UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

ALISSON DE LIMA BRITO

Modelos de Fragilidade Compartilhada: Uma Abordagem Paramétrica e Semi-paramétrica

Campina Grande - PB

Agosto de 2018

ALISSON DE LIMA BRITO

Modelos de Fragilidade Compartilhada: Uma Abordagem Paramétrica e Semi-paramétrica

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Tiago Almeida de Oliveira

Campina Grande - PB

Agosto de 2018

B862m Brito, Alisson de Lima.
Modelos de fragilidade compartilhada [manuscrito]: uma abordagem paramétrica e semi-paramétrica / Alisson de Lima Brito. - 2018.
56 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2018.
"Orientação : Prof. Dr. Tiago Almeida de Oliveira , Corrdenação do Curso de Estatística - CCT."

1. Retinopatia diabética. 2. Análise de sobrevivência.
3. Efeito aleatório. I. Título

21. ed. CDD 519.53

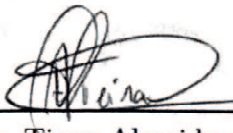
ALISSON DE LIMA BRITO

Modelos de Fragilidade Compartilhada: Uma Abordagem Paramétrica e Semi-paramétrica

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

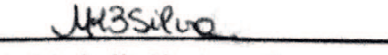
Trabalho aprovado em 29 de Agosto de 2018.

BANCA EXAMINADORA



Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba

Prof. Me. Ednário Barbosa de Mendonça
Universidade Estadual da Paraíba



Profª. Drª. Michelli Karinne Barros da Silva
Universidade Federal de Campina Grande

A Deus primeiramente, por me proporcionar muitas oportunidades tais como esta que vivo.
Aos meus pais, por me ajudarem a obter essa vitória, principalmente a minha querida mãe, que nos momentos difíceis também sofreu junto comigo.
Aos meus avós maternos, que nos momentos de dificuldade me ajudaram e acompanharam minha jornada até aqui.
A minha noiva e companheira, que sempre esteve presente em minha trajetória, me ajudando, apoiando e incentivando para conclusão deste curso.

Com amor, dedico.

Agradecimentos

A professora Ana Patrícia Bastos Peixoto por sempre acreditar em minha capacidade, sempre me ajudar no que foi possível, sempre me estender a mão quando precisei e me apoiar.

Ao professor Tiago Almeida de Oliveira que também sempre me ajudou no que esteve ao seu alcance, contribuiu para o desenvolvimento deste trabalho, se dedicando comigo aos estudos voltados para o mesmo.

Ao professor Ricardo Alves de Olinda por sempre me incentivar em relação aos estudos, acreditar no meu potencial e me ajudar diversas vezes quando me encontrava necessitado.

A todos os professores e funcionários do departamento de estatística pelo incetivo aos alunos do curso, pelas lições de vida, pelo empenho e comprometimento com o trabalho, pela total disponibilidade em resolver os problemas relacionados ao curso. Além do meu agradecimento de coração deixo também meus parabéns pelo belo trabalho.

A todos os integrantes da empresa Open Data Analysis Júnior, com os quais tive o prazer de passar muitos momentos, poder aprender também como pessoa.

Aos colegas e amigos que pude fazer durante minha carreira acadêmica e passar muitos momentos bons, Leandro Valter, Arthur Oliveira, Rodolfo Crystian, obrigado pelo apoio, pela ajuda em diversas situações e pelo companheirismo.

A todos os citados acima, posso lhes dizer que foi um grande prazer participar dessa grande família que encontrei na Universidade Estadual da Paraíba.

Enfim, a todos deixo meu sincero obrigado!

"Uma resposta aproximada para o problema certo vale muito mais do que uma resposta exata para um problema aproximado."
(John Wilder Tukey)

"A estatística é a gramática da ciência."
(Karl Pearson)

"A única coisa que se coloca entre um homem e o que ele quer na vida é normalmente meramente a vontade de tentar e a fé para acreditar que aquilo é possível."
(Richard M. Devos)

Resumo

Em análise de sobrevivência diversos modelos podem ser utilizados para verificar a influência de covariáveis na variável dada pelo tempo até a ocorrência do evento de interesse. A grande maioria desses modelos, no entanto, não levam em consideração a heterogeneidade presente entre os indivíduos no estudo. Sendo assim, modelos de sobrevivência que consideram essa variável não observável foram propostos na literatura para incorporar o fato de indivíduos possuírem diferentes fragilidades de sofrer o evento. O objetivo deste trabalho foi fazer uma análise comparativa entre os ajustes dos modelos clássicos de sobrevivência sem a presença da fragilidade e os ajustes com a inclusão do efeito aleatório no modelo (modelos de fragilidade), afim de verificar a contribuição da fragilidade nas estimativas dos parâmetros, fazendo uma abordagem paramétrica e semi-paramétrica através do modelo de Cox. Para tanto, foi utilizado um banco de dados de pacientes com Retinopatia diabética que faziam um tratamento à *laser* na Irlanda do Norte. Utilizou-se também os testes Log-rank e Peto para comparação entre as curvas de sobrevivência, descobrindo que o tratamento efetuado pelos pacientes para retardamento da cegueira realmente surtiu efeito, e que a Diabetes do tipo 2 se mostrou mais agressiva do que a Diabetes do tipo 1. Após os ajustes dos modelos foi identificado que todos os modelos de fragilidades tiveram um melhor ajuste, quando comparados aos demais modelos sem a presença da fragilidade, demostrando a grande contribuição dessa variável no modelo. Dentre os modelos testados, o modelo de fragilidade lognormal com distribuição do risco lognormal foi o que melhor se adequou aos dados, sendo portanto o melhor modelo a ser utilizado para as inferências.

Palavras-chaves: Retinopatia diabética, Análise de sobrevivência, Efeito aleatório.

Abstract

In Survival Analysis he different models may be used to see the influence the covariates in the variate time until the occurrence of the event. However, these models do not take into account the present heterogeneity between individuals in the study. Thus, survival models that considerate this unobserved variable were literature proposed to incorporate the frailties to fell in fail. The objective this monograph was made a comparative analysis between the adjusts the survival classical models and the adjust with frailty term in the model (frailty model), for to verify the contribution the component frailty in the parameter estimates, make the parametric and semiparametric approach. For this purpose was utilized datasets with patients diabetic retinopathy that making laser treatment in North Ireland. Was used too Log-rank and Peto tests for comparison between survival curves was the knowledge that treatment for diabetic retinopathy patients has an effect and that type 2 diabetic was more aggressive than type 1 diabetic. All frailty models have been satisfactory adjust when comparisons with classic models, demonstrate the efficiency the frailty variable. The lognormal frailty with risk distribution lognormal was better adjust the data, is, therefore, the best model for inferences in this case.

Key-words: Diabetic retinopathy, Survival analysis, Random effect.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

DM1	Diabetes Mellitus Tipo 1
DM2	Diabetes Mellitus Tipo 2
OMS	Organização Mundial da Saúde
SBD	Sociedade Brasileira de Diabetes
MPSF	Modelo Paramétrico sem a Fragilidade
MPFCL	Modelo Paramétrico de Fragilidade Compartilhada Lognormal
MSSF	Modelo semi-paramétrico sem a fragilidade
MFCSG	Modelo de fragilidade compartilhada semi-paramétrico gama

Lista de símbolos

γ	Letra Grega minúscula gama
Γ	Letra Grega maiúscula Gama
λ	Letra Grega minúscula lambda
Λ	Letra Grega maiúscula Lambda
δ	Letra Grega minúscula delta
Δ	Letra Grega maiúscula Delta
σ	Letra Grega minúscula sigma
π	Letra Grega minúscula pi
θ	Letra Grega minúscula teta
ψ	Letra Grega minúscula psi
ξ	Letra Grega minúscula xi
τ	Letra Grega tau
ϕ	Letra Grega minúscula fi
Φ	Letra Grega maiúscula fi
μ	Letra Grega mu
α	Letra Grega alpha
ν	Letra Grega nu
β	Letra Grega beta
η	Letra Grega eta

Sumário

1 Introdução

Diabetes Mellitus é uma doença caracterizada pela elevação da glicose no sangue. Pode ocorrer devido a defeitos na secreção ou na ação do hormônio insulina, que é produzido no pâncreas. A função principal da insulina é promover a entrada de glicose para as células do organismo de forma que ela possa ser aproveitada para as diversas atividades celulares. A falta da insulina ou um defeito na sua ação resulta portanto em acúmulo de glicose no sangue, o que chamamos de hiperglicemia (SBEM, 2017).

Segundo a OMS[1] (Organização Mundial da Saúde), 422 milhões de adultos ao redor do mundo têm diabetes e, cerca de 1,5 milhões das mortes ocorridas a cada ano estão ligadas a essa doença. Como esse número vem crescendo, a instituição já classifica a doença como uma epidemia. A cada ano, sete milhões de indivíduos entram nessa lista. No Brasil, a SBD (Sociedade Brasileira de Diabetes) estima que 12 milhões de pessoas tenham a doença, sendo que metade delas não sabe disso. O Diabetes Mellitus está divido em dois grupos: Diabetes tipo 1 (DM1) e Diabetes tipo 2 (DM2).

De acordo com a Sociedade Brasileira de Endocrinologia e Metabologia SBEM (2017), a DM1 é resultado da formação de anticorpos pelo próprio organismo contra as células beta pancreáticas, levando a deficiência de insulina. Já na DM2, é observado um quadro de resistência insulínica, onde a produção de insulina pelas células beta está sendo dificultada, o que leva a um aumento da produção de insulina para tentar manter a glicose em níveis normais. A DM2 ocorre em cerca de 90% dos pacientes notificados com a Diabetes.

Retinopatia Diabética é uma complicação que ocorre quando o excesso de glicose no sangue danifica os vasos sanguíneos dentro da retina, geralmente está ligada à maneira inadequada de se tratar a diabetes. Caso o paciente não busque tratamento, a visão pode ficar seriamente comprometida. Segundo Silva (2012), a Retinopatia Diabética é uma das complicações crônicas da diabetes, que 20 anos após o início da patologia está presente em 99% dos pacientes com Diabetes tipo 1 e em 60% dos pacientes com Diabetes tipo 2, constituindo a principal causa de cegueira em adultos.

A Retinopatia Diabética pode surgir sem que o paciente note diferença em sua visão. Com o passar do tempo, porém, a visão passa a piorar, podendo até mesmo chegar à cegueira, caso não seja tratada. A doença apresenta 4 fases, em que as três primeiras corresponde à fase mais moderada da patologia. Na quarta fase, forma mais agressiva da doença, se faz necessário um tratamento à *laser*. Nesse tratamento, os vasos sanguíneos

[1] https://avozdaserra.com.br/noticias/estima-se-que-12-milhoes-de-brasileiros-tenham-diabetes-mas-metade-nao-sabe-disso

neoformados e as áreas sem oxigenação são fotocoagulados. Normalmente, são necessárias duas ou mais sessões de aplicação a *laser*. Caso haja hemorragia severa, o paciente deve ser submetido à um procedimento cirúrgico chamado vitrectomia, para remover o sangue do olho.

Os estudos voltados a área da saúde ou das ciências médicas, buscam muitas vezes, como melhorar a qualidade de vida de pacientes, como descobrir qual é o melhor tratamento a ser executado para um determinado tipo de problema. Neste contexto, a análise de sobrevivência se faz presente e de muita importância para responder essas perguntas. Em estudos de sobrevivência se está interessado em observar o tempo até a ocorrência de um determinado evento ou desfecho. Com uso desta técnica, pode-se calcular as probabilidades pertinentes a ocorrência do evento, uma vez que este ainda não tenha ocorrido.

Neste estudo foi observado o tempo até a cegueira total de pacientes com Retinopatia Diabética que estavam sendo submetidos (ou não) a um tratamento à *laser*. Os dados obtidos para este trabalho foram provenientes de um estudo realizado na Irlanda do Norte em 1976 por Blair et al. (1980). Onde foi observado o tempo até a cegueira total de pacientes tratados ou não por determinados tipos de *lasers*, fazendo uso de técnicas de análise de sobrevivência, contudo, neste estudo não foi estimado um modelo para explicar a possível influência de covariáveis na cegueira destes pacientes.

Diante disto, objetivou-se neste trabalho fazer uma comparação entre os modelos de fragilidade com a modelagem paramétrica da função de risco e o modelo de riscos proporcionais de Cox com a introdução dos efeitos aleatórios (fragilidade), buscando o melhor ajuste que pudesse explicar de forma efetiva a influência de covariáveis na ocorrência do evento em estudo (cegueira) com aplicação aos dados de pacientes com Retinopatia Diabética. A presença da fragilidade modelada neste trabalho, se dá através dos grupos em estudo (tratado e controle), onde foram ajustados modelos de fragilidade compartilhada.

2 Fundamentação Teórica

O estudo de Análise de Sobrevivência é de grande valia para estudar o tempo até a ocorrência de um determinado evento e as probabilidades associadas a este evento. O estudo desta área da estatística é essencial quando se tem informações parciais ou incompletas sobre os indivíduos em estudo.

Neste capítulo é apresentada uma fundamentação teórica para dar embasamento a análise de um conjunto de dados de paciente com Retinopatia Diabética. Mais especificamente, é apresentada uma revisão sobre o estimado produto de Kaplan-Meier, testes de comparação entre curvas de sobrevivência, tais como Log-rank e Peto, modelos probabilísticos para modelagem do tempo de sobrevivência, o modelo de riscos proporcionais de Cox e modelos de fragilidade compartilhada.

2.1 Marco Histórico

O levantamento e observação de dados para se obter uma resposta de um determinado evento vem sido empregado a milhares de anos. Atualmente dar-se a essas ações o nome de análise estatística. O campo das análises estatísticas vem se desenvolvendo e aprimorando cada dia mais com a presença de novas técnicas para a análise de dados e o avanço tecnológico. Em particular, a Análise de Sobrevivência é uma das áreas mais antigas no campo da estatística. Os primeiros estudos de sobrevivência foram desenvolvidos nas ciências atuariais e demografia no século XVII. A primeira tabela de vida foi apresentada por Graunt (1662).

