Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística
Mestrado em Estatística

Modelo de Cox para Dados com Censura Intervalar

Autor .....: Lívia Menezes da Paz

Orientador : Prof. Dr. Antônio Eduardo Gomes

Belo Horizonte

Estado de Minas Gerais - Brasil

Março de 2005

Lívia Menezes da Paz

Modelo de Cox para Dados com Censura Intervalar

Dissertação apresentada ao Curso de Mestrado em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Título de Mestre em Estatística.

Orientador: Prof. Dr. Antônio Eduardo Gomes
Instituto de Ciências Exatas - ICEX - UFMG

Belo Horizonte

Estado de Minas Gerais - Brasil

Março de 2005

À minha mãe (in memorian), verdadeiro alicerce de minha vida e que continua mais viva do que nunca em minha mente e em meu coração. À meu pai e aos meus irmãos.

"Quando a gente pensa que
sabe todas as respostas...
Vem a vida e muda
todas as perguntas..."

by Emilly Paull

Agradecimentos

Este trabalho é uma construção coletiva. É o resultado de inúmeras relações humanas. Relações que permanecem. Embora algumas não se façam mais presentes, estão aconchegadas no peito. Outras estão vividas, acalentadas pelo carinho, pela amizade, pelo afeto. Por isso, agradeço a todos pelo aprendizado adquirido, pela convivência, pela palavra, pela dúvida, pelo esclarecimento, pelo estímulo.

Ao Profº. Antônio Eduardo Gomes, pela orientação, pela competência e dedicação dirigidas a este trabalho, sem os quais não seria possível a realização desta pesquisa, e também pela amizade, incentivo e atenção demonstrados ao longo desta jornada.

Aos professores do Departamento de Estatística, em especial Profª. Sueli Mingoti, Profº. Enrico Colosimo e Profº. Renato Assunção, pelo incentivo, pelo apoio e por acreditarem na minha capacidade.

Aos professores que participaram da comissão examinadora.

À minha família, pelo apoio incondicional, ainda que distante.

Aos meus amigos queridos de Salvador, que compreenderam a minha ausência e recusa a tantos convites.

As minhas amigas do Mestrado de Estatística, Janaína Giovani Noronha, Tânia Lúcia Hojo e Jacira Miranda, pela compreensão, pelo apoio, pelos momentos de descontração e pelo companheirismo demonstrados em cada etapa de nosso trajeto.

Aos amigos da UFMG, Ana Paula Travassos, Denise Nacle, Juliana Mambrini, Stella Mares, pelo permanente incentivo nesta caminhada.

Aos meus amigos da UFBA, Cristiane Mêrces, Silvia Patrícia Barreto e Gilson Dourado, pela amizade, pelas sugestões, pela paciência e pelos momentos de diversão.

Aos funcionários do Departamento de Estatística, em especial Rogéria Ferreira Figueiredo, pela amizade, pelo auxílio e presteza ao longo desta jornada.

À CAPES, pelo apoio financeiro.

A Deus, por ter me dado as condições físicas e espirituais para que eu chegasse até aqui.

Enfim, a todos os professores, funcionários e alunos do Mestrado em Estatística que, direta ou indiretamente, contribuíram, e muito, para a realização desta dissertação.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

A literatura estatística tem refletido bastante interesse em métodos para análise de dados de tempos de sobrevivência. Entretanto, existem poucos métodos disponíveis implementados computacionalmente para descrever a distribuição de tempos de sobrevivência na presença de covariáveis quando os tempos de falha são dados com censura intervalar.

Dados com censura intervalar surgem quando o tempo até a ocorrência de um evento de interesse somente é conhecido por ter ocorrido dentro de um intervalo de tempo. Tais dados surgem, por exemplo, em experimentos clínicos ou longitudinais em que pacientes podem ser acompanhados somente através de exames periódicos.

Uma metodologia aplicada para avaliar o efeito de covariáveis no risco de falha é o modelo de riscos proporcionais de Cox quando os dados são exatos ou de censura à direita. Entretanto, existem poucos métodos para ajustar este modelo quando os tempos de falha possuem censura intervalar.

Este trabalho tem por objetivo descrever um método para a regressão de Cox com dados de censura intervalar, a fim de comparar os resultados quanto a vício, precisão e tamanho de amostra. O método é baseado na imputação múltipla através de dois algorítmos muito utilizados no contexto de regressão de dados censurados (Pan, 2000b). O desempenho dos algorítmos será examinado no estudo de simulação. Além disso, pretende-se ampliar o trabalho fazendo uma suavização da curva de sobrevivência base estimada na metodologia proposto por Pan.

Palavras Chave: Modelo de riscos proporcionais; Censura Intervalar; PMDA; Imputação Múltipla; Suavização.

Abstract

The interest in the development of survival data analysis techniques has been growing in the past few years. However, very few of them have been implemenented computationally when data are interval censored.

Interval censored data come up when the event of interest is known to have happened in a time interval, as for example, in clinical trials or longitudinal studies where patients are observed only in periodical examinations.

The Cox's proportional hazards model is widely used to evaluate the effects of covariates on the failure risk when the failure time is observed or right censored. However, there are few methods to fit that model under interval censoring.

The goal of this work is to describe a method to fit the proportional hazards model for interval censored data, and compare the results regarding bias, variability and sample size. The method is based on multiple imputation and two algorithms used for regression procedures with censored data were applied (Pan, 2000b). The performance of the algorithms was examined in the simulation study. Also, we smoothed the baseline survival function for the imputation process in the method proposed by Pan.

Key words: proportional hazards model, interval censoring, PMDA, multiple imputation, smoothing.

Capítulo 1

Introdução

A análise de sobrevivência tem grande importância em diversas áreas de pesquisa, tais como medicina, biologia, saúde pública, epidemiologia, engenharia, economia, direito, dentre outras. Embora os métodos estatísticos sejam aplicados em todas essas áreas, o enfoque desse trabalho será voltado para a área médica.

A análise de sobrevivência é o conjunto de técnicas e modelos estatísticos usados na análise do comportamento de variáveis positivas, tais como: tempo decorrido entre o início do tratamento até a morte do paciente, período de remissão de uma doença, tempo até o desenvolvimento de uma doença, ou simplesmente tempo até a morte. A variável resposta é o tempo até a ocorrência de um evento de interesse, denominado tempo de falha.

Uma característica importante dos dados de tempos de falha é a presença de censura, que representa a observação parcial da resposta. Isto refere-se às situações em que alguns indivíduos encontram-se livres do evento por terem, por exemplo, sido retirados mais cedo do estudo ou pelo término do experimento. Enquanto comum na prática, a censura impede o uso de alguns procedimentos estatísticos convencionais. Em consequência, métodos nessa área têm sido desenvolvidos e o interesse pelo assunto, crescido significativamente após a publicação do artigo de Cox (1972).

A distribuição dos tempos de sobrevivência é usualmente caracterizada por três funções: a função densidade, a função de sobrevivência e a função taxa de falha. Para modelar os dados de sobrevivência, basicamente, existem dois tipos de modelos: os paramétricos e os semiparamétricos, sendo o modelo semiparamétrico de maior interesse para esse trabalho.

O modelo de riscos proporcionais de Cox é a metodologia mais aplicada para avaliar o efeito das covariáveis no risco de falha. Este método tem sido extensivamente desenvolvido e aplicado no caso em que tempos de falha são exatos ou de censura à direita. Entretanto, existem poucos métodos implementados computacionalmente desenvolvidos para ajustar um modelo de riscos proporcionais quando os tempos de falha possuem censura intervalar, ou seja, o tempo de falha somente é conhecido por ter ocorrido dentro de algum intervalo de tempo.

1.1 Motivação

A motivação do tema de estudo surgiu pelo fato de existirem poucos métodos disponíveis na literatura estatística para descrever a distribuição de tempos de sobrevivência quando os tempos de falha apresentam censura intervalar. Pan (2000b) propôs um método baseado na imputação múltipla para regressão de Cox com dados de censura intervalar. A idéia básica consiste em imputar tempos de sobrevivência exatos para dados com censura intervalar finita usando dois algoritmos para análise de dados de regressão censurados. Após a imputação inicial, tem-se dados com censura à direita. Pode-se, então, estimar o vetor de coeficientes β e a função de sobrevivência base S_0 pelos procedimentos computacionais usuais e utilizar estas estimativas iterativamente no processo de imputação dos dados.

1.2 Objetivos

O objetivo é aplicar a proposta de análise apresentada por Pan (2000b), a fim de comparar os resultados quanto a vício e variabilidade, para vários tamanhos de amostra, com aqueles por ele encontrados.

Além disso, pretende-se ampliar o trabalho fazendo uma suavização da curva de sobrevivência base estimada na metodologia proposta por Pan (2000b), a fim de eliminar o problema da diminuição progressiva nos pontos de salto a cada iteração, e que pode ser grave para amostras pequenas.

Capítulo 2

Revisão da Literatura

Peto (1973) foi o primeiro a propor um método não paramétrico para estimação da distribuição de sobrevivência sob censura intervalar. Este artigo descreve um algoritmo de Newton-Raphson para a estimativa não paramétrica de máxima verossimilhança da função de distribuição acumulada.

Turnbull (1976) obteve o mesmo estimador que Peto, mas usando uma aproximação diferente na estimação. Ele deriva o mesmo estimador usando um algoritmo iterativo de auto-consistência, o algoritmo EM. Gentleman e Geyer (1994) mostraram que este estimador de auto-consistência não é sempre o estimador de máxima verossimilhança (EMV), e que o EMV não é necessariamente único. Os autores discutem, ainda, condições sob os quais este estimador pode ser determinado.

Lindsey e Ryan (1998) compararam os modelos paramétricos, semiparamétricos e não paramétricos que estão disponíveis em softwares estatísticos padrões ou que podem ser facilmente programados. As autoras concluíram que os modelos paramétricos apresentam melhor desempenho. Kaplan e Meier (1958) desenvolveram métodos não paramétricos para analisar dados com observações incompletas.

Finkelstein e Wolfe (1985) propuseram uma técnica semiparamétrica para análise de regressão para dados com censura intervalar que modela a função de probabilidade do tempo de resposta, T, e uma covariável aleatória, Z, como o produto de

um modelo paramétrico, $P(Z|T)$, e uma distribuição não paramétrica, $P(T)$, para o tempo de resposta. Finkelstein (1986) também considerou ajuste do modelo de riscos proporcionais para dados com censura intervalar utilizando o algoritmo de Newton-Raphson.

Huang e Wellner (1995) estabeleceram a normalidade assintótica do estimador não paramétrico de máxima verossimilhança (ENPMV) do coeficiente de regressão. Kooperberg e Clarkson (1997) modelaram a função de risco base usando splines. Satten *et al.* (1998) propuseram aproximação marginal para imputar os ranks dos tempos de falha censurados.

Honda (2004) propôs um estimador de regressão não paramétrico para dados de status corrente (caso 1 de censura intervalar) fazendo uma modificação no estimador de correlação de rank máximo (MRC), que foi desenvolvido por Han (1987), para modelos lineares. Esta modificação foi feita devido à falta de eficiência do estimador MRC (Sherman, 1993).

Lin *et al.* (1998) estudaram métodos semiparamétricos para analisar dados de status corrente sob o modelo de regressão de riscos aditivos. Este modelo especifica que a função de risco para o tempo de falha associado com um conjunto de covariáveis dependentes do tempo é a soma de uma função de risco base arbitrária e o vetor de covariáveis.

Shiboski (1998) propôs uma aproximação semiparamétrica para estimação em modelos de regressão de dados de status corrente usando técnicas de modelagem aditiva generalizada e de regressão isotônica. Este procedimento fornece estimativas simultâneas da distribuição de tempos de falha e efeitos de covariáveis.

Huang e Rossini (1997) descreveram o comportamento de amostras finitas de um estimador de máxima verossimilhança restrita para o modelo de chances proporcionais com dados de censura intervalar. Rabinowitz *et al.* (2000) propuseram uma aproximação para ajustar o modelo de regressão de chances proporcionais para dados de censura intervalar. A aproximação é baseada em rotinas de regressão

logística condicional em pacotes estatísticos padrão.

Kooperberg *et al.* (1995a) introduziram a regressão de risco (HARE) para estimar a função log-risco condicional baseada em modelos lineares. Kooperberg e Clarkson (1997) também extendem a metodologia HARE para acomodar dados de censura intervalar.

Li e Zhang (1998) desenvolveram M-estimadores para os coeficientes de regressão assim como estimadores assintoticamente eficientes.

Self e Grossman (1986) descreveram a estatística de rank linear para diferença entre grupos quando os dados são de censura intervalar. Esta estatística é aproximadamente relacionada àquela descrita por Prentice (1978) para dados de censura à direita.

Rabinowitz *et al.* (1995) propuseram ajustar o modelo de tempo de falha acelerado para dados de censura intervalar utilizando uma classe de estatística score que pode ser usada para procedimentos de estimação e de intervalos de confiança.

Uma característica do modelo de Cox é que, para o caso de tempos exatos e de censura à direita, o método não fornece especificação ou estimação do risco base. Embora uma forma funcional seja assumida para o efeito de covariáveis, a estimação depende somente do rank ordenado dos tempos de falha exatos e censurados. Para dados de censura intervalar, não é possível identificar o rank exato dos tempos de falha. Assim, a metodologia que tem sido desenvolvida para dados exatos e de censura à direita não pode ser diretamente aplicada ao caso de censura intervalar. Goggins *et al.* (1998) propuseram ajustar o modelo de riscos proporcionais para tempos de falha com censura intervalar usando um algoritmo EM com métodos de Monte Carlo via Cadeia de Markov.

Como os dados de censura intervalar podem ser vistos como dados incompletos, uma aproximação para o problema de regressão é derivar a estimativa de máxima verossimilhança marginal do modelo de riscos proporcionais usando o algoritmo EM (Dempster *et al.*, 1977). Outros autores têm proposto métodos de simulação que

não são baseados no algoritmo EM. Satten (1996) estima o modelo de Cox para dados de censura intervalar usando uma aproximação estocástica baseada no processo Robbins-Munro e uma amostragem de Gibbs para amostras de ranks consistentes com os dados observados.

Frequentemente, dados de censura intervalar são analisados substituindo cada intervalo do tempo de falha com um único valor. Um método muito utilizado é o da imputação do ponto médio, que toma o ponto médio do intervalo de cada indivíduo como sendo o tempo exato. Este método, entretanto, pode conduzir a estimativas viciadas se os intervalos forem grandes e variados (Law e Brookmeyer, 1992). Além disso, ele subestima os erros-padrões dos coeficientes pois trata os tempos de sobrevivência como conhecidos, quando na verdade não são. Outra aproximação comum é usar o limite superior dos intervalos como tempo de falha. Este estimador tem desvantagens similares àquela da imputação do ponto médio.

Goetghebeur e Ryan (2000) desenvolveram uma aproximação semiparamétrica para análise de regressão de riscos proporcionais de dados com censura intervalar. Um algoritmo EM baseado em verossimilhança aproximada que envolve maximizar uma verossimilhança parcial de Cox padrão para a estimativa dos coeficientes de regressão e assim usar o estimador de Breslow para o risco base desconhecido. A imputação múltipla parece produzir estimativas da variância melhores do que os métodos assintóticos para conjuntos de dados de tamanho pequeno a moderado.

Pan (1999) propôs uma extensão do algoritmo iterativo do minorante convexo para ajuste do modelo de Cox para dados com censura intervalar. Ele reformulou o algoritmo iterativo do minorante convexo (ICM), proposto por Groeneboom e Wellner (1992), como um método de projeção do gradiente generalizado (GGP). Pan e Chappell (2002) mostraram que o estimador não paramétrico de máxima verossimilhança (ENPMV) do coeficiente de regressão e a sobrevivência base funciona bem para o modelo de riscos proporcionais de Cox com dados de censura intervalar e dados truncados à esquerda, utilizando duas alternativas: a aproximação

de verossimilhança marginal e a estimativa de máxima verossimilhança monótona.

Taylor *et al.* (2002) descreveram procedimentos de imputação múltipla não paramétrica para tratar tempos de falha faltantes para observações censuradas no contexto de testes e estimação da sobrevivência não paramétrica.

Wei e Tanner (1991) apresentaram dois algoritmos semiparamétricos para a análise de dados de regressão censurada: o algoritmo de dados aumentados do "pobre homem" (PMDA) e o algoritmo de dados aumentados da normal assintótica (ANDA). Estes algoritmos foram motivados pelo algoritmo de dados aumentados encontrado em Tanner e Wong (1987a) para análise de problemas de dados faltantes. Bebchuck e Betensky (2000) apresentaram um algoritmo de dados aumentados para estimar a função risco baseado em dados de censura intervalar. Os autores aplicam métodos de imputação múltipla e de verossimilhança local para obter estimativas não paramétricas suavizadas para a função de risco.

Métodos mais complexos também têm sido propostos para analisar dados de tempos de falha com censura intervalar. Brookmeyer e Goedert (1989) propõem um modelo de regressão de dois estágios para análise de tempos latentes de censura intervalar, mas este depende das suposições paramétricas. Sinhá *et al.* (1994) usaram o algoritmo EM de Monte Carlo (MCEM) para estimar os parâmetros do modelo de Cox com dados de tempos de falha agrupados.

Capítulo 3

Análise de Sobrevivência

Neste capítulo, são apresentadas algumas definições importantes na Análise de Sobrevivência, além de uma descrição do modelo de riscos proporcionais.

3.1 Conceitos Básicos

Os conjuntos de dados de sobrevivência são caracterizados pelos tempos de falha e pelas censuras. Estes dois componentes constituem a resposta. O tempo de falha é determinado pelo tempo inicial de observação, pela escala de medida e pelo evento de interesse, comumente associado à falha.

3.1.1 Censura

Em geral, nos métodos de análise de sobrevivência, existem três mecanismos de censura que são diferenciados em estudos médicos:

• Censura Tipo I: ocorre quando o tempo para a realização de um estudo é pré-estabelecido;

• Censura Tipo II: ocorre quando o tempo de censura é determinado por um número pré-estabelecido de falhas a serem observadas;

• Censura Aleatória: ocorre quando o indivíduo é retirado no decorrer do estudo

sem ter ocorrido a falha, sendo este o tipo de censura que mais ocorre na prática médica.

