2.4}$$

Seja $\mathbf{D} = (n, \mathbf{y}, \nu, \mathbf{X})$ o conjunto dos dados observados. Uma vez que (2.4) inclui as variáveis latentes N, trabalhamos com o logaritmo da função de verossimilhança marginal (obtida fazendo-se o somatório nas variáveis não observadas N), dada pela equação (vide Apêndice B)

$$l(\phi; D) = \sum_{i=1}^{n} \Big\{ \nu_i x_i^{'} \beta + \nu_i \ln f(y_i|\lambda) - exp(x_i^{'}\beta)[1 - S(y_i|\lambda)] \Big\}. \tag{2.5}$$

Através dela, obtemos as estimativas de máxima verossimilhança dos parâmetros envolvidos.

No próximo Capítulo, iremos desenvolver a teoria acerca do método do *escore corrigido*, que será usado aqui para lidar com erro de medida em covariáveis, com base em uma correção na equação (2.5).

Capítulo 3

Modelo com Fração de Cura e Erro de Medida

3.1 Suposições sobre o Erro de Medida

Suponhamos que no conjunto de covariáveis do modelo com fração de cura, uma delas não seja medida com precisão. Desta forma, para cada indivíduo i, $i = 1, 2, \ldots, n$, teremos $p + 1$ covariáveis, sendo

- z_i: covariável medida com erro;
- $x_i^{'} = (x_{i1}, x_{i2}, \ldots, x_{ip})$: vetor de p covariáveis medidas sem erro.

Consideramos um modelo de erro nas covariáveis com estrutura aditiva. Assim, em vez de observarmos a covariável z_i, observamos w_i, $i = 1, 2, \ldots, n$, e elas se relacionam da seguinte forma:

$$w_i = z_i + u_i, \tag{3.1}$$

sendo u_i, para $i = 1, 2, \ldots, n$, variáveis aleatórias i.i.d. representando o erro de medida, com $u_i \sim N(0, \sigma_u^2)$, independentes de z_i, y_i e ν_i. Consideramos um modelo funcional, ou seja, assumimos que z_i são constantes desconhecidas.

3.2 Função de Verossimilhança Marginal

Considerando os verdadeiros valores da covariável medida com erro, $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$, o logaritmo da função de verossimilhança marginal do modelo com fração de cura, dado por (2.5), é agora representado por

$$l(\phi; D) = \sum_{i=1}^{n} \Big\{ \nu_i \Big(x_i' \beta_x + z_i \beta_z \Big) + \nu_i \ln f(y_i|\lambda) - exp\Big(x_i' \beta_x + z_i \beta_z \Big) [1 - S(y_i|\lambda)] \Big\}, \tag{3.2}$$

sendo $\phi = (\beta', \lambda')'$ com $\beta = (\beta_x', \beta_z)'$ e $\mathbf{D} = (n, \mathbf{y}, \nu, \mathbf{z}, \mathbf{X})$. Se substituírmos as verdadeiras covariáveis $\mathbf{z}$ pelas covariáveis observadas $\mathbf{w}= (w_1, w_2, \ldots, w_n)'$, o logaritmo da função de verossimilhança marginal recebe a denominação de *naive*, sendo a expressão dada por

$$l(\phi; \overline{D}) = \sum_{i=1}^{n} \Big\{ \nu_i \Big(x_i' \beta_x + w_i \beta_z \Big) + \nu_i \ln f(y_i|\lambda) - exp\Big(x_i' \beta_x + w_i \beta_z \Big) [1 - S(y_i|\lambda)] \Big\}, \tag{3.3}$$

com $\overline{D} = (n, \mathbf{y}, \nu, \mathbf{w}, X)$.

As estimativas obtidas a partir de (3.3), chamadas estimativas *naive*, podem ser assintoticamente viciadas (Stefanski 1985). Em geral, o estimador *naive* é atenuado pela presença de erro de medição, fato que verificamos no próximo Capítulo através de simulações computacionais.

3.3 Método do Escore Corrigido

O método do escore corrigido (Nakamura 1990, Gimenez & Bolfarine 1997) é um método para estimação em modelos com erro de medida nas variáveis, que possibilita a obtenção de estimadores consistentes e assintoticamente normais. O método pressupõe o conhecimento da variância do erro de medida, ou uma estimativa dela, e pode ser utilizado tanto em modelos funcionais como estruturais.

Este método propõe calcular inicialmente uma correção no logaritmo da função de verossimilhança observada ou naive $l(\phi; \overline{D})$, que é função das covariáveis observadas $\mathbf{w}$, obtendo assim o que denominamos de logaritmo da função de verossimilhança corrigida l^*, que deve satisfazer a equação

$$E[l^*(\phi; \overline{D})|D] = l(\phi; D). \tag{3.4}$$

Desta forma, a esperança do logaritmo da função de verossimilhança corrigida $l^*(\phi; \overline{D})$, condicionada à D, é igual ao logaritmo da função de verossimilhança dos dados reais, sem erro de medida, $l(\phi; D)$, dada pela expressão (3.2).

Quando l^* é diferenciável, definimos a função escore corrigido U^* como sendo a sua derivada em relação aos parâmetros do modelo, ou seja,

$$U^*(\phi; \overline{D}) = \frac{\partial l^*(\phi; \overline{D})}{\partial \phi}. \tag{3.5}$$

As estimativas do método do escore corrigido são definidas como a solução da equação $U^*(\phi; \overline{D}) = 0$, sendo $-\frac{\partial U^*(\phi; \overline{D})}{\partial \phi}$ uma matriz positiva definida.

Quando a função escore corrigido satisfaz a condição

$$E[U^*(\phi; \overline{D})|D] = U(\phi; D), \tag{3.6}$$

com $U(\phi; D) = \frac{\partial l(\phi; D)}{\partial \phi}$, e sob certas condições de regularidade, propriedades assintóticas dos estimadores podem ser obtidas usando resultados apresentados em Gimenez & Bolfarine (1997). Suponhamos então válida a relação (3.6) e denotemos

$$U^*(\phi; \overline{D}) = \sum_{i=1}^{n} U_i^*(\phi, y_i, \nu_i, w_i, X_i) = \sum_{i=1}^{n} \frac{\partial l_i^*(\phi, y_i, \nu_i, w_i, X_i)}{\partial \phi}$$

e

$$H^*(\phi; \overline{D}) = \sum_{i=1}^{n} H_i^*(\phi, y_i, \nu_i, w_i, X_i) = \sum_{i=1}^{n} \frac{\partial U_i^*(\phi, y_i, \nu_i, w_i, X_i)}{\partial \phi}$$

(para simplificar a notação, no que segue denotamos as funções acima por $U^*(\phi)$, $U_i^*(\phi)$, $H^*(\phi)$ e $H_i^*(\phi)$). Sejam então

$$\overline{H}_n^*(\phi) = \frac{1}{n}\sum_{i=1}^{n} H_i^*(\phi),$$

$$\overline{\Lambda}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} E\Big[- H_i^*(\phi)\Big],$$

$$\overline{\Gamma}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} E\Big[U_i^*(\phi)U_i^*(\phi)^{'}\Big].$$

Denotando por ϕ_0 o verdadeiro valor do parâmetro segue que, sob certas condições de regularidade, $\widehat{\phi}_{ec}$ segue uma distribuição assintótica normal com média ϕ_0 e matriz de covariâncias $n^{-1}\Omega_n$, com

$$\Omega_n = \Big\{\overline{\Lambda}_n(\phi_0)\Big\}^{-1}\overline{\Gamma}_n(\phi_0)\Big\{\overline{\Lambda}_n(\phi_0)^{'}\Big\}^{-1}. \tag{3.7}$$

Uma estimativa consistente para (3.7) é dada por

$$\widehat{\Omega}_n = \Big\{ - \overline{H}_n^*(\widehat{\phi}_{ec})\Big\}^{-1}\overline{G}_n(\widehat{\phi}_{ec})\Big\{ - \overline{H}_n^*(\widehat{\phi}_{ec})^{'}\Big\}^{-1}, \tag{3.8}$$

com

$$\overline{G}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} U_i^*(\phi)U_i^*(\phi)^{'}.$$

Testes estatísticos baseados nas propriedades assintóticas dos estimadores escore corrigido podem ser construídos para testar hipóteses de interesse usando resultados apresentados em Gimenez et al. (2000). No Capítulo 5 realizamos uma aplicação envolvendo um conjunto de dados supostamente com erro de medida e utilizamos estes resultados para calcular o erro padrão e o p-valor associados às estimativas escore corrigido.

3.4 Correção para o Modelo com Fração de Cura

Seja $F(\cdot|\lambda)$ a distribuição de probabilidade das variáveis latentes (R_i's), com $\lambda = (\lambda_1, \ldots, \lambda_k)$ um vetor de parâmetros da distribuição considerada. Obtemos a correção a partir do logaritmo da função de verossimilhança marginal naive (3.3), calculando a sua esperança levando em conta a relação (3.1). Desta forma, temos

$$\begin{aligned} E(l(\phi;\overline{D})|D) &= \sum_{i=1}^{n}\Big\{E\Big[\nu_i\Big(x_i'\beta_x + w_i\beta_z\Big)\Big|D\Big] + \nu_i \ln f(y_i|\lambda) \\ &\quad -E\Big[exp\Big(x_i'\beta_x + w_i\beta_z\Big)\Big|D\Big][1 - S(y_i|\lambda)]\Big\} \\ &= \sum_{i=1}^{n}\Big\{\nu_i\Big(x_i'\beta_x + E(w_i\beta_z|D)\Big) + \nu_i \ln f(y_i|\lambda) \\ &\quad -E[exp(w_i\beta_z)|D]exp(x_i'\beta_x)[1 - S(y_i|\lambda)]\Big\}. \end{aligned}$$

Mas $E(w_i\beta_z|D) = z_i\beta_z$ e $E[exp(w_i\beta_z)|D] = exp\Big(z_i\beta_z + \dfrac{\beta_z^2\sigma_u^2}{2}\Big)$, $i = 1, \ldots, n$. Note que para calcularmos $E[exp(w_i\beta_z)|D]$, com $w_i \sim N(z_i, \sigma_u^2)$, basta utilizarmos as propriedades da função geradora de momentos da distribuição normal, $M_{w_i}(t)$, com $t = \beta_z$. Para maiores detalhes, recomendamos o livro do Magalhães (2006). Então,

$$\begin{aligned} E(l(\phi;\overline{D})|D) &= \sum_{i=1}^{n}\Big\{\nu_i\Big(x_i'\beta_x + z_i\beta_z\Big) + \nu_i \ln f(y_i|\lambda) \\ &\quad -exp(x_i'\beta_x + z_i\beta_z + \frac{\beta_z^2\sigma_u^2}{2})[1 - S(y_i|\lambda)]\Big\}, \end{aligned}$$

que difere da expressão (3.2) apenas pelo termo $\eta = \dfrac{\beta_z^2\sigma_u^2}{2}$. Portanto, o logaritmo da função de verossimilhança marginal corrigida é dado por

$$
\begin{aligned}
l^*(\phi;\overline{D}) &= \sum_{i=1}^{n}\Big\{\nu_i\Big(x_i'\beta_x + w_i\beta_z\Big) + \nu_i \ln f(y_i|\lambda) \\
&\quad -exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)[1 - S(y_i|\lambda)]\Big\},
\end{aligned} \tag{3.9}
$$

Com a expressão (3.9), calculamos o vetor escore corrigido U^*, obtido derivando-se l^* em relação a todos estes parâmetros, possuindo dimensão $p + k + 1$ ($p = dim(\beta_x)$, $k = dim(\lambda)$ e $dim(\beta_z) = 1$). Assim,

$$
U^*(\phi) = (U_1^{\beta_x}, \ldots, U_p^{\beta_x}, U^{\beta_z}, U_1^{\lambda}, \ldots, U_k^{\lambda}), \tag{3.10}
$$

sendo

$$
\begin{aligned}
U_j^{\beta_x} &= \sum_{i=1}^{n}\Big\{x_{ij}\Big[\nu_i - exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)[1 - S(y_i|\lambda)]\Big]\Big\}, \\
&\quad j = 1, \ldots, p, \\
U^{\beta_z} &= \sum_{i=1}^{n}\Big\{\nu_i w_i - (w_i - \beta_z\sigma_u^2)exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)[1 - S(y_i|\lambda)]\Big\}, \\
U_j^{\lambda} &= \sum_{i=1}^{n}\Big\{\frac{\nu_i}{f(y_i|\lambda)}\frac{\partial f(y_i|\lambda)}{\partial\lambda_j} + exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)\frac{\partial S(y_i|\lambda)}{\partial\lambda_j}\Big\}. \\
&\quad j = 1, \ldots, k.
\end{aligned}
$$

3.5 Especificando a Distribuição $F(\cdot)$

Na expressão (3.9) é necessário que especifiquemos a distribuição $F(\cdot)$ das variáveis latentes (R_i's), que representam os tempos associados aos riscos competitivos. A seguir determinamos a expressão do logaritmo da verossimilhança marginal corrigida para diferentes distribuições.