Não é atoa que essa área da Estatística recebeu o nome de "Análise de Sobrevivência", uma vez que foi desenvolvida justamente para se poder calcular a real chance de sobrevivência de indivíduos sob chance de morte. Existe ainda um certo desconhecimento por parte de alguns do potencial das técnicas da Análise de Sobrevivência. Contudo, a área possui atualmente um sentido muito mais amplo, podendo ser aplicada para o estudo do tempo até a ocorrência de qualquer tipo de evento.

Sua utilização se faz de grande importância, já que as demais técnicas utilizadas no ramo da estatística não foram desenvolvidas para análise de dados ou observações incompletas, enquanto que a teoria da Análise de Sobrevivência possui propriedades que nos permite a análise desse tipo de dado. Essas observações incompletas ocorrem quando o indivíduo em estudo deixa de ser observado por algum motivo qualquer ao longo do tempo de estudo. Conforme Moore (2016), a Análise de Sobrevivência permite estimar valores com propriedades ótimas quando se tem perda da observação ao longo do tempo.

Um grande avanço no campo da análise de sobrevivência ocorreu a partir da década de 1950. A inauguração desta nova fase é representada pelo artigo de Kaplan e Meier (1958) que propõem um estimador para a curva de sobrevivência. Este é um dos artigos mais citados na história da Estatística. Enquanto que o método da tabela de vida clássica foi baseado em uma divisão grosseira do tempo em intervalos fixos, o método desenvolvido por Kaplan e Meier se adequava bem para intervalos curtos quanto para intervalos longos de tempo (AALEN et al., 2009). De fato, o artigo publicado por Kaplan e Meier em 1958 repercutiu bastante, e o método proposto no artigo ainda é atualmente a maneira mais utilizada de se estimar a função de sobrevivência de forma não-paramétrica, por suas ótimas propriedades assintóticas.

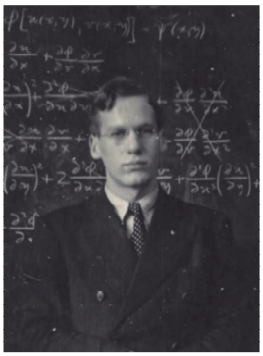


Figura 1 – Edward L. Kaplan

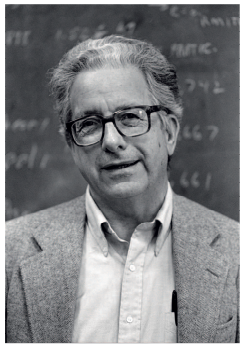


Figura 2 – Paul Meier

Outro grande marco na história da análise de sobrevivência foi a publicação do artigo do estatístico David Roxbee Cox em 1972 (COX, 1972), sendo este o segundo artigo mais citado da história, perdendo apenas para o artigo de Kaplan e Meier (1958). Cox apresentou em seu artigo um modelo de regressão que poderia ser utilizado para estudar o tempo até a ocorrência de um determinado evento de interesse, ajustando por covariáveis. O modelo de sobrevivência de Cox é também conhecido por modelo de riscos proporcionais devido a pressuposição de que as funções de riscos entre os grupos devem ser proporcionais.

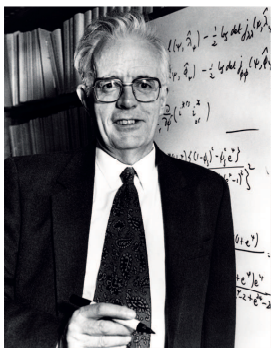


Figura 3 – Sir David Roxbee Cox

Mais tarde, Vaupel, Manton e Stallard (1979) introduziram um efeito aleatório no modelo de sobrevivência enquanto trabalhavam em um estudo demográfico, e chamaram

esse efeito de "fragilidade". Em seu estudo, eles afirmaram que a taxa do risco de mortalidade de um indivíduo aumenta mais rapidamente com a idade, quando comparada à taxa do risco de mortalidade da população, e assim, cada indivíduo possui uma taxa diferente do risco de morte. Esses modelos ficariam conhecidos por serem uma extensão do modelo de riscos proporcionais de Cox, também chamados de "Modelos de Fragilidade". Nesta teoria dos modelos de fragilidade é levado em consideração a heterogeneidade presente nos indivíduos, ou seja, o modelo leva em consideração que um determinado indivíduo é mais "suscetível" à sofrer um evento de interesse do que outro.

2.2 Análise de Sobrevivência

Análise de sobrevivência é uma classe de métodos estatísticos para se estudar o tempo até a ocorrência de um determinado evento de interesse. Estes métodos são mais frequentemente aplicados para o estudo de tempo de vida. De fato, eles foram originalmente desenvolvidos para esta finalidade, o que explica o nome. Por conta disto, há um aumento na visão restrita das potenciais aplicações desses métodos. Análise de Sobrevivência é extremamente utilizada para vários estudos de diferentes tipos de eventos em ambas as ciências naturais e sociais, incluindo início de doença, falha de equipamentos, terremotos, acidentes automobilísticos, falhas no mercado de ações, nascimentos, casamentos, divórcios, promoções, aposentadorias, entre outros (ALLISON, 2010).

A Análise de Sobrevivência pode ser conhecida como diversos outros nomes, a depender do campo de aplicação da técnica, tais como análise de confiabilidade (engenharia), análise de duração (economia), análise de tempo de falha (engenharia), análise de transição (economia), análise de eventos históricos (sociologia). Contudo, a alteração no nome, não afeta a construção dos métodos desenvolvidos (ALLISON, 2010).

2.2.1 Tempo

Segundo Colosimo e Giolo (2006), T é uma variável aleatória, não-negativa, usualmente contínua, que representa o tempo de falha. Essa variável é especificada em Análise de Sobrevivência pelas funções densidade de probabilidade $f(t)$, de sobrevivência $S(t)$, de risco $\lambda(t)$ e risco acumulado $\Lambda(t)$.

2.2.2 Censura

Em Análise de Sobrevivência estuda-se o tempo até a falha, sendo esta um determinado evento de interesse. Nestes estudos, é usual que nem todos os indivíduos sob investigação tenham sofrido o evento de interesse ao término do período de estudo. Dessa forma, uma característica decorrente em estudo de sobrevivência é a presença de observações incompletas ou parciais que são chamadas de censuras.

De acordo com Colosimo e Giolo (2006), em estudos clínicos alguns tipos de censura são diferenciáveis. Censura do tipo I é aquela em que o estudo será terminado após um período de tempo pré-estabelecido. Censura do tipo II é aquela em que o número de falhas é pré-fixado para o término do estudo. E censura aleatória é aquela em que o indivíduo observado é retirado do estudo sem que tenha ocorrido a falha, ou este indivíduo deixa de ser observado por outro motivo qualquer, que não esteja ligado a objetivo de estudo. Ainda segundo os autores, uma representação simples do mecanismo de censura aleatória é feita utilizando duas variáveis aleatórias independentes T e C, representando o tempo de falha de um indivíduo e o tempo de censura associado a este mesmo indivíduo respectivamente. Neste caso, observa-se o seguinte para este indivíduo,

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

e

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

em que, δ é o indicador de falha.

Sendo assim, os pares (T_i, C_i), para $i = 1, \ldots, n$, formam uma amostra aleatória de n indivíduos. E se todo $C_i = C$, uma constante fixa determinada pelo pesquisador, temos a censura do tipo I. Ou seja, a censura do tipo I é um caso particular da censura aleatória (COLOSIMO; GIOLO, 2006).

Esses mecanismos de censura são os tipos de censuras mais comum na prática e são conhecidos como censura à direita, pois o tempo de falha é maior que o tempo registrado. Existe ainda dois tipos de censura, que são a censura à esquerda e censura intervalar. A censura à esquerda ocorre quando o tempo de falha é menor que o tempo registrado e a censura intervalar ocorre quando não se sabe o tempo exato de ocorrência do evento de interesse, sabe-se que ele ocorreu em um intervalo especificado. Na Figura 4 é ilustrado os mecanismos de censura que foram descritos.

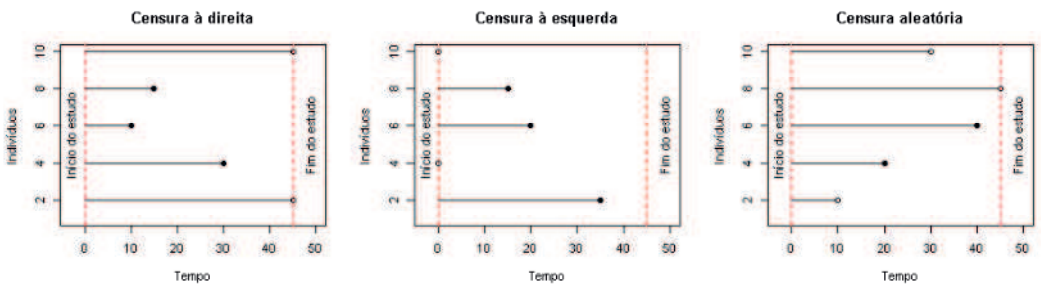


Figura 4 – Representação gráfica de censura, em que • representa falha e ∘ censura.

As censuras podem ainda ser classificadas em: informativa (aquela em que a perda da informação está ligada ao evento de interesse) e não informativa (aquela em que a perda da informação não está ligada ao evento de interesse).

Segundo Strapasson (2007), para análise de sobrevivência é necessário que as observações sejam representadas por um vetor (t_i, δ_i, x_i) em que, t_i é o tempo observado de falha ou censura e δ_i uma variável indicadora de falha, em que $\delta_i = 1$, se o tempo observado corresponde a uma falha ou $\delta_i = 0$, se corresponde a uma censura. Para cada indivíduo observado tem-se uma covariável x_i, em que $i = 1, \ldots, n$.

2.2.3 Truncamento

Alguns estudos de sobrevivência apresentam também outra característica peculiar, que chamamos de truncamento. Diferente da censura, no truncamento as informações ou observações não são perdidas devido à fatores associados (ou não) ao evento de interesse. Nestes estudos, o pesquisador delimita quais indivíduos devem pertencer a amostra. Um exemplo de dados truncados é considerar que, para que um determinado indivíduo entre no estudo o mesmo deverá experimentar um determinado evento. Um exemplo bem simples desse tipo de truncamento seria estudarmos adolescentes do sexo feminino que já tiveram a menarca (primeira menstruação), neste caso só deveriam ser incluídas na amostra as garotas que já tiveram a primeira menstruação. Outra forma de truncamento é fazer um "corte" na janela de tempo. Esta segunda forma de truncamento é mais comum em estudos retrospectivos, onde o pesquisador delimita a janela de tempo que deseja observar.

2.2.4 Função Densidade de Probabilidade

Em Análise de Sobrevivência a função de densidade de probabilidade é definida como o limite da probabilidade da falha ocorrer em um dado intervalo de tempo por unidade de tempo. Assim, se fixarmos o intervalo de tempo $[t, t + \Delta t]$ a função densidade de probabilidade é então expressa como em Diniz e Louzada (2012), da seguinte forma

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

em que, $f(t) \geq 0$ para todo t, e tem a área abaixo da curva igual a 1.

2.2.5 Função de Sobrevivência

Um das funções mais utilizadas em estudos de sobrevivência e de muita importância é a função de sobrevivência. De acordo com Moore (2016), é umas das maneiras chave para se descrever ou especificar a distribuição de sobrevivência, sendo definida como a probabilidade de sobreviver além do tempo t. Formalmente ou matematicamente é escrita como

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t), \tag{2.2}$$

em que $F(t)$ é função de distribuição acumulada.

Esta função recebe valor 1 no tempo 0, e decresce ou permanece constante ao longo do tempo, e a curva de sobrevivência nunca cai abaixo do valor zero (MOORE, 2016). Ou seja,

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

2.2.6 Função de Risco

Uma outra função bastante utilizada na análise de sobrevivência para modelar a taxa de falha dos indivíduos em estudo é a função de risco ou taxa de falha, a qual descreve a taxa de falha instantânea dos indivíduos envolvidos.

A função de risco $\lambda(t)$ é a taxa de falha instantânea no tempo t condicionada à sobrevivência até o tempo t. Essa função é bastante útil para descrever a distribuição do tempo de vida de pacientes, descrevendo a forma em que a taxa instantânea de falha muda ao longo do tempo.

Deste modo, fixando um intervalo de tempo t e $t + \Delta t$. A função taxa de falha ou função de risco é então definida da seguinte maneira

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.3}$$

2.2.7 Função de Risco Acumulado

A função de risco acumulado, assim como a função de sobrevivência e a função de risco vistas anteriormente, também caracteriza um papel muito importante na técnica de análise de sobrevivência. Assim como o próprio nome sugere, esta função nos fornece o risco acumulado ou taxa de falha acumulada dos indivíduos, e é definida como

$$\Lambda(t) = \int_0^t \lambda(u) du. \tag{2.4}$$

A função taxa de falha acumulada, $\Lambda(t)$, não tem uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é taxa de falha, $\lambda(t)$. Isto acontece essencialmente na estimação não paramétrica em que $\Lambda(t)$ apresenta um estimador com propriedades ótimas e $\lambda(t)$ é difícil de ser estimada (COLOSIMO; GIOLO, 2006).

2.2.8 Relações entre as Funções

As funções densidade de probabilidade, de sobrevivência, de risco e risco acumulado são matematicamente equivalentes, ou seja, dada uma função, as demais podem ser derivadas. A seguir será discutido sobre a relação dessas funções.

