Universidade de Brasília
Departamento de Estatística

Aplicação de Técnicas de Análise de Sobrevivência para analisar Eventos Recorrentes

por

Thaís Alvares de Carvalho Oliveira

Monografia apresentada para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2016

Thaís Alvares de Carvalho Oliveira

Aplicação de Técnicas de Análise de Sobrevivência para analisar Eventos Recorrentes

Orientadora:
Profa. Dra. **Juliana Betini Fachini Gomes**

Monografia apresentada para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2016

Dedicatória

À minha mãe,
Veranilce *por sempre acreditar no meu potencial. Por ser incansável na tarefa de me educar e interceder por mim, junto à nossa mãe Virgem Maria.*

Ao meu pai,
Sebastião*, por ser meu exemplo de profissionalismo e racionalidade. Por fazer o impossível na busca de garantir que os obstáculos encontrados por suas filhas fossem infinitamente menores que os que ele mesmo superou.*

Agradecimentos

À Deus, que mesmo nos momentos que questionei minha capacidade me mostrou a luz no fim do túnel.

À orientadora, "psicóloga"e amiga Juliana, pela disposição e paciência desde meu primeiro semestre na UnB. Por toda sua atenção, preocupação e carinho, se mostrando sempre compreensiva nos momentos que mais precisei.

Aos meus pais, pelo suporte financeiro e emocional, todo o amparo e amor incondicional nas situações mais difíceis.

A todos os professores da UnB que contribuíram de alguma forma para a minha formação acadêmica e pessoal, em especial o professor Gladston, que sempre me mostrou os caminhos dentro do curso e foi responsável por despertar dentro de mim a paixão pela estatística.

À minha irmã Érika, por me pressionar durante todos esses anos, ser mais mãe que irmã, muitas vezes me tirando do sério, mas sempre servindo de modelo do que devo ou não seguir.

À minha irmã Aurilene, por ter sempre se mostrado compreensiva e disposta a me ajudar no que eu precisasse em todas as fases da minha vida.

À minha "única e verdadeira amiga"Daniela, que tem se mostrado uma irmã nesses quinze anos de amizade. Por todos os conselhos, brigas e "endoidadas".

Ao meu melhor amigo e companheiro Marcus, por sempre me ouvir e acreditar na minha inteligência, ser minha válvula de escape nas situações mais intensas e ter continuado ao meu lado no momento que mais precisei.

Aos amigos e colegas de curso que encontrei nessa jornada, em especial Letícia, João Gustavo, Brenda, Lena e João Marcos. Pessoas que estiveram sempre presentes ao longo da minha graduação fosse nos grupos de estudos, noites em claro em véspera de provas ou simples tardes de ócio nas dependências da UnB.

A todos que de alguma forma contribuíram para a realização deste trabalho e formação acadêmica, deixo aqui o meu muito obrigada.

"Yes, there are two paths you can go by, but in the long run, there's still time to change the road you're on."

Stairway to Heaven - Led Zeppelin

Sumário

Resumo

Aplicação de Técnicas de Análise de Sobrevivência para analisar Eventos Recorrentes

Esta monografia utiliza uma modelagem marginal como aplicação de técnicas de análise de sobrevivência a dados com eventos recorrentes. Os modelos Andersen & Gill e Peterson Williams e Peterson foram ajustados aos dados de diarreia infantil para avaliar o efeito da suplementação de vitamina A na redução da taxa da morbidade por diarreia. Os parâmetros do modelo foram estimados pelo método de máxima verossimilhança parcial e a proporcionalidade dos riscos foram testadas utilizando os resíduos padronizados de Schoenfeld. Foi utilizado um conjunto de dados reais para exemplificar a utilização dos modelos marginais, porém esses modelos não se comportaram da maneira esperada para os dados escolhidos. Os impasses foram devidamente explicados e foi proposta, ainda, outra abordagem.

Palavras-chave: Análise de Sobrevivência; modelos marginais; eventos recorrentes; riscos proporcionais; máxima verossimilhança parcial.

Abstract

Survival Analysis Applied to Multiple Events

This report aims to study the effect of vitamin A supplementation on diarrhea morbidity rate and to do so, AG and PWP models were adjusted to a real data set. The models' parameters were estimated by the partial maximum likelihood and the proportional hazards assumption was tested by the padronized Schoenfeld residuals. It is important to highlight that the models did not produce the expected results and it was not possible to adjust one of the models. These obstacles were properly explained, and another approach was suggested.

Keywords: Survival Analysis, marginal models; multiple events; proportional hazards; partial maximum likelihood.

Capítulo 1

Introdução

A análise de sobrevivência é um ramo da estatística há muito conhecido por sua vasta e notável aplicabilidade no campo da saúde, o que gerou, e ainda gera, frutos benéficos devido aos estudos clínicos desenvolvidos na área médica, como estimação de tempo de vida de pacientes e comparação de tratamentos, por exemplo.

Além dessa tradicional abordagem, não se deve deixar de lado a ascendente utilização dessas técnicas nos mais variados contextos como o previdenciário, analisando taxa de mudança de emprego, promoção e aposentadoria; no direito, na tentativa de entender o tempo de julgamento dos processos judiciais de maneira a reduzi-lo; nas ciências sociais, com indicadores socioeconômicos e educacionais, dentre outros.

Esse tipo de abordagem da estatística tem como foco principal estudar o tempo, denominado tempo de falha, até a ocorrência de um fenômeno, comumente especificado como o óbito. Em outras palavras, analisa-se o tempo decorrido entre um momento no qual o indivíduo se encontra em uma condição específica e o instante em que esse estado se modifica.

Como enfoque mais específico, pode-se considerar estudos de AIDS, onde o objetivo é verificar a suposição de que a infecção do vírus HIV aumenta o risco de contaminação ou desenvolvimento de uma doença específica. Ou seja, a falha nesse caso é o aparecimento da enfermidade e os tempos de falha são calculados para cada indivíduo, a contar do início do estudo.

Em geral alguns indivíduos não são observados até a ocorrência da falha, o que deixa o tempo de observação incompleto. Essa perda é denominada censura e ocorre quando o indivíduo falha por um motivo diferente do que está em estudo, é impossibilitado de permanecer na análise ou o evento de interesse não ocorreu até o fim da pesquisa.

Nos casos em que o evento de interesse não consiste na morte do indivíduo, não existe a necessidade de finalizar o acompanhamento daqueles que já experimentaram a

falha. Pode ser interessante estudar novas ocorrências do evento, o que faz com que indivíduos que já falharam, mas que podem falhar novamente, continuem sob observação. Essa particularidade caracteriza os eventos recorrentes, que são o foco de análise desta monografia.

Dito isso, define-se o objetivo principal do presente trabalho como a utilização de procedimentos e modelos de análise de sobrevivência para analisar dados que consideram tempo de falha, censura e eventos recorrentes em sua estrutura. Para tanto foi selecionado um banco de dados reais, proveniente de um estudo sobre o efeito de suplementação de vitamina A para o tratamento de diarreia infantil, e toda a modelagem e análise dos dados foi feita no *software* estatístico R.

Capítulo 2

Revisão de Literatura

2.1 Conceitos básicos

A Análise de Sobrevivência é um ramo da estatística que consiste em analisar dados cuja variável de interesse representa o tempo decorrido entre um instante inicial até a ocorrência de um evento de interesse, definido tempo de falha. Com vasta aplicabilidade na área da saúde, essa técnica permite estudar tais tempos de sobrevivência e não raro encontra-se informações incompletas dentre os dados, o que é denominado censura. Nesse âmbito, tal acontecimento denota a característica fundamental dessa vertente. A seguir serão definidos com mais detalhes essas e outras particularidades da Análise de Sobrevivência.

2.1.1 Tempo de Falha

A variável resposta, tempo de falha, consiste no tempo até a ocorrência de um evento de interesse previamente especificado e é composta por três elementos. Um desses componentes se refere à escala de medida, que usualmente é o tempo real mas que também pode ser número de ciclos, quilometragem de um carro ou quilogramas.

Neste contexto é imprescindível que o instante inicial de acompanhamento seja muito bem definido para que os indivíduos possam ser comparados na origem do estudo. Na prática, esse momento é demarcado por um acontecimento comum aos participantes do estudo, como é o caso de diagnóstico de uma doença, data do começo de um tratamento ou de início de uma greve.

Como último elemento, destaca-se o evento de interesse, ou falha, que em geral consiste na morte de um paciente, todavia com o desenvolvimento de estudos em Análise de Sobrevivência é cada vez mais comum deparar-se com tais episódios não ligados à área

da saúde, tampouco limitado apenas a essa fatalidade. Dentre esses destaca-se o caso das greves trabalhistas - considerando o tempo inicial ou final da mesma -, início ou fim do período de desemprego ou instante de desistência em um teste de esforço físico.

Em geral os indivíduos são retirados do estudo após a ocorrência da falha, seja por interesse do pesquisador - primeira manifestação de certa doença ou primeira mudança de emprego - ou por uma impossibilidade natural - morte do paciente ou produto impróprio para o uso quando atingido o prazo de validade.

Vale ressaltar a importância da clareza quanto à definição da falha, uma vez que a mesma pode ser facilmente confundida com uma censura, implicando em uma coleta de dados incorreta e consequentemente prejuízos na análise dos mesmos.

2.1.2 Eventos Recorrentes

Ainda que seja comum retirar os indivíduos do estudo após a ocorrência da falha, por vezes é de interesse do pesquisador que os mesmos permaneçam sob análise (e também sob risco de falha). Isso se deve ao fato de alguns eventos de interesse não serem terminais, podendo, então, ocorrer mais de uma vez (Tomazzela, 2003). Esse contexto caracteriza os eventos múltiplos e o interesse e a necessidade de aplicar a análise de sobrevivência nesse tipo de dados foi destacada por Therneau e Grambsch (2000).

Estudos envolvendo eventos múltiplos por indivíduo podem ser definidos em eventos do mesmo tipo, como infecções recorrentes em pacientes com AIDS; ou de tipos diferentes, a exemplo de múltiplas sequelas distintas em portadores de uma determinada doença. Outros casos que podem configurar eventos recorrentes são ocorrência de diversos tumores para um indivíduo, reincidência criminal e uma peça de equipamento falhar em repetidas experiências (Tomazzela, 2003).

Os casos que consistem em múltiplos eventos do mesmo tipo caracterizam os denominados eventos recorrentes, que serão o foco deste trabalho e uma melhor explanação será feita mais adiante, na seção Metodologia. Seguindo essa linha, serão definidas a seguir outras propriedades e particularidades fundamentais para a modelagem desse tipo de evento.

2.1.3 Censura

A censura consiste na perda de informação dos indivíduos em análise e impossibilidade dos mesmos experimentarem o evento de falha. Pode-se exemplificar tais acontecimentos como a morte antes da manifestação da doença em estudo ou por uma razão diferente da estudada.