Os mecanismos apresentados são conhecidos por censura à direita, quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado; por censura à esquerda, quando o evento de interesse já aconteceu quando o indivíduo foi observado. Neste trabalho, o foco principal é o tipo de censura mais geral, que é o da censura intervalar, quando se conhece somente que o evento de interesse ocorreu em um certo intervalo, ou seja, para cada observação, temos censura à direita, ou censura à esquerda, ou ambas. Este tipo de censura surge em experimentos clínicos e estudos longitudinais, e será discutido no próximo capítulo.

Uma representação do mecanismo de censura à direita é feita utilizando-se duas variáveis aleatórias. Seja T uma variável aleatória representando o tempo de falha e seja C uma outra variável aleatória, independente de T, representando o tempo de censura. Os dados são representados por

$$t = min(T, C)$$

e

$$\delta = \begin{cases} 1 & \text{se} \quad T \leq C \\ 0 & \text{se} \quad T > C \end{cases}$$

onde $\delta = 1$ representa uma observação não censurada e $\delta = 0$ representa a presença de censura.

A suposição de independência entre as variáveis T e C simplifica a análise estatística dos dados e é razoável em vários estudos clínicos.

3.1.2 Funções de Sobrevivência

A distribuição dos tempos de sobrevivência é usualmente caracterizada por três funções:

• **A função densidade** é definida em termos da probabilidade de um indivíduo falhar em um curto intervalo de tempo (Δt):

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

Através da função densidade, podemos obter a probabilidade para qualquer intervalo de tempo.

• **A função de sobrevivência** é definida como a probabilidade de um indivíduo sobreviver até o tempo t.

$$S(t) = P(T \geq t) = 1 - F(t) = 1 - \int_0^t f(u)du$$

$S(t)$ é uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência.

• **A função taxa de falha ou de risco** é definida em termos da probabilidade de um indivíduo falhar durante um intervalo $[t, t + \Delta t)$ dado que ele não tenha falhado até o tempo $t > 0$.

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

A função de risco representa a razão de falha instantânea para um indivíduo que sobreviveu até o tempo t.

Essas funções são matematicamente equivalentes, ou seja, se uma delas é dada, as demais podem ser obtidas facilmente. O relacionamento entre as funções bem como as propriedades de cada uma delas podem ser encontradas em diversos textos na área de análise de sobrevivência, como Lawless (1982), Lee (1992), Collet (1994) e Klein e Moeschberger (1997).

3.2 Descrevendo Dados de Sobrevivência

Em análise de sobrevivência, a descrição dos dados é feita através de técnicas não paramétricas que permitem incorporar censuras à análise. Existem diversos métodos para estimar a função de sobrevivência, sendo o estimador produto-limite (ou de Kaplan-Meier) o mais conhecido entre eles (Colosimo, 2001) para o caso em que censuras à direita são observadas.

- **Estimador de Kaplan-Meier**

O estimador de Kaplan-Meier é um estimador não paramétrico da função de sobrevivência. Este estimador é, também, conhecido como estimador produto-limite.

Na sua construção, considera-se um número de intervalos de tempo igual ao número de falhas distintas, sendo os limites dos intervalos os tempos de falha da amostra. Para obter o estimador Kaplan-Meier, os tempos de sobrevivência observados devem estar ordenados de tal forma que $t_1 < t_2 < \ldots < t_n$. O estimador da função de sobrevivência é dado por:

$$\hat{S}(t) = \prod_{i/t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

onde n_i é o número de indivíduos sob risco no tempo t_i, e d_i é o número de falhas observadas no tempo t_i.

Em geral, o estimador de Kaplan-Meier se aplica a qualquer tamanho de amostra, mas quando essa amostra é muito grande, é mais conveniente agrupar os tempos de falha em intervalos e realizar a análise através da Tabela de Vida. Assim, podemos dizer que o Kaplan-Meier é um caso especial da Tabela de Vida, onde cada intervalo contém uma observação distinta (na ausência de empates).

As propriedades desse estimador estão detalhadas em Collet (1994) e em Klein e Moeschberger (1997). Para estimar a função de sobrevivência, existem também os estimadores da Tabela de Vida e de Nelson-Aalen.

• **Estimador da Tabela de Vida**

O estimador da Tabela de Vida é uma das técnicas mais antigas para análise de dados de sobrevivência e é utilizado basicamente em grandes amostras. Este estimador também é conhecido como método atuarial.

A construção de uma tabela de vida consiste em dividir o eixo do tempo em k intervalos $I_i = [t_{i-1}, t_i)$, $i = 1, \ldots, k$ em que $t_0 = 0$ e $t_k = \infty$. O estimador da Tabela de Vida para $S(t)$ é similar ao proposto por Kaplan-Meier com um estimador ligeiramente diferente para $q_i = \frac{d_i}{n_i}$ uma vez que, nesse caso, tem-se:

- d_i=nº de falhas no intervalo $I_i = [t_{i-1}, t_i)$ e
- n_i=[nº de indivíduos em risco em t_{i-1}]-[$\frac{1}{2}$ x (nº de censuras em $I_i = [t_{i-1}, t_i)$)].

O estimador da Tabela de Vida é, portanto, dado por:

$$\hat{S}(t_{I_i}) = \begin{cases} 1 & i = 1 \\ \hat{S}(t_{I_{(i-1)}})(1 - \hat{q}_i) & i = 2, \ldots, k. \end{cases}$$

Para maiores detalhes sobre este estimador, pode-se consultar, por exemplo, Lee (1980).

• **Estimador de Nelson-Aalen**

Este estimador é mais recente do que o de Kaplan-Meier e baseia-se na função de sobrevivência expressa por:

$$S(t) = exp(-H(t))$$

onde $H(t)$ é a função de risco acumulada, isto é $H(t) = \int_0^t h(u)du$. Como $S(t)$ pode ser estimada a partir de $H(t)$, Nelson (1972) propôs e, posteriormente, Aalen (1978) provou as propriedades assintóticas do seguinte estimador para $H(t)$:

$$\tilde{H}(t) = \sum_{i/t_i<t} \left(\frac{d_i}{n_i}\right)$$

em que n_i e d_i são definidos como no estimador de Kaplan-Meier.

Desse modo, o estimador de Nelson-Aalen para a função de sobrevivência $S(t)$ é dada por

$$\tilde{S}(t) = exp(-\tilde{H}(t))$$

Os estimadores de Nelson-Aalen e de Kaplan-Meier apresentam, na maioria das vezes, estimativas muito próximas para $S(t)$. A vantagem do estimador Kaplan-Meier é que é possível encontrá-lo implementado em diversos pacotes estatísticos, o que não acontece com o estimador de Nelson-Aalen.

3.3 Modelo de Regressão de Cox

O modelo de regressão de Cox, proposto por Cox (1972), é um dos mais populares na análise de dados de sobrevivência. Este modelo permite que a análise dos tempos de vida até a ocorrência de um evento seja realizada considerando-se as covariáveis de interesse. Cox (1975), assim como outros autores, propõe modelar dados de sobrevivência, na presença de covariáveis, por meio da função de risco.

Assume-se, nesse modelo, que os tempos t_i, $i = 1, \ldots, n$, são independentes e que o risco do indivíduo i é dado por:

$$h(t|\mathbf{Z}_i) = h_0(t)g(\mathbf{Z}_i'\beta) \tag{3.1}$$

em que $h_0(t)$ é conhecida como a função de risco base, ou seja, o risco de um indivíduo com covariáveis $\mathbf{Z} = \mathbf{0}$, β é o vetor de dimensão p de coeficientes de regressão desconhecidos, e $\mathbf{Z}_i$ é o vetor de dimensão p de covariáveis observadas para o indivíduo i.

O componente não paramétrico $h_0(t)$ não é especificado, e é uma função não-negativa do tempo. Devido à presença deste componente no modelo, o componente paramétrico é geralmente usado na forma multiplicativa na função de risco:

$$g(\mathbf{Z}_i'\beta) = exp(\mathbf{Z}_i'\beta) = exp(\beta_1\mathbf{Z}_1 + \ldots + \beta_p\mathbf{Z}_p) \tag{3.2}$$

Uma das vantagens deste modelo está na comparação entre indivíduos ou grupos distintos. A denominação desse modelo como sendo de riscos proporcionais se deve ao fato que a razão entre as funções de risco de dois indivíduos,

$$\frac{h_i(t|\mathbf{Z}_i)}{h_j(t|\mathbf{Z}_j)} = \frac{h_0(t)exp(\mathbf{Z}_i'\beta)}{h_0(t)exp(\mathbf{Z}_j'\beta)} = \frac{exp(\mathbf{Z}_i'\beta)}{exp(\mathbf{Z}_j'\beta)} = exp(\mathbf{Z}_i'\beta - \mathbf{Z}_j'\beta) \tag{3.3}$$

$(i, j = 1, \ldots, n \quad \text{e} \quad i \neq j)$, não depende do tempo t.

A suposição básica para o uso do modelo de riscos proporcionais de Cox é, portanto, que as taxas de falha sejam proporcionais.

A função taxa de falha base acumulada, bem como a correspondente função de sobrevivência são também de interesse, e estas relacionam-se com a função de risco

base por, respectivamente,

$$H_0(t) = \int_0^t h_0(u)du$$

e

$$S(t|\mathbf{Z}) = exp(-H_0(t)exp(\mathbf{Z}'\beta)) = [S_0(t)]^{exp(\mathbf{Z}'\beta)}$$

onde

$$S_0(t) = exp(-H_0(t)).$$

O principal interesse é estimar β (e eventualmente $h_0(\cdot)$), com base nos dados que possivelmente contêm censuras. Cox (1975) propôs o método de máxima verossimilhança parcial, no qual elimina a função de base $h_0(t)$ e leva em consideração os tempos de sobrevivência censurados. A metodologia usada na estimaçao de máxima verossimilhança parcial será apresentada a seguir.

3.3.1 Estimação dos parâmetros no Modelo de Cox

A função de verossimilhança incorporando censura é expressa da seguinte forma

$$L(\beta) = \prod_{i=1}^{n} [f(t_i, \mathbf{Z}_i, \beta)]^{\delta_i} [S(t_i, \mathbf{Z}_i, \beta)]^{1-\delta_i}$$

para um conjunto de dados observados $(t_i, \mathbf{Z}_i, \delta_i)$, $i = 1, 2, \ldots, n$. Assim, substituindo $f(t_i, \mathbf{Z}_i, \beta)$ pela relação existente entre as funções

$$f(t, \mathbf{Z}, \beta) = h(t, \mathbf{Z}, \beta)S(t, \mathbf{Z}, \beta),$$

teremos

$$L(\beta) = \prod_{i=1}^{n} [h(t_i, \mathbf{Z}_i, \beta)]^{\delta_i} [S(t_i, \mathbf{Z}_i, \beta)].$$

Com este resultado, pode-se construir a função de verossimilhança parcial proposta por Cox.

Para a construção da função de verossimilhança parcial, considere uma amostra composta de n indivíduos em que $k \leq n$ falhas distintas ocorrem nos tempos $t_1 \leq t_2 \leq \ldots \leq t_k$, e que a probabilidade condicional do i-ésimo indivíduo vir a falhar no tempo t_i, $i = 1, 2, \ldots, k$, dado que se conhece quais indivíduos que estão sob risco em t_i, é:

$$\frac{h_i(t_i|\mathbf{Z}_i)}{\sum\limits_{j \in R(t_i)} h_j(t_j|\mathbf{Z}_j)} = \frac{h_0(t)exp(\mathbf{Z}_i'\beta)}{\sum\limits_{j \in R(t_i)} h_0(t)exp(\mathbf{Z}_j'\beta)} = \frac{exp(\mathbf{Z}_i'\beta)}{\sum\limits_{j \in R(t_i)} exp(\mathbf{Z}_j'\beta)} \tag{3.4}$$

em que $R(t_i)$ é o conjunto de todos os indivíduos ainda em risco no tempo t_i. Pode-se observar que o componente não paramétrico $h_0(t)$ desaparece em (3.4).

A função de verossimilhança $L(\beta)$ a ser usada para fazer inferências no modelo é, então, formada pelo produto da expressão (3.4) para todos os tempos de falha, ou seja,

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{exp(\mathbf{Z}_i'\beta)}{\sum\limits_{j \in R(t_i)} exp(\mathbf{Z}_j'\beta)} \right]^{\delta_i} \tag{3.5}$$

em que δ_i é o indicador de falha da i-ésima observação.

Os indivíduos censurados entram na função de verossimilhança parcial $L(\beta)$ através do conjunto de risco $R(t_i)$, e contribuem para esta função somente enquanto permanecem sob risco.

Os estimadores de máxima verossimilhança parcial são encontradas maximizando $L(\beta)$, e são obtidos pela resolução do sistema de equações simultâneas definido por $U(\beta) = 0$, em que $U(\beta)$ é o vetor escore, composto das primeiras derivadas da função $l(\beta) = log(L(\beta))$, isto é,

$$U(\beta) = \frac{\partial log L(\beta)}{\partial \beta} = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum\limits_{j \in R(t_i)} Z_j exp(\mathbf{Z}'_j \beta)}{\sum\limits_{j \in R(t_i)} exp(\mathbf{Z}'_j \beta)} \right]$$

em que $l(\beta)$ é o logaritmo da função de verossimilhança.

Os estimadores de máxima verossimilhança parcial podem ser obtidos por processos iterativos, como por exemplo, o método de Newton-Raphson.

A função de verossimilhança parcial assume que os tempos de sobrevivência são contínuos, e consequentemente, não pressupõe a possibilidade de empates dos valores observados. Quando ocorrem empates entre falhas e censuras, utiliza-se a convenção de que a censura ocorreu após a falha, o que define as observações a serem incluídas no conjunto de risco em cada tempo de falha.

Para incorporar as observações de falhas empatadas, caso estiverem presentes, a função de verossimilhança parcial deve ser modificada. Existem duas aproximações que são encontradas com frequência nos pacotes estatísticos. A verossimilhança proposta por Breslow e Crowley (1974) é dada por

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(s'_i \beta)}{\left[\sum\limits_{j \in R(t_i)} exp(\mathbf{Z}'_j \beta) \right]^{d_i}} \tag{3.6}$$

onde s_i é o vetor dado pela soma das p covariáveis para os indivíduos que falham no tempo t_i, $i = 1, \ldots, k$ e d_i é o número de falhas neste tempo. Esta verossimilhança é bastante simples de calcular, e é uma aproximação adequada quando o número de observações empatadas não é grande. Por estas razões, este método é usualmente implementado em pacotes estatísticos para análise de sobrevivência.

Efron (1977) propôs

$$L(\beta) = \prod_{i=1}^{k} \frac{exp(s'_i \beta)}{\prod_{k=1}^{d_i} \left[\sum\limits_{j \in R(t_i)} exp(\mathbf{Z}'_j \beta) - \frac{k-1}{d_k} \sum\limits_{j \in D(t_i)} exp(\mathbf{Z}'_j \beta) \right]} \tag{3.7}$$

como uma verossimilhança aproximada para o modelo de riscos proporcionais, em que $D(t_i)$ é o conjunto de todos os indivíduos que falham em t_i. Esta é uma melhor aproximação para a função de verossimilhança (3.5) do que aquela proposta por Breslow, embora na prática, ambas as aproximações frequentemente forneçam resultados similares quando o número de empates é pequeno.

3.3.2 Modelo de Regressão de Cox com dados de censura intervalar

Como visto, anteriormente, a função de sobrevivência $S(t|\mathbf{Z})$ para o modelo de riscos proporcionais é dada por

$$S(t|\mathbf{Z}_i) = S_0(t)^{exp(\mathbf{Z}_i'\beta)}$$

onde S_0 é a função de sobrevivência base desconhecida, $\mathbf{Z}_i$ é um vetor de covariáveis e β é o vetor de coeficientes de regressão. Devido à censura, não observa-se o tempo de falha T_i diretamente. Para cada T_i, sabe-se somente que ele é interno ao intervalo $(U_i, V_i]$. A censura à direita é equivalente a termos $V_i = \infty$. A censura à esquerda equivale a termos $U_i = 0$. É usual assumir que T_i é independente do mecanismo de censura. Para diferenciar da censura à direita, refere-se a uma observação com V_i finita como censura intervalar finita.

Então, baseado nas n observações $(U_1, V_1, Z_1), \ldots, (U_n, V_n, Z_n)$, deseja-se estimar o coeficiente de regressão β, seu erro-padrão, e a função de sobrevivência base S_0. Uma alternativa para encontrar-se tais estimativas será utilizando o método da imputação múltipla, detalhado no próximo capítulo.

3.3.3 Estimação da Função de Risco Base

Os coeficientes de regressão β são de grande interesse no modelo de Cox. No entanto, funções relacionadas com $h_0(t)$, tais como as funções $H_0(t)$ e $S_0(t)$ apresentadas na Seção 2.3, são também importantes.

Se $h_0(t)$ fosse especificada parametricamente, seria possível estimá-la por meio da função de verossimilhança. Entretanto, na construção da função de verossimilhança parcial, $h_0(t)$ é eliminada completamente e, desse modo, os estimadores propostos serão de natureza não paramétrica.

Um estimador simples proposto para a função de risco base acumulada $H_0(t)$, referenciado como estimador de Nelson-Aalen-Breslow, ou simplesmente de estimador de Breslow, é uma função escada, com saltos nos distintos tempos de falha, e expresso por:

$$\hat{H}_0(t_i) = \sum_{t_i \leq t} \frac{d_i}{\sum\limits_{j \in R(t_i)} exp(\mathbf{Z}'_j \beta)}$$

em que d_i é o número de ocorrências no tempo t_i, e $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_{\mathbf{p}})$ é o estimador de máxima verossimilhança parcial.

Note que, se não existirem covariáveis, então

$$\hat{H}_0(t_i) = \sum_{t_i \leq t} \left(\frac{d_i}{n_i} \right)$$

isto é, o estimador de Nelson-Aalen visto anteriormente.

Uma vez estimada a função de risco base acumulada, pode-se obter a função de sobrevivência base acumulada, isto é,

$$\hat{S}_0(t) = exp(-\hat{H}_0(t))$$

e a partir desta, obter a função de sobrevivência estimada que, para um indivíduo

com vetor de covariáveis Z_i, é expressa por:

$$\hat{S}(t|\mathbf{Z}_i) = [\hat{S}_0(t)]^{exp(\mathbf{Z}_i'\hat{\beta})} \tag{3.8}$$

e será, também, uma função escada decrescendo com o tempo.