Colosimo e Giolo (2006) afirmam que a probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência da seguinte forma

$$S(t_1) - S(t_2).$$

Assim, a taxa de falha no intervalo $[t_1, t_2)$ é definida como a razão entre a probabilidade da falha ocorrer neste intervalo de tempo dado que não ocorreu antes de t_1 e o comprimento do intervalo. Assim sendo, a taxa de falha no intervalo pode ser expressa como

$$\frac{P[t_1 \leq T < t_2 | T > t_1]}{(t_2 - t_1)} = \frac{P[t_1 \leq T < t_2]}{(t_2 - t_1)S(T > t_1)} = \frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \tag{2.5}$$

Se considerarmos um intervalo geral $[t, t + \Delta t]$ a equação (2.5) pode ser reescrita como

$$\frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{2.6}$$

Deste modo, a taxa de falha instantânea no tempo t condicionada à sobrevivência até o tempo t, é dada por

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{2.7}$$

E dado que a função densidade de probabilidade pode ser obtida como a derivada de primeira ordem da função de distribuição acumulada, temos que

$$f(t) = \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t}. \tag{2.8}$$

Tendo em vista que a função de sobrevivência pode ser obtida em termos da função de distribuição acumulada como $S(t) = 1 - F(t)$, a função de risco em (2.7) pode então ser reescrita como

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)} = \frac{1}{S(t)} \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t}.$$

E portanto,

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{2.9}$$

Obtendo a função densidade de probabilidade em termos da função de distribuição acumulada, temos

$$f(t) = \frac{d}{dt}F(t).$$

Mas a função de distribuição acumulada $F(t) = 1 - S(t)$, logo

$$f(t) = \frac{d}{dt}\Big[1 - S(t)\Big] = -S'(t),$$

que substituindo na equação (2.9), temos

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}\log S(t).$$

Integrando a função acima, de zero a t e sabendo que $S(0) = 1$, temos que

$$\int_0^t \lambda(x)dx = -\log S(t),$$

ou seja,

$$\Lambda(t) = -\log S(t),$$

o que implica que

$$S(t) = \exp[-\Lambda(t)].$$

Daí, segue que as funções acima podem então ser obtidas pelas seguintes igualdades:

1. $f(t) = -S'(t)$;
2. $S(t) = \exp[-\Lambda(t)]$;
3. $\lambda(t) = \dfrac{f(t)}{S(t)}$;
4. $\Lambda(t) = -\log S(t)$.

2.2.9 Estimador de Kaplan-Meier

A análise descritiva consiste essencialmente em encontrar medidas de tendência central e variabilidade. Com a presença de censuras, este tipo de tratamento deve ser realizado através da análise de sobrevivência. O objetivo de uma análise estatística envolvendo dados de sobrevivência está relacionado com a identificação de fatores de prognóstico para uma certa doença ou à comparação de tratamentos em estudos clínicos.

Como a função densidade de probabilidade e a função de risco a partir dos dados amostrais não permitem a presença de observações censuradas as quais são comuns os dados de sobrevivência e confiabilidade, Kaplan e Meier (1958) propuseram um estimador em que as estimativas podem ser obtidas a partir de métodos não-paramétricos, que não supõem nenhuma distribuição conhecida. Este estimador é conhecido na literatura por estimador Kaplan-Meier ou estimador limite-produto, no qual permite a presença de observações censuradas como relatado por Colosimo e Giolo (2006).

O estimador não paramétrico de Kaplan-Meier é atualmente o mais utilizado na literatura para estimar a função de sobrevivência na presença de censura. Segundo e Colosimo e Giolo (2006) o estimador $\widehat{S(t)}$ de Kaplan-Meier, é definido por,

$$\widehat{S(t)} = \left(\frac{n_1 - d_1}{n_1}\right) \times \cdots \times \left(\frac{n_{t_0} - d_{t_0}}{n_{t_0}}\right) = \prod_{j,t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j,t_j<t} \left(1 - \frac{d_j}{n_j}\right), \quad (2.10)$$

em que,

i) $t_1 \leq t_2 \leq t_3 \leq \ldots \leq t_k$, os k tempos distintos e ordenados de falha;

ii) d_j: número de falhas até o tempo t_j, *j= 1, 2,. . . ,k*;

iii) n_j: número de indivíduos sob risco, ou seja, os indivíduos que não falharam e não censurados até t_j e

iv) t_0 é o maior tempo de falha menor que t.

Breslow e Crowley (1974) estudaram as principais propriedades do estimador, são elas: ele é não viciado para amostras grandes, é fracamente consistente para pequenas amostras, converge assintoticamente para um processo gaussiano e é estimador de máxima verossimilhança de $S(t)$.

2.2.10 Estimador de Nelson-Aalen

Uma outra alternativa para estimar a função de sobrevivência na presença de censura é o estimador de Nelson-Aalen, que é um estimador mais eficiente para uma pequena quantidade de obervações. De acordo Colosimo e Giolo (2006), este é um estimador baseado na função de sobrevivência, como segue,

$$S(t) = \exp\{-\Lambda(t)\}, \quad (2.11)$$

em que, $\Lambda(t)$ é a função de risco acumulado.

Um estimador para $\Lambda(t)$ foi inicialmente proposto por Nelson (1972) e retomado por Aalen (1978), que provou suas propriedades assintóticas usando processos de contagem (COLOSIMO; GIOLO, 2006). Na literatura, este estimador é conhecido como estimador de Nelson-Aalen e é dado da seguinte maneira:

$$\widetilde{\Lambda}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j}\right), \tag{2.12}$$

em que, d_j e n_j são os número de falhas e o número de indivíduos sob risco no tempo t_j respectivamente. Dessa forma, com base no estimador de Nelson-Aalen, um estimador para a função de sobrevivência é dado da forma:

$$\widetilde{S}(t) = \exp\left\{-\widetilde{\Lambda}(t)\right\}. \tag{2.13}$$

2.2.11 Estimador da Tabela de Vida

A tabela de vida foi o primeiro método desenvolvido para estimar a função de sobrevivência na presença de censura (KALBFLEISCH; PRENTICE, 2011). Este estimador é um resumo dos dados de sobrevivência agrupados em intervalos convenientes. Em alguns casos, os dados são coletados um em cada forma de agrupamento. Em outros casos, os dados podem ser agrupados para obter uma simples apresentação e maior facilidade de compreensão.

Supondo, por exemplo, que os dados são agrupados em intervalos $I_i, \ldots, I_k$ tal que

$$I_j = (t_0 + \cdots + t_{j-1}, t_0 + \cdots + t_j), \tag{2.14}$$

de largura t_j, com $t_0 = 0$. A tabela de vida apresenta então o número de falhas e censuras nos tempos de sobrevivência caindo em cada intervalo (KALBFLEISCH; PRENTICE, 2011).

Supondo que c_j represente os tempos de censura e d_j os tempos de falha no intervalo I_j, e seja $n_j = \sum_{i\geq j}(d_i + c_i)$ o número de indivíduos sob risco no início do j-ésimo intervalo. O estimador da tabela de vida padrão da probabilidade condicional de falha em I_j, é $\hat{q}_j = 1$ se $n_j = 0$ e

$$\hat{q}_j = \frac{d_j}{n_j - c_j/2}, \tag{2.15}$$

caso contrário. O termo $c_j/2$ no denominador é usado em uma tentativa de ajustar o fato de que nem todos n_j indivíduos estão sob risco para todo I_j. O estimador da tabela de vida correspondente a função de sobrevivência do ij-ésimo intervalo é finalmente, segundo Kalbfleisch e Prentice (2011), dado por

$$\widehat{S(t)} = \prod_{i=1}^{j}(1 - \hat{q}_i). \tag{2.16}$$

2.2.12 Comparação Entre as Curvas de Sobrevivência

Muitas vezes em estudos de sobrevivência é interessante testar se existe diferença significativa entre as probabilidades de sobrevivência em determinados grupos. Por exemplo, quando se quer verificar a eficiência de uma determinada droga sob pacientes que possuem uma determinada doença é interessante verificar se há diferenças entre o grupo que está recebendo a droga (tratado) e o grupo que não recebe a droga (controle). Para este fim, vários testes e generalizações foram propostos por muitos autores na literatura, tais como: Mantel (1966), Gehan (1965), Peto e Peto (1972), Prentice (1978), entre outros. Dois testes bastante conhecidos e muito utilizados para comparação entre as curvas de sobrevivência são os testes *Log-rank* e *Peto*, os quais serão abordados a seguir.

2.2.12.1 Teste Log-Rank

Este é o teste mais utilizado na literatura para comparação entre curvas de sobrevivência, sua aplicação é apropriada quando os grupos em estudo possuem a propriedade de riscos proporcionais. A hipótese nula relacionada a este teste é: $H_0 : S_1(t) = S_2(t)$. De acordo com Colosimo e Giolo (2006) a estatística do teste *log-rank* é dada por

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}\left(V_j\right)^2}, \tag{2.17}$$

em que d_{2j} caracteriza a falha dos indivíduos do grupo 2 no tempo j. w_{2j} e V_j é a média e a variância de d_{2j}, respectivamente, obtidos a partir da distribuição de d_{2j} o qual segue uma distribuição hipergeométrica com parâmetros n_j, n_{1j} e d_j, ou seja, $d_{2j} \sim hipergeom(n_j, n_{1j}, d_j)$, em que n_{1j} denota os indivíduos sob risco no grupo 1 no tempo j. Sob a hipótese nula $H_0 : S_1(t) = S_2(t)$, para todo t no período de acompanhamento, T tem uma distribuição qui-quadrado com 1 grau de liberdade para grandes amostras.

Colosimo e Giolo (2006) mostram uma generalização do teste *log-rank* para a comparação entre r curvas de sobrevivências nos r grupos distintos, em que a estatística de teste também segue uma distribuição qui-quadrado, com $(r - 1)$ graus de liberdade para este caso.

2.2.12.2 Teste Peto

No caso particular da comparação de duas funções de sobrevivência, a seguinte forma geral inclui os testes mais importantes na literatura e generaliza a estatística T da equação (2.17) [ver(Colosimo e Giolo (2006))]:

$$S = \frac{\left[\sum_{j=1}^{k} u_j(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j \left(V_j\right)^2}, \tag{2.18}$$

em que u_j são os pesos que especificam os testes. Sob a hipótese nula de que as funções de sobrevivência não diferem entre os grupos, a estatística S tem distribuição qui-quadrado com 1 grau de liberdade para grandes amostras (COLOSIMO; GIOLO, 2006). Em particular o teste *Log-rank* é obtido quando consideramos que $u_j = 1$, com $j = 1, 2, \ldots, k$.

Peto e Peto (1972), Prentice (1978) e Colosimo e Giolo (2006) sugerem utilizar uma função do peso que depende diretamente da expressão passada de sobrevivência observada das duas amostras combinadas. A função do peso é uma modificação do estimador de Kaplan-Meier e é definido de tal forma que seu valor é conhecido antes da falha ocorrer. O estimador da função de sobrevivência é

$$\widetilde{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j + 1 - d_j}{n_j + 1}\right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j + 1}\right), \tag{2.19}$$

e os pesos utilizados são:

$$u_j = \widetilde{S}(t_j - 1)\frac{n_j}{n_j + 1}. \tag{2.20}$$

Este estimador é conhecido como Peto-Prentice (COLOSIMO; GIOLO, 2006). A principal diferença entre os testes *Log-rank* e *Peto* é que no teste *Peto* é feita uma ponderação relativa à experiência de sobrevivência anterior. O que não ocorre no teste *Log-rank*.

2.3 Modelos Probabilísticos em Análise de Sobrevivência

Conforme Colosimo e Giolo (2006), os modelos de regressão para modelagem paramétrica do tempo de sobrevivência dos indivíduos são também conhecidos de uma maneira geral como modelo de tempo de vida acelerado. Uma vez que a função das covariáveis é acelerar ou desacelerar o tempo de vida.

Nesta seção serão abordados os modelos paramétricos utilizados para modelar o tempo de sobrevivência neste trabalho. Serão apresentadas as funções densidade de probabilidade, de sobrevivência, de risco e risco acumulado pertinentes a cada modelo.

2.3.1 O Modelo Exponencial

Este é um dos mais simples modelos utilizados para a modelagem do tempo de falha em estudos de sobrevivência. O modelo exponencial assume que o risco de sofrer o evento é constante ao longo do tempo. Por esse fato essa distribuição não é frequentemente

utilizada neste tipo de estudo. Contudo, em algumas situações o modelo consegue modelar de forma efetiva a taxa de risco com característica constante.

Se uma variável aleatória T que representa o tempo de falha segue uma distribuição exponencial a qual denotamos por ($T \sim \text{Exp}(\lambda)$), Então as funções densidades de probabilidade, de sobrevivência, de risco e risco acumulado são, respectivamente:

Função densidade de probabilidade:

$$f(t) = \lambda e^{-\lambda t}, \qquad \lambda > 0. \tag{2.21}$$

Função de sobrevivência:

$$S(t) = e^{-\lambda t}. \tag{2.22}$$

Função de risco:

$$\lambda(t) = \lambda. \tag{2.23}$$

Função de risco acumulado:

$$\Lambda(t) = \lambda t. \tag{2.24}$$

Com média e variância, dadas respectivamente por:

$$\begin{aligned} E(T) &= \frac{1}{\lambda}, \\ Var(T) &= \frac{1}{\lambda^2}. \end{aligned}$$

A modelagem do tempo de sobrevivência T considerando a presença de p covariáveis (ou variáveis explicativas) pode ser feita através do modelo de regressão com resposta exponencial da seguinte forma:

$$T = \exp\left\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right\} \epsilon, \tag{2.25}$$

em que, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ é o vetor de parâmetros associado as covariáveis $\mathbf{x}^\top = (x_1, x_2 \ldots, x_p)$ e ϵ é o termo de erro aleatório que segue uma distribuição exponencial.

Note que o modelo de regressão descrito em (2.25) é linearizável se tomarmos o logaritmo de T, ou seja, o modelo pode ser reescrito da seguinte maneira:

$$\log(T) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon, \tag{2.26}$$

onde, $\varepsilon = \log(\epsilon)$. Dessa forma, ε segue uma distribuição do valor extremo. Segundo Colosimo e Giolo (2006) essa distribuição é bastante utilizada em Análise de Sobrevivência, pois caracteriza de forma adequada a distribuição do logaritmo de certos tempos de vida.