Quando o tempo de censura registrado é maior que o tempo de falha diz-se que a mesma é uma censura à esquerda e indica que o evento de interesse ocorreu antes do indivíduo ser observado. Analogamente, a censura à direita é definida em situações em que o tempo de falha se encontra à direita do tempo observado. Esse último é o caso mais comum em estudos de análise de sobrevivência e pode ser classificado em alguns tipos.

De acordo com sua forma de ocorrência, a censura à direita pode ser definida como aleatória, do tipo I ou do tipo II. Cada um se traduz, respectivamente, em situações em que o paciente é retirado da análise antes da ocorrência de falha por algum motivo alheio à pesquisa; o mesmo não experimenta o evento até o fim do estudo, que é determinado de antemão pelo pesquisador; o estudo se encerra após atingir um número previamente delimitado de falhas.

Existe ainda um terceiro tipo de censura, definida como intervalar. Ela se caracteriza como um tipo mais geral onde não se detém informação do tempo exato de falha, apenas de um intervalo no qual a falha ocorreu (Colosimo e Giolo, 2006). Essa forma de censura ocorre comumente em estudos onde os indivíduos não são acompanhados continuamente, mas em visitas periódicas onde se registra a ocorrência ou não da falha no intervalo acompanhado.

No tratamento dos dados é importante que as censuras estejam presentes, uma vez que mesmo incompletas, ainda fornecem informações importantes sobre o estudo em questão e sua omissão pode implicar conclusões enviesadas e pouco confiáveis e nesta monografia, será considerada censura à direita aleatória. Devido a isso, se faz necessária a introdução de uma variável extra na análise, algo que indique se o valor do tempo de falha de um indivíduo específico foi observado ou não e tal componente é conhecida como variável indicadora de falha ou censura.

Dessa forma, a variável resposta em análise de sobrevivência é representada pelo par (t, δ), sendo t o tempo registrado e δ a variável indicadora de falha ou censura, como é representado abaixo, e desta forma a variável resposta é representada por duas colunas no banco de dados.

$$\delta = \begin{cases} 1, & \text{se } t \text{ é um tempo de falha,} \\ 0, & \text{se } t \text{ é um tempo censurado.} \end{cases}$$

2.2 Tempo de Sobrevivência

Um dos principais problemas associados à presença de censura é a dificuldade, ou por vezes até a impossibilidade de se aplicar os métodos tradicionais da estatística para o tratamento dos dados. Sendo assim, em estudos de sobrevivência destaca-se a importância

de se definir a variável aleatória não negativa T, usualmente contínua, como o tempo de falha.

A distribuição dessa variável pode ser caracterizada pela função de sobrevivência, pela função de densidade de probabilidade e pela função de risco, que serão todas detalhadas a seguir.

2.2.1 Função de Sobrevivência e Função de Densidade

A função densidade de probabilidade é definida como a probabilidade de um indivíduo falhar em um intervalo de tempo específico, representado por Δt na função a seguir:

$$f(t) = \lim_{\triangle t \to 0} \frac{P(t \leqslant T < t + \Delta t)}{\Delta t}. \tag{2.1}$$

Cabe ressaltar que a função $f(t)$ é sempre positiva para todo t e $\int_0^\infty f(t) = 1$.

A função de distribuição acumulada da variável T é definida por:

$$F(t) = \int_0^t f(u)du, \tag{2.2}$$

e estabelece relações úteis com a função de sobrevivência.

A função de sobrevivência consiste na probabilidade de ocorrência do evento de interesse após o instante t, ou seja, do indivíduo sobreviver (não falhar) até o tempo t e é definida da seguinte forma:

$$S(t) = P(T > t), t \geq 0. \tag{2.3}$$

Outra forma de encontrar a função de sobrevivência é utilizando a função de densidade de probabilidade acumulada que representa, portanto, a probabilidade de ocorrência do acontecimento de interesse até o instante t.

$$F(t) = P(T \leqslant t), t \geq 0,$$

$$F(t) = 1 - P(T > t),$$

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

A função $F(t)$ carrega informações importantes que permitem tirar conclusões sobre o conjunto de dados, como o percentual de indivíduos que não falharam até um tempo de interesse, a distribuição dos mesmos ao logo do tempo e até mesmo comparar o tempo de

vida de observações de dois ou mais grupos.

2.2.2 Função de Risco

A função de risco também é conhecida como função de taxa de falha e representa o risco de uma falha ocorrer em um intervalo específico, sabendo que não ocorreu antes do instante inicial e sua forma geral é representada a seguir:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.4}$$

Em situações que Δt é muito pequeno, $\lambda(t)$ representa a taxa instantânea de falha de um indivíduo no tempo t, dado que o mesmo não falhou até esse momento. Vale ressaltar que essas taxas de falha são números positivos, porém sem limite superior.

Essa função se faz muito útil quando se deseja representar a distribuição do tempo de vida de indivíduos pois descreve a evolução da probabilidade instantânea de falha ao longo do tempo. Outra característica importante é de sua informação ser, por vezes, de natureza qualitativa, o que pode colaborar na seleção de um modelo adequado para representar os dados de acordo com a forma que a função de risco pode assumir.

Situações nas quais os indivíduos são observados num período no qual ocorre um envelhecimento gradual configuram função de risco monótona crescente, uma vez que a taxa de falha dos mesmos aumenta com o decorrer do tempo. Como consequência, a proporção de indivíduos que falham em um determinado instante cresce no decorrer do tempo e é também o tipo mais comum em Análise de Sobrevivência.

Menos comuns, mas ainda assim possíveis, são os casos de função de risco monótona decrescente: quanto mais tempo sem experimentar a falha, menor é a probabilidade de o indivíduo falhar (indivíduos operados para correção de um defeito que representava o principal risco de morte, por exemplo); constante: quando a taxa de falha não muda com o passar do tempo; unimodal: quando a taxa de falha começa crescente e depois decresce; e em forma de U ou banheira: ocorre em populações onde indivíduos são acompanhados desde o nascimento até à morte, caso inverso da unimodal.

Outra função usada no contexto de análise de sobrevivência é a de taxa de falha acumulada, que é expressa na forma:

$$\Lambda(t) = \int_0^t \lambda(u) du. \tag{2.5}$$

Essa função se mostra bastante útil na avaliação da função de taxa de falha $\lambda(t)$, essencialmente na estimação não paramétrica em que $\Lambda(t)$ apresenta um estimador com

ótimas propriedades e a função de risco é difícil de ser estimada.

2.2.3 Relações entre as Funções

As funções supracitadas se relacionam matematicamente, o que pode ser útil em processos de estimação além de auxiliar, por exemplo, no conhecimento das demais funções quando não se tem definida alguma delas. Tais relações são estabelecidas a seguir.

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

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t)$$

e

$$S(t) = \exp\{-\Lambda(t)\} = \exp\left\{-\int_0^t \lambda(u)du\right\}.$$

2.3 Técnicas de Análise de Dados de Sobrevivência

Um dos principais problemas associados à presença de censura é a dificuldade, ou por vezes até a impossibilidade de se aplicar os métodos tradicionais da estatística para o tratamento dos dados. A habitual análise exploratória serve de base para diagnósticos futuros e visa resumir informações acerca das observações em forma de medidas de tendência central, dispersão e de posição, por exemplo.

Uma abordagem clássica que ainda pode se fazer útil em estudos de sobrevivência é o exame de gráficos de dispersão de cada covariável quantitativa com a resposta, o que explicita possível relação entre as variáveis ainda que com a presença de informações parciais ou incompletas. Em contrapartida, devido ao fato de dados de sobrevivência apresentarem censuras, as técnicas estatísticas usuais não são adequadas e podem implicar em uma análise enviesada das observações.

Dito isso, deseja-se estabelecer uma estimativa das funções de sobrevivência e de risco e para tanto adota-se um estimador não paramétrico, ou seja, que independe da distribuição de probabilidade dos dados, denominado Estimador de Kaplan-Meier.

2.3.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier é o mais utilizado para estimar a função de sobrevivência, que muitas vezes é difícil de ser calculada devido à presença de censura nos estudos. Essa formulação foi proposta por Kaplan e Meier (1958) e consiste em uma adaptação da forma empírica de $S(t)$, formulada a partir do princípio de ausência de censura como segue:

$$S(t) = \frac{\text{n}^\text{o} \text{ de observações que não falharam até o tempo } t}{\text{n}^\text{o} \text{ total de observações no estudo}}. \tag{2.6}$$

Analogamente, é possível desenvolver uma equação numérica seguindo a ideia estabelecida nessa equação.

Supõe-se, então, que uma amostra compreenda n pacientes no estudo e que dentre essas medidas haja somente $k(\leqslant n)$ falhas distintas - d_j - nos tempos $t_1 < t_2 < \ldots < t_k$. Considerando n_j o número de indivíduos sob risco em t_j, aqueles que não falharam e nem foram censurados, Colosimo e Giolo (2006) definem o estimador de Kaplan-Meier como:

$$\widehat{S}(t) = \prod_{j\,:\,t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j\,:\,t_j<t} \left(1 - \frac{d_j}{n_j}\right).$$

Dentre suas propriedades, Colosimo e Giolo (2006) destacam a qualidade de ser não viciado, fracamente consistente, convergir assintoticamente para uma distribuição normal e ser estimador de máxima verossimilhança de $S(t)$.

2.4 Modelos de Regressão

Além do tempo de sobrevivência e da variável indicadora de censura, ainda é possível que sejam observadas outras variáveis no banco de dados indicando as mais diversas características da população, tais como sexo, idade, tipo de tratamento ao qual o paciente foi submetido, dentre outros. Essas componentes são conhecidas como variáveis explicativas ou covariáveis.

Por vezes, o objetivo de estudos de sobrevivência se concentra em entender a relação que se estabelece entre o tempo de falha e algumas dessas variáveis, bem como quantificá-las. Para tanto, a forma mais eficiente de acomodar o efeito das mesmas é utilizar um modelo de regressão apropriado para dados censurados.

Em análise de sobrevivência existem duas classes de modelos: os paramétricos e os semiparamétricos. A primeira categoria é considerada mais eficiente, porém menos flexível, ao passo que a segunda, também conhecida como modelo de Cox, além de sua

versatilidade traz, ainda, a vantagem de incorporar com facilidade covariáveis dependentes do tempo, ou seja, cujos valores ao final do experimento podem não ser os mesmos observados no início. Este trabalho abordará os modelos de regressão semiparamétricos e suas propriedades serão definidas a seguir.

2.4.1 O Modelo de Regressão de Cox

O modelo de Cox (Cox, 1972) é usado extensivamente em estudos médicos e uma razão para sua popularidade é o fato de dispensar a necessidade de se formular um modelo probabilístico que se adeque aos dados, como é o caso dos modelos paramétricos. Além disso, sua formulação conta com a presença de um componente não-paramétrico que o torna bastante flexível.

