# MODELOS DE RISCOS PROPORCIONAIS E ADITIVOS PARA O TRATAMENTO DE COVARIÁVEIS DEPENDENTES DO TEMPO

TARCIANA LIBERAL PEREIRA

Orientador: Prof. Dr. Enrico Antônio Colosimo
Orientadora: Profa. Dra. Maria Cristina Falcão Raposo
Área de concentração: Estatística Aplicada

Dissertação submetida como requerimento parcial para obtenção do grau de Mestre em Estatística pela Universidade Federal de Pernambuco

Recife, janeiro de 2004

Dedico este trabalho aos meus pais Agamenon e Vilma, as minhas irmãs Keyla e Dayse, a Luiz e a Kamylle.

# Agradecimentos

- A Deus por me iluminar e fortalecer em todos os momentos da minha vida.

- Aos meus pais pela criação maravilhosa, por estarem sempre presentes na minha vida, me incetivando, orientando e me apoiando.

- A Luiz por sempre acreditar em mim, pelo companheirismo me acolhendo e incentivando nas horas boas e nas difíceis, pela segurança que tenho ao seu lado e por seu amor que tanto me fortalece.

- A minhas irmãs por toda força, incentivo e carinho que me deram e que tanto me ajudou.

- A minha sobrinha linda Kamylle por todo carinho com que sempre me recebeu, pela alegria transmitida me dando força para continuar.

- À tia Zezé, tio Luizito, Vovó e Poly por me acolherem, pelo crédito e proteção e por tudo que sempre fizeram por mim.

- À professora Maria Cristina pela sua dedicada orientação, amizade e confiança em mim depositada. Por sempre nos mostrar que independente do obstáculo podemos sempre ir em frente, pelo seu exemplo de força e grandioso espírito de solidariedade humana.

- Ao professor Enrico Colosimo pela confiança em mim depositada, pela orientação segura e objetiva e por toda atenção dispensada no desenvolvimento desta dissertação.

- Aos meus eternos amigos da graduação Tati, Rodrigo, Cézar, Paty, Ângela, Juliana e Andréa por todos os momentos e por todas as experiências vivenciadas.

- Aos queridos amigos do mestrado Tati, Sílvia, Paty, Keila, Gilson, Moisés, Felipe, Raydonal, João Marcelo e Bartolomeu por juntos termos superados as dificuldades e fortalecido ainda mais nossos conhecimentos. Por todos os momentos que passamos, pelas experiências trocadas e pelas lições aprendidas com cada um.

- À professora Jacira Guiro por todo conhecimento passado durante a iniciação científica, pela convivência maravilhosa, pela amizade, por toda força e por tudo que me ensinou como aluna e pessoa.

- Ao professor Enivaldo Carvalho da Rocha pelo estímulo, confiança e por ter me incentivado a fazer parte do Diretório Acadêmico de Estatística onde tive um imenso aprendizado.

- À professora Cláudia, professor Manoel e o professor Sylvio pela preocupação demonstrada, incentivo, amizade e por suas contribuições à minha formação como estatística.

- Aos professores Edmilson, Viviana, André Toom, Isaac, Klaus, Cribari, Audrey, Francisco, Cristiano, Carla e Getúlio pelos conhecimentos transmitidos.

- À Adriana, Cícero, Tonho e Valéria pela competência profissional, amizade e por todo apoio dado ao longo das minhas atividades acadêmicas.

- Ao professor Pedro Israel do Departamento de Nutrição pela concessão do banco de dados utilizado no Capítulo 5 desta dissertação e pelas sugestões e esclarecimentos prestados para sua utilização.

- Aos amigos da graduação em especial a Catarina, Hugo, Hemílio, Manú e Oscar.

- Ao pessoal da turma nova do mestrado: Tatiane, Renata, Lenaldo, Sandra, Sandrinha, André, Cherubino, Júnior e Gesy.

- À CAPES, pelo apoio financeiro.

# Resumo

Freqüentemente em análise de sobrevivência quando covariáveis são incorporadas na análise os seus valores registrados são aqueles medidos na origem do tempo ou no início do estudo. Contudo em muitos estudos que envolvem dados de sobrevivência existem covariáveis que são monitoradas durante o estudo e seus valores mudam neste período. Estas covariáveis cujos valores se alteram com o tempo são conhecidas como *Covariáveis Dependentes do Tempo* e têm muita utilidade na análise de dados de sobrevivência pois podem ser utilizadas tanto para acomodar medidas que variam com o tempo durante o estudo como também para modelar o efeito de indivíduos que mudam de grupo durante um tratamento. Análises que consideram estas covariáveis podem fornecer resultados mais precisos e a não inclusão delas pode acarretar em sérios vícios de estimação. Um modelo bastante flexível e extensivamente usado em análise de sobrevivência por incorporar o efeito de covariáveis fixas é o modelo de riscos proporcionais de Cox que pode ser generalizado para incorporar o efeito de covariáveis dependentes do tempo. Apesar de ainda não serem muito utilizados na prática, modelos alternativos ao de Cox têm sido sugeridos ao longo dos anos. Aalen propôs um modelo de risco aditivo ou linear que fornece uma alternativa útil para o modelo de Cox. Este modelo tem mostrado freqüentemente vantagens práticas especialmente quando as covariáveis têm efeitos variando no tempo pois permite a observação de mudanças no tempo na influência de cada covariável separadamente. Neste trabalho estes dois modelos são apresentados e é mostrado o uso na presença de covariáveis dependentes do tempo. Dois bancos de dados reais são utilizados para ilustrar os ajustes destes dois modelos. Na primeira aplicação estes modelos são utilizados para avaliar fatores que podem estar relacionados com a duração do aleitamento materno. Na segunda aplicação é verificado se a infecção pelo HIV é fator de risco para o desenvolvimento da sinusite.

# Abstract

Frequently in survival analysis when covariates are incorporated in the analysis its registered values are those measured in the origin of the time or the beginning of the study. However in many studies that involve data of survival covariates exist that are monitored during the study and its values move in this period. These covariveis whose values if modify with the time are known as Covariates Dependent of the Time and have much utility in the analysis of survival data therefore can be used in such a way to accomodate measured that vary with the time during the study as also to model the effect of individuals that move of group during a treatment. Analyses that consider these covariates can supply resulted more necessary and the inclusion of them cannot cause serious vices of estimation. A sufficiently flexible and extensively used model in analysis of survival for incorporating the effect of fixed covariates is the model of proportional risks of Cox that can be generalized to incorporate the effect of dependent covariates of the time. Although not to be very used in the practical one, alternative models to the one of Cox have been suggested to the long one of the years. Aalen considered a model of additive or linear risk that supplies a useful alternative the model of Cox. This model has shown frequently practical advantages especially when the covariates have effect varying in the time therefore separately allow the control of the influence of the variation in the time of each covariate. In this work these two models are presented and are shown to the use in the presence of dependents covariates of the time. Two real data bases are used to compare the adjustments of these two models. In the first application these models are used to evaluate factors that can be related with the duration of the maternal breastfeeding. In the second application it is verified if the infection for the HIV is factor of risk for the development of the sinusite.

# Conteúdo

# Capítulo 1

# Introdução

Em muitos estudos na área médica ou industrial a variável resposta é o tempo transcorrido até a realização de algum evento de interesse. A análise de sobrevivência é um conjunto de modelos e técnicas estatísticas adequados a lidar com dados deste tipo. O termo sobrevivência é usado porque o primeiro uso destas técnicas surgiu de uma empresa de seguro que estava desenvolvendo métodos de custo de prêmios de seguro de vida. Era necessário conhecer o risco, ou tempo de sobrevivência médio, associado a um tipo particular de cliente. O risco foi estimado a partir de um grande grupo de indivíduos com uma idade particular, por sexo e outras características pertinentes.

O evento de interesse é freqüentemente referido como "falha" embora este evento possa ser, por exemplo, ocorrência de nascimento, casamento, mudança de residência, promoção em uma empresa ou a execução de uma tarefa em um experimento de psicologia. Todavia as áreas de maiores aplicações são a engenharia, cujo ramo de estudo é denominado confiabilidade, em que produtos ou componentes são colocados sob teste até a falha e, a área médica, no estudo de doenças crônicas em que o evento de interesse pode ser o desenvolvimento da doença ou a morte de um paciente e ,neste caso, é chamada de análise de sobrevivência.

Estudos de sobrevivência incluem freqüentemente indivíduos cujo tempo até a ocorrência do evento, denominado tempo de sobrevivência ou falha, não é conhecido totalmente. Embora não tenha ocorrido o evento de interesse, existe disponível a informação de que o tempo de sobrevivência é maior do que este tempo observado. Estas observações são referidas como censuradas. Uma outra característica é que os dados de sobrevivência geralmente não são distribuídos simetricamente e portanto não é razoável assumir que dados deste tipo tenham uma distribuição normal. Este problema pode ser resolvido aplicando uma transformação aos dados, como por exemplo a logarítimica, para obter uma distribuição aproximadamente simétrica. Entretanto, uma abordagem mais satisfatória é adotar um modelo distribucional alternativo para os dados originais. Por tais razões estudos nessa área não são tratados por procedimentos estatísticos padrões.

A quantidade de estudos de sobrevivência têm crescido muito durante os últimos anos e novos métodos estatísticos têm sido desenvolvidos neste período. Um dos instrumentos

mais antigos usado pelas companhias de seguros para estimar a sobrevivência é a tabela de vida que foi primeiramente desenvolvida pelo astrônomo E. Halley no século XVII. Uma aproximação simplificada deste método é descrita por Hill (1984). Kaplan e Meier desenvolveram, em 1958, um estimador que é o mais usado em estudos clínicos para estimar a função de sobrevivência, dando início dessa forma a uma nova fase de pesquisas relacionada aos métodos não-paramétricos.

Na engenharia esta técnica passou a ser estudada com mais detalhes por engenheiros e estatísticos a partir dos anos 50 onde a confiabilidade dos produtos era avaliada através de testes de vida. Nos anos 60 e 70, foram introduzidas uma grande variedade de novas aplicações bem como novas técnicas de análise destinadas a auxiliar a solução de problemas de confiabilidade. Em 1970, Nelson utilizou modelos estatísticos para lidar com a confiabilidade de produtos, denominados testes de vida acelerados. Estes testes são uma forma de obter informações sobre a confiabilidade de um produto de maneira mais rápida. Eles consistem em submeter o produto a condições estressantes, por exemplo níveis altos de temperatura, pressão ou voltagem, observando o seu comportamento para as condições do projeto. (Nelson, 1990).

As vezes o interesse não é apenas na distribuição do tempo de sobrevivência. É comum a comparação de tempos de sobrevivência de dois ou mais grupos bem como a verificação da influência de outros fatores nesse tempo tanto na engenharia quanto, principalmente, na medicina. Por exemplo, o tempo de vida de um componente pode ser influenciado pela pressão exercida nele ou a temperatura a que está exposto e, a idade de um paciente pode influenciar no aparecimento de dores na coluna. A maneira mais eficiente de solucionar o problema e incorporar o efeito dessas covariáveis no estudo é utilizar um modelo estatístico de regressão. Existem duas classes de modelos populares para tal situação, os modelos paramétricos e os semi-paramétricos. A segunda classe de modelos, introduzida por Cox em 1972, também chamada simplesmente de modelo de regressão de Cox é bastante flexível e extensivamente usada em dados de interesse na área de saúde. Freqüentemente em estudos de sobrevivência o efeito das covariáveis de interesse pode variar ao longo do tempo de duração do estudo. O modelo de Cox pode ser generalizado para incorporar o efeito destas covariáveis conhecidas como covariáveis dependentes do tempo.

O modelo de riscos proporcionais de Cox é referido como a maior estrutura de análise de regressão para dados de sobrevivência. A aproximação estabelecida por Cox, incluíndo sua idéia de verossimilhança parcial, é extremamente útil e representa um dos maiores avanços na análise de dados censurados. Algumas alternativas para o modelo de Cox têm sido sugeridas na literatura ao longo dos anos. Em 1980, Aalen propôs

um modelo de risco aditivo que fornece uma alternativa útil para o modelo de Cox. Este modelo tem mostrado freqüentemente vantagens práticas especialmente quando as covariáveis têm efeitos variando no tempo, pois em vários estudos em que as covariáveis são acompanhadas ao longo do tempo os seus valores podem ser modificados durante o estudo. Aalen (1980, 1989, 1993) mostrou que seu modelo é útil tanto como uma alternativa bem como uma ferramenta de diagnóstico para o modelo de Cox.

Considerando a grande utilidade no uso de modelos de regressão na análise de dados censurados e, a real necessidade de analisar modelos contendo covariáveis dependentes no tempo pretende-se com este trabalho apresentar teoricamente o modelo de riscos proporcionais de Cox e o modelo aditivo de Aalen e fazer duas aplicações a dados com algumas covariáveis cujos valores variam no tempo.

A presente dissertação de mestrado está dividida em seis capítulos. Uma descrição dos conceitos básicos indispensáveis ao desenvolvimento deste trabalho é apresentada no Capítulo 2.

O Capítulo 3 descreve o modelo de riscos proporcionais de Cox e os testes para avaliar o seu desempenho. No Capítulo 3 está também apresentada a extensão do modelo de Cox levando em consideração a inclusão de covariáveis dependentes do tempo.

O modelo aditivo de Aalen é o tema do Capítulo 4 em que, além da descrição do modelo é apresentado um gráfico para verificar o comportamento dos efeitos das covariáveis no tempo.

No Capítulo 5 são apresentadas duas aplicações dos modelos descritos a dados incluindo covariáveis dependentes do tempo. Os dados utilizados na primeira aplicação tratam da duração do aleitamento materno em áreas urbanas na zona da mata meridional do estado de Pernambuco onde foram avaliados fatores que podem estar relacionados com o tempo de aleitamento, tais como condição sócio-econômica e avaliação antropométrica. Na segunda aplicação os dados utilizados fazem parte de um estudo relacionado a manifestações otorrinolaringológicas em pacientes com AIDS. Pretende-se identificar se a infecção pelo HIV influencia na ocorrência de sinusite. A covariável que indica o grupo de classificação quanto a infecção pelo HIV é dependente do tempo.

Para realização deste trabalho foram utilizados os softwares `R`, `SPSS`, `EPI-info` e a linguagem de tipografia TEX. O software `R` é interpretado como uma linguagem computacional designada para análise de dados estatísticos que se caracteriza pelo compromisso entre a grande flexibilidade oferecida pelas linguagens compiladas, tais como `C`, `C++` e `FORTRAN` e a conveniência de softwares estatísticos tradicionais. Inclui uma ampla variedade de métodos estatísticos tradicionais e modernos com uma característica importante é que o `R` é um software gratuito e portanto pode ser obtido e

distribuído sem custo. Para mais detalhes sobre esta linguagem de programação ver Cribari–Neto e Zarkos (1999). A versão utilizada neste trabalho foi a 1.7.0 e está disponível no endereço `http://www.r-project.org`. O programa `R` foi utilizado neste trabalho para obtenção de todos os resultados que estão apresentados no Capítulo 5. Entre eles pode-se destacar as curvas de sobrevivência, os testes e os ajustes dos modelos de Cox e de Aalen para covariáveis dependentes do tempo. Para aplicar o modelo de Aalen foi necessário utilizar a função `addreg` que está disponível no endereço `http://www.med.uio.no/imb/stat/addreg/`.

O `SPSS` é um software estatístico freqüentemente utilizados para análise e manipulação de dados em diversas áreas do conhecimento. Na elaboração desta dissertação, o `SPSS`, na versão 11.0, foi usado para a criação dos bancos de dados e manipulação das seguintes covariáveis para os dados do aleitamento materno: o tempo inicial do acompanhamento, o tempo final do acompanhamento, a renda per-capta e o índice de massa corporal da mãe. Para a avaliação nutricional das crianças foi usado o `EPI-info`, freqüentemente usado nas ciências da saúde. As medidas antropométricas foram construídas atendendo às recomendações da Organização Mundial da Saúde (OMS). Os índices padronizados construídos, medidos em escore Z, foram o WAZ (peso por idade) e o HAZ (altura por idade) descritos no Capítulo 5 e cujo padrão de referência utilizado para comparação foi o do National Center Health Statistics (NCHS) de uso recomendado pela OMS.

Para elaboração do texto da presente dissertação foi usada a linguagem de tipografia TEX amplamente utilizada para publicações científicas. Esta linguagem se destaca pela flexibilidade e qualidade de apresentação e permite a definição de novos comandos para obtenção do efeito desejado no texto. É também de domínio público podendo ser obtida gratuitamente no endereço `http://www.miktex.org`. Maiores detalhes sobre o TEX podem ser encontrados em Knuth (1986).

# Capítulo 2

# Conceitos Básicos

## 2.1 Tempo de Sobrevivência

Em estudos de análise de sobrevivência a característica importante não é apenas o resultado do evento, tal como a morte ou a ocorrência de promoção em uma empresa, mas o tempo até a ocorrência deste evento denominado tempo de sobrevivência, de vida ou de falha. Para determinar este tempo de sobrevivência são necessários três elementos básicos: o tempo inicial, a escala de medida e o evento falha.

O tempo de origem ou tempo inicial deve ser precisamente definido para cada indivíduo ou elemento em estudo. Em um estudo clínico aleatorizado a data de aleatorização é a escolha natural para o tempo inicial. A data do diagnóstico ou do início do tratamento de doenças também são outras possíveis alternativas em estudos médicos.

Usualmente a escala de medida é o tempo de relógio ou tempo real embora possam surgir outras possibilidades, tais como, o número de ciclos, a quilometragem de um carro ou o tamanho medido até encontrar o primeiro defeito em um fio de tecido.

O evento falha que é a denominação do evento de interesse também deve ser definido de forma clara e precisa. Na área médica a falha pode representar a morte de um paciente por uma causa específica, a recaída de uma doença ou a incidência de uma nova doença. Em outras aplicações podem ser: a obtenção de um emprego, o divórcio ou a recaída de uma interrupção de fumo. O tempo de sobrevivência ou tempo de falha vai do tempo inicial até a ocorrência do evento falha usando a escala de medida já definida anteriormente.