O modelo exponencial foi um dos primeiros modelos a serem amplamente utilizados, envolvendo a confiabilidade de componentes eletrônicos e sistemas técnicos. Perdendo competitividade com outros modelos probabilísticos, devido a não flexibilidade de sua função de risco.

2.3.2 O Modelo Log-normal

No modelo log-normal ($T \sim \log N(\mu, \sigma^2)$), o logaritmo natural $\ln(T)$ do tempo de vida T é assumido ser normalmente distribuído ($\ln(T) \sim N(\mu, \sigma^2)$). A função densidade de probabilidade de uma variável aleatória T que segue uma distribuição log-normal com parâmetros $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ é dada por

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t) - \mu}{\sigma}\right)^2\right\}, \tag{2.27}$$

com média e variância dadas respectivamente por [verWienke (2010)]

$$\begin{aligned} E(T) &= \exp\left(\mu + \frac{\sigma^2}{2}\right), \\ Var(T) &= \exp\left(2\mu + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right]. \end{aligned}$$

Segundo Colosimo e Giolo (2006) no modelo log-normal as funções de sobrevivência e de risco não apresentam um forma analítica explícita e podem ser representadas por

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right) \quad \text{e} \quad \lambda(t) = \frac{f(t)}{S(t)}, \tag{2.28}$$

em que $\Phi(t)$ denota a função de distribuição acumulada de uma distribuição normal padrão.

Apesar de ser matematicamente mais complexa, esta distribuição tem sido amplamente utilizada como distribuição da falha em diversas situações, tais como análise de insolação elétrica ou tempo de ocorrência de câncer de pulmão entre fumantes.

De acordo com Colosimo e Giolo (2006) o modelo de regressão log-normal é obtido generalizando a equação (2.26), incluindo um parâmetro de escala na formulação do modelo. Isto é equivalente a assumir que o termo ε segue uma distribuição normal com parâmetro de escala σ^2. Assim sendo, o modelo de regressão fica dado por

$$\log(T) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \sigma\varepsilon = \mathbf{x}^\top \boldsymbol{\beta} + \sigma\varepsilon, \tag{2.29}$$

onde, T segue uma distribuição log-normal com parâmetro de forma μ e parâmetro de escala σ^2.

2.3.3 O Modelo Log-logístico

De acordo com Wienke (2010) a distribuição log-logística tem uma forma funcional bastante flexível, é um dos modelos de tempo de vida paramétricos em que a taxa de risco pode ser crescente ou decrescente.

Colosimo e Giolo (2006) definem as funções densidade, sobrevivência, risco e risco acumulado para o modelo log-logístico da seguinte forma:

Função densidade de probabilidade:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}\Big(1 + (t/\alpha)^{\gamma}\Big)^{-2}, \tag{2.30}$$

com $\alpha > 0$ e $\gamma > 0$, os parâmetros de forma e escala respectivamente.

Função de sobrevivência:

$$S(t) = \frac{1}{1 + (t/\alpha)^{\gamma}}. \tag{2.31}$$

Função de risco:

$$\lambda(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha\Big[1 + (t/\alpha)^{\gamma}\Big]}. \tag{2.32}$$

Função de risco acumulado:

$$\Lambda(t) = \ln(1 + (t/\alpha)^{\gamma}). \tag{2.33}$$

A média e a variância da variável aleatória T são dadas respectivamente por

$$\begin{aligned} E(T) &= \frac{\pi\alpha Csc(\pi/\gamma)}{\gamma}, \\ Var(T) &= \frac{2\pi\alpha^2 Csc(2\pi/\gamma)}{\gamma} - E(T)^2. \end{aligned}$$

A forma geral da função de risco de uma distribuição log-logística é bastante similar a forma da função de risco na distribuição log-normal. A distribuição log-logística pode ser obtida como mistura de distribuições Gompertz (variável que segue uma distribuição Gama com média e variância iguais a 1) (WIENKE, 2010).

Para a distribuição log-logística o modelo de regressão pode ser obtido como visto em (2.29), em que o tempo de sobrevivência T tem distribuição log-logística e de maneira correspondente o termo ε segue uma distribuição logística.

2.4 Modelo de Riscos Proporcionais de Cox

Em muitas situações é difícil ajustarmos um modelo paramétrico que represente de forma efetiva o tempo de sobrevivência dos indivíduos em determinados estudos. Devido à

esse problema, Cox (1972) propôs um novo modelo de regressão para análise de dados de sobrevivência em que a pressuposição de que o tempo de sobrevivência T não necessitasse seguir uma distribuição de probabilidade conhecida. Com isso, o modelo de regressão de Cox passou a ser um dos modelos mais utilizados na literatura para modelagem de dados de sobrevivência, devido à sua grande flexibilidade.

Uma das principais pressuposições para utilização do modelo de regressão de Cox é que os riscos dos indivíduos de cada grupo seja proporcional ao longo do tempo. Por esse motivo o modelo de Cox é também conhecido como modelo de riscos proporcionais. Esse modelo permite a análise de dados em que a variável resposta é o tempo até a ocorrência de um determinado evento de interesse, sendo ajustado por covariáveis.

Colosimo e Giolo (2006) definem o modelo de Cox da seguinte maneira. Considere p covariáveis, de modo que $\mathbf{x}$ seja um vetor com os componentes $\mathbf{x} = (x_1, \ldots, x_p)^\top$. A expressão geral do modelo de regressão de Cox considera:

$$\lambda(t) = \lambda_0(t) g(\mathbf{x}^\top \boldsymbol{\beta}), \tag{2.34}$$

em que g é uma função não-negativa que deve ser especificada, tal que $g(0) = 1$. Este modelo é composto de dois componentes, um não-paramétrico e outro paramétrico. O componente não-paramétrico, $\lambda_0(t)$, não é especificado e é uma função não-negativa do tempo. Este componente é usualmente chamado de função de base, pois $\lambda(t) = \lambda_0(t)$ quando $\mathbf{x} = \mathbf{0}$. O componente paramétrico é frequentemente usado na seguinte forma multiplicativa:

$$g(\mathbf{x}^\top \boldsymbol{\beta}) = \exp(\mathbf{x}^\top \boldsymbol{\beta}) = \exp(\beta_1 x_1 + \ldots + \beta_p x_p), \tag{2.35}$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis. Esta forma garante que $\lambda(t)$ seja sempre não-negativa. Outras formas para a função $g(\mathbf{x}^\top \boldsymbol{\beta})$ foram propostas por (Storer et al, 1983). Todavia, a forma multiplicativa é a mais utilizada.

2.5 Modelo de Fragilidade Compartilhada

Todos os métodos estatísticos definidos anteriormente para análise de dados de sobrevivência consideram a suposição de que os tempos de sobrevivência de indivíduos distintos são independentes, essa suposição é válida para muitas situações. Entretanto, existem situações em que a suposição de independência nos tempos de sobrevivência não é válida como, por exemplo, quando estamos estudando determinados grupos com características semelhantes entre si, é esperado que o comportamento do tempo de sobrevivência dos indivíduos de um determinado grupo apresente certas semelhanças que não são observáveis

em indivíduos de um grupo distinto. Neste caso, é razoável pensarmos que exista associação entre os tempos de sobrevivência de indivíduos de um mesmo grupo.

Nas situações em que supõe-se que existe uma associação nos tempos de sobrevivência dos indivíduos, o mesmo é caracterizado como dados de sobrevivência multivariados. Neste contexto, para considerar a existência dessa possível associação entre os tempos de sobrevivência, um modelo que tem sido usado com frequência é o modelo de fragilidade (*Frailty model*). Nesse modelo, um efeito aleatório denominado fragilidade, é introduzido na função de risco para descrever essa possível associação (COLOSIMO; GIOLO, 2006).

Existem diversas situações em que se faz necessário o uso dos modelos de fragilidade, em muitas delas a modelagem desse efeito aleatório se faz de forma univariada, onde se está modelando a heterogeneidade entre os indivíduos em estudo. Outros modelos bastante utilizados são os modelos de fragilidade compartilhada, que entra em um contexto de análise multivariada dessas obervações. Nos modelos de fragilidade compartilhada a ideia é modelar a fragilidade existente entre os grupos.

Wienke (2010) define o modelo de fragilidade compartilhada da seguinte maneira: suponha que existem n grupos e que o i-ésimo grupo possui n_i indivíduos, os quais estão associados com uma fragilidade não observada Z_i, $(1 \leq i \leq n)$. Um vetor $\boldsymbol{X}_{ij}$, $(1 \leq i \leq n, 1 \leq j \leq n_i)$ está associado com o j-ésimo indivíduo do i-ésimo grupo.

O modelo de fragilidade compartilhada é dado por

$$\lambda_{ij}(t) = z_i \lambda_0(t) \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta}) = z_i \lambda_{ij,c}(t), \tag{2.36}$$

onde z_i é o fator de risco comum para todos os indivíduos no grupo i, $\lambda_0(t)$ é o risco de base comum para todos os indivíduos, $\exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta})$ é o fator de especificações dos indivíduos que contribui para o risco, e $\lambda_{ij,c}(t)$ é o risco do indivíduo j do grupo i depois do efeito da fragilidade do grupo ter sido avaliado (DUCHATEAU; JANSSEN, 2008).

Quando a informação da covariável do indivíduo j do grupo i, $\mathbf{x}_{ij}$ é igual a $\mathbf{x}$, podemos reescrever a equação (2.36) da seguinte forma

$$\lambda_{ij}(t) = z_i \lambda_{x,c}(t). \tag{2.37}$$

Assume-se que as fragilidades Z_i são variáveis aleatórias independentes e identicamente distribuídas com um função de densidade comum $f(z, \theta)$, em que θ é o parâmetro da distribuição das fragilidades. Um modelo de fragilidade compartilhada semi-paramétrico é um modelo de fragilidade com uma função de risco de base não paramétrica $\lambda_0(t)$ (WIENKE, 2010).

De acordo com Duchateau e Janssen (2008), este modelo induz correlação entre os tempos de eventos de indivíduos de um mesmo grupo. É investigada a função de

sobrevivência conjunta para um grupo, neste contexto a transformação de Laplace é uma ferramenta matemática bastante utilizada. O termo de fragilidade tem um efeito na função de sobrevivência populacional, devido ao fato da experiência de indivíduos mais frágeis na média de eventos anteriores, as alterações na estrutura da população ao longo do tempo com um impacto óbvio na função de sobrevivência populacional.

Assumindo $j = 1, \ldots, n_i$, então a função de sobrevivência condicional conjunta para o grupo i é dada por

$$S_i(\boldsymbol{t}_{ni}) = \exp\left[-z_i(\Lambda_0(t_1)\exp(\mathbf{x}_{i1}^\top\boldsymbol{\beta}) + \ldots + \Lambda_0(t_{ni})\exp(\mathbf{x}_{in_i}^\top\boldsymbol{\beta}))\right], \tag{2.38}$$

com $\boldsymbol{t}_{ni} = (t_1, \ldots, t_{ni})$.

Usando a notação da equação (2.37) a função de sobrevivência conjunta para um grupo de tamanho n com informação da covariável $\mathbb{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top)^\top$ é obtida da função de sobrevivência condicional conjunta integrando com respeito a distribuição da fragilidade, da seguinte maneira

$$\begin{aligned} S_{\mathbb{X},f}(\boldsymbol{t}_n) &= \int_{-\infty}^{+\infty} \exp(-z\Lambda_{\mathbb{X},c}(\boldsymbol{t}_n)) f_Z(z)dz \\ &= E[\exp(-Z\Lambda_{\mathbb{X},c}(\boldsymbol{t}_n))], \end{aligned} \tag{2.39}$$

com $Z \sim f_Z$.

Note que a última linha da equação (2.39) é a transformação de Laplace de Z, $\mathcal{L}(s) = E[\exp(-Zs)]$, em que $s = \Lambda_{\mathbb{X},c}(\boldsymbol{t}_n)$, assim podemos escrever a função de sobrevivência conjunta para um grupo i em função da transformação de Laplace, da seguinte forma:

$$S_{\mathbb{X},f}(\boldsymbol{t}_n) = \mathcal{L}(\Lambda_{\mathbb{X},c}(\boldsymbol{t}_n)). \tag{2.40}$$

A transformação de Laplace tem um importante papel no estudo dos modelos de fragilidade (DUCHATEAU; JANSSEN, 2008).

Outra função que se faz necessário sua obtenção é a função de sobrevivência populacional, a qual não leva em consideração a presença de grupos entre os indivíduos. A partir desta, é possível descrever como a função de sobrevivência evolui ao longo do tempo para a escolha aleatória de um indivíduo com informação da covariável $\mathbf{x}$, sem levar em consideração os dados agrupados. Obviamente, desde que mais indivíduos frágeis morram mais cedo, a população sofre alterações ao longo do tempo, e isto tem um importante impacto na função de sobrevivência populacional (DUCHATEAU; JANSSEN, 2008). A função de sobrevivência populacional para os indivíduos tendo informação da covariável $\mathbb{X}$ é

obtida a partir da função de sobrevivência condicional $S_{ij}(t) = \exp(-z_i\Lambda_{\mathbf{x},c}(t))$ integrando sob a fragilidade com respeito a função densidade da fragilidade

$$S_{\mathbf{x},f}(t) = \int_0^\infty \exp(-z_i\Lambda_{\mathbf{x},c}(t))f_Z(z)dz. \tag{2.41}$$

A função de sobrevivência populacional pode também ser obtida por meio da transformação de Laplace da seguinte forma:

$$S_{\mathbf{x},f}(t) = \mathcal{L}(\Lambda_{\mathbf{x},c}(t)). \tag{2.42}$$

Uma vez que consideramos a existência de associação nos tempos de sobrevivência de indivíduos do mesmo grupo, é razoável pensarmos em mensurar essa dependência nos tempos de sobrevivência desses indivíduos. Uma medida bastante utilizada na literatura para medir a dependência de valores é o τ de Kendall.

