Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Modelo de regressão log de sobrevivências proporcionais para dados discretos com presença de censura

por

Tiago Chandiona Ernesto Franque

Brasília, 13 de setembro de 2024

Modelo de regressão log de sobrevivências proporcionais para dados discretos com presença de censura

por

Tiago Chandiona Ernesto Franque

Dissertação de mestrado apresentado ao Departamento de Estatística da Universidade de Brasília, como um dos requisitos para obtenção do grau de mestrado em Estatística

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Brasília, 13 de setembro de 2024

Proportional log survival model for censored discrete time-to-event data

by

Tiago Chandiona Ernesto Franque

Master's dissertation presented to the Department of Statistics of the University of Brasilia, as one of the requirements for obtaining the degree of Master of Statistics.

Advisor: Prof. Dr. Eduardo Yoshio Nakano

Brasilia, September 13th of 2024

Dissertação de mestrado submetido ao Programa de Pós-Graduação em Estatística do Departamento de Estatística da Universidade de Brasília, como requisito de obtenção do grau de mestrado em Estatística

Texto aprovado por:

Prof. Dr. Eduardo Yoshio Nakano

Orientador, EST/UnB

Prof. Dr. Frederico Machado Almeida

EST/UnB

Profa. Dra. Agatha Sacramento Rodrigues

DEST/UFES

Eu não vim até aqui pra desistir agora.

(Engenheiros do Hawaii)

Dedico este trabalho a mim mesmo, minha esposa Rosa da Felicidade Eugénio Benzane, as minhas filhas Wendy e Wynny e meu orientador Professor Eduardo Yoshio Nakano.

Agradecimentos

Primeiramente, a Deus, por todas as conquistas que já me fez alcançar não deixando em nenhum momento me faltar saúde, paz e disposição para realizar as minhas tarefas diárias. Certamente sem sua presença em minha vida nada seria possível, muito menos a execução deste trabalho. Agradeço a minha família, aos meus irmãos e a minha esposa Rosa da felicidade Eugénio Benzane que sempre me apoiaram em minhas decisões. Também agradeço as minhas filhas Wendy e Wynny, que são melhores presentes que Deus me concedeu e a motivação para enfrentar com disposição momentos difíceis. Aos professores Eduardo Yoshio Nakano e Frederico Machado Almeida pela ótima orientação e por dedicar bastante tempo compartilhando seus conhecimentos comigo, sempre com muita paciência e disposição. Aos meus amigos e colegas de Pós-graduação em especial ao Melquisadec, Leonardo, João, Moisés, Rebeca e Estevão pela amizade e companherismo indispensáveis. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Por fim, à todos que direta ou indiretamente contribuiram para que esse objetivo fosse alcançado.

Resumo

Um dos modelos de regressão mais populares na análise de dados de sobrevivência é o modelo de riscos proporcionais de Cox, cuja principal característica é considerar que as covariáveis atuem multiplicativamente na função de risco. No entanto, essa característica não pode ser satisfeita quando os tempos de sobrevivência são discretos, devido ao fato da função de risco ser limitada no intervalo (0,1). Neste contexto, este trabalho apresenta o desenvolvimento do modelo log de sobrevivências proporcionais para dados de sobrevivência discretos. Inferências dos parâmetros do modelo foram formuladas considerando dados com censura à direita e a distribuição Weibull discreta como distribuição basal dos dados. Estudos de simulação foram realizados para verificar as propriedades assintóticas dos estimadores. Ademais, procedimentos para a verificação da suposição de proporcionalidade do logaritmo da função de sobrevivência foram propostos e o modelo foi ilustrado por meio de um conjunto de dados sobre o tempo de sobrevivência de pacientes com leucemia.

Palavras-Chave: Análise de sobrevivência; modelo de regressão; distribuição discreta.

Abstract

One of the most popular regression models in survival data analysis is the Cox proportional hazards model, which considers that the covariates act multiplicatively on the risk function. However, this characteristic cannot be satisfied when survival times are discrete, due to the fact that the hazard function is limited in the interval (0,1). In this context, this work presents the development of the proportional log survival model for discrete time-to-event data. Inferences of the models parameters were formulated considering the presence of right censoring and the discrete Weibull baseline distribution. Simulation studies were carried out to check the asymptotic properties of the estimators. In addition, procedures for checking the proportional log survival assumption were proposed, and the proposed model was illustrated using a dataset on the survival time of patients with leukemia.

Keywords: Survival analysis; regression models; discrete distribution.

Sumário

1 Introdução **1**

2 Conceitos básicos **4**

2.1 Censura 4

2.2 Descrição do Comportamento do Tempo de Sobrevivência 6

2.2.1 Variáveis aleatórias contínuas (Tempos Contínuos) 6

2.2.2 Variáveis aleatórias discretas (tempos discretos) 10

2.3 Estimação Não Paramétrica 13

2.4 Estimação Paramétrica 14

2.4.1 Método de Máxima Verossimilhança 14

2.4.2 Intervalo de Confiança 15

2.4.3 Teste de Hipóteses 17

3 Revisão de literatura **20**

3.1 Alguns modelos de regressão em Análise de Sobrevivência 20

3.1.1 O Modelo de Riscos Proporcionais 20

3.1.2 O Modelo de Falha Acelerada 21

3.1.3 O Modelo Híbrido 22

3.1.4 O Modelo de Chances de Sobrevivência Proporcionais 22

3.2 Breve discussão sobre tempos discretos 25

3.2.1 Vantagens de Uso de Tempos Discretos em relação aos contínuos . . . 25
3.2.2 Discretização de Tempos Contínuos . . . 26
3.2.3 Discretização a Partir da Função de Distribuição Acumulada Contínua . 27
3.2.4 Distribuição Weibull Discreta . . . 28
3.2.5 Distribuição Log-Normal Discreta . . . 30
3.2.6 Distribuição Log-Logística Discreta . . . 31

4 Modelo de Regressão Log de Sobrevivências Proporcionais Para Dados Discretos 33
4.1 Formulação do Modelo . . . 33
4.1.1 Comportamento da função de risco . . . 35
4.2 Verificação da Suposição do MRLSP . . . 38
4.3 Procedimento para estimação pontual dos parâmetros . . . 40
4.4 Procedimentos para estimação intervalar dos parâmetros . . . 42
4.5 Procedimento para teste de hipóteses . . . 44

5 Simulações Computacionais 45
5.1 Tempo discretos sem presença de censura . . . 46
5.2 Tempos discretos com presença de censura . . . 48

6 Aplicação em Dados Reais 52
6.1 Dados de Leucemia . . . 52
6.1.1 Verificação da suposição de log de sobrevivência Proporcionais . . . 55
6.1.2 Comparação do MRLSP-WD com o Modelo de Regressão Chance de sobrevivência Proporcionais WD e Modelo de Regressão WD . . . 56

7 Considerações Finais 60

Referências Bibliográficas 61

Referências Bibliográficas **65**

Lista de Tabelas

3.1 Relações entre os modelos RP, FA e Híbrido. 22

5.1 Casos das simulações. 46

5.2 Média das estimativas, EQM e PC dos parâmetros do MRLSP-WD considerando os casos da simulação e diversos tamanhos amostrais na ausência de censuras. 47

5.3 Média das estimativas, EQM e PC dos parâmetros do MRLSP-WD considerando os casos da simulação e diversos tamanhos amostrais na presença de censuras para $n = 30$ e $n = 50$. 49

5.4 Média das estimativas, EQM e PC dos parâmetros do MRLSP-WD considerando os casos da simulação e diversos tamanhos amostrais na presença de censuras para $n = 100$ e $n = 500$. 50