3.5.1 Modelo Exponencial

Substituindo as relações (1.7) e (1.8) em (3.9), e lembrando que o tempo t para cada indivíduo é representado pelas observações y_i, $i = 1, \ldots, n$, obte-

mos

$$l^*(\phi;\overline{D}) = \sum_{i=1}^{n}\Big\{\nu_i\Big(x_i'\beta_x + w_i\beta_z + \ln\lambda - \lambda y_i\Big) - exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)[1 - exp(-\lambda y_i)]\Big\}, \tag{3.11}$$

3.5.2 Modelo Weibull

Substituindo as relações (1.11) e (1.12) em (3.9), obtemos

$$l^*(\phi;\overline{D}) = \sum_{i=1}^{n}\Big\{\nu_i\Big(x_i'\beta_x + w_i\beta_z + \gamma + \ln(\rho y_i^{\rho-1}) - y_i^{\rho}e^{\gamma}\Big) - exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)[1 - exp(-y_i^{\rho}e^{\gamma})]\Big\}, \tag{3.12}$$

expressão obtida por Mizoi et al. (2007).

3.5.3 Modelo Exponencial por Partes

O logaritmo da função de verossimilhança marginal corrigida, dado por (3.9), aliado à relação (1.5), pode ser reescrito da seguinte forma:

$$l^*(\phi;\overline{D}) = \sum_{i=1}^{n}\Big\{\nu_i\Big(x_i'\beta_x + w_i\beta_z + \ln h(y_i|\lambda) + \ln S(y_i|\lambda)\Big) - exp\Big(x_i'\beta_x + w_i\beta_z - \frac{\beta_z^2\sigma_u^2}{2}\Big)[1 - S(y_i|\lambda)]\Big\}. \tag{3.13}$$

Para o MEP, o valor da função risco depende do intervalo em que a observação está localizada. Assim, para uma observação y_i, $i = 1, \ldots, n$, o risco é dado por

$$h(y_i|\lambda) = \prod_{j=1}^{k}\lambda_j^{b_{ij}} \tag{3.14}$$

sendo

$$b_{ij} = \begin{cases} 1 & \text{se } y_i \in I_j; \\ 0 & \text{caso contrário.} \end{cases}$$

A função de sobrevivência para o MEP, $S(y_i|\lambda)$, é dada pela relação (1.14). Portanto, o logaritmo da função de verossimilhança marginal corrigida para o MEP é obtido substituindo as relações (1.14) e (3.14) em (3.13).

Temos então a expressão do logaritmo da função de verossimilhança corrigida para um conjunto de dados com uma covariável medida com erro e utilizando o MEP para a distribuição de probabilidade das variáveis latentes. No próximo Capítulo realizamos simulações para avaliar o desempenho do MEP em tal situação, utilizando um procedimento análogo ao de Mizoi et al. (2007), que utiliza a distribuição exponencial.

Capítulo 4

Simulação

4.1 Simulando o Conjunto de Dados

Suponha que temos duas covariáveis (x e z) associadas a cada indivíduo, e que apenas uma delas (z) é medida com erro. As covariáveis serão consideradas como valores fixos para cada indívíduo, embora seus valores sejam gerados a partir da distribuição Normal. Assim, para $i = 1, \ldots, n$, o conjunto de dados será gerado da seguinte forma:

$$x_i \sim N(0, 1),$$

$$z_i \sim N(0, 1),$$

$$w_i = z_i + u_i,$$

com $u_i \sim N(0, \sigma_u^2)$.

Temos ainda que:

$$\theta_i = exp(\beta_x x_i + \beta_z z_i), \text{ com } \beta_x = \beta_z = 0,3 \ ,$$
$$N_i \sim Poisson(\theta_i),$$

Consideramos para as variáveis latentes, representando os tempos associados aos riscos competitivos (R_i's), uma distribuição exponencial caracterizada no Capítulo 1 pelas expressões (1.7) e (1.8). Usaremos a parametrização

$\lambda = exp(\gamma)$, devido à vantagem computacional na estimação do parâmetro γ que pode assumir qualquer valor real. Portanto, as variáveis R_{ij} serão geradas considerando $\gamma = -1,5$, com

$$R_{ij} \sim Exp(e^{\gamma}),\ j = 1, \ldots, N_i.$$

Assim, para os indivíduos não imunes ($N > 0$), temos

$$T_i = min\{R_{i1}, \ldots, R_{iN_i}\}.$$

Além disso, simulamos censuras aleatórias da seguinte forma:

$$C_i \sim Exp(e^{\mu}),$$

sendo $\mu = \mu(p_c)$ obtido por simulação de acordo com a proporção de censura fixada para a população de não-curados p_c, através da relação

$$p_c = P(T > C).$$

E por fim, temos

$$Y_i = min\{T_i, C_i\}.$$

Todo o processo foi implementado com o software R [1], utilizando sua linguagem de programação, os geradores de números aleatórios *rnorm*, *rpois* e *rexp*, a função *optim* para maximizar as funções envolvidas (verossimilhança e verossimilhança corrigida), dentre outras funções. Os principais algoritmos implementados estão descritos no Apêndice C.

[1] http://www.r-project.org/

4.2 Efeito do Erro de Medida na Covariável Medida sem Erro

Baseado nos dados simulados, maximizaremos as expressões (3.3) e (3.9) para obtermos respectivamente as estimativas Naive e EC (escore corrigido). Mizoi et al. (2007) utiliza este mesmo procedimento de simulação, porém com apenas uma covariável (z) que é medida com erro. Adicionamos uma segunda covariável (x), sem erro de medida, para analisar o comportamento da estimativa do coeficiente de regressão associado a ela (β_x) na presença de outra covariável medida com erro em amostras finitas, utilizando o método do escore corrigido. Para esta análise, geramos amostras sem proporção de censura ($p_c = 0\%$), com tamanhos $n = 50$, $n = 100$ e $n = 300$ e variâncias do erro de medida $\sigma_u^2 = 0,1$, $\sigma_u^2 = 0,5$ e $\sigma_u^2 = 1$. Para cada combinação destes valores, geramos 1500 réplicas e apresentamos os resultados empíricos para a média, o erro padrão (EP) e a raiz quadrada do erro quadrático médio (REQM) na Tabela 4.1.

		n=50			n=100			n= 300		
Estimador		média	EP	REQM	média	EP	REQM	média	EP	REQM
β_x		0,3131	0,1977	0,1981	0,3064	0,1310	0,1311	0,3000	0,0751	0,0751
β_z		0,3168	0,1942	0,1949	0,2994	0,1345	0,1345	0,2999	0,0739	0,0739
γ		-1,5098	0,2125	0,2126	-1,5058	0,1480	0,1481	-1,5016	0,0849	0,0848
$\sigma_u^2 = 0,1$										
	β_x	0,3097	0,1989	0,1990	0,3152	0,1376	0,1384	0,2994	0,0732	0,0731
Naive	β_z	0,2871	0,1937	0,1940	0,2825	0,1302	0,1313	0,2738	0,0719	0,0765
	γ	-1,5135	0,2025	0,2029	-1,5072	0,1455	0,1457	-1,5010	0,0828	0,0828
	β_x	0,3122	0,2011	0,2014	0,3168	0,1386	0,1396	0,3006	0,0735	0,0735
EC	β_z	0,3229	0,2221	0,2232	0,3147	0,1469	0,1476	0,3031	0,0806	0,0807
	γ	-1,5188	0,2036	0,2044	-1,5111	0,1460	0,1464	-1,5040	0,0831	0,0831
$\sigma_u^2 = 0,5$										
	β_x	0,3137	0,1954	0,1958	0,3011	0,1374	0,1374	0,3046	0,0748	0,0749
Naive	β_z	0,2093	0,1592	0,1831	0,2033	0,1067	0,1439	0,1992	0,0595	0,1171
	γ	-1,4954	0,2088	0,2087	-1,4967	0,1486	0,1485	-1,4964	0,0859	0,0860
	β_x	0,3208	0,2091	0,2101	0,3069	0,1431	0,1432	0,3092	0,0768	0,0774
EC	β_z	0,3446	0,2960	0,2992	0,3295	0,1955	0,1977	0,3076	0,0980	0,0983
	γ	-1,5166	0,2158	0,2164	-1,5132	0,1522	0,1527	-1,5078	0,0871	0,0874
$\sigma_u^2 = 1$										
	β_x	0,3017	0,1843	0,1842	0,3035	0,1336	0,1336	0,2948	0,0749	0,0751
Naive	β_z	0,1550	0,1420	0,2029	0,1490	0,0905	0,1761	0,1514	0,0524	0,1576
	γ	-1,4910	0,2092	0,2093	-1,4865	0,1440	0,1446	-1,4845	0,0870	0,0884
	β_x	0,3224	0,3141	0,3147	0,3140	0,1484	0,1490	0,3039	0,0807	0,0808
EC	β_z	0,3013	0,4018	0,4016	0,3264	0,2372	0,2386	0,3235	0,1281	0,1302
	γ	-1,5226	0,2227	0,2237	-1,5103	0,1506	0,1509	-1,5034	0,0901	0,0902

Tabela 4.1: Média, erro padrão (EP) e raiz quadrada do erro quadrático médio (REQM) empíricos das estimativas naive e escore corrigido (NAIVE e EC) para dados simulados do modelo exponencial, com $\beta_x = 0,3$, $\beta_z = 0,3$ e $\gamma = -1,50$, sem censura, com proporção média de imunes 0,36 e para diferentes valores da variância do erro de medida (σ_u^2), com base em 1500 réplicas.

Assim como Mizoi et al. (2007), vimos claramente a atenuação do estimador naive de β_z, coeficiente de regressão da covariável medida com erro, de acordo com o aumento da variância do erro de medida. Quanto ao coeficiente de regressão da covariável medida sem erro, β_x, não percebemos através da Tabela 4.1 alterações significativas, mas apenas que a sua estimativa naive é sempre inferior a estimativa escore corrigido. Para melhor avaliar esta diferença e determinar se há vício das estimativas em relação ao verdadeiro valor, $\beta_x = 0,3$, propomos a utilização das quantidades

$$d_x = \frac{\widehat{\beta}_{x_{naive}} - \beta_x}{\beta_x}, \tag{4.1}$$

$$d_{xEC} = \frac{\widehat{\beta}_{x_{EC}} - \beta_x}{\beta_x}, \tag{4.2}$$

$$d = \widehat{\beta}_{x_{EC}} - \widehat{\beta}_{x_{naive}}. \tag{4.3}$$

Contruimos os gráficos destas 3 quantiades em função da variância do erro de medida, para diferentes tamanhos de amostra ($n = 50$, $n = 100$ e $n = 300$). Para cada combinação, geramos 1000 réplicas para obtermos as médias. Os resultados são representados graficamente nas Figuras (4.1), (4.2) e (4.3).

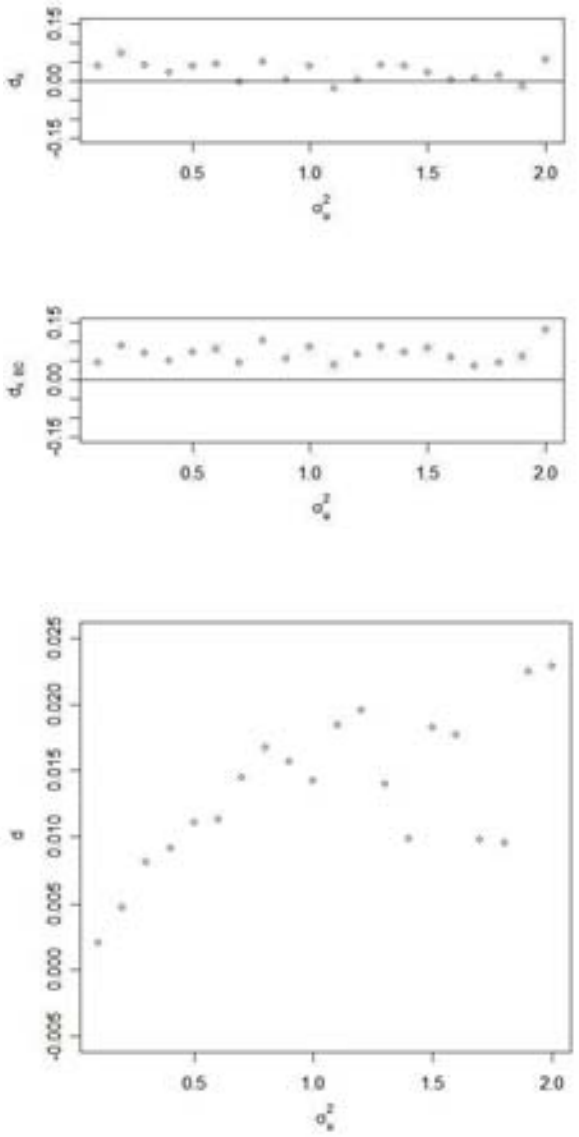

Figura 4.1: Variação de d_x, d_{xEC} e d em função de σ_u^2 para $n = 50$.