Para dois grupos i e k de tamanho dois aleatoriamente escolhidos, os tempos de eventos são (T_{i1} e T_{i2}) e (T_{k1} e T_{k2}). Uma suposição típica para dados bivariados de sobrevivência é que a informação da covariável é a mesma em cada grupo, i.e., $\mathbb{X} = (\mathbf{x}_{i1}, \mathbf{x}_{i2}) = (\mathbf{x}_1, \mathbf{x}_2) = \mathbb{X}$, e novamente nós caímos ao subíndice $\mathbb{X}$ e $\mathbf{x}$ nas funções de sobrevivência conjunta e populacional (DUCHATEAU; JANSSEN, 2008).

A medida τ de Kendall (KENDALL, 1938) é definida da seguinte maneira

$$\tau = E[sinal((T_{i1} - T_{k1})(T_{i2} - T_{k2}))], \tag{2.43}$$

em que $sinal(x) = \{-1, 0, 1\}$ para $x < 0, x = 0, x > 0$. Uma formulação alternativa para distribuições contínuas é dado por

$$\begin{aligned} \tau &= P((T_{i1} - T_{k1})(T_{i2} - T_{k2}) > 0) - P((T_{i1} - T_{k1})(T_{i2} - T_{k2}) < 0) \\ &= 2P((T_{i1} - T_{k1})(T_{i2} - T_{k2}) > 0) - 1 \\ &= 2p - 1. \end{aligned} \tag{2.44}$$

Em estudos da fragilidade em modelos de regressão para dados de sobrevivência o τ de Kendall possui um recurso intuitivo forte. Podendo ser utilizado para medir a associação dos tempos de sobrevivência dos indivíduos em estudo.

2.5.1 Modelos de Fragilidade Paramétricos

Nesta subseção serão apresentados alguns modelos da componente de fragilidade com uma abordagem paramétrica. Serão mostradas algumas distribuições úteis para modelagem desse efeito aleatório incluso no modelo e as características inerentes a estes modelos. Outras distribuições para modelagem do componente aleatório presente no modelo (fragilidade) é apresentado em Duchateau e Janssen (2008).

O Modelo de Fragilidade Gama

Aqui serão expostos algumas características básicas da distribuição gama. A função densidade de probabilidade bi-paramétrica é dada por

$$f_Z(z) = \frac{\gamma^\delta z^{\delta-1}\exp(-\gamma z)}{\Gamma(\delta)}, \tag{2.45}$$

com $\delta > 0$ o parâmetro de forma e $\gamma > 0$ o parâmetro de escala. A transformação de Laplace é

$$\mathcal{L}(s) = \int_0^{+\infty} \exp(-zs) f_Z(z) dz = \gamma^\delta (s+\gamma)^{-\delta}. \tag{2.46}$$

Nesse caso, é possível obtermos a média e a variância da distribuição por meio da primeira e segunda derivada da transformação de Laplace, como segue

$$\mathcal{L}^{(1)}(s) = -\delta\gamma^\delta (s+\gamma)^{\delta-1}$$

e

$$\mathcal{L}^{(2)}(s) = \delta(\delta+1)\gamma^\delta (s+\gamma)^{-\delta-2}.$$

Avaliando essas derivadas com $s = 0$, temos

$$E(Z) = (-1)\mathcal{L}^{(1)}(0) = \delta/\gamma$$

e

$$Var(Z) = \mathcal{L}^{(2)}(0) - \left(-\mathcal{L}^{(1)}(0)\right)^2 = \delta/\gamma^2.$$

Segundo Duchateau e Janssen (2008) na modelagem da fragilidade a escolha típica dos parâmetros da distribuição gama é $\delta = \gamma$. Usando θ como notação para a variância de Z, temos que $E(Z) = 1$ e $Var(Z) = \theta = 1/\gamma$. Esta distribuição com parâmetros $(1/\theta, 1/\theta)$ é conhecida como distribuição gama uni-paramétrica com variância θ e sua função densidade é dada por

$$f_Z(z) = \frac{z^{1/\theta-1}\exp(-z/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}, \tag{2.47}$$

com a seguinte transformação de Laplace

$$\mathcal{L}(s) = (1+\theta s)^{-1/\theta}. \tag{2.48}$$

As funções de sobrevivência conjunta e populacional são dadas respectivamente por

$$S_{\mathbb{X},f}(t_1,\ldots,t_n) = (1+\theta\Lambda_{\mathbb{X},c}(\boldsymbol{t}_n))^{-1/\theta} \tag{2.49}$$

e

$$S_{\mathbf{x},f}(t) = \mathcal{L}(\Lambda_{\mathbf{x},c}(t)) = (1+\theta\Lambda_{\mathbf{x},c}(t))^{-1/\theta}. \tag{2.50}$$

Para o modelo Gama o τ de Kendall que é uma medida de dependência global é calculado utilizando a primeira e a segunda derivada da transformação de Laplace. Sendo dado por

$$\tau = \frac{\theta}{\theta+2}. \tag{2.51}$$

Mais detalhes de como se obter a medida τ de Kendall para o modelo de fragilidade Gama e outros modelos de fragilidade são expostos em Duchateau e Janssen (2008).

O Modelo de Fragilidade Lognormal

McGilchrist (1993) desenvolveu a metodologia para ajustar modelos de fragilidade em paralelo com a teoria dos modelos clássicos mistos. Assim ele propôs o modelo $\lambda_{ij}(t)\exp(\mathbf{x}_{ij}^{\top}\boldsymbol{\beta}+w_i)$, em que w_i é o valor atual do efeito aleatório W_i seguindo uma distribuição normal com média zero e variância γ. A fragilidade correspondente tem uma distribuição lognormal (DUCHATEAU; JANSSEN, 2008). Ou seja,

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\gamma}}\exp\left(-\frac{(\log u)^2}{2\gamma}\right), \tag{2.52}$$

com $\gamma > 0$. A média e a variância da fragilidade são dadas, respectivamente, por

$$\begin{aligned} E(Z) &= \exp(\gamma/2), \\ Var(Z) &= \exp(2\gamma)-\exp(\gamma). \end{aligned}$$

Para estudo das distribuições da fragilidade é frequente a escolha de uma parametrização em que $E(Z)=1$. Entretanto, para a distribuição lognormal, é natural assumirmos uma distribuição normal com média zero para o efeito aleatório W. Como consequência a média da fragilidade Z não é um.

No modelo de fragilidade lognormal a avaliação da transformação de Laplace não é explícita devido à complicações na obtenção das suas funções de sobrevivência e de risco. Outra medida que também não possui uma forma fechada para este modelo é o τ de Kendall, sendo este obtido através da transformação de Laplace.

2.5.2 Modelos de Fragilidade Semi-paramétricos

Aqui será discutido a extensão do modelo de riscos proporcionais de Cox, também conhecidos como modelos de fragilidade semi-paramétricos, nestes modelos existe a presença de um efeito aleatório. Serão apresentados os modelos Gama e Lognormal para modelagem da componente da fragilidade.

O ajuste dos modelos de fragilidade semi-paramétricos com distribuição Gama e Lognormal para a componente de fragilidade é feito através do método da verossimilhança parcial penalizada aproximada para as estimativas dos parâmetros, sendo que o modelo Gama pode ser ajustado por um método iterativo chamado algoritmo EM (Esperança Maximização). McGilchrist e Aisbett (1991) usaram um expressão de representação alternativa para o modelo para poder derivar a verossimilhança parcial penalizada. O modelo é representado da seguinte forma

$$\lambda_{ij}(t) = \lambda_0(t) \exp\left(\mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + w_i\right), \tag{2.53}$$

em que $w_i = \log z_i$. Onde w_i denota o efeito aleatório presente no modelo com variância γ e z_i denota a fragilidade com variância θ.

Fragilidades Gama

No modelo de fragilidade Gama consideramos que os efeitos aleatórios w_i em (2.53) seguem a seguinte distribuição

$$f_W(w) = \frac{\eta^{\nu} \exp(\nu w - \eta \exp(w))}{\Gamma(\nu)}, \qquad -\infty < w < \infty.$$

Isso nos leva a uma distribuição Gama bi-paramétrica para as fragilidades z_i, representadas pela seguinte função densidade de probabilidade, com parâmetros $\nu > 0$ e $\eta > 0$.

$$f_Z(z) = \frac{\eta^{\nu} z^{\nu-1} \exp(-\eta z)}{\Gamma(\nu)}, \qquad 0 \leq z < \infty. \tag{2.54}$$

Porém, é comum considerarmos uma distribuição Gama uni-paramétrica para as fragilidades, com média 1 e variância θ, cuja densidade pode ser vista em (2.47). Além disso, a modelagem é baseada no efeito aleatório $W = \log(Z)$ ao invés das fragilidades. Dada a função densidade para os $Z_i's$ em (2.47), então a função densidade para os $W_i's$ corresponde a

$$f_W(w) = \frac{(\exp(w))^{1/\theta} \exp(-\exp(w)/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}. \tag{2.55}$$

Fragilidades Lognormal

Se considerarmos a situação em que os efeitos aleatórios w_i seguem uma distribuição normal com média μ e variância γ. As fragilidades z_i seguem dessa forma uma distribuição lognormal com os parâmetros μ e γ representada pela função densidade de probabilidade dada pela equação (2.56), ver[Duchateau e Janssen (2008)].

$$f_Z(z) = \frac{1}{z\sqrt{2\pi\gamma}} \exp\left(-\frac{1}{2\gamma}(\log z - \mu)^2 \right), \tag{2.56}$$

com

$$\begin{aligned} E(Z) &= \exp(\mu + \gamma/2), \\ Var(Z) &= \exp(2\mu + \gamma)(\exp(\gamma) - 1). \end{aligned}$$

Se consideramos uma função densidade normal com média zero, chegaremos na equação dada em (2.52), com as respectivas média e variância para as fragilidades,

$$\begin{aligned} E(Z) &= \exp(\gamma/2), \\ Var(Z) &= \exp(\gamma)(\exp(\gamma) - 1). \end{aligned}$$

2.6 Estimação dos Parâmetros

Esta seção apresentará os métodos de estimação utilizados para estimar os parâmetros dos modelos de fragilidade compartilhada paramétricos e semi-paramétricos.

2.6.1 Estimação nos Modelos Paramétricos

De acordo com Munda et al. (2012) na análise de dados de sobrevivência agrupados com a presença de censura à direita, a observação por indivíduos $j \in J_i = \{1, \ldots, n_i\}$ do grupo $i \in I = \{1, \ldots, n\}$ é o par $\boldsymbol{u}_{ij} = (y_{ij}, \delta_{ij})$, em que $y_{ij} = \min(t_{ij}, c_{ij})$, ou seja, o mínimo entre o tempo de sobrevivência t_{ij} e o tempo de censura c_{ij}, e onde $\delta_{ij} = I(t_{ij} \leq c_{ij})$ é o indicador de falha. É também possível a inclusão da informação da covariável nessa estrutura, neste caso, $\boldsymbol{u}_{ij} = (y_{ij}, \delta_{ij}, \mathbf{x}_{ij})$ em que $\mathbf{x}_{ij}$ denota o vetor de covariáveis para a ij-ésima observação. Se a truncagem à esquerda também se faz presente, tempos truncados τ_{ij} são representados pelo vetor $\boldsymbol{\tau}$.

No ajuste paramétrico, a estimação é baseada na verossimilhança marginal em que as fragilidades tem sido integradas sob a média da verossimilhança condicional com a respectiva distribuição da fragilidade. Sob a suposição de censura à direita não informativa e independência entre o tempo de censura e o tempo de sobrevivência das variáveis aleatórias, dada a informação da covariável, a log-verossimilhança marginal dos dados observados $\boldsymbol{u} = \{\boldsymbol{u}_{ij}; i \in I, j \in J_i\}$ pode ser escrita como

$$\begin{aligned}\ell_{marg}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\gamma};\boldsymbol{u}|\boldsymbol{\tau}) &= \sum_{i=1}^{n}\left\{\left[\sum_{j=1}^{n_i}\delta_{ij}\Big(\log(\lambda_0(y_{ij}))+\mathbf{x}_{ij}^{\top}\boldsymbol{\beta}\Big)\right]\right. \\ &+ \log\left[(-1)^{d_i}\mathcal{L}^{(d_i)}\left(\sum_{j=1}^{n_i}\Lambda_0(y_{ij})\exp(\mathbf{x}_{ij}^{\top}\boldsymbol{\beta})\right)\right] \\ &- \left.\log\left[\mathcal{L}\left(\sum_{j=1}^{n_i}\Lambda_0(\tau_{ij})\exp(\mathbf{x}_{ij}^{\top}\boldsymbol{\beta})\right)\right]\right\},\end{aligned} \tag{2.57}$$

com $d_i = \sum_{j=1}^{n_i}\delta_{ij}$ o número de eventos no i-ésimo grupo, e $\mathcal{L}^{(q)}(.)$ a q-ésima derivada da transformação de Laplace da distribuição da fragilidade.

As estimativas de $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ são obtidas maximizando a log-verossimilhança, isto pode ser facilmente feito se for possível calcular as derivadas de ordem superior da transformação de Laplace $\mathcal{L}^{(q)}(.)$ até $q = \max\{d_1,\ldots,d_n\}$ (MUNDA et al., 2012).

2.6.2 Estimação nos Modelos Semi-paramétricos

A estimação dos parâmetros nos modelos de fragilidade semi-paramétricos com distribuições da componente de fragilidade Gama e Lognormal é feita por meio da verossimilhança parcial penalizada aproximada. No modelo Gama também é possível a estimação por meio do algoritmo EM (Esperança Maximização). Duchateau e Janssen (2008) mostram que as estimativas geradas pelo método da verossimilhança parcial penalizada são as mesmas encontradas pelo algoritmo EM.