Capítulo 4

Censura Intervalar

4.1 Introdução

Dados censurados aparecem por uma variedade de razões e, para fazer distinção entre os diversos tipos de censura (à direita, à esquerda, intervalar, etc), deve-se considerar a forma que os dados são obtidos. O termo censura intervalar é usado para a descrição de situações em que se sabe que o tempo de sobrevivência de um indivíduo, T_i, ocorre entre dois valores, ou seja, $T_i \in (U_i, V_i]$.

É bastante comum dados de sobrevivência com censura intervalar ocorrerem em ensaios clínicos e estudos longitudinais, como por exemplo, em estudos da Síndrome de Imunodeficiência Adquirida (AIDS) referente a infecção do Vírus de Imunodeficiência Humana (HIV) e o tempo de incubação da AIDS (o tempo de infecção do HIV ao diagnóstico da AIDS). Neste caso, se o indivíduo é HIV positivo no início do estudo, o tempo de infecção do HIV é usualmente determinado por um estudo retrospectivo da história do indivíduo. Assim, somente um intervalo dado pelo último teste HIV negativo e pelo primeiro teste HIV positivo é conhecido para o tempo de infecção do HIV.

Em uma aplicação de análise de sobrevivência padrão, indivíduos são acompa-

nhados com relação à ocorrência de um evento específico. Se o evento observado ocorrer, os dados são registrados como o tempo de falha, T, e o indicador de censura δ, é considerado como valor 1. Se no final do período da observação o evento não for observado, a observação é considerada como censura à direita, o valor de T seria considerado como o último tempo de observação e δ teria o valor 0. Se um indivíduo foi selecionado para um experimento e o evento de interesse já tenha ocorrido, seu dado seria censura à esquerda, e δ seria o valor 0. Se o evento não tiver ocorrido antes de uma visita (tempo U), mas já houver ocorrido quando de uma seguinte visita (tempo V), então T é sabido somente pertencer ao intervalo $[U, V)$. Estes dados são conhecidos como dados com censura intervalar.

Observe que os tempos exatos de falha tanto quanto os tempos de censura à direita e à esquerda, são casos particulares de dados com censura intervalar com $U_i = V_i$ para os tempos exatos de falha, $V_i = \infty$ para censuras à direita e $U_i = 0$ para censuras à esquerda. Assim, a função de verossimilhança para os vários tipos de censura pode ser escrita incorporando as seguintes contribuições:

I. tempos exatos: $f(T)$

II. tempos de censura à direita: $S(U)$

III.tempos de censura à esquerda: $1 - S(V)$

IV. tempos ocorrendo em intervalos: $S(U) - S(V)$.

onde $f(\cdot)$ denota a função densidade de probabilidade e $S(\cdot)$ a função de sobrevivência.

Devido à falta de metodologia estatística, uma comum aproximação é assumir que o evento ocorreu no final (ou início ou ponto médio) de cada intervalo, e então aplicar métodos para dados de tempos de falha. Entretanto, esta aproximação pode levar a invalidar a inferência, e em particular, tenderá a subestimar os erros-padrões dos parâmetros estimados.

Um importante caso especial de dados de censura intervalar é "current status data" (dados de status corrente). Nesta situação, cada indivíduo é observado so-

mente uma vez para o estado da ocorrência do evento de interesse no tempo de observação. Esta estrutura de dados é também conhecida como caso 1 de censura intervalar, e o caso geral é conhecido como caso 2 de censura intervalar.

Diversos autores assumem diferentes convenções para representar dados de sobrevivência com censura intervalar. Peto (1973) e Turnbull (1976), por exemplo, assumem intervalos de tempos fechados, isto é, $[U_i, V_i]$ para facilitar a acomodação de tempos exatos, ou seja, de observações em que $U_i = V_i$. Outros autores como, por exemplo, Finkelstein (1986), assumem intervalos semi-abertos e estabelecem que a contribuição de qualquer tempo exato de falha, T_i, é $S(T_i)$. Bons argumentos têm sido utilizados para a escolha de qualquer uma das convenções para definir intervalos censurados. Na prática, a escolha terá pouco impacto e qualquer convenção pode ser adotada (Lindsey & Ryan, 1998).

Considerando a convenção de intervalos semi-abertos, e fazendo $\delta_i = 1_{\{T_i \leq U_i\}}$ e $\gamma_i = 1_{\{T_i \in (U_i, V_i]\}}$ tem-se, para esse estudo, que a contribuição para a função de verossimilhança dos indivíduos com $\delta_i = 1$ e $\gamma_i = 0$, assumindo serem censurados à direita, é de $S(U_i)$, a contribuição dos indivíduos com $\delta_i = 0$ e $\gamma_i = 1$, é de $S(U_i) - S(V_i)$, e a contribuição dos indivíduos com $\delta_i = 0$ e $\gamma_i = 0$, assumindo serem censurados à esquerda, é de $1 - S(V_i)$. Assim, a função de verossimilhança é definida por:

$$L(S, \beta) = \prod_{i=1}^{n} [S(U_i, Z_i, \beta)]^{\delta_i} [S(U_i, Z_i, \beta) - S(V_i, Z_i, \beta)]^{\gamma_i} [1 - S(V_i, Z_i, \beta)]^{1-\delta_i-\gamma_i}$$

em que Z_i é o vetor de covariáveis ao qual está associado a um vetor de parâmetros desconhecidos β. Assim, o logaritmo da função de verossimilhança é definida por:

$$\begin{aligned} l = logL &= \sum_{i=1}^{n} \delta_i log[S(U_i, Z_i, \beta)] + \gamma_i log[S(U_i, Z_i, \beta) - S(V_i, Z_i, \beta)] \\ &\quad + (1 - \delta_i - \gamma_i) log[1 - S(V_i, Z_i, \beta)] \end{aligned} \tag{4.1}$$

A estimação dos parâmetros é feita maximizando-se (4.1), de modo que programas estatísticos podem ser utilizados para essa finalidade.

Para maximizar a função de verossimilhança sem a presença de covariáveis, uma forma simples e comum é usar o algoritmo de auto-consistência proposto por Turnbull (1976), que pode ser visto como uma aplicação do algoritmo EM. Esta aproximação é fácil de ser implementada, mas é sabido ter uma taxa de convergência lenta. Uma alternativa é aplicar o algoritmo iterativo do minorante convexo, introduzido por Groeneboom e Wellner (1992), que converge mais rápido que o algoritmo de auto-consistência. Todos os algoritmos citados são iterativos e não existe uma forma fechada para o ENPMV de F.

4.2 Caracterização dos Estimadores Não Paramétricos de Máxima Verossimilhança

Como mencionado anteriormente, existem dois casos de censura intervalar, descritos a seguir:

Caso 1. Seja $(T_1, U_1), \ldots, (T_n, U_n)$ uma amostra de variáveis aleatórias em $\Re_+^2$, em que T_i (tempo de sobrevivência) e U_i (tempo de observação) são variáveis aleatórias independentes (não-negativas) com funções de distribuição F e G, respectivamente. Suponha que somente observamos as variáveis U_i e $\delta_i = 1_A$ onde $A = \{T_i \leq U_i\}$. O log da verossimilhança para F é dado pela seguinte função

$$l(F) = logL(F) = \sum_{i=1}^{n} \{\delta_i logF(U_i) + (1 - \delta_i) log(1 - F(U_i))\}. \tag{4.2}$$

Caso 2. Seja $(T_1, U_1, V_1), \ldots, (T_n, U_n, V_n)$ uma amostra de variáveis aleatórias em $\Re_+^3$, em que T_i é uma variável aleatória (não-negativa) com função de distribuição F, e em que U_i e V_i são variáveis aleatórias (não-negativas), independentes de T_i, com função de distribuição conjunta H e tal que $U_i \leq V_i$ com probabilidade 1. Suponha que somente observamos as variáveis (U_i, V_i), que são os tempos de observação, e $\delta_i = 1_{\{T_i \leq U_i\}}$, $\gamma_i = 1_{\{T_i \in (U_i, V_i]\}}$. Neste caso, o log da verossimilhança para F é dado

pela função

$$l(F) = logL(F) = \sum_{i=1}^{n}\{\delta_i logF(U_i) + \gamma_i log(F(V_i) - F(U_i)) + (1 - \delta_i - \gamma_i)log(1 - F(U_i))\}. \quad (4.3)$$

O segundo caso de censura intervalar guarda relação com o caso de "censura dupla", onde o valor de T_i é conhecida se $U_i \leq T_i \leq V_i$ (Chang, 1990), enquanto que, na censura intervalar, o valor de T_i é desconhecido, mesmo sabendo que $U_i \leq T_i \leq V_i$. As partes restantes dos modelos para censura dupla e para censura intervalar, caso 2, são as mesmas.

Para encontrarmos tais estimativas, utilizaremos um procedimento proposto por Groeneboom e Wellner (1992) que baseia-se em um algoritmo para cálculo de regressões isotônicas.

4.3 Regressão Isotônica

Primeiro será mostrada a caracterização do estimador não paramétrico de máxima verossimilhança (ENPMV), usando os conceitos da regressão isotônica, para censura intervalar, caso 1.

Definição 4.3.1. Seja T o conjunto finito $\{t_1, \ldots, t_k\}$ com a ordenação simples $t_1 \prec t_2 \prec \ldots \prec t_k$. Uma função real f em um conjunto Q simplesmente ordenado é isotônica se $t \prec y$ implica $f(t) \leq f(y)$ $\forall t, y \in Q$.

Definição 4.3.2. Seja g uma função qualquer em Q e w uma função positiva em Q. Uma função g^* em Q é uma regressão isotônica de g com pesos w se, e somente se, minimiza, na classe de funções isotônicas f em Q, a soma

$$\sum_{i=1}^{k}[g(t_i) - f(t_i)]^2 w(t_i).$$

Assumindo ainda a ordenação simples $t_1 \prec t_2 \prec \ldots \prec t_k$, considere as somas acumuladas

$$G_j = \sum_{i=1}^{j} g(t_i)w(t_i) \quad \text{e} \quad W_j = \sum_{i=1}^{j} w(t_i), \quad j = 1, 2, \ldots, k.$$

A regressão isotônica de g é dada pela inclinação do minorante convexo máximo do diagrama de soma acumulada formado pelos pontos $P_j = (W_j, G_j)$ no plano cartesiano. O minorante convexo máximo é o supremo de todas as funções convexas cujos gráficos se encontram abaixo do diagrama de soma acumulada (Figura 4.1).

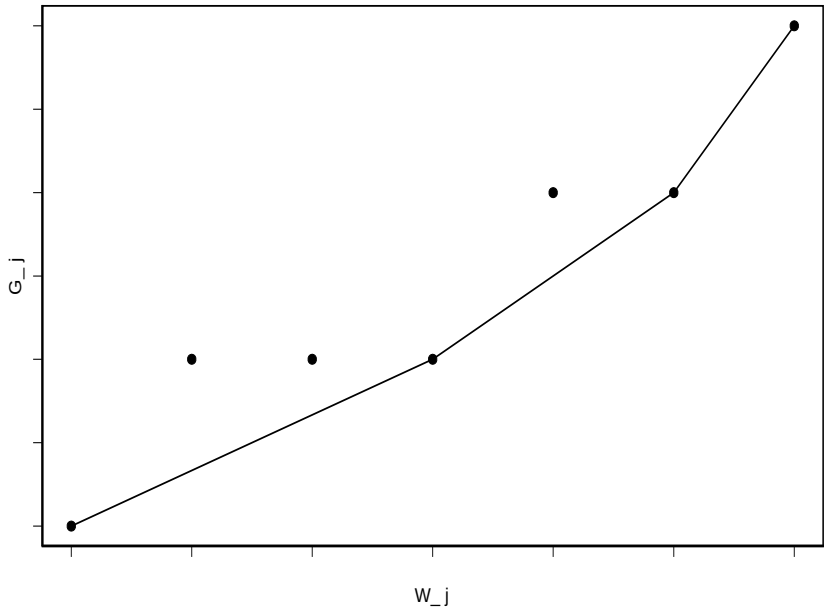

Figura 4.1: "Diagrama de Soma Acumulada" e respectivo "Minorante Convexo Máximo".

Teorema 4.3.1. Se f é isotônica em Q, a imagem de f está em $I \subset \Re$, e se ϕ é uma função convexa, então g^* maximiza

$$\sum_t \{\phi(f(t)) + (g(t) - f(t))\varphi(f(t))\}w(t)$$

em que $\varphi(t) = \phi'(t) = d\phi(t)/dt$.

Note que a expressão (4.2) pode ser escrita na forma acima com

$$\phi(t) = t\log t + (1-t)\log(1-t), \quad g(U_i) = \delta_i$$

$$\varphi(t) = \phi'(t) = \log(t) - \log(1-t), \quad w(U_i) = 1$$

Portanto, o estimador não paramétrico de máxima verossimilhança de F é dado pela regressão isotônica g^* e calculado da seguinte forma:

$$y_m = \max_{i\leq m}\min_{k\geq m}\frac{\sum_{i\leq j\leq k}\delta_{(j)}}{k-i+1}$$

em que $y_m = F(\hat{U}_m)$, $m = 1,\ldots,n$.

Essa solução é chamada da "fórmula max-min" para o problema de maximização. A solução pode ser encontrada graficamente plotando os pontos $(0,0)$ e $(i, \sum_{j\leq i}\delta_{(j)})$, $i = 1,\ldots,n$, no plano cartesiano, através do diagrama de soma acumulada, e calculando o minorante convexo máximo dos pontos em $[0,n]$.

Definição 4.3.3. O minorante convexo máximo é definido como a função $H^* : [0,n] \to \Re$ tal que

$$H^*(u) = sup\{H(u); H(i) \leq \sum_{j\leq i}\delta_j, 0 \leq i \leq n, H(0) = 0, \quad H \quad \text{é convexa}\}$$

Note que, se $\delta_i = 0$, $1 \leq i \leq k_1$ e $\delta_i = 1$, $k_2 \leq i \leq n$, para quaisquer $0 < k_1 < k_2 < n$, então $y_i = 0$, $1 \leq i \leq k_1$ e $y_i = 1$, $k_2 \leq i \leq n$. O valor y_i é dado pela derivada à esquerda de H^* em i para $i = 1,\ldots,n$.

A caracterização do estimador não paramétrico de máxima verossimilhança (ENPMV) para censura intervalar, caso 2, será apresentada na próxima seção com mais detalhes.

4.4 Algoritmo Iterativo do Minorante Convexo

Em muitos problemas estatísticos, o cálculo da estimativa não paramétrica resume-se a maximizar (minimizar) uma certa função côncava (convexa) $\mathcal{L}^*$ sobre

um cone convexo C em $\Re^n$. Para este tipo de problema de otimização, condições suficientes e necessárias serão discutidas.

Definição 4.4.1. Seja $\mathcal{L}^* : \Re^n \to (-\infty, \infty]$ uma função convexa contínua tal que $\mathcal{L}^*$ é continuamente diferenciável no conjunto $\{\underset{\sim}{y} \in \Re^n : \mathcal{L}^* < \infty\}$. Seja $C \subset \Re^n$ um cone convexo. Então

$$\underset{\sim}{\hat{y}} = \arg\min_{\underset{\sim}{y} \in C} \mathcal{L}^*(\underset{\sim}{y}). \tag{4.4}$$

Seja $S = \{U_j : \delta_j = 1 \quad \text{ou} \quad \gamma_j = 1\} \cup \{V_j : \delta_j = 0\}$ e R_j=j-ésimo ponto ordenado do conjunto S, $j = 1, \ldots, m = n + \sum_{j=1}^{n} \gamma_j$. Definindo $y_j = F(R_j)$, queremos encontrar $\underset{\sim}{y} = (y_1, y_2, \ldots, y_m)$ que minimiza

$$\mathcal{L}(F) = -\sum_{i=1}^{n} \delta_i log F(U_i) + \gamma_i log(F(V_i) - F(U_i)) + (1 - \delta_i - \gamma_i) log(1 - F(V_i))$$

tal que $\underset{\sim}{y} \in C$, $C = \{(x_1, x_2, \ldots, x_m) : 0 \leq x_1 \leq \ldots \leq x_m \leq 1\}$. Note que, a menos da truncagem no valor 1, C é um cone convexo.

Considere a expansão de Taylor de $\mathcal{L}(F)$ em torno de um ponto $\underset{\sim}{y}^{(k)} \in C$.

$$\begin{aligned} \mathcal{L}^*(\underset{\sim}{y}) &= \mathcal{L}(\underset{\sim}{y}^{(k)}) + (\underset{\sim}{y} - \underset{\sim}{y}^{(k)})^T \nabla \mathcal{L}(\underset{\sim}{y}^{(k)}) + \frac{1}{2}(\underset{\sim}{y} - \underset{\sim}{y}^{(k)})^T D^{(k)} (\underset{\sim}{y} - \underset{\sim}{y}^{(k)}) \\ &= c + \frac{1}{2}\{\underset{\sim}{y} - \underset{\sim}{y}^{(k)} + [D^{(k)}]^{-1} \nabla \mathcal{L}(\underset{\sim}{y}^{(k)})\}^T D^{(k)} \{\underset{\sim}{y} - \underset{\sim}{y}^{(k)} + [D^{(k)}]^{-1} \nabla \mathcal{L}(\underset{\sim}{y}^{(k)})\} \end{aligned}$$

onde $D^{(k)} = D(\underset{\sim}{y}^{(k)})$ é uma matriz diagonal positiva definida e c não depende de $\underset{\sim}{y}$.