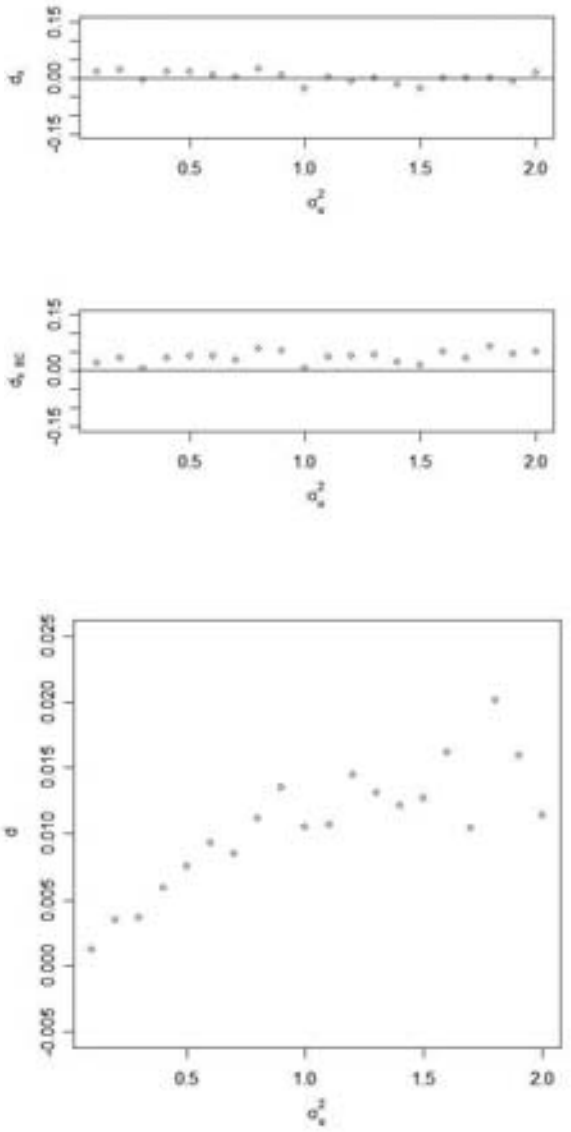

Figura 4.2: Variação de d_x, d_{xEC} e d em função de σ_u^2 para $n = 100$.

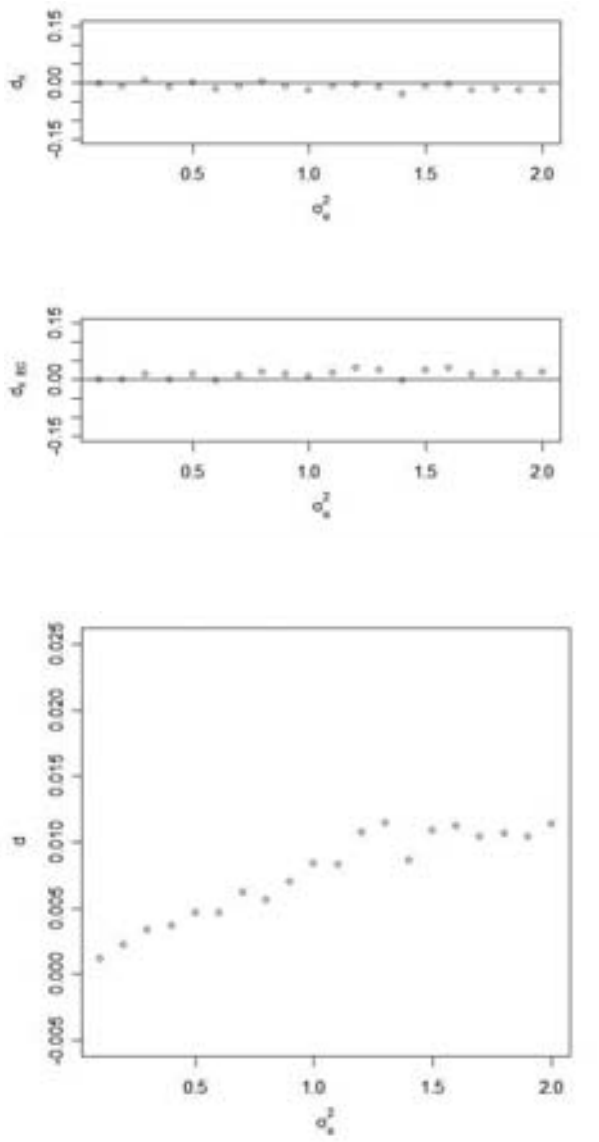

Figura 4.3: Variação de d_x, d_{xEC} e d em função de σ_u^2 para $n = 300$.

Das Figuras (4.1), (4.2) e (4.3), podemos extrair alguns resultados. Para amostras de tamanho $n = 50$ e $n = 100$, o estimador escore corrigido mostrou-se viciado em relação ao verdadeiro valor, fornecendo em média para este exemplo valores acima do verdadeiro valor. Este vício não é abalado pela variação do valor da variância do erro de medida, mas sim pelo tamanho da amostra. À medida que n cresce, o vício diminui. O mesmo não acontece com o estimador naive, que fornece, desde $n = 50$, estimativas que variam em torno do verdadeiro valor, e que não são afetadas pelo aumento da variância do erro de medida da covariável medida com erro. Quanto à diferença entre os dois estimadores, vimos que o escore corrigido fornece valores sempre diferentes, neste caso superiores, dos valores fornecidos pelo estimador naive. Essa diferença cresce à medida que a variância do erro de medida cresce, pelo menos até $\sigma_u^2 = 1$, e diminui com o aumento do tamanho da amostra, como esperado.

4.3 Adequação do MEP ao Modelo com Fração de Cura e Erro de Medida

Ao utilizarmos o MEP, devemos inicialmente definir a partição do eixo do tempo. Tal partição em geral é feita de maneira arbitrária, de forma que, em situações práticas, diferentes partições proporcionam resultados distintos.

Como o modelo exponencial possui uma função risco constante em todo o intervalo de tempo, podemos considerá-lo como um caso particular do MEP com apenas um subintervalo em $\Re^+$. Portanto, estaremos gerando valores de uma exponencial particionada com apenas 1 subintervalo. Entretanto, para verificar a flexibilidade do modelo em ajustar funções risco, consideraremos para fins de estimação um modelo exponencial por partes com 4 subintervalos.

O conjunto de dados é gerado de maneira similar ao exposto no início deste Capítulo, porém utilizando apenas uma covariável (z), que é medida

com erro. Iremos considerar uma partição com 4 valores (determinando 4 subintervalos) determinados pelos quartis da distribuição exponencial, ou seja, $\{s_1, s_2, s_3, s_4\}$ tais que $P(T < s_1) = 0,25$, $P(T < s_2) = 0,50$, $P(T < s_3) = 0,75$ e $s_4 = max\{\delta_i y_i, i = 1, \ldots, n\}$, sendo δ_i um indicador de não-imunidade ($\delta_i = 0$ se o indivíduo for imune, $\delta_i = 1$ caso contrário). Assim, além do β_z, temos 4 parâmetros do modelo a serem estimados: $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. Utilizamos a parametrização $\lambda_j = exp(\gamma_j)$, $j = 1, 2, 3$ e 4, pela vantagem no processo de maximização onde o γ pode assumir qualquer valor real.

Para esta análise, geramos amostras com proporções de censura $p_c = 0\%$, $p_c = 15\%$, $p_c = 25\%$ e $p_c = 50\%$ sobre os indivíduos não curados, com tamanhos de amostra de $n = 50$, $n = 100$ e $n = 300$ e variâncias do erro de medida $\sigma_u^2 = 0,1$, $\sigma_u^2 = 0,5$ e $\sigma_u^2 = 1$. Para cada combinação destes valores, geramos 1500 réplicas e apresentamos a média, o erro padrão (EP) e a raiz quadrada do erro quadrático médio (REQM) nas Tabelas 4.2 à 4.5.

σ_u^2	Par	n=50 média	n=50 EP	n=50 REQM	n=100 média	n=100 EP	n=100 REQM	n= 300 média	n= 300 EP	n= 300 REQM
	γ_1	-1,5419	0,3399	0,3424	-1,5443	0,2354	0,2394	-1,5050	0,1286	0,1286
	γ_2	-1,5361	0,3810	0,3825	-1,5043	0,2621	0,2621	-1,5116	0,1497	0,1501
-	γ_3	-1,5413	0,4561	0,4578	-1,5067	0,3139	0,3138	-1,5021	0,1692	0,1692
	γ_4	-1,4044	0,5462	0,5543	-1,4394	0,3501	0,3551	-1,4826	0,1925	0,1933
	β_z	0,3116	0,1959	0,1961	0,3052	0,1369	0,1370	0,3012	0,0752	0,0752
	$\gamma_{1_{naive}}$	-1,5417	0,3254	0,3280	-1,5188	0,2320	0,2327	-1,5040	0,1284	0,1285
	$\gamma_{2_{naive}}$	-1,5349	0,3900	0,3914	-1,5257	0,2726	0,2738	-1,4997	0,1487	0,1486
	$\gamma_{3_{naive}}$	-1,5281	0,5126	0,5132	-1,5152	0,3184	0,3187	-1,5044	0,1718	0,1718
	$\gamma_{4_{naive}}$	-1,3685	0,5538	0,5691	-1,4271	0,3744	0,3813	-1,4861	0,1920	0,1925
	$\beta_{z_{naive}}$	0,2787	0,1870	0,1881	0,2769	0,1331	0,1350	0,2745	0,0693	0,0738
0,1	$\gamma_{1_{EC}}$	-1,5491	0,3277	0,3312	-1,5242	0,2330	0,2342	-1,5083	0,1288	0,1290
	$\gamma_{2_{EC}}$	-1,5397	0,3910	0,3929	-1,5293	0,2732	0,2747	-1,5027	0,1488	0,1488
	$\gamma_{3_{EC}}$	-1,5308	0,5153	0,5161	-1,5175	0,3187	0,3191	-1,5062	0,1719	0,1720
	$\gamma_{4_{EC}}$	-1,3694	0,5542	0,5692	-1,4279	0,3743	0,3811	-1,4867	0,1921	0,1925
	$\beta_{z_{EC}}$	0,3125	0,2136	0,2139	0,3081	0,1501	0,1503	0,3037	0,0775	0,0776
	$\gamma_{1_{naive}}$	-1,5401	0,3342	0,3365	-1,4977	0,2261	0,2260	-1,4888	0,1263	0,1267
	$\gamma_{2_{naive}}$	-1,5358	0,3730	0,3746	-1,5070	0,2560	0,2560	-1,5018	0,1500	0,1499
	$\gamma_{3_{naive}}$	-1,5343	0,5993	0,6001	-1,5049	0,3063	0,3062	-1,4898	0,1758	0,1761
	$\gamma_{4_{naive}}$	-1,3774	0,5537	0,5669	-1,4397	0,3399	0,3451	-1,4790	0,1947	0,1957
	$\beta_{z_{naive}}$	0,1964	0,1487	0,1811	0,1994	0,1067	0,1466	0,2000	0,0602	0,1167
0,5	$\gamma_{1_{EC}}$	-1,5737	0,3474	0,3551	-1,5205	0,2311	0,2320	-1,5055	0,1278	0,1278
	$\gamma_{2_{EC}}$	-1,5565	0,3789	0,3829	-1,5221	0,2579	0,2587	-1,5131	0,1509	0,1514
	$\gamma_{3_{EC}}$	-1,5453	0,5838	0,5854	-1,5134	0,3075	0,3077	-1,4963	0,1764	0,1764
	$\gamma_{4_{EC}}$	-1,3842	0,5572	0,5689	-1,4426	0,3401	0,3448	-1,4811	0,1946	0,1954
	$\beta_{z_{EC}}$	0,3364	0,3169	0,3189	0,3178	0,1860	0,1867	0,3078	0,0981	0,0984
	$\gamma_{1_{naive}}$	-1,5340	0,3255	0,3272	-1,4971	0,2252	0,2252	-1,4838	0,1325	0,1334
	$\gamma_{2_{naive}}$	-1,5160	0,3735	0,3737	-1,5038	0,2539	0,2539	-1,4866	0,1474	0,1479
	$\gamma_{3_{naive}}$	-1,5010	0,4678	0,4677	-1,5074	0,2953	0,2953	-1,4934	0,1698	0,1699
	$\gamma_{4_{naive}}$	-1,3902	0,5585	0,5690	-1,4548	0,3411	0,3439	-1,4769	0,1912	0,1925
	$\beta_{z_{naive}}$	0,1339	0,1185	0,2040	0,1441	0,0867	0,1783	0,1476	0,0523	0,1611
1	$\gamma_{1_{EC}}$	-1,5760	0,3429	0,3512	-1,5341	0,2417	0,2440	-1,5103	0,1361	0,1364
	$\gamma_{2_{EC}}$	-1,5420	0,3848	0,3869	-1,5277	0,2590	0,2604	-1,5047	0,1492	0,1492
	$\gamma_{3_{EC}}$	-1,5150	0,4719	0,4720	-1,5209	0,2975	0,2981	-1,5039	0,1704	0,1704
	$\gamma_{4_{EC}}$	-1,3944	0,5544	0,5642	-1,4590	0,3415	0,3439	-1,4804	0,1915	0,1924
	$\beta_{z_{EC}}$	0,3306	0,5055	0,5062	0,3285	0,2774	0,2787	0,3127	0,1271	0,1277

Tabela 4.2: Estimativas para dados simulados do MEP, com $\beta_z = 0,30$, $\gamma_j = -1,50$, para $j = 1, \ldots, 4$, 1500 réplicas, $\pi_{\text{médio}} = 0,36$ (sem censura).