6.1 Tempo de sobrevivência e Idade de 30 pacientes com AML. 53

6.2 Estimativas dos parâmetros do MRLSP-WD (dados da Tabela 6.1). 55

6.3 Estimativas dos parâmetros do MRCSP-WD (dados da Tabela 6.1). 58

6.4 Estimativas dos parâmetros do MRWD (dados da Tabela 6.1). 58

6.5 Erros máximos provenientes da estimação dos modelos MRLSP-WD, MRCSP-WD e MRWD (dados da Tabela 6.1). 59

Lista de Figuras

3.1 Riscos da distribuição WD para diferentes valores dos parâmetros. 30

3.2 Riscos da distribuição LND para diferentes valores dos parâmetros. 31

3.3 Riscos da distribuição LLD para diferentes valores dos parâmetros. 32

4.1 Taxas de falhas e respectivas razões entre os riscos do MRLSP-WD para risco base crescente e diversos valores de λ ($\lambda = 1$ representa a distribuição base). . . 36

4.2 Taxas de falha do MRLSP-WD para risco base constante e diversos valores de λ ($\lambda = 1$ representa a distribuição base). 37

4.3 Taxas de falhas e respectivas razões entre os riscos para risco base decrescente e diversos valores de λ ($\lambda = 1$ representa a distribuição base). 38

6.1 Função de sobrevivência via Estimador de Kaplan-Meier para o conjunto de dados da Tabela 6.1 54

6.2 Função de sobrevivência estimada do MRLSP-WD para cada nível da covariável Idade do conjunto de dados da Tabela 6.1. 56

6.3 Verificação de suposição de log de sobrevivência proporcionais para a covariável Idade. 57

6.4 Função de sobrevivência estimada pelo MRLSP-WD, MRCSP-WD e MRWD para cada nível da covariável Idade do conjunto de dados da Tabela 6.1. 59

Capítulo 1

Introdução

A análise de sobrevivência é uma das áreas da estatística responsável pela análise de dados cuja variável resposta é o tempo até a ocorrência de um evento de interesse (falha), geralmente atribuído ao óbito de um indivíduo (em dados clínicos) ou falha de um equipamento (em estudos de experimentos industriais). Uma das características desse ramo da estatística é a presença de observações incompletas ou parciais da variável resposta denominados dados censurados que são importantes e úteis nas análises, uma vez que, a omissão destes pode acarretar conclusões viciadas (Cardial, Cobre e Nakano, 2024).

Um aspecto comum dos estudos de sobrevivência é a presença de variáveis explicativas que representam a heterogeneidade, tais como, idade, sexo e tipo de tratamento. A forma mais eficiente de acomodar o efeito das variáveis explicativas, propondo uma relação entre os tempos de sobrevivência e as variáveis explicativas de interesse, é por meio de um modelo de regressão adequado para dados de sobrevivência.

Um dos modelos de regressão mais popular na análise de dados de sobrevivência é o modelo de Riscos Proporcionais (RP) de Cox, proposto por Cox (1972), que no referido artigo seminal também introduziu o modelo para tempos discretos, em que as covariáveis agem multiplicativamente na chance (odds) da função de risco. Um tratamento abrangente da metodologia estatística foi feito por Tutz e Schmid (2016) , para regressão semi-paramétrica e recentes ex-

tensões têm sido propostas em trabalhos como Berguer e Schmid (2018).

Uma limitação do modelo de RP está justamente no fato do mesmo impor a proporcionalidade dos riscos. Como forma de ultrapassar essa limitação, surgiram várias outras propostas de modelos, como o modelo de tempo de Falha Acelerado (FA) proposto por Kalbfleisch e Prentice (1980), o modelo híbrido, proposto por Ciampi e Etezadi-Amoli (1985), o modelo híbrido estendido (HE) proposto por Louzada-Neto (1997) e modelo de Cox com covariáveis dependentes no tempo (Cox e Oakes, 1984). Ademais, um modelo de chances de riscos proporcionais para dados discretos em sua versão paramétrica foi proposto em Vieira et al. (2023).

Apesar dos modelos de RP serem bem populares, o Modelo de Chances Proporcionais (MCP), em que covariáveis agem multiplicativamente na chance (odds) de sobrevivência, sendo denominado como Modelo de Chances de Sobrevivência Proporcionais (MCSP), é potencialmente um concorrente do modelo de RP, e também tem uma história bastante longa com sua utilização atrelada a dados contínuos. Tendo sido apresentado pela primeira vez em uma estrutura paramétrica por Bennett (1983), um extensivo estudo para demonstração das propriedades do modelo foi realizado por Murphy, Rossini e Vaart (1997). Posteriormente foi proposto por Yang e Prentice (1999) a inferência semi-paramétrica e as aplicações fazendo uso do referido modelo foram realizadas em Royston e Parmar (2002), Wang e Wang (2022) e Zhou, Zhang e Lu (2022). Ademais, Cardial, Cobre e Nakano (2024) apresentam a versão paramétrica do MCSP para dados discretos.

Os modelos (de risco, de chance de risco ou de chance de sobrevivência) proporcionais destacados anteriormente têm as suas suposições para um possível ajustamento de um conjunto de dados, e estas suposições não são sempre verificadas (satisfeitas). Nesse contexto, o objetivo deste trabalho é de apresentar mais uma alternativa de um modelo de regressão com estrutura proporcional para dados discretos com presença de censuras. Mais especificamente, a proposta é de apresentar um modelo de regressão log de sobrevivências proporcionais (MRLSP), que assume que as covariáveis agem multiplicativamente no logaritmo da função de sobrevivência.

O presente trabalho está dividido em sete capítulos. No Capítulo 2 são apresentados os

conceitos básicos de Análise de Sobrevivência, no Capítulo 3 apresenta-se a revisão de literatura com os principais modelos de regressão em Análise de Sobrevivência e uma breve discussão sobre tempos discretos. No Capítulo 4 é apresentado o modelo proposto que será designado por Modelo de Regressão Log de Sobrevivências Proporcionais. Neste capítulo é apresentado a formulação do modelo, procedimentos inferenciais e também propostas de verificação da suposição de proporcionalidade do modelo. O Capítulo 5 apresenta um estudo de simulação com o objetivo de verificar as propriedades assintóticas dos estimadores dos parâmetros do modelo. Uma aplicação do modelo em dados reais é apresentada no Capítulo 6, e no Capítulo 7 apresenta-se as considerações finais do trabalho.

Capítulo 2

Conceitos básicos

Neste capítulo abordar-se-á sobre os conceitos básicos de análise de sobrevivência com base em algumas referências bibliográficas. Na primeira seção deste capítulo debruçar-se-á sobre os principais conceitos de análise de sobrevivência assim como as funções que caraterizam este ramo de Estatística e suas relações mas considerando a variável aleatória contínua e discreta. Também ao longo deste capítulo faz se a caraterização do tempo discreto até a ocorrência do evento de interesse ou tempos contínuos discretizados tal como a sua tratativa considerando a sua distribuição de probabilidade. Neste capítulo também são apresentadas as vantagens de utilização de tempos discretos em relação aos contínuos. Assim, fez se também a representação gráfica das suas funções de risco discretizadas para compreender melhor as suas monotonicidades em função dos seus parâmetros. Por fim, é apresentando a inferência estatística numa abordagem clássica (frequentísta), apresentando a função de verossimilhança, intervalos de confiança e teste de hipóteses dos parâmetros do modelo.

2.1 Censura

A Análise de Sobrevivência é um ramo de Estatística que trata um conjunto de procedimentos estatísticos para análise de dados relacionados a varável resposta tempo até a ocorrência de

um determinado evento de interesse. Ela tem uma larga aplicação em diferentes áreas de conhecimentos como por exemplo: nas ciências de saúde e industriais, na qual os indivíduos em estudo são acompanhados até a ocorrência de um evento de interesse, que geralmente denotado como falha (ou morte).