## 2.2 Censura

Os estudos em análise de sobrevivência envolvem uma resposta temporal e são freqüentemente prospectivos e de longa duração. Porém podem terminar antes que ocorra o evento de interesse para todos os casos da amostra. Uma característica importante decorrente destes estudos é a presença de observações incompletas ou parciais do

tempo de sobrevivência denominadas censuras. Em um exemplo de estudo de hepatite (Soares e Colosimo, 1995), no grupo controle, que não recebeu o tratamento testado, oito pacientes não haviam morrido quando o estudo terminou e o acompanhamento de outros cinco foi perdido no decorrer do estudo ou seja dos 15 pacientes deste grupo 13 foram censurados.

Pode-se observar que toda informação obtida por uma observação censurada é que o seu tempo de falha é superior ao tempo registrado. É importante notar que mesmo censuradas, todas as observações de um estudo de sobrevivência devem ser usadas na análise estatística pois mesmo incompletas fornecem informações sobre o tempo de falha e, a omissão destas observações no cálculo das estatísticas de interesse provavelmente resultarão em conclusões viciadas.

Existem três conhecidos mecanismos de censura. A censura do tipo I, também denominada censura à direita, ocorre quando o estudo é terminado após um período pré-estabelecido de tempo. As observações cujo evento de interesse não foi observado até este tempo são ditas censuradas. Um outro tipo de censura, a do tipo II é aquela onde o estudo será terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos. O terceiro tipo que é a do tipo aleatória é o mecanismo de censura mais comum em estudos médicos e pode ocorrer se a observação for retirada no decorrer do estudo sem ter ocorrido o evento de interesse. Por exemplo em um estudo médico o paciente após entrar no estudo decide não ir até o fim, seja porque ele mudou de local de residência, de hospital ou simplesmente porque perdeu o interesse no estudo. Neste caso a censura aleatória ocorre porque há perda de acompanhamento. Uma outra forma de ocorrer este tipo de censura é se o evento de interesse ocorrer por uma razão diferente da estudada. Em um estudo de câncer onde o evento falha é a morte do paciente, se ele morrer, por exemplo, de um acidente automobilístico esta observação é dita censurada.

Para representar o processo de censura aleatório é necessário o uso de duas variáveis aleatórias. Suponha que o tempo de falha de uma observação seja representado pela variável aleatória $T$ e seja $C$ uma variável aleatória independente de $T$ representando o tempo de censura associado a esta observação. Então os dados observados consistem em $t = \min(T, C)$ e o indicador de censura é dado por

$$\delta = \begin{cases} 0, & \text{se o tempo de sobrevivência é censurado,} \\ \\ 1, & \text{para tempo de sobrevivência não censurado, ou seja, } T > C. \end{cases}$$

## 2.3 Funções do Tempo de Sobrevivência

O tempo de sobrevivência de um indivíduo é uma variável aleatória $T$ que pode assumir valores não negativos. Estes valores que $T$ pode assumir têm uma distribuição de probabilidade que pode ser especificada de várias formas, algumas das quais são particularmente úteis e bastante usadas para ilustrar diferentes aspectos dos dados em aplicações de sobrevivência: a função de sobrevivência, a função de risco e a função de risco acumulada.

### 2.3.1 Função de Sobrevivência

Suponha que a variável aleatória $T$ tenha uma distribuição de probabilidade com função densidade de probabilidade $f(t)$. A função de distribuição de $T$ é então dada por

$$F(t) = P(T \leq t) = \int_0^t f(u)du,$$

e representa a probabilidade de que o tempo de sobrevivência seja menor que algum valor $t$.

A função de sobrevivência denotada por $S(t)$ é definida então como a probabilidade do tempo de sobrevivência ser maior ou igual de que um certo tempo $t$. Em termos probabilísticos isto é escrito como

$$S(t) = P(T \geq t).$$

Escrevendo em termos da função de distribuição tem–se que

$$S(t) = 1 - F(t),$$

ou seja, em um estudo médico onde o evento de interesse é a morte, a função de sobrevivência fornece a probabilidade de um indivíduo sobreviver além de um tempo $t$.

A função de sobrevivência é uma função não crescente no tempo com as propriedades de que a probabilidade de sobreviver pelo menos ao tempo zero é um e a probabilidade de sobreviver no tempo infinito é zero. Isto é,

$$S(t) = \begin{cases} 1, & \text{para } t = 0, \\ \\ 0, & \text{para } t = \infty. \end{cases}$$

Para descrever a função de sobrevivência é geralmente utilizada uma representação gráfica de $S(t)$, ou seja, o gráfico de $S(t)$ versus $t$ que é chamado de curva de sobrevivência. Uma curva íngreme representa razão de sobrevivência baixa ou curto tempo de sobrevivência e uma curva de sobrevivência gradual ou plana representa taxa de sobrevivência alta ou sobrevivência longa.

A curva de sobrevivência pode ser usada para comparar distribuições de sobrevivência de dois ou mais grupos e também para determinar quantidades relevantes tal como a mediana e outros percentis. É importante salientar que tratando de distribuições de sobrevivência assimétricas, a média não deve ser usada para descrever a tendência central da distribuição. Neste caso a mediana deve ser utilizada devido a influência que valores extremos, tempos de vida muito curtos ou longos, proporcionam na média.

### 2.3.2 Função de Risco

As funções $F(t)$ e $f(t)$ fornecem duas formas, matematicamente equivalentes, de especificar a distribuição de uma variável aleatória contínua não-negativa, contudo existem outras funções equivalentes que podem ser usadas. Uma função especial bastante utilizada, devido a sua interpretação em análise de sobrevivência, é a função de risco denotada por $h(t)$.

A função de risco do tempo de sobrevivência $T$ fornece a taxa de falha condicional, ou seja, é definida como a taxa de falha em um intervalo pequeno de tempo ($\Delta t$) assumindo que o indivíduo tenha sobrevivido até o início do intervalo. Para se obter uma definição formal da função de risco considere um intervalo de tempo $[t, t + \Delta t)$ e expresse a probabilidade de uma observação falhar neste intervalo em termos da função de sobrevivência como

$$S(t) - S(t + \Delta t).$$

A taxa de falha no intervalo $[t, t+\Delta t)$ é definida como a probabilidade de que a observação falhe neste intervalo, dado que não falhou antes de $t$, dividida pelo comprimento do intervalo. Dessa forma a taxa de falha no intervalo $[t, t + \Delta t)$ é expressa por

$$\frac{S(t) - S(t + \Delta t)}{[(t + \Delta t) - t]S(t)}.$$

Assim $h(t)$ pode ser escrita como

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Para $\Delta t$ pequeno, $h(t)$ apresenta a taxa de falha instantânea no tempo $t$ e é também denominada de função de taxa de falha ou taxa de mortalidade condicional.

A função de risco desempenha um papel importante na análise de dados de sobrevivência sendo bastante útil para especificar a distribuição do tempo de vida pois descreve a forma em que a taxa instantânea de falha muda com o tempo. A função de risco pode então ser definida como

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

Pode-se escrever a função de risco em termos da função de distribuição $F(t)$ e da função densidade de probabilidade $f(t)$ da seguinte forma

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}.$$

### 2.3.3 Função de Risco Acumulada

A função de Risco Acumulada é bastante utilizada em procedimentos de análise gráfica para verificação da adequação de modelos estatísticos (Nelson, 1982).

Esta função pode ser definida a partir da função de risco por

$$H(t) = \int_0^t h(x)dx \tag{2.1}$$

Substituíndo $h(x)$ na Equação (2.1) tem-se que

$$\begin{aligned} H(t) &= \int_0^t \frac{f(x)}{[1 - F(x)]} dx = -\log[1 - F(x)] \mid_0^t \\ &= -\log[1 - F(t)] + \log[1 - F(0)] = -\log[1 - F(t)] \\ &= -\log[S(t)]. \end{aligned}$$

Assim

$$H(t) = -\log[S(t)]$$

e

$$S(t) = \exp^{-H(t)}. \tag{2.2}$$

## 2.4 Estimação da Função de Sobrevivência

Um passo inicial nos estudos de tempo de vida é usualmente a estimação da função de sobrevivência. Estes estudos freqüentemente apresentam observações censuradas, o que requer técnicas estatísticas especializadas para acomodar a informação contida nestas observações. Algumas técnicas estatísticas podem ser utilizadas para analisar dados de tempo de sobrevivência na presença de censura. Podem ser citados três estimadores não-paramétricos, que serão apresentados a seguir, usados para estimação da função de sobrevivência: a tabela de vida, o estimador de Kaplan-Meier e o estimador de Nelson-Aalen. Estes estimadores são conhecidos como não-paramétricos pois usam os próprios dados para estimar as quantidades necessárias da análise, sem fazer uso de suposições a respeito da forma da distribuição dos tempos de sobrevivência.

### 2.4.1 Tabela de Vida

A tabela de vida que também é conhecida como método atuarial é um dos instrumentos estatísticos mais antigos utilizados pelas companhias de seguro desde o século XVII. Berkson e Gage (1950), Cutler e Ederer (1958) e Gehan (1969) desenvolveram métodos para estimação da função de sobrevivência. A tabela de vida é considerada como um procedimento que mostra a distribuição do tempo de sobrevivência para grupos homogêneos de indivíduos, requerendo um número grande de observações de no mínimo 30 para que os tempos possam ser agrupados em intervalos.

Para construir uma tabela de vida primeiramente divide-se o período total de observação em um certo número de intervalos e para cada intervalo estima-se o valor da taxa de falha e a partir da obtenção desses valores estima-se a função de sobrevivência. A taxa de falha ou função de risco foi definida anteriormente na Seção 2.3.2 como a probabilidade de uma observação falhar em um certo intervalo de tempo dado que ela não falhou até o início deste intervalo. Esta função pode ser estimada na tabela de vida a partir de dados censurados por

$$\hat{h}(t_{i-1}) = \frac{N^{\underline{o}} \quad falhas \quad em \quad [t_{i-1}, t_i)}{(N^{\underline{o}} \quad sob \quad risco \quad em \quad t_{i-1}) - (N^{\underline{o}} \quad censuras \quad em \quad [t_{i-1}, t_i))/2}, \tag{2.3}$$

em que $i = 1, \ldots, n$, $t = t_1, \ldots, t_n$ e $t_0 = 0$. Verifica-se na Equação (2.3) que observações censuradas no intervalo $[t_{i-1}, t_i)$ são tratadas como se estivessem sob risco durante a metade do intervalo considerado, dado que no denominador da Equação (2.3) o número de censuras no intervalo $[t_{i-1}, t_i)$ é dividido por dois. Suponha um estudo iniciado com $n$ indivíduos, o risco de falhar até $t_1$ é $\hat{h}(t_1)$, ou seja dos $n$ indivíduos $n[\hat{h}(t_1)]$ não chegarão a $t_1$. Assim no final do primeiro período $n[1 - \hat{h}(t_1)]$ indivíduos ainda estarão vivos. Dessa maneira a função de sobrevivência, que é a probabilidade de sobreviver além de $t_1$ pode então ser estimada por

$$\hat{S}(t_1) = \frac{n[1 - \hat{h}(t_1)]}{n} = 1 - \hat{h}(t_1).$$

De forma análoga, dos $n[1 - \hat{h}(t_1)]$ indivíduos que sobreviveram ao final do primeiro período apenas $n[1 - \hat{h}(t_1)][1 - \hat{h}(t_2)]$ chegarão ao final do segundo período. Portanto

$$\hat{S}(t_2) = [1 - \hat{h}(t_1)][1 - \hat{h}(t_2)].$$

Assim, de uma forma geral, para qualquer tempo $t$ o estimador atuarial da função de sobrevivência é definido por

$$\hat{S}_{TV}(t) = \prod_{i/t_i<t} [1 - \hat{h}(t_i)]. \tag{2.4}$$

Uma estimativa gráfica da função de sobrevivência será uma função escada, com valores constantes da função em cada intervalo de tempo.

### 2.4.2 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier é sem dúvida o mais utilizado em estudos clínicos. Foi proposto por Kaplan e Meier em 1958 e é também conhecido como estimador produto-limite. A construção do estimador de Kaplan-Meier considera o número de intervalos igual ao número de falhas em tempos distintos e os limites dos intervalos são os próprios tempos de falhas da amostra. Sejam $t_1, t_2, \ldots, t_n$ os tempos de falhas de maneira que $t_1 \leq t_2 \leq \ldots \leq t_n$.

O estimador de Kaplan-Meier é então definido como

$$\hat{S}_{KM}(t) = \prod_{i/t_i<t} \frac{n_i - d_i}{n_i}, \tag{2.5}$$

onde $d_i$ é o número de falhas no tempo $t_i$ e $n_i$ é o número de indivíduos que não falharam e não foram censurados até o tempo $t_i$ (exclusive). Pode-se verificar que o estimador de Kaplan-Meier pode ser obtido a partir da Equação (2.4) considerando a função de risco estimada igual a $d_i/n_i$. Em seu artigo original Kaplan e Meier justificaram a Equação (2.5) apresentada acima mostrando que ela é o estimador de máxima verossimilhança da função de sobrevivência $S(t)$.

As propriedades assintóticas destes dois estimadores descritos anteriormente foram estudadas por alguns autores tais como, Kaplan e Meier (1958), Breslow e Crowley (1974), Efron (1967), Meier (1975) e Aalen (1976). Estes estudos mostraram que o estimador de Kaplan-Meier é não-viciado em grandes amostras e em amostras de tamanhos menores existem algumas evidências empíricas da superioridade deste estimador em relação a tabela de vida. A principal diferença entre a tabela de vida e o estimador de Kaplan-Meier é o número de intervalos utilizados na construção dos mesmos. Na tabela de vida os tempos de falhas são agrupados em intervalos de forma arbitrária enquanto que o estimador de Kaplan-Meier é baseado em um número de intervalos igual ao número de tempos de falha distintos. Usualmente o estimador de Kaplan-Meier considera um número de intervalos maior que o número de intervalos da tabela de vida, confirmando a superioridade do mesmo dado que quanto maior o número de intervalos melhor a aproximação para a verdadeira distribuição do tempo de falha. Para esta dissertação os limites de confiança utilizados para o estimador de Kaplan-Meier foram construídos de acordo com Klein e Moeschberger (1997).

### 2.4.3 Estimador de Nelson - Aalen

O estimador de Kaplan-Meier é o mais usado freqüentemente para estimar a função de sobrevivência. Contudo um estimador alternativo referido como estimador de Nelson-Aalen foi sugerido por Nelson (1972) e estudado em seguida por Aalen (1978). Assim como o estimador de Kaplan-Meier, este estimador não-paramétrico requer apenas uma ordenação dos tempos de falhas ou censuras e não inclue o efeito de covariáveis. O estimador de Nelson-Aalen pode ser obtido usando a teoria de processos de contagem, o que permite a derivação de suas propriedades (Andersen et al., 1993; Fleming e Harrington, 1991).

Seja $t_1 \leq t_2 \leq \ldots \leq t_n$ os tempos de falhas ordenados, com função de sobrevivência $S(t)$. O estimador de Nelson-Aalen da função de risco acumulada $H(t)$ é dado por

$$\hat{H}_{NA}(t) = \sum_{i/t_i<t} \Big(\frac{d_i}{n_i}\Big).$$

Através da relação entre a função de risco acumulada e a função de sobrevivência, apresentada na Equação (2.2), o estimador de Nelson-Aalen da função de Sobrevivência é então dado por

$$\hat{S}_{NA}(t) = \exp(-\hat{\mathrm{H}}_{\mathrm{NA}}(\mathrm{t})).$$

A obtenção da função de risco acumulada usando a relação entre essa função e a função de sobrevivência gera alguns problemas em amostras pequenas quando se utiliza o estimador de Kaplan-Meier sendo então aconselhável o uso do estimador de Nelson-Aalen para obter esta função.

Através de um estudo de simulação Colosimo, et al. (2002) mostraram que o estimador de Nelson-Aalen é melhor que o estimador de Kaplan-meier para obter estimativas da fração de sobrevivência, especialmente quando o método de interpolação é usado. Já na estimação do percentil o estimador de Kaplan-Meier apresentou um melhor desempenho para taxas de falhas decrescentes ao passo que o estimador de Nelson-Aalen forneceu resultados melhores para taxas de falha crescentes. Alguns autores têm proposto uma interpolação linear usando ambos estimadores para obter estimativas da sobrevivência (Lee, 1992).

# Capítulo 3

# Modelo de Riscos Proporcionais de Cox

## 3.1 Forma e Estimação do Modelo

Os estudos em análise de sobrevivência muitas vezes envolvem covariáveis que podem estar relacionadas com o tempo de sobrevivência. Essas covariáveis devem ser incluídas na análise estatística dos dados para explicar seu possível efeito no tempo de sobrevivência. Uma das alternativas metodológicas que incorpora informações no estudo do tempo de sobrevivência através da introdução de covariáveis é o modelo de riscos proporcionais. Uma família de riscos proporcionais é uma classe de modelos com a propriedade de que diferentes indivíduos têm funções de riscos proporcionais. Ou seja, a razão entre duas funções de riscos para dois indivíduos distintos não varia com o tempo. Isto implica que a função de risco no tempo $t$, dado $x$, pode ser escrita na forma

$$h(t/x) = h_0(t)g(x, \beta) \tag{3.1}$$

em que $h_0(t)$ é uma função arbitrária de risco padrão ou de base, $x$ é o vetor de covariáveis fixas, $g$ é uma função que deve ser especificada e $\beta$ é o vetor de parâmetros regressores associado com as covariáveis. Sob a suposição de riscos proporcionais, Cox propôs em 1972 o *Modelo de Riscos Proporcionais de Cox* onde a parte paramétrica do modelo $g(x, \beta)$ é geralmente tomada como $\exp(x'\beta)$ (Cox, 1972).

O conjunto de valores das covariáveis no modelo de riscos proporcionais de Cox será representado pelo vetor $x$, tal que $x = (x_1, x_2, \cdots, x_p)'$. Seja $t_i$ o tempo de sobrevivência do $i$-ésimo indivíduo que possivelmente depende do valor dessas $p$ covariáveis. Dessa maneira o principal interesse em problemas como este é avaliar como estas covariáveis influenciam $t_i$. Então, no modelo de riscos proporcionais de Cox a função de risco do $i$-ésimo indivíduo pode ser escrita como

$$h_i(t/x_{1i}, \ldots, x_{pi}) = h_0(t) \exp(\beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

ou de forma equivalente

$$h_i(t/x_i) = h_0(t)\exp(x_i'\beta),$$

em que $\beta' = (\beta_1, \ldots, \beta_p)$ é um vetor de parâmetros desconhecidos e $x_i' = (x_{1i}, \ldots, x_{pi})$.