O elemento não-paramétrico é representado pelo componente $\lambda_0(t)$, que é uma função positiva não especificada. Ela é também conhecida como função de risco de base, pois é comum a todos os indivíduos.

De forma geral, em um estudo composto por p covariáveis, pode-se definir $\mathbf{x} = (x_1, \ldots, x_p)'$ como o vetor de covariáveis que tem dimensão $p \times 1$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, com dimensão $1 \times p$, o vetor de parâmetros associados às covariáveis, a expressão do modelo é dada por:

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) g(\mathbf{x}'\boldsymbol{\beta}). \tag{2.7}$$

Ainda que outras formas para a função $g(\mathbf{x}'\boldsymbol{\beta})$ tenham sido propostas, a relação $g(\mathbf{x}'\boldsymbol{\beta}) = \exp\{\mathbf{x}'\boldsymbol{\beta}\}$ atua de forma multiplicativa e garante que $\lambda(t \mid \mathbf{x})$ seja sempre não-negativa. Essa será a associação adotada nesta monografia e essa propriedade multiplicativa da função exponencial se evidencia no formato:

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp\{\mathbf{x}'\boldsymbol{\beta}\} = \exp\{\beta_1 x_1 + \ldots + \beta_p x_p\}. \tag{2.8}$$

Uma particularidade do modelo de Cox é a ausência da constante β_0, presente nos modelos de regressão paramétricos e indicadora do impacto na variável resposta, quando as demais variáveis assumem valor nulo. Essa característica se deve ao fato do modelo em questão já possuir um componente não-paramétrico, o que absorve esse termo constante (Colosimo e Giolo, 2006).

O modelo proposto por Cox é também conhecido como modelo de taxas proporcionais, que significa que a razão das taxas de falha de dois indivíduos i e j quaisquer é sempre constante no tempo e é dada por:

$$\frac{\lambda(t \mid \mathbf{x}_i)}{\lambda(t \mid \mathbf{x}_j)} = \frac{\lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\lambda_0(t)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \exp\{\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_j'\boldsymbol{\beta}\}.$$

Sendo assim, a suposição básica para o uso do modelo de Cox é que as taxas de falha sejam proporcionais (Colosimo e Giolo, 2006). Essa propriedade se aplica também aos modelos que derivam do modelo de Cox e sua aplicação será evidenciada mais adiante na seção Resultados.

2.4.2 Ajuste do Modelo de Cox

Os coeficientes $\boldsymbol{\beta}$'s que medem os efeitos que as covariáveis exercem sobre a função de risco devem ser estimados a partir das observações amostradas para que o modelo fique determinado. É necessário, então, que algum método de estimação seja utilizado e que esses valores obtidos sejam devidamente interpretados.

O método mais utilizado para essa estimação, o de máxima verossimilhança, não é apropriado devido à presença do componente não paramétrico. Uma solução para esse empecilho leva em consideração as falhas e censuras passadas, foi proposta formalmente por Cox em 1975 e é denominado método de máxima verossimilhança parcial.

- **Método de Máxima Verossimilhança Parcial**

Colosimo e Giolo (2006) destacam que em uma amostra de n indivíduos onde existam $k \leqslant n$ falhas distintas nos tempos $t_1 < t_2 < \ldots < t_k$, sendo $\mathbf{R}_i = \mathbf{R}(t_i)$ o conjunto de indivíduos sob risco no tempo t_i, diz-se que a probabilidade da i-ésima observação falhar no tempo t_i, conhecendo as observações que estão sob risco em t_i é tal que:

$$\frac{P[\text{indivíduo falhar em } t_i \mid \text{sobreviveu a } t_i \text{ e história até } t_i]}{\text{P[uma falha} \mid \text{história até } t_i]} =$$

$$\frac{\lambda_i(t \mid \mathbf{x}_i)}{\sum_{j\in\mathbf{R}(t_i)}\lambda_i(t \mid \mathbf{x}_j)} = \frac{\lambda_0(t)\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\lambda_0(t)\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}}.$$

O produto de todos os termos dessa equação, associados aos tempos distintos de falha, constitui a função de verossimilhança desejada, como segue:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k}\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}} = \prod_{i=1}^{n}\left(\frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{\sum_{j\in R(t_i)}\exp\{\mathbf{x}_j'\boldsymbol{\beta}\}}\right)^{\delta_i}. \tag{2.9}$$

Aplicando o logaritmo na equação (2.9), obtém-se a seguinte equação:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i'\boldsymbol{\beta} - \log \left(\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\} \right) \right].$$

Desta maneira, os valores de $\boldsymbol{\beta}$ que maximizam a função de verossimilhança parcial, $L(\boldsymbol{\beta})$, são obtidos resolvendo-se o sistema de equações definido por $U(\boldsymbol{\beta}) = 0$, em que $U(\boldsymbol{\beta})$ é o vetor escore de derivadas de primeira ordem da função $l(\boldsymbol{\beta})$. Ou seja:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\mathbf{x}_i - \frac{\sum_{j \in R(t_i)} \mathbf{x}_i \exp\{\mathbf{x}_j'\hat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\hat{\boldsymbol{\beta}}\}} \right] = 0.$$

Essa função assume que os tempos de sobrevivência são contínuos, o que não contempla a possibilidade de empates dos valores observados, sendo que na prática isso pode ocorrer devido à escala de medida (Colosimo e Giolo, 2006). Na ocorrência de tempos iguais para falha e censura de um mesmo indivíduo, por convenção é considerado que a censura ocorreu após a falha, o que leva a incluir tais observações no conjunto de risco em cada tempo de falha.

É preciso, então, que seja feita uma modificação na função de verossimilhança parcial, $L(\boldsymbol{\beta})$, para incorporar tais observações empatadas quando as mesmas estão presentes, o que foi proposto por Breslow (1972) e Peto (1972). Desta maneira, considera-se s_i o vetor formado pela soma das correspondentes p covariáveis para os indivíduos que falham no mesmo tempo t_i $(i = 1, \ldots, k)$ e d_i o número de falhas neste mesmo tempo, logo a aproximação considera a seguinte função de verossimilhança parcial:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\{\mathbf{s}_i'\boldsymbol{\beta}\}}{\left[\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j'\boldsymbol{\beta}\}\right]^{d_i}}.$$

Vale ressaltar que essa aproximação é adequada quando o número de observações empatadas não é grande.

Mais adiante serão descritas a base de dados utilizada nessa monografia e as técnicas de análise multivariada das observações, ideal para eventos múltiplos. Essas técnicas são de certa forma uma extensão do modelo de Cox e os modelos que serão descritos na Metodologia se sustentam nas demonstrações apresentadas até então.

- **Interpretação dos coeficientes do modelo**

O efeito das covariáveis que compõem a expressão (2.7) pode acelerar ou desacelerar a função taxa de falha, porém a propriedade de taxas de falha proporcionais do modelo deve ser usada para interpretar os coeficientes estimados (Colosimo e Giolo, 2006). Toma-se, então, a razão das taxas de falha de dois indivíduos, i e j, que têm os mesmos valores

para as covariáveis, com exceção da l-ésima, obtendo:

$$\frac{\lambda(t \mid \mathbf{x}_i)}{\lambda(t \mid \mathbf{x}_j)} = \exp\{\beta_l(x_{il} - x_{jl})\}.$$

Essa razão é constante para todo o acompanhamento e supondo, por exemplo, que x_l seja uma variável dicotômica indicando o sexo feminino do indivíduo, diz-se que a taxa de falha entre mulheres é $\exp\{\beta_l\}$ vezes a de indivíduos do sexo masculino.

Analogamente, nos casos de variáveis contínuas, um efeito de $\exp\{\beta\} = 1,306$ significa que ao aumentar em uma unidade a variável x, a taxa de falha é aumentada em $30,6\%$.

2.4.3 Adequação do Modelo de Cox

Apesar de ser bastante flexível, o modelo de Cox não se adequa a qualquer situação, o que requer uma avaliação do ajuste do mesmo. Para tanto, existe uma série de técnicas gráficas e estatísticas que amparam essa apreciação.

A seguir serão descritos os métodos que serão utilizados nesta monografia como parte do processo de análise do banco de dados selecionado. Cabe ressaltar que existem outros procedimentos para essa avaliação, porém este trabalho fará uso apenas daqueles que serão descritos adiante.

- **Avaliação da Suposição de Riscos Proporcionais**

Como mencionado anteriormente, o modelo de Cox assume que as taxas de falha são proporcionais e a violação dessa suposição pode acarretar vícios de estimação dos coeficientes do modelo (Struthers e Kalbfleisch, 1986). Para tanto, existem alguns artifícios capazes de verificar se esse pressuposto é válido ou não para a situação em estudo.

Um gráfico descritivo pode ser elaborado através da divisão do banco de dados em estratos de uma variável de interesse, e após estimar $\hat{\Lambda}_{0j}(t)$ por meio da expressão (2.10), analisa-se as curvas do logaritmo de $\hat{\Lambda}_{0j}(t)$ *versus* t ou $\log(t)$. A suposição de riscos proporcionais é, então, considerada válida se as curvas projetadas de cada estrato apresentarem diferença aproximadamente constantes ao longo do tempo.

Caso essas curvas não estejam paralelas, há indícios de violação da suposição de taxas de falha proporcionais. Dessa maneira, deve-se considerar elaborar esse gráfico para cada covariável incluída no estudo, o que pode auxiliar a identificar aquela que descumpre o pressuposto de riscos proporcionais.

Mais uma vez essa decisão depende da interpretação dos gráficos, que nem sempre

levam à mesma conclusão. Em contrapartida, existem outros artifícios menos subjetivos que auxiliam essa avaliação como os resíduos padronizados de Schoenfeld (1982).

Com crescente aplicação, Colosimo e Giolo (2006) destacam que esses resíduos são definidos para cada falha do conjunto de dados, onde o i-ésimo indivíduo que possui vetor de covariáveis $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ e foi observado falhar, tem vetor de resíduos de Schoenfeld $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{ip})'$ em que cada componente r_{iq}, para $q = 1, \ldots, p$, é definido por:

$$r_{iq} = x_{iq} - \frac{\sum_{j \in R(t_i)} x_{jq} \exp\{\mathbf{x}_j' \hat{\boldsymbol{\beta}}\}}{\sum_{j \in R(t_i)} \exp\{\mathbf{x}_j' \hat{\boldsymbol{\beta}}\}}. \tag{2.10}$$

Esses resíduos são definidos para cada falha e não são definidos para censuras, como destacam Colosimo e Giolo (2006). Ainda, para cada uma das p covariáveis incluídas no modelo, existe um correspondente resíduo para o indivíduo i, e como os mesmos são definidos para cada falha do estudo, é gerada uma matriz com d linhas e p colunas, sendo d o número de falhas e p o número de variáveis no modelo.