Os autores como Cardial, Cobre e Nakano (2024), Colosimo e Giolo (2006) afirmam que a principal caraterística do estudo de Análise de Sobrevivência é a presença de dados censurados, que é a existência de observações incompletas ou parciais ou ainda quando não se observa o evento de interesse durante o tempo de acompanhamento, uma vez que, tal observação para todos os elementos de um determinado estudo nem sempre é possível. Sendo assim, há necessidade da introdução de uma variável que indique se o tempo de vida foi ou não observado. Essa variável é definida na literatura como variável indicadora de falha. É preciso lembrar que, mesmo censurados, todos os resultados provenientes de um estudo devem ser usados na análise estatística, pois a omissão da censura no ajuste de modelo poderá acarretar em estimativas viciadas, Colosimo e Giolo (2006).

As censuras são classificadas em três (3) tipos, sendo censuras à direita, à esquerda e intervalar. A seguir descreve-se cada tipo de censura:

(i) **Censura à direita**: quando o evento de interesse encontra - se à direita do tempo registrado;

(ii) **Censura à esquerda**: quando o evento de interesse está à esquerda do tempo registrado; e

(iii) **Censura intervalar**: quando não se sabe o tempo exato de ocorrência do evento de interesse, por exemplo os indivíduos são acompanhados em visitas periódicas, sabe-se apenas que o evento de interesse ocorreu em algum momento entre duas visitas.

Por outro lado, pode-se subclassificar a censura à direita em três subgrupos distintos, a saber:

- **Censura do tipo I**: o estudo termina após um período pré-estabelecido de tempo;

- **Censura do tipo II**: o estudo termina após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos estudados;
- **Censura do tipo aleatória**: ocorre quando um indivíduo é retirado no decorrer do estudo sem ter ocorrido a falha. Também ocorre, por exemplo, se o indivíduo falhar por uma razão diferente do evento de interesse.

No presente trabalho, será considerada apenas a censura aleatória, que é a mais utilizada em trabalhos de análise de sobrevivência. A caracterização do mecanismo de censura aleatória é descrita a seguir.

Considere T^* uma variável aleatória que representa o tempo até a falha e C uma variável aleatória, independente de T^* que representa o tempo até a censura. Dessa forma, $t = \min(C, T^*)$ denota o tempo observado do indivíduo com $\delta = I_{[0,C]}(T^*)$ representando a variável indicadora de falha (ou censura) que será igual a 1 se o tempo de sobrevida for observado e 0 se for censurado:

$$\delta = \begin{cases} 1, & \text{se ocorrer a falha} \quad T^* < C \\ 0, & \text{se ocorrer a censura} \quad T^* > C. \end{cases} \tag{2.1}$$

2.2 Descrição do Comportamento do Tempo de Sobrevivência

Na análise de sobrevivência, o comportamento da variável aleatória não-negativa que descreve o tempo até a falha (até a ocorrência do evento de interesse), é geralmente especificada pela sua função de sobrevivência ou pela função de risco (ou taxa de falha). Essas duas funções são relacionadas matematicamente, de forma que se uma delas é conhecida, a outra pode ser obtida.

2.2.1 Variáveis aleatórias contínuas (Tempos Contínuos)

O comportamento de uma variável aleatória contínua não-negativa, $T \geqslant 0$, pode ser caracterizada por meio de várias funções matematicamente relacionados. Entre elas tem-se a função

densidade de probabilidade, a função de sobrevivência e a função de risco. Essas funções serão descritas em detalhes a seguir.

Função densidade de probabilidade

De acordo com Nicolis, Meyer-Kress e Haubs (1983), a função densidade de probabilidades de T, denotada por $f(t)$, é aquela que satisfaz as seguintes condições:

$f(t) \geqslant 0$, para todo $t \geqslant 0$ e

$\int_0^{+\infty} f(t)dt = 1$

Essa função pode ser vista como o limite da probabilidade de um indivíduo experimentar o evento de interesse no intervalo de tempo $[t, t + \Delta t)$, dividida pelo comprimento do intervalo e pode ser expressa por:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t)}{\Delta t}, \quad t \geqslant 0. \tag{2.2}$$

Função de Sobrevivência

A função de sobrevivência, denotada por $S(t)$, é definida como a probabilidade de um indivíduo não falhar até um determinado tempo t, ou seja, a probabilidade desse indivíduo sobreviver além de t. Esta função é uma das principais funções probabilísticas para representar o tempo de sobrevivência e é definida por:

$$S(t) = P[T > t] = \int_t^{+\infty} f(t)dt, \quad \forall t \geqslant 0. \tag{2.3}$$

Função de Risco (Taxa de Falha) e Risco Acumulado (Taxa de Falha Acumulada)

A Função de Risco (ou função taxa de falha), denotada por $h(t)$, representa o risco instantâneo que o indivíduo tem de experimentar o evento de interesse em um determinado tempo t. No caso de uma variável aleatória contínua, esta função é definida como a razão do limite da

probabilidade condicional de um indivíduo experimentar o evento de interesse no intervalo de tempo $[t, t + \Delta t)$ dado que o mesmo não tenha experimentado o evento de interesse antes de t, sobre o intervalo de tempo Δt. A função de risco, denotado por $h(t)$, é definida por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}, \quad t \geqslant 0. \tag{2.4}$$

É importante notar que para variáveis aleatórias contínuas, a função de risco $h(t)$ é uma função que assume valores reais positivos e essa função não é limitada superiormente. A função $h(t)$ descreve como o risco (taxa de falha) se modifica com o passar do tempo. Por esse motivo, essa função é muito utilizada para descrever o comportamento do tempo de sobrevivência. Alguns autores consideram que a função taxa de falha é mais informativa que a função de sobrevivência, pois, diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de risco podem diferir drasticamente como afirma Colosimo e Giolo (2006).

Uma outra função importante que pode ser obtida a partir da função $h(t)$ é a Função de Risco Acumulada ou Taxa de Falha Acumulada, denotada por $H(t)$. A função $H(t)$ não tem uma interpretação direta, mas ela tem procedimentos de estimação não-paramétrica e na seleção de um modelo mais apropriado para ajustar um determinado conjunto de dados. A função $H(t)$ fornece o risco acumulado do indivíduo no tempo t e para o caso de uma variável contínua é definida por:

$$H(t) = \int_0^t h(u)du. \tag{2.5}$$

Relações importantes entre *f(t), S(t), h(t) e H(t)*

Matematicamente existe relações entre as funções $f(t)$, $S(t)$ e $h(t)$. Essas relações podem ser utilizadas para a obtenção de uma dessas funções quando uma outra é especificada.Se T é uma variável aleatória contínua, a função de risco pode ser obtida a partir das funções de

probabilidade e de sobrevivência a partir da seguinte relação:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P\left[(t \leqslant T < t + \Delta t) \cap (T \geqslant t)\right]}{\Delta t P(T \geqslant t)}$$

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

Logo a função de risco pode ser dada por:

$$h(t) = \frac{f(t)}{S(t)}. \tag{2.6}$$

A função densidade de probabilidades, $f(t)$, é definida como a derivada da Função de Distribuição Acumulada, $F(t)$, isto é,

$$f(t) = \frac{d}{dt} F(t). \tag{2.7}$$

Existe uma relação entre a Função de Distribuição Acumulada e a função de sobrevivência que é: $F(t) = 1 - S(t)$, substituindo a função $F(t)$ por $1 - S(t)$ na equação (2.7), tem - se:

$$f(t) = \frac{d}{dt}[(1 - S(t)] = -\frac{d}{dt} S(t) = -S'(t). \tag{2.8}$$

Substituindo $f(t)$ por $S'(t)$ na equação (2.6) tem se:

$$h(t) = -\frac{S'(t)}{S(t)}. \tag{2.9}$$

A equação (2.9) representa o simétrico da derivada de logaritmo natural (logaritmo de base euler) da função de sobrevivência em ordem a t, assim,

$$h(t) = -\frac{d}{dt} \log(S(t)). \tag{2.10}$$

Integrando ambos os membros da equação (2.10), obtém se:

$$\int_0^t h(u)du = -\int_0^t \frac{d}{du}\log(S(u))du. \tag{2.11}$$

Desenvolvendo a equação (2.11) para livrar se das integrais, tem- se:

$$-\int_0^t h(u)du = \log(S(t)) \Leftrightarrow S(t) = e^{-\int_0^t h(u)du},$$

que resulta em

$$S(t) = e^{-H(t)}. \tag{2.12}$$

Conhecidas as funções de risco ou risco acumulados, é possível desenvolver procedimentos de estimação somente a partir delas (sem necessidade de obter a função de sobrevivência), de (2.6), temos que $f(t) = h(t)S(t)$, que resulta em

$$f(t) = h(t)e^{-H(t)} \tag{2.13}$$

ou

$$f(t) = h(t)e^{-\int_0^t h(u)du}. \tag{2.14}$$

2.2.2 Variáveis aleatórias discretas (tempos discretos)

Ao considerar T como sendo variável aleatória que assume valores inteiros não negativos, é possível especificar as funções que a representam em análise de sobrevivência.