Este modelo é chamado de riscos proporcionais devido a propriedade de que a razão das taxas de falha de dois indivíduos diferentes é constante no tempo. Ou seja, a razão das funções de risco para dois indivíduos $i$ e $j$ é dada por

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)\exp(\mathrm{x_i'}\beta)}{h_0(t)\exp(\mathrm{x_j'}\beta)} = \exp(\mathrm{x_i'}\beta - \mathrm{x_j'}\beta).$$

Esta razão não depende do tempo, isto é, o risco de falha de um indivíduo em relação ao outro é constante para todos os tempos de acompanhamento.

Os dois componentes multiplicativos do modelo são de naturezas distintas, um não-paramétrico e o outro paramétrico sendo esta a razão do modelo ser do tipo semi-paramétrico o que o torna bastante flexível. O componente não-paramétrico, $h_0(t)$, não especificado, é uma função não negativa no tempo geralmente chamada de função de base pois $h(t) = h_0(t)$ quando $x = 0$. O componente paramétrico é em geral usado em termo multiplicativo e, por ser na forma exponencial, garante que $h(t)$ será positiva. Um exemplo da flexibilidade deste modelo é possuir alguns modelos conhecidos como casos particulares tal como o modelo de regressão Weibull (Kalbfleisch e Prentice, 1980).

O modelo de regressão de Cox é caracterizado pelos coeficientes $\beta$ que medem o efeito das covariáveis sobre a função de risco. Dessa maneira é necessário um método de estimação para se fazer inferência no modelo. O método de máxima verossimilhança usual, bastante conhecido e freqüentemente usado, não pode ser utilizado aqui, pois a presença do componente não-paramétrico $h_0(t)$ na função de verossimilhança torna este método inapropriado. Frente a tal dificuldade, Cox (1975) propôs o método de máxima verossimilhança parcial que condiciona a verossimilhança à história dos tempos de sobrevivência e censuras anteriores e desta forma elimina a função de base desconhecida $h_0(t)$.

## Verossimilhança Parcial

Nos intervalos onde nenhuma falha ocorre não existe nenhuma informação sobre o vetor de parâmetros $\beta$ pois $h_0(t)$ pode, teoricamente, ser identicamente igual a zero em tais intervalos. Uma vez que é necessário um método de análise válido para todas $h_0(t)$ possíveis, a consideração de uma distribuição condicional é necessária.

Considere uma amostra de $n$ indivíduos, onde se tem $k(\leq n)$ falhas distintas nos tempos $t_1 \leq t_2 \ldots \leq t_k$. A probabilidade condicional da $i$-ésima observação vir a falhar no tempo $t_i$, conhecendo quais observações estão sob risco em $t_i$ é

$$\frac{h_i(t_i)}{\sum\limits_{j \in R(t_i)} h_j(t_i)} = \frac{h_0(t_i)\exp(x_i'\beta)}{\sum\limits_{j \in R(t_i)} h_0(t_i)\exp(x_j'\beta)} = \frac{\exp(x_i'\beta)}{\sum\limits_{j \in R(t_i)} \exp(x_j'\beta)}, \tag{3.2}$$

em que $R(t_i)$ é o conjunto dos índices dos indivíduos sob risco no tempo $t_i$. Pode-se verificar que ao utilizar a probabilidade condicional, o componente não-paramétrico $h_0(t)$ desaparece da Equação (3.2).

A função de verossimilhança parcial $L(\beta)$ é obtida fazendo o produto dessas probabilidades condicionais, associadas aos distintos tempos de falha, ou seja,

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(x_i'\beta)}{\sum\limits_{j \in R(t_i)} \exp(x_j'\beta)} \tag{3.3}$$

$$= \prod_{i=1}^{n} \left[ \frac{\exp(x_i'\beta)}{\sum\limits_{j \in R(t_i)} \exp(x_j'\beta)} \right]^{\delta_i},$$

em que

$$\delta_i = \begin{cases} 0, & \text{se o i-ésimo tempo de sobrevivência é censurado,} \\ 1, & \text{caso contrário} \end{cases}.$$

A função $l(\beta)$ é obtida pelo logarítmo da função de verossimilhança parcial, ou seja, $l(\beta) = \log(L(\beta))$ e $U(\beta)$ é o vetor escore composto das primeiras derivadas da função $l(\beta)$. Estimadores para o vetor de parâmetros $\beta$ podem ser obtidos maximizando o logaritmo da função de verossimilhança parcial (3.3), ou seja, resolvendo o sistema de equações definido por $U(\beta) = 0$. Isto é o equivalente a

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[ x_i'\beta - \log \sum_{j \in R(t_i)} \exp(x_j'\beta) \right] = 0. \tag{3.4}$$

O procedimento de estimação requer um método iterativo que é geralmente o método de Newton-Raphson, pois as equações encontradas em (3.4) não apresentam forma fechada.

Cox (1975) mostra informalmente que o método usado para construir esta verossimilhança gera estimadores que são consistentes e assintoticamente normalmente distribuídos, com matriz de covariâncias assintóticas estimadas consistentemente pelo inverso do negativo da matriz de segundas derivadas parciais do logaritmo da função de verossimilhança. Provas formais destas propriedades foram apresentadas mais tarde por Tsiatis (1981) e Andersen e Gill (1982).

A função de verossimilhança parcial dada em (3.3) é utilizada para tempos de sobrevivência contínuos e, portanto, não considera a possibilidade de empates dos valores observados. Entretanto, na prática, podem ocorrer empates nos tempos de falhas ou censuras devido à escala de medida. No caso em que ocorrem empates entre falhas e censuras, ou seja os tempos de falhas são iguais mas um deles é censurado, para definir quais observações serão incluídas no conjunto de risco em cada tempo de falha usa-se a convenção de que a censura ocorreu após a falha.

No caso de empates entre falhas, a função de verossimilhança parcial (3.3) deve ser modificada para incorporar tais observações. A aproximação proposta por Breslow(1972) e Peto(1972) é frequentemente usada nos softwares estatísticos. Considere $s_i$ o vetor composto pela soma das $p$ covariáveis para os indivíduos que falham no tempo $t_i$, $i = 1, \cdots, k$ e $d_i$ é o número de falhas neste mesmo tempo. Esta aproximação considera a seguinte função de verossimilhança parcial

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(s_i'\beta)}{\left[\sum_{j \in R(t_i)} \exp(s_j'\beta)\right]^{d_i}}.$$

Quando o número de observações empatadas em qualquer tempo é grande não é adeqüiado o uso desta aproximação. Para estes casos é aconselhável utilizar o modelo de regressão de Cox para dados agrupados (Lawless, 1982; Prentice e Gloeckler, 1978).

## 3.2 Estimação da Função de Sobrevivência através de $h_0(t)$

Considerando que para um determinado indivíduo todas as covariáveis têm valores iguais a zero, pode-se então obter a função de sobrevivência padrão expressa por

$$S_0(t) = \exp\Big( - \int_0^{\mathrm{t}} \mathrm{h_0(u)du}\Big),$$

ou seja

$$S_0(t) = \exp[-\mathrm{H_0(t)}],$$

em que $H_0(t)$ é a função de taxa de falha de base acumulada.
Assim a função de sobrevivência pode ser definida como

$$S(t) = \exp\left\{ - \int_0^{\mathrm{t}} \mathrm{h(u/x)du}\right\},$$

substituindo a função de risco tem-se que

$$\begin{aligned} S(t) &= \exp\left\{ - \int_0^{\mathrm{t}} \exp(\mathrm{x}'\beta)\mathrm{h_0(u)du}\right\} \\ &= \exp\left\{ -\exp(\mathrm{x}'\beta) \int_0^{\mathrm{t}} \mathrm{h_0(u)du}\right\}, \end{aligned}$$

assim $S(t)$ pode ser expressa por

$$S(t) = [S_0(t)]^{\exp(\mathrm{x}'\beta)}.$$

## 3.3 Testes para os Parâmetros do Modelo de Cox

O interesse do pesquisador freqüentemente está relacionado a verificar a associação de covariáveis ao tempo de sobrevivência. A hipótese nula pode então ser definida de maneira que todas as variáveis consideradas não explicam a variação no tempo de sobrevivência. Em outras palavras,

$$H_0 = \beta_1 = \beta_2 = \ldots = \beta_p = 0 \tag{3.5}$$

Três testes podem ser usados para verificar esta hipótese nula global: o teste da razão de verossimilhança, o teste de Wald e o teste Escore que são descritos a seguir. Maiores detalhes sobre estes testes podem ser encontrados em Collett (1994), Kalbfleisch e Prentice (1980) e Le (1997).

### 3.3.1 Teste da Razão de Verossimilhança

Para comparar modelos encaixados ou verificar se um modelo particular é adequado, o uso de uma estatística de teste é requerido. Visto que a função de verossimilhança resume a informação contida nos dados sobre os parâmetros desconhecidos, uma estatística adequada é o valor da função de verossimilhança quando os parâmetros são substituídos pelas suas estimativas de máxima verossimilhança. Isto é a verossimilhança maximizada sob o modelo assumido. Seja $\hat{L}$ a verossimilhança maximizada para um dado modelo. É mais conveniente usar menos duas vezes o logarítmo da verossimilhança maximizada como estatística de teste. Dessa maneira a estatística de interesse é dada por $-2\log\hat{\mathrm{L}}$. Dado que $\hat{L}$ é, na realidade, o produto de várias probabilidades condicionais, sendo dessa forma menor que 1, então $-2\log\hat{\mathrm{L}}$ será sempre positiva e para um certo conjunto de dados quanto menor o valor de $-2\log\hat{\mathrm{L}}$, melhor o modelo. Da mesma forma, quanto maior o valor da verossimilhança maximizada melhor é o ajuste do modelo.

Esta estatística é utilizada para comparar modelos distintos ajustados para os mesmos dados. Dessa forma, para verificar a adeqüação de um determinado modelo, ou seja, verificar a hipótese (3.5) é necessário a definição de um modelo de Cox onde nenhuma covariável tenha influência na sobrevivência e todos os indivíduos tenham o mesmo risco $h_0(t)$, ou seja, todos os coeficientes de regressão sejam iguais a zero. Este modelo, denominado de modelo nulo, tem verossimilhança maximizada associada denotada por $\hat{L_0}$. Por outro lado define-se $\hat{L_\nu}$ como a verossimilhança maximizada do modelo que contém $\nu$ coeficientes de regressão estimados pelo método de máxima verossimilhança parcial. A estatística do teste da razão de verossimilhança parcial (RV) para testar o ajuste de cada modelo é definida como

$$RV = -2\log(\hat{\mathrm{L}}_0/\hat{\mathrm{L}}_\nu) = -2(\log\hat{\mathrm{L}}_0 - \log\hat{\mathrm{L}}_\nu).$$

Sob a hipótese nula (3.5) de que os coeficientes são iguais a zero, esta estatística tem assintoticamente distribuição qui-quadrado com número de graus de liberdade igual a quantidade $\nu$ de coeficientes de regressão estimados.

Para comparar os ajustes de dois modelos encaixados ao mesmo banco de dados, um com $(\nu + k)$ coeficientes regressores e o outro com $\nu$ coeficientes regressores, a estatística dada na equação acima torna-se então

$$RV = -2(\log\hat{\mathrm{L}}_{\nu+\mathrm{k}} - \log\hat{\mathrm{L}}_\nu),$$

que também tem distribuição qui-quadrado mas com $(\nu + k) - \nu = k$ graus de liberdade.

A hipótese nula então é a de que nenhuma melhora no ajuste do modelo foi verificada com a inclusão dos $k$ coeficientes.

Os testes de Wald e o Escore também podem ser utilizados para o teste simultâneo de várias covariáveis. Apesar do teste da razão de verossimilhança ser preferível por questões de consistência e estabilidade nos métodos de cálculos associados, em amostras de tamanhos grandes os testes se tornam equivalentes.

### 3.3.2 Teste de Wald

É usado principalmente para verificar se um coeficiente particular é significativamente igual a zero na presença dos outros termos do modelo. Por exemplo, suponha que um modelo contenha três variáveis explicativas $X_1, X_2$ e $X_3$ com coeficientes dados respectivamente por $\beta_1, \beta_2$ e $\beta_3$. A estatística de teste $(\hat{\beta}_1/\mathrm{DP}(\hat{\beta}_1))$ é então usada para testar a hipótese nula $\beta_1 = 0$ na presença de $\beta_2$ e $\beta_3$. Caso não existam evidências para rejeitar esta hipótese, conclui-se que a variável $X_1$ não é necessária no modelo na presença de $X_2$ e $X_3$. O resultado isolado do teste de hipótese para um coeficiente particular pode não ser fácil de interpretar pois em geral as estimativas individuais $\hat{\beta}_1$,$\hat{\beta}_2, \ldots$ em um modelo de riscos proporcionais não são independentes umas das outras. Assim a hipótese nula $\beta = 0$ pode ser testada utilizando a estatística

$$Z = \frac{\hat{\beta}}{\sqrt{\widehat{\mathrm{VAR}}(\hat{\beta})}}, \tag{3.6}$$

em que $\sqrt{\widehat{\mathrm{VAR}}(\hat{\beta})}$ é o erro padrâo estimado de $\hat{\beta}$ e $\widehat{\mathrm{VAR}}(\hat{\beta}) \approx -\mathrm{E}\left\{\frac{\partial^2(\log \mathrm{L}(\beta))}{\partial \beta^2}\right\}^{-1}$. Sob $H_0$, a estatística (3.6) tem uma distribuição normal padrão. Equivalentemente pode-se utilizar o quadrado desta estatística.

$$W = Z^2 = \frac{\hat{\beta}^2}{\widehat{\mathrm{VAR}}(\hat{\beta})},$$

que sob a hipótese nula tem distribuição qui-quadrado com 1 grau de liberdade. Valores de $W$ superiores ao valor tabelado da distribuição qui-quadrado indicam que a covariável associada a $\beta$ é importante para explicar a variação da resposta.

### 3.3.3 Teste Escore

A estatística do teste escore, assim como a do teste da razão de verossimilhança, é baseada diretamente na função de verossimilhança. Esta estatística denominada de $S$ é definida, para testar a hipótese (3.5), por

$$S = \frac{u^2(0)}{IF(0)},$$

em que

$$u(\beta) = \frac{\partial(\text{logL}(\beta))}{\partial\beta},$$

é o vetor escore eficiente de ordem $p \times 1$ e

$$IF(\beta) = -\text{E}\Big(\frac{\partial^2(\text{logL}(\beta))}{\partial\beta^2}\Big),$$

é a matriz de informação de Fisher de ordem $p \times p$. Sob a hipótese nula (3.5), $S$ tem uma distribuição qui-quadrado com $p$ graus de liberdade e valores de $S$ maiores do que o valor tabelado da distribuição qui-quadrado implicam que se deve rejeitar $H_0$. O teste escore tem uma forma aparentemente complexa. Entretanto, de maneira mais resumida, este teste pode ser definido como a razão entre o quadrado da primeira derivada do logarítmo da verossimilhança, com os parâmetros de interesse iguais a zero e a segunda derivada do logarítmo da verossimilhança, também avaliada com os parâmetros de interesse iguais a zero.

## 3.4 Covariáveis Dependentes do Tempo

### 3.4.1 Introdução

Quando covariáveis são registradas para modelar dados de sobrevivência, os valores tomados para tais covariáveis são em geral aqueles medidos na origem do tempo ou no início do estudo. Por exemplo no estudo para comparar dois tratamentos de câncer de próstata (Collet, 1994), a idade dos pacientes, nível de hemoglobina, tamanho do tumor, grupo de tratamento e o valor de um índice combinando o estágio do tumor e

grau, conhecido como índice Gleason foram registrados na data em que o paciente entrou no estudo. O impacto dessas covariáveis no risco de morte foi então avaliado.

Entretanto em muitos estudos que envolvem dados de sobrevivência existem outras covariáveis que são monitoradas durante o estudo e seus valores mudam neste período. No exemplo de câncer de próstata o tamanho do tumor muda durante o tratamento e pode ser medido de forma regular. Se estes valores forem incorporados na análise estatística é possível fornecer uma melhor previsão do tempo de sobrevivência do paciente. Isto é, valores mais recentes do tamanho do tumor podem fornecer uma melhor indicação da expectativa futura de vida do que aqueles valores registrados na origem do tempo.

Estas covariáveis cujos valores se alteram com o tempo são conhecidas como *Covariáveis Dependentes do Tempo*. Análises que consideram estas covariáveis podem fornecer resultados mais precisos e a não inclusão destes valores pode acarretar em sérios vícios. Estas covariáveis têm muita aplicação em análise de sobrevivência pois podem ser utilizadas tanto para acomodar medidas que variam com o tempo durante um estudo como também podem ser úteis para modelar o efeito de indivíduos que mudam de grupo durante um tratamento.

Tais covariáveis podem ser consideradas dentro de duas amplas classificações referidas como covariáveis internas e covariáveis externas (Kalbfleisch e Prentice,1980).

Covariáveis internas são aquelas que caracterizam um indivíduo sob estudo e podem ser medidas apenas enquanto o paciente sobrevive. Os valores observados levam informação sobre o tempo de sobrevivência do correspondente indivíduo (paciente). Um exemplo pode ser a quantidade de glóbulos brancos no sangue.

Por outro lado, covariáveis externas são variáveis que não necessariamente requerem a sobrevivência do paciente para sua existência. Um tipo de variável externa é aquela que muda de tal forma que seus valores serão conhecidos avançando em um tempo futuro. Existem alguns exemplos tais como a dose de uma droga que pode variar de maneira pré-determinada durante o estudo e, a idade de um paciente, uma vez que a idade no início do tratamento é conhecida, a idade do paciente em algum tempo futuro pode ser obtida de forma exata.

### 3.4.2 Modelo de Cox com Covariáveis Dependentes do Tempo

Os diferentes tipos de covariáveis dependentes do tempo apresentados na Seção 3.4.1 podem ser incorporados ao modelo de regressão de Cox, generalizando-o como

$$h_i(t) = h_0(t)\exp(x_i'(t)\beta). \tag{3.7}$$

É importante verificar que definindo desta forma, o modelo dado pela Equação (3.7) não é mais de riscos proporcionais. Os valores das covariáveis $x_i'(t)$ dependem do tempo $t$ e a razão das funções de risco no tempo $t$ para dois indivíduos $i$ e $j$ dada por

$$\frac{h_i(t)}{h_j(t)} = \exp(x_i'(t)\beta - x_j'(t)\beta),$$

é também dependente do tempo e a interpretação dos coeficientes do modelo deve considerar o tempo $t$.