σ_u^2	Par	n=50 média	n=50 EP	n=50 REQM	n=100 média	n=100 EP	n=100 REQM	n= 300 média	n= 300 EP	n= 300 REQM
	γ_1	-1,3710	0,1989	0,3897	-1,3271	0,2413	0,2968	-1,3287	0,1365	0,2190
	γ_2	-1,3380	0,5842	0,6060	-1,2861	0,3050	0,3724	-1,2557	0,1650	0,2948
-	γ_3	-1,2567	1,3902	1,4109	-1,1850	0,3881	0,4998	-1,1950	0,2129	0,3720
	γ_4	-1,1886	1,9122	1,9368	-1,0388	0,9958	1,0971	-1,0860	0,3105	0,5174
	β_z	0,2981	0,1989	0,1989	0,2889	0,1319	0,1324	0,2817	0,0753	0,0775
	$\gamma_{1_{naive}}$	-1,3850	0,3495	0,3678	-1,3548	0,2281	0,2703	-1,3195	0,1334	0,2245
	$\gamma_{2_{naive}}$	-1,2882	0,4226	0,4726	-1,2745	0,2982	0,3737	-1,2619	0,1721	0,2938
	$\gamma_{3_{naive}}$	-1,2446	0,8634	0,9001	-1,2083	0,3825	0,4809	-1,1711	0,2188	0,3949
	$\gamma_{4_{naive}}$	-1,1534	2,2204	2,2466	-1,0023	0,9013	1,0294	-1,0779	0,3418	0,5431
	$\beta_{z_{naive}}$	0,2700	0,1878	0,1901	0,2687	0,1303	0,1340	0,2570	0,0707	0,0827
0,1	$\gamma_{1_{EC}}$	-1,3932	0,3520	0,3677	-1,3611	0,2292	0,2679	-1,3242	0,1339	0,2209
	$\gamma_{2_{EC}}$	-1,2935	0,4237	0,4712	-1,2789	0,2989	0,3718	-1,2653	0,1724	0,2912
	$\gamma_{3_{EC}}$	-1,2585	1,0528	1,0798	-1,2108	0,3828	0,4797	-1,1731	0,2189	0,3933
	$\gamma_{4_{EC}}$	-1,1310	1,4679	1,5131	-0,9922	0,8041	0,9508	-1,0786	0,3420	0,5426
	$\beta_{z_{EC}}$	0,3036	0,2174	0,2173	0,2994	0,1476	0,1475	0,2846	0,0792	0,0806
	$\gamma_{1_{naive}}$	-1,3628	0,3431	0,3694	-1,3335	0,2340	0,2872	-1,3088	0,1364	0,2349
	$\gamma_{2_{naive}}$	-1,3014	0,4448	0,4870	-1,2765	0,2919	0,3676	-1,2509	0,1645	0,2985
	$\gamma_{3_{naive}}$	-1,2072	0,9121	0,9576	-1,2012	0,3904	0,4915	-1,1742	0,2133	0,3894
	$\gamma_{4_{naive}}$	-1,1639	1,3061	1,3482	-0,9997	0,6467	0,8175	-1,0778	0,3342	0,5384
	$\beta_{z_{naive}}$	0,1870	0,1459	0,1845	0,1924	0,1104	0,1541	0,1859	0,0596	0,1287
0,5	$\gamma_{1_{EC}}$	-1,3995	0,3626	0,3762	-1,3596	0,2425	0,2802	-1,3269	0,1389	0,2220
	$\gamma_{2_{EC}}$	-1,3246	0,4532	0,4858	-1,2942	0,2965	0,3608	-1,2637	0,1657	0,2886
	$\gamma_{3_{EC}}$	-1,2305	1,0531	1,0867	-1,2114	0,3922	0,4868	-1,1816	0,2141	0,3837
	$\gamma_{4_{EC}}$	-1,1612	1,2882	1,3316	-1,0099	0,6960	0,8511	-1,0801	0,3342	0,5366
	$\beta_{z_{EC}}$	0,3265	0,3904	0,3912	0,3075	0,1948	0,1948	0,2860	0,0962	0,0972
	$\gamma_{1_{naive}}$	-1,3329	0,3379	0,3769	-1,3201	0,2356	0,2964	-1,3087	0,1348	0,2339
	$\gamma_{2_{naive}}$	-1,2839	0,4430	0,4928	-1,2526	0,2955	0,3853	-1,2410	0,1708	0,3102
	$\gamma_{3_{naive}}$	-1,1886	0,6953	0,7616	-1,2038	0,4031	0,5001	-1,1713	0,2143	0,3923
	$\gamma_{4_{naive}}$	-1,1237	1,2704	1,3246	-0,9838	0,6802	0,8537	-1,0815	0,3247	0,5296
	$\beta_{z_{naive}}$	0,1228	0,1178	0,2128	0,1352	0,0875	0,1866	0,1394	0,0521	0,1689
1	$\gamma_{1_{EC}}$	-1,3689	0,3551	0,3784	-1,3597	0,2491	0,2859	-1,3384	0,1399	0,2137
	$\gamma_{2_{EC}}$	-1,3081	0,4586	0,4969	-1,2786	0,3041	0,3761	-1,2616	0,1735	0,2948
	$\gamma_{3_{EC}}$	-1,2041	0,7554	0,8111	-1,2188	0,4055	0,4933	-1,1828	0,2153	0,3834
	$\gamma_{4_{EC}}$	-1,1412	1,3099	1,3578	-0,9881	0,6981	0,8654	-1,0851	0,3250	0,5270
	$\beta_{z_{EC}}$	0,3165	0,6609	0,6609	0,3033	0,2930	0,2929	0,2956	0,1273	0,1273

Tabela 4.2: Estimativas para dados simulados do MEP, com $\beta_z = 0,30$, $\gamma_j = -1,50$, para $j = 1, \ldots, 4$, 1500 réplicas, $\pi_{\text{médio}} = 0,36$ (15% de censura).

σ_u^2	Par	n=50 média	n=50 EP	n=50 REQM	n=100 média	n=100 EP	n=100 REQM	n= 300 média	n= 300 EP	n= 300 REQM
	γ_1	-1,2659	0,3771	0,4437	-1,2198	0,2497	0,3753	-1,2088	0,1447	0,3252
	γ_2	-1,1081	0,5110	0,6438	-1,0956	0,3582	0,5401	-1,0718	0,1945	0,4703
-	γ_3	-0,7936	0,9733	1,2023	-0,8928	0,5700	0,8327	-0,9165	0,2745	0,6448
	γ_4	-1,5247	1,8480	1,8475	-0,9939	1,4966	1,5794	-0,6909	0,8273	1,1570
	β_z	0,2859	0,2013	0,2017	0,2703	0,1388	0,1419	0,2681	0,0780	0,0842
	$\gamma_{1_{naive}}$	-1,2574	0,3768	0,4480	-1,2178	0,2481	0,3757	-1,1975	0,1439	0,3349
	$\gamma_{2_{naive}}$	-1,1306	0,4826	0,6076	-1,0857	0,3532	0,5443	-1,0707	0,1931	0,4707
	$\gamma_{3_{naive}}$	-0,7225	0,9898	1,2583	-0,9166	0,5896	0,8293	-0,9251	0,2782	0,6386
	$\gamma_{4_{naive}}$	-1,5178	1,5827	1,5823	-0,9752	1,4537	1,5451	-0,7203	0,6462	1,0126
	$\beta_{z_{naive}}$	0,2627	0,1904	0,1940	0,2462	0,1254	0,1364	0,2429	0,0750	0,0943
0,1	$\gamma_{1_{EC}}$	-1,2667	0,3799	0,4457	-1,2242	0,2493	0,3717	-1,2028	0,1446	0,3305
	$\gamma_{2_{EC}}$	-1,1370	0,4842	0,6050	-1,0904	0,3538	0,5412	-1,0746	0,1935	0,4673
	$\gamma_{3_{EC}}$	-0,7221	0,9717	1,2445	-0,9193	0,5898	0,8275	-0,9273	0,2784	0,6367
	$\gamma_{4_{EC}}$	-1,5287	1,6025	1,6022	-0,9762	1,4430	1,5347	-0,7198	0,6321	1,0040
	$\beta_{z_{EC}}$	0,2955	0,2194	0,2194	0,2740	0,1415	0,1439	0,2688	0,0838	0,0894
	$\gamma_{1_{naive}}$	-1,2385	0,3604	0,4452	-1,2153	0,2436	0,3746	-1,1906	0,1436	0,3411
	$\gamma_{2_{naive}}$	-1,1306	0,5217	0,6391	-1,0846	0,3612	0,5504	-1,0626	0,1965	0,4795
	$\gamma_{3_{naive}}$	-0,7487	0,9352	1,1994	-0,8801	0,6150	0,8730	-0,9207	0,2801	0,6434
	$\gamma_{4_{naive}}$	-1,4767	1,5537	1,5533	-0,9911	1,3219	1,4161	-0,7073	0,6666	1,0356
	$\beta_{z_{naive}}$	0,1746	0,1533	0,1980	0,1801	0,1116	0,1638	0,1807	0,0604	0,1337
0,5	$\gamma_{1_{EC}}$	-1,2770	0,3791	0,4397	-1,2438	0,2537	0,3605	-1,2116	0,1460	0,3232
	$\gamma_{2_{EC}}$	-1,1551	0,5310	0,6331	-1,1037	0,3659	0,5393	-1,0778	0,1979	0,4663
	$\gamma_{3_{EC}}$	-0,7573	0,9135	1,1771	-0,8944	0,6500	0,8882	-0,9291	0,2808	0,6362
	$\gamma_{4_{EC}}$	-1,4908	1,5066	1,5061	-1,0348	1,3959	1,4709	-0,7113	0,6686	1,0338
	$\beta_{z_{EC}}$	0,2963	0,3419	0,3419	0,2961	0,2512	0,2511	0,2799	0,0998	0,1018
	$\gamma_{1_{naive}}$	-1,2069	0,3511	0,4572	-1,2074	0,2615	0,3923	-1,1840	0,1437	0,3471
	$\gamma_{2_{naive}}$	-1,1031	0,5249	0,6579	-1,0822	0,3654	0,5550	-1,0543	0,1988	0,4879
	$\gamma_{3_{naive}}$	-0,7279	0,8865	1,1754	-0,9050	0,6066	0,8496	-0,9265	0,2760	0,6364
	$\gamma_{4_{naive}}$	-1,5060	1,5229	1,5224	-0,9867	1,4066	1,4969	-0,7114	0,6203	1,0032
	$\beta_{z_{naive}}$	0,1137	0,1210	0,2222	0,1319	0,0873	0,1894	0,1367	0,0514	0,1712
1	$\gamma_{1_{EC}}$	-1,2421	0,3982	0,4743	-1,2518	0,2809	0,3748	-1,2189	0,1484	0,3178
	$\gamma_{2_{EC}}$	-1,1246	0,5446	0,6613	-1,1146	0,3794	0,5407	-1,0792	0,2021	0,4668
	$\gamma_{3_{EC}}$	-0,7415	0,8943	1,1724	-0,9235	0,6119	0,8406	-0,9401	0,2776	0,6249
	$\gamma_{4_{EC}}$	-1,5590	1,5862	1,5868	-1,0084	1,3966	1,4801	-0,7123	0,5990	0,9894
	$\beta_{z_{EC}}$	0,3024	0,7019	0,7017	0,3081	0,3383	0,3383	0,2936	0,1288	0,1289

Tabela 4.2: Estimativas para dados simulados do MEP, com $\beta_z = 0,30$, $\gamma_j = -1,50$, para $j = 1, \ldots, 4$, 1500 réplicas, $\pi_{\text{médio}} = 0,36$. (25% de censura)