Função de Probabilidade

Para o caso em que T assume valores inteiros não negativos, a função de probabilidade, $p(.)$ é uma função que atribui uma probabilidade a cada um dos possíveis valores assumidos pela

variável, e é definida por:

$$p(t) = P(T = t), \quad t = 0, 1, 2, \ldots \quad . \tag{2.15}$$

Segundo Magalhães (2006) e Cardial, Cobre e Nakano (2024), a equação (2.15) deve satisfazer as seguintes condições:

(i) $0 \leq p(t) \leq 1, \quad t = 0, 1, 2, \ldots$ e
(ii) $\sum_{t=0}^{\infty} p(t) = 1$.

Função de Sobrevivência

A função de sobrevivência no tempo t, é a capacidade de um indivíduo sobreviver até o tempo t, e é definida como:

$$S(t) = P(T \geq t) = \sum_{k=t+1}^{\infty} P(T = k), \quad t = 0, 1, 2, \ldots \quad . \tag{2.16}$$

Função de Risco (Taxa de Falha) e Função de Risco Acumulado (Taxa de Falha Acumulada)

A taxa de falha, também conhecida como função de risco, é representada por $h(t)$ e especifica a taxa de falha instantânea de um indivíduo, ou seja, a probabilidade do mesmo vir a falhar no tempo t dado que este não falhou até esse tempo t. Quanto a variável aleatória discreta, a função de risco é igual a zero, exceto nos pontos em que pode ocorrer uma falha. A função de risco para dados discretos é definida no intervalo $0 < h(t) < 1$ e é expressa por:

$$h(t) = P(T = t | T \geq t), \quad t = 0, 1, 2, \ldots \quad . \tag{2.17}$$

A partir da função de risco (taxa de falha), pode-se definir a função de risco acumulado (taxa de falha acumulada) para variável aleatória discreta:

$$H(t) = \sum_{k=0}^{t} h(k). \tag{2.18}$$

Relações entre p(t), S(t) e h(t)

Tal como no caso em que T é contínuo é possível estabelecer relações entre as funções de probabilidade, sobrevivência e risco descrito a seguir. A função de sobrevivência pode se relacionar em relação a função de risco através da seguinte relação:

$$S(t) = \prod_{k=0}^{t} [1 - h(k)], \quad t = 0, 1, 2, ... \quad . \tag{2.19}$$

Ademais, a equação (2.20) estabelece a relação entre a função de sobrevivência e a função de risco

$$h(t) = \begin{cases} 1 - S(0), & \text{se} \quad t = 0 \\ \\ 1 - \frac{S(t)}{S(t-1)}, & \text{se} \quad t = 1, 2, ... \quad . \end{cases} \tag{2.20}$$

Também pode-se relacionar a função de probabilidade e função de sobrevivência por:

$$p(t) = \begin{cases} 1 - S(0), & \text{se} \quad t = 0 \\ \\ S(t-1) - S(t), & \text{se} \quad t = 1, 2, ... \quad . \end{cases} \tag{2.21}$$

A equação (2.21) estabelece a relação entre a função de probabilidade e de sobrevivência. Ademais, a função de probabilidade pode ser obtida a partir de $h(t)$ e $S(t)$ a partir da seguinte

relação como mostra-se a seguir:

$$h(t) = P(T = t|T \geq t) = \frac{P[(T = t), (T \geq t)]}{P(T \geq t)} = \frac{P(T = t)}{P(T \geq t)} = \frac{p(t)}{p(t) + S(t)}$$

Isolando $p(t)$ pelos os princípios de equivalência obtém-se:

$$p(t) = \frac{h(t)}{1 - h(t)} S(t). \tag{2.22}$$

As funções e relações apresentadas pelas equações (2.17), (2.18), (2.19), (2.20) e (2.21) são mais detalhadas em Tutz e Schmid (2016) e Nakano (2017).

2.3 Estimação Não Paramétrica

De acordo com Colosimo e Giolo (2006), o uso de técnicas não paramétricas são importantes para descrever os dados de sobrevivência pela sua simplicidade e facilidade de aplicação. Antes de aplicar modelos avançados, recomenda-se para uma análise inicial dos dados. Seja $t_{(j)}$ o j-ésimo tempo distinto de falha, isto é, $t_{(1)} < t_{(2)} < ... < t_{(m)}$. Assim, o estimador não paramétrico da função de sobrevivência de Kaplan-Meier (Kaplan e Meier, 1958) é definido por:

$$\widehat{S}_{KM}(t) = \prod_{j:t_{(j)} \leq t} \left[1 - \frac{d_j}{n_j}\right], \tag{2.23}$$

em que, d_j é o número de indivíduos que experimentaram o evento de interesse no tempo $t_{(j)}$ e n_j é o número de indivíduos que estão sob o risco no tempo $t_{(j)}, j = 1, 2, \ldots, m$. O estimador de Kaplan-Meier é considerado uma das ferramentas mais utilizadas para estimar de forma não paramétrica a função de sobrevivência na presença de censuras e será adotado no presente trabalho.

2.4 Estimação Paramétrica

O processo de estimação segundo a abordagem clássica pode ser realizado a partir do método de máxima verossimilhança, método dos momentos ou por método dos mínimos quadrados, sendo que estes dois últimos não podem ser adotados na presença de censura. Neste trabalho será apresentado o método de máxima verossimilhança para obtenção das estimativas dos parâmetros na presença de censura, assim como os procedimentos para estimação intervalar e teste de hipóteses.

2.4.1 Método de Máxima Verossimilhança

Segundo Migon, Gamerman e Louzada (2014), o método de máxima verossimilhança é atualmente o método de estimação mais utilizado na inferência clássica. Seja $\mathbf{t} = (t_1, t_2, \ldots, t_n)^T$ uma amostra aleatória observada de uma variável aleatória T; $f(t_i, \boldsymbol{\phi})$ representa a função densidade de probabilidade (ou função de probabilidade se T for discreta) e $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_h)^T$. A função de verossimilhança para $\boldsymbol{\phi}$ na ausência de censura é dada por:

$$L(\boldsymbol{\phi}; t) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\phi}). \tag{2.24}$$

No entanto, na presença de censuras (principal caraterística em análise de sobrevivência), os dados censurados devem ser separados dos elementos que registaram o evento de interesse. Dessa forma, as observações são reordenadas e divididas em dois subconjuntos, as r primeiras observações são as não censuradas, sendo a sua contribuição para a função de verossimilhança dada por $f(t_i, \phi)$ e as $n - r$ observações seguintes são as censuradas a direita, sendo sua contribuição para função de verossimilhança dada pela função de sobrevivência, $S(t_i, \phi)$. Assim, a função de verossimilhança ganha a seguinte forma:

$$L(\boldsymbol{\phi}; \mathbf{t}) \propto \prod_{i=1}^{r} f(t_i; \boldsymbol{\phi}) \prod_{i=r+1}^{n} S(t_i; \boldsymbol{\phi}). \tag{2.25}$$