O coeficiente $\beta_j$, com $j = 1, \ldots, p$, pode portanto ser interpretados como o logarítimo da razão de riscos para dois indivíduos cujo valor da $j$-ésima covariável no tempo $t$ difere de uma unidade quando as outras covariáveis assumem o mesmo valor neste tempo.

Para obter as estimativas dos parâmetros do modelo de regressão de Cox com covariáveis dependentes do tempo basta estender a função escore parcial para

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left[ x_i'(t_i)\beta - \log \sum_{j \in R(t_i)} \exp(x_j'(t_i)\beta) \right],$$

que é uma extensão da Equação 3.4, considerando covariáveis dependentes do tempo.

Para construir intervalos de confiança e testar hipóteses sobre os coeficientes do modelo são necessárias propriedades assintóticas dos estimadores de máxima verossimilhança parcial. As provas mais gerais das propriedades para covariáveis dependentes do tempo foram apresentadas por Andersen e Gill (1982). Desta forma pode-se usar as estatísticas dos testes, apresentadas na Seção 3.3, para fazer inferências no modelo de regressão de Cox com covariáveis dependentes do tempo.

# Capítulo 4

# Modelo Aditivo de Aalen

## 4.1 Introdução

Na teoria clássica de regressão a esperança das variáveis respostas é o objeto principal de modelagem. Em análise de sobrevivência freqüentemente a função de risco é a base da modelagem de regressão. O risco é uma função natural para descrever a distribuição do tempo de vida. Informalmente a função de risco mede o risco de um evento ocorrer em um dado tempo condicional a sobrevivência ao tempo imediatamente anterior. Existem várias possibilidades de modelos de regressão que tem como base a função de risco. Uma delas é o tão conhecido modelo de riscos proporcionais de Cox com a sua verossimilhança parcial apresentada no Capítulo 3. Este modelo tem as vantagens de uma simples interpretação dos resultados e de estar disponível em vários softwares computacionais. Entretanto Aalen em 1989 citou algumas limitações do modelo de Cox. A primeira delas é que as suposições do modelo podem não valer, as vezes o modelo de Cox é usado na literatura sem que suas propriedades sejam checadas e também não é claro se satisfazendo as propriedades usuais de proporcionalidade garantem a adequação do modelo de Cox. Em segundo lugar mudanças ao longo do tempo na influência das covariáveis não são facilmente descobertas e o modelo de Cox não é adaptado para uma descrição detalhada de efeitos de covariáveis ao longo do tempo. Por último a suposição de proporcionalidade do risco é vulnerável à mudanças no número de covariáveis modeladas. Se algumas covariáveis são retiradas de um modelo ou medidas com um diferente nível de precisão, a proporcionalidade é geralmente afetada. Portanto Aalen verificou uma falta de consistência do modelo de Cox a este respeito.

Estas limitações conduziram a uma ampla variedade de modelos que generalizam o modelo de Cox. Uma alternativa, baseada no seu modelo de risco multiplicativo para processo de contagem (Aalen, 1978), foi sugerida originalmente por Aalen (em 1980). Este modelo apresentado de forma mais simples em 1989 (Aalen, 1989) é um modelo de risco aditivo para análise de regressão de dados censurados. Este modelo aditivo de Aalen fornece uma alternativa útil ao modelo de riscos proporcionais de Cox pois

permite que ambos os parâmetros e os vetores de covariáveis variem com o tempo. Já que efeitos temporais não são assumidos serem proporcionais para cada covariável, o modelo de Aalen é capaz de fornecer informações detalhadas a respeito da influência temporal de cada covariável. Os modelos de Cox e Aalen diferem fundamentalmente, o de Cox tem uma função básica não-paramétrica, mas o efeito das covariáveis é modelado parametricamente. Por outro lado, o modelo de Aalen é completamente não-paramétrico no sentido de que funções são ajustadas e não parâmetros. Ou seja, na estimação dos parâmetros o modelo de Aalen usa apenas informação local o que faz este modelo bastante flexível. Os estimadores propostos por Aalen generalizam o tão conhecido estimador de Nelson-Aalen que é o estimador natural no caso de população homogênea. Aplicações foram apresentadas por Mau (1986) e (1988) e Andersen e Vaeth (1989) e resultados teóricos foram feitos por McKeague (1986), McKeague e Utikal (1988) e Huffer e McKeague (1987) indicando que o modelo pode ser útil e é sem dúvida razoável para explorar vantagens da linearidade analogamente a teoria clássica de modelo linear.

## 4.2 Definição

Em um estudo típico, um número de indivíduos são observados ao longo do tempo para verificar a ocorrência de um determinado evento. O acontecimento deste evento é assumido independente entre os indivíduos. Como no modelo de risco multiplicativo, tem-se um tempo até a ocorrência do evento para cada indivíduo, cuja distribuição depende de um vetor dado por $x_i(t) = (1, x_{1i}(t), x_{2i}(t), \ldots, x_{pi}(t))'$ onde $x_{ij}(t)$, com $j = 1, \ldots, p$, são os valores observados, para o $i$-ésimo indivíduo, das covariáveis que podem variar no tempo. Seja $n$ o número de indivíduos, $p$ o número de covariáveis na análise e $h_i(t)$ a função de risco para o tempo de sobrevivência $t_i$ de um indivíduo $i$. O modelo mais geral para $h_i(t)$ que parece ser acessível a análises estatísticas é

$$h_i(t) = \alpha(t, x_i(t)), \tag{4.1}$$

em que $\alpha$ é uma função do tempo geral e desconhecida. Apesar desse modelo ser atrativo do ponto de vista teórico, a exigência de tamanhos de amostras grandes torna-o difícil de ser utilizado na prática.

Assumindo que $\alpha(t, 0) = 0$ e ignorando todos os termos de ordens maiores da expansão de Taylor de primeira ordem de $\alpha(t, X)$ sobre $X = 0$, ou seja, isto é o primeiro termo da expansão da série de Taylor de uma função de risco geral sobre o vetor de covariáveis igual a zero, então o modelo (4.1) se reduz ao modelo de risco aditivo de Aalen dado por

$$h_i(t) = \alpha_0(t) + \sum_{j=1}^{p} \alpha_j(t) x_{ij}(t).$$

Considerando a forma matricial

$$h(t) = \alpha(t) Y(t),$$

em que $\alpha(t) = (\alpha_0(t), \alpha_1(t), \ldots, \alpha_p(t))'$ é um vetor de funções do tempo desconhecidas, cujo primeiro elemento $\alpha_0(t)$ é interpretado como uma função de parâmetro básica, enquanto que $\alpha_j(t)$, $j = 1, \ldots, p$, chamados aqui funções de regressão medem a influência das respectivas covariáveis. A matriz $Y(t)$ de ordem $n \times (p+1)$ é construída da seguinte maneira: se o evento considerado ainda não ocorreu para o i-ésimo indivíduo e ele não é censurado então a i-ésima linha de $Y(t)$ é o vetor $x_i(t) = (1, x_{i1}(t), x_{i2}(t), \ldots, x_{ip}(t))'$. Caso contrário, se o indivíduo não está sob risco no tempo $t$, então a linha correspondente de $Y(t)$ contém apenas zeros.

Este modelo é considerado não-paramétrico pois nenhuma forma paramétrica particular é assumida para as funções de regressão. Como visto, estas funções podem variar arbitrariamente com o tempo, revelando mudanças na influência das covariáveis. Esta é uma das vantagens do modelo acima bem como a não exigência de tamanho de amostra extremamente grande.

## 4.3 Estimação

O modelo de riscos proporcionais assume que os efeitos das covariáveis agem multiplicativamente na função de risco. Os coeficientes estimados da estrutura de regressão são constantes desconhecidas cujos valores não mudam com o tempo. No modelo de Aalen assume-se que as covariáveis agem de maneira aditiva na função de risco e os coeficientes de riscos desconhecidos podem ser funções do tempo, ou seja, os efeitos das covariáveis podem variar durante o estudo. Dessa forma os estimadores dos parâmetros são baseados nas técnicas de mínimos quadrados. A derivação desses estimadores é similar a derivação do estimador de Nelson-Aalen da função de risco acumulada apresentado na Seção 2.4.3.

A aproximação para estimação depende das suposições sobre a forma funcional das funções de regressão que neste caso são não-paramétricas. A estimação direta das funções de regressão é difícil na prática sendo mais fácil a estimação da função de regressão

acumulada. Isto ocorre pelo mesmo motivo que é mais fácil estimar a função de distribuição acumulada do que a função de densidade de probabilidade. Considera-se então a estimação do vetor coluna $A(t)$ com elementos $A_j(t)$ dados por

$$A_j(t) = \int_0^t \alpha_j(s)ds$$

Sejam $t_1 < t_2 < \ldots < t_k$ os tempos de falhas ordenados. Aalen considerou um estimador razoável de $A(t)$, denominado estimador de mínimos quadrados de Aalen, que é dado por.

$$A^*(t) = \sum_{t_k \leq t} Z(t_k) I_k, \tag{4.2}$$

em que $I_k$ é um vetor de zeros que assume o valor 1 para o indivíduo cujo evento ocorre no tempo $t_k$. Enquanto que $Z(t)$ é a inversa generalizada de $Y(t)$. Em princípio, $Z(t)$ pode ser qualquer inversa generalizada de $Y(t)$. Uma escolha simples pode ser baseada no princípio de mínimos quadrados local, ou seja

$$Z(t) = [Y(t)'Y(t)]^{-1}Y(t)'.$$

Esta inversa usada comumente em modelos de regressão, em geral, pode não ser ótima. Uma escolha ótima dependerá do conhecimento dos verdadeiros valores dos parâmetros. Huffer e McKeague (1987) sugeriram o uso de uma outra inversa definindo assim o estimador de mínimos quadrados ponderados. Neste trabalho será usada a inversa de mínimos quadrados.

É importante notar que o estimador de $A(t)$ é definido apenas sobre um intervalo de tempo onde $Y(t)$ tem posto completo, ou seja, a estimação termina no tempo onde $Y(t)$ perde o posto completo, que é uma consequência do princípio não paramétrico. Ramlau-Hansen (1983) mostrou que também é possível estimar a função mais diretamente utilizando métodos de estimação da densidade de probabilidade.

Os componentes de $A^*(t)$ convergem assintoticamente, sob condições apropriadas, para um processo gaussiano (Aalen 1989). Então um estimador da matriz de covariância de $A^*(t)$ é dado por

$$\Omega^*(t) = \sum_{t_k \leq t} Z(t_k) I_k^D Z(t_k)',$$

em que $I_k^D$ é uma matriz diagonal com $I_k$ como diagonal.

Não é difícil verificar, como consequência dos resultados obtidos anteriormente, que pode-se estimar o risco acumulado e a função de sobrevivência correspondentes dados os valores das covariáveis. Seja $x(t) = (1, x_1(t), x_2(t), \ldots, x_p(t))'$ o conjunto de valores das covariáveis no tempo $t$. O estimador do risco acumulado $H^*(t)$ é dado por

$$H^*(t) = A^*(t)'x(t).$$

De acordo com a relação apresentada no Capítulo 2, entre a função de sobrevivência e a função de risco acumulada, a função de sobrevivência é estimada então por

$$S^*(t) = \exp(-H^*(t)). \tag{4.3}$$

Alternativamente, baseada no estimador de Kaplan-Meier, a função de sobrevivência pode ser estimada como

$$S^{**}(t) = \prod_{t_k \leq t} [1 - (Z(t_k)I_k)'x(t)].$$

A função de sobrevivência estimada não é necessariamente monótona sobre todo o período de observação. Ela pode aumentar para alguns valores de $t$ e de acordo com a equação (4.3) decrescer para algum $t$.

## 4.4 Teste para os efeitos das covariáveis

É freqüentemente de interesse testar se uma covariável específica tem algum efeito na função de risco total. Para o modelo aditivo de Aalen isto corresponde a testar a hipótese nula de que não existe efeito da covariável no risco. A hipótese nula para algum $j \geq 1$ é estabelecida como

$$H_j : \alpha_j(t) = 0, \qquad t \in [0, T]$$

É importante lembrar que no contexto não-paramétrico a hipótese nula acima pode apenas ser testada sobre intervalos de tempo onde $Y(t)$ tem posto completo. Dentro da estrutura do modelo, Aalen (1980, 1989) desenvolveu para todo tempo de falha uma estatística de teste para $H_j$ dada pelo j−ésimo elemento $U_j$ do vetor

$$U = \sum_{t_k} K(t_k)Z(t_k)I_k, \tag{4.4}$$

em que $K(t)$, uma função peso não negativa, é uma matriz diagonal $(p+1)\times(p+1)$. A estatística de teste da Equação (4.4) surge como uma combinação ponderada da soma do estimador de $A_j(t)$ apresentado na equação (4.2). Os elementos diagonais de $K(t)$ são funções pesos e suas escolhas podem depender das alternativas para a hipótese nula de interesse.

Uma escolha ótima da função peso necessitará do conhecimento das verdadeiras variâncias dos estimadores, entretanto isto dependerá de funções de parâmetros desconhecidas. Aalen considerou duas escolhas para a função peso. A primeira possibilidade é considerar cada função peso igual ao número de observações que permanecem no conjunto de risco em algum tempo dado. Neste caso a matriz $K(t)$ é substituída por um escalar $K_1(t_k)$ dado por

$$K_1(t_k) = \sum_{i=1}^{n} K_{1i}(t),$$

com $K_{1i} = \begin{cases} 1, & \text{se o i-ésimo indivíduo está sob risco no tempo t,} \\ 0, & \text{caso contrário.} \end{cases}$

Uma segunda escolha é tomar $K_2(t) = \{\text{diag}[(\text{Y(t)}'\text{Y(t)})^{-1}]\}^{-1}$, em que $K_2(t)$ é dada como a inversa de uma matriz diagonal tendo a mesma diagonal principal da matriz $(Y(t)'Y(t))^{-1}$. Este peso é escolhido por analogia ao problema da regressão de mínimos quadrados em que as variâncias dos estimadores são proporcionais aos elementos diagonais da matriz $(Y'Y)^{-1}$ sendo $Y$ o desenho da matriz. Estudos preliminares parecem indicar a escolha da segunda opção que pode ser mais poderosa em algumas situações. Neste trabalho foi utilizada esta última opção como função peso.

Um estimador da matriz de covariância de $U$ dado pela Equação (4.4) é

$$V = \sum_{t_k} K(t_k) Z(t_k) I_k^D Z(t_k)' K(t_k)'.$$

Suponha que se queira testar simultâneamente todos $H_j$ para $j$ em algum subconjunto $A$ de $\{1,\ldots,p\}$ consistindo de $s$ elementos. Seja $U_A$ definido como o subvetor correspondente de $U$ e $V_A$ a submatriz correspondente de $V$, isto é, $V_A$ é a matriz de covariâncias estimadas de $U_A$. A estatística de teste normalizada $U_A' V_A^{-1} U_A$ é assintoticamente distribuída como uma qui-quadrado com $s$ graus de liberdade quando $H_j$ vale para todo $j$ em $A$. Se o interesse é testar apenas uma das hipóteses $H_j$, então é usada a estatística de teste $U_j V_{jj}^{-1/2}$. Esta estatística tem uma distribuição assintótica normal padrão sob a hipótese nula.

Através da escolha de diferentes pesos, Lee e Weissfeld (1998) derivaram quatro novas estatísticas de testes para o modelo de riscos aditivos. A primeira função peso contém $K_1(t)$ como caso especial e é dada por uma função quadrada, contínua e integrável em $[0, 1]$. A segunda função peso derivada é uma combinação da primeira função peso proposta e de $K_2(t)$ e a terceira é baseada na estimativa de Kaplan-Meier. Por último a quarta função peso proposta combina esta última função e a função peso $K_2(t)$. Estas estatísticas foram comparadas com as duas estatísticas propostas por Aalen usando simulação de Monte Carlo e duas aplicações a dados reais. Das estatísticas propostas os autores verificaram que uma é superior para detectar diferenças no risco para tempos de sobrevivência grandes e uma outra é superior para detectar diferenças claras no risco e risco cruzado.

## 4.5 Gráfico das Funções de Regressão Acumuladas

Em estudos Clínicos ou tratamentos médicos a significância de uma covariável pode mudar durante o período de acompanhamento. Através do modelo de Aalen é possível estimar a contribuição das covariáveis para a função de risco em cada tempo de falha. O resumo desta contribuição sobre o tempo produz uma função de regressão para cada covariável que pode ser plotada em relação ao tempo. Ou seja, $A_j^*(t)$ pode ser considerada como uma função empírica descrevendo a influência da $j$-ésima covariável.

A inclinação do gráfico da função de regressão acumulada contra o tempo fornece informação sobre a influência de cada covariável, sendo possível verificar se uma covariável particular tem um efeito constante ou varia com o tempo ao longo do período de estudo. Por exemplo, se $\alpha_j(t)$ é constante, então o gráfico deve aproximar-se de uma linha reta. Inclinações positivas ocorrem durante períodos em que aumentos dos valores das covariáveis são associados com aumentos na função de risco. Por outro lado, inclinações negativas ocorrem em períodos quando crescimentos nos valores das covariáveis estão associados com decréscimos na função de risco. As funções de regressão acumuladas têm inclinações aproximadamente iguais a zero em períodos em que as covariáveis não influenciam a função risco. Portanto o gráfico das funções de regressão do modelo linear de Aalen pode ser recomendado também como um instrumento para detectar efeitos de covariáveis dependentes do tempo bem como uma técnica de diagnóstico que pode extrair informações adicionais úteis. Através de um estudo, onde foram apresentados os resultados de dois exemplos, Mau (Mau, 1986) mostrou que as funções de regressão podem fornecer informações importantes que devem ser perdidas quando apenas o modelo de Cox é aplicado.