σ_u^2	Par	n=50 média	n=50 EP	n=50 REQM	n=100 média	n=100 EP	n=100 REQM	n= 300 média	n= 300 EP	n= 300 REQM
	γ_1	-1,1235	0,5264	0,6470	-1,0396	0,3587	0,5836	-0,9878	0,2096	0,5534
	γ_2	0,1285	1,5401	2,2411	-0,0319	1,1481	1,8634	-0,5555	0,7128	1,1831
-	γ_3	-2,1377	1,4094	1,5465	-2,1774	1,6647	1,7967	-2,2632	2,8224	2,9229
	γ_4	-1,9582	1,3099	1,3873	-1,9648	1,3571	1,4341	-1,9454	1,9707	2,0197
	β_z	0,2861	0,2631	0,2634	0,2802	0,1674	0,1686	0,2669	0,0959	0,1015
	$\gamma_{1_{naive}}$	-1,1235	0,5575	0,6726	-1,0304	0,3718	0,5989	-0,9788	0,2045	0,5598
	$\gamma_{2_{naive}}$	0,1804	1,3057	2,1278	0,0023	1,0694	1,8438	-0,5478	0,6946	1,1785
	$\gamma_{3_{naive}}$	-2,1933	1,2731	1,4493	-2,1521	1,4570	1,5958	-1,8142	2,2508	2,2719
	$\gamma_{4_{naive}}$	-1,9335	1,1409	1,2201	-2,0038	1,3226	1,4149	-1,9774	1,6266	1,6947
	$\beta_{z_{naive}}$	0,2640	0,2445	0,2470	0,2437	0,1629	0,1723	0,2442	0,0886	0,1047
0,1	$\gamma_{1_{EC}}$	-1,1422	0,5741	0,6763	-1,0413	0,3762	0,5931	-0,9879	0,2059	0,5519
	$\gamma_{2_{EC}}$	0,1624	1,3113	2,1171	-0,0080	1,0731	1,8376	-0,5580	0,6953	1,1707
	$\gamma_{3_{EC}}$	-2,2302	1,3366	1,5226	-2,1019	1,3466	1,4746	-1,7930	2,2306	2,2491
	$\gamma_{4_{EC}}$	-1,9538	1,1804	1,2642	-2,0649	1,1720	1,3006	-2,0391	1,6187	1,7056
	$\beta_{z_{EC}}$	0,3025	0,2925	0,2924	0,2734	0,1865	0,1883	0,2714	0,0998	0,1038
	$\gamma_{1_{naive}}$	-1,0765	0,5270	0,6759	-1,0168	0,3652	0,6057	-0,9553	0,1998	0,5801
	$\gamma_{2_{naive}}$	0,2440	1,3397	2,1989	0,0008	1,1138	1,8687	-0,5047	0,7074	1,2210
	$\gamma_{3_{naive}}$	-2,2073	1,3524	1,5257	-2,0543	1,4723	1,5728	-1,8937	2,3234	2,3558
	$\gamma_{4_{naive}}$	-1,9201	1,1740	1,2465	-1,9648	1,1997	1,2862	-1,9716	1,6977	1,7615
	$\beta_{z_{naive}}$	0,1678	0,1745	0,2189	0,1855	0,1346	0,1766	0,1763	0,0746	0,1445
0,5	$\gamma_{1_{EC}}$	-1,1319	0,5837	0,6899	-1,0749	0,3954	0,5805	-0,9900	0,2076	0,5506
	$\gamma_{2_{EC}}$	0,1843	1,3792	2,1767	-0,0560	1,1098	1,8210	-0,5432	0,7073	1,1897
	$\gamma_{3_{EC}}$	-2,2046	1,3085	1,4857	-2,0588	1,4453	1,5491	-1,9213	2,4318	2,4672
	$\gamma_{4_{EC}}$	-1,9950	1,2688	1,3615	-2,0444	1,2879	1,3979	-1,9420	1,8412	1,8929
	$\beta_{z_{EC}}$	0,3262	0,6998	0,7001	0,3143	0,2624	0,2627	0,2784	0,1274	0,1292
	$\gamma_{1_{naive}}$	-1,0492	0,5323	0,6974	-0,9640	0,3371	0,6331	-0,9552	0,2031	0,5814
	$\gamma_{2_{naive}}$	0,2220	1,3621	2,1953	-0,0190	1,1097	1,8505	-0,5254	0,6922	1,1952
	$\gamma_{3_{naive}}$	-2,1586	1,3722	1,5216	-2,0612	1,4498	1,5542	-1,7864	2,1739	2,1919
	$\gamma_{4_{naive}}$	-1,9599	1,3118	1,3896	-2,0312	1,2955	1,3998	-2,0047	1,6881	1,7613
	$\beta_{z_{naive}}$	0,0986	0,1427	0,2469	0,1134	0,0981	0,2108	0,1270	0,0616	0,1836
1	$\gamma_{1_{EC}}$	-1,0307	0,6539	0,8047	-1,0125	0,3720	0,6131	-1,0095	0,2198	0,5375
	$\gamma_{2_{EC}}$	0,2235	1,3900	2,2139	-0,0585	1,1136	1,8213	-0,5839	0,7005	1,1531
	$\gamma_{3_{EC}}$	-2,2366	1,4533	1,6289	-2,0921	1,4159	1,5343	-1,8629	2,2627	2,2909
	$\gamma_{4_{EC}}$	-2,0501	1,4113	1,5143	-2,0228	1,1181	1,2339	-1,9845	1,6305	1,7005
	$\beta_{z_{EC}}$	0,3577	1,1929	1,1939	0,2877	0,4991	0,4991	0,2795	0,1557	0,1569

Tabela 4.2: Estimativas para dados simulados do MEP, com $\beta_z = 0,30$, $\gamma_j = -1,50$, para $j = 1, \ldots, 4$, 1500 réplicas, $\pi_{\text{médio}} = 0,36$ (50% de censura).

Em relação ao parâmetro β_z, vimos que o método do escore corrigido fornece estimativas melhores que as estimativas naive, que são atenuadas pela presença do erro de medida. O erro padrão das estimativas naive são menores que das estimativas escore corrigido, pois não levam em conta o aumento da variância ocasionado pela presença do erro de medida.

Vimos que, quando não há censura, os 4 parâmetros do MEP convergem para o mesmo valor do parâmetro do modelo exponencial utilizado para gerar os dados, mostrando que o modelo ajustado se aproxima do teórico. Porém, ao introduzirmos censuras, suas estimativas são seriamente afetadas.

Além disso, observamos que o erro quadrático médio das estimativas diminui com o aumento do tamanho da amostra, e aumenta com o crescimento da proporção de censura.

Capítulo 5

Aplicação

5.1 Caso 1: Evasão Escolar

5.1.1 Caracterização do Conjunto de Dados

Analisamos um conjunto de dados, coletados entre os anos de 1997 e 2004, correspondente à 354 alunos do curso de estatística da UFRN, que na ocasião de inscrição para o vestibular responderam um questionário sócio-econômico contendo 25 questões, das quais escolhemos 4 para relacionar ao modelo. Os dados foram obtidos de Freire & Valença (2005), e contêm informações sobre o tempo (em semestres) que os alunos ultrapassaram o período de 6 semestres para se graduar (o mínimo exigido são 7 semestres) e a situação quanto ao término do curso, graduados (falha ou evento de interesse) ou não-graduados (censura), até 2004.

As covariáveis consideradas foram:

1. Sexo (categorizada, 2 níveis):
   $x_1 = 1$ se for homem;
   $x_1 = 0$ se for mulher.

2. Idade (contínua): x_2

3. Grau de instrução da mãe (categorizada, 4 níveis):
   $x_{31} = 1$: Desconhece, analfabeta ou ensino fundamental incompleto;

$x_{32} = 1$: Ensino fundamental completo ou ensino médio incompleto;

$x_{33} = 1$: Ensino médio completo;

$x_{31} = x_{32} = x_{33} = 0$: Ensino superior incompleto, completo ou pós-graduação.

4. Meio de transporte que mais utiliza (categorizada, 2 níveis):

   $x_4 = 1$: transporte coletivo;

   $x_4 = 0$: carro próprio ou da família, outros.

5.1.2 Caracterização do Modelo

Consideramos para as variáveis latentes representando os tempos associados aos riscos competitivos (R_i's) o modelo paramétrico Weibull, descrito no Capítulo 1, com funções densidade de probabilidade e sobrevivência dadas por (1.11) e (1.12). A expressão do logaritmo da função de verossimilhança marginal (2.5) para o modelo weibull é dada por

$$\begin{aligned} l^*(\phi; \overline{D}) &= \sum_{i=1}^{n} \Big\{ \nu_i \Big(x_i^{'} \beta_x + \gamma + \ln(\rho y_i^{\rho-1}) - y_i^{\rho} e^{\gamma} \Big) - \\ &\quad exp\Big(x_i^{'} \beta_x \Big) [1 - exp(-y_i^{\rho} e^{\gamma})] \Big\}, \end{aligned}$$

sendo $\phi = (\beta_x^{'}, \lambda^{'})^{'}$, com $\lambda = (\rho, \gamma)^{'}$ os parâmetros da distribuição Weibull e $\beta_x = (\beta_1, \beta_1, \beta_{31}, \beta_{32}, \beta_{33}, \beta_4)^{'}$ os coeficientes de regressão das covariáveis $(x_1, x_2, x_{31}, x_{32}, x_{33}, x_4)$ envolvidas no modelo.

5.1.3 Resultados

Utilizamos a função *optim* do software R, associada ao método Quasi-Newton de Broyden, Fletcher, Goldforb e Shanno (BFGS), para obter as estimativas de máxima verossimilhança. Os resultados são apresentados na Tabela 5.1.

Tabela 5.1 - Estimativas de máxima verossimilhança. Dados de evasão do curso de Estatística UFRN, $n = 354$ alunos, modelo Weibull.

Descrição	**Parâmetro**	**Estimativa**	**EP**	**P-valor**
sexo	β_1	$-0,4419$	$0,1996$	$0,0269$
idade	β_2	$-0,0305$	$0,0139$	$0,0278$
grau de	β_{31}	$-0,4187$	$0,3493$	
instrução	β_{32}	$-0,1242$	$0,2877$	$0,0089$
da mãe	β_{33}	$-0,3947$	$0,2782$	
transporte	β_4	$0,5611$	$0,2498$	$0,0250$
parâmetros	γ	$-3,1404$	$0,2557$	$< 0,0001$
Weibull	$\ln \rho$	$0,5983$	$0,0763$	$< 0,0001$

Com base nas estimativas dadas na Tabela 5.1 e na relação (2.2), a fração de cura fica relacionada às covariáveis através da equação

$$\begin{aligned} \pi(\mathbf{x}) &= exp\Big(- exp(- 0,4419x_1 - 0,0305x_2 - 0,4187x_{31} \\ &\quad -0,1242x_{32} - 0,3947x_{33} + 0,5611x_4)\Big). \end{aligned} \tag{5.1}$$

A fração de cura para a amostra estudada possui a distribuição de frequência sumarizada na Tabela 5.2. e representada na Figura 5.1.

Tabela 5.2 - Sumário de estatísticas para π.
Dados de evasão escolar.

Min.	1º Qu	Mediana	Média	3º Qu	Max.
0,3524	0,5244	0,6057	0,6037	0,6783	0,9127

Como exemplo, imaginemos os sequintes perfis:

- **Perfil 1:** um homem de 23 anos, tendo mãe que possui ensino fundamental completo ou médio incompleto e acostumado a andar de transporte coletivo. Em termos de covariáveis, esse perfil é representado pelo vetor $\mathbf{x} = (1, 23, 0, 1, 0, 1)$. Para este perfil, o modelo fornece, através da relação (5.1), $\pi = 0,6106$.

- **Perfil 2:** uma moça de 17 anos, tendo mãe que possui ensino superior (completo ou incompleto) ou pós-graduação e acostumada a andar de transporte coletivo. Temos $\mathbf{x} = (0, 17, 0, 0, 0, 1)$, $\pi(\mathbf{x}) = 0,3522$.

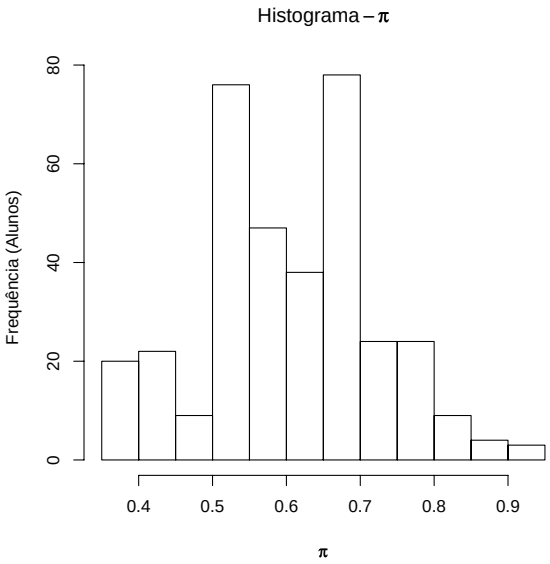

Figura 5.1: Histograma dos valores da fração de cura (π) dos perfis presentes na amostra.

- **Perfil 3:** um rapaz de 17 anos, tendo mãe que possui ensino superior (completo ou incompleto) ou pós-graduação e acostumado a andar de transporte coletivo. Temos $\mathbf{x} = (1, 17, 0, 0, 0, 1)$, $\pi(\mathbf{x}) = 0,5113$.

- **Perfil 4:** Um homem de 49 anos, tendo mãe que é analfabeta ou possui o ensino fundamental incompleto e acostumado a andar de carro próprio ou da família. Temos $\mathbf{x} = (1, 49, 1, 0, 0, 0)$, $\pi(\mathbf{x}) = 0,9095$.

O modelo mostrou uma adequação bastante satisfatória ao conjunto de dados, retratando com fidelidade a real situação do problema de evasão que ocorre no curso de estatística da UFRN. Comparando o valor que a função de sobrevivência estimada (Kaplan-meier - Figura 1.1) se acomoda com a média das proporções de imunidade fornecida pelo modelo, vimos que são valores bem próximos, em torno de 0,60. Vimos tambem que é possível traçar perfis e analisá-los quanto a chance de evasão.

Analisando os perfis exemplificados, constatamos algumas particularidades. Comparando os perfis 2 e 3, percebemos que a chance de um homem abandonar o curso e nunca se formar é maior que a de uma mulher sob condições sócio-econômicas equivalentes. Determinamos os perfis com maior (Perfil 4) e menor (Perfil 2) chance de evasão dentro da amostra.

Portanto, o modelo se mostrou uma ferramenta adequada para tratar com o problema de estimação da taxa de evasão do curso de graduação.

5.2 Caso 2: Câncer de Mama

5.2.1 Caracterização do Conjunto de Dados

Analisamos um conjunto de dados, obtidos de Macedo & Valença (2005), contendo informações a respeito de $n = 355$ pacientes com câncer de mama atendidas no Hospital Prof. Dr. Luiz Antônio no período de 1991 à 1995. Este Hospital é uma unidade filantrópica que compõe a Liga Norte-Rio-Grandense contra o Câncer.

A coleta de dados foi realizada de forma retrospectiva a partir dos prontuários de 485 pacientes, dos quais 130 foram excluídos por não atenderem critérios de inclusão do estudo original, registrados no Hospital Dr. Luiz Antônio com diagnóstico de câncer de mama comprovado através de exame anátomo-patológico. Todas as pacientes se submeteram a tratamento cirúrgico para retirada do tumor e a medida de interesse foi o tempo decorrido (em meses) entre a remissão e a recidiva do câncer de mama, ou seja, o tempo livre da doença. A Figura 5.2 apresenta a curva de sobrevivência estimada (estimador Kaplan-Meier) para o conjunto de dados de câncer de mama, mostrando uma alta proporção de censura à direita.

Consideramos as seguintes covariáveis na nossa análise:

1. EST: estadiamento do tumor ressecado (categorizada, 2 níveis):
   $x_1 = 1$ se estadiamento IIIA ou IIIB;
   $x_1 = 0$ se estadiamento 0, 1, IIA ou IIB.