Queremos encontrar y que minimize $\mathcal{L}^*(\underset{\sim}{y})$ em C. Jongbloed (1995) mostra que

$$\underset{\sim}{\hat{y}} = \arg\min_{\underset{\sim}{y} \in C} \mathcal{L}^*(\underset{\sim}{y})$$

se e somente se

$$\sum_{j=l}^{m} \frac{\partial \mathcal{L}^*(\underset{\sim}{y})}{\partial \underset{\sim}{y_j}} \begin{cases} \geq 0 \quad \text{para} \quad \forall 1 \leq l \leq m \\ = 0 \quad \text{se} \quad \hat{y}_l > \hat{y}_{l-1} \quad \text{ou} \quad l = 1 \end{cases}$$

Em particular, se $\mathcal{L}^*(\underset{\sim}{y})$ é da forma

$$\mathcal{L}^*(\underset{\sim}{y}) = \frac{1}{2}\sum_{j=1}^{m}(y_j - \xi_j)^2 d_j = \frac{1}{2}(\underset{\sim}{y} - \underset{\sim}{\xi})^T D(\underset{\sim}{y} - \underset{\sim}{\xi})$$

onde $\underset{\sim}{\xi} \in \Re^m$ é fixo, $D = diag(d_j)$ e $d_j < 0$, $1 \leq j \leq m$, temos que

$$\underset{\sim}{\hat{y}} = \arg\min_{\underset{\sim}{y} \in C} \frac{1}{2}(\underset{\sim}{y} - \underset{\sim}{\xi})^T D(\underset{\sim}{y} - \underset{\sim}{\xi})$$

se e somente se

$$\sum_{j=l}^{m} d_j(\hat{y}_j - \xi_j) \begin{cases} \leq 0 & \text{para,} \quad 1 \leq l \leq m \\ = 0 & \text{se} \quad \hat{y}_l > \hat{y}_{l-1} \quad \text{ou} \quad l = 1 \end{cases} \tag{4.5}$$

Considere os pontos P_j, $0 \leq j \leq m$, em $\Re^2$ definidos da seguinte forma $P_j = (\sum_{i=1}^{j} d_i, \sum_{i=1}^{j} d_i\xi_i)$, para $1 \leq j \leq m$, e $P_0 = (0, 0)$. Estes pontos formam um "diagrama de soma acumulada". O "minorante convexo máximo" deste diagrama de soma acumulada é definido como sendo a maior função convexa abaixo dos pontos P_j, $0 \leq j \leq m$.

É fácil ver que a caracterização (4.5) tem uma interpretação geométrica. Para $1 \leq l \leq m$, $\underset{\sim}{y_l}$ é a derivada à esquerda do minorante convexo máximo do diagrama de soma acumulada formado pelos pontos $\{P_j; 0 \leq j \leq m\}$ avaliada em P_l.

De fato, a desigualdade em (4.5) mostra que o minorante convexo é um minorante do diagrama de soma acumulada. Além disso, como a derivada à esquerda do minorante convexo só pode mudar (crescer) nos pontos P_l onde o minorante convexo iguala o diagrama de soma acumulada, a igualdade em (4.5) ocorre nestes pontos.

Considere $\underset{\sim}{\xi} = \underset{\sim}{y}^{(k)} - [D^{(k)}]^{-1}\nabla\mathcal{L}(\underset{\sim}{y}^{(k)})$. Então, $\mathcal{L}^*(\underset{\sim}{y})$ é maximizada fazendo y_l igual a derivada à esquerda em P_l do minorante convexo máximo do diagrama de soma acumulada $\{P_j; 0 \leq j \leq m\}$, com $P_j = (\sum_{i=1}^{j} d_i, \sum_{i=1}^{j} d_i\xi_i)$. Podemos tomar $d_i = \partial^2\mathcal{L}(\underset{\sim}{y})/\partial y_i^2$.

Note que $\underset{\sim}{\xi}$ depende de $\underset{\sim}{y}^{(k)}$ (que é a estimativa corrente de $\underset{\sim}{\hat{y}}$). Portanto, temos que utilizar um algoritmo iterativo para obter $\underset{\sim}{\hat{y}}$.

0) Faça $k = 0$ e ϵ pequeno.

1) Escolha $\underset{\sim}{y}^{(k)} \in C$.

2) Construa o diagrama de soma acumulada com $\{P_j; 0 \leq j \leq m\}$ a partir de $\underset{\sim}{y}^{(k)}$.

3) Obtenha $\underset{\sim}{y}^{(k+1)}$ através da derivada à esquerda do minorante convexo máximo do diagrama de soma acumulada.

4) Se $\frac{||\underset{\sim}{y}^{(k+1)} - \underset{\sim}{y}^{(k)}||}{||\underset{\sim}{y}^{(k)}||} > \epsilon$, faça $\underset{\sim}{y}^{(k)} = \underset{\sim}{y}^{(k+1)}$ e vá para 2. Caso contrário, faça $\underset{\sim}{\hat{y}} = \underset{\sim}{y}^{(k+1)}$.

Capítulo 5

Modelo de Cox via Métodos de Imputação

5.1 Introdução

Dados faltantes frequentemente tornam mais difíceis análises de dados para investigações científicas, principalmente na área médica. A imputação é uma técnica comum para tratar conjuntos de dados com valores faltantes. O desenvolvimento de métodos estatísticos para tratar dados faltantes tem sido uma área de pesquisa ativa nas últimas décadas. Existem alguns fatores de interesse que tipicamente surgem nessa situação: perda de eficiência, dificuldade em tratar e analisar os dados, e vício devido às diferenças entre os dados observados e não observados.

Existem dois tipos de imputação que são encontrados com frequência na literatura: simples e múltipla. O método da imputação simples substitui cada valor faltante por um único valor imputado. A principal desvantagem da imputação simples é que ela trata os valores faltantes como se fossem valores imputados, subestimando a verdadeira variabilidade. A imputação múltipla supera esta deficiência substituindo cada valor faltante com dois ou mais valores prováveis. Consequente-

mente, a imputação múltipla cria os conjuntos de dados imputados, sendo cada qual analisado separadamente por um método padronizado, e os resultados combinados.

Neste trabalho, o método de imputação será baseado na imputação múltipla para regressão de Cox com dados de censura intervalar. A idéia básica é imputar tempos de sobrevivência exatos para dados com censura intervalar finita e aproveitar métodos desenvolvidos para dados com censura à direita.

5.2 Imputação Múltipla

Rubin (1987) descreve a imputação múltipla como um processo de três passos. Primeiramente, os conjuntos de valores plausíveis para as observações faltantes são criados refletindo a incerteza sobre o modelo de não-resposta. Cada um destes conjuntos de valores plausíveis pode ser usado para substituir os valores faltantes e cria um conjunto de dados completo. Segundo, cada um destes conjuntos de dados pode ser analisado usando métodos de dados completos. Finalmente, os resultados são combinados permitindo que a incerteza da imputação seja considerada.

Para imputar os tempos de sobrevivência de dados com censura intervalar, foram utilizados dois algoritmos semiparamétricos: "Poor Man's Data Augmentation" (PMDA) e "Asymptotic Normal Data Augmentation" (ANDA), para análise de regressão de dados censurados. Em seguida, serão descritos os dois algoritmos.

5.2.1 Algoritmo PMDA para Dados de Censura Intervalar

O PMDA é usado para imputar tempos de sobrevivência de dados com censura intervalar envolvendo a criação de conjuntos de dados imputados múltiplos num algoritmo iterativo. Serão usados o sobrescrito (j) e o subscrito (k) para representar a j-ésima iteração e o k-ésimo conjunto de dados imputados, respectivamente. Os

parâmetros a serem estimados incluem o coeficiente de regressão β e a função de sobrevivência base S_0. A seguir, descreve-se o algoritmo PMDA:

(1) Supor estimativas iniciais do coeficiente de regressão $\hat{\beta}^{(j)}$ e da função de sobrevivência base $\hat{S}_0^{(j)}$.

(2) Gerar m conjuntos de possíveis observações de censura à direita $\{X_{(1)}, \delta_{(1)}, Z\}$, $\ldots, \{X_{(m)}, \delta_{(m)}, Z\}$ como segue: para cada observação (U_i, V_i, Z_i), $i = 1, \ldots, n$ e $k = 1, \ldots, m$, gera-se o valor T_i a partir da distribuição $[\hat{S}_0^{(j)}]^{exp(Z_i\hat{\beta}^{(j)})}$, condicionado a $\{U_i < T_i \leq V_i\}$, de tal forma que, se

$$V_i < \infty \begin{cases} X_{(k),i} = T_i \\ \delta_{(k),i} = 1 \end{cases}, \quad \text{e se} \quad V_i = \infty \begin{cases} X_{(k),i} = U_i \\ \delta_{(k),i} = 0 \end{cases}$$

(3) Usar cada $\{X_{(k)}, \delta_{(k)}, Z\}$ para ajustar o modelo de Cox para obter a estimativa $\hat{\beta}_{(k)}^{(j)}$ e suas covariâncias estimadas $\hat{\Sigma}_{(k)}^{(j)}$.

(4) Baseado em $\{X_{(k)}, \delta_{(k)}, Z\}$ e $\hat{\beta}_{(k)}^{(j)}$, calcular a estimativa de Breslow da função de sobrevivência base $\hat{S}_{0,(k)}^{(j)}$ para $k = 1, \ldots, m$.

(5) Seja

$$\hat{\beta}^{(j+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\beta}_{(k)}^{(j)}, \quad \hat{S}_0^{(j+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{S}_{0,(k)}^{j}$$

$$\hat{\Sigma}^{(j+1)} = \frac{1}{m}\sum_{k=1}^{m} \hat{\Sigma}_{(k)}^{(j)} + \left(1 + \frac{1}{m}\right) \frac{\sum_{k=1}^{m} [\hat{\beta}_{(k)}^{(j)} - \hat{\beta}^{(j+1)}]^2}{m-1} \tag{5.1}$$

(6) Fazer $j \leftarrow j + 1$. Ir para o passo 2 até $\hat{\beta}^{(j)}$ convergir.

Os $\hat{\beta}^{(j)}$, $\Sigma^{(j)}$ e $\hat{S}_0^{(j)}$ na convergência são as estimativas finais.

No passo 1, como valor inicial para o processo, considera-se $\beta^{(0)} = 0$. Para a estimativa de sobrevivência base inicial $S_0^{(0)}$, é necessário primeiro gerar conjuntos de dados imputados, mantendo as observações de censura à direita, para cada observação de censura intervalar $(U_i, V_i]$. O tempo de falha exato T_i é extraído aleatoriamente da distribuição uniforme $U(U_i, V_i)$ em que $X_{(k),i} = T_i$ e $\delta_{(k),i} = 1$

quando $V_i < \infty$, para $k = 1, \ldots, m$ e $i = 1, \ldots, n$. A estimativa de Breslow da função de sobrevivência base é dada por

$$\hat{S}_0^{(0)} = \sum_{k=1}^{m} \frac{\hat{S}_{0,(k)}^{(0)}}{m}.$$

No passo 2, são criados m conjuntos de dados imputados. A experiência geral é que frequentemente m não necessita ser grande. Usa-se um tamanho moderado, $m = 10$. Supondo que, no intervalo $(U_i, V_i]$, $[\hat{S}_0^{(j)}]^{exp(Z_i\hat{\beta}^{(j)})}$ tem massa de probabilidade $\{p_1, \ldots, p_{k_i}\}$ nos pontos $\{x_1, \ldots, x_{k_i}\}$, então T_i é extraído aleatoriamente de $\{x_1, \ldots, x_{k_i}\}$ com probabilidades proporcionais a $\{p_1, \ldots, p_{k_i}\}$.

O passo 3 pode ser realizado facilmente usando os métodos desenvolvidos para dados de censura à direita, tal como o método da verossimilhança parcial, que é aplicado aos dados imputados para atualizar as estimativas.

No passo 5, obtém-se a estimativa da covariância do coeficiente de regressão estimado em (5.1). Ela é a soma das variâncias dentro (intra) e entre (inter) imputações. O segundo termo mede a variabilidade extra devido à censura intervalar finita. Um fator de inflação ($\frac{1}{m}$) é usado quando estima-se a variância entre (inter) imputações para incluir o efeito do número finito de imputações (Rubin, 1987; Tanner e Wong, 1987b; Schenker e Welsh, 1988). Na simulação, o critério de convergência utilizado foi $|\beta^{(j+1)} - \hat{\beta}^{(j)}| < 0.01$ ou $j > 50$.

5.2.2 Algoritmo ANDA para Dados de Censura Intervalar

Sabe-se que o PMDA pode subestimar a verdadeira variabilidade quando a proporção de dados com censura à direita é grande, enquanto que o algoritmo ANDA fornece resultados mais precisos (Wei e Tanner, 1991). O algoritmo ANDA pode ser implementado modificando dois passos no algoritmo PMDA como segue:

(1) No passo 5, a estimativa do coeficiente de regressão é aproximada por uma combinação de distribuições normais,

$$g^{(j+1)}(\beta) = \frac{1}{m}\sum_{k=1}^{m} N\left(\hat{\beta}_{(k)}^{(j)}, \hat{\Sigma}_{(k)}^{(j)}\right)$$

(2) No passo 2, primeiramente, amostre m tempos de $g^{(j)}(\beta)$ para obter $\beta_{(k)}^{(j)}$, $k = 1, \ldots, m$; então, para cada $k = 1, \ldots, m$ e cada observação com censura intervalar finita (U_i, V_i, Z_i), amostre T_i da distribuição $[\hat{S}_0^j]^{exp(Z_i\hat{\beta}_{(k)}^{(j)})}$, condicionado a $\{U_i < T_i \leq V_i\}$, mantendo as observações com censura à direita.

Também usa-se o mesmo valor inicial $S_0^{(0)}$, $\hat{\beta}_{(k)}^{(0)} = 0$, e $\hat{\Sigma}_{(k)}^{(0)} = 0$. As partes restantes do algoritmo ANDA serão as mesmas do algoritmo PMDA.

5.3 Suavização da Função de Sobrevivência

Um problema constante na Estatística é a obtenção da função densidade de probabilidade de uma variável aleatória. Quando não é possível encontrar uma distribuição conhecida, utiliza-se a estimação não paramétrica de densidade. Nestes casos, existem conjuntos de técnicas que podem tornar possível a estimação de uma função sem a determinação de uma estrutura paramétrica para a mesma.

A suavização pelo método do núcleo é um difundido método não paramétrico de estimação de uma função densidade de probabilidade. Uma questão crucial na aplicação deste método é a determinação do parâmetro de suavização ou janela, freqüentemente chamada h, que controla a quantidade de suavização a ser feita. Na literatura, a escolha do h é ampla, sendo os métodos de validação cruzada e plug-in os mais estudados.

5.3.1 Núcleo Estimador

Suponha um processo contínuo univariado com distribuição desconhecida, mas bem-definida, e seja a distribuição subjacente a ser estimada. Uma solução simples para o problema de estimação é o histograma.

Seja $f(t)$ a densidade desconhecida e $\hat{f}(t)$ a estimativa suavizada dos n eventos nos pontos T_i:

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{t - T_i}{h}\right) \tag{5.2}$$

onde

$$w(t) = \begin{cases} \frac{1}{2}, & |t| < 1 \\ 0, & \text{caso contrário} \end{cases}$$

Observe que na equação (5.2), a função $\hat{f}(t)$ não é contínua. Isso pode ser evitado substituindo a função $w(t)$ por uma função núcleo K que satisfaça as condições $K(t) \geq 0$, $\forall t \in \Re$, e

$$\int_{-\infty}^{\infty} K(t)dt = 1. \tag{5.3}$$

Suponha uma amostra $T_1, \ldots, T_n$ de observações i.i.d de uma distribuição univariada contínua com função densidade de probabilidade f. O núcleo estimador de f avaliado no ponto t é definido como

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{t - T_i}{h}\right) \tag{5.4}$$

onde h é denominado parâmetro de suavidade ou janela. A função K é chamada de função núcleo, normalmente uma densidade de probabilidade simétrica. A janela h controla a quantidade de suavização que é feita. O fator de normalização n^{-1} faz de $\hat{f}(t)$ uma função densidade de probabilidade. Dado que K é não-negativa e satisfaz a condição (5.4), conclui-se que $\hat{f}$ é uma função densidade de probabilidade. Além disso, $\hat{f}$ herda as propriedades de diferenciabilidade e continuidade de K, devido à sua forma aditiva.

5.3.2 Propriedades do Núcleo Estimador

A seguir são definidas as propriedades do núcleo estimador para a função densidade utilizando algumas medidas do desempenho do estimador $\hat{f}$.

Uma qualidade da estimativa pode ser medida pelo cálculo do Erro Quadrático Médio (MSE) em t, definido como:

$$MSE_t(\hat{f}) = E\{\hat{f}(t) - f(t)\}^2 \tag{5.5}$$

Esta medida pode ser reescrita em função do vício e da variância de $\hat{f}$. Observa-se que uma diminuição do vício implica no aumento da variância do estimador e vice-versa.

O Erro Quadrático Integrado Médio (MISE) é uma medida global da discrepância de $\hat{f}$ com relação a f, e é definida como:

$$MISE(\hat{f}) = E \int \{\hat{f}(t) - f(t)\}^2 dt$$

O MISE pode ser escrito como a soma do vício quadrático integrado e da variância integrada.

Outra medida muito utilizada é o Erro Quadrático Integrado (ISE) definido como:

$$ISE(\hat{f}) = \int \{\hat{f}(t) - f(t)\}^2 dt. \tag{5.6}$$

Vale ressaltar que o vício e a variância do núcleo estimador não podem ser facilmente calculados para o caso geral. Silverman (1986) mostra algumas suposições que permitem obter aproximações.

5.3.3 Escolha da Função Núcleo

A escolha da função núcleo é uma questão bem discutida na literatura. Utilizando-se qualquer núcleo razoável obtém-se resultados aproximadamente ótimos e a função núcleo pode ser escolhida com base em propriedades como a diferenciabilidade.

Alguns núcleos amplamente utilizados são apresentados a seguir:

Epanechnikov $K(t) = 0.75(1 - t^2)I_{[-1,1]}(t)$;

Uniforme $K(t) = \frac{1}{2}I_{[-1,1]}(t)$;

Quártico $K(t) = 0.9375(1 - x^2)^2 I_{[-1,1]}(t)$;

Normal $K(t) = (\sqrt{2\pi})^{-1} exp(-\frac{t^2}{2}), -\infty < t < \infty$.

5.3.4 Escolha da Janela Ótima

A escolha de h tem mais importância para o bom desempenho do núcleo estimador que a escolha da função núcleo, K. A escolha do valor da janela h é fundamental para o bom desempenho do núcleo estimador, tornando-se importante o desenvolvimento de métodos automáticos para cálculo da janela ótima. Dentre os mais usados, destacam-se os métodos plug-in e o de validação cruzada por mínimos quadrados para estimação da função de densidade.

Pan (2000a) propôs um estimador não paramétrico de máxima verossimilhança (ENPMV) para estimar a função de sobrevivência com dados de censura intervalar usando um método de suavização através de núcleo estimadores.

Suponha que $\hat{S}$ é o estimador não paramétrico de máxima verossimilhança (ENPMV) da função de sobrevivência, então o núcleo estimador de sua densidade é

$$f(t;h) = -\frac{1}{h}\int K\left(\frac{t-u}{h}\right) d\hat{S}(u) \tag{5.7}$$

em que K é uma função núcleo e h é a janela ótima. O núcleo estimador da função de sobrevivência $S(t;h)$ é então obtido integrando o estimador da densidade $f(t;h)$.