Introduzindo a variável indicadora de falha δ_i, que assume valor 1 se o tempo t_i for de falha e 0 se for censura à direita na função de verossimilhança, a expressão (2.25) assume a seguinte forma:

$$L\left(\boldsymbol{\phi};\boldsymbol{t},\boldsymbol{\delta}\right) \propto \prod_{i=1}^{n}\left[f(t_i;\boldsymbol{\phi})\right]^{\delta_i}\left[S(t_i;\boldsymbol{\phi})\right]^{1-\delta_i}, \tag{2.26}$$

em que $f\left(\cdot;\boldsymbol{\phi}\right)$ e $S\left(\cdot;\boldsymbol{\phi}\right)$ são, respectivamente, a função densidade de probabilidade (se T é contínua) ou função de probabilidade (se T é discreta) e função de sobrevivência da distribuição adotada. Ao aplicar o logaritmo ambos os membros na função de verossimilhança na equação (2.26), obtém-se a seguinte equação:

$$\ell\left(\boldsymbol{\phi};\boldsymbol{t},\boldsymbol{\delta}\right) = \sum_{i=1}^{n}\left[\delta_i \log\left[f(t_i;\boldsymbol{\phi})\right] + (1-\delta_i)\log\left[S(t_i;\boldsymbol{\phi})\right]\right] + \xi, \tag{2.27}$$

em que ξ é uma constante que não depende de $\boldsymbol{\phi}$.

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\phi}$ que maximizam $L\left(\boldsymbol{\phi},\boldsymbol{t},\boldsymbol{\delta}\right)$ ou equivalentemente que maximizam a função $\ell\left(\boldsymbol{\phi},\boldsymbol{t},\boldsymbol{\delta}\right)$, geralmente representados por $\hat{\boldsymbol{\phi}}$ e são obtidos resolvendo o sistema de equações:

$$\frac{\partial \ell\left(\boldsymbol{\phi},\boldsymbol{t},\boldsymbol{\delta}\right)}{\partial \boldsymbol{\phi}} = \mathbf{0}_{\boldsymbol{h}\times\mathbf{1}}. \tag{2.28}$$

2.4.2 Intervalo de Confiança

Os intervalos de confiança fornecem um método para adicionar mais informações a um estimador $\widehat{\boldsymbol{\phi}}$, quando deseja-se estimar um parâmetro desconhecido $\boldsymbol{\phi}$ (Cardial, Cobre e Nakano, 2024). Assim, pode-se encontrar um intervalo em que almeja-se ter alta probabilidade de conter $\boldsymbol{\phi}$.

A obtenção da matriz de variância e covariância de $\widehat{\boldsymbol{\phi}}$, $\widehat{\boldsymbol{\phi}} = (\widehat{\phi_1}, \widehat{\phi_2}, \ldots, \widehat{\phi_h})^T$, é dada por:

$$I_f(\widehat{\boldsymbol{\phi}}) = -\left.\frac{\partial^2 \ell\left(\boldsymbol{\phi},\boldsymbol{t}|\boldsymbol{\delta}\right)}{\partial \boldsymbol{\phi}^2}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}. \tag{2.29}$$

Sob certas condições de regularidade, pode-se obter a variância que é assintoticamente igual inversa da matriz de informação de Fisher observada:

$$\widehat{Var}\left(\widehat{\boldsymbol{\phi}}\right) \approx \left[I_f(\widehat{\boldsymbol{\phi}})\right]^{-1}, \tag{2.30}$$

em que $I_f(\cdot)$ é a informação de Fisher observada da amostra dada pela equação (2.29).

Sob certas condições de regularidades, tem-se que:

$$\widehat{\boldsymbol{\phi}} \overset{a}{\sim} N_h\left(\boldsymbol{\phi}, \widehat{Var}(\widehat{\boldsymbol{\phi}})\right), \tag{2.31}$$

isto é, $\widehat{\boldsymbol{\phi}}$ converge assintoticamente para uma distribuição normal multivariada com média $\boldsymbol{\phi}$ e matriz de variância e covariância $\widehat{Var}(\widehat{\boldsymbol{\phi}})$, em que h é a dimensão do vetor de parâmetro $\widehat{\boldsymbol{\phi}}$.

Assim, um intervalo aproximado de $(1-\alpha)\times 100\%$ de confiança para o parâmetro ϕ_j é dado por:

$$\left[\widehat{\phi}_j - Z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\widehat{\phi}_j)}; \widehat{\phi}_j + Z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\widehat{\phi}_j)}\right], \quad j = 1, 2, \ldots, h, \tag{2.32}$$

em que $\widehat{\phi}_j$ é o estimador de máxima verossimilhança de ϕ_j, $Z_{(1-\alpha/2)}$ é o quantil $1-\alpha/2$ de uma distribuição normal padrão e $\widehat{Var}(\widehat{\phi}_j)$ é a estimativa da variância do estimador $\widehat{\phi}_j$, dado por (2.30).

Ademais, um intervalo de confiança para uma função real $\pi(\boldsymbol{\phi})$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_h)^T$, pode ser obtido por meio do método delta multivariado descrito a seguir.

A estimativa da função real $\widehat{\pi(\boldsymbol{\phi})}$ segue assintoticamente uma distribuição normal com média $\pi(\boldsymbol{\phi})$ e variância $\widehat{Var}\left[\pi(\widehat{\boldsymbol{\phi}})\right]$, isto é,

$$\widehat{\pi(\boldsymbol{\phi})} \overset{a}{\sim} N\left(\pi(\boldsymbol{\phi}), \widehat{Var}\left[\pi(\widehat{\boldsymbol{\phi}})\right]\right), \tag{2.33}$$

em que,

$$\widehat{Var}\left[\pi(\widehat{\boldsymbol{\phi}})\right] = \sum_{s=1}^{h}\sum_{j=1}^{h}\widehat{Cov}\left(\widehat{\phi}_s;\widehat{\phi}_j\right)\left(\left.\frac{\partial\pi(\boldsymbol{\phi})}{\partial\phi_s}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}\right)\times\left(\left.\frac{\partial\pi(\boldsymbol{\phi})}{\partial\phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}}\right).$$

A expressão (2.33) é utilizada quando há o interesse em construir intervalos de confiança para funções dos parâmetros do modelo.

2.4.3 Teste de Hipóteses

Teste da Razão de verossimilhanças

Para um modelo com um vetor de parâmetros $\boldsymbol{\Phi} = (\phi_1, \phi_2, ..., \phi_h)^T$, em algumas situações, deseja-se testar hipóteses relacionadas a este vetor ou a um subvetor dele. O teste da razão de verossimilhanças (TRV) é baseado na função de verossimilhança, e envolve a comparação dos valores do logaritmo da função de verossimilhança maximizada e sob a hipótese nula $H_0 : \boldsymbol{\phi} \in \boldsymbol{\Phi}_0$ vs $H_1 : \boldsymbol{\phi} \in \boldsymbol{\Phi}_1$ com $\boldsymbol{\Phi} = \boldsymbol{\Phi}_0 \cup \boldsymbol{\Phi}_1$ e $\boldsymbol{\Phi}_0 \cap \boldsymbol{\Phi}_1 = \oslash$, (Rohatgi e Saleh, 2015). Neste caso a estatística de teste é dada pela seguinte expressão:

$$TRV = -2\log\left[\frac{L(\boldsymbol{\phi_0};\boldsymbol{t},\boldsymbol{\delta})}{L(\widehat{\boldsymbol{\phi}};\boldsymbol{t},\boldsymbol{\delta})}\right]. \tag{2.34}$$

Aplicando a propriedade de quociente do logaritmo tem-se:

$$TRV = 2\left[\log L(\widehat{\boldsymbol{\phi}};t,\delta) - \log L(\boldsymbol{\phi_0};\boldsymbol{t},\delta)\right], \tag{2.35}$$

que sob H_0, segue aproximadamente uma distribuição qui-quadrado com ρ graus de liberdade. Aqui, ρ representa a diferença do número de parâmetros dos modelos (modelo completo e modelo sob H_0). Para amostras suficientemente grandes, H_0 é rejeitada, a um nível de significância de $\alpha\times100\%$ se $TRV > \chi^2_{\rho,1-\alpha}$. Aqui, $\chi^2_{\rho,1-\alpha}$ é o quantil $1-\alpha$ de uma distribuição qui-quadrado com ρ graus de liberdade.