2. NLR: número de linfonodos ressecados (discreta).

3. NLC: número de linfonodos comprometidos (discreta).

4. TNC: tratamentos não-cirúrgicos (categorizada, 4 níveis):
   $x_{41} = 1$: QH (quimio-hormonioterapia);
   $x_{42} = 1$: RH (radio-hormonioterapia);

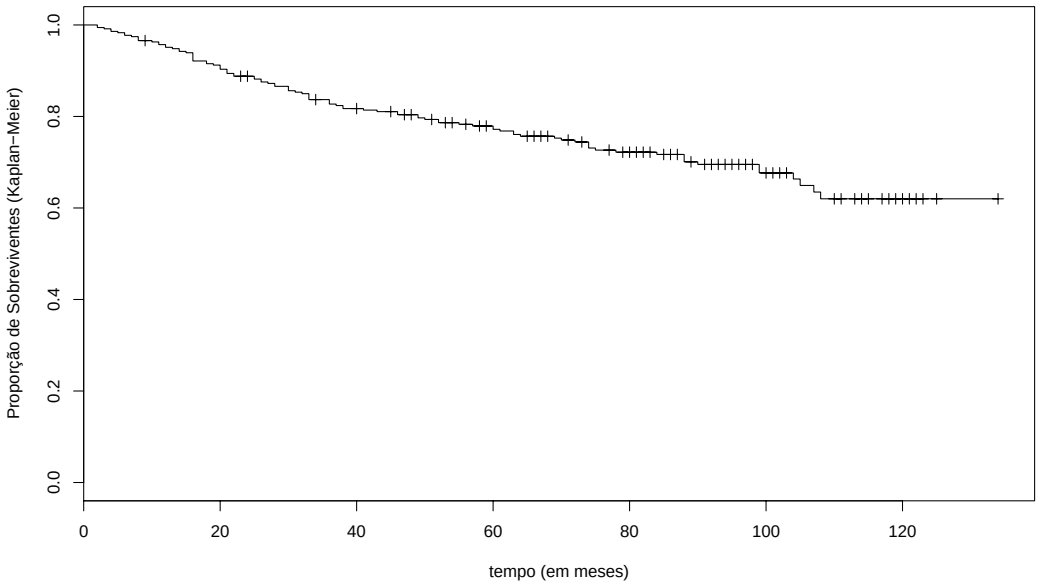

Figura 5.2: Estimativas Kaplan-Meier para os dados de tempo até a recidiva do câncer de mama. Amostra com $n = 355$ pacientes.

$x_{43} = 1$: QRH (quimio-radio-hormonioterapia).
$x_{41} = x_{42} = x_{43} = 0$: nenhum ou outros tipos.

5. TTUM: tamanho do tumor (contínua):

A covariável TTUM (tamanho do tumor) é uma medida feita indiretamente por equipamentos de ressonância magnética. Além disso, o formato de um tumor é irregular e algumas suposições precisam ser feitas para tal medição. Por exemplo, neste conjunto de dados, considerou-se o tamanho do tumor como sendo a dimensão de maior diâmetro. Portanto, supomos que a covariável TTUM é afetada por erros de medição. Em nossa análise, consideramos a metodologia descrita no Capítulo 3 (estrutura aditiva e normalidade do erro), supomos alguns valores para a variância do erro de medida e verificamos o comportamento das estimativas.

5.2.2 Caracterização do Modelo

Consideramos para as variáveis latentes representando os tempos associados aos riscos competitivos (R_i's) o modelo paramétrico Weibull, descrito no Capítulo 1, com funções densidade de probabilidade e sobrevivência dadas por (1.11) e (1.12), e a expressão do logaritmo da função de verossimilhança marginal corrigida dada pela equação (3.12), sendo $\phi = (\beta', \lambda')'$, com $\lambda' = (\rho, \gamma)$ e $\beta' = (\beta_x', \beta_z)'$ um vetor 7×1 de coeficientes de regressão, $\beta_x = (\beta_{EST}, \beta_{NLR}, \beta_{NLC}, \beta_{TNC,QH}, \beta_{TNC,RH}, \beta_{TNC,QRH})'$ e $\beta_z = \beta_{TTUM}$, o coeficiente de regressão da covariável TTUM, supostamente medida com erro.

5.2.3 Resultados

Apresentamos na Tabela 5.3 os resultados das estimativas escore corrigido, variando o valor da variância do erro de medida sobre a covariável TTUM, nos valores $\sigma_u^2 = 0$ (representando, de acordo com a suposição da existência do erro de medida, as estimativas naive), $\sigma_u^2 = 0,01$, $\sigma_u^2 = 0,10$ e $\sigma_u^2 = 0,25$. Utilizamos novamente a função *optim* do software R, associada ao método Quasi-Newton de Broyden, Fletcher, Goldforb e Shanno (BFGS), para obter as estimativas escore corrigido. Para obtenção do erro padrão e p-valor das estimativas, utilizamos os resultados apresentados na Seção 3.3.

Em relação às estimativas de β_{TTUM} (coeficiente de regressão da covariável medida com erro), observamos que à medida que aumentamos a variância do erro de medida, o seu valor é aumentado (0,0912 à 0,0988). Além disso, o erro padrão aumenta mas o P-valor diminui, aumentando o seu nível de significância. As outras covariáves supostas medidas sem erro sofrem pequenas alterações com o aumento da variância do erro de medida. O coeficiente de regressão da covariável EST (β_{EST}) sofre uma redução, enquanto que o da covariável NLC (β_{NLC}) sofre um aumento. Em ambos os casos, observa-se a diferença entre as estimativas naive e escore corrigido em relação às covariáveis medidas sem erro.

σ_u^2	**Parâmetro**	**Estimativa**	**EP**	**P-valor**
0	β_{EST}	$0,5968$	$0,2799$	$0,0407$
	β_{NLR}	$-0,0427$	$0,0152$	$0,0082$
	β_{NLC}	$0,0716$	$0,0242$	$< 0,0001$
	$\beta_{TNC,QH}$	$-1,0590$	$0,6021$	
	$\beta_{TNC,RH}$	$-1,1640$	$0,3926$	$0,0106$
	$\beta_{TNC,QRH}$	$-0,7137$	$0,2533$	
	β_{TTUM}	$0,0912$	$0,0547$	$0,1050$
	γ	$-5,8504$	$0,5751$	$0,0352$
	ρ	$1,3562$	$0,0984$	$0,0004$
0,01	β_{EST}	$0,5954$	$0,2803$	$0,0414$
	β_{NLR}	$-0,0427$	$0,0152$	0.0082
	β_{NLC}	$0,0716$	$0,0242$	$< 0,0001$
	$\beta_{TNC,QH}$	$-1,0694$	$0,6061$	
	$\beta_{TNC,RH}$	$-1,1685$	$0,3931$	$0,0102$
	$\beta_{TNC,QRH}$	$-0,7160$	$0,2533$	
	β_{TTUM}	$0,0918$	$0,0549$	$0,1039$
	γ	$-5,8522$	$0,5760$	$0,0353$
	ρ	$1,3569$	$0,0985$	$0,0004$
0,10	β_{EST}	$0,5905$	$0,2822$	$0,0442$
	β_{NLR}	$-0,0430$	$0,0152$	0.0079
	β_{NLC}	$0,0718$	$0,0242$	$< 0,0001$
	$\beta_{TNC,QH}$	$-1,0614$	$0,6021$	
	$\beta_{TNC,RH}$	$-1,1690$	$0,3939$	$0,0105$
	$\beta_{TNC,QRH}$	$-0,7177$	$0,2540$	
	β_{TTUM}	$0,0941$	$0,0563$	$0,1042$
	γ	$-5,8525$	$0,5752$	$0,0352$
	ρ	$1,3562$	$0,0985$	$0,0004$
0,25	β_{EST}	$0,5804$	$0,2861$	$0,0503$
	β_{NLR}	$-0,0433$	$0,0152$	0.0041
	β_{NLC}	$0,0721$	$0,0242$	$< 0,0001$
	$\beta_{TNC,QH}$	$-1,0653$	$0,6022$	
	$\beta_{TNC,RH}$	$-1,1769$	$0,3960$	$0,0103$
	$\beta_{TNC,QRH}$	$-0,7243$	$0,2553$	
	β_{TTUM}	$0,0988$	$0,0589$	$0,1030$
	γ	$-5,8559$	$0,5756$	$0,0352$
	ρ	$1,3563$	$0,0986$	$0,0004$

Tabela 5.3 - Estimativas escore corrigido. Modelo Weibull, dados de câncer de mama, $n = 355$ pacientes.

Com base nas estimativas sem erro de medida ($\sigma_u^2 = 0$) dadas na Tabela 5.3 e na relação (2.2), a fração de cura fica relacionada às covariáveis através da equação

$$\begin{aligned}\pi(\mathbf{x}) &= exp\Big(- exp(0,5968x_{EST} - 0,0427x_{NLR} + 0,0716x_{NLC} - \\ & \quad 1,0590x_{TNC,QH} - 1,1640x_{TNC,RH} - 0,7137x_{TNC,QRH} + \\ & \quad 0,0912x_{TTUM})\Big). \end{aligned} \tag{5.2}$$

A fração de cura para a amostra estudada possui a distribuição de frequência sumarizada na Tabela 5.4. e representada na Figura 5.3.

Tabela 5.4 - Sumário de estatísticas para π.
Dados de câncer de mama.

Min.	1º Qu	Mediana	Média	3º Qu	Max.
0,0048	0,5120	0,6806	0,6099	0,7731	0,9173

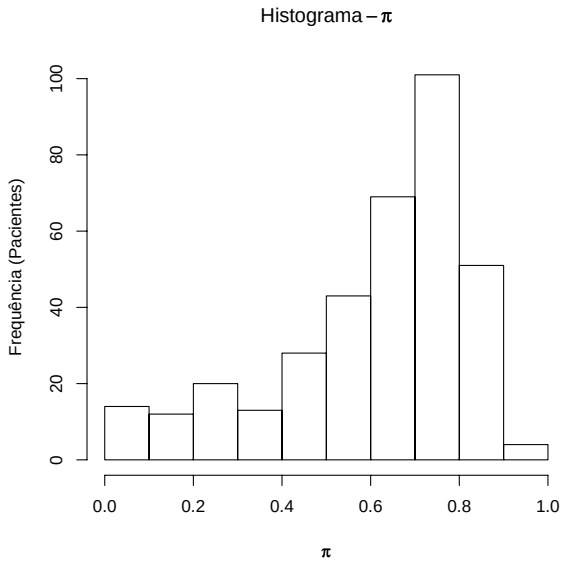

Figura 5.3: Histograma dos valores da fração de cura (π) dos perfis presentes na amostra.

Assim como no Caso 1, podemos traçar perfis e observar os resultados fornecidos pelo modelo. Como exemplo, imaginemos os seguintes:

- **Perfil 1:** pacientes com 32 linfonodos ressecados, sendo que nenhum deles se apresentou comprometido (metástase), com estadiamento 0, 1, IIA ou IIB, tendo realizado o tratamento não-cirúrgico de radio-hormonioterapia, e com tumor medindo 0,9 cm. Em termos de covariáveis, esse perfil é representado pelo vetor $\mathbf{x} = (0, 32, 0, 0, 1, 0, 0.9)$. Para este perfil, o modelo fornece, através da relação (5.2), $\pi = 0,9173$.

- **Perfil 2:** pacientes com 19 linfonodos ressecados, sendo que nenhum deles se apresentou comprometido, com estadiamento IIIA ou IIIB, tendo realizado o tratamento não-cirúrgico de quimio-hormonioterapia, e com tumor medindo 3,5 cm. Temos $\mathbf{x} = (1, 19, 0, 1, 0, 0, 3.5)$, $\pi = 0,6806$.

- **Perfil 3:** pacientes com 38 linfonodos ressecados, sendo que 32 deles se apresentaram comprometidos, com estadiamento IIIA ou IIIB, não tendo realizado qualquer tipo de tratamento não-cirúrgico, e com tumor medindo 4,5 cm. Temos $\mathbf{x} = (1, 38, 32, 0, 0, 0, 4.5)$, $\pi = 0,0048$.

Assim como no Caso 1, o modelo mostrou uma adequação satisfatória ao conjunto de dados e diversas conclusões podem ser tiradas a respeito da influência das covariáveis consideradas no tempo de remissão.

Capítulo 6

Considerações Finais

6.1 Conclusões

Estudamos nesse trabalho um modelo de sobrevivência que suporta casos em que os dados de sobrevivência possuem uma fração de curados, ou seja, indivíduos que nunca apresentarão o evento de interesse. Além disso, avaliamos através de simulação, situações especiais para o caso em que covariáveis associadas ao modelo de sobrevivência com fração de cura sejam medidas com erro, utilizando o método do escore corrigido que fornece estimadores consistentes.

No estudo de simulação, analisamos o comportamento dos estimadores naive e escore corrigido quando uma das covariáveis é medida com erro. Em relação ao coeficiente de regressão da covariável medida com erro, verificamos uma atenuação do estimador naive e valores do estimador escore corrigido mais próximo do verdadeiro. Quanto às covariáveis medidas sem erro, verificamos a influência que os seus estimadores de regressão (naive e escore corrigido) sofrem quando da presença de uma outra covariável medida com erro, mostrando através de gráficos os possíveis vícios do estimador escore corrigido em relação ao verdadeiro valor da estimativa para amostras de tamanho moderado ($n = 50$ e $n = 100$). Variamos o tamanho da amostra e a variância do erro de medida, e verificamos a existência de diferença entre

as estimativas naive e escore corrigido, diferença esta que diminui com o aumento do tamanho da amostra. Verificamos também através de simulação a adequação do modelo semi-paramétrico exponencial por partes ao modelo de sobrevivência com fração de cura e erro de medida, onde as propriedades do estimador escore corrigido para amostras finitas foram mantidas.