Será aplicada uma aproximação de validação cruzada que consiste primeiro em dividir o conjunto de dados originais $D = \{U_i, V_i\}$, $i = 1, \ldots, n$, aleatoriamente em V conjuntos de igual tamanho $D^{(v)}$, $v = 1, \ldots, V$. Suponha que $\hat{S}^{(-v)}(\cdot;h)$ é o núcleo estimador da função de sobrevivência com base nos dados $D - D^{(v)}$. Então

a janela ótima h_0 é definida como

$$h_0 = \arg\max_h \sum_{v=1}^{V} L(S^{(-v)}(\cdot; h)|D^{(v)})$$

onde o núcleo estimador da função de sobrevivência é $\hat{S}(\cdot) = \hat{S}(\cdot; h_0)$.

Note que, na seleção de uma janela h apropriada, os métodos usuais baseados nas distribuições assintóticas para dados completos ou com censura à direita, como os estimadores "plug-in", não são disponíveis em parte devido ao fato que geralmente não existe forma explícita para o ENPMV dos dados de censura intervalar. Por exemplo, não é claro como escrever o erro quadrático integrado médio (MISE) do núcleo estimador com janela h. Além disso, o método de validação cruzada é o mais confiável para amostras pequenas. Computacionalmente este método é lento e instável, devido a este fato utilizou-se uma outra metodologia que foi proposta por Silverman (1986), que será descrita em seguida.

Silverman (1986) propôs uma aproximação natural para escolha da janela ótima, h, através da minimização do erro quadrático médio aproximado que pode ser mostrado pelo cálculo simples,

$$h = k_2^{-2/5} \left\{ \int K(t)^2 dt \right\}^{1/5} \left\{ \int f''(x)^2 dx \right\}^{-1/5} n^{-1/5}. \tag{5.8}$$

Uma aproximação muito fácil e natural é usar uma família padrão de distribuições atribuindo um valor ao termo $\int f''(x)^2 dx$ na expressão (5.8) para a janela ótima ideal. Por exemplo, a distribuição normal com variância σ^2 tem, sendo ϕ a densidade normal padrão,

$$\begin{aligned} \int f''(x)^2 dx &= \sigma^{-5} \int \phi''(x)^2 dx \\ &= \frac{3}{8}\pi^{-1/2}\sigma^{-5} \approx 0.212\sigma^{-5}. \end{aligned} \tag{5.9}$$

se o núcleo normal é usado, então a janela ótima será encontrada substituindo o valor (5.9),

$$h = 1.06\sigma n^{-1/5}. \tag{5.10}$$

Consequentemente, uma maneira fácil de escolher o parâmetro de suavização é estimar σ a partir dos dados e então substituir em (5.10). O desvio-padrão da amostra usual ou um estimador mais robusto de σ poderiam ser usados. A equação (5.10) trabalha bem se a população for normalmente distribuída, mas pode suavizar pouco se a população for multimodal. A fim de investigar a sensibilidade da janela ótima, melhores resultados podem ser obtidos usando uma medida robusta de dispersão. A equação (5.10) escrito em termos da amplitude interquartílica A da distribuição normal torna-se

$$h = 0.79An^{-1/5}. \tag{5.11}$$

O uso de (5.11) pode também não suavizar adequadamente quando a população for bimodal. Para evitar isto, uma alternativa para (5.10) e (5.11) seria usar a estimativa adaptável de dispersão, que é definida com

$$A = \min(\text{desvio-padrão}, \text{amplitude interquartílica}/1.34). \tag{5.12}$$

em vez de σ na equação (5.10). Isto servirá para as densidades unimodais e também para o caso de densidades bimodais.

5.3.5 Problemas de Fronteira

Problemas de fronteira podem afetar os núcleo estimadores quando o suporte da função a ser estimada é limitado. Sabe-se que todo estimador de f é viciado, mas perto das fronteiras do intervalo, o vício do estimador é ainda maior que no seu interior, com tendência a subestimação. Na literatura, existem alguns métodos para correção deste problema.

Nas situações em que a variável de interesse tem suporte limitado em um ou nos dois extremos (por exemplo, assume somente valores positivos), uma maneira

simples de eliminação do efeito de fronteira é realizar uma transformação na variável de interesse de modo que ela não tenha mais domínio limitado.

O motivo da ocorrência dos problemas de fronteira é que a curva a ser estimada tem uma descontinuidade no ponto de fronteira, de modo que a expansão do vício, que depende de suposições de suavidade, não é mais válida. O efeito da fronteira também pode afetar a seleção da janela ótima, que é fundamental para o bom desempenho do núcleo estimador.

Diggle e Marron (1988) adaptaram o método de reflexão (ou "imagem espelhada") para a correção do método da estimação de densidade. Essa correção toma as funções núcleo que se estendem além da fronteira e "dobra-as" na fronteira, de modo que toda a sua massa esteja dentro do intervalo. Este método é válido para estimação de funções densidade e de intensidade.

Pan (2000b) propôs usar a estimativa de Breslow da sobrevivência base (veja passo 4 na Seção 5.2.1), podendo haver algumas questões se a estimativa de Breslow conduz a uma diminuição do suporte da estimativa de sobrevivência base com o procedimento de iteração. Isto não pareceu ser um problema no estudo de simulação. Entretanto, isto poderia transformar-se numa questão quando o tamanho da amostra é pequeno.

Para investigar isto, Pan (2000b) fez um estudo de simulação usando o algoritmo PMDA similar ao passo 3, mas com um tamanho de amostra muito pequeno ($n = 40$). Em 42 dentre as 1000 simulações independentes, o suporte da estimativa de sobrevivência base reduziu-se somente a um ponto, levando os coeficientes de regressão estimados próximos a zero. Para remediar isto, foi proposto suavizar a sobrevivência base estimada. Em vez de usar a estimativa de Breslow, foi utilizado a estimativa de Link, que é uma suavização linear da estimativa de Breslow, podendo evitar uma redução do suporte da estimativa de sobrevivência base e assim corrigir o vício. Contudo, devido ao bom desempenho com um tamanho de amostra razoável, usando a estimativa de Breslow pode ser preferível a menos que o tamanho

de amostra seja muito pequeno.

Neste estudo, para a escolha da janela ótima no processo de suavização de S_0 via núcleo estimadores, apresentamos dois métodos. A metodologia proposta por Pan (2000a) não pode ser utilizado pois tornou-se computacionalmente lenta e instável. Já a metodologia utilizando a escolha de h ótimo sugerida por Silverman (1986), além de ser uma aproximação natural e fácil, é bastante viável computacionalmente.

Capítulo 6

Simulações

Para investigar o desempenho para amostras finitas, foi realizado um estudo de simulação. Nesse estudo, os dados foram gerados do modelo de Cox usando S-plus, sendo a distribuiçao de Weibull base com parâmetro de forma 2 e parâmetro de escala 1, com tamanho de amostra n. Foi considerado uma única covariável (Z_i), binária $\{0, 1\}$ (isto é, com probabilidade igual) ou uma uniforme contínua $U(0, 2)$. O verdadeiro coeficiente de regressão (β_0) no modelo de Cox foi zero ou um. O comprimento do intervalo de tempo entre duas examinações contínuas (len) é constante $(len = 0.5)$ ou como um número aleatório de uma distribuição uniforme $U(0.2, 0.8)$.

Supondo $k + 1$ examinações, para gerar um conjunto de dados simulados, foram repetidos os seguintes passos:

(1) Gerar a covariável Z_i da distribuição especificada.

(2) Gerar o tempo de sobrevivência T_i de acordo com o modelo de Cox especificado.

(3) Gerar o limite superior do primeiro intervalo de tempo de exame $Y_i \sim U(0, 1)$.

(4) Dos intervalos $(0, Y_i], (Y_i, Y_i+len], \ldots, (Y_i+k*len, \infty)$ escolhe-se $(U_i, V_i]$ como o único que satisfaz $U_i < T_i \leq V_i$. A Tabela 6.1 mostra os 9 casos considerados.

Tabela 6.1: Casos considerados para simulação.

#	β_0	Z	len	k
1	0	{0,1}	0.5	2
2	1	{0,1}	0.5	2
3	1	$U(0,2)$	0.5	2
4	0	{0,1}	0.5	1
5	1	{0,1}	0.5	1
6	1	$U(0,2)$	0.5	1
7	0	{0,1}	$U(0.2,0.8)$	2
8	1	{0,1}	$U(0.2,0.8)$	2
9	1	$U(0,2)$	$U(0.2,0.8)$	2

Foi feita a implementação dos dois algoritmos semiparamétricos: PMDA e ANDA, conforme metodologia apresentada no Capítulo 5. Como um dos objetivos é suavizar a curva de sobrevivência base estimada, o procedimento de suavização foi implementado pelo método do núcleo estimador com objetivo de estimar a sobrevivência base, utilizando o núcleo normal, apresentado na Seção 5.3.6 do Capítulo 5. Para cada tamanho de amostra n, foram considerados m conjuntos de dados imputados, $m = 1, \ldots, 10$.

6.1 Resultados

Para testar a eficiência do algoritmo, será usado o recurso da simulação similar ao utilizado por Pan (2000b), a fim de comparar os resultados com aqueles encontrados por ele. Será mostrado o desempenho dos algoritmos PMDA e ANDA.

Foram utilizadas médias de Monte Carlo dos coeficientes de regressão estimados

e seus erros-padrões estimados (com seus desvios-padrões de Monte Carlo entre parênteses) e as duas implementações de imputação múltipla: PMDA e ANDA. Além disso, utilizou-se o núcleo estimador para suavizar a sobrevivência base para tamanhos de amostras pequenos.

Foram geradas 1000 amostras independentes de tamanhos 25, 50 e 100, e 400 amostras independentes de tamanhos 25 e 50, com número de imputações variando de 1 até 10, já que é encontrado na literatura que o número de imputação igual a 10 é considerado um tamanho moderado. O verdadeiro coeficiente de regressão no modelo de Cox utilizado foi zero ou um, para os casos considerados na simulação (# 1,2,...,9), como está descrito na Tabela 6.1.

6.1.1 Resultados encontrados por Pan(2000b)

Os resultados encontrados pelo método proposto por Pan (2000b) mostram que os coeficientes de regressão estimados apresentam vício pequeno. Note que o desempenho dos dois estimadores são próximos. Posteriormente, considerando o erro-padrão estimado, é verificado que o PMDA desempenha razoavelmente bem em muitos casos, mas às vezes subestima ligeiramente a variabilidade se a proporção de dados com censura à direita é grande (por exemplo, nos casos 8 e 9). Verificou-se que o ANDA tem um desempenho satisfatório em todos os casos. O erro-padrão médio estimado é próximo do desvio-padrão de Monte Carlo do coeficiente de regressão estimado (Tabela 6.2).

Tabela 6.2: Estimativas de β via Monte Carlo-Pan(n=100, 1000 amostras, m=10).

	PMDA		ANDA	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	0.00 (0.23)	0.23 (0.01)	0.01 (0.23)	0.23 (0.01)
2	1.00 (0.25)	0.25 (0.02)	1.00 (0.25)	0.25 (0.02)
3	0.99 (0.24)	0.23 (0.02)	0.98 (0.25)	0.26 (0.04)
4	0.00 (0.27)	0.27 (0.01)	0.00 (0.27)	0.27 (0.01)
5	1.00 (0.28)	0.27 (0.02)	1.00 (0.27)	0.27 (0.02)
6	0.99 (0.25)	0.24 (0.03)	0.98 (0.25)	0.26 (0.04)
7	-0.01 (0.20)	0.21 (0.01)	0.00 (0.20)	0.21 (0.01)
8	1.01 (0.27)	0.26 (0.02)	1.00 (0.26)	0.26 (0.02)
9	0.98 (0.25)	0.24 (0.02)	0.97 (0.25)	0.26 (0.04)

6.1.2 Resultados para o Algoritmo PMDA

Considerando os n casos, pode-se observar que os resultados encontrados para o algoritmo PMDA não são muito diferentes daqueles encontrados por Pan (2000b). O coeficiente de regressão estimado ($\hat{\beta}$) apresenta vício muito pequeno para o tamanho de amostra considerado. O erro-padrão médio estimado é bem próximo do desvio-padrão de Monte Carlo do coeficiente de regressão estimado. Verifica-se que a medida que o número de imputações aumenta, o coeficiente de regressão estimado ($\hat{\beta}$) apresenta vício pequeno. Note que nos casos 3, 6 e 9, os coeficientes de regressão estimados apresentam vício e variabilidade muito maior que nos outros casos. Isto poderia ter sido influenciado pelo fato da covariável ser uniforme contínua (Tabelas 6.3, 6.4 e 6.5).

Tabela 6.3: Estimativas de β via Monte Carlo-PMDA (n=100, 1000 amostras).

	$m = 2$		$m = 3$		$m = 4$	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	-0.01 (0.29)	0.28 (0.01)	0.00 (0.29)	0.28 (0.01)	-0.01 (0.29)	0.28 (0.01)
2	1.03 (0.27)	0.26 (0.02)	1.02 (0.27)	0.26 (0.01)	1.04 (0.27)	0.26 (0.01)
3	1.11 (0.25)	0.23 (0.02)	1.10 (0.24)	0.22 (0.01)	1.10 (0.25)	0.22 (0.02)
4	0.00 (0.34)	0.33 (0.02)	0.00 (0.34)	0.33 (0.02)	-0.02 (0.33)	0.33 (0.02)
5	1.04 (0.31)	0.30 (0.02)	1.02 (0.30)	0.30 (0.02)	1.04 (0.31)	0.30 (0.02)
6	1.10 (0.25)	0.24 (0.02)	1.10 (0.26)	0.24 (0.02)	1.12 (0.26)	0.24 (0.02)
7	0.00 (0.27)	0.27 (0.01)	0.01 (0.27)	0.27 (0.01)	0.00 (0.28)	0.27 (0.01)
8	1.04 (0.26)	0.26 (0.01)	1.03 (0.26)	0.25 (0.01)	1.03 (0.26)	0.25 (0.01)
9	1.11 (0.24)	0.22 (0.02)	1.13 (0.24)	0.22 (0.02)	1.13 (0.25)	0.22 (0.02)

Tabela 6.4: Estimativas de β via Monte Carlo-PMDA (n=100, 1000 amostras).

	$m = 5$		$m = 6$		$m = 7$	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	-0.01 (0.28)	0.28 (0.01)	0.01 (0.28)	0.28 (0.01)	0.00 (0.28)	0.28 (0.01)
2	1.02 (0.26)	0.26 (0.01)	1.02 (0.27)	0.26 (0.01)	1.03 (0.28)	0.26 (0.01)
3	1.10 (0.25)	0.22 (0.02)	1.11 (0.25)	0.22 (0.02)	1.13 (0.25)	0.22 (0.02)
4	-0.01 (0.34)	0.33 (0.02)	-0.01 (0.33)	0.33 (0.02)	0.02 (0.34)	0.33 (0.02)
5	1.03 (0.31)	0.30 (0.02)	1.04 (0.31)	0.29 (0.02)	1.04 (0.30)	0.30 (0.02)
6	1.10 (0.26)	0.24 (0.02)	1.09 (0.27)	0.24 (0.02)	1.10 (0.27)	0.24 (0.02)
7	0.00 (0.27)	0.27 (0.01)	-0.02 (0.26)	0.27 (0.01)	0.00 (0.28)	0.27 (0.01)
8	1.04 (0.26)	0.25 (0.01)	1.04 (0.27)	0.25 (0.01)	1.04 (0.27)	0.25 (0.01)
9	1.14 (0.25)	0.22 (0.02)	1.13 (0.25)	0.22 (0.02)	1.13 (0.25)	0.22 (0.02)

Tabela 6.5: Estimativas de β via Monte Carlo-PMDA (n=100, 1000 amostras).

	$m = 8$		$m = 9$		$m = 10$	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	0.00 (0.29)	0.28 (0.01)	0.00 (0.29)	0.28 (0.01)	0.01 (0.28)	0.28 (0.01)
2	1.03 (0.27)	0.26 (0.01)	1.03 (0.27)	0.26 (0.01)	1.04 (0.27)	0.26 (0.01)
3	1.11 (0.25)	0.22 (0.02)	1.13 (0.26)	0.22 (0.02)	1.12 (0.26)	0.22 (0.02)
4	0.00 (0.34)	0.33 (0.02)	0.00 (0.32)	0.33 (0.02)	0.00 (0.35)	0.33 (0.02)
5	1.00 (0.29)	0.29 (0.02)	1.03 (0.31)	0.30 (0.02)	1.04 (0.31)	0.30 (0.02)
6	1.11 (0.27)	0.24 (0.02)	1.11 (0.26)	0.24 (0.02)	1.12 (0.26)	0.24 (0.02)
7	-0.02 (0.26)	0.27 (0.01)	-0.01 (0.27)	0.27 (0.01)	-0.01 (0.26)	0.27 (0.01)
8	1.04 (0.27)	0.25 (0.01)	1.02 (0.26)	0.25 (0.01)	1.04 (0.28)	0.25 (0.01)
9	1.12 (0.24)	0.22 (0.02)	1.12 (0.26)	0.22 (0.02)	1.14 (0.26)	0.22 (0.02)

Quando diminui o tamanho de amostra de 100 para 25, observa-se que a precisão do coeficiente de regressão estimado diminui e aumenta a variabilidade dos erros-padrões estimados. Já considerando os desvios-padrões estimados de Monte Carlo, verifica-se que os resultados ficam bem diferentes daqueles encontrados quando o tamanho de amostra é igual a 100 (Tabelas 6.6 e 6.10).

Na Tabela 6.7 pode-se observar que os coeficientes de regressão nos casos 2, 5 e 8 são ligeiramente viciados e a variabilidade é maior comparando com as amostras de tamanho 100.

Tabela 6.6: Estimativas de β via Monte Carlo-PMDA ($n = 25$, 1000 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	-0.02 (0.72)	0.67 (1.43)
7	0.00 (0.65)	0.62 (1.19)

Tabela 6.7: Estimativas de β via Monte Carlo-PMDA ($n = 50$, 1000 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	-0.02 (0.42)	0.41 (0.03)
2	1.07 (0.41)	0.38 (0.03)
4	0.00 (0.49)	0.48 (0.05)
5	1.08 (0.45)	0.43 (0.05)
7	0.00 (0.38)	0.39 (0.03)
8	1.06 (0.42)	0.37 (0.03)

Tabela 6.8: Estimativas de β via Monte Carlo-PMDA ($n = 50$, 400 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	-0.01 (0.42)	0.41 (0.03)
2	1.08 (0.41)	0.38 (0.03)
3	1.16 (0.36)	0.33 (0.03)
4	0.02 (0.49)	0.48 (0.06)
5	1.07 (0.42)	0.43 (0.05)
7	-0.01 (0.39)	0.39 (0.03)
8	1.02 (0.39)	0.37 (0.03)

Tabela 6.9: Estimativas de β via Monte Carlo-PMDA ($n = 50$, 400 amostras, $m = 5$).