Alternativamente, o teste pode ser realizado por meio do nível descritivo (valor-p) do teste, que neste caso é definido por:

$$valor - p = P\left(\chi^2_\rho > TRV\right), \tag{2.36}$$

em que χ^2_ρ denota uma variável aleatória com distribuição qui-quadrado com ρ graus de liberdade. Neste caso, a H_0 será rejeitada à favor de H_1 se valor-p < α.

Teste de Wald

O teste de Wald sugere quais variáveis do modelo estão contribuindo com algo significativo. Para um modelo com um vetor de parâmetros $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_h)^T$, em algumas situações, deseja-se testar hipóteses relacionadas a um único parâmetro, isto é, H_0: $\phi_j = \phi_{j0}$ vs H_a: $\phi_j \neq \phi_{j0}$, $j = 1, 2, \ldots, h$.

Neste caso a estatística de teste sob a ideia da significância $\phi_{j0} = 0$ é dada pela seguinte expressão:

$$W = \frac{(\widehat{\phi_j} - \phi_{j0})^2}{Var(\widehat{\phi_j})}, \tag{2.37}$$

que sob H_0 segue aproximadamente uma distribuição qui-quadrado com 1 grau de liberdade. Para amostras suficientemente grandes, H_0 é rejeitada, a um nível de significância de $\alpha \times 100\%$ se $W > \chi^2_{1,1-\alpha}$. Aqui, $\chi^2_{1,1-\alpha}$ é o quantil ao nível de $1 - \alpha$ de confiança de uma distribuição qui-quadrado com 1 grau de liberdade.

Alternativamente, o teste pode ser realizado por meio do nível descritivo (valor-p) do teste, que neste caso é definido por:

$$valor - p = P\left(\chi^2_1 > W\right), \tag{2.38}$$

em que χ^2_1 denota uma variável aleatória com distribuição qui-quadrado com 1 grau de liber-

dade. Neste caso, a H_0 será rejeitada à favor de H_1 se valor-$p < \alpha$.

Capítulo 3

Revisão de literatura

Modelos de regressão têm sido amplamente utilizados para modelar dados de sobrevivência. Um dos modelos mais empregado é o de Cox (1972), que foi primeiro modelo proposto para modelar dados de sobrevivência na presença de covariáveis (Parreira, 2007). A seguir apresentamos a evolução dos modelos de regressão a partir do modelo de riscos proporcionais, assim como outros modelos que podem ser usados para modelar os dados de sobrevivência como os modelos de chances de sobrevivência proporcionais. Ademais, é apresentada uma breve discussão sobre dados de sobrevivência discretos, assim como procedimentos para obtenção de distribuições discretas análogas às distribuições de variáveis aleatórias contínuas.

3.1 Alguns modelos de regressão em Análise de Sobrevivência

3.1.1 O Modelo de Riscos Proporcionais

No modelo de Riscos Proporcionais (RP), considera-se que a função de risco pode ser apresentada em dois fatores, sendo um representa o efeito das covariáveis e outro do tempo (contínuo). Seja $h(t|\mathbf{Z})$ a função de risco no tempo t e para um individuo com o vetor de covariáveis $\mathbf{Z}$, o modelo de risco proporcionais proposto por Cox (1972) é dado por:

$$h(t|\mathbf{Z}) = g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta})h_0(t), \tag{3.1}$$

em que $g(.)$ é uma função de ligação positiva e igual a 1 se o seu argumento for igual a zero, $h_0(t)$ é a função de risco base e $\boldsymbol{\beta}$ representa o vetor de coeficientes de regressão associados a $\mathbf{Z}$. Na teoria são várias formas de funções que podem ser empregadas para função $g(.)$, mas geralmente tem se recorrido a função $\exp(.)$ para $g(.)$, portanto, ao longo deste trabalho será adotada a mesma.

Parreira (2007) afirma que o modelo proposto por Cox (1972) assume que o vetor de covariáveis $\mathbf{Z}$ tem um efeito multiplicativo (um múltiplo) na função de risco. Isto implica que na sua estrutura impõe proporcionalidade entre funções de riscos de diferentes níveis de covariáveis, não permitindo que elas se interceptem e nem dependam do tempo t.

O modelo de riscos proporcionais proposto por Cox (1972) tem uma grande limitação que é suposição de proporcionalidades entre funções de riscos em diferentes níveis de covariável, uma vez que na prática não tem sido tão fácil encontrar um conjunto de dados que atendem esta suposição.

3.1.2 O Modelo de Falha Acelerada

Em situações em que o modelo de Cox não é adequado, tem-se o Modelo de Falha Acelerada (FA) como alternativa. Proposto por Kalbfleisch e Prentice (1980), o modelo FA é dado por:

$$h(t|\mathbf{z}) = g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta})h_0\left(g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta})t\right). \tag{3.2}$$

O modelo FA tem uma larga vantagem em relação ao modelo de riscos proporcionais de Cox, visto que ele permite modelar dados com cruzamento de funções de riscos. Pois, neste modelo, o vetor uma de covariável $\mathbf{Z}$ tem um efeito multiplicativo não apenas na função de risco, mas também no tempo t. Desta forma, a covariável $\mathbf{Z}$ afeta o tempo de sobrevivência

causando deformações na escala do tempo. Louzada-Neto (1997) afirma que os modelos de riscos proporcionais de RP e FA compreendem famílias diferentes em termos interpretativos, razão pelo qual devem ser tratados de forma diferentes.

3.1.3 O Modelo Híbrido

Com o propósito de acomodar ambos os modelos (de RP e FA) em uma família, Ciampi e Etezadi-Amoli (1985) propuseram um modelo híbrido (RP/FA) dado por:

$$h(t|\mathbf{z}) = g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_1})h_0\left(g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})t\right), \tag{3.3}$$

em que $\boldsymbol{\beta_1}$ e $\boldsymbol{\beta_2}$ são vetores de coeficientes de regressão associados a $\mathbf{Z}$. Para $\boldsymbol{\beta_2} = \mathbf{0}$, obtém-se o modelo de riscos proporcionais (RP) e para $\boldsymbol{\beta_1} = \boldsymbol{\beta_2}$ tem-se o modelo de falha acelerada (FA).

As relações entre os modelos (3.1), (3.2) e (3.3) são apresentadas na Tabela 3.1 (Parreira, 2007):

Tabela 3.1: Relações entre os modelos RP, FA e Híbrido.