A relação $\sum_i \mathbf{r}_i = 0$ também é valida para esses resíduos, e como descrito por Therneau e Grambsch (2000), uma forma padronizada dos resíduos de Schoenfeld permite que a estrutura de correlação dos mesmos seja considerada e é definida por:

$$\mathbf{s}_i^* = [\mathcal{I}(\hat{\boldsymbol{\beta}})]^{-1} \times \mathbf{r}_i, \tag{2.11}$$

com $\mathcal{I}(\hat{\boldsymbol{\beta}})$ representando a matriz de informação observada.

Esses resíduos são analisados graficamente pela projeção de $\mathbf{s}_{iq}^* + \hat{\beta}_q$ *versus* t, para a covariável $q = 1, \ldots, p$, como é sugerido por Therneau e Grambsch(1994). Gráficos com tendências aproximadamente horizontais indicam proporcionalidade dos riscos, ao passo que tendências mais acentuadas podem ser indicativos de violação dessa suposição.

Como reforçado anteriormente, a interpretação desses gráficos é subjetiva. Entretanto pode ser utilizada em conjunto com outras técnicas estatísticas e testes de hipóteses para conclusões mais exatas. O coeficiente de correlação de Pearson - ρ - entre $\mathbf{s}_{iq}^*$ e $g(t)$ para cada covariável, por exemplo, pode indicar que a suposição não foi violada, quando os valores de ρ estão próximos de zero (Colosimo e Giolo, 2006).

Colosimo e Giolo (2006) apresentam um teste para a hipótese global de proporcionalidade que assume $g_q(t) = g(t)$ e utiliza a estatística-teste:

$$T = \frac{(g - \bar{g})' S^* \mathcal{I} S^{*'} (g - \bar{g})}{d \sum_k (g_k - \bar{g})^2}, \tag{2.12}$$

em que $S^* = dR\mathcal{I}^{-1}$, sendo R a matriz $d \times p$ dos resíduos de Schoenfeld não-padronizados.

Sob a hipótese nula de taxas de falha proporcionais a estatística T tem aproximadamente distribuição qui-quadrado com p graus de liberdade e valores de $T > \chi^2_{p,1-\alpha}$ apontam evidências contra a hipótese nula.

Analogamente, a hipótese de proporcionalidade para a q-ésima covariável pode ser testada pela estatística do teste:

$$T_q = \frac{d\Big(\sum_k (g_k - \bar{g}) s^*_{qk}\Big)^2}{\mathcal{I}_q^{-1} \sum_k (g_k - \bar{g})^2},$$

em que as propriedades anteriores se mantêm e por consequência, valores de $T_q > \chi^2_{1,1-\alpha}$ apontam evidências contra a suposição de taxas de falha proporcionais para a variável q.

Capítulo 3

Metodologia

3.1 Material

No fim do século XX a relação entre suplementação de vitamina A e morbidade infantil era bastante estudada em diversas partes do mundo, o que gerou inúmeras pesquisas sobre o tema. No Brasil, essa questão foi abordada por Barreto e cols. (1994) que conduziu e elaborou um estudo no nordeste do país.

O conjunto de dados fruto do referido estudo foi cedido pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia, e consiste no material a ser analisado nesta monografia. Os dados foram coletados entre dezembro de 1990 e dezembro de 1991 e uma discussão detalhada desses dados está publicada em Andreozzi (2002).

O estudo consiste em um ensaio comunitário randomizado placebo-controlado - o que permite inferência confiável entre eventos da natureza em questão - e como proposto pela comunidade internacional, teve como principal interesse avaliar o impacto da suplementação com vitamina A na redução da morbidade por diarreia e infecção respiratória.

O arquivo utilizado na análise contém uma coorte de 860 crianças entre 6 e 48 meses, das quais 426 pertencem ao grupo placebo e 434 integram o grupo vitamina A, o que gera 5.592 registros e 11 variáveis. Outra característica desse banco de dados é a presença de múltiplos eventos por indivíduo, os diversos episódios de diarreia por criança, particularidade de interesse deste trabalho.

O arquivo contendo os dados estão disponíveis na página referente a Análise de Sobrevivência da Fiocruz, no endereço http://sobrevida.fiocruz.br/diarreia.html . A seguir a estrutura do mesmo, bem como as variáveis que o compõe são descritas na Tabela 3.1.

Tabela 3.1: Variáveis do banco de dados *diarreia.csv*

Variável	Descrição
numcri	identificador da criança
grupo	vit = receberam vitamina A, pla = placebo
sexo	fem = feminino, masc = masculino
idade	idade em meses no início do estudo
ini	data do início do período **sem** diarreia
fim	data do fim do período **sem** diarreia
diasant	duração em dias do episódio de diarreia anterior
mediadej	média de dejeções líquidas ou semilíquidas do episódio de diarreia anterior
enum	numeração das observações para cada criança
status	0 = censura, 1 = evento (diarreia)
tempo	dias sem diarreia (*fim* - *ini*)

3.2 Métodos

Devido à natureza dos dados escolhidos, surge a necessidade de utilizar técnicas de análise diferenciadas das descritas anteriormente. Na revisão de literatura foi introduzido o modelo de regressão de Cox (1972) e as demonstrações e fórmulas apresentadas anteriormente serão úteis para descrever os modelos que seguem nas próximas seções.

3.2.1 Análise de Sobrevivência Multivariada

Em Análise de Sobrevivência, o estudo de eventos cujo desfecho acontece apenas uma vez durante o período de acompanhamento é extremamente comum e existe uma infinidade de modelos estatísticos na literatura capazes de tratar problemas que surgem desses dados. Ainda que nesses acompanhamentos exista a possibilidade do evento de interesse ocorrer mais de uma vez, o indivíduo sai do grupo de risco após experimentar a falha (Carvalho et al, 2011).

Ainda que eventos únicos sejam bastante comuns, por vezes o interesse do estudo é o tempo até a ocorrência de um fenômeno repetidas vezes ou fenômenos de diferentes tipos. As técnicas habituais de Análise de Sobrevivência não são adequadas para estudar esses eventos, denominados múltiplos ou recorrentes, e por isso devem ser analisados por modelos diferenciados (Tomazzela, 2003).

É preciso levar em consideração fatores de risco que podem estar associados aos tempos de falha de um indivíduo e mesmo nesse contexto de eventos múltiplos, assim como

no de eventos únicos, existe uma série de abordagens possíveis. Portanto, é importante entender a natureza dos dados a fim de encontrar a melhor maneira de trabalhá-los.

É possível, por exemplo, tratar os eventos de cada indivíduo de maneira independente uns dos outros; escolher apenas um tipo de ocorrência (segunda ocorrência ou uma falha específica, por exemplo); ou analisar os múltiplos eventos conjuntamente, mesmo quando a ocorrência de um exclui a do outro, os quais são denominados eventos competitivos.

Existem situações em que é razoável supor uma associação entre os tempos de sobrevivência, como é o caso dos dados escolhidos para serem analisados nesta monografia, e nessas circunstâncias existem duas abordagens possíveis e bastante utilizadas. Uma dessas estratégias consiste nos modelos de fragilidade, que ajusta e condiciona cada indivíduo a um efeito aleatório - ou fragilidade - e a partir disso encara os eventos como independentes (Carvalho et al, 2011).

Essa abordagem é aplicável apenas a eventos do mesmo tipo e a fragilidade é usada com o intuito de pesar a correlação existente entre os múltiplos tempos observados (Colosimo e Giolo, 2006). A segunda abordagem consiste nos modelos marginais, extensões do modelo de Cox que são bastante flexíveis e capazes de ajustar dados nas mais diversas situações.

O presente trabalho se concentra em aplicar essa modelagem aos dados de diarreia e as técnicas utilizadas para tanto, bem como outras definições serão detalhadas a seguir.

3.2.2 Modelos Marginais

Como citado anteriormente, esses modelos consistem em extensões do modelo de Cox e com isso, muito do que foi descrito na Revisão de Literatura se mostrará útil nesta seção. Ainda que seja uma extensão de outro modelo, por se tratar de eventos relativamente diferenciados, é importante observar algumas questões para análise do tempo envolvendo eventos múltiplos.

Essa cautela é bastante útil para orientar a escolha do modelo e é fundamental que se defina de maneira clara a população em risco, o risco basal λ_0, a estrutura temporal ou ordenação, e a estrutura de dependência entre os eventos (Carvalho et al, 2011).

Ao se considerar que a população em risco muda a cada ocorrência de um novo evento, é razoável supor que λ_0 possa variar de um evento para outro. No modelo de Cox esse valor era o mesmo para todos os indivíduos e seu valor não era estimado, porém essa suposição não deve ser regra ao lidar com eventos recorrentes.

A formatação do banco de dados define quem está sob risco na ocorrência de cada

evento (Carvalho et al, 2011). Nos dados que serão utilizados nesta monografia, ao considerar os fatores associados à morbidade por diarreia, estarão sob risco de novo episódio de diarreia apenas crianças que apresentam quadro de dejeção normal. Crianças vivenciando episódio de dejeções líquidas não podem ser incluídas na população de risco.

Os eventos do banco de dados escolhido são caracterizados como **ordenados**, uma vez que a sucessão dos tempos segue uma ordem delimitada pela data de início e fim de cada evento (Carvalho et al, 2011). Três indivíduos, o 65º, o 66º e o 67º, apenas com as covariáveis *grupo* e *sexo* estão apresentados na Tabela 3.2 para exemplificar a entrada dos dados no pacote R.

Tabela 3.2: Entrada dos dados de diarreia no R.

numcri	ini	fim	status	tempo	grupo	sexo
65	1	17	1	16	pla	fem
65	19	25	1	6	pla	fem
65	27	379	0	352	pla	fem
66	1	217	1	216	pla	masc
66	219	267	0	48	pla	masc
67	1	54	1	53	vit	masc
67	56	267	0	211	vit	masc

Combinando-se então as definições de grupo sob risco e estrutura do risco basal, é possível identificar dois tipos de eventos ordenados: **independentes**, semelhante a afirmar que a criança sempre volta ao grupo de risco após finalizar cada episódio de diarreia, e que o momento de ocorrência de cada evento independe dos tempos anteriores; e **estruturados**, que assume que a criança só está sob risco de sofrer o n-ésimo evento depois que o evento de ordem $n-1$ tiver ocorrido.

Esses tipos de eventos consistem nas formulações propostas por Andersen e Gill e por Prentice, Williams e Peterson. Existe ainda uma terceira formulação marginal proposta por Wei, Lin e Weissfeld e todas essas serão devidamente abordadas a seguir.