	$m = 5$	
#	$\hat{\beta}$	Erro padrão
1	0.01 (0.44)	0.41 (0.03)
2	1.02 (0.40)	0.38 (0.03)
3	1.16 (0.34)	0.33 (0.03)
4	0.02 (0.51)	0.48 (0.06)
5	1.07 (0.43)	0.43 (0.05)
6	1.10 (0.35)	0.35 (0.04)
7	0.00 (0.35)	0.39 (0.02)
8	1.08 (0.41)	0.37 (0.03)
9	1.14 (0.33)	0.32 (0.02)

Tabela 6.10: Estimativas de β via Monte Carlo-PMDA ($n = 25$, 400 amostras, $m = 5$).

	$m = 5$	
#	$\hat{\beta}$	Erro padrão
1	0.03 (0.60)	0.60 (0.07)
4	-0.02 (0.96)	0.94 (3.03)
7	-0.02 (0.75)	0.68 (2.00)

6.1.3 Resultados para o Algoritmo ANDA

Nas Tabelas 6.11, 6.12 e 6.13 verifica-se que os coeficientes de regressão estimados apresentam vício pequeno. A variabilidade destas estimativas se torna estável a medida que o número de imputações aumenta. No entanto, percebe-se uma ligeira mudança quando o tamanho amostral passa de 100 para tamanhos 50 e 25 aumentando o vício do coeficiente de regressão estimado.

Tabela 6.11: Estimativas de β via Monte Carlo-ANDA (n=100, 1000 amostras).

	$m = 2$		$m = 3$		$m = 4$	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	-0.01 (0.28)	0.28 (0.01)	0.01 (0.27)	0.28 (0.01)	-0.01 (0.28)	0.28 (0.02)
2	1.03 (0.27)	0.26 (0.01)	1.02 (0.27)	0.26 (0.01)	1.02 (0.26)	0.26 (0.01)
3	1.08 (0.25)	0.22 (0.02)	1.10 (0.24)	0.22 (0.02)	1.11 (0.25)	0.22 (0.02)
4	-0.03 (0.34)	0.33 (0.02)	-0.01 (0.33)	0.33 (0.02)	-0.00 (0.34)	0.33 (0.02)
5	1.03 (0.30)	0.30 (0.02)	1.04 (0.31)	0.30 (0.02)	1.04 (0.31)	0.30 (0.02)
6	1.08 (0.25)	0.24 (0.02)	1.10 (0.26)	0.24 (0.02)	1.10 (0.26)	0.24 (0.02)
7	0.01 (0.26)	0.27 (0.01)	0.00 (0.26)	0.26 (0.01)	-0.01 (0.27)	0.27 (0.01)
8	1.03 (0.27)	0.25 (0.01)	1.03 (0.26)	0.25 (0.01)	1.03 (0.27)	0.25 (0.01)
9	1.09 (0.24)	0.22 (0.02)	1.12 (0.24)	0.22 (0.02)	1.12 (0.25)	0.22 (0.02)

Tabela 6.12: Estimativas de β via Monte Carlo-ANDA (n=100, 1000 amostras).

	$m = 5$		$m = 6$		$m = 7$	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	-0.02 (0.29)	0.28 (0.01)	0.00 (0.27)	0.28 (0.01)	0.00 (0.28)	0.28 (0.01)
2	1.03 (0.27)	0.26 (0.01)	1.04 (0.27)	0.26 (0.01)	1.04 (0.28)	0.26 (0.01)
3	1.12 (0.25)	0.22 (0.02)	1.14 (0.25)	0.22 (0.02)	1.11 (0.25)	0.22 (0.02)
4	0.00 (0.33)	0.33 (0.02)	0.01 (0.34)	0.33 (0.02)	-0.01 (0.33)	0.33 (0.02)
5	1.04 (0.31)	0.30 (0.02)	1.06 (0.31)	0.30 (0.02)	1.02 (0.30)	0.29 (0.02)
6	1.12 (0.26)	0.24 (0.02)	1.11 (0.27)	0.24 (0.02)	1.11 (0.27)	0.24 (0.02)
7	0.01 (0.28)	0.27 (0.01)	-0.01 (0.28)	0.27 (0.01)	0.01 (0.26)	0.27 (0.01)
8	1.03 (0.27)	0.25 (0.01)	1.03 (0.27)	0.25 (0.01)	1.04 (0.28)	0.25 (0.01)
9	1.12 (0.24)	0.22 (0.02)	1.13 (0.25)	0.22 (0.02)	1.15 (0.25)	0.22 (0.02)

Tabela 6.13: Estimativas de β via Monte Carlo-ANDA (n=100, 1000 amostras).

	$m = 8$		$m = 9$		$m = 10$	
#	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão	$\hat{\beta}$	Erro padrão
1	0.00 (0.27)	0.28 (0.01)	0.01 (0.28)	0.28 (0.01)	-0.01 (0.29)	0.28 (0.01)
2	1.03 (0.29)	0.26 (0.01)	1.01 (0.27)	0.26 (0.01)	1.03 (0.28)	0.26 (0.01)
3	1.12 (0.25)	0.22 (0.02)	1.13 (0.26)	0.22 (0.02)	1.12 (0.25)	0.22 (0.02)
4	-0.01 (0.33)	0.33 (0.02)	0.01 (0.34)	0.33 (0.02)	0.01 (0.34)	0.33 (0.02)
5	1.04 (0.31)	0.30 (0.02)	1.05 (0.31)	0.30 (0.02)	1.05 (0.31)	0.30 (0.02)
6	1.10 (0.26)	0.24 (0.02)	1.11 (0.26)	0.24 (0.02)	1.11 (0.26)	0.24 (0.02)
7	0.01 (0.27)	0.27 (0.01)	0.00 (0.27)	0.27 (0.01)	0.00 (0.28)	0.27 (0.01)
8	1.03 (0.27)	0.25 (0.01)	1.05 (0.26)	0.25 (0.01)	1.05 (0.30)	0.25 (0.01)
9	1.13 (0.24)	0.22 (0.02)	1.12 (0.25)	0.22 (0.02)	1.12 (0.25)	0.22 (0.02)

No caso 3, observa-se que quando o número de imputações é igual a 5, o coeficiente de regressão estimado aumenta, diminuindo assim a sua precisão. Também pode-se observar que quanto menor o tamanho amostral maior será a variabilidade e os desvios-padrões estimados de Monte Carlo apesar deles não serem tão diferentes dos erros-padrões estimados (Tabela 6.15).

Tabela 6.14: Estimativas de β via Monte Carlo-ANDA ($n = 50$, 1000 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	-0.02 (0.40)	0.41 (0.03)
2	1.05 (0.40)	0.38 (0.03)
3	1.14 (0.35)	0.32 (0.03)
4	0.00 (0.48)	0.48 (0.05)
6	1.13 (0.36)	0.35 (0.04)
7	0.01 (0.38)	0.39 (0.03)
8	1.07 (0.41)	0.37 (0.03)
9	1.05 (0.34)	0.32 (0.03)

Tabela 6.15: Estimativas de β via Monte Carlo-ANDA ($n = 25$, 400 amostras, $m = 5$).

	$m = 5$	
#	$\hat{\beta}$	Erro padrão
1	0.01 (0.63)	0.61 (0.08)
3	1.17 (0.49)	0.48 (0.08)
7	-0.01 (0.66)	0.58 (0.08)

Tabela 6.16: Estimativas de β via Monte Carlo-ANDA ($n = 50$, 400 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	-0.02 (0.38)	0.41 (0.03)
2	1.04 (0.40)	0.38 (0.03)
3	1.14 (0.36)	0.32 (0.03)
4	0.01 (0.48)	0.48 (0.06)
5	1.10 (0.46)	0.43 (0.05)
6	1.12 (0.36)	0.35 (0.04)
7	-0.01 (0.40)	0.39 (0.03)
8	1.09 (0.42)	0.37 (0.03)
9	1.13 (0.35)	0.32 (0.03)

Tabela 6.17: Estimativas de β via Monte Carlo-ANDA ($n = 50$, 400 amostras, $m = 5$).

	$m = 5$	
#	$\hat{\beta}$	Erro padrão
1	0.00 (0.42)	0.41 (0.03)
2	1.03 (0.40)	0.38 (0.03)
3	1.13 (0.33)	0.32 (0.03)
4	0.04 (0.51)	0.48 (0.05)
5	1.07 (0.44)	0.43 (0.04)
6	1.08 (0.37)	0.35 (0.04)
7	0.01 (0.41)	0.39 (0.03)
8	1.05 (0.38)	0.37 (0.02)
9	1.14 (0.34)	0.32 (0.03)

6.1.4 Resultados utilizando a Suavização por Núcleo Estimadores

Utilizando a metodologia proposto por Silverman (1986) para escolha da janela ótima, h, observa-se que os resultados não são muito desejáveis quando o tamanho de amostra é pequena pois os coeficientes de regressão estimados apresentam vício maior e a variabilidade é mais elevada do que aqueles resultados encontrados pelos algoritmos PMDA e ANDA.

Na Tabela 6.20 pode-se observar que no caso 4, apesar do coeficiente de regressão estimado estar bem próximo do verdadeiro valor, sua variabilidade é grande e o desvio-padrão estimado de Monte Carlo é bem diferente do erro-padrão estimado.

Destaca-se também que nos casos 3, 6 e 9, os coeficientes de regressão estimados apresentam valores bem mais elevados que os outros casos (Tabela 6.19 e 6.20).

Tabela 6.18: Estimativas de β via Monte Carlo por Suavização ($n = 25$, 1000 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	0.00 (0.62)	0.60 (0.08)
3	1.30 (0.68)	0.52 (0.11)

Tabela 6.19: Estimativas de β via Monte Carlo por Suavização ($n = 50$, 400 amostras, $m = 10$).

	$m = 10$	
#	$\hat{\beta}$	Erro padrão
1	0.00 (0.40)	0.41 (0.04)
2	1.05 (0.40)	0.38 (0.03)
3	1.18 (0.43)	0.33 (0.04)
4	0.02 (0.46)	0.48 (0.05)
5	1.05 (0.46)	0.43 (0.05)
6	1.19 (0.48)	0.36 (0.05)
7	-0.01 (0.42)	0.39 (0.03)
8	1.02 (0.38)	0.36 (0.03)
9	1.20 (0.42)	0.33 (0.04)

Tabela 6.20: Estimativas de β via Monte Carlo por Suavização ($n = 25$, 400 amostras, $m = 5$).

	$m = 5$	
#	$\hat{\beta}$	Erro padrão
1	0.02 (0.64)	0.61 (0.09)
3	1.25 (0.64)	0.52 (0.10)
4	-0.02 (1.16)	1.21 (4.20)
6	1.30 (0.77)	0.57 (0.14)
7	0.02 (0.58)	0.57 (0.06)
9	1.27 (0.69)	0.52 (0.11)

Tabela 6.21: Estimativas de β via Monte Carlo por Suavização ($n = 50$, 400 amostras, $m = 5$).

	$m = 5$	
#	$\hat{\beta}$	Erro padrão
1	0.00 (0.42)	0.41 (0.03)
2	1.04 (0.37)	0.38 (0.02)
3	1.17 (0.41)	0.33 (0.04)
4	0.02 (0.49)	0.49 (0.06)
5	1.07 (0.47)	0.43 (0.05)
6	1.15 (0.43)	0.36 (0.05)
7	0.04 (0.38)	0.39 (0.03)
8	1.04 (0.40)	0.37 (0.03)
9	1.23 (0.42)	0.34 (0.05)

6.2 Comparação dos Métodos

Nesta seção, serão apresentados os resultados graficamente para a comparação do desempenho dos seguintes métodos: Algoritmo PMDA, Algoritmo ANDA e suavização por Núcleo Estimador. Todos os resultados foram baseados na geração de 400 amostras de tamanhos 50 e 25.

A Figura 6.1 apresenta os coeficientes de regressão estimados com seus respectivos erros-padrões estimados. Nota-se que não há quase nenhuma diferença quando comparamos os resultados entre os três métodos em estudo pois eles se aproximam do verdadeiro valor do coeficiente de regressão do modelo de Cox, mesmo para tamanhos de amostra 25 e 50. Na Figura 6.2, os coeficientes de regressão estimados são quase não viciados e os erros-padrões apresentam pouca variabilidade. Observa-se que quando o número de imputações é igual a 10, o algoritmo PMDA tem coeficiente de regressão estimado mais elevado do que os outros.

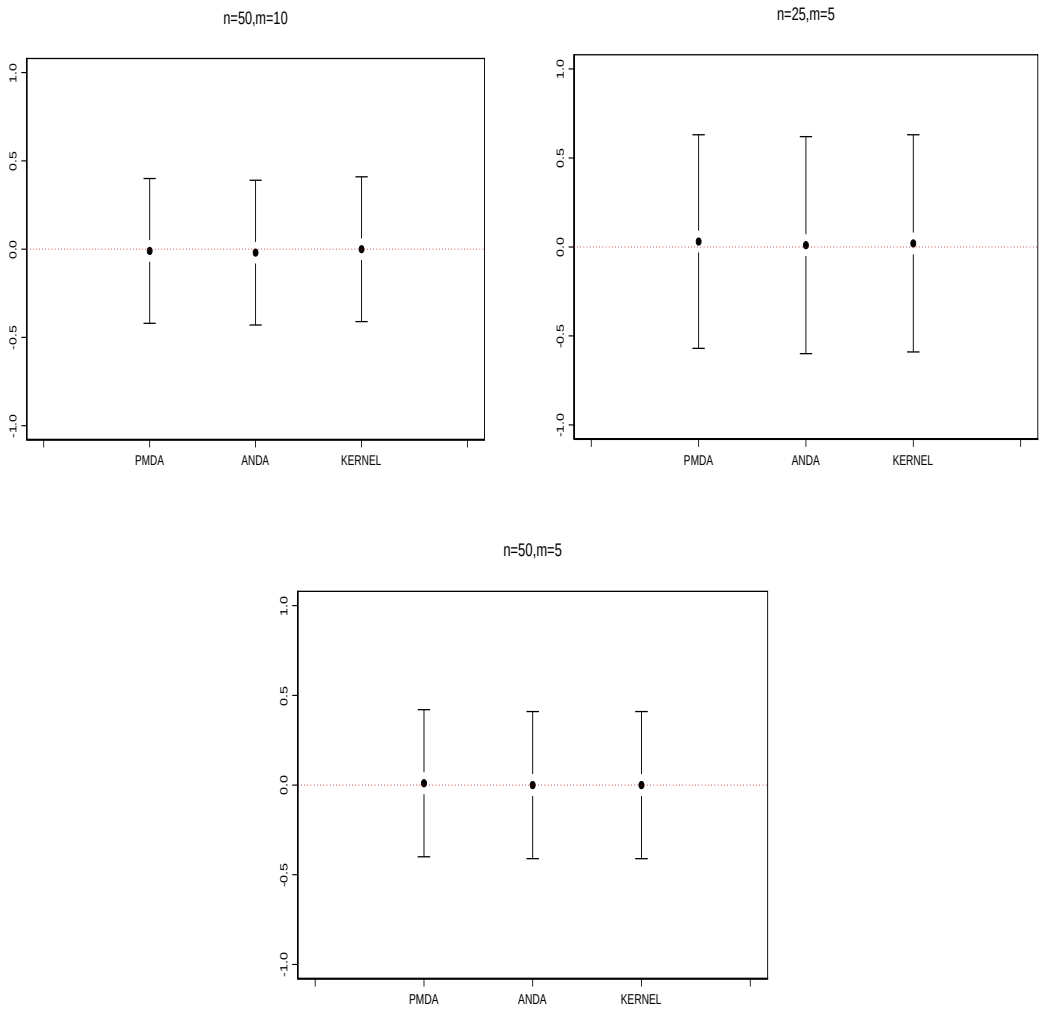

Figura 6.1: "Estimativas de β com seus respectivos erros-padrões (# 1)".

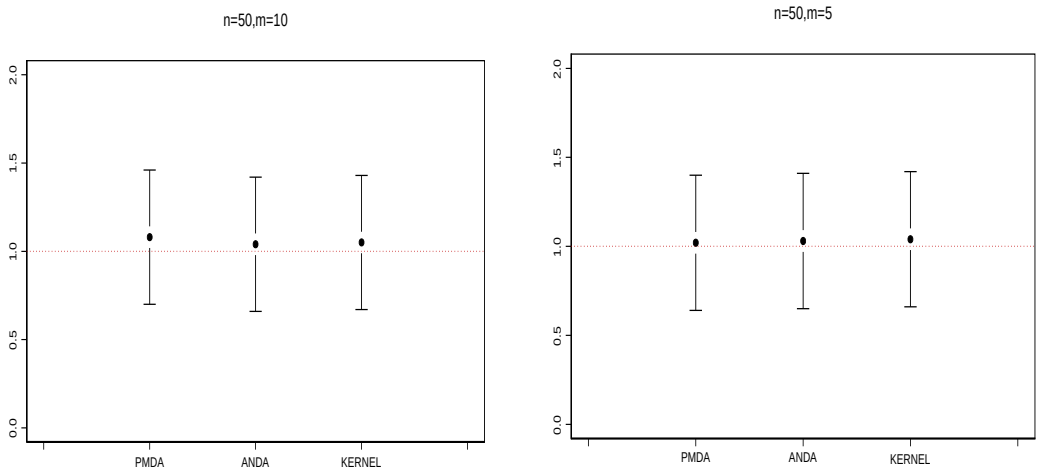

Figura 6.2: "Estimativas de β com seus respectivos erros-padrões (# 2)".

Pode-se observar que no caso 3, os coeficientes de regressão estimados são viciados e que os erros-padrões dos métodos em estudo não variam muito entre eles, mesmo mudando o número de imputação de 10 para 5. Também pode ser verificado que os resultados do algoritmo PMDA e do método pelo núcleo estimador são mais próximos que do algorimo ANDA (Figura 6.3).