# Capítulo 5

# Aplicações

## 5.1 Dados de Aleitamento Materno

### 5.1.1 Introdução

O aleitamento materno se constitui no principal componente de oportunidade nutricional, associando elementos fundamentais da nutrição correta: alimento, saúde e cuidados. No Brasil, devido o aumento das taxas de mortalidade infantil por desnutrição tendo como fator determinante a disseminação do aleitamento artificial, várias pesquisas de âmbito nacional foram feitas sobre a questão alimentar e nutricional do país. Evidenciou-se que havia uma tendência decrescente na prevalência da amamentação, conseqüência de mudanças de hábitos e comportamentos introduzidos pela vida moderna, tais como a participação da mulher no mercado de trabalho e a utilização do leite artificial. A zona da mata meridional de Pernambuco é uma área socialmente mais vulnerável em vários aspectos, entre eles na ocorrência da desnutrição, com características muito comuns no Nordeste Brasileiro. A principal atividade econômica é a monocultura açucareira, imprimindo características próprias à organização política e social no meio rural e urbano. A fragilidade dos sistemas produtivos alternativos (fruticultura, pequenas lavouras), a sazonalidade do desemprego, o subemprego em atividades subsidiárias de comércio urbano e da prestação de serviços avulsos contribuem ainda mais para que esta área seja mais exposta aos problemas de saúde e nutrição que resultam de condições desfavoráveis de vida. Uma das conseqüências decorrente dos fatos citados acima e que ocorre nesta região é a curta duração do aleitamento materno e a introdução precoce de leite artificial. Os fatos citados acima motivaram a realização de um estudo de coorte realizado com crianças acompanhadas do nascimento até os 18 meses de vida em áreas urbanas da zona da mata meridional do estado de Pernambuco. Foi utilizado um estudo de coorte pois possibilita o acompanhamento das ocorrências, das mudanças e o controle confiável dos fatores de risco relacionados às situações de interesse. Mesmo propiciando um custo

elevado e uma operacionalização freqüentemente difícil os estudos de seguimento populacional resultam em bancos de dados ricos descritivamente e analiticamente.

Este projeto de pesquisa foi realizado pelos Departamentos de Nutrição, Materno-Infantil e de Fisiologia e Farmacologia da Universidade Federal de Pernambuco em colaboração com a Universidade de Londres (LSHTM) e a Universidade de Motpelier-França. Trata-se de uma coorte de 652 crianças selecionadas através de amostragem sistemática (de um total de 1909 nascidos vivos) em 6 maternidades nas áreas urbanas dos municípios pernambucanos de Palmares, Catende, Água Preta e Joaquim Nabuco. O acompanhamento dessas crianças, selecionadas no período de setembro de 1997 a agosto de 1998, foi de setembro de 1997 a fevereiro de 2000. Para cada criança do sexo masculino nascida com peso inferior a 3100g, foi selecionada uma nascida logo a seguir com peso superior ou igual a 3100g e para as crianças do sexo feminino o ponto de corte foi de 3000g. Os critérios de exclusão adotados foram as malformações congênitas, hipóxia perinatal e gemelaridade. Nas primeiras 24 horas de vida foram avaliadas as medidas antropométricas e idade gestacional dos recém-nascidos. Também foram avaliadas as condições sócio-econômicas, ambientais e demográficas das famílias, assistência no pré-natal, características reprodutivas maternas, bem como o estado nutricional $post-partum$ através de indicadores antropométricos (peso e altura). O acompanhamento da morbidade e do aleitamento materno foi realizado através de visitas domiciliares realizadas duas vezes por semana nos primeiros 12 meses e uma vez por semana dos 12 aos 18 meses, sendo as informações prestadas pelas mães. As medidas antropométricas de peso, comprimento, perímetro cefálico e toráxico foram realizadas durante as visitas às residências aos 2, 4, 6, 9, 12, 15 e 18 meses de vida. O estado nutricional foi avaliado através dos indicadores peso/idade, comprimento/idade e peso/comprimento, usando-se o ponte de corte abaixo de $-2$ dos escores $Z$ para classificar déficit nutricional.

Para utilizar as técnicas de Análise de Sobrevivência descritas nos capítulos anteriores a variável de interesse selecionada foi o tempo de amamentação que é o tempo desde o nascimento até a criança parar de mamar. Dessa forma crianças que nunca mamaram foram retiradas do banco. Assim, o banco de dados utilizado corresponde a informações de 642 crianças das quais 118 (18.4%) são censuradas, ou seja, ainda mamavam quando terminou o estudo ou houve perda de acompanhamento.

O banco de dados utilizado nesta aplicação construído a partir dos arquivos da pesquisa é composto das seguintes colunas:

- NÚMERO: Número de identificação da criança.
- START: Covariável contínua que indica o tempo em dias das visitas para realização

das medidas de interesse.

• STOP: Covariável contínua que indica o tempo final em dias das visitas, sendo o último o tempo de falha ou censura da criança.

• EVENTO: Covariável dicotômica que assume o valor 0 em tempos cujo evento ainda não ocorreu, no caso da criança ainda estar mamando e assume o valor 1 no tempo em que ocorre a falha, ou seja, a criança parou de mamar. Para pacientes censurados esta variável sempre assume o valor 0.

• WAZ: Covariável contínua que informa o estado nutricional da criança, medida em cada visita. É a relação entre o peso observado e o peso considerado normal ou de referência por idade. Esta variável varia no tempo e assume valores entre $-4$ e 4. Para valores $< -2$ a criança é considerada com desnutrição grave ou moderada, valores no intervalo $[-2, -1)$ indicam desnutrição leve e valores $\geq -1$ caracterizam as crianças que se encontram em condições nutricionais normais. Neste trabalho essa covariável é tratada como contínua.

• HAZ: Covariável contínua que informa o estado nutricional da criança. É a relação entre a altura observada e a altura de referência por idade. Esta variável contínua também varia no tempo e a classificação nutricional das crianças é a mesma descrita acima para a covariável WAZ.

• SEXO: Covariável dicotômica que indica o sexo da criança, toma o valor 1 se a criança é do sexo masculino e 2 se a criança é do sexo feminino.

• QUANCONS: Variável discreta que mede a quantidade de consultas de pré-natal feitas durante a gravidez da criança em estudo.

• TRABGRAV: Variável categorizada que assume valor 1 se a mãe da criança trabalhou durante a gravidez e 2 se ela não trabalhou.

• CIGARROS: Variável contínua que expressa a quantidade média de cigarros fumados por dia durante a gravidez.

• BEBEU: Variável categorizada que assume valor 1 se a mãe ingeriu bebida alcoólica na maioria dos dias durante a gravidez e 2 caso contrário.

• IDMAE: Variável contínua que informa a idade da mãe em anos.

• ALTMAE: Variável contínua que indica a altura da mãe em cm.

• PESOMAE: Variável contínua que informa o peso da mãe em kg.

• ESTUDMAE: Variável discreta que expressa quantos anos a mãe tem de escolaridade.

• ESTUDPAI: Variável discreta que informa quantos anos o pai tem de escolaridade.

• RPERCAPT: Variável contínua que indica a renda per capta de cada domicílio. Esta variável é a renda total dos moradores do domicílio dividido pelo número total de pessoas que residem neste domicílio.

• IMCMAE: Variável contínua que mede o índice de massa corporal da mãe, calculado pela relação entre o peso (kg) e o quadrado da altura (m).

• TIPOCASA: Variável categorizada que expressa o regime de ocupação da residência, assumindo o valor 1 se a casa é própria ou alugada e 2 se a casa é cedida, invadida ou outros tipos.

• ÁGUA: Variável categorizada que informa de onde vem a água usada em casa. Assume o valor 1 se é da rede geral com ou sem canalização interna e 2 para outras opções como poço, nascente ou chafariz.

• LIXO: Variável categorizada que indica o destino do lixo assumindo o valor 1 se é coleta direta e 2 se é coleta indireta, queimado, enterrado, colocado em terreno baldio, entre outros.

• GELAD: Variável categorizada que toma o valor 1 se existe geladeira funcionando na residência e 2 se não existe.

### 5.1.2 Resultados Numéricos

Inicialmente foi utilizado o banco de dados com o acompanhamento das crianças até os 18 meses. Para uma análise preliminar foi feita a curva de sobrevivência do tempo de aleitamento, através do estimador de Kaplan-Meier. Pode-se observar (Figura 5.1) que a sobrevivência final é de aproximadamente $0,106$, ou seja, a probabilidade de uma criança continuar mamando além dos 18 meses é de $10,6\%$.

Ao aplicar os modelos de regressão de Cox e de Aalen foi verificado que as variáveis dependentes do tempo não foram significativas. Como nos 9 primeiros meses de acompanhamento $77,7\%$ das crianças já haviam parado de mamar, o banco de dados foi truncado aos 9 meses para que influências no tempo de aleitamento pudessem ser melhor percebidas. Através da análise da curva de sobrevivência apresentada na Figura 5.2 pode-se verificar que a sobrevivência final é de aproximadamente 0,223 ou seja a probabilidade de uma criança continuar mamando além dos 9 meses é de $22,3\%$. Uma outra informação que pode ser retirada desta figura é que apenas $32,4\%$ das crianças continuaram mamando após os 6 meses, o que corresponde a 183 dias aproximadamente.

**Figura 5.1: Curva de Sobrevivência para o Tempo de Aleitamento até os 18 meses.**

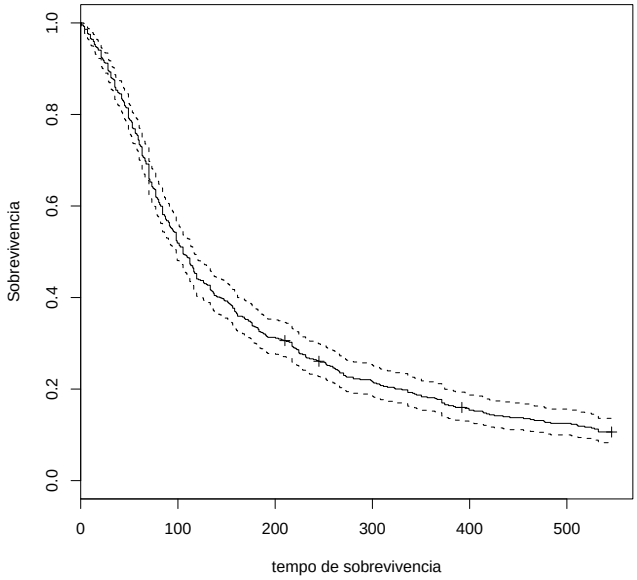
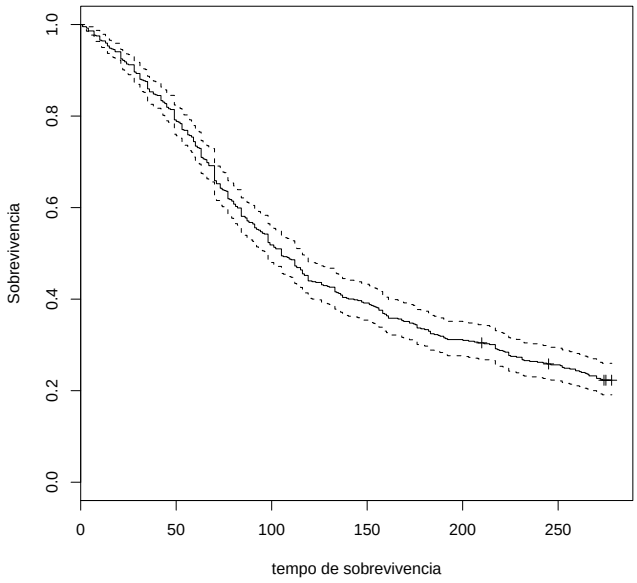



**Figura 5.2: Curva de Sobrevivência para o Tempo de Aleitamento até os 9 meses.**



Devido a presença da covariável dependente do tempo "waz", o risco relativo não é mais constante no tempo, isto é, foi violada a suposição de riscos proporcionais. Como o modelo de riscos proporcionais não é adequado para estes dados foi utilizado o modelo de regressão de Cox incluíndo covariáveis dependentes do tempo, a seguir descrito.

## Modelo de Cox com Covariáveis Dependentes do Tempo

Para identificar quais covariáveis dentre as pesquisadas influenciam no tempo de aleitamento foi utilizado o modelo de regressão de Cox com covariáveis dependentes do tempo. O procedimento utilizado foi o "backward", ou seja, iniciou-se com o modelo com todas as covariáveis explicativas, retirando-se uma a uma as covariáveis não significativas até chegar ao modelo final onde todas as covariáveis foram significativas. Observa-se nesse primeiro ajuste (Quadro 5.1) que a variável que se apresentou mais insignificante foi "idmae" que indica a idade da mãe devendo portanto ser a primeira covariável retirada do modelo.

**Quadro 5.1: Resultados do primeiro ajuste do Modelo de Cox com Covariáveis Dependentes do Tempo para os Dados do Aleitamento.**

| Covariável | Coeficiente | Erro Padrão | P-valor | Risco Relativo I.C (95%) |
|---|---|---|---|---|
| **waz** | $-0,204$ | 0,090 | 0,024 | 0,815(0,683;0,973) |
| **haz** | 0,128 | 0,096 | 0,180 | 1,137(0,942;1,372) |
| **sexo** | 0,0261 | 0,1084 | 0,810 | 1,026(0,830;1,269) |
| **quancons** | $-0,0345$ | 0,0258 | 0,180 | 0,966(0,918;1,016) |
| **trabgrav** | $-0,0387$ | 0,1311 | 0,770 | 0,962(0,744;1,244) |
| **cigarros** | 0,0265 | 0,0125 | 0,033 | 1,027(1,002;1,052) |
| **bebeu** | $-0,2953$ | 0,3016 | 0,330 | 0,744(0,412;1,344) |
| **idmae** | $-0,0004$ | 0,0101 | 0,970 | 1,000(0,980;1,020) |
| **estudmae** | 0,0309 | 0,0199 | 0,120 | 1,031(0,992;1,072) |
| **estudpai** | 0,0011 | 0,0194 | 0,950 | 1,001(0,964;1,040) |
| **rpercapt** | $-0,0005$ | 0,0006 | 0,430 | 1,000(0,998;1,001) |
| **imcmae** | 0,0152 | 0,0155 | 0,330 | 1,015(0,985;1,047) |
| **tipocasa** | $-0,2726$ | 0,1892 | 0,150 | 0,761(0,525;1,103) |
| **agua** | $-0,0704$ | 0,2242 | 0,750 | 0,932(0,601;1,446) |
| **lixo** | $-0,1034$ | 0,1293 | 0,420 | 0,902(0,700;1,162) |
| **gelad** | 0,1717 | 0,1270 | 0,180 | 1,187(0,926;1,523) |

Foram então sendo retiradas sucessivamente as covariáveis menos significantes e os resultados do modelo de regressão de Cox obtido após 15 ajustes estão apresentados no Quadro 5.2. Pode-se verificar que há evidências estatísticas de que as covariáveis "cigarros" e "tipocasa" foram identificadas como fatores influentes no tempo de aleitamento. Através dos dados apresentados no referido quadro pode-se verificar que se a criança mora em casa própria ou alugada (tipocasa=1), o risco dela parar de mamar é maior do que se ela mora em casa invadida ou cedida, o que indica que crianças com melhor condição financeira mamam menos. Nota-se também que se por exemplo a mãe fumou 20 cigarros por dia durante a gravidez o risco da criança parar de mamar é $1,64$ vezes o risco de uma criança cuja mãe não fumou durante a gravidez. Quanto maior o número de cigarros que a mãe fumou, menor é o tempo de aleitamento da criança.

**Quadro 5.2: Resultados do Ajuste Final do Modelo de Cox com Covariáveis Dependentes do Tempo para os Dados do Aleitamento.**

| Covariável | Coeficiente | Erro Padrão | P-valor | Risco Relativo I.C(95%) |
|---|---|---|---|---|
| **cigarros** | 0,0246 | 0,0098 | 0,012 | 1,025(1,01;1,04) |
| **tipocasa** | -0,2899 | 0,1571 | 0,065 | 0,748(0,55;1,02) |

No quadro 5.3 estão apresentados os testes estatísticos utilizados para a avaliação do modelo ajustado. De acordo com os três testes aplicados pode-se verificar que o modelo ajustado foi significativo com mais de 95% de confiança, ou seja, o modelo obtido explica bem os dados de aleitamento materno.

**Quadro 5.3: Testes para os parâmetros do modelo de Cox.**

| Teste | Valor | Graus de liberdade | P-valor |
|---|---|---|---|
| **Razão de Verossimilhança** | 9,06 | 2 | 0,0108 |
| **Wald** | 9,73 | 2 | 0,0077 |
| **Escore** | 9,87 | 2 | 0,0072 |

## Modelo Aditivo de Aalen

Como uma alternativa ao modelo de Cox foi utilizado o modelo aditivo de Aalen. Semelhante ao modelo de risco proporcional de Cox iniciou-se o ajuste partindo do modelo com todas as covariáveis até chegar ao modelo reduzido onde todas as covariáveis foram significativas. Verifica-se que no ajuste inicial, cujos resultados encontram-se apresentados no quadro 5.4, a variável "água" foi a que se apresentou menos significativa sendo assim a primeira variável a ser retirada do modelo. Pode-se observar também que os coeficientes das covariáveis "sexo" e "água" apareceram com sinais diferentes no primeiro ajuste do modelo de Cox (Quadro 5.1). Isto ocorre devido ao fato de que a não significância do parâmetro corresponde a aceitação da hipótese de que o mesmo é estatísticamente igual a zero não havendo assim problema de que estes coeficientes tenham sinais diferentes em ambos os modelos.