- **Formulação de Andersen & Gill (AG)**

  Proposto em 1982, esse modelo considera que o risco de base é igual em todos os intervalos de tempos analisados, com a peculiaridade do indivíduo retornar ao grupo de risco após cada evento e assume que os eventos em cada intervalo disjunto são independentes. Devido a essa particularidade, a entrada dos dados é feita de maneira diferenciada: cada indivíduo é representado por uma série de linhas com os respectivos intervalos de tempos indicados por (tempo de entrada no estudo, tempo do primeiro evento], (segundo tempo de entrada, tempo do segundo evento], e assim

por diante (Colosimo e Giolo, 2006).

Indivíduos com nenhum evento será representado por apenas uma linha, e dependendo da escala de medida usada para o tempo, a primeira observação poderá ou não começar no zero. Se a mesma iniciar no tempo de entrada, o modelo para o i-ésimo indivíduo é representado por:

$$\lambda_i = \lambda_0(t) \exp\{\mathbf{x}_i'(t)\boldsymbol{\beta}\}, \tag{3.1}$$

em que os componentes são os mesmos utilizados anteriormente para formular o modelo de Cox, e $\mathbf{x}_i'(t)$ é o vetor de dimensão $1 \times p$ de covariáveis observadas para o i-ésimo indivíduo.

A principal diferença entre esse modelo e o de Cox está justamente no fato de um indivíduo permanecer em risco mesmo após experimentar o evento de interesse. O modelo AG assume que existe independência entre os múltiplos eventos de um mesmo indivíduo, de maneira que o risco desse indivíduo falhar pela primeira vez é o mesmo risco dele falhar na n-ésima vez. Dessa maneira, seu histórico não afeta o risco presente, ou seja, o risco dele falhar em um instante t qualquer independe do número de falhas experimentadas pelo indivíduo até esse instante t.

Com a finalidade de testar o pressuposto de independência, é recomendado verificar se a variância robusta das estimativas dos parâmetros deste modelo são um pouco maiores que as do modelo de Cox usual (Carvalho et al, 2010). O caso em que esses valores diferem muito pode ser um indicativo de que o modelo AG não é adequado.

- **Formulação de Wei, Lin e Weissfeld (WLW)**

  Essa formulação, proposta em 1989, trata as respostas de um conjunto de dados ordenados como se fosse um problema de riscos competitivos com respostas não ordenadas. Isto significa que o indivíduo no início do período de observação é considerado estar sob risco de sofrer m eventos e o tempo é sempre contado a partir do zero. Desta maneira, a função de taxa de falha para o m-ésimo evento do i-ésimo indivíduo é expressa da seguinte forma:

$$\lambda_{im}(t) = \lambda_{0m}(t) \exp\{\mathbf{x}_i'(t)\boldsymbol{\beta}_m\}. \tag{3.2}$$

  Como esse modelo pressupõe que todos os indivíduos iniciam o estudo em risco de sofrer m eventos, é necessário criar observações fictícias para os que ainda não experimentaram este número. Ou seja, todos os indivíduos do banco de dados deverão ser representados por m linhas (sendo m o máximo de eventos experimentado por um indivíduo qualquer), independentemente do número de eventos que cada um tenha experimentado.

Uma característica interessante desse modelo é o fato de possibilitar a utilização de uma função de taxa de falha separada para cada evento, assim como para cada estrato, o que é mostrado pela notação $\boldsymbol{\beta}_m$. Além disso, essa formulação tem foco no tempo de sobrevivência total de um estudo até a ocorrência do m-ésimo evento de interesse, o que é recomentado em casos que os eventos ocorreram em diferentes ordens, bem como diferentes tipos de evento.

Ao entrar no estudo, o indivíduo está concomitantemente em risco de sofrer o primeiro, segundo, terceiro, m-ésimo evento (Carvalho et al, 2011). Esse raciocínio não se aplica aos dados de diarreia e portanto essa modelagem foi descartada do processo de análise dos dados, mas foi incluída nesta seção por fazer parte dos modelos marginais e ser útil em outros estudos.

- **Formulação de Prentice, Williams e Peterson (PWP)**

  Em sua formulação proposta em 1981, esse modelo parte do princípio de que um indivíduo não pode estar sob risco para o m-ésimo evento sem que tenha experimentado o evento $m-1$. Dessa maneira, identifica-se uma similaridade com os dados de diarreia infantil, uma vez que não existe a possibilidade de uma criança experimentar o segundo evento de diarreia sem antes ter experimentado o primeiro. Esse modelo é também denominado condicional, justamente pela característica de existir essa relação de dependência entre os tempos de falha de um mesmo indivíduo.

  Devido a essa suposição, o modelo exige que a análise seja feita de forma separada nos diferentes estratos dependentes do tempo, o que implica na possibilidade de variação da função de risco de um evento para outro - o modelo AG não permite que isso ocorra. A função de taxa de falha do modelo PWP é representada, então, por:

$$\lambda_{im}(t) = \lambda_{0m}(t)\exp\{\mathbf{x}_i'(t)\boldsymbol{\beta}_m\}. \tag{3.3}$$

  As funções de risco dos modelo PWP e WLW são formalmente idênticas, contudo no primeiro o indivíduo necessariamente deve ter $m-1$ eventos para experimentar o m-ésimo, enquanto que no modelo WLW ele está em um conjunto de m eventos no tempo t.

  Cabe ressaltar que a entrada dos dados para os modelos AG e PWP é feita da mesma maneira que foi apresentada na Tabela 3.2. No modelo WLW essa introdução é feita de maneira diferenciada, de maneira que cada indivíduo seja representado por tantas linhas quanto os eventos que está sujeito a sofrer. Devido a essa diferença na entrada dos dados, ainda que os modelos PWP e WLW tenham a mesma fórmula, a análise desses dados produz resultados diferentes.

3.2.3 Estimação dos Parâmetros

A estimação dos parâmetros dos modelos citados anteriormente, assim como no modelo de Cox, é feita através do método de máxima verossimilhança parcial, ignorando a correlação existente entre as observações. Mota (2013) apresenta essa estimação de maneira análoga à descrita anteriormente na revisão de literatura.

Supondo-se n indivíduos, que cada um possa experimentar m falhas e que a censura não interfere na probabilidade de sobrevivência em um tempo futuro. A função de verossimilhança parcial para o modelo AG pode ser representada como:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{m=1}^{k_i} \left(\frac{\exp\{\mathbf{x}'_{mi}\boldsymbol{\beta}\}}{\sum_{j=1}^{n} \sum_{l=1}^{k} \exp\{\mathbf{x}'_{lj}\boldsymbol{\beta}\}} \right)^{\delta_i}.$$

Já para os modelos PWP e WLW, a função é dada por:

$$L(\boldsymbol{\beta}_m) = \prod_{i=1}^{n} \prod_{m=1}^{k_i} \left(\frac{\exp\{\mathbf{x}'_{mi}\boldsymbol{\beta}_m\}}{\sum_{j=1}^{n} \exp\{\mathbf{x}'_{mj}\boldsymbol{\beta}_m\}} \right)^{\delta_{mi}}.$$

Com a finalidade de alcançar uma estimativa robusta, é feita a correção na variância dos $\hat{\boldsymbol{\beta}}$ através de uma aproximação da estimativa jackknife, que consiste na remoção de uma ou mais amostras do conjunto total observado, recalculando-se o estimador a partir dos valores restantes (Mota, 2013). Esse procedimento gera uma estimativa não viciada da variância para dados correlacionados sempre que a observação deixada de fora for independente das observações que entram. Em outras palavras, esse processo sendo agrupado por indivíduo deixa de fora um sujeito em um tempo, em vez de uma observação no tempo.

Os resíduos de jackknife são utilizados para obter uma estimativa de jackknife agrupado por indivíduo e os mesmos são definidos como:

$$\mathbf{J}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)},$$

em que $\hat{\boldsymbol{\beta}}_{(i)}$ é o resultado do ajuste que inclui todas as observações exceto o indivíduo i.

Uma forma de calcular os valores dos resíduos de jackknife é feita utilizando o método de Newton-Raphson, foi proposta por Therneau & Grambsch (2000) e é reescrita da seguinte maneira:

$$\Delta\beta = 1'(\mathbf{U}\mathcal{I}^{-1}) = 1'\mathbf{D},$$

sendo $\mathbf{U}$ a matriz de escore residual, então a mudança em $\hat{\boldsymbol{\beta}}$ em cada iteração é soma

da coluna da matriz $\mathbf{D}$, definida como escore residual dimensionada pela matriz $\mathcal{I}^{-1}$, que corresponde à variância dos $\hat{\boldsymbol{\beta}}$.

Agrupada por indivíduo, essa estimativa pode ser escrita da seguinte forma:

$$V_j = \frac{n-1}{n}(\mathbf{J} - \overline{\mathbf{J}})'(\mathbf{J} - \overline{\mathbf{J}}),$$

em que $\overline{\mathbf{J}}$ é a matriz de médias das colunas de $\mathbf{J}$. A variância passa a ser escrita como $\mathbf{D}'\mathbf{D} = \mathcal{I}^{-1}(\mathbf{U}'\mathbf{U})\mathcal{I}^{-1}$, que pode ser vista como um estimador sanduíche ABA, em que $A = \mathcal{I}^{-1}$ é a estimativa usual da variância e $\mathbf{U}'\mathbf{U}$ é o termo de correção.

Capítulo 4

Resultados e Discussões

4.1 Análise Descritiva

A fim de descrever e resumir o conjunto de dados, foi feita uma análise descritiva dos mesmos através das estimativas de Kaplan-Meier tanto para entender o contexto geral, quanto para identificar o comportamento de algumas variáveis. Esse processo é importante pois os resultados obtidos podem influenciar escolhas no processo de modelagem.

A Figura 4.1 apresenta a estimativa para a função de sobrevivência e a análise do gráfico permite formular algumas suposições. Até o centésimo dia, por exemplo, a probabilidade de uma criança apresentar episódio de diarreia diminui consideravelmente à medida que o tempo passa. Nos dias subsequentes esse ritmo é menor, visto que a curva é mais suave após essa marca.

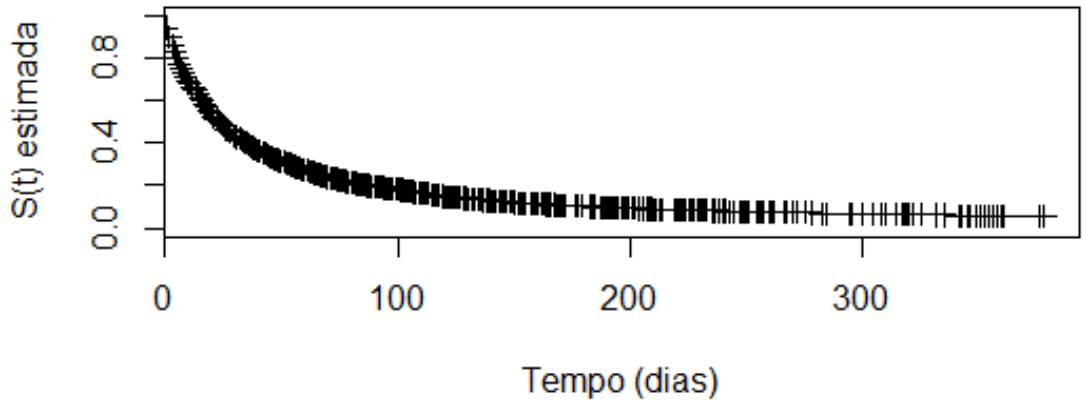

Figura 4.1: Estimativa de Kaplan-Meier para os tempos até a ocorrência de diarreia.