Realizamos aplicações do modelo em dois conjuntos de dados reais, ambos caracterizados por uma grande proporção de censura ao final do tempo de observação que, em conjunto com a natureza dos fenômenos envolvidos (tempo de conclusão de um curso de graduação e tempo de remissão de um câncer), indicam a presença de indivíduos imunes ao evento de interesse.

Na primeira aplicação, consideramos dados de evasão escolar de um curso de graduação, sem erro de medida, e testamos a adequação do modelo de sobrevivência com fração de cura na estimação da taxa de evasão do curso.

Na segunda, utilizamos um conjunto de dados de um estudo clínico sobre o câncer de mama, onde o evento de interesse é o tempo entre a remissão e a recidiva do tumor. Consideramos que uma das covariáveis (tamanho do tumor) é medida com erro, e aplicamos o modelo de sobrevivência com fração de cura e erro de medida formulado nas seções anteriores, testando-o para diferentes valores de variância do erro de medida e analisando as estimativas obtidas.

6.2 Pesquisas Futuras

Propomos possíveis pesquisas futuras que podem ser desenvolvidas com base neste trabalho e nas suas referências.

1. Ao utilizarmos o MEP, devemos sempre definir a partição do eixo do tempo. Tal partição em geral é feita de maneira arbitrária, de forma que, em situações práticas, diferentes partições proporcionam resultados distintos. Demarqui (2006) propõe uma abordagem Bayesiana para o problema dessa escolha, sendo os valores da partição parâmetros a

serem estimados. Então uma proposta de trabalho é usar a proposta de Demarqui (2006) no modelo de sobrevivência com fração de cura e erro de medida, utilizando o MEP como distribuição da variáveis latentes (R_i's).

2. Considerar a situação onde existem covariáveis omissas no conjunto de dados. Estudar como podemos lidar com esse tipo de problema, incorporando as covariáveis que possuem omissões no modelo com fração de cura e erro de medida.

3. Utilizar outros métodos, além do escore corrigido, para obtenção de estimadores consistentes em conjunto de dados com erro de medida, utilizando o modelo com fração de cura e erro de medida. Comparar os diferentes métodos em relação ao modelo.

Apêndice A

Função de Verossimilhança

Considerando o conjunto de dados completo dado por $D_c = (n, \mathbf{y}, \nu, \mathbf{n}, X)$, desejamos obter a função de verossimilhança (2.3), de acordo com as suposições feitas na seção 2.3. Como n e X são dados a serem fornecidos pelo problema, iremos calcular a densidade conjunta dos vetores $\mathbf{y} = (y_1, \ldots, y_n)$, $\nu = (\nu_1, \ldots, \nu_n)$ e $\mathbf{n} = (n_1, \ldots, n_n)$. Ressaltamos a diferença entre n (tamanho da amostra) e $\mathbf{n}$ (vetor com os valores assumidos pela variável aleatória $N_i \sim Poisson(\theta)$, $i = 1, \ldots, n$, representando o número de variáveis latentes para cada indivíduo da amostra). Assim,

$$\begin{aligned} f(\mathbf{y}, \nu, \mathbf{n}) &= \prod_{i=1}^{n} f(y_i, \nu_i, n_i) \\ &= \prod_{i=1}^{n} f(y_i, \nu_i | n_i) f(n_i). \end{aligned} \tag{A.1}$$

Lembrando que $N_i \sim Poisson(\theta)$, temos

$$f(n_i) = P(N_i = n_i) = \frac{\theta^{n_i} e^{-\theta}}{n_i!}. \tag{A.2}$$

Temos que $y_i = min\{T_i, C_i\}$, com $T_i = min\{R_{i0}, R_{i1}, \ldots, R_{iN_i}\}$.

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

Então,

$$\begin{aligned} S_T(t|n_i) = P(T_i \geq t|N_i = n_i) &= P(min\{R_{i0}, R_{i1}, \ldots, R_{iN_i}\} \geq t) \\ &= P(R_{i0} \geq t)P(R_{i1} \geq t)\ldots P(R_{in_i} \geq t) \\ &= 1S(t|\lambda)\ldots S(t|\lambda) \\ &= S(t|\lambda)^{n_i}. \end{aligned} \tag{A.3}$$

Assim, utilizando a relação (1.4), temos

$$f_T(t|n_i) = -\frac{dS_T(t|n_i)}{dt} = n_i f(t|\lambda)S(t|\lambda)^{n_i-1}. \tag{A.4}$$

Podemos particionar a expressão $f(y_i, \nu_i|N_i = n_i)$ em dois casos disjuntos: $\nu_i = 0$ ou $\nu_i = 1$. Desta forma, e utilizando as relações (A.3) e (A.4), temos

$$\begin{aligned} P(y_i = t, \nu_i = 0|N_i = n_i) &= P(C_i = t, T_i > C_i|N_i = n_i) \\ &= P(T_i > C_i|C_i = t, N_i = n_i)P(C_i = t) \\ &= S_T(t|n_i)g(t) \\ &= S(t|\lambda)^{n_i}g(t) \end{aligned} \tag{A.5}$$

e

$$\begin{aligned} P(y_i = t, \nu_i = 1|N_i = n_i) &= P(T_i = t, T_i \leq C_i|N_i = n_i) \\ &= P(T_i \leq C_i|T_i = t, N_i = n_i)P(T_i = t|N_i = n_i) \\ &= G(t)f_T(t|n_i) \\ &= G(t)n_i f(t|\lambda)S(t|\lambda)^{n_i-1} \end{aligned} \tag{A.6}$$

Portanto, de (A.5) e (A.6), a distribuição de (y_i, ν_i) dado $N_i = n_i$, $i = 1, \ldots, n$, sob a suposição de censura não-informativa é

$$f(y_i, \nu_i|n_i) \; \alpha \begin{cases} S(y_i|\lambda)^{n_i} & \text{se } \nu_i = 0, \\ n_i f(y_i|\lambda)S(y_i|\lambda)^{n_i-1} & \text{se } \nu_i = 1. \end{cases} \tag{A.7}$$

A equação (A.7) pode ser reescrita de forma mais sintética, e com o abuso de notação de trocarmos o sinal de proporcionalidade pelo sinal de igualdade, da forma

$$\begin{aligned} f(y_i, \nu_i | n_i) &= [S(y_i|\lambda)^{n_i}]^{1-\nu_i}[n_i f(t|\lambda) S(y_i|\lambda)^{n_i-1}]^{\nu_i} \\ &= S(y_i|\lambda)^{n_i-\nu_i}[n_i f(y_i|\lambda)]^{\nu_i} \end{aligned} \tag{A.8}$$

Substituindo (A.2) e (A.8) em (A.1), temos

$$\begin{aligned} f(\mathbf{y}, \nu, \mathbf{n}) &= \prod_{i=1}^{n} S(y_i|\lambda)^{n_i-\nu_i}[n_i f(y_i|\lambda)]^{\nu_i} \frac{\theta^{n_i} e^{-\theta}}{n_i!} \\ &= \prod_{i=1}^{n} S(y_i|\lambda)^{n_i-\nu_i}[n_i f(y_i|\lambda)]^{\nu_i} exp\Big\{ \sum_{i=1}^{n} [n_i \ln\theta - \ln n_i! - \theta] \Big\}. \end{aligned}$$

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

$$L(\theta, \lambda; D_c) = \prod_{i=1}^{n} S(y_i|\lambda)^{N_i-\nu_i}[N_i f(y_i|\lambda)]^{\nu_i} exp\Big\{ \sum_{i=1}^{n} [N_i \ln\theta - \ln(N_i!) - \theta] \Big\}.$$

Introduzindo covariáveis no modelo através do parâmetro θ através da relação $\theta \equiv \theta(x_i'\beta) = exp(x_i'\beta)$, temos que

$$\begin{aligned} L(\beta, \lambda; D_c) &= \Big\{ \prod_{i=1}^{n} S(y_i|\lambda)^{N_i-\nu_i}[N_i f(y_i|\lambda)]^{\nu_i} \Big\} \times \\ &\quad exp\Big\{ \sum_{i=1}^{n} [N_i x_i'\beta - \ln(N_i!) - exp(x_i'\beta)] \Big\}, \end{aligned}$$

que é a expressão da função de verossimilhança (2.3).

Apêndice B

Logaritmo da Função de Verossimilhança Marginal

Vamos mostrar a obtenção do logaritmo da função de verossimilhança marginal (2.5).

Do logaritmo da função de verossimilhança dos dados completos (2.4), e lembrando a relação $\theta_i = exp(x_i'\beta)$, temos para cada indivíduo i, $i = 1, \ldots, n$, a seguinte parcela sendo somada:

$$\begin{aligned} l_i(\theta, \lambda; D_c) &= (N_i - \nu_i) \ln S(y_i|\lambda) + \nu_i \ln N_i + \nu_i \ln f(y_i|\lambda) + N_i \ln \theta_i - \\ &\quad \ln(N_i!) - \theta_i \\ &= \ln S(y_i|\lambda)^{N_i} - \nu_i \ln S(y_i|\lambda) + \ln N_i^{\nu_i} + \nu_i \ln f(y_i|\lambda) + \\ &\quad \ln \theta_i^{N_i} - \ln(N_i!) - \theta_i \\ &= -\nu_i \ln S(y_i|\lambda) + \nu_i \ln f(y_i|\lambda) - \theta_i + \ln \left\{ \frac{[S(y_i|\lambda)\theta_i]^{N_i} N_i^{\nu_i}}{N_i!} \right\} \\ &= K_N + \ln \left\{ \frac{[S(y_i|\lambda)\theta_i]^{N_i} N_i^{\nu_i}}{N_i!} \right\} \end{aligned} \tag{B.1}$$

sendo $K_N = -\nu_i \ln S(y_i|\lambda) + \nu_i \ln f(y_i|\lambda) - \theta_i$, uma constante em relação a N_i. Aplicando exponencial a (B.1), obtemos os fatores multiplicados para para cada indivíduo i, $i = 1, \ldots, n$, da função de verossimilhança:

$$L_i(\theta, \lambda; D_c) = \exp(K_N) \frac{[S(y_i|\lambda)\theta_i]^{N_i} N_i^{\nu_i}}{N_i!} \tag{B.2}$$

Para obter a função de verossimilhança marginal, vamos somar (B.2) em todos os valores que N_i pode assumir.

$$\begin{aligned}
L_i(\theta,\lambda;D) &= \sum_{N_i=0}^{\infty}\exp(K_N)\frac{[S(y_i|\lambda)\theta_i]^{N_i}N_i^{\nu_i}}{N_i!} \\
&= \exp(K_N)\sum_{N_i=0}^{\infty}\left\{\nu_i S(y_i|\lambda)\theta_i\frac{[S(y_i|\lambda)\theta_i]^{N_i-1}}{(N_i-1)!}+(1-\nu_i)\frac{[S(y_i|\lambda)\theta_i]^{N_i}}{(N_i)!}\right\} \\
&= \exp(K_N)\Bigg\{\nu_i S(y_i|\lambda)\theta_i\sum_{N_i=1}^{\infty}\left\{\frac{[S(y_i|\lambda)\theta_i]^{N_i-1}}{(N_i-1)!}\right\}+(1-\nu_i)\times \\
&\quad \sum_{N_i=0}^{\infty}\left\{\frac{[S(y_i|\lambda)\theta_i]^{N_i}}{(N_i)!}\right\}\Bigg\} \\
&= \exp(K_N)\Big[\nu_i S(y_i|\lambda)\theta_i\exp\Big(S(y_i|\lambda)\theta_i\Big)+(1-\nu_i)\exp\Big(S(y_i|\lambda)\theta_i\Big)\Big] \\
&= \exp(K_N)\exp\Big(S(y_i|\lambda)\theta_i\Big)\Big[\nu_i S(y_i|\lambda)\theta_i+1-\nu_i\Big]
\end{aligned} \tag{B.3}$$

Aplicando o logaritmo à (B.3) e substituindo a expressão de K_N, obtemos a parcela do logaritmo da função de verossimilhança marginal

$$\begin{aligned}
l_i(\theta,\lambda;D) &= \ln\left\{\exp(K_N)\exp\Big(S(y_i|\lambda)\theta_i\Big)\Big[\nu_i S(y_i|\lambda)\theta_i+1-\nu_i\Big]\right\} \\
&= K_N+S(y_i|\lambda)\theta_i+\ln\Big[\nu_i S(y_i|\lambda)\theta_i+1-\nu_i\Big] \\
&= -\nu_i\ln S(y_i|\lambda)+\nu_i\ln f(y_i|\lambda)-\theta_i+S(y_i|\lambda)\theta_i+\nu_i\ln\Big(S(y_i|\lambda)\theta_i\Big) \\
&= -\nu_i\ln S(y_i|\lambda)+\nu_i\ln f(y_i|\lambda)-\theta_i+S(y_i|\lambda)\theta_i+\nu_i\ln(S(y_i|\lambda)+\nu_i\ln\theta_i \\
&= \nu_i\ln\theta_i+\nu_i\ln f(y_i|\lambda)-\theta_i\Big[1-S(y_i|\lambda)\Big]
\end{aligned} \tag{B.4}$$

O logaritmo da função de verossimilhança marginal (2.5) é então obtido somando (B.4) em i, $i=1,\ldots,n$, e substituindo a relação $\theta_i = exp(x_i^{'}\beta)$,

$$l(\beta,\lambda;D) = \sum_{i=1}^{n}\Big\{\nu_i x_i^{'}\beta+\nu_i\ln f(y_i|\lambda)-exp(x_i^{'}\beta)[1-S(y_i|\lambda)]\Big\}.$$

Apêndice C

Algoritmos em R

C.1 Gerando os Dados Simulados

```
x <- rnorm(n,0,1)
z <- rnorm(n,0,1)
u <- rnorm(n,0,sqrt(sigmau))
w <- z + u
teta <- exp(0.3*x + 0.3*z)
N <- numeric(n)
for(i in 1:n){N[i] <- rpois(1,teta[i])}
v <- rep(1,n) #indicador de falha;
vi <- rep(0,n) #Indicador de imunidade.
C <- rexp(n, exp(mi)) #Gera censuras aleatórias.
y <- numeric(n)
for(i in 1:n){
N[i] <- rpois(1,teta[i]) #número de variáveis latentes.
if (N[i] > 0){
y[i] <- min(rexp(N[i],exp(-1.5)),C[i]) #Observações (não imunes).
if(y[i] == C[i]) v[i] <- 0
}
else{
y[i] <- 10^4  #Tempo de grande ordem de grandeza para os imunes.
vi[i] <- 1
v[i] <- 0
}}
```

Obs.: n é o tamanho da amostra, *sigmau* a variância do erro de medida e *mi* obtido por simulação de acordo com a proporção de censura desejada para o conjunto de dados.