**Quadro 5.4: Resultados do Primeiro Ajuste do Modelo Aditivo de Aalen para os Dados do Aleitamento.**

| Covariável | Coeficiente | Erro Padrão | P-valor | I.C (95%) |
|---|---|---|---|---|
| **constante** | 1,673 | 1,451 | 0,116 | $(-1,171;4,517)$ |
| **waz** | $-0,265$ | 0,132 | 0,042 | $(-0,524;-0,005)$ |
| **haz** | 0,148 | 0,147 | 0,215 | $(-0,140;0,437)$ |
| **sexo** | $-0,007$ | 0,176 | 0,722 | $(-0,352;0,338)$ |
| **quancons** | $-0,042$ | 0,038 | 0,135 | $(-0,115;0,032)$ |
| **trabgrav** | $-0,206$ | 0,255 | 0,740 | $(-0,705;0,293)$ |
| **cigarros** | 0,075 | 0,054 | 0,169 | $(-0,030;0,180)$ |
| **bebeu** | $-0,128$ | 0,564 | 0,435 | $(-1,234;0,977)$ |
| **idmae** | $-0,006$ | 0,015 | 0,792 | $(-0,036;0,025)$ |
| **estudmae** | 0,039 | 0,033 | 0,156 | $(-0,026;0,105)$ |
| **estudpai** | 0,040 | 0,037 | 0,678 | $(-0,032;0,112)$ |
| **rpercapt** | $-0,001$ | 0,001 | 0,432 | $(-0,003;0,001)$ |
| **imcmae** | 0,019 | 0,025 | 0,443 | $(-0,030;0,068)$ |
| **tipocasa** | $-0,415$ | 0,216 | 0,186 | $(-0,839;0,008)$ |
| **agua** | 0,176 | 0,329 | 0,833 | $(-0,470;0,822)$ |
| **lixo** | $-0,141$ | 0,221 | 0,549 | $(-0,575;0,293)$ |
| **gelad** | 0,255 | 0,216 | 0,210 | $(-0,168;0,677)$ |

Após 15 ajustes obteve-se o modelo final. Analisando-se os resultados apresentados no Quadro 5.5 constata-se que neste modelo as covariáveis "waz" e "tipocasa" apresentaram-se significativas, logo estas covariáveis influenciam no tempo de aleitamento. Pode-se observar que analogamente ao modelo de Cox crianças que moram em casa própria ou alugada mamam menos do que crianças que moram em casa invadida ou cedida. Constata-se também que quanto maior o valor da covariável "waz" menor o risco da criança parar de mamar, ou seja, crianças desnutridas mamam menos do que as crianças que se encontram em condições nutricionais normais.

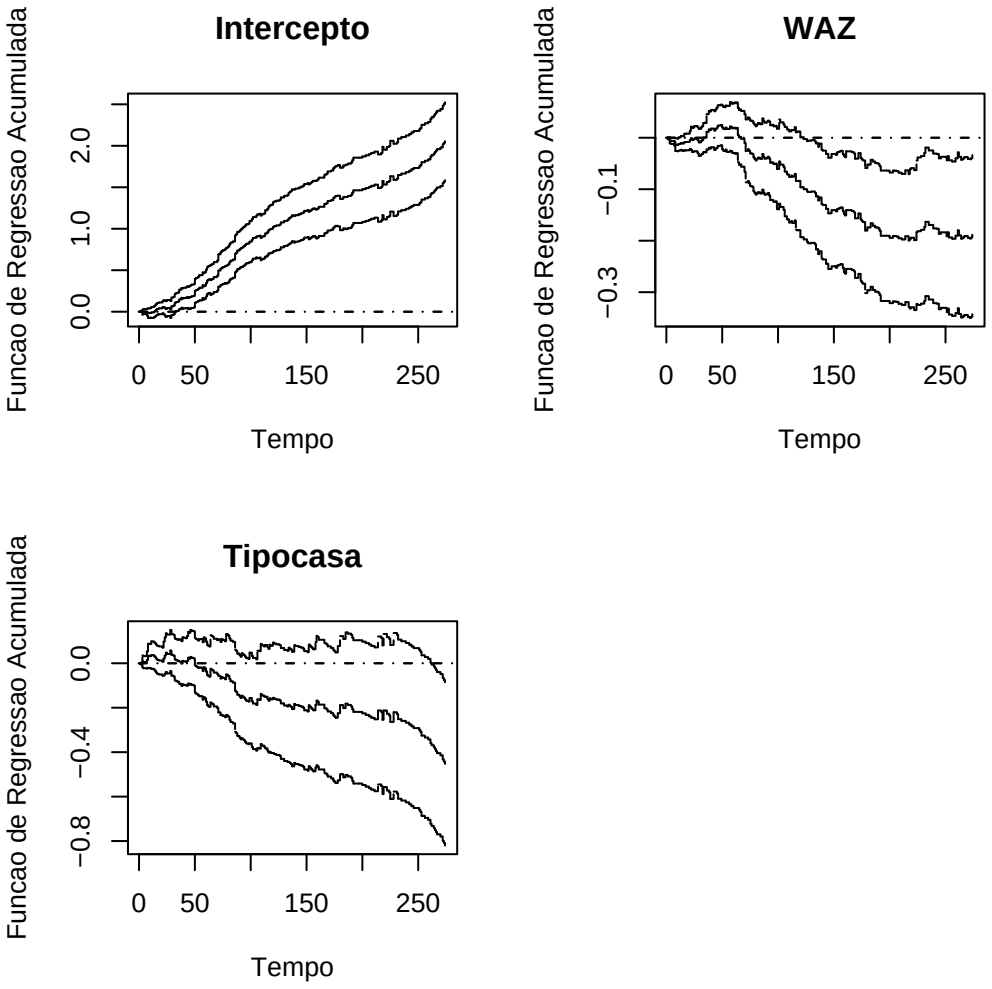
**Quadro 5.5: Resultados do Ajuste Final do Modelo Aditivo de Aalen para os Dados do Aleitamento.**

| **Covariável** | **Coeficiente** | **Erro Padrão** | **P-valor** | **I.C (95%)** |
|---|---|---|---|---|
| **constante** | 2,056 | 0,240 | 0,000 | (1,586;2,526) |
| **waz** | -0,188 | 0,079 | 0,049 | $(-0,343;-0,034)$ |
| **tipocasa** | -0,453 | 0,187 | 0,066 | $(-0,820;-0,085)$ |

## Gráfico das Funções de Regressão Acumuladas

Uma vantagem do modelo aditivo de Aalen é que ele é capaz de fornecer informações detalhadas a respeito da influência temporal de cada covariável. A Figura 5.3 mostra as funções de regressão acumuladas e os respectivos intervalos com 95% de confiança para o intercepto e para as covariáveis significativas no modelo de Aalen. A função de regressão acumulada para a covariável "waz" é não linear e apresenta evidências de que o efeito desta covariável diminue com o tempo e parece desaparecer após aproximadamente 6 meses de acompanhamento quando a função tem um comportamento paralelo ao eixo do tempo. O mesmo acontece com a covariável "tipocasa" contudo o seu efeito parece estabilizar em menos de 4 meses.

**Figura 5.3: Estimativas das Funções de Regressão Acumuladas com Intervalo de 95% de Confiança para os Dados do Aleitamento.**



## 5.2 Dados de Sinusite em Pacientes com Aids

### 5.2.1 Introdução

Estudos na área médica realizados em pacientes com AIDS ("Acquired Immnodeficiency Sindrome") indicam que a infecção pelo HIV ("Human Immunodeficiency") é fator de risco para o desenvolvimento de doenças otorrinolaringológicas (ORL). Os primeiros estudos não mencionaram a sinusite ou citaram apenas casos isolados de sinusite crônica. Sample, Lenahan e Serwonska et al. (1989) realizaram um estudo sobre manifestações ORL em pacientes com AIDS. Nos resultados obtidos 30% dos pacientes com AIDS apresentaram sinusite, sendo assim demonstrado que a sinusite era uma manifestação ORL mais freqüente do que citado em estudos anteriores. A sinusite se apresenta na AIDS com elevada freqüência e precária resposta aos tratamentos administrados. Entretanto a literatura mundial ainda não determinou a importância da sinusite no contexto geral da síndrome e outros fatores de riscos relacionados a estas manifestações em pacientes com AIDS têm sido pouco estudados.

Os dados utilizados nesta aplicação fazem parte de um estudo para avaliar a incidência de manifestações ORL em pacientes infectados pelo HIV realizado no Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG) (Gonçalves, 1995). Os pacientes que fizeram parte do estudo foram acompanhados no período de março de 1993 a fevereiro de 1995, considerando apenas os pacientes que entraram no estudo até julho de 1994.

Para entrar no estudo os pacientes tinham que ter idade superior a 15 anos, ter um exame prévio HIV positivo ou pertencer a grupos de comportamento de risco para adquirir o HIV, tais como, indivíduos que têm relações sexuais sem o uso de preservativo com parceiro desconhecido ou que possa ser portador do vírus, indivíduos que usam drogas endovenosas compartilhando agulhas e indivíduos que sofrem transfusões de sangue. Os pacientes incluídos no estudo foram encaminhados ao Centro de Treinamento e Referência em Doenças Infecto-Parasitárias (CTR-DIP) da cidade de Belo Horizonte, Minas Gerais. Após a primeira consulta os pacientes foram encaminhados ao serviço de Otorrinolaringologia do Hospital das Clínicas da UFMG.

As doenças ORL avaliadas baseadas nos estudos de prevalência destas manifestações na literatura em pacientes infectados pelo HIV foram: afta; candidíase oral; herpes labial; oral ou nasal; leucoplasia pilosa; sinusite; sarcoma de kaposi; otite aguda e serosa; lin-

fadenopatia cervical no cavum e na glândula parótida. Entre as enfermidades citadas foi utilizada neste trabalho apenas a sinusite. A classificação dos pacientes com relação a infecção pelo HIV foi de acordo com os critérios do CDC ("Centers of Disease Control", 1987) onde os pacientes foram classificados como HIV soronegativo, HIV soropositivo assintomático, pacientes com ARC ("AIDS Related Complex") ou pacientes com AIDS. Pacientes HIV soronegativo são aqueles que não possuem o HIV, constituindo o grupo controle do estudo. Pacientes HIV soropositivo assintomático são aqueles que possuem o vírus mas não desenvolveram o quadro clínico de AIDS e que apresentam um perfil imunológico estável. Os pacientes com ARC apresentam baixa imunidade e outros indicadores clínicos que antecedem o quadro clínico de AIDS. E, por último, os pacientes com AIDS são aqueles que já desenvolveram infecções oportunistas que segundo o critério do CDC de 1987 definem o quadro clínico de AIDS. O acompanhamento foi feito através de consultas trimestrais e o número mediano de consultas para cada paciente foi igual a quatro. A cada consulta a classificação do paciente foi reavaliada. Deste modo esta covariável indicadora do grupo depende do tempo, dado que os pacientes podem mudar de grupo ao longo do estudo.

Fizeram parte do estudo 112 pacientes, sendo 91 pacientes HIV positivo e 21 HIV negativo, dos quais aproximadamente 75% foram censurados. A variável de interesse foi o tempo desde a primeira consulta até a ocorrência de sinusite. A seguir estão descritas as covariáveis indicadoras de tempo, censura e as demais covariáveis incluídas no estudo que podem ou não ser consideradas como fator de risco para a ocorrência de sinusite.

O banco de dados utilizado nesta aplicação é composto das seguintes colunas:

- CÓDIGO: Número de identificação de cada paciente.

- START: Covariável contínua que indica o tempo em dias que os pacientes desenvolveram sinusite e mudaram de classificação quanto a infecção pelo HIV. Esta covariável é igual a zero quando o paciente entra no estudo.

- STOP: Covariável contínua que indica o tempo desde a primeira consulta até a morte, ocorrência de sinusite ou censura destes pacientes.

- EVENTO: Covariável dicotômica que assume o valor 0 em tempos cujo evento ainda não ocorreu, ou seja, o paciente não adquiriu a sinusite, e assume o valor 1 no tempo em que ocorre a falha. Para pacientes censurados esta variável sempre assume valor 0.

- SEXO: Covariável dicotômica que indica o sexo do paciente, toma o valor 0 se o paciente é do sexo masculino e 1 se é do sexo feminino.

- IDADE: Covariável contínua que informa a idade do paciente em anos.

• GRUPO: Covariável categorizada que informa o nível de infecção pelo HIV de acordo com a classificação do CDC (1987). Assume o valor 1 para pacientes HIV soronegativo, 2 para pacientes HIV soropositivo assintomático, 3 para pacientes com ARC e 4 para pacientes com AIDS.

Para verificar a influência de cada grupo de classificação quanto a infecção pelo HIV foram construídas as seguintes variáveis indicadoras:

$$x_{2i}(t) = \begin{cases} 1, & \text{se o i-ésimo indivíduo no tempo t está no grupo 2;} \\ 0, & \text{caso contrário} \end{cases}$$

$$x_{3i}(t) = \begin{cases} 1, & \text{se o i-ésimo indivíduo no tempo t está no grupo 3;} \\ 0, & \text{caso contrário} \end{cases}$$

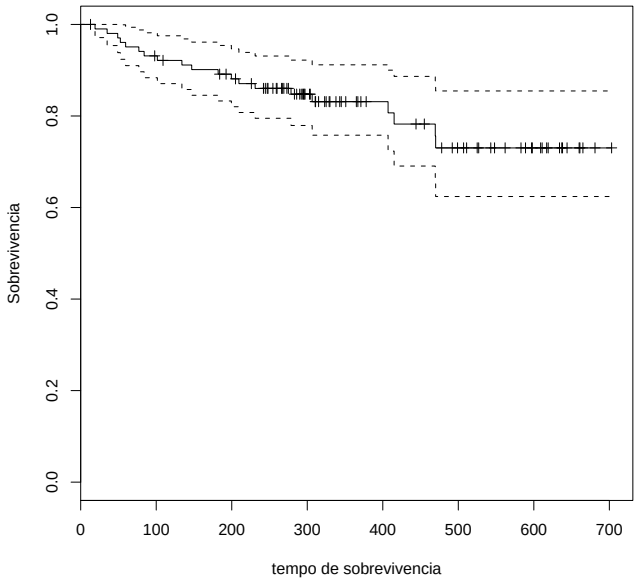
$$x_{4i}(t) = \begin{cases} 1, & \text{se o i-ésimo indivíduo no tempo t está no grupo 4;} \\ 0, & \text{caso contrário} \end{cases}$$

É bom lembrar que se o paciente estiver no grupo 1 em um determinado tempo, as três variáveis acima assumirão o valor zero neste tempo.

### 5.2.2 Resultados Numéricos

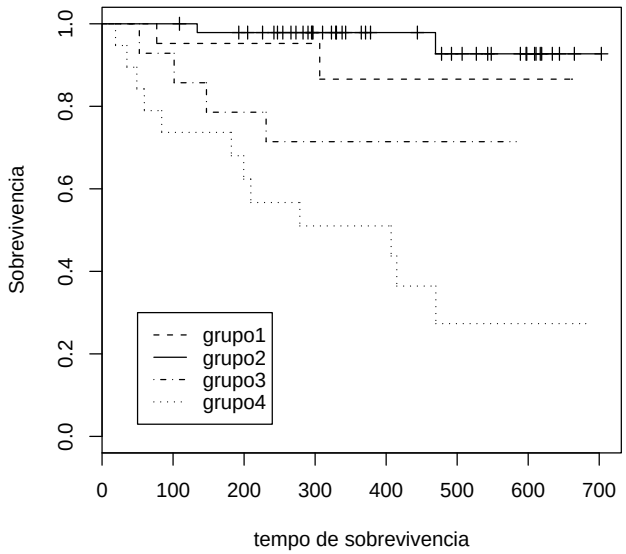
Inicialmente foi estimada a curva de sobrevivência, através do estimador de Kaplan-Meier, que pode ser visualizada na figura 5.5 a seguir. A sobrevivência final é de aproximadamente $0,73$, ou seja, o paciente tem uma probabilidade de 73% de não adquirir sinusite até o final do estudo que é caracterizado pelo maior acompanhamento que foi de aproximadamente de 700 dias.

**Figura 5.5: Curva de Sobrevivência para o Tempo até a Ocorrência de Sinusite em pacientes com AIDS.**




Foi verificado que dos pacientes incluídos no estudo 19, aproximadamente 14%, mudaram de grupo de classificação durante o período de acompanhamento. Assim, considerando que o grupo de classificação quanto a infecção pelo HIV pode ser um fator de risco para ocorrência de sinusite foram construídas curvas de sobrevivência para os quatro grupos de classificação. As curvas a seguir (Figura 5.6) referem-se aos pacientes do grupo 1, 2, 3 e 4 respectivamente. Pode-se constatar que os pacientes com AIDS (grupo = 4) tem a menor sobrevivência o que significa que estes pacientes têm uma maior probabilidade de adquirir sinusite do que os pacientes dos demais grupos. Os pacientes que fazem parte do grupo HIV soronegativo e HIV soropositivos assintomáticos, grupos 1 e 2 respectivamente não diferem muito com relação a sobrevivência e os que fazem parte do grupo de pacientes com ARC (grupo 3) apresentam uma sobrevivência final aproximada de $0,71$. O que significa que a probabilidade de um paciente do grupo 3 adquirir sinusite após o final do estudo é de aproximadamente 71%.

**Figura 5.6: Curvas de Sobrevivência para os diferentes níveis de infecção pelo HIV.**




Devido a presença da covariável "grupo", que varia no tempo, a suposição de riscos proporcionais foi violada. Como o modelo de riscos proporcionais não é adequado para estes dados foi então ajustado o modelo de regressão de Cox incluíndo covariáveis dependentes do tempo.

## Modelo de Cox com Covariáveis Dependentes do Tempo.

O modelo de regressão de Cox com covariáveis dependentes do tempo foi aplicado aos dados de sinusite, utilizando-se o procedimento "backward" como método de seleção das covariáveis. Observa-se no quadro 5.6 que a covariável "sexo" é não significativa, o que implica que será feito um novo ajuste sem a presença dessa covariável.

No Quadro 5.7 encontram-se os resultados do modelo de regressão de Cox ajustado no qual a idade e os grupos de riscos foram identificados como fatores que influenciam no desenvolvimento da sinusite. Apesar de não ser significativa a covariável "$x_2$", que indica que o paciente é HIV soropositivo assintomático, continuou no modelo devido ao fato de que esta covariável representa um dos grupos de classificação quanto a infecção pelo HIV e sua interpretação tem relevância na prática em relação aos demais grupos de classificação. Assim o risco de desenvolver sinusite em pacientes HIV soropositivo

**Quadro 5.6: Resultados do Primeiro Ajuste do Modelo de Cox com Covariáveis Dependentes do Tempo para os Dados de Sinusite em Pacientes com Aids.**

| Covariável | Coeficiente | Erro Padrão | P-valor | Risco Relativo I.C. (95%) |
|---|---|---|---|---|
| **idade** | $-0,0793$ | 0,0324 | 0,01400 | 0,924(0,8670;0,984) |
| **sexo** | 0,1546 | 0,4910 | 0,7500 | 1,167(0,4459;3,055) |
| **HIV Assint.**($x_2$) | $-0,7225$ | 1,0009 | 0,4700 | 0,486(0,0683;3,453) |
| **com ARC** ($x_3$) | 2,3102 | 0,8463 | 0,0063 | 10,077(1,9183;52,934) |
| **com AIDS** ($x_4$) | 2,7030 | 0,8112 | 0,0009 | 14,925(3,0438;73,184) |

assintomáticos não difere significativamente (nível de significância - $p = 0,47$) do grupo HIV soronegativo. Contudo pacientes que fazem parte do grupo com ARC ou com AIDS têm um risco maior de desenvolver a sinusite do que os pacientes HIV soronegativos. No grupo com ARC este risco é cerca de 10 vezes ($\exp(2,273)$) o risco daqueles no grupo HIV soronegativo e, no grupo com AIDS, o risco chega a ser 14 vezes ($\exp(2,649)$). Entretanto a precisão associada aos riscos dos grupos com ARC e com AIDS é bastante reduzida, o que pode ser comprovado ao se observar a grande amplitude de seus intervalos de confiança. Constata-se também que valores mais altos da idade do paciente diminuem o risco da ocorrência de sinusite, o que significa que pacientes mais jovens estão mais sujeitos a esta infecção.