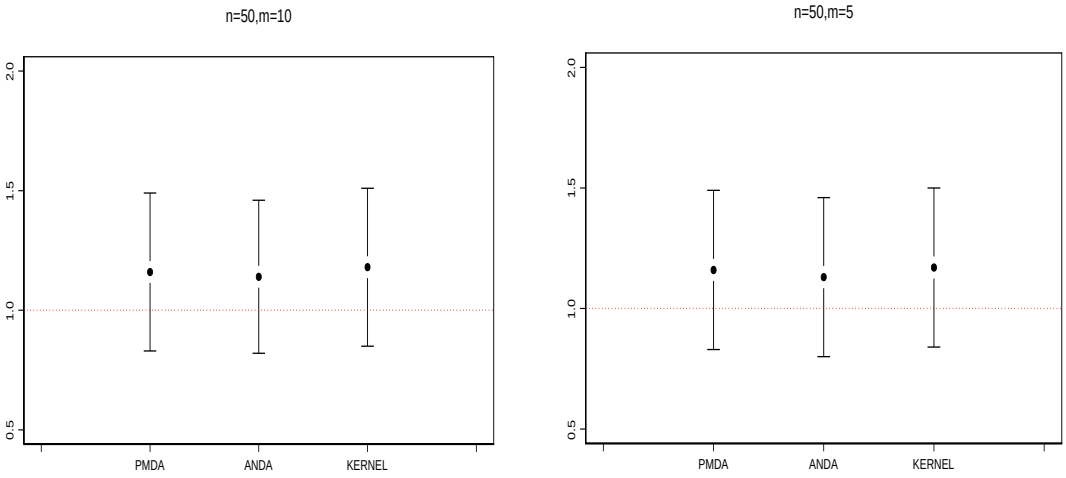

Figura 6.3: "Estimativas de β com seus respectivos erros-padrões (# 3)".

A Figura 6.4 apresenta os coeficientes de regressão estimados que são ligeiramente viciados e seus erros-padrões têm uma variabilidade similar entre os métodos que estão em análise. Na Figura 6.5, observa-se que os algoritmos PMDA e ANDA, e o método pelo núcleo estimador apresentam coeficientes de regressão estimados com vício pequeno.

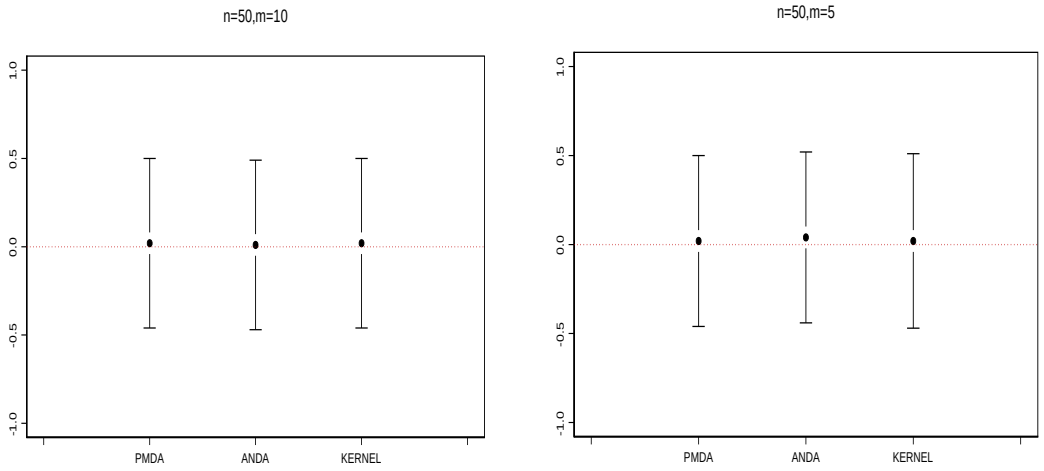

Figura 6.4: "Estimativas de β com seus respectivos erros-padrões (# 4)".

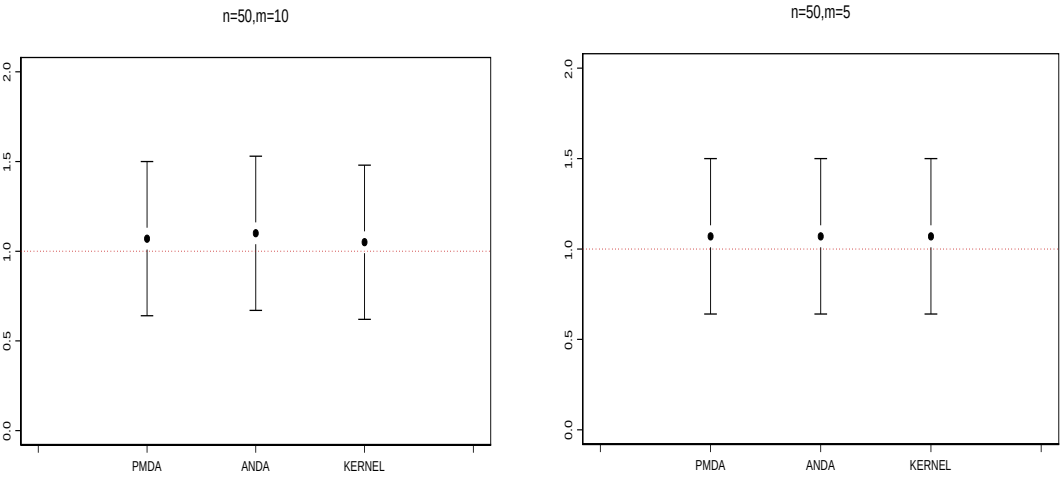

Figura 6.5: "Estimativas de β com seus respectivos erros-padrões (# 5)".

No caso 6, observa-se que os coeficientes de regressão e erros-padrões estimados dos algoritmos PMDA e ANDA apresentam semelhança, enquanto que pelo método do núcleo estimador mostra que o coeficiente de regressão estimado tem um vício ligeiramente mais elevado do que os resultados encontrados utilizando os algoritmos PMDA e ANDA (Figura 6.6).

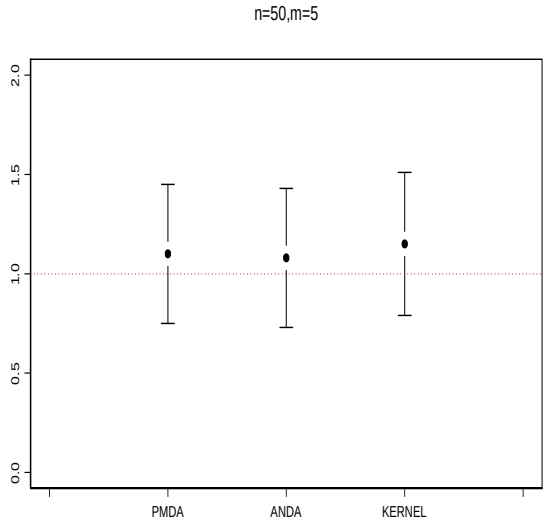

Figura 6.6: "Estimativas de β com seus respectivos erros-padrões (# 6)".

A Figura 6.7 mostra que nas 400 amostras de tamanho 25 e 50 para número de imputação $m = 5, 10$, os coeficientes de regressão estimados estão bem próximos do valor verdadeiro, apresentando vício pequeno. No entanto, quando o tamanho da amostra é igual a 25, os erros-padrões são mais elevados com variabilidade muito grande. Considerando o caso 8, pode-se verificar que o algoritmo ANDA tem o coeficiente de regressão estimado um pouco mais viciado que os outros métodos, quando o número de imputações é igual a 10 (Figura 6.8).

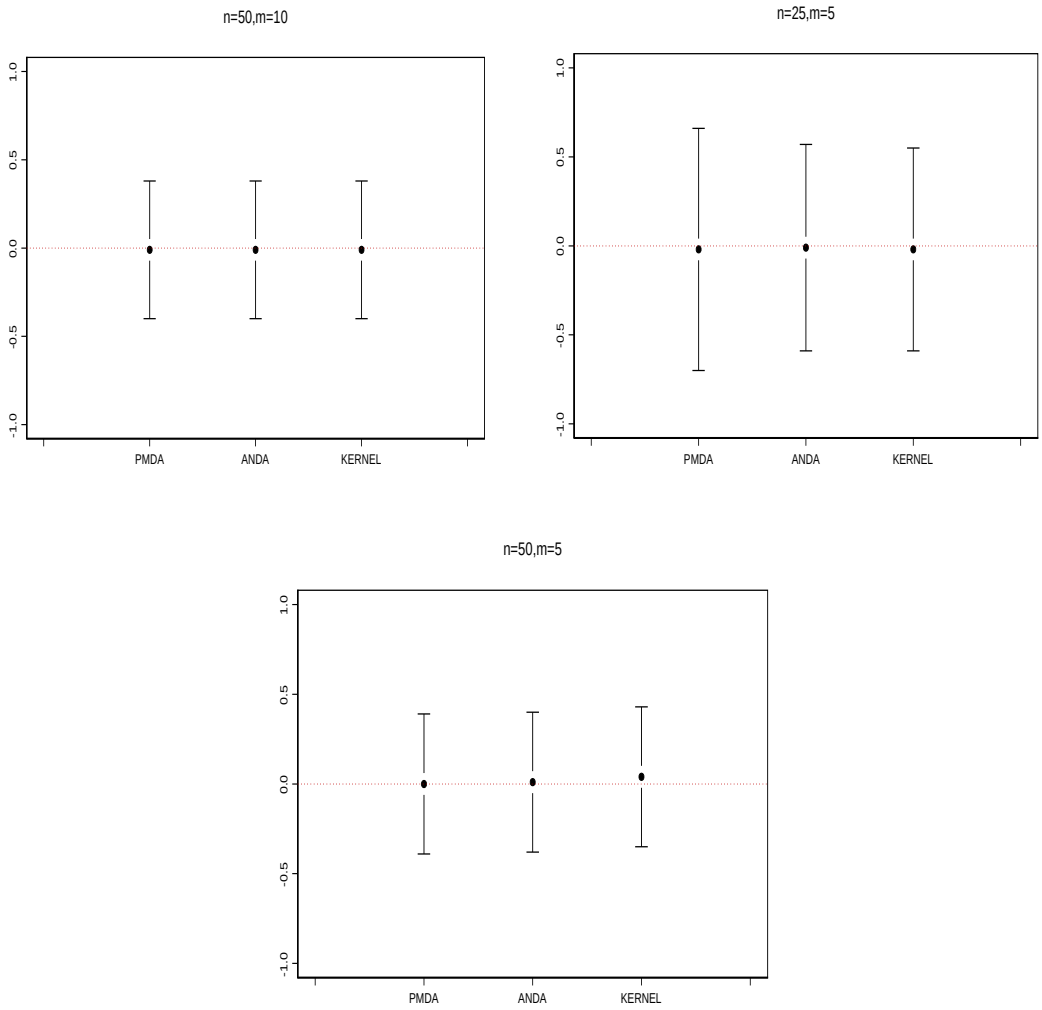

Figura 6.7: "Estimativas de β com seus respectivos erros-padrões (# 7)".

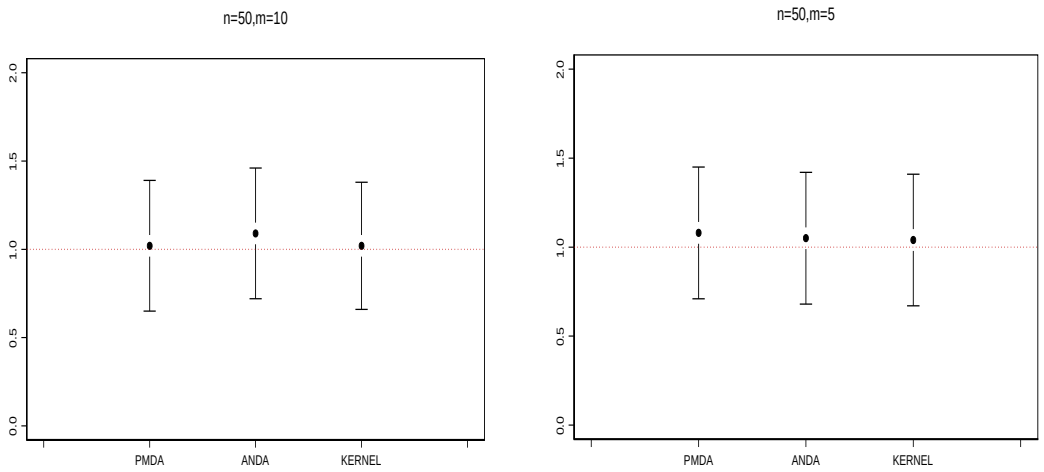

Figura 6.8: "Estimativas de β com seus respectivos erros-padrões (# 8)".

Na Figura 6.9 observamos que os valores dos coeficientes de regressão estimados estão bastante afastados de seu valor verdadeiro, ou seja, são viciados, aumentando a sua variabilidade. Nota-se que o método pelo núcleo estimador destaca-se pois o vício e sua variabilidade são maiores que o dos algoritmos PMDA e ANDA.

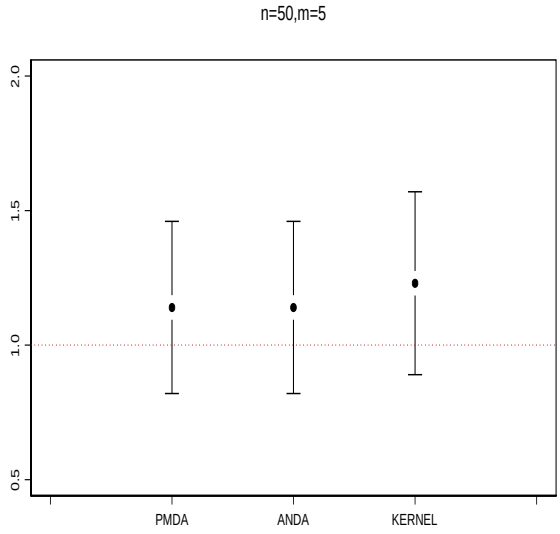

Figura 6.9: "Estimativas de β com seus respectivos erros-padrões (# 9)".

No intuito de verificar o problema de convergência, foi feito um estudo nos métodos em estudo. Para ilustrar, utilizamos um critério de convergência, $|\hat{\beta}^{(j+1)} - \hat{\beta}^{(j)}| < 10^{-6}$ ou $j > 500$, diferente daquele proposto por Pan (2000b). Os resultados parecem indicar que os algoritmos PMDA e ANDA, e o método pelo núcleo estimador não convergem.

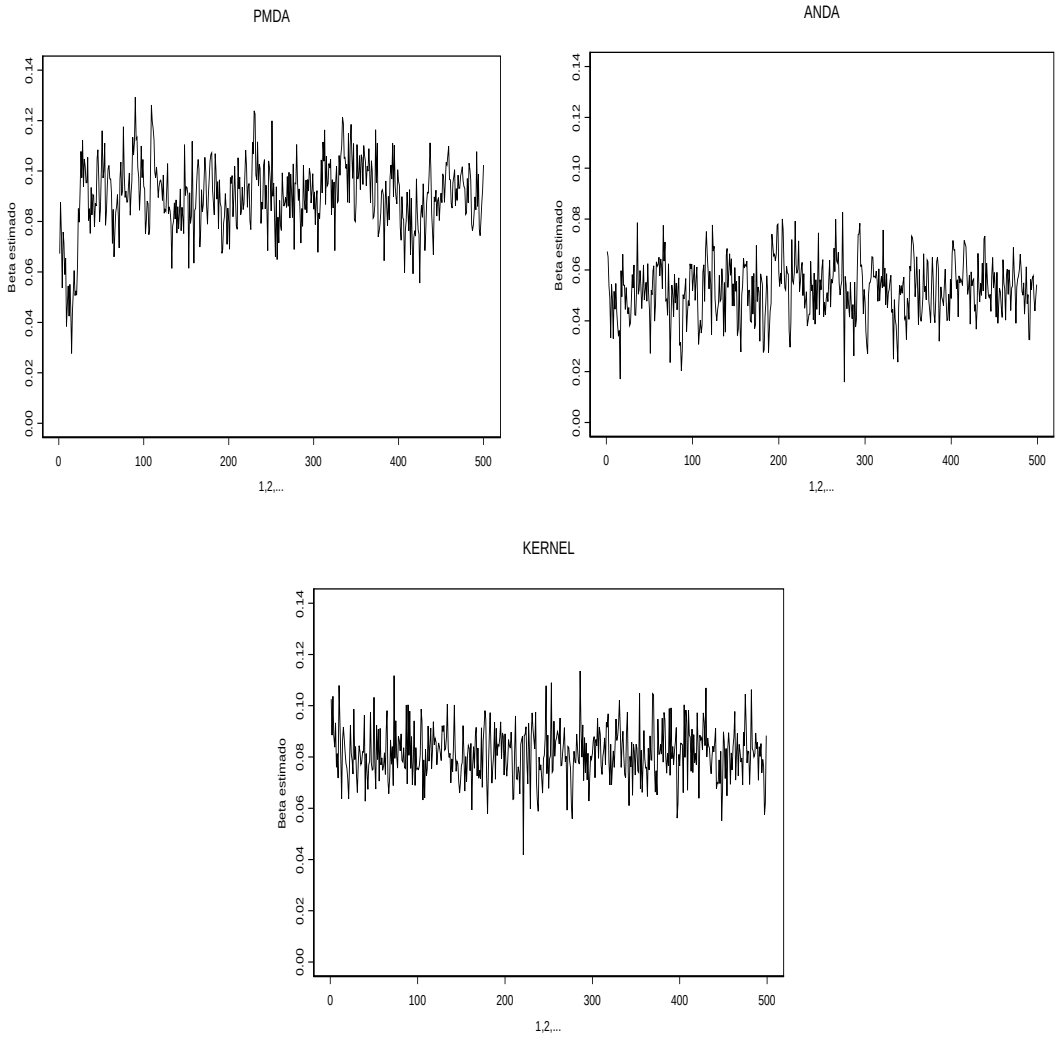

Figura 6.10: "Beta Estimado dos métodos em estudo".

A Figura 6.11 apresenta o número de pontos de salto a cada iteração para tamanhos de amostra 20 e 100, com número de imputações 5 e 10 para o algoritmo PMDA. Quando o tamanho de amostra é igual a 100, os pontos de salto aumentam, enquanto para o tamanho 20 diminui o número dos pontos de salto. Quanto ao número de imputações, este parece ter pouca influência. Foi com intuito de eliminar o problema da diminuição progressiva nos pontos de salto a cada iteração que utilizamos a suavização da curva de sobrevivência base estimada na metodologia proposta por Pan (2000b).