Modelo	**Restrição**	**Risco**	**Sobrevivência**
RP/FA	—	$g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_1})h_0\left(g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})t\right)$	$\left[S_0(g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})t)\right]^{\frac{g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_1})}{g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})}}$
FA	$\boldsymbol{\beta_1} = \boldsymbol{\beta_2}$	$g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})h_0\left(g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})t\right)$	$S_0(g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_2})t)$
RP	$\boldsymbol{\beta_2} = \mathbf{0}$	$g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_1})h_0(t)$	$[S_0(t)]^{g(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\beta_1})}$

Fonte: Elaborado pelo autor

3.1.4 O Modelo de Chances de Sobrevivência Proporcionais

Cardial, Cobre e Nakano (2024) afirma que um concorrente do modelo de riscos proporcionais de Cox é o Modelo de Chances Proporcionais (MCP) proposto por Bennett (1983), que

introduziu a estrutura do modelo semi-paramétrico em um contexto de análise de sobrevivência utilizando como distribuição base a distribuição Log-logística. Murphy, Rossini e Vaart (1997) demonstraram as propriedades deste modelo, entre elas que o estimador do coeficiente de regressão é assintoticamente normal com variância eficiente. A inferência semi-paramétrica foi desenvolvida por Yang e Prentice (1999) e aplicações em tempos de sobrevivência contínuos são apresentados em Royston e Parmar (2002), Wang e Wang (2022) e Zhou, Zhang e Lu (2022). Segundo Cardial, Cobre e Nakano (2024), neste modelo as covariáveis agem multiplicativamente na chance (odds) de sobrevivência. Dessa maneira, esse modelo é denominado por Modelo de Chances de Sobrevivência Proporcionais (MCSP). Ao longo dos anos esse modelo, que tem em sua proposta inicial a sua utilização como modelo de regressão, é utilizado em uma escala menor quando comparado ao modelo de riscos proporcionais de Cox.

Na presença de um vetor de covariáveis $\mathbf{Z}$ e vetor de coeficientes de regressão $\boldsymbol{\beta}$. O MCSP é dado por:

$$\frac{S(t|\mathbf{Z})}{1-S(t|\mathbf{Z})} = g(\boldsymbol{Z^T\beta})\frac{S_0(t)}{1-S_0(t)}, \tag{3.4}$$

em que $S(\boldsymbol{t}|\boldsymbol{Z})$ é a função de sobrevivência no tempo t na presença da covariável $\mathbf{Z}$, $S_0(.)$ é a função de sobrevivência base e $g(.)$ é uma função de ligação não negativa que é igual a 1 quando o seu argumento é nulo.

A partir de (3.4) obtemos a função de sobrevivência do MCSP, que é dada por:

$$S(t|\mathbf{Z}) = \frac{g(\boldsymbol{Z^T\beta})S_0(t)}{1+[g(\boldsymbol{Z^T\beta})-1]S_0(t)}. \tag{3.5}$$

Bennett (1983) traça um paralelo do MCSP ao modelo de riscos proporcionais, comparando a razão das funções de risco para o caso contínuo. O autor afirma que, no modelo de Cox, as taxas de risco para grupos separados de pacientes têm uma relação de proporcionalidade entre si, enquanto no MCSP estas taxas de risco convergem com o tempo. No MCSP a função base é

conhecida, por exemplo se escolhermos a função de sobrevivência base ou risco base da distribuição Weibull, o modelo passa ser chamado de MCSP-Weibull; Se a função de sobrevivência ou risco base forem da distribuição Log-normal, o modelo é chamado MCSP-Log-normal e se função de sobrevivência ou risco base forem da distribuição Log-logística, o modelo passa ser chamado de MCSP-Log-logística.

Ao contrário do modelo de RP que assumem tempos contínuos, o MCSP também podem ser obtidos a partir de distribuições discretas. Quando os tempos são contínuos é possível ter, dentre outros, os seguintes MCSP:

- **Modelo de Chances de Sobrevivências Proporcionais Weibull Contínuo (MCSP-WC)** - Quando a distribuição base é a distribuição Weibull Contínua;
- **Modelo de Chances de Sobrevivências Proporcionais Log-Normal Contínuo (MCSP-LNC)** - Quando a distribuição base é a distribuição Log-Normal contínuo;
- **Modelo de Chances de Sobrevivências Proporcionais Log-Logística Contínuo (MCSP-LLC)** - Quando a distribuição base é a distribuição Log-Logística contínuo;

E quando os tempos são discretos é possível ter os seguintes MCSP:

- **Modelo de Chances de Sobrevivências Proporcionais Weibull Discreto (MCSP-WD)** - Quando a distribuição base é a distribuição Weibull discreta. (Cardial, Cobre e Nakano, 2024);
- **Modelo de Chances de Sobrevivências Proporcionais Log-Normal Discreto (MCSP-LND)** - Quando a distribuição base é a distribuição Log-Normal discreto;
- **Modelo de Chances de Sobrevivências Proporcionais Log-Logística Discreto (MCSP-LLD)** - Quando a distribuição base é a distribuição Log-Logística discreto.

3.2 Breve discussão sobre tempos discretos

Conforme Berguer e Schmid (2018), na maioria das vezes, é assumido nessas análises que o tempo de sobrevivência é dado por uma variável aleatória medida em uma escala contínua, havendo uma extensiva literatura para esse caso. No entanto, na prática, as medições de tempo costumam ser discretas, havendo situações em que a hora exata do evento pode não ser conhecida, mas apenas o intervalo durante o qual ocorreu o evento de interesse.

De uma forma geral, Tutz e Schmid (2016) afirmam que o tempo até ocorrência do evento de interesse discreto ocorre como:

- Medições intrinsecamente discretas;
- Dados agrupados.

Os dados agrupados representam eventos em intervalos de tempo subjacentes, e a variável resposta refere-se a um intervalo, que pode ter tamanhos iguais ou diferentes. Há exemplos utilizados em trabalhos como:

- Brunello e Nakano (2015) conduziram um estudo sobre o tempo em meses até a morte de homens diagnosticados com Síndrome de Imunodeficiência Adquirida;
- Cardial, Fachini-Gomes e Nakano (2020) realizaram uma pesquisa do estudo sobre o tempo em meses de pacientes com câncer de cabeça e pescoço.

3.2.1 Vantagens de Uso de Tempos Discretos em relação aos contínuos

De acordo com Tutz e Schmid (2016), métodos estatísticos desenvolvidos para tempo até ocorrência do evento de interesse discreto utilizados para esse fim possuem uma série de vantagens:

- A consideração de modelos para tempos discretos tem a vantagem de que os riscos podem ser formulados como probabilidades condicionais. Portanto, eles são muito mais acessíveis para interpretação do que funções de risco contínuos;

- Na prática, muitos tempos de evento são intrinsecamente discretos ou são observados em um escala discreta. Consequentemente, usar modelos para tempos discretos é mais apropriado do que a aproximação dos dados observados por um modelo de sobrevivência contínuo;

- Em contraste com os modelos de sobrevivência para tempo contínuo, modelos para eventos discretos não causam problemas com empates (Se duas observações distintas têm o mesmo valor, recebendo assim a mesma classificação, são consideradas empatadas);

- Modelos para tempos discretos podem ser incorporados a estrutura de um Modelo Linear Generalizado (MLG). Consequentemente, a estimativa é facilmente obtida usando um software padrão para a estimativa de MLG's (em caso de ausência de censuras);

- Incorporação na estrutura de MLG's permite usar a metodologia também para modelos avançados. Por exemplo, ao incluir parâmetros subjetivos específicos nos chamados modelos de fragilidade.

Nakano e Carrasco (2006), utilizando tempos discretos aplicados ao modelo exponencial e ao análogo discreto deste modelo (modelo geométrico) mostraram haver um melhor ajuste dos dados ao utilizar o modelo discreto. O mesmo raciocínio foi seguido para simulações computacionais, observando aspectos como variabilidade dos dados, tamanho da amostra e percentual de censura, constatando-se que nem sempre é aceitável a utilização de um modelo contínuo para a análise de dados discretos.

3.2.2 Discretização de Tempos Contínuos

Cardial, Cobre e Nakano (2024) afirma que, em diversas aplicações, variáveis originais podem ser contínuas por natureza, mas discretas por observação e, portanto, é razoável e conveniente modelar a situação por uma distribuição discreta apropriada gerada a partir dos modelos

contínuos subjacentes preservando um ou mais traços (caraterísticas) importantes da distribuição contínua.