**Quadro 5.7: Resultados do Ajuste Final do Modelo de Cox com Covariáveis Dependentes do Tempo para os Dados de Sinusite em Pacientes com Aids.**

| Covariável | Coeficiente | Erro Padrão | P-valor | Risco Relativo I.C. (95%) |
|---|---|---|---|---|
| **idade** | $-0,077$ | 0,0313 | 0,0140 | 0,926(0,8709;0,984) |
| **HIV Assint.**($x_2$) | $-0,730$ | 1,0006 | 0,4700 | 0,482(0,0678;3,424) |
| **com ARC** ($x_3$) | 2,273 | 0,8371 | 0,0066 | 9,705(1,8812;50,064) |
| **com AIDS** ($x_4$) | 2,649 | 0,7897 | 0,0008 | 14,141(3,0082;66,473) |

Os resultados apresentados para os três testes aplicados (Quadro 5.8) indicam que com 95% de confiança o modelo obtido foi significativo, ou seja, o modelo se ajusta bem aos dados de sinusite.

**Quadro 5.8: Testes para Avaliar o Ajuste do Modelo.**

| Teste | Valor | Graus de Liberdade | P-valor |
|---|---|---|---|
| **Razão de Verossimilhança** | 38,6 | 4 | 8,37E-08 |
| **Wald** | 25,9 | 4 | 3,32E-05 |
| **Escore** | 39,2 | 4 | 6,41E-08 |

## Modelo Aditivo de Aalen.

Os resultados do ajuste do modelo aditivo de Aalen para os dados de sinusite encontram-se no Quadro 5.9. Nota-se que a variável "sexo" não é significativa e portanto deve ser retirada do modelo.

**Quadro 5.9: Resultados do Primeiro Ajuste do Modelo Aditivo de Aalen para os Dados de Sinusite em Pacientes com Aids.**

| Covariável | Coeficiente | Erro Padrão | P-valor | I.C (95%) |
|---|---|---|---|---|
| **constante** | 1,051 | 0,415 | 0,005 | (0,238;1,865) |
| **idade** | $-0,033$ | 0,015 | 0,013 | $(-0,062;-0,004)$ |
| **sexo** | 0,063 | 0,195 | 0,889 | $(-0,319;0,445)$ |
| **HIV Assint. ($x_2$)** | 0,004 | 0,140 | 0,463 | $(-0,270;0,278)$ |
| **com ARC ($x_3$)** | 0,917 | 0,371 | 0,020 | (0,189;1,644) |
| **com AIDS ($x_4$)** | 1,566 | 0,545 | 0,000 | (0,497;2,635) |

As covariáveis idade do paciente e os grupos de riscos foram consideradas como fatores influentes na ocorrência da sinusite (ver Quadro 5.10). Assim como no modelo de Cox a covariável "$x_2$" que indica o grupo HIV soropositivo assintomático não apresentou coeficiente significativo mas permaneceu no modelo por representar um dos grupos de classificação quanto a infecção pelo HIV, portanto o risco de desenvolver sinusite em pacientes HIV soropositivo assintomáticos não difere significativamente (nível de significância - $p = 0,498$) do grupo HIV soronegativo. Pacientes que fazem parte do grupo com AIDS têm um risco maior de desenvolver a sinusite do que os pacientes dos demais grupos de classificação. Comparando-se com o grupo HIV soronegativo este risco é de

aproximadamente 1,6 vezes. Verifica-se também, por exemplo, que um aumento de 20 anos na idade do paciente diminue em $0,66$ vezes o risco de ocorrência da sinusite, o que confirma que quanto maior a idade do paciente menor o risco de desenvolvimento desta doença. O parâmetro não ser significativo corresponde a aceitar a hipótese de que o mesmo é estatísticamente igual a zero, portanto não existe problema no fato de que o coeficiente estimado da variável "$x_2$" tenha sido negativo no modelo de Cox e positivo no modelo de Aalen.

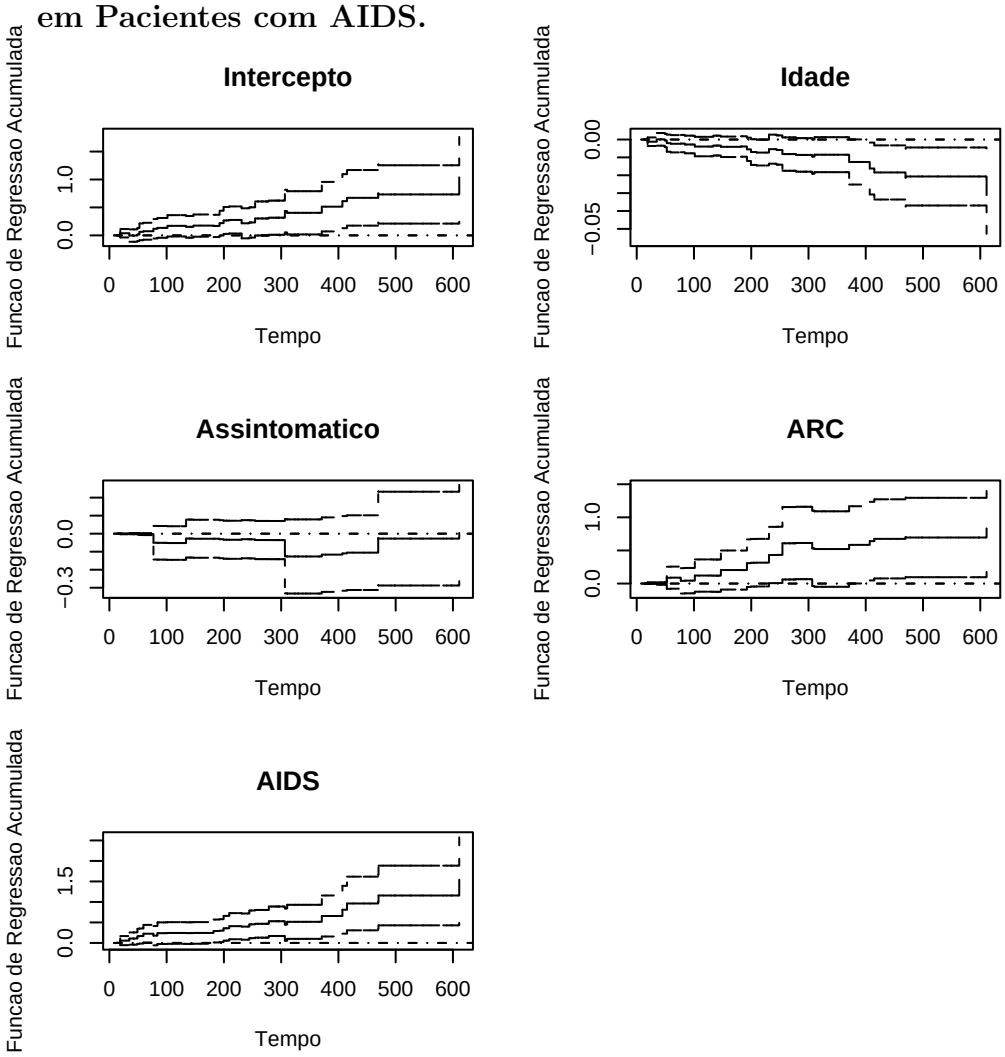
**Quadro 5.10: Resultados do Ajuste Final do Modelo Aditivo de Aalen para os Dados de Sinusite em Pacientes com Aids.**

| **Covariável** | **Coeficiente** | **Erro Padrão** | **P-valor** | **I.C (95%)** |
|---|---|---|---|---|
| **constante** | 1,038 | 0,405 | 0,004 | (0,244;1,831) |
| **idade** | $-0,031$ | 0,013 | 0,011 | $(-0,057;-0,005)$ |
| **HIV Assint.** ($x_2$) | 0,004 | 0,136 | 0,498 | $(-0,263;0,271)$ |
| **com ARC** ($x_3$) | 0,833 | 0,336 | 0,020 | (0,175;1,491) |
| **com AIDS** ($x_4$) | 1,544 | 0,536 | 0,000 | (0,493;2,595) |

## Gráfico das Funções de Regressão Acumuladas.

Através da análise gráfica das funções de regressão acumuladas contra o tempo, apresentadas na Figura 5.7, pode-se observar o comportamento do efeito de cada covariável significativa no modelo de Aalen. A função de regressão acumulada para a idade tem uma inclinação consistentemente negativa e seu efeito no risco da ocorrência da sinusite diminue razoalvemente com o tempo. Isto indica que crescimentos nos valores da idade, neste período, estão associados com decréscimos na função de risco. A covariável que indica o grupo com ARC parece ter uma influência clara e crescente por cerca de 10 meses com uma influência menor que parece desaparecer depois desse período.

**Figura 5.7: Estimativas das Funções de Regressão Acumuladas com Intervalo de 95% de Confiança para os Dados de Sinusite em Pacientes com AIDS.**



# Capítulo 6

# Conclusões

Em estudos de sobrevivência o modelo de riscos proporcionais de Cox é um método padrão de análise. O modelo aditivo de Aalen é uma alternativa ao de Cox bastante útil principalmente na presença de covariáveis dependentes do tempo pois permite que tanto as covariáveis como os parâmetros variem no tempo fornecendo assim informações detalhadas a respeito da influência temporal de cada covariável. Neste trabalho foram apresentados estes dois modelos e duas aplicações envolvendo dados censurados na presença de covariáveis cujos valores variam no tempo. Os modelos de Cox e Aalen foram utilizados para analisar estes dados.

Na primeira aplicação foi estudado o tempo de aleitamento materno. Foi observado que aos 6 meses aproximadamente apenas 32% das crianças continuaram mamando e que ao final dos 9 meses de acompanhamento aproximadamente 78% das crianças já haviam parado de mamar. Pode-se concluir então que ocorre uma curta duração do aleitamento materno pois segundo as recomendações da Organização Mundial da Saúde (OMS) as crianças devem mamar sem a introdução de nenhuma outra alimentação até os 6 meses no mínimo e isto não ocorre na região estudada. Foi usado o modelo de Cox generalizado para incluir o efeito de covariáveis dependentes do tempo na duração do aleitamento materno. Os resultados obtidos a partir da análise estatística mostram que tanto o tipo de casa que a criança mora quanto a quantidade de cigarros que a mãe fumou por dia durante a gravidez influenciam na duração do aleitamento materno. Por outro lado utilizando o modelo aditivo de Aalen observou-se que a variável que indica o tipo de casa continuou sendo significativa mas ao invés da quantidade de cigarros a variável que apresentou-se influente no tempo de aleitamento foi a variável dependente do tempo "waz" (escore $z$ peso/idade) que indica a condição nutricional da criança. Este fato confirma a suposição de que o modelo de Aalen é mais sensível a efeitos das covariáveis que variam no tempo do que o modelo de Cox. Uma diferença importante entre os dois modelos é que como no modelo de Aalen os parâmetros também variam no tempo é possível construir um gráfico das funções de regressão acumuladas onde é possível verificar o comportamento da influência de cada covariável no tempo. Foi verificado que tanto a situação nutricional quanto a variável tipo de casa diminuem sua influência no tempo de aleitamento materno com o passar do tempo.

Uma outra aplicação foi feita para verificar quais os fatores de risco para o desenvolvimento da sinusite, entre eles a infecção pelo HIV. Foi observado que pacientes com AIDS têm uma maior probabilidade de desenvolver sinusite pois os mesmos apresentaram a menor sobrevivência. Por exemplo pacientes com AIDS tem 27% de risco de não desenvolver a sinusite ao final do estudo enquanto que pacientes HIV soronegativos têm aproximadamente 87%. Com relação aos modelos de Cox e de Aalen foi verificado que tanto a idade quanto a infecção pelo HIV são fatores de risco para o desenvolvimento da sinusite e que a progressão da imunodeficiência aumenta de forma significante este risco, ou seja a contribuição no risco para pacientes com AIDS é maior do que a contribuição dos pacientes HIV soropositivos assintomáticos. Através do gráfico das funções de regressão acumuladas do modelo de Aalen foi constatado que a influência da idade no risco de desenvolvimento da sinusite diminue com o tempo ao passo que o efeito dos grupos quanto a infecção pelo HIV aumenta com o tempo. Este gráfico representa uma ferramenta importante na análise dos dados pois covariáveis medidas no início do período de observação podem freqüentemente perder sua influência no tempo sendo útil ter um método que revela isto.

# Apêndice A

# Bancos de Dados

Este apêndice apresenta partes dos bancos de dados organizados na forma que devem ser lidos no R.

- Parte do Banco de Dados do Aleitamento Materno

| numero | start | stop | evento | waz | haz | sexo | quancons | trabgrav | cigarros | bebeu | idmae |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A001B | 0 | 56 | 0 | -0.68 | -0.08 | 1 | 4 | 2 | 0 | 2 | 16 |
| A001B | 56 | 94 | 1 | 0.32 | -0.02 | 1 | 4 | 2 | 0 | 2 | 16 |
| A001R | 0 | 42 | 1 | -0.53 | -1.13 | 1 | 4 | 2 | 0 | 2 | 17 |
| A001S | 0 | 56 | 0 | -0.63 | -0.56 | 1 | 2 | 2 | 0 | 2 | 16 |
| A001S | 56 | 67 | 1 | -1.14 | -1.39 | 1 | 2 | 2 | 0 | 2 | 16 |
| A003T | 0 | 56 | 0 | -0.21 | -0.12 | 1 | 3 | 2 | 0 | 2 | 20 |
| A003T | 56 | 105 | 1 | 0.32 | -0.69 | 1 | 3 | 2 | 0 | 2 | 20 |
| A005A | 0 | 56 | 0 | -1.36 | -1.39 | 1 | 7 | 2 | 5 | 2 | 38 |
| A005A | 56 | 119 | 0 | -0.90 | -1.31 | 1 | 7 | 2 | 5 | 2 | 38 |
| A005A | 119 | 182 | 0 | -0.02 | -0.92 | 1 | 7 | 2 | 5 | 2 | 38 |
| A005A | 182 | 273 | 0 | 0.17 | -1.04 | 1 | 7 | 2 | 5 | 2 | 38 |
| A005B | 0 | 56 | 0 | -0.83 | -1.00 | 1 | 3 | 2 | 0 | 2 | 16 |
| A005B | 56 | 105 | 1 | -0.42 | -0.61 | 1 | 3 | 2 | 0 | 2 | 16 |
| A005C | 0 | 56 | 0 | -1.75 | -1.65 | 1 | 99999 | 2 | 20 | 2 | 15 |
| A005C | 56 | 116 | 1 | -1.62 | -2.49 | 1 | 99999 | 2 | 20 | 2 | 15 |
| A009A | 0 | 57 | 0 | -1.12 | -1.48 | 1 | 6 | 2 | 0 | 2 | 31 |
| A009A | 57 | 120 | 0 | -0.33 | -0.41 | 1 | 6 | 2 | 0 | 2 | 31 |
| A009A | 120 | 183 | 0 | -0.60 | -0.91 | 1 | 6 | 2 | 0 | 2 | 31 |
| A009A | 183 | 264 | 1 | -0.56 | -0.68 | 1 | 6 | 2 | 0 | 2 | 31 |
| A009B | 0 | 57 | 0 | -1.36 | -1.39 | 1 | 2 | 2 | 0 | 2 | 23 |
| A009B | 57 | 92 | 1 | 0.15 | -0.26 | 1 | 2 | 2 | 0 | 2 | 23 |
| A009T | 0 | 56 | 0 | 0.80 | 1.10 | 1 | 0 | 2 | 0 | 2 | 19 |
| A009T | 56 | 59 | 1 | -0.90 | -0.49 | 1 | 0 | 2 | 0 | 2 | 19 |
| A010R | 0 | 55 | 0 | -0.10 | -0.12 | 2 | 6 | 2 | 0 | 2 | 26 |
| A010R | 55 | 118 | 0 | 0.57 | 0.81 | 2 | 6 | 2 | 0 | 2 | 26 |
| A010R | 118 | 181 | 0 | 0.03 | 0.67 | 2 | 6 | 2 | 0 | 2 | 26 |
| A010R | 181 | 272 | 0 | -0.67 | 0.55 | 2 | 6 | 2 | 0 | 2 | 26 |
| A011A | 0 | 56 | 0 | -1.80 | -1.26 | 1 | 6 | 1 | 0 | 2 | 29 |
| A011A | 56 | 119 | 0 | -0.90 | -0.96 | 1 | 6 | 1 | 0 | 2 | 29 |
| A011A | 119 | 182 | 0 | -0.88 | -1.30 | 1 | 6 | 1 | 0 | 2 | 29 |
| A011A | 182 | 222 | 1 | -0.80 | -1.04 | 1 | 6 | 1 | 0 | 2 | 29 |
| A011R | 0 | 56 | 0 | 0.74 | 0.18 | 1 | 6 | 2 | 0 | 2 | 18 |
| A011R | 56 | 73 | 1 | 0.97 | -0.10 | 1 | 6 | 2 | 0 | 2 | 18 |

- Parte do Banco de Dados da Incidência de Sinusite em Pacientes com AIDS

| código | idade | sexo | grupo | start | stop | evento | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|---|---|---|---|---|
| 113079 | 22 | 1 | 2 | 0.00 | 378.00 | 0 | 1 | 0 | 0 |
| 144278 | 32 | 0 | 4 | 0.00 | 84.00 | 1 | 0 | 0 | 1 |
| 151746 | 36 | 0 | 2 | 0.00 | 109.00 | 0 | 1 | 0 | 0 |
| 232207 | 34 | 0 | 2 | 0.00 | 134.00 | 1 | 1 | 0 | 0 |
| 286787 | 29 | 0 | 2 | 0.00 | 338.00 | 0 | 1 | 0 | 0 |
| 301326 | 29 | 1 | 3 | 0.00 | 311.00 | 0 | 0 | 1 | 0 |
| 318240 | 38 | 0 | 4 | 0.00 | 182.00 | 1 | 0 | 0 | 1 |
| 341597 | 32 | 1 | 1 | 0.00 | 77.00 | 1 | 0 | 0 | 0 |
| 360135 | 30 | 1 | 1 | 0.00 | 184.00 | 0 | 0 | 0 | 0 |
| 365435 | 33 | 0 | 2 | 0.00 | 543.00 | 0 | 1 | 0 | 0 |
| 370233 | 35 | 1 | 1 | 0.00 | 286.00 | 0 | 0 | 0 | 0 |
| 378530 | 41 | 0 | 4 | 0.00 | 470.00 | 1 | 0 | 0 | 1 |
| 385496 | 31 | 0 | 4 | 0.00 | 407.00 | 1 | 0 | 0 | 1 |
| 386739 | 48 | 1 | 3 | 0.00 | 231.00 | 1 | 0 | 1 | 0 |
| 389169 | 31 | 0 | 2 | 0.00 | 205.00 | 0 | 1 | 0 | 0 |
| 398075 | 21 | 1 | 1 | 0.00 | 637.00 | 0 | 0 | 0 | 0 |
| 401254 | 22 | 1 | 1 | 0.00 | 345.00 | 0 | 0 | 0 | 0 |
| 402682 | 32 | 0 | 1 | 0.00 | 638.00 | 0 | 0 | 0 | 0 |
| 407776 | 37 | 1 | 1 | 0.00 | 292.00 | 0 | 0 | 0 | 0 |
| 408401 | 25 | 0 | 0 | 0.00 | 294.00 | 0 | 0 | 0 | 0 |
| 408971 | 99 | 0 | 2 | 0.00 | 471.50 | 0 | 1 | 0 | 0 |
| 408971 | 99 | 0 | 4 | 471.50 | 507.00 | 0 | 0 | 0 | 1 |
| 411263 | 34 | 1 | 3 | 0.00 | 141.50 | 0 | 0 | 1 | 0 |
| 411263 | 34 | 1 | 4 | 141.50 | 244.50 | 1 | 0 | 0 | 1 |
| 415002 | 31 | 0 | 4 | 0.00 | 49.00 | 1 | 0 | 0 | 1 |
| 415438 | 27 | 0 | 4 | 0.00 | 511.00 | 0 | 0 | 0 | 1 |

# Apêndice B

# Métodos Computacionais

Este apêndice apresenta os comandos do R utilizados na elaboração desta dissertação de mestrado.