Censura	0%	15%	25%	50%
mi	< -10	-2.307	-1.533	0.335

C.2 Simulação com Modelo Exponencial

```
#Funções de Bz, Gama e Bx.

Fec <- function(p){ #Verossimilhança corrigida.
Bz <- p[1]
Gama <- p[2]
Bx <- p[3]
for(i in 1:n){
L[i] <- v[i]*(x[i]*Bx + w[i]*Bz + Gama - y[i]*exp(Gama))- exp(x[i]*Bx +
w[i]*Bz-(Bz^2)*(sigmau)/2)*(1-exp(-y[i]*exp(Gama)))
}
LL <- sum(L)
return(-LL)
}

Fnaive <- function(p){ #Verossimilhança naive.
Bz <- p[1]
Gama <- p[2]
Bx <- p[3]
for(i in 1:n){
L[i] <- v[i]*(x[i]*Bx + w[i]*Bz + Gama - y[i]*exp(Gama))- exp(x[i]*Bx +
w[i]*Bz)*(1-exp(-y[i]*exp(Gama)))
}
LL <- sum(L)
return(-LL)
}

L <- numeric(n)
p <- c(0.3,-1.5, 0.3)
optim(p,Fec,method='BFGS')
optim(p,Fnaive,method='BFGS')
```

Obs.: A função *optim* por default minimiza as funções, por isso retornamos $-LL$ no algoritmo.

C.3 Simulação com MEP

```
Nint = 4
yni <- numeric() #Vetor com observações dos não imunes.
u <- 1
for(i in 1:length(y)){
if(y[i]!= max(y)){
yni[u] <- y[i]
u <- u+1}}
yni <- sort(yni)

comp <- 1/Nint
lim <- seq(comp,1,comp)
s <- numeric()
for(i in 1:Nint-1){
s[i] <- -log(1-lim[i])/exp(-1.5) #Valores dos quartis para a partição.
}
s[Nint] <- max(yni)

pos <- numeric() #Vetor indicando o subintervalo de cada observações.
for(i in 1:length(y)){
j <- 1
while(y[i] > s[j] & j != Nint + 1){
j <- j+1}
pos[i] <- j
}

Fmep <- function(p){ #Verossimilhança naive.
lamb1 <- p[1]
lamb2 <- p[2]
lamb3 <- p[3]
lamb4 <- p[4]
Bz <- p[5]
S <- function(i){
if(pos[i] == 1) Sob <- exp(-y[i]*exp(lamb1))
if(pos[i] == 2) Sob <- exp(-exp(lamb2)*(y[i]-s[1]) - exp(lamb1)*s[1])
if(pos[i] == 3) {Sob <- exp(-exp(lamb3)*(y[i]-s[2]) - exp(lamb1)*s[1] -
exp(lamb2)*(s[2]-s[1]))}
if(pos[i] == 4) {Sob <- exp(-exp(lamb4)*(y[i]-s[3]) - exp(lamb1)*s[1] -
exp(lamb2)*(s[2]-s[1])- exp(lamb3)*(s[3]-s[2]))}
if(pos[i] == 5) Sob <- 10^-10 #indivíduo imune.
return(Sob)}
for(i in 1:n){
```

```
if(pos[i] == 1) L[i]<-v[i]*(w[i]*Bz+log(exp(lamb1)*S(i)))-exp(w[i]*Bz)*(1-S(i))
if(pos[i] == 2) L[i]<-v[i]*(w[i]*Bz+log(exp(lamb2)*S(i)))-exp(w[i]*Bz)*(1-S(i))
if(pos[i] == 3) L[i]<-v[i]*(w[i]*Bz+log(exp(lamb3)*S(i)))-exp(w[i]*Bz)*(1-S(i))
if(pos[i] == 4) L[i]<-v[i]*(w[i]*Bz+log(exp(lamb4)*S(i)))-exp(w[i]*Bz)*(1-S(i))
if(pos[i] == 5) L[i]<- -exp(w[i]*Bz)*(1-S(i))}
LL <- sum(L)
return(-LL)}

FmepEC <- function(p){ #Verossimilhança corrigida.
lamb1 <- p[1]
lamb2 <- p[2]
lamb3 <- p[3]
lamb4 <- p[4]
Bz <- p[5]
S <- function(i){
if(pos[i] == 1) Sob <- exp(-y[i]*exp(lamb1))
if(pos[i] == 2) Sob <- exp(-exp(lamb2)*(y[i]-s[1]) - exp(lamb1)*s[1])
if(pos[i] == 3) {Sob <- exp(-exp(lamb3)*(y[i]-s[2]) - exp(lamb1)*s[1] -
exp(lamb2)*(s[2]-s[1]))}
if(pos[i] == 4) {Sob <- exp(-exp(lamb4)*(y[i]-s[3]) - exp(lamb1)*s[1] -
exp(lamb2)*(s[2]-s[1])- exp(lamb3)*(s[3]-s[2]))}
if(pos[i] == 5) Sob <- 10^-10
return(Sob)}
for(i in 1:n){
if(pos[i] == 1) {L[i] <- v[i]*(w[i]*Bz + log(exp(lamb1)*S(i))) - exp(w[i]*Bz -
(Bz^2)*sigmau/2)*(1-S(i))}
if(pos[i] == 2) {L[i] <- v[i]*(w[i]*Bz + log(exp(lamb2)*S(i))) - exp(w[i]*Bz -
(Bz^2)*sigmau/2)*(1-S(i))}
if(pos[i] == 3) {L[i] <- v[i]*(w[i]*Bz + log(exp(lamb3)*S(i))) - exp(w[i]*Bz -
(Bz^2)*sigmau/2)*(1-S(i))}
if(pos[i] == 4) {L[i] <- v[i]*(w[i]*Bz + log(exp(lamb4)*S(i))) - exp(w[i]*Bz -
(Bz^2)*sigmau/2)*(1-S(i))}
if(pos[i] == 5) L[i] <- - exp(w[i]*Bz - (Bz^2)*sigmau/2)*(1-S(i))}
LL <- sum(L)
return(-LL)}

L <- numeric(n)
p <- c(rep(-1.5,Nint), 0.3)
optim(p,Fmep)
optim(p,FmepEC)
```

Obs.: Nint é o número de subintervalos de acordo com a partição escolhida. Neste trabalho, utilizamos $Nint = 4$ intervalos.

Referências Bibliográficas

[1] Berkson, J. & Gage, R. (1952). Survival curve for cancer patients following treatment. *Journal of American Statistical and Probability Letters* **29**: 271-278.

[2] Boag, J. W. (1949). Maximum likelihood estimations of the proportion of patients cured by cancer therapy. J. *Roy. Statist. Soc.* **B11**: 15-53.

[3] Bolfarine, H. & Sandoval, M. C. (2001). Introdução a Inferência Estatística. Coleção Matemática Aplicada. Sociedade Brasileira de Matemática.

[4] Carrol, R.J., Ruppert, D. & Stefanski, L. A. (1995). Measurement Error in Nonlinear Models. *Monographs on Statistics and Applied Probability* **63**.

[5] Chen, M.-H., Harrington, D. & Ibrahim, J. (2002). Bayesian cure rate models for malignant melanoma: A case study of ECOG Trial E1690. *Apllied Statistics* **51**: 135-150.

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

[7] Chen, M.-H., Ibrahim, J. & Lipsitz, S. (2002). Bayesian methods for missing covariates in cure rate models. *Lifetime Data Analysis* **8**: 117-146.

[8] Chen, M.-H., Ibrahim, J. & Sinha, D. (1999). A new bayesian model for survival data with a surving fraction. *Journal of American Statistical Association* **94**: 909-919.

[9] Chen, M.-H., Ibrahim, J. & Sinha, D. (2002). Bayesian inference for multivariate survival data with a surving fraction. *Journal of Multivariate Analysis* **80**: 101-126.

[10] Colosimo, E. A. & Giolo, S. R. (2006). Análise de Sobrevivência Aplicada. ABE - Projeto Fisher. São Paulo: Edgard Blucher.

[11] Demarqui, F. N. (2006). Modelo Exponencial por Partes via Modelo Partição Produto. *Dissertação de Mestrado - Departamento de Estatística - Instituto de Ciências Exatas - Universidade Federal de Minas Gerais.*

[12] Farewell, V. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**.

[13] Farewell, V. (1986). Mixture Models in Survival Analysis: 'Are They Worth the Risk?'. *Statistical Society of Canada*, **14**, 257-262.

[14] Freire, M. P. da S. & Valença, D. M. (2005). Modelo de Sobrevivência para Estudar o Tempo até a Conclusão de um Curso de Graduação. *CD da 38ª Reunião Regional da ABE.*

[15] Friedman, M. (1982). Piecewise Exponential Models for Survival Data with Covariates. *The Annals of Statistics* **10**: 101-113.

[16] Gimenez, P. & Bolfarine, H. (1997). Corrected score functions in classical error-in-variables and incidental parameter models. *Australian Journal of Statistics* **39**: 325-344.

[17] Gimenez, P., Bolfarine, H. & Colosimo, E. (2000). Hypotheses testing for error-in-variables models. *Annals of the Institute of Statistical Mathematics* **52**: 698-711.

[18] Goldman, A. (1984). Survivorship analysis when cure is a possibility: a Monte Carlo study. *Statistics in Medicine* **3**: 153-163.

[19] Gray, R. J. & Tsiatis, A. A. (1989). A Linear Rank Test for Use When the Main Interest Is in Differences in Cure Rates. *Biometrics* **45**: 899-904.

[20] Greenhouse, J. & Wolfe, R. (1984). A competing risks derivation of a mixture model for the analysis of survival data. *Communications in Statistics - Theory Meth.* **13**: 3133-3154.

[21] Halpern, J. & Brown, B. (1987). Cure rate models: Power of the log-rank and generalized Wilcoxon tests. *Statistics in Medicine* **6**: 483-489.

[22] Kim, J. S. & Proschan, F. (1991). Piecewise Exponential Estimation of the Survival Function. *IEEE Translations on Reliability* **40**: 134-139.

[23] Kuk, A. C. & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**: 531-541.

[24] Laska, E. M. & Meisner, M. J. (1992). Nonparametric Estimation and Testing in a Cure Model. *Biometrics* **48**: 1223-1234.

[25] Lawless, J. (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley and Sons, New York.

[26] Lindsey, J. C. & Ryan, L. M. (1993). A Three-state Multiplicative Model for Rodent Tumorigenicity Experiments. *Applied Statistics* **42**, 283-300.

[27] Macedo, C. P. C. de & Valença, D. M. (2005). Análise de Sobreviência Envolvendo Pacientes com Câncer de Mama. *Revista Brasileira de Estatística* (a ser publicado).

[28] Magalhães, M. N. (2006). Probabilidade e Variáveis Aleatórias. 2ª ed. São Paulo: EDUSP.

[29] Maller, R. & Zhou, X. (1996). Survival Analysis with Long-Term Survivors, John Wiley and Sons, New York.

[30] Mizoi, M., Bolfarine, H. & Lima, A. C. P. (2007). Cure rate model with Measurement Error. *Communications in Statistics - Simulation and Computation*, **36**: 185 - 196.

[31] Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77**: 127-137.

[32] R Development Core Team (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[33] Sposto, R., Sather, H. & Baker, S. (1992). A comparison os tests of the difference in the proportion of patients who are cured. *Biometrics* **48**: 87-99.

[34] Stefanski, L. (1985). The effects of measurement error on parameter estimation. *Biometrika* **72**: 583-592.

[35] Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics* **54**: 138-146.

[36] Yakovlev, A., Asselain, B., Bardou, V., Fourquet, A., Hoang. T., Rochefediere, A. & Tsodikov, A. (1993). A simple stochastic model of

tumor recurrence and its applications to data on premenopausal breast cancer, *in* B. Asselain, M. Boniface, C. Duby, C. Lopez, J. Masson & J. Tranchefort (eds). *Biometric et Analyse de Donnees Spatio-Temporelles* **12**: 66-82.