Verossimilhança Parcial Penalizada

De acordo com Duchateau e Janssen (2008) na abordagem da verossimilhança parcial penalizada a verossimilhança completa dos dados consiste de duas partes. A primeira parte é a verossimilhança condicional dos dados dado as fragilidades, enquanto que a segunda parte corresponde a distribuição das fragilidades.

Nesta abordagem a segunda parte da verossimilhança é considerada ser um termo de penalidade. Se o valor atual do efeito aleatório está longe da média zero, o valor absoluto do logaritmo da função densidade avaliada neste valor será tipicamente grande e o termo de penalidade tem tipicamente uma contribuição negativa para verossimilhança completa dos dados. Adicionalmente, tomando os efeitos aleatórios como outro conjunto de parâmetros na primeira parte da verossimilhança, esta verossimilhança pode ser transformada em uma expressão de verossimilhança parcial (DUCHATEAU; JANSSEN, 2008). O que resulta em

$$l_{ppl}(\boldsymbol{\gamma},\boldsymbol{\beta},\boldsymbol{w}) = l_{part}(\boldsymbol{\beta},\boldsymbol{w}) - l_{pen}(\gamma,\boldsymbol{w}), \tag{2.58}$$

em que, $l_{ppl}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{w})$ é a verossimilhança completa dos dados, $l_{part}(\boldsymbol{\beta}, \boldsymbol{w})$ é a verossimilhança condicional dos dados dado as fragilidades e $l_{pen}(\gamma, \boldsymbol{w})$ é a função de penalidade. Com $\eta_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + w_i$ e $\boldsymbol{\eta} = (\eta_{11}, \ldots, \eta_{sn_s})$, temos que

$$l_{part}(\boldsymbol{\beta}, \boldsymbol{w}) = \sum_{i=1}^{s}\sum_{j=1}^{n_i} \delta_{ij}\left[\eta_{ij} - \log\left(\sum_{qw \in R(y_{ij})} \exp(\eta_{qw})\right)\right] \tag{2.59}$$

e

$$l_{pen}(\gamma, \boldsymbol{w}) = -\sum_{i=1}^{s} \log f_W(w_i), \tag{2.60}$$

onde, $f_W(w_i)$ é a função densidade para os efeitos aleatórios $w_i's$.

Função de penalidade no modelo com efeito aleatório normal

Para os efeitos aleatórios $w_i, i = 1, \ldots, s$, tendo uma densidade normal com média zero e variância γ, temos que a função de penalidade no modelo é dada por

$$l_{pen}(\gamma, \boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{s}\left(\frac{w_i^2}{\gamma} + \log(2\pi\gamma)\right). \tag{2.61}$$

Segundo Duchateau e Janssen (2008) a maximização da log-verossimilhança parcial penalizada é feita por meio de uma algoritmo iterativo em que é avaliado em dois passos: *"inner loop"* e *"outer loop"*. No *"inner loop"* é usado o processo de Newton-Raphson para maximizar, para um valor provisório de γ, $l_{ppl}(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{w})$ para $\boldsymbol{\beta}$ e $\boldsymbol{w}$ (melhores preditores linear imparciais). No *"outer loop"*, a maximização restrita da verossimilhança do estimador de γ é obtido usando os preditores. O processo é iterado até a convergência.

Função de penalidade no modelo com fragilidade gama

Considerando que os efeitos aleatórios são representados pela função densidade $f_W(w)$ dada em (2.55). A verossimilhança parcial penalidade para o modelo com fragilidade gama pode ser escrito como observado em (2.58), com a mesma expressão para a primeira parte da verossimilhança parcial dada em (2.59), mas agora com uma função de penalidade dada por

$$l_{pen}(\theta, \boldsymbol{w}) = -\frac{1}{\theta}\sum_{i=1}^{s}\Big(w_i - \exp(w_i)\Big). \tag{2.62}$$

A maximização da log-verossimilhança no modelo com fragilidade gama é similar à maximização vista para o modelo com efeitos aleatórios seguindo uma distribuição normal no passo *"inner loop"*. O que os diferencia é apenas a função de penalidade. No modelo gama, para um valor fixado de θ, as estimativas de $\boldsymbol{\beta}$ e $\boldsymbol{w}$ são obtidas.

A abordagem da verossimilhança parcial penalizada para o modelo com fragilidade gama usa um *"outer loop"* diferente do usado no modelo com efeito aleatório com densidade normal. No modelo gama, o *"outer loop"* é baseado na maximização de uma versão perfil da verossimilhança marginal para θ. Mais detalhes sobre a estimação dos parâmetros no modelo gama são encontrados em Duchateau e Janssen (2008).

Estimação via algoritmo EM

A estimação dos parâmetros no modelo semi-paramétrico sem a presença da fragilidade é realizada tomando a proposta da verossimilhança parcial, comentada na Seção 2.4. Para o modelo de fragilidade semi-paramétrico é preciso levar em consideração a contribuição dos termos de fragilidades não observadas. Duchateau e Janssen (2008) sugere que uma solução para isto, é considerarmos a verossimilhança parcial em combinação com o algoritmo Esperança-Maximização (algoritmo EM), o qual é tipicamente utilizado na presença de informações não observadas.

O algoritmo itera entre um passo de esperança e um passo de maximização. No passo esperança, os valores esperados das fragilidades não observadas condicionadas à informação observada de uma estimativa atual do parâmetro são obtidos. No passo maximização, esses valores esperados são considerados ser uma informação verdadeira, e novas estimativas dos parâmetros de interesse são obtidas pela maximização da verossimilhança, dado os valores esperados (DUCHATEAU; JANSSEN, 2008).

A utilização do algoritmo para um determinado problema depende de duas condições. A primeira é que deve ser fácil a obtenção dos valores esperados da informação não observada. A segunda é que a maximização da verossimilhança condicionada aos valores esperados da informação não observada deve ser simples, já que o algoritmo é baseado na execução desses dois passos iterativos. Estas condições são satisfeitas para o modelo de fragilidade semi-paramétrico gama. Mais detalhes sobre a aplicação dos princípios do algoritmo no modelo de fragilidade gama são dados em Duchateau e Janssen (2008).

2.7 Seleção de Modelos

2.7.1 Critério de Informação de Akaike

A ideia básica no Critério de Informação de Akaike (AIC) é ajustar o modelo mais parcimonioso possível, ou seja, que tenha um menor número de parâmetros em comparação com o modelo contendo todos os parâmetros (modelo saturado), mas que consiga explicar ou descrever o fenômeno tão bem ou até mesmo melhor que o modelo saturado.

Para Moore (2016) uma das melhores maneiras de se avaliar os modelos estatísticos é através do cálculo de AIC, que consiste em avaliar a verossimilhança do modelo, penalizado

pelo número de parâmetros. O objetivo é encontrarmos o modelo tal que a quantidade abaixo seja minimizada.

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}) + 2k, \tag{2.63}$$

em que, $\ell(\hat{\boldsymbol{\beta}})$ é a log-verossimilhança do modelo e k é o número de parâmetros.

Segundo Klein e Moeschberger (2005) a inclusão de variáveis no modelo causa uma diminuição no valor de AIC, contudo, em algum ponto, o critério passa a aumentar, indicando que a inclusão de determinadas variáveis é desnecessário e não contribuirá para as estimativas dos parâmetros.

3 Material e Métodos

3.1 Material

O conjunto de dados utilizado neste trabalho é derivado do trabalho feito por Blair et al. em 1976, na Irlanda do Norte. A base de dados contém 394 observações de 197 pacientes com retinopatia diabética que faziam um tratamento de fotocoagulação à *laser*. Para cada paciente foi aleatorizado um dos olhos para receber o tratamento e o outro olho foi tido como controle. As variáveis presentes no banco de dados são: *id* (que é uma variável identificadora do indivíduo), *olho*, *status*, *tratamento*, *idade*, *tipo de laser* e *tipo de diabetes*. Esse conjunto de dados é apenas uma amostra aleatória do conjunto de dados original descrito no trabalho de Blair et al. (1980). É possível ter acesso à esses dados através do comando `data(rms)` no software R.

3.2 Métodos

O *software* R é atualmente uma das ferramentas mais utilizadas para análises estatísticas, abrangendo todas as técnicas disponíveis neste seguimento. A ferramenta permite também a modelagem da fragilidade em análise de sobrevivência, com implementações nos modelos de fragilidade paramétricos e semi-paramétricos. O ajuste dos modelos pode ser feito com auxílio de vários pacotes disponíveis no software, entre eles podemos citar: `survival` Therneau (2015), `frailtySurv` de Monaco, Gorfine e Hsu (2015), Gorfine, Zucker e Hsu (2006), `frailtyEM` Balan e Putter (2017), `parfm` Munda et al. (2012), `frailtyHL` Ha, Lee e Song (2001), `phmm` Donohue e Xu (2010), Donohue et al. (2011), Vaida, Xu et al. (2000), `coxme` Therneau (2015), Therneau e Grambsch (2000), `frailtypack` Rondeau, Mazroui e Gonzalez (2012).

Neste trabalho foram avaliadas as curvas de sobrevivência de Kaplan-Meier, os testes Log-rank e Peto para comparação entre essas curvas, bem como ajustes de modelos de regressão como o modelo de riscos proporcionais de Cox e os modelos de fragilidade compartilhada (paramétrico e semi-paramétrico). As análises foram feitas por meio do software R (R Core Team, 2018) na versão 3.5.0. Foram utilizados os pacotes `parfm` para o ajuste do modelo de fragilidade paramétrico, `frailtySurv` e `frailtyEM` para o ajuste do modelo semi-paramétrico. O gráfico da curva de sobrevivência foi plotado com o auxílio do pacote `surminer` e foram utilizados os pacotes `muhaz` e `rms` para plotar as curvas de Kaplan-Meier com a presenta de subgrupos.

4 Resultados e Discussão

4.1 Análise Descritiva

O início da análise se deu com o cálculo de algumas medidas descritivas do tempo de sobrevivência dos pacientes submetidos ou não ao tratamento à *laser* por meio do estimador de Kaplan-Meier. Dessa forma a Tabela 1 mostra um resumo das estimativas de Kaplan-Meier obtidas para os grupos tratado e controle.

Tabela 1 – Resumo das estimativas de Kaplan-Meier para os grupos tratado e controle.

Grupo	**N**	**Eventos**	**Mediana**	$\boldsymbol{IC_{95\%}}$
Controle	197	101	43,7	$[31,6; 59,8]$
Tratado	197	54	NA	NA

Por meio da Tabela 1 é possível observar que o número de indivíduos que sofreram o evento de interesse (cegueira) no grupo controle foi praticamente o dobro do número de indivíduos que sofreram o evento de interesse no grupo tratado, o que nos leva a acreditar que o tratamento à *laser* parece exercer um efeito positivo a cerca da cegueira desses pacientes. Ver-se também que para o grupo tratado, o estimador de Kaplan-Meier não forneceu a estimativa do tempo mediano de sobrevivência, isso se deu pelo fato de que o tempo de observação terminou sem que 50% dos pacientes do grupo tratado tenham sofrido o evento de interesse (cegueira). Hammes et al. (2015), Lee et al. (2017) também utilizaram o estimador de Kaplan-Meier para avaliar as curvas de sobrevivência de diferentes grupos.

Na Figura 5 é possível observar as curvas de sobrevivência obtidas pelo estimador de Kaplan-Meier para os grupos tratado e controle. Nesta figura, se faz disponível também a observação de um gráfico contendo o número de indivíduos em risco.

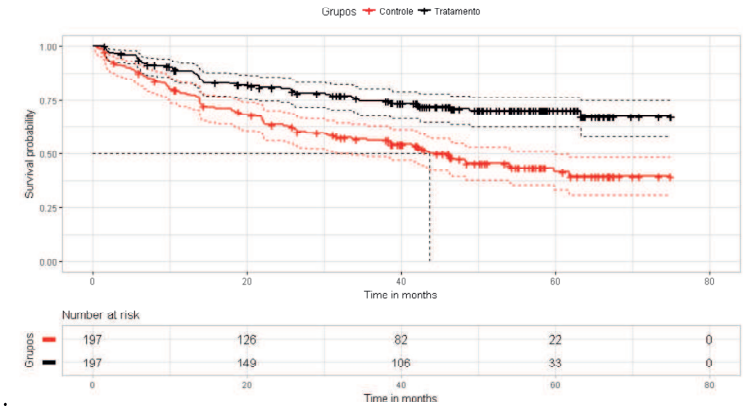


Figura 5 – Curvas de sobrevivência de Kaplan-Meier para os grupos tratado e controle.

A Figura 5 mostra que a curva de sobrevivência do grupo tratado permanece acima da curva de sobrevivência do grupo controle durante todo o período de estudo, ou seja, os indivíduos do grupo controle tinham uma maior probabilidade de sofrer o evento (cegueira). Nota-se também, no gráfico do número de indivíduos sob risco que, ao longo do tempo, os indivíduos do grupo controle sofriam mais o evento do que os indivíduos do grupo tratado, chegando ao final do estudo com mais pacientes cegos.

A Figura 6 mostra as curvas do risco acumulado para os grupos tratado e controle. Assim como esperado, observa-se que o grupo controle possui, durante todo o período de estudo, uma curva de risco superior à curva do grupo tratado. Reforçando a ideia de que o tratamento parece ter um controle sob a cegueira desses pacientes.

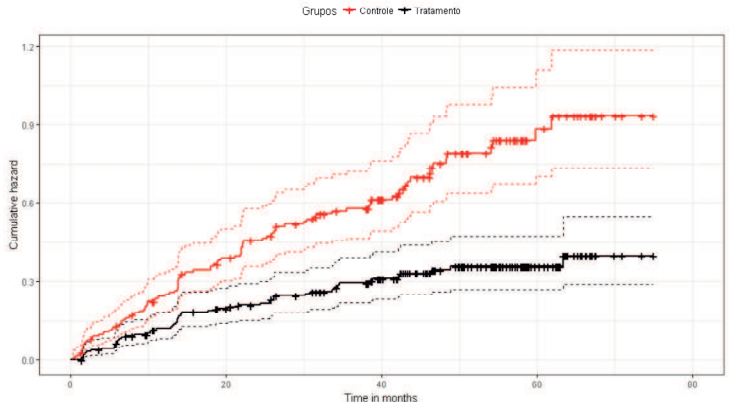


Figura 6 – Curvas de risco acumulado para os grupos tratado e controle.