Muitos fatores podem estar associados a esse ritmo, então é preciso saber o peso de cada variável no conjunto dos dados. Observando os gráficos das estimativas de Kaplan-Meier das covariáveis sexo e grupo de suplementação apresentados na Figura 4.2, observa-se que as curvas não apresentam diferenças muito aparentes, indicando que possivelmente o tempo de sobrevivência não está sendo influenciado pelas variáveis em questão.

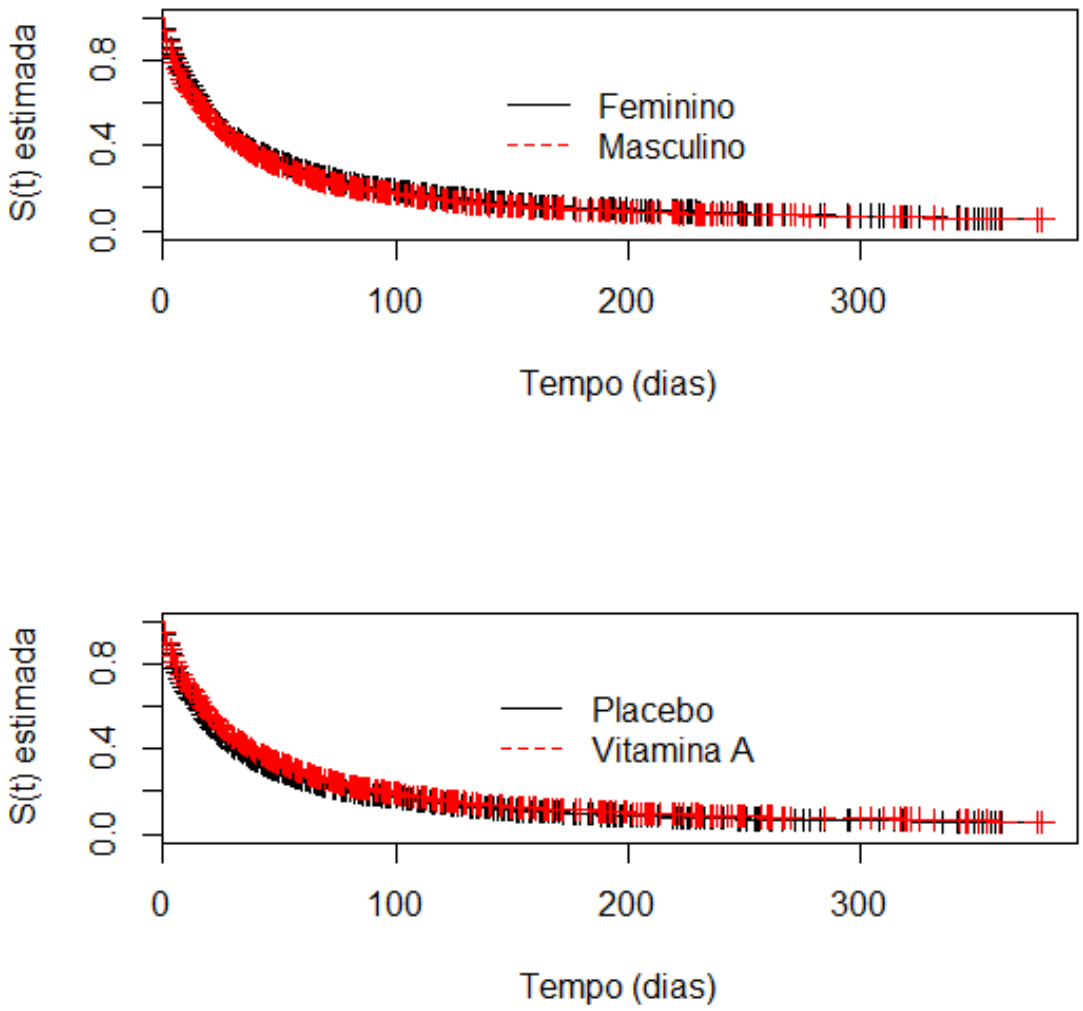

Figura 4.2: Estimativas de Kaplan-Meier para o sexo da criança e o grupo de suplementação.

Essa característica é de certa forma intrigante, uma vez que o interesse principal no estudo desse banco de dados é entender o impacto da variável suplementação na diarreia infantil. Ainda que as curvas sejam muito semelhantes, é possível notar um leve desvio

naquela referente ao placebo, na parte inferior esquerda do gráfico. Como enfatizado na Revisão de Literatura, análises gráficas são subjetivas e é preciso avaliar se esse desvio (pouco evidente) é significativo ou não.

Por ora, não é possível obter conclusões muito confiáveis acerca dessa influência, portanto é necessário que seja feita uma modelagem dos dados para entender melhor a relação existente entre cada covariável e o tempo até a reincidência de diarreia.

Da mesma maneira, foram analisadas as variáveis quantitativas: idade, média de dejeções líquidas e duração em dias do episódio de diarreia anterior. Como essas variáveis são quantitativas, uma alternativa para estimas as suas curvas de sobrevivência é categorizá-las. Dessa maneira, optou-se por construir os gráficos com categorias que consistem nos intervalos de classe delimitados pelos quartis de cada variável.

A Figura 4.3 apresenta diferença evidente entre as faixas etárias dos indivíduos em estudo e aparentemente as curvas não se cruzam. Essas características podem ser indicativos de que a idade da criança exerce alguma influência no tempo até a reincidência de diarreia e que os riscos dessa variável são proporcionais, indícios que serão testados na próxima seção.

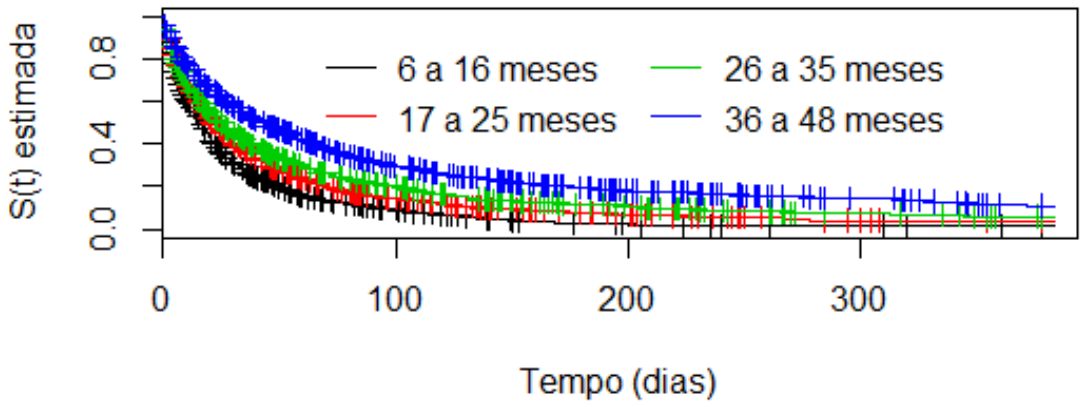

Figura 4.3: Estimativas de Kaplan-Meier para a idade das crianças no início do estudo.

Os gráficos contidos na Figura 4.4 apresentam diferença visível entre a primeira curva e as demais, porém dentre essas últimas não se observa diferença muito aparente. Isso pode significar que só existe diferença significativa entre dois grupos principais dentro de cada variável: mais ou menos que duas dejeções líquidas em média; e mais ou menos que um dia com quadro de diarreia no episódio anterior.

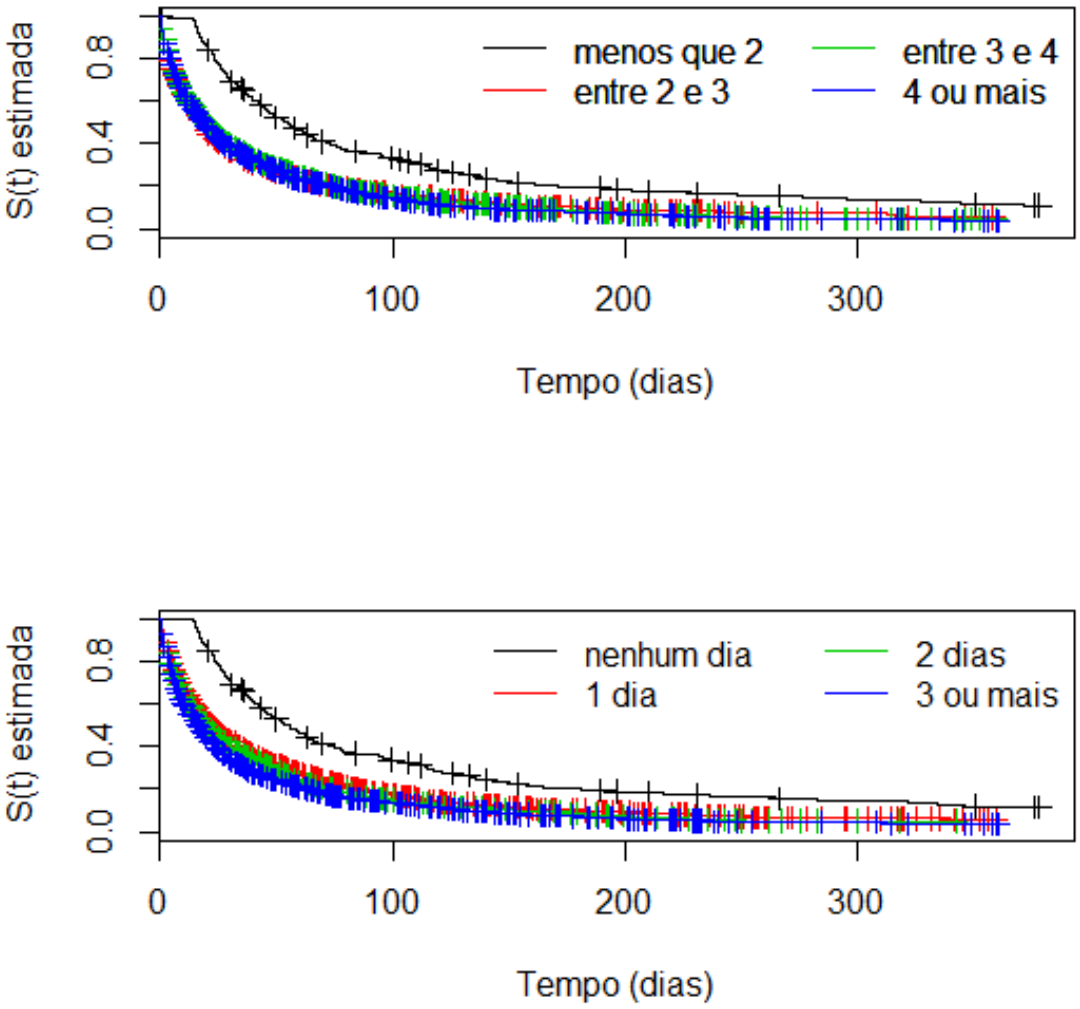

Figura 4.4: Estimativas de Kaplan-Meier para a média de dejeções líquidas no episódio de diarreia anterior e a quantidade de dias com quadro de diarreia no episódio anterior.