- Banco de Dados Referente ao Aleitamento Materno

```
# Ler os Dados #

bsimples61<-read.table("c:/alunos/tarci/tese/aleitamento/bsimples61.dat"
,header=F)

# Nomear as Variaveis #

numero61<-bsimples61[,1]
tempo61<-bsimples61[,2]
censura61<-bsimples61[,3]

# Curva de Sobrevivencia #

plot(survfit(Surv(tempo61,censura61)),conf.int=T,xlab="tempo de sobrevivencia",
ylab="Sobrevivencia")


# Ler o Banco com as Covariaveis Dependentes do Tempo #

bancotese4<-read.table("c:/alunos/tarci/tese/aleitamento/banco6meses1.dat",
header=F)

# Nomear as Variaveis #

numero61<-bancotese4[,1]
start61<-bancotese4[,2]
stop61<-bancotese4[,3]
event61<-bancotese4[,4]
pesonew61<-bancotese4[,5]
compnew61<-bancotese4[,6]
waznew61<-bancotese4[,7]
haznew61<-bancotese4[,8]
sexo61<-bancotese4[,9]
quancons61<-bancotese4[,10]
```

```
trabgrav61<-bancotese4[,11]
cigarros61<-bancotese4[,12]
bebeu61<-bancotese4[,13]
maele61<-bancotese4[,14]
paile61<-bancotese4[,15]
totmora61<-bancotese4[,16]
idmae61<-bancotese4[,17]
altmae61<-bancotese4[,18]
pesomae61<-bancotese4[,19]
estudmae61<-bancotese4[,20]
estudpai61<-bancotese4[,21]
rpercapt61<-bancotese4[,22]
imcmae61<-bancotese4[,23]
tipocasa61<-bancotese4[,24]
agua61<-bancotese4[,25]
lixo61<-bancotese4[,26]
gelad61<-bancotese4[,27]
tempo61<-bancotese4[,28]
censura61<-bancotese4[,29]

# Colocando valor de missing como NA para trabalhar no R #

paile61[paile61==99999]<-NA
summary(paile61)
quancons61[quancons61==99999]<-NA
summary(quancons61)
cigarros61[cigarros61==99999]<-NA
summary(cigarros61)
estudpai61[estudpai61==99999]<-NA
summary(estudpai61)
rpercapt61[rpercapt61==99999]<-NA
summary(rpercapt61)
consucod61[consucod61==99999]<-NA
summary(consucod61)
espaicod61[espaicod61==99999]<-NA
summary(espaicod61)
rpcapcod61[rpcapcod61==99999]<-NA
summary(rpcapcod61)

# Colocando como data frame para Ajustar o Modelo de Cox #

str(d6meses1 <- data.frame(cbind(numero61,start61, stop61,event61,waznew61,
haznew61,sexo61,quancons61,trabgrav61,cigarros61,bebeu61,idmae61,estudmae61,
estudpai61,rpercapt61,imcmae61,tipocasa61,agua61,lixo61,gelad61,tempo61,
censura61)))

# Ajustando o Modelo de Cox com Covariaveis Dependentes do Tempo #
```

```
cox1<-coxph( Surv(start61, stop61, event61) ~waznew61+haznew61+sexo61+
quancons61+trabgrav61+cigarros61+bebeu61+idmae61+estudmae61+estudpai61+
rpercapt61+imcmae61+tipocasa61+agua61+lixo61+gelad61 , data=d6meses1)
summary(cox1)

cox2<-update(cox1,~.-totmora)
summary(cox2)

cox3<-update(cox2,~.-altmae)
summary(cox3)

cox4<-update(cox3,~.-trabgrav)
summary(cox4)

# Ajustando o Modelo de Aalen #

aalen6meses1<-addreg(Surv(start61, stop61, event61) ~waznew61+haznew61+sexo61+
quancons61+trabgrav61+cigarros61+bebeu61+idmae61+estudmae61+estudpai61+
rpercapt61+imcmae61+tipocasa61+agua61+lixo61+gelad61 , data=d6meses1)

aalen6meses15<-addreg(Surv(start61, stop61, event61) ~waznew61+tipocasa61,
data=d6meses1)

# Grafico dos Coeficientes de Regressao Acumulados
do Modelo de Aalen #

plot(aalen6meses15,xlab="Tempo",ylab="Funcao de Regressao Acumulada")


# Grafico dos Residuos para Avaliar o Ajuste do Modelo de Aalen #

tempomamar<-aalen6meses15$times
coefmamar<-aalen6meses15$increments
inter<-rep(1,1547)
covmamar<-cbind(inter,waz61,casa61)
covmamart<- t(covmamar)
riscomamar<-coefmamar%*%covmamart
diagonal<-diag(riscomamar)
riscoaleit<-diagonal
for(i in 1:1546) {
riscoaleit[i+1]<-riscoaleit[i+1]+riscoaleit[i] }
plot(tempomamar,riscoaleit,xlab="Tempo",ylab="Risco Acumulado")
```

- Banco de Dados Referente a Incidência da Sinusite em Pacientes com AIDS

```
# Ler os Dados #

aidsimples<-matrix(scan("c:/alunos/tarci/tese/aleitamento/aidsimples1.dat"),
ncol=2,byrow=T)

# Nomear as Variaveis #

tempoaids<-aidsimples[,1]
censaids<-aidsimples[,2]

# Curva de Sobrevivencia #

Saids<-survfit(Surv(tempoaids,censaids))
summary(Saids)
plot(Saids,xlab="tempo de sobrevivencia",ylab="Sobrevivencia")

# Curva de Sobrevivencia para cada grupo #

aidsgrupo1<-matrix(scan("c:/alunos/tarci/tese/aleitamento/aidsgrupo1.dat")
,ncol=2,byrow=T)
tempoaidsg1<-aidsgrupo1[,1]
censaidsg1<-aidsgrupo1[,2]
sobrevag1<-survfit(Surv(tempoaidsg1,censaidsg1))


aidsgrupo2<-matrix(scan("c:/alunos/tarci/tese/aleitamento/aidsgrupo2.dat"),
ncol=2,byrow=T)
tempoaidsg2<-aidsgrupo2[,1]
censaidsg2<-aidsgrupo2[,2]
sobrevag2<-survfit(Surv(tempoaidsg2,censaidsg2))

aidsgrupo3<-matrix(scan("c:/alunos/tarci/tese/aleitamento/aidsgrupo3.dat"),
ncol=2,byrow=T)
tempoaidsg3<-aidsgrupo3[,1]
censaidsg3<-aidsgrupo3[,2]
sobrevag3<-survfit(Surv(tempoaidsg3,censaidsg3))

aidsgrupo4<-matrix(scan("c:/alunos/tarci/tese/aleitamento/aidsgrupo4.dat"),
ncol=2,byrow=T)
tempoaidsg4<-aidsgrupo4[,1]
censaidsg4<-aidsgrupo4[,2]
sobrevag4<-survfit(Surv(tempoaidsg4,censaidsg4))

plot(sobrevag2,xlab="tempo de sobrevivencia",ylab="Sobrevivencia",col=1)
lines(sobrevag1,col=2)
lines(sobrevag4,col=3)
```

```
lines(sobrevag3,col=4)
legend(100,0.2,c("grupo1","grupo2","grupo3","grupo4"),col=c(2,1,4,3))


# Ler o Banco com as Covariaveis Dependentes do Tempo #

aids1<-read.table("c:/alunos/tarci/tese/aleitamento/aids1.txt",header=T)

# Nomear as Variaveis #

idade1<-aids1[,1]
sexo1<-aids1[,2]
grupo1<-aids1[,3]
start1<-aids1[,4]
stop1<-aids1[,5]
evento1<-aids1[,6]
x21<-aids1[,7]
x31<-aids1[,8]
x41<-aids1[,9]

# Colocando valor de missing como NA para trabalhar no R #

idade1[idade1==99]<-NA
summary(idade1)

# Colocando como data frame para Ajustar o Modelo de Cox #

str(daids1 <- data.frame(cbind(idade1,sexo1,grupo1,start1,
stop1,evento1,x21,x31,x41)))

# Ajustando o Modelo de Cox com Covariaveis Dependentes do Tempo #

coxaids1<-coxph( Surv(start1, stop1, evento1) ~idade1+sexo1+x21+x31+x41,
data=daids1)
summary(cox1)

coxaids2<-update(cox1,~.-sexo1)
summary(cox2)

# Ajustando o Modelo de Aalen #

aalenaids1<-addreg( Surv(start1, stop1, evento1) ~idade1+sexo1+x21+x31+x41,
data=daids1)
aalenaids2<-addreg( Surv(start1, stop1, evento1) ~idade1+x21+x31+x41,
data=daids1)


# Grafico dos Coeficientes de Regressao Acumulados
```

```
do Modelo de Aalen #

plot(aalenaids2,xlab="Tempo",ylab="Funcao de Regressao Acumulada",
labelofvariable=c("Intercepto","Idade","Assintomatico","ARC","AIDS"))

# Grafico dos Residuos para Avaliar o Ajuste do Modelo de Aalen #

tempoaids<-aalenaids2$times
coefaids<-aalenaids2$increments
interaids<-rep(1,112)
covaids<-cbind(interaids,idade2,x22,x32,x42)
covaidst<- t(covaids)
riscoaids<-coefaids%*%covaidst
haids<-diag(riscoaids)
riscoaids1<-diag(riscoaids)
for(i in 1:111) {
riscoaids1[i+1]<-riscoaids1[i+1]+riscoaids1[i] }
plot(tempoaids,riscoaids1,xlab="Tempo",ylab="Risco Acumulado")
```

# Referências

[1] Aalen, O. O. (1976). Nonparametric Inference in Connection with Multiple Decrement Models. *Scandinavian Journal Statistics*, **3**, 15–27.

[2] Aalen, O. O. (1978). Nonparametric Inference for a Family of Counting Processes. *Annals of Statistics*, **6**, 701–726.

[3] Aalen, O. O. (1980). A Model for Nonparametric Regression Analysis of Counting Processes.*Lecture Notes in Statistics* Springer, New York, **2**, 1–25.

[4] Aalen, O. O. (1989). A Linear Regression Model for the Analysis of Life Times. *Statistics in Medicine*, **8**, 907–925.

[5] Aalen, O. O. (1993). Further Results on the Non-Parametric Linear Regression Model in Survival Analysis. *Statist. Med.*, **12**, 1569–1588.

[6] Andersen, P. K. & Gill, R.(1982). Coxs Regression Model for a Counting Processes: A Large Sample Study. *Ann. Statistics*, **10**, 1100–1200.

[7] Andersen, P. K. & Vaeth, M.(1989). Simple parametric and non-parametric models for excess and relative mortality. *Biometrics*, **45** , 523–535.

[8] Andersen, P. K., Borgan, O., Gill, R.D. & Keiding, N.(1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

[9] Berkson, J. & Gage, R.R. (1950). Calculation of Survival Rates for Cancer. *Proceedings of Staff Meetings, Mayo Clinic*, **25**, 250.

[10] Breslow, N.(1972). Contribuição à Discussão do Artigo de D. R. Cox. *Journal of the Royal Statistical Society B*, **34**, 216–217.

[11] Breslow, N. & Crowley, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *Annals of Statistics*, **2**, 437–453.

[12] Collet, D. (1994). *Modelling Survival Data in Medical Research*. Londom: Chapman & Hall.

[13] Colosimo, E.A., Ferreira, F.F., Oliveira, M.D. & Souza, C.B. (2002). Empirical Comparisions Between Kaplan-Meier and Nelson-Aalen Survival Function. *Journal Stat. Comput. Sim.*, **72**, 299–308.

[14] Cox, D.R.(1972). Regression Models and Life Tables(with discussion). *Journal of the Royal Statistics Soc. B*, **34**, 187–220.

[15] Cox, D.R.(1975). Partial Likelihood. *Biometrika*, **62**, 269–276.

[16] Cribari-Neto, F. & Zarkos, S. G. (1999).R Yet Another Econometric Programming Environment. *Journal Appl. Econ.*, **14**, 319–329.

[17] Cutler, S.J. & Ederer, F. (1958). Maximum Utilization of the Life Table Method in Analysing Survival. *Journal of Chronic Diseases*, **8**, 699–712.

[18] Efron, B. (1967). The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, **72**, 557–565.

[19] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis.* John Wiley, New York.

[20] Gehan, E.A. (1969). Estimating Surviaval Functions from the Life Table. *Journal of Chronic Diseases*, **21**, 629–644.

[21] Gonçalves, D.U.(1995). Incidência, Marcadores de Prognóstico e Fatores de Risco Relacionados às Manifestações Otorrinolaringológicas em Pacientes Infectados pelo HIV. *Dissertação de Mestrado - Faculdade de Medicina/UFMG, Belo Horizonte.*

[22] Hill, A. B. (1984). *A Short Textbook of Medical Statistics.* London: Hodder e Stoughton.

[23] Huffer, F. W. & McKeague, I.W. (1987). *Survival analysis using additive risk models.* Technical Report 396, Department of Statistics, Stanford University.

[24] Kalbfleisch & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* Wiley Series in Probability and Mathematical Statistics.

[25] Kaplan, E.L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457–481.

[26] Klein, J. P. & Moeschberger, M.L. (1997). *Survival Analysis: Tecniques for Censored and Truncated Data.* Springer-Verlag, New York.

[27] Knuth, D.E. (1986). *The TEXbook*. New York: Addison–Wesley.

[28] Lee, E.T. (1992). *Statistical Methods for Survival Data Analysis.* John Wiley, New York.

[29] Lee, E. & Weissfeld, L. A. (1998). Assessment of Covariate Effects in Aalen's Additive Hazard Model. *Statistics in Medicine*, **17**, 983–998.

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

[31] Le, C.T. (1997). *Applied Survival Analysis.* John Wiley e Sons, New York.

[32] Mau, J. (1986). On a graphical method for the detection of time-dependent effects of covariates in survival data. *Applied Statistics*, **35**, 245–255.

[33] Mau, J. (1988). A Comparison of counting process models for complicated life histories. *Applied Stochastic Models and Data Analysis*, **4**, 283–298.

[34] McKeague, I. W. (1986). Estimation for a semimartingale regression model using the method of sieves. *Annals of Statistics*, **14**, 579–589.

[35] Meier, P. (1975). Statistics and Medical Experimentation. *Biometrics*, **31**, 511–529.

[36] Nelson, W.B. (1970). *Statistical Methods for Accelerated Life Test Data - The Inverse Power Law Model. General Eletric Corporate Research and Development.* T15 Report 71-C-011.

[37] Nelson, W. (1972). Theory and Application of Hazard Plotting for Censored Survival Data. *Biometrics*, **14**, 945–966.

[38] Nelson, W. (1982). *Applied Life Data Analysis.* John Wiley and Sons, New York - NY.

[39] Nelson, W. (1990). *Statistical Models, Test Plans, and Data Analysis.* John Wiley and Sons, New York - NY.

[40] Peto, R.(1972). Contribuio a Discusso do Artigo de D. R. Cox. *Journal of the Royal Statistical Society B*, **34**, 205–207.

[41] Prentice, R.L. & Gloeckler, L. A. (1978). Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data. *Biometrics*, **34**, 57–67.

[42] Ramlau-Hansen, H. (1983). Smoothing Counting Process Intensities by Means of Kernel Functions. *Annals of Statistics*, **11**, 453–466.

[43] Sample, S., Lenahan, G.A.,Serwonska, M.H. et al. (1989). Allergic Diseases and Sinusitis in Acquired Immunodeficiency Syndrome. *Journal Allergy Clin. Immunol.*, **83**, 190.

[44] Soares, J. F. & Colosimo, E. A.(1995). *Métodos Estatísticos na Pesquisa Clínica*. 40ª RBRAS e 6º SEAGRO.

[45] Tsiatis, A.A. (1981). A Large Sample Study of Cox's Regression Model. *Annals of Statistics*, **9**, 93–108.