Durante o período de estudo os pacientes foram submetidos à um tratamento com dois tipos diferentes de *lasers*: *xenon* e *argon*. Objetivou-se também descobrir se existe diferença entre as curvas de sobrevivência entre os pacientes que faziam tratamento com esses tipos de *lasers*. Na Figura 7 são expostas essas curvas de sobrevivência.

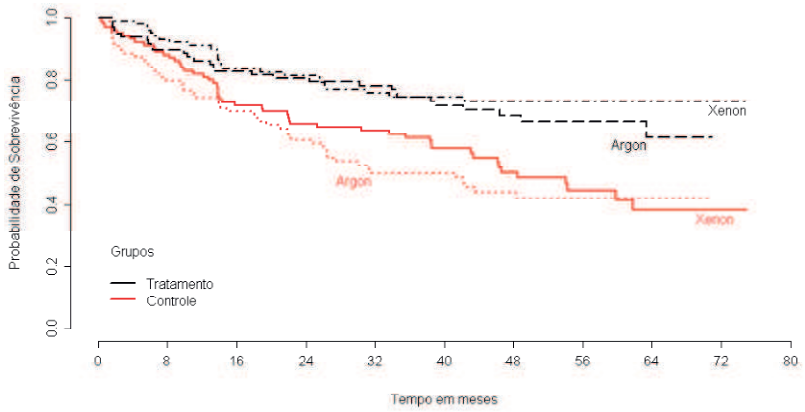


Figura 7 – Curvas de sobrevivência para os tipos de *lasers*.

A partir da Figura 7, pode-se perceber que para o grupo tratado, os pacientes que fizeram tratamento com o tipo de *laser xenon* tiveram uma probabilidade de sobrevivência superior aos pacientes que fizeram o tratamento com o *laser argon*. Enquanto que, para o grupo controle, os pacientes que tinham o olho tratado com o *laser xenon* tinham uma maior probabilidade de sobrevivência do que os pacientes que tinham o olho tratado com o *laser argon*, com este resultado alternando no final do estudo.

Um outro questionamento que se fez presente na formulação deste trabalho é se existia diferença entre as curvas de sobrevivência dos pacientes que tinham diabetes do tipo 1 e do tipo 2. Deste modo, a Figura 8 apresenta as curvas de Kaplan-Meier para esses grupos.

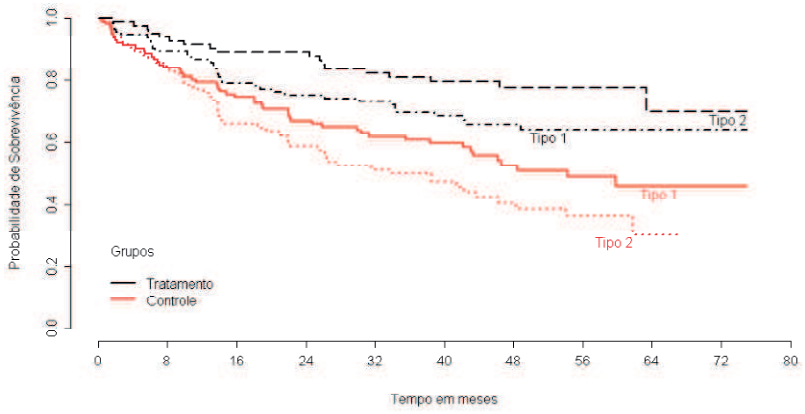


Figura 8 – Curvas de sobrevivência para os pacientes com diabetes do tipo 1 e do tipo 2.

Na Figura 8 é possível observar que para o grupo dos pacientes tratados, os indivíduos que tinham diabetes do tipo 1 eram mais suscetíveis a sofrer o evento quando comparados aos indivíduos que tinham diabetes do tipo 2. No grupo dos pacientes que não faziam tratamento, ver-se que os indivíduos que tinham diabetes do tipo 2 eram mais suscetíveis a sofrer o evento, ao invés dos indivíduos que tinham diabetes do tipo 1. Huster, Brookmeyer e Self (1989), também utilizaram o estimador de Kaplan-Meier para avaliar as funções de sobrevivência dos indivíduos com diabetes do tipo 2 nos grupos tratado e controle no estudo da Retinopatia Diabética, chegando aos mesmos resultados aqui apresentados. Entretanto, neste trabalho não foi utilizado um teste para verificar a diferença entre as curvas de sobrevivência desses grupos.

Os pacientes submetidos a este experimento, faziam tratamento em apenas um dos olhos, enquanto que o outro olho era controle, alguns faziam tratamento no olho direito, outros no olho esquerdo. A Figura 9 mostra as curvas de sobrevivência para os dois tipos de olho tratado, a fim de verificar se existe diferença entre eles.

As Figuras 5, 7, 8 e 9 tratam das diferenças entre as curvas de sobrevivência dos

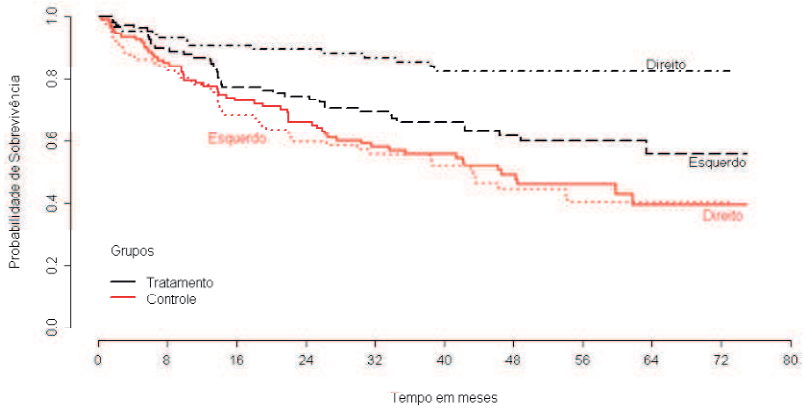


Figura 9 – Curvas de sobrevivência para os pacientes que faziam tratamento no olho direito e no olho esquerdo.

grupos em estudo, essas figuras indicaram que existiu uma diferença entre as curvas desses grupos, resta saber se essas diferenças são de fato significativas, ou seja, estatisticamente diferentes. Para tanto, foram utilizados os testes *Log-rank* e *Peto* para esses quatro grupos. A Tabela 2 mostra os resultados obtidos.

Tabela 2 – Testes *Log-rank* e *Peto* para diferença entre as curvas de sobrevivência dos grupos estudados.

	Log-rank		**Peto**	
Grupo	χ^2	**valor-p**	χ^2	**valor-p**
Tratado vs *Controle*	22,2	$< 0,0001$	20,7	$< 0,0001$
Laser xenon vs *Laser argon*	22,4	$< 0,0001$	20,9	$< 0,0001$
DM tipo 1 vs *DM tipo 2*	22,5	$< 0,0001$	20,8	$< 0,0001$
Olho direito vs *Olho esquerdo*	24,0	$< 0,0001$	23,1	$< 0,0001$

Na Tabela 2 é possível observar que para a primeira comparação entre as curvas de sobrevivência houve diferença significativa, o que significa dizer que a probabilidade de sobrevivência no grupo tratado é estatisticamente diferente da probabilidade de sobrevivência no grupo controle, ou seja, o tratamento à *laser* surtiu realmente efeito positivo para evitar ou retardar a cegueira nestes pacientes. Ver-se também que houve diferença significativa para os tipos de *lasers*, o que nos leva a entender que um dos *lasers* é mais eficiente quanto ao tratamento contra a cegueira, na Figura 7 é possível ver que o *laser xenon* foi mais eficaz. Houve diferença significativa também no grupo de pacientes que tinha diabetes do tipo 1 e do tipo 2, o que significa que um dos tipos de diabetes é mais agressiva quanto a obter a cegueira, na Figura 8 é possível ver que a DM2 se mostrou mais agressiva. Silva (2012) também afirma em seu trabalho que a diabetes do tipo 2 é responsável por um maior número de pacientes cegos pela retinopatia diabética. Também houve diferença entre as curvas para o olho tratado (esquerdo ou direito), ou seja, para um

dos olhos é mais fácil obter a cegueira, esta relação pode ser vista na Figura 9. Hammes et al. (2015) também utilizaram o teste *Log-rank* para verificar as diferenças entre as curvas de sobrevivência de diferentes tipos de retinopatia em pacientes com diabetes do tipo 2 na Alemanha e na Áustria.

4.2 Modelagem Estatística

Nesta seção serão apresentadas as modelagens paramétricas e semi-paramétricas, fazendo uma comparação interna com o ajuste do modelo sem a consideração da existência da fragilidade, bem como uma comparação entre as duas abordagens distintas (paramétrica e semi-paramétrica) entre os modelos de fragilidade compartilhada.

4.2.1 Ajustes Paramétricos

O primeiro passo para esta etapa do trabalho foi identificar os melhores modelos que pudessem explicar a influência de covariáveis na variável resposta (tempo até a cegueira). Para tanto, foi utilizado o critério de informação de Akaike, no objetivo de verificar as melhores combinações entre as funções de risco e as distribuições para modelagem da fragilidade. Deste modo a Figura 10 mostra os valores de AIC para os modelos em questão.

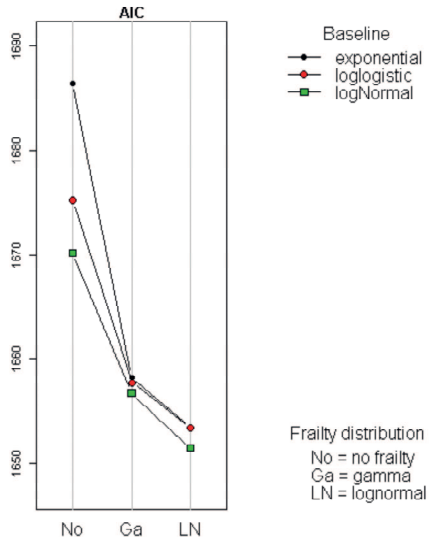


Figura 10 – Valores de AIC para as distribuições do risco e da fragilidade.

Percebe-se que em todos os casos o ajuste dos modelos sem a consideração da presença da fragilidade gerou os maiores valores de AIC, comprovando que o termo fragilidade é de grande importância para melhor predição dos valores reais da variável resposta. É possível ver também que o modelo que obteve o menor valor de AIC foi a combinação entre o risco com distribuição lognormal com a fragilidade tendo distribuição

também lognormal. Sendo assim, este foi o modelo escolhido para ser utilizado, desta forma foi dado prosseguimento com as análises com esse modelo.

A seguir é exposto na Tabela 3 as estimativas dos parâmetros das variáveis que foram significativas para os modelos, considerando o modelo sem a fragilidade e o modelo de fragilidade compartilhada, com seus respectivos valores de AIC.

Tabela 3 – Estimativas dos parâmetros para o modelo paramétrico sem a fragilidade e o modelo de fragilidade compartilhada.

MPSF (AIC = 1664,885)				
Covariável	**Coef**	**exp(Coef)**	**SE**	**valor-p**
Tratamento_tratado	-0,491	0,612	0,219	0,025 (<0,05)
Olho_esquerdo	0,319	1,376	0,163	0,050
Diabetes_tipo 2	0,310	1,363	0,199	0,119
Tratado_diabetes tipo 2	-0,782	0,457	0,352	0,026 (<0,05)
MPFCL (AIC = 1645,542)				
Covariável	**Coef**	**exp(Coef)**	**SE**	**valor-p**
Tratamento_tratado	-0,656	0,519	0,240	0,006 (<0,01)
Olho_esquerdo	0,490	1,632	0,185	0,008 (<0,01)
Diabetes_tipo 2	0,389	1,475	0,268	0,147
Tratado_diabetes tipo 2	-0,908	0,403	0,371	0,015 (<0,05)

MPSF: Modelo Paramétrico sem a Fragilidade.

MPFCL: Modelo Paramétrico de Fragilidade Compartilhada Lognormal.

Na Tabela 3 notas-se que o modelo de fragilidade se adequou melhor que o modelo sem a fragilidade, corroborando com o estudo de Swain e Grover (2016) que utilizaram os modelos de fragilidade gama e inversa gaussiana para dados de pacientes com HIV/AIDS na Índia, tendo melhores resultados através dos modelos de fragilidade. É possível observar que após o processo de seleção de variáveis houve uma diminuição no valor do AIC para o modelo de fragilidade compartilhada, onde anteriormente com o modelo saturado tínhamos um valor de AIC igual a 1652 (Figura 10). Nota-se também que não houve efeito significativo da variável *Tipo de diabetes*, entretanto, houve um efeito significativo para a interação dessa variável com a variável *Tratamento*, por isso, esta permanece no modelo. Observa-se que na presença do tratamento há uma diminuição em cerca de 48,1% na cegueira desses pacientes e que as chances de ficar cego do olho esquerdo são maiores que no olho direito, ver-se também que os pacientes que tinham diabetes do tipo 2 e faziam o tratamento tinham maiores chances de não cegarem, com uma diminuição de 59,7% de obter a cegueira, isto foi visto também na Figura 8, em que os pacientes que tinham diabetes do tipo 2 tiveram maiores probabilidade de sobrevivência. O τ de Kendall para o modelo paramétrico foi estimado em 0,316 indicando a presença de associação entre os tempos de sobrevivência dos grupos (tratado e controle).

4.2.2 Ajustes Semi-paramétricos

Assim como nos modelos paramétricos, para os modelos semi-paramétricos foi-se em busca das melhores distribuições que pudessem modelar a fragilidade, neste caso, não foi preciso investigar as distribuições para o risco, uma vez que estamos utilizando o modelo semi-paramétrico de Cox. A Tabela 4 mostra os valores de AIC para as distribuições investigadas.