Esses gráficos servem principalmente para obter uma ideia de como os dados estão distribuídos e, de maneira mais rústica, entender a forma que cada variável explicativa está se relacionando com a variável resposta. O sentido dessa relação, bem como sua intensidade e significância serão quantificados no processo de modelagem desses dados a seguir.

4.2 Modelagem

Após ter realizado a análise descritiva dos dados e ter um entendimento básico da maneira que cada covariável está se relacionando com o tempo até a reincidência de

diarreia, inicia-se o processo de modelagem desses dados. Primeiramente os dados serão tratados pelo modelo de Cox simples, e em seguida será utilizada a abordagem marginal.

O ajuste com o modelo de Cox serviu de amparo para identificar variáveis significativas e analisar a adequabilidade dos modelos. Os valores estimados nessa etapa não serão apresentados, com exceção do erro padrão das variáveis que compõem o modelo final, que serão utilizados para avaliar a qualidade do mesmo.

- **Modelo AG**

Foram incluídas no modelo, respectivamente, as variáveis *grupo* (x_1), *sexo* (x_2), *idade* (x_3), *diasant* (x_4) e *mediadej* (x_5). A significância e a proporcionalidade dos riscos de cada uma delas dentro do modelo foram avaliadas e aquelas que não se ajustaram foram devidamente retiradas.

Para realização do teste de significância de cada uma das variáveis, foram utilizadas as hipóteses:

$$\begin{cases} H_0\text{: A covariável não é significativa para o modelo } (\beta = 0); \\ H_1\text{: A covariável é significativa para o modelo } (\beta \neq 0). \end{cases}$$

A partir da Tabela 4.1 e a um nível de significância de 5%, o resultado $p - valor = 0,285$ indica que a variável *sexo* é a menos significativa para o modelo. No que tange as demais covariáveis, o teste mostra evidências de que todas têm influência sobre o tempo de sobrevivência, ainda que a variável *grupo* esteja relativamente próxima do nível de 5%.

Tabela 4.1: Estimativas do modelo AG para todas as variáveis.

Parâmetros	Estimativas	Erro	Erro Robusto	Z	p-valor
$grupo : \beta_1$	$-0,1190$	$0,0292$	$0,0588$	$-2,024$	$0,043$
$sexo : \beta_2$	$0,0631$	$0,0292$	$0,0591$	$1,069$	$0,285$
$idade : \beta_3$	$-0,0279$	$0,0013$	$0,0024$	$-11,466$	$< 0,0001$
$diasant : \beta_4$	$0,0501$	$0,0037$	$0,0050$	$9,973$	$< 0,0001$
$mediadej : \beta_5$	$0,1269$	$0,0076$	$0,0152$	$8,372$	$< 0,0001$

No contexto da proporcionalidade das taxas de falha, as hipóteses avaliadas foram:

$$\begin{cases} H_0\text{: Os riscos são proporcionais;} \\ H_1\text{: Os riscos não são proporcionais.} \end{cases}$$

e a Tabela 4.2 indica que apenas a variável *diasant*, com $p - valor = 0,865$, não viola a suposição de riscos proporcionais, a um nível de significância de 5%. As demais covariáveis quebram esse pressuposto e influenciam para que o modelo como um todo também viole.

Tabela 4.2: Testes da proporcionalidade das taxas de falha no modelo AG.

Covariáveis	ρ	χ^2	p-valor
vitamina A	$0,0441$	$37,7104$	$< 0,0001$
masculino	$0,0299$	$17,5396$	$< 0,0001$
idade	$-0,0429$	$34,3305$	$< 0,0001$
diasant	$-0,0019$	$0,0289$	$0,865$
mediadej	$-0,0338$	$14,6331$	$0,0001$
Global	$-$	$100,3036$	$< 0,0001$

O fato das demais variáveis violarem o pressuposto de riscos proporcionais indica que não devem ser incluídas no modelo, porém é possível que a presença ou ausência de algumas delas interfira nesse resultado. A variável *grupo*, por exemplo, é de extrema importância para o estudo, e caso ela não faça parte do modelo final, ainda que se mostre um bom modelo, o mesmo não estará considerando o efeito de maior interesse.

Com o intuito de evitar esse impasse, nas etapas de seleção e exclusão das covariáveis, foi sempre priorizada a inclusão da variável *grupo*. Nos casos em que nenhuma outra covariável além dessa poderia ser excluída e ela se mostrou não significativa para o modelo, a decisão de retirá-la foi tomada.

O teste de proporcionalidade para o modelo global também foi rejeitado, reforçando a suspeita de que as variáveis que violam esse pressuposto influenciam no risco geral do modelo. É possível que apenas uma delas esteja causando esse prejuízo, portanto foi feita uma análise minuciosa do impacto de cada variável.

Após algumas etapas de análise, chegou-se ao modelo final composto apenas pelas covariáveis $diasant(x_4)$ e $mediadej(x_5)$. Com as estimativas apresentadas na Tabela 4.3, identifica-se uma relação positiva entre a variável *diasant* e o tempo até uma nova ocorrência de diarreia. Com $\hat{\beta}_1 = 0,0623$, a cada aumento de um dia na duração da diarreia, estima-se que a taxa de ocorrência de um novo episódio aumente em $6,43\%$ $(\exp\{0,0623\} = 1,0643)$.

Foi observada também uma relação positiva entre a média de dejeções líquidas do episódio de diarreia, e a reincidência desse acontecimento. A cada aumento de uma dejeção líquida, estima-se um aumento de $15,88\%$ $(\exp\{0,1474\} = 1,1588)$ no risco de nova ocorrência de diarreia.

Em outras palavras, o modelo traz a informação de que quanto maior a duração e a quantidade de dejeções líquidas da criança em um episódio de diarreia, maior é o risco dessa criança sofrer o evento novamente.

Tabela 4.3: Estimativas do modelo AG para as variáveis *diasant* e *mediadej*.

Parâmetros	Estimativas	Erro	Erro Robusto	Z	p-valor
$diasant : \beta_4$	$0,0623$	$0,0035$	$0,0051$	$12,046$	$< 0,0001$
$mediadej : \beta_5$	$0,1474$	$0,0069$	$0,0182$	$8,081$	$< 0,0001$

Os testes para a proporcionalidade dos riscos associados às variáveis *diasant* e *mediadej* indica haver rejeição da suposição feita na hipótese nula. Optou-se por incluir a Tabela 4.4, ainda que a mesma indique a violação do pressuposto de riscos proporcionais e os valores contidos nela servirão de comparação com os do modelo PWP mais adiante.

Tabela 4.4: Testes da proporcionalidade das taxas de falha no modelo AG com as covariáveis *diasant* e *mediadej*.

Covariáveis	ρ	χ^2	p-valor
diasant	$-0,0164$	$2,44$	$0,119$
mediadej	$-0,0356$	$22,67$	$< 0,0001$
Global	$-$	$29,61$	$< 0,0001$

Mesmo com os devidos cuidados durante a seleção das variáveis, não foi possível chegar a um modelo com valores aceitáveis dentro da estrutura da modelagem AG. Com os valores encontrados para o teste de proporcionalidade das variáveis restantes, o próximo passo consistiria em excluir *mediadej* e ao realizar novamente o teste, *diasant* estaria associado a $p - valor = 0$.

Isso é um indicativo de que esse modelo não é apropriado para os dados em estudo e o motivo surge provavelmente do fato de que não existe independência entre os eventos de diarreia de cada criança. O modelo AG, por outro lado, parte justamente do princípio de independência entre os eventos.

Outro indício de que o modelo AG não é adequado para os dados escolhidos reside justamente no fato do valor estimado do erro padrão robusto (4ª coluna da Tabela 4.3) de cada variável ser significativamente diferente do erro padrão estimado no modelo de Cox (3ª coluna da Tabela 4.3). Como é esperado que essas estimativas se mantenham razoavelmente similares, desvios podem significar falta de adequação do modelo utilizado.

Dessa maneira, como os valores da Tabela 4.4 mostraram, conclui-se que não é possível utilizar essa modelagem nos dados de diarreia infantil. O próximo passo consiste, então, em utilizar o modelo PWP e avaliar o ajuste e adequabilidade do mesmo.

- **Modelo PWP**

Assumindo as mesmas hipóteses e nível de significância para a realização dos tes-

tes dentro do modelo PWP, além da mesma ordem de inclusão das variáveis no modelo, observa-se na Tabela 4.5 que as variáveis *grupo* e *sexo* não influenciam o tempo de sobrevivência. Como citado na modelagem anterior, durante o processo de seleção das variáveis foi sempre priorizada a inclusão da variável *grupo* por ser considerada a mais importante dentro dos dados.

Tabela 4.5: Estimativas do modelo PWP para todas as variáveis.

Parâmetros	Estimativas	Erro	Z	p-valor
$grupo : \beta_1$	$-0,0370$	$0,0351$	$-1,055$	$0,2914$
$sexo : \beta_2$	$0,0454$	$0,0352$	$1,287$	$0,1980$
$idade : \beta_3$	$-0,0154$	$0,0015$	$-10,244$	$< 0,0001$
$diasant : \beta_4$	$0,0278$	$0,0048$	$5,807$	$< 0,0001$
$mediadej : \beta_5$	$0,0329$	$0,0137$	$2,400$	$0,0164$

O teste de proporcionalidade apresentado na Tabela 4.6 indica haver uma violação do pressuposto associado às variáveis *grupo* e *idade*, o que visivelmente gerou impacto no teste para o modelo como um todo. Não há evidências para rejeitar as demais variáveis, porém é possível que o resultado para o mesmo teste com um modelo com menos variáveis seja diferente.

Tabela 4.6: Testes da proporcionalidade das taxas de falha no modelo PWP.