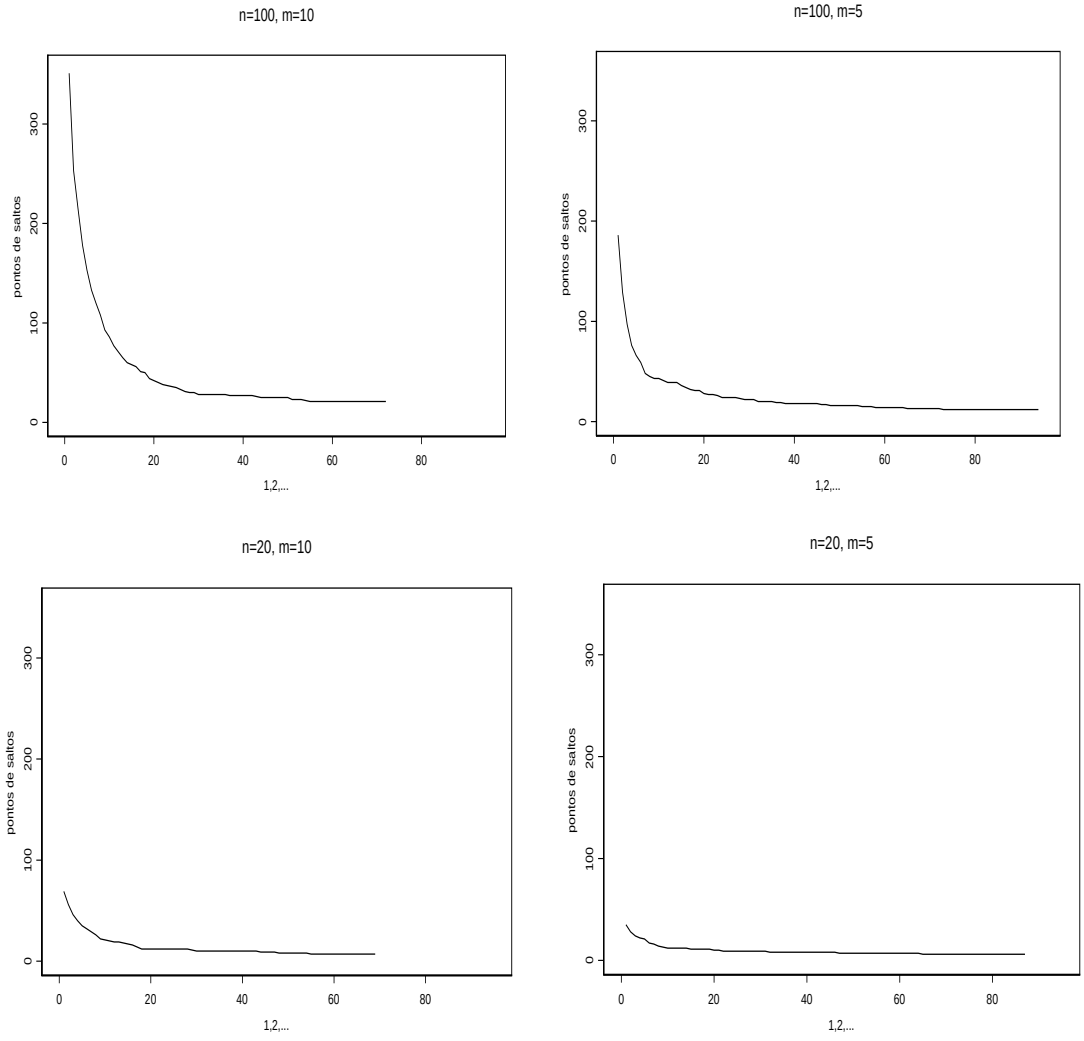

Figura 6.11: "Pontos de salto para o algoritmo PMDA".

6.3 Aplicação

Foi realizado um estudo retrospectivo considerando os dados do estudo de câncer de mama (Finkelstein e Wolfe, 1985; Finkelstein, 1986). Haviam 94 pacientes de câncer de mama que tinham sido tratados por dois tratamentos após tumorectomia, isto é, os pacientes tratados somente com radioterapia e os tratados com radioterapia primária com quimioterapia. O objetivo deste estudo é investigar qual tratamento tem melhor efeito "cosmético" à longo prazo, ou seja, comparar os pacientes que receberam a radioterapia primária com quimioterapia àqueles que receberam somente o tratamento de radioterapia para determinar se a quimioterapia afeta a taxa de deterioração do estado cosmético. Neste estudo, os pacientes foram assistidos em visitas na clínica a cada 4 a 6 meses e para aqueles que viviam em lugares mais afastados da clínica os intervalos de seguimento foram frequentemente mais longos. Tomamos o grupo somente com radioterapia como o grupo de referência com covariável $Z_i = 0$, e radioterapia primária com quimioterapia com $Z_i = 1$. Nossa aproximação da imputação múltipla proposta usando PMDA e ANDA produz os coeficientes de regressão (erro-padrão) 0.87 (0.29) e 0.86 (0.29), respectivamente. Quando utilizamos o método do núcleo estimador, a estimativa do coeficiente de regressão do modelo de Cox é 0.88 com erro-padrão estimado igual a 0.29. Percebe-se que os resultados encontrados pelos algoritmos e suavização por núcleo estimador são bem próximos. Já os resultados encontrados por Pan (2000b) produz, para os algoritmos PMDA e ANDA, 0.90 (0.29) e 0.92 (0.29), respectivamente. Portanto, podemos concluir que os nossos resultados não diferem significativamente daqueles encontrados por Pan (2000b).

Capítulo 7

Conclusões

Foi proposto usar imputação múltipla como uma metodologia geral para regressão de Cox com dados de censura intervalar, com objetivo de verificar o desempenho dos algoritmos PMDA e ANDA, comparando os resultados quanto a vício e variabilidade para vários tamanhos de amostra, com os resultados encontrados por Pan (2000b), além de suavizar a curva de sobrevivência base estimada através do núcleo estimador a fim de eliminar o problema de diminuição nos pontos de salto para amostras pequenas.

A partir dos métodos propostos, de suas aplicações a dados reais e dos estudos de simulações, as seguintes conclusões foram obtidas:

• Foi utilizado um procedimento estatístico padrão que é fácil de implementar para dados com censura à direita, tal como o método de verossimilhança parcial;

• Através da simulação, observamos que os algoritmos PMDA e ANDA apresentam bons resultados estimando o coeficiente de regressão no modelo de riscos proporcionais para amostras médias e pequenas;

• A suavização pelo núcleo estimador apresenta bons resultados, no entanto, mostra-se inferior aos algoritmos PMDA e ANDA. Uma possível explicação para isso é o fato de que, o coeficiente de regressão e seu erro-padrão estimado apresentarem vício e variabilidade elevados;

• Para os métodos apresentados, o número de imputações não pareceu ter grande influência, mesmo modificando o tamanho da amostra;

• O número de pontos de saltos a cada iteração para vários tamanhos de amostra decresce significativamente, mas os resultados indicam que isto não afeta as estimativas obtidas com os algoritmos PMDA e ANDA. Quanto maior o tamanho de amostra melhor será a curva de sobrevivência base para o algoritmo PMDA;

• Em alguns casos, o programa no S-plus não gerou alguns betas, apresentando problemas de convergência no ajuste do modelo de riscos proporcionais de Cox;

• Utilizamos um critério bem menor do que aquele proposto por Pan (2000b) e verificamos que existe o problema de convergência mesmo aumentando o número de iterações para todos os algoritmos e tamanhos de amostra.

Referências Bibliográficas

[1] AALEN, O.O. (1978). *Nonparametric Inference for a Family of Counting Processes.* Annals of Statistics, vol. 6, pp. 701-726.

[2] BRESLOW, N.E., CROWLEY, J. (1974). *A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship.* Annals of Statisitics, vol. 2, pp. 437-453.

[3] BARLOW, R.E., BARTHOLOMEW, D.J., and BRUNK, H.D. (1972). *Statistical Inference under Order Restrictions-The Theory and Application of Isotonic Regression.* John Wiley & Sons.

[4] BEBCHUCK, J.D., and BETENSKY, R.A. (2000). *Multiple Imputation for Simple Estimation of the Hazard Function Based on Interval Censored Data.* Statistics in Medicine, vol. 19, pp. 405-419.

[5] BROKMEYER, R., and GOEDERT, J.J. (1989). *Censoring in an Epidemic with an Application to Hemophilia-Associated AIDS.* Biometrics, vol. 45, pp. 325-335.

[6] CHANG, M.N. (1990). *Weak Convergence of a Self-Consistent Estimator of the Survival Function With Doubly Censored Data.* Annals of Statistics, vol. 18, pp. 391-404.

[7] COLLET, A. (1994). *Modelling Survival Data in Medical Research.* Chapman and Hall, London.

[8] COLOSIMO, E.A. (2001). *Análise de Sobrevivência Aplicada.* 46ª Reunião Anual da RBRAS e 9º SEAGRO. Departamento de Estatística, Belo Horizonte, UFMG.

[9] COX, D.R. (1972). *Regression Models and Life-Tables (with discussion).* Journal of the Royal Statistical Society, Series B, vol. 34, pp. 187-220.

[10] COX, D.R. (1975). *Partial Likelihood.* Biometrika, March, vol. 62, pp. 269-276.

[11] DEMPSTER, A.P., LAIRD, N. M., and RUBIN, D. B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, Series B, vol. 39, pp. 1-38.

[12] DIGGLE, P., and MARRON, J.S. (1988). *Equivalence of Smoothing Parameter Selectors in Density and Intensity Estimation.* Journal of the American Statistical Association, vol. 83, pp. 793-800.

[13] EFRON, B. (1977). *The Eficiency of Cox's Likelihood Function for Censored Data.* Journal of the American Statistical Association, vol. 72, pp. 557-565.

[14] FINKELSTEIN, D.M., and WOLFE, R.A. (1985). *A Semiparametric Model for Regression Analysis of Interval Censored Failure Time Data.* Biometrics, vol. 41, pp. 933-945.

[15] FINKELSTEIN, D.M. (1986). *A Proportional Hazards Model of Interval Censored Failure Time Data.* Biometrics, vol. 42, pp. 845-854.

[16] GENTLEMAN, R., and GEYER, C.J. (1994). *Maximum Likelihood for Interval Censored Data: Consistency and Computation.* Biometrika, vol. 81, pp. 618-623.

[17] GIOLO, S.R. (2003). *Introdução à Análise de Sobrevivência.* Relatório Técnico, Departamento de Estatística da UFPA.

[18] GOETGHEBEUR, E., and RYAN, L. (2000). *Semiparametric Regression Analysis of Interval Censored Data.* Biometrics, vol. 56, pp. 1139-1144.

[19] GOGGINS, W.B., FINKELSTEIN, D.M., SCHOENFELD, D.A., and ZASLAVSKY, A.M. (1998). *A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Failure Data under the Cox Proportional Hazards Model.* Biometrics, vol. 54, pp. 1498-1507.

[20] GROENEBOOM, P., and WELLNER, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birhauser Verlag.

[21] HAN, A.K. (1987). *Nonparametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimation.* Journal of Econometrics, vol. 35, pp. 303-316.

[22] HUANG, J., and WELLNER, J.A. (1995). *A Interval-Censored Failure Data: A review of recent progress. In Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis.* D.Y.Lin and T.R.Fleming (eds), New York: Springer-Verlag, pp. 123-169.

[23] HUANG, J., and ROSSINI, A.J. (1997). *Sieve Estimation for the Proportional Odds Failure Time regression Model with Interval Censoring.* Journal of the American Statistical Association, vol. 92, pp. 960-967.

[24] HONDA, T. (2004). *Nonparametric Regression with Current Status Data.* Annals of the Institute of Statistical Mathematics, vol. 56, pp. 49-72.

[25] JONGBLOED, G. (1995). *Three Statistical Inverse Problems.* PhD Dissertation, Delft University of Technology, Nertherlands.

[26] JONGBLOED, G. (1998). *The Iterative Convex Minorant Algorithm for Nonparametric Estimation.* Journal of Computational and Graphical Statistics, vol. 7, pp. 310-321.

[27] KALBLEISCH, J.D., and PRENTICE, R.L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

[28] KAPLAN, E.L., and MEIER, P. (1958). *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association, vol. 53, pp. 457-481.

[29] KLEIN, J.P., and MOESCHBERGER, M. (1997). *Survival Analysis.* New York: Springer Verlag.

[30] KOOPERBERG, C., STONE, C.J., and TRUONG, Y.K. (1995a). *Hazard Regression.* journal of the American Statistical Association, vol. 90, pp. 78-94.

[31] KOOPERBERG, C., and CLARKSON, D.B. (1997). *Hazards Regression with Interval Censored Data.* Biometrics, vol. 53, pp. 1485-1494.

[32] LAMPORT, L. (1986). *Latex - A Document Preparation System.* Addison-Wesley Publishing Company.

[33] LAW, C.G., and BROOKMEYER, R. (1992). *Effects of Mid-point Imputation on the Analysis of Doubly Censored Data.* Statistics in Medicine, vol. 11, pp. 1569-1578.

[34] LAWLESS, J.F. (1982). *Statistical Models and Methods for lifetime Data.* John Wiley and Sons, New York.

[35] LEE, E. (1980). *Statistical Methods for Survival Data Analysis.* Belmont: Lifetime Learning Publications.

[36] LI, G., and ZHANG, C.H. (1998). *Linear Regression with Interval Censored Data.* The Annals of Statistics, vol. 26, pp. 1306-1327.

[37] LIN, D.Y., OAKES, D., and YING, Z. (1998). *Additive Hazards Regression with Current Status Data.* Biometrika, vol. 85, pp. 289-298.

[38] LINDSEY, J.C., and RYAN, L.M. (1998) *Tutorial in bioestatistics: Methods for Interval Censored Data.* Statistics in Medicine, vol. 17 pp. 219-238.

[39] NELSON, W. (1972). *Theory and Application of Hazard Plotting for Censored Failure Data.* Technometrics, vol. 14, pp. 945-965.

[40] PAN, W. (1999). *Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval Censored Data.* Journal of Computational and Graphical Statistics, vol. 8, pp. 109-120.

[41] PAN, W. (2000a). *Smooth Estimation of the Survival Function for Interval Censored Data.* Statistics in Medicine, vol. 19, pp. 2611-2624.

[42] PAN, W. (2000b). *A Multiple Imputation Approach to Cox Regression with Interval Censored Data.* Biometrics, vol. 56, pp. 199-203.

[43] PAN, W., and CHAPPELL, R. (2002). *Estimation in the Cox Proportional Hazards Model with Left-Truncadet and Interval Censored Data.* Biometrics, vol. 58, pp. 64-70.

[44] PETO, R. (1973). *Experimental Survival Curves for Interval Censored Data.* Applied Statistics, vol. 22, pp. 86-91.

[45] PRENTICE, R.L. (1978). *Linear Rank Tests with Right Censored Data.* Biometrika, vol. 65, pp. 167-179.

[46] RABINOWITZ, D., TSIATIS, A., and ARAGON, J. (1995). *Regression with Interval Censored Data.* Biometrika, vol. 82, pp. 501-513.

[47] RABINOWITZ, D., BETENSKY, R.A., and TSIATIS, A.A. (2000). *Using Conditional Logistic Regression to Fit Proportional Odds Models to Interval Censored Data.* Biometrics, vol. 56, pp. 511-518.

[48] RAO, C.R. (1973). *Linear Statistical Inference and Its Applications.* 2º Edição New York: Wiley.

[49] RUBIN, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

[50] SATTEN, G.A. (1996). *Rank based Inference in the Proportional Hazards Model for Interval Censored Data.* Biometrika, vol. 83, pp. 355-370.

[51] SATTEN, D.B., DATTA, S., and WILLIAMSON, J.M. (1998). *Inference based on Imputed Failure Times for the Proportional Hazards Model with Interval Censored Data.* Journal of the American Statistical Association, vol. 93, pp. 318-327.

[52] SCHENKER, N. and WELSH, A.H. (1998). *Asymptotic Results for Multiple Imputation.* Annals of Statistics, vol. 16, pp. 1550-1566.

[53] SELF, S.G., and GROSSMAN, E.A. (1986). *Linear Rank Tests for Interval Censored Data with Application to PCB Levels in Adipose Tissue of Transformer Repair Workers.* Biometrics, vol. 42, pp. 521-530.

[54] SHERMAN, R.P., (1993). *The Limiting Distribution of the Maximum Rank Correlation Estimator.* Econometrica, vol. 61, pp. 123-137.

[55] SHIBOSKI, S.C. (1998). *Generalized Additive Models for Current Status Data.* Lifetime Data Analysis, vol. 4, pp. 29-50.

[56] SILVERMAN, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman & Hall.

[57] SINHÁ, D., TANNER, M.A., and HALL, W.J. (1994). *Maximizing the Marginal Likelihood from Grouped Survival Data.* Biometrika, vol. 81, pp. 53-60.

[58] SUN, J. (1998). *Interval Censoring.* Encyclopedia of Biostatistics, First Edition, John Wiley & Ltd., pp. 2090-2095.

[59] TAYLOR, J.M.G, MURRAY, S., and HSU, C.H. (2002). *Survival Estimation and Testing via Multiple Imputation.* Statisitcs & Probability Letters, vol. 58, pp. 221-232.

[60] TANNER, M.A., and WONG, W.H. (1987a). *The Calculation of Posterior Distributions by Data Augmentation.* Journal of the American Statistical Association, vol. 82, pp. 528-549.

[61] TANNER, M.A., and WONG, W.H. (1987b). *An Application of Imputation to an Estimation Problem in Grouped Lifetime Analysis.* Technometrics, vol. 29, pp. 23-32.

[62] TRAVASSOS, A.P.A. (2003). *Problemas de Fronteira dos Núcleo Estimadores e suas Abordagens.* Dissertação de Mestrado, Departamento de Estatística da UFMG.

[63] TURNBULL, B.W. (1976). *The Empirical Distribution Function with Arbitralily Grouped, Censored and Truncated Data.* journal of the Royal Statistical Society, Series B, vol. 38, pp. 290-295.

[64] WEI, G.C.G., and TANNER, M.A. (1991). *Applications of Multiple Imputation to the Analysis of Censored Regression Data.* Biometrics, vol. 47, pp. 1297-1309.

Livros Grátis

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

Milhares de Livros para Download:

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