Tabela 4 – Valores de AIC para os modelos de fragilidade compartilhada semi-paramétricos.

Modelos	**AIC**
Gama	1697,694
Lognormal	2003,371
Nofrailty	1711,551

Na Tabela 4 percebe-se que o modelo gama gerou um valor de AIC inferior ao obtido para o modelo lognormal, o que implica em um melhor ajuste, sendo assim, trabalharemos com o modelo de fragilidade compartilhada semi-paramétrico gama daqui em diante. Neste seguimento a Tabela 5 mostra as estimativas dos parâmetros para o modelo de fragilidade gama e o modelo de cox convencional (sem a consideração da fragilidade).

Tabela 5 – Estimativas dos parâmetros para o modelo semi-paramétrico sem a fragilidade e o modelo de fragilidade compartilhada semi-paramétrico gama.

MSSF (AIC = 1711,551)				
Covariáveis	**Coef**	**exp(Coef)**	**SE**	**valor-p**
Tratamento_tratado	-0,482	0,617	0,220	0,011 (<0,05)
Olho_esquerdo	0,322	1,380	0,163	0,022 (<0,05)
Diabetes_tipo 2	0,311	1,365	0,200	0,116
Tratado_diabetes tipo 2	-0,785	0,456	0,352	0,0097 (<0,01)
MFCSG (AIC = 1697,694)				
Covariáveis	**Coef**	**exp(Coef)**	**SE**	**valor-p**
Tratamento_tratado	-0,635	0,530	0,233	0,008 (<0,01)
Olho_esquerdo	0,467	1,595	0,179	0,011 (<0,05)
Diabetes_tipo 2	0,366	1,442	0,270	0,174
Tratado_diabetes tipo 2	-0,916	0,400	0,366	0,013 (<0,05)

MSSF: Modelo semi-paramétrico sem a fragilidade.

MFCSG: Modelo de fragilidade compartilhada semi-paramétrico gama.

É possível observar que as estimativas dos parâmetros obtidas para o modelo semi-paramétrico gama são praticamente as mesmas que foram obtidas para o modelo paramétrico com fragilidade e risco lognormal, levando as mesas conclusões vistas anteriormente. Observa-se que o modelo com a presença da fragilidade gerou os menores valores de AIC para ambos os casos (paramétrico e semi-paramétrico), o que mostra a importância de se utilizar os efeitos aleatórios nos modelos de sobrevivência para podermos

gerar estimativas mais fidedignas. Para o modelo semi-paramétrico também foi calculado o τ de Kendall, que resultou em 0,348, corroborando com o resultado do modelo paramétrico e indicando a presença de associação nos tempos de sobrevivência dos grupos.

4.3 Testando a Adequabilidade dos Modelos

As duas próximas seções tratarão de uma análise de diagnóstico para os modelos ajustados nos casos paramétrico e semi-paramétrico. A partir disto é possível avaliar a adequação dos modelos aos dados e verificar qual dos modelos se ajustou melhor.

4.3.1 Adequação do Modelo Paramétrico

Devido à limitações encontrados na implementação do pacote `parfm` não foi possível avaliar o ajuste do modelo paramétrico. Sendo esta, uma perspectiva futura do estudo.

4.3.2 Adequação do Modelo Semi-paramétrico

Esta etapa do trabalho consiste em avaliar os resíduos do modelo ajustado a fim de verificar se os valores preditos pelo modelo se aproximaram bem dos valores observados. Sendo assim, na Figura 11 podemos ver uma gráfico dos resíduos de Martingal *vs* os valores ajustados para o modelo de fragilidade semi-paramétrico gama. Onde é possível observar que, não foram obtidos altos valores para os resíduos do modelo e um comportamento assimétrico é observado, bem como esperado para os resíduos de Martingal, o que nos dá a ideia da possibilidade de um bom ajuste.

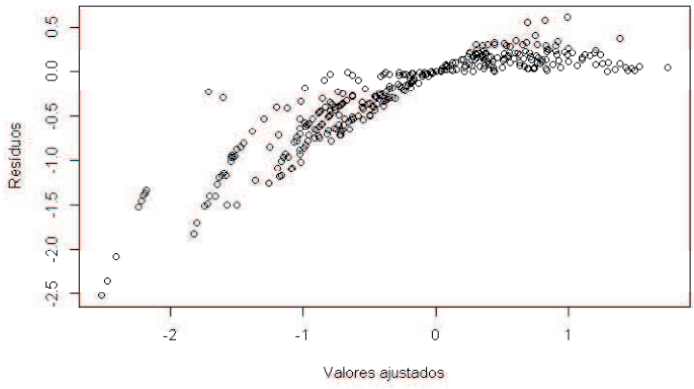


Figura 11 – Resíduos *vs* valores ajustados para o modelo de fragilidade semi-paramétrico gama.

De acordo com os resultados encontrados neste trabalho, os modelos de fragilidade compartilhada paramétrico e semi-paramétrico obtiveram melhores resultados do que a

modelagem clássica, baseado nos valores de AIC (Tabelas 3 e 5). E apesar de obtermos bons resultados para a adequabilidade do modelo semi-paramétrico, e de não se ter verificado a adequabilidade do modelo paramétrico, este último se mostrou mais competitivo.

5 Conclusão

Diante do exposto pode-se concluir que os modelos de fragilidade compõe uma ferramenta estatística ainda mais poderosa do que os modelos convencionais de sobrevivência para poder-se avaliar a influência de covariáveis na variável resposta quando existe a presença de associação nos tempos de sobrevivência dos indivíduos. Neste trabalho ficou claro que, em todos os casos, os modelos de fragilidade se comportaram melhor quanto à adequabilidade do ajuste, quando comparados aos modelos convencionais, gerando menores valores de AIC, e por conseguinte melhores estimavas para os parâmetros do modelo.

Por meio do teste Log-rank foi possível comprovar que o *laser xenon* teve maior eficácia para diminuir o risco da cegueira nos pacientes e que a diabetes do tipo 2 quando tratada obteve melhores chances de não obter a cegueira. É visível também neste estudo que, em ambos os casos, com e sem a presença da fragilidade, os modelos paramétricos se comportaram melhor que os modelos semi-paramétricos (via ajuste por Cox). Foi também visto que o modelo de fragilidade compartilhada lognormal com função de risco de base também lognormal obteve o melhor ajuste, sendo este portanto, o melhor modelo que explica o fenômeno aqui estudado. Os modelos de fragilidade conseguiram captar os mesmos resultados encontrados através dos testes Log-rank e Peto para o retardamento da cegueira nos pacientes.

Perspectivas futuras em relação a este trabalho é apresentar uma análise comparativa entre os modelos de fragilidade univariados e os modelos de fragilidade compartilhada (fragilidade multivariada). Bem como fazer um aprofundamento junto a análise residual para esses modelos.

Referências

AALEN, O. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, JSTOR, p. 701–726, 1978. Citado na página 25.

AALEN, O. O. et al. History of applications of martingales in survival analysis. *Electronic Journal for History of Probability and Statistics*, v. 5, n. 1, p. 1–28, 2009. Citado na página 17.

ALLISON, P. D. *Survival analysis using SAS: a practical guide*. 2. ed. North Carolina: Sas Institute, 2010. Citado na página 18.

BALAN, T. A.; PUTTER, H. *frailtyEM: An R Package for Estimating Semiparametric Shared Frailty Models*. 2017. Citado na página 43.

BLAIR, A. et al. The 5-year prognosis for vision in diabetes. *The Ulster medical journal*, Ulster Medical Society, v. 49, n. 2, p. 139, 1980. Citado 2 vezes nas páginas 15 e 43.

BRESLOW, N.; CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*, JSTOR, p. 437–453, 1974. Citado na página 24.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. 1. ed. São Paulo: Edgard Blucher, 2006. Citado 13 vezes nas páginas 18, 19, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31 e 32.

COX, D. R. Regression models and life tables (with discussion). *Journal Royal of Statistical Society, B*, v. 34, p. 187–220, 1972. Citado 2 vezes nas páginas 17 e 31.

DINIZ, C.; LOUZADA, F. Modelagem estatística para risco de crédito. *ABE, São Paulo-SP*, 2012. Citado na página 20.

DONOHUE, M. et al. Conditional akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, Oxford University Press, v. 98, n. 3, p. 685–700, 2011. Citado na página 43.

DONOHUE, M.; XU, R. phmm: proportional hazards mixed-effects model (phmm). *R package version 0.6*, v. 3, 2010. Citado na página 43.

DUCHATEAU, L.; JANSSEN, P. *The Frailty Models*. 1. ed. New York: Springer Science & Business Media, 2008. Citado 9 vezes nas páginas 32, 33, 34, 35, 36, 38, 39, 40 e 41.

GEHAN, E. A. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, Oxford University Press, v. 52, n. 1-2, p. 203–224, 1965. Citado na página 26.

GORFINE, M.; ZUCKER, D. M.; HSU, L. Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika*, Oxford University Press, v. 93, n. 3, p. 735–741, 2006. Citado na página 43.

GRAUNT, J. Natural and political observations upon the bills of mortality. *London, printed by Martyn J*, 1662. Citado na página 16.

HA, I. D.; LEE, Y.; SONG, J.-k. Hierarchical likelihood approach for frailty models. *Biometrika*, Oxford University Press, v. 88, n. 1, p. 233–233, 2001. Citado na página 43.

HAMMES, H.-P. et al. Risk factors for retinopathy and dme in type 2 diabetes—results from the german/austrian dpv database. *PloS one*, Public Library of Science, v. 10, n. 7, p. e0132492, 2015. Citado 2 vezes nas páginas 44 e 48.

HUSTER, W. J.; BROOKMEYER, R.; SELF, S. G. Modelling paired survival data with covariates. *Biometrics*, JSTOR, p. 145–156, 1989. Citado na página 46.

KALBFLEISCH, J. D.; PRENTICE, R. L. *The statistical analysis of failure time data*. 2. ed. New Jersey: John Wiley & Sons, 2011. v. 360. Citado na página 25.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado 2 vezes nas páginas 17 e 24.

KENDALL, M. G. A new measure of rank correlation. *Biometrika*, JSTOR, v. 30, n. 1/2, p. 81–93, 1938. Citado na página 34.

KLEIN, J. P.; MOESCHBERGER, M. L. *Survival Analysis: Techniques for Censored and Truncated Data*. 2. ed. New York: Springer Science & Business Media, 2005. Citado na página 42.

LEE, C. S. et al. The united kingdom diabetic retinopathy electronic medical record users group: report 3: baseline retinopathy and clinical features predict progression of diabetic retinopathy. *American journal of ophthalmology*, Elsevier, v. 180, p. 64–71, 2017. Citado na página 44.

MANTEL, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, v. 50, p. 163–170, 1966. Citado na página 26.

MCGILCHRIST, C.; AISBETT, C. Regression with frailty in survival analysis. *Biometrics*, JSTOR, v. 47, p. 461–466, 1991. Citado na página 37.

MCGILCHRIST, C. A. Reml estimation for survival models with frailty. *Biometrics*, JSTOR, v. 49, p. 221–225, 1993. Citado na página 36.

MONACO, J.; GORFINE, M.; HSU, L. frailtysurv: General semiparametric shared frailty model. *R package version*, v. 1, n. 2, 2015. Citado na página 43.

MOORE, D. F. *Applied survival analysis using R*. New Jersey: Springer, 2016. Citado 4 vezes nas páginas 16, 20, 21 e 41.

MUNDA, M. et al. Parfm: parametric frailty models in r. *Journal of Statistical Software*, Foundation for Open Access Statistics, v. 51, n. 11, p. 1–20, 2012. Citado 3 vezes nas páginas 38, 39 e 43.

NELSON, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, Taylor & Francis, v. 14, n. 4, p. 945–966, 1972. Citado na página 25.

PETO, R.; PETO, J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, JSTOR, p. 185–207, 1972. Citado 2 vezes nas páginas 26 e 27.

PRENTICE, R. L. Linear rank tests with right censored data. *Biometrika*, Oxford University Press, v. 65, n. 1, p. 167–179, 1978. Citado 2 vezes nas páginas 26 e 27.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado na página 43.

RONDEAU, V.; MAZROUI, Y.; GONZALEZ, J. R. frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation. *J Stat Softw*, v. 47, n. 4, p. 1–28, 2012. Citado na página 43.

SBEM. *O que é Diabetes?* 2017. Acesso em: 17 de out. 2017. Disponível em: <https://www.endocrino.org.br/o-que-e-diabetes/>. Citado na página 14.

SILVA, C. S. C. d. *Retinopatia diabética: relatório de estágio*. Tese (Doutorado) — Universidade da Beira Interior, Lisboa, jun. 2012. Citado 2 vezes nas páginas 14 e 47.

STRAPASSON, E. *Comparaçao de modelos com censura intervalar em análise de sobrevivência*. Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 20.

SWAIN, P. K.; GROVER, G. Accelerated failure time shared frailty models: Application to hiv/aids patients on anti-retroviral therapy in delhi, india. *Turkiye Klinikleri J Biostat*, v. 8, n. 1, p. 13–20, 2016. Citado na página 49.

THERNEAU, T. M. *A Package for Survival Analysis in S*. [S.l.], 2015. Version 2.38. Disponível em: <https://CRAN.R-project.org/package=survival>. Citado na página 43.

THERNEAU, T. M.; GRAMBSCH, P. M. *Modeling Survival Data: Extending the Cox Model*. New York: Springer, 2000. Citado na página 43.

VAIDA, F.; XU, R. et al. Proportional hazards model with random effects. *Statistics in medicine*, Citeseer, v. 19, n. 24, p. 3309–3324, 2000. Citado na página 43.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, Springer, v. 16, n. 3, p. 439–454, 1979. Citado na página 17.

WIENKE, A. *Frailty Models in Survival Analysis*. Boca Raton: CRC Press, 2010. Citado 3 vezes nas páginas 29, 30 e 32.