Covariáveis	ρ	χ^2	p-valor
grupo	$0,0515$	$17,684$	$< 0,0001$
sexo	$0,0183$	$2,267$	$0,132$
idade	$0,0434$	$12,145$	$0,0005$
diasant	$-0,0077$	$0,371$	$0,543$
mediadej	$-0,0042$	$0,102$	$0,749$
Global	$-$	$31,335$	$< 0,0001$

Foi realizada novamente uma seleção das variáveis candidatas a fazer parte do modelo final pelo método *backward* e os valores obtidos estão expostos na Tabela 4.7. Mesmo que tenha sido feito um esforço no sentido de evitar excluir a variável *grupo*, não foi possível garantir que a mesma estivesse presente no modelo final, que terminou composto apenas por $diasant(x_4)$ e $mediadej(x_5)$.

Ainda assim é possível extrair informações importantes dos valores calculados, como a relação positiva entre as covariáveis e a variável resposta. Com $\hat{\beta}_1 = 0,0322$, a cada dia adicional com diarreia, espera-se que a ocorrência de um novo episódio aumente $3,27\%$ ($\exp\{0,0322\} = 1,0327$). Sobre a média de dejeções líquidas do episódio anterior, estima-

se que cada aumento de uma dejeção líquida gere um aumento de $3,6\%$ ($\exp\{0,0354\} = 1,036$) no risco de nova ocorrência de diarreia.

Essa relação pode não ser muito forte, porém o modelo aponta que é significativa, e carrega a mesma informação desse teste no modelo AG: quanto maior a duração e a quantidade de dejeções líquidas da criança em um episódio de diarreia, maior é o risco dessa criança sofrer o evento novamente.

Tabela 4.7: Estimativas do modelo PWP para as variáveis *diasant* e *mediadej*.

Parâmetros	Estimativas	Erro Padrão	Z	p-valor
$diasant : \beta_4$	$0,0322$	$0,0048$	$6,693$	$< 0,0001$
$mediadej : \beta_5$	$0,0354$	$0,0139$	$2,537$	$0,0112$

Diferente do que foi encontrado no modelo AG, o teste da proporcionalidade dos riscos para os dados utilizando o modelo PWP indicou que os riscos são, de fato, proporcionais. A Tabela 4.8 indica que não há evidências para rejeitar a hipótese nula para nenhuma das variáveis e nem para o modelo global.

Tabela 4.8: Testes da proporcionalidade das taxas de falha no modelo PWP com as variáveis *diasant* e *mediadej*.

Covariáveis	ρ	χ^2	p-valor
diasant	$-0,0139$	$1,232$	$0,267$
mediadej	$-0,0095$	$0,549$	$0,459$
Global	$-$	$1,739$	$0,419$

- **Comparação entre os dois modelos**

Como os modelos AG e PWP chegaram às mesmas variáveis após o processo de seleção, é possível comparar qual dos dois se ajusta melhor aos dados. Ainda que tenha sido explicado anteriormente o motivo do modelo AG não ser apropriado, foram utilizados para fazer essa comparação os critérios AIC e BIC, cujas definições se encontram na seção Anexo.

Esses critérios consistem em medidas de avaliação do modelo que causa menor perda de informações ao tentar explicar o fenômeno em estudo. Nesse sentido, optou-se por utilizar os critérios de informação de Akaike e Bayesiano, vastamente utilizados em estudos estatísticos e apresentados na Tabela 4.9.

Por definição, quanto menor o valor desses critérios, melhor é o modelo utilizado. Como esperado e devidamente justificado anteriormente, o modelo PWP se mostra o

mais indicado para os dados de diarreia. Entretanto, é importante destacar algumas características que podem ter contribuído para isso.

Em sua estrutura, o modelo PWP leva em consideração que o risco de ocorrer o primeiro evento é diferente dos subsequentes, o que é uma suposição razoável em casos de reinternações hospitalares, infartos e, como demonstrado, diarreia infantil. ou seja, o segundo evento está condicionado à ocorrência do primeiro evento e assim sucessivamente (Carvalho et al, 2011).

Tabela 4.9: Valores de AIC e BIC para os modelos AG e PWP.

	Modelos	Marginais
Critério	AG	PWP
AIC	$61924,37$	$39025,96$
BIC	$61937,63$	$39039,22$

Com os valores calculados nas Tabelas 4.7 e 4.8, optou-se por definir o modelo PWP com as covariáveis *diasant* e *mediadej* como modelo final. Essa configuração não leva em consideração a covariável que, no momento da escolha do banco de dados, acreditava-se ser a mais relacionada à variável resposta, porém sua exclusão se mostrou necessária.

Capítulo 5

Considerações Finais

Nesta monografia foi feito um estudo da aplicação dos modelos marginais a dados de diarreia infantil e os resultados encontrados, apesar de não estarem de acordo com o esperado no começo do trabalho, fornecem informações importantes.

O imprevisto mais penoso consistiu no modelo ajustado aos dados não levar em consideração a principal covariável do estudo, que caso fosse significativa para o modelo, explicaria justamente a relação estabelecida entre a suplementação com vitamina A e o tempo até a ocorrência de diarreia. Mesmo com esse impasse, foi possível alcançar um modelo mais simples e válido para os dados.

Não se pode dizer que o modelo PWP apresentado na Tabela 4.7 é o melhor dentre todos os possíveis, até porque esta monografia tinha como objetivo entender melhor os modelos marginais e se possível ajustá-los a um banco de dados reais, sem antes saber se o ajuste seria bom. Ainda assim o modelo final pode ser útil no entendimento de outros fatores associados a esse tempo e foi possível entender, por exemplo, que existe uma relação positiva entre as variáveis dias de duração e média de dejeções líquidas ou semilíquidas do episódio de diarreia anterior e o tempo até a ocorrência de um novo episódio de diarreia.

É provável também que um modelo de fragilidade seja mais adequado aos dados de diarreia e, principalmente, capaz de incluir a variável *grupo*. Não foi possível testar a qualidade e adequabilidade do mesmo para os dados em questão nesta monografia, porém essa modelagem pode ser uma aplicação interessante das técnicas de sobrevivência para trabalhos futuros.

Como comentado, os resultados obtidos não eram esperados, o que faz parte do processo de modelagem de dados na área da estatística. Ainda que adversidades dessa natureza sejam desagradáveis, é fundamental extrair o máximo de informação do processo, de maneira que seja útil em experiências futuras.

Referências Bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Assis, A. M. O. and Barreto, M. L. (2002). Suplementação com vitamina A: impacto na morbidade e efeitos adversos. *Revista Brasileira de Epidemiologia*, 5:84 – 92.

Borges, A. I. M. (2014). Análise de sobrevivência com o r. Master's thesis, Universidade da Madeira.

Breslow, N. E. (1972). Discussion of professor cox?s paper. *J. Royal Stat. Soc. B*, (34):216–217.

Bustamante-Teixeira, M. T., Faerstein, E., and do Rosário Latorre, M. (2002). Técnicas de análise de sobrevida. *Cad. Saúde Pública*, 18(3):579–594.

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S., and Shimamura, S. E. (2005). *Análise de Sobrevivência: teoria e aplicações em saúde*. Fiocruz, Rio de Janeiro.

Castro, M. S. M. and Carvalho, M. S. (2005). Agrupamento da classificação internacional de doenças para análise de reinternações hospitalares. *Cad. Saúde Pública*, 21(1):317–323.

Cobre, J. (2010). *Modelos de Sobrevivência na Presença de Eventos Recorrentes e Longa Duração*. PhD thesis, Universidade Federal de São Carlos. Departamento de Estatística.

Colosimo, E. A. and Giolo, S. R. (2006). *Análise de Sobrevivência Aplicada*. Edgard Blücher, São Paulo. ABE - Projeto Fisher.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276. Department of Mathematics, Imperial College, London.

Giolo, S. R. Modelos de riscos proporcionais. Universidade Federal do Paraná. Departamento de Estatística.

Grambsh, P. M. and Therneau, T. M. (1972). Proportional hazards tests and diagnostics based onweighted residuals. *Biometrika*, 81(3):515–526.

Hosmer, D. W. and Lemeshow, S. (1999). Applied survival analysis. New York.

Konishi, S. and Kitagawa, G. (1978). Information criteria and statistical modeling. New York: Springer.

Mota, T. S. (2013). Modelagem em análise de sobrevivência com eventos recorrentes aplicada a dados da Área médica. Master's thesis, Universidade Estadual Paulista.

Peto, R. (1972). Rank tests of maximal power against lehmann-type alternatives. *Biometrika*, (59):472–475.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society*, (Series A 135):185–207.

Porfírio, L. V. (2015). Um modelo de regressão log-weibull com fração de cura para dados de pacientes com aids. Universidade de Brasília. Instituto de Ciências Exatas. Departamento de Estatística. Trabalho de conclusão de graduação.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379.

Rocha, C. S. (1995). Modelos de sobrevivência. Universidade de Lisboa. Faculdade de Ciências.

Schwarz, G. (1978). Estimating the dimensional of a model. *Annals of Statistics, Hayward*, 6(2):461–464.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73(2):363–369.

Terry M. Therneau and Patricia M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

Tomazella, V. L. D. (2003). *Modelagem de Dados de Eventos Recorrentes via Processo de Poisson com Termo de Fragilidade*. PhD thesis, Universidade de São Paulo. Instituto de Ciências Matemáticas e Computação.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073.

Anexo

A.1 Critérios de Informação

Para comparar n modelos $g_1(x \mid \boldsymbol{\theta_1})$, $g_2(x \mid \boldsymbol{\theta_2})$, ..., $g_n(x \mid \boldsymbol{\theta_n})$, deve-se comparar as magnitudes da função $L(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}})$, porém esse método não fornece verdadeira comparação, uma vez que não se conhece o verdadeiro modelo $g(x)$.

Dessa maneira, os critérios de informação são construídos para avaliar e corrigir o viés introduzido no momento da estimação. Segundo Konishi e Kitagawa (2008), um critério de informação tem a forma:

$$CI(\mathbf{X}_n, \hat{G}) = -2\sum_{i=1}^{n} \log f(X_i \mid \hat{\boldsymbol{\theta}}(X_n)) + 2(b(G)) \tag{A.1}$$

Critério de Informação de Akaike

Akaike (1974) mostrou que o viés é dado assintoticamente por p, em que p é o número de parâmetros a serem estimados no modelo, e definiu seu critério de informação como:

$$\text{AIC} = -2\log L(\hat{\boldsymbol{\theta}}) + 2p \tag{A.2}$$

Critério de Informação Bayesiano

Proposto por Schwarz (1978), considera que $f(x \mid \boldsymbol{\theta})$ é o modelo escolhido, p o número de parâmetros a serem estimados, n o número de observações da amostra e é definido por:

$$\text{BIC} = -2\log f(x \mid \boldsymbol{\theta}) + p\log n \tag{A.3}$$