Chakraborty (2015) apresenta um levantamento completo dos métodos e construções para gerar análogos discretos de distribuições de probabilidade contínuas. Por outro lado a dupla Jayakumar e Babu (2018) apresentaram diferentes metodologias de discretização de distribuições contínuas, a saber:

- Discretizar a função de distribuição acumulada contínua;
- Discretizar a função densidade de probabilidade contínua;
- Discretizar a função de risco contínua;
- Obter distribuição discreta de tempo de vida a partir da taxa de falha alternativa.

A seguir explica-se como se processa a primeira metodologia.

3.2.3 Discretização a Partir da Função de Distribuição Acumulada Contínua

Considere X uma variável aleatória contínua que assume valores não negativos. A variável aleatória discreta é dada por $T = \lfloor X \rfloor$, em que $\lfloor X \rfloor$ denota "a parte inteira de X", isto é, T denota o maior inteiro menor ou igual a X.

Se $F_X(.)$ é a função de distribuição acumulada de X, a distribuição de probabilidade de T, representada por $p(t)$, pode ser escrita como:

$$p(t) = P(T = t) = P(t \leq X < t + 1) = F_X(t + 1) - F_X(t), \quad t = 0, 1, 2, ... \quad . \tag{3.6}$$

Nesse contexto, algumas publicações em análise de sobrevivência lidam com a utilização dessa metodologia para determinar o análogo discreto das distribuições contínuas e consequentemente obter novas distribuições de probabilidade discretas ou atuar com os análogos discretos provenientes da referida metodologia em aplicações como:

- Nakagawa e Osaki (1975) que obtiveram a distribuição Weibull discreta (WD);
- Jayakumar e Babu (2018) que obtiveram a distribuição Weibull Geométrica discreta;
- Vieira et al. (2023) que atua com a distribuição Log-logística discreta;
- Biazatti e Nakano (2020) que atuam com a distribuição Log-Normal discreta;
- Cardial, Fachini-Gomes e Nakano (2020), que atuam com a distribuição Weibull discreta exponenciada (WDE);
- Sarhan (2017) que obteve a distribuição banheira de dois parâmetros discreta (DTPBT) e
- Chakraborty (2015) que obteve a distribuição Gumbel discreta.

3.2.4 Distribuição Weibull Discreta

A distribuição Weibull é o modelo probabilístico mais aceito e utilizado na modelagem de dados de sobrevivência devido à flexibilidade que a mesma apresenta. Essa versatilidade se deve aos dois parâmetros que a distribuição apresenta: o parâmetro de escala e o parâmetro de forma, que proporcionam uma variedade de formas e devido a sua função taxa de falha ser monótona, ou seja, ela é crescente, decrescente ou constante. Por ter essas vantagens, o caso discreto dessa distribuição vem sendo estudado ao longo dos anos, e há, conforme Vila, Nakano e Saulo (2018), pelo menos três versões conhecidas da distribuição Weibull discreta (WD):

- Distribuição Weibull discreta de tipo I , que discretiza a função de distribuição acumulada contínua (Nakagawa e Osaki, 1975);
- Distribuição Weibull discreta de tipo II, que discretiza a função de risco contínua (Stein e Dattero, 1984); e
- Distribuição Weibull discreta de tipo III, que obtém a distribuição discreta de tempo de vida a partir a taxa de falha alternativa (Padgett e Spurrier, 1985).

Neste trabalho será considerada a distribuição Weibull discreta do tipo I que tem os parâmetros $\theta > 0$ e $0 < b < 1$ com $t = 0, 1, 2, \ldots$ cuja função de probabilidade, de sobrevivência e risco são, respectivamente:

$$p(t|b, \theta) = b^{t^\theta} - b^{(t+1)^\theta}, \tag{3.7}$$

$$S(t|b, \theta) = b^{(t+1)^\theta} \text{e} \tag{3.8}$$

$$h(t|b, \theta) = \frac{b^{t^\theta} - b^{(t+1)^\theta}}{b^{t^\theta}}. \tag{3.9}$$

Com relação a função de risco (3.9) da distribuição Weibull discreta do tipo I, temos que:

- Se $\theta > 1$, então $h(t)$ é estritamente crescente;
- Se $\theta = 1$, então $h(t)$ é constante (a distribuição WD se reduz a distribuição Geométrica);
- Se $0 < \theta < 1$, então $h(t)$ é estritamente decrescente;

A Figura [3.1] apresenta o comportamento da função risco da distribuição WD.

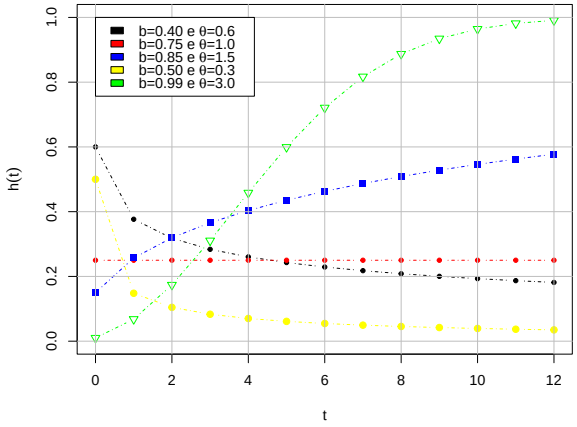


Figura 3.1: Riscos da distribuição WD para diferentes valores dos parâmetros.
Fonte: Elaborado pelo autor

A seguir apresenta-se as distribuições log-Normal e log-Logística discretizadas sem os passos intermediários para sua discretização a partir das suas respectivas distribuições contínuas.

3.2.5 Distribuição Log-Normal Discreta

Seguindo os passos análogos e aplicando a equação (3.6), que foi aplicada para discretizar a distribuição Weibull contínuo, obtém-se a distribuição Log-Normal discreto (LND) caracterizada pelos parâmetros μ e σ^2. Sendo $\mu \in \mathbb{R}$ o parâmetro de locação e $\sigma^2 > 0$ de escala com $t = 0, 1, 2, \ldots$ cuja funções de probabilidade, sobrevivência e risco são, respetivamente:

$$p(t|\mu,\sigma) = \Phi\left(\frac{\log(t+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\log t-\mu}{\sigma}\right), \tag{3.10}$$

$$S(t|\mu,\sigma) = 1 - \Phi\left(\frac{\log(t+1)-\mu}{\sigma}\right) \quad \text{e} \tag{3.11}$$

$$h(t|\mu,\sigma) = \frac{\Phi\left(\frac{\log(t+1)-\mu}{\sigma}\right) - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)}{1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right)}, \tag{3.12}$$

em que $\Phi(.)$ é a função de distribuição acumulada da normal padrão.

A Figura 3.2 apresenta o comportamento da função de risco da distribuição Log-Normal discreta para diversos valores de μ e σ.

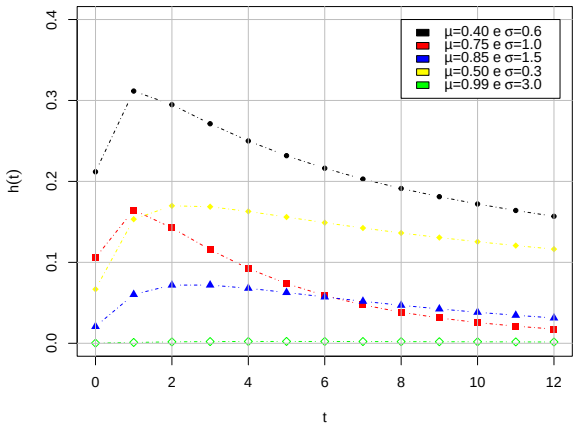


Figura 3.2: Riscos da distribuição LND para diferentes valores dos parâmetros.
Fonte: Elaborado pelo autor

Uma característica importante da distribuição Log-Normal na modelagem de tempos de sobrevivência está relacionada ao fato dela permitir acomodar funções de risco unimodais, o que pode ser adequado em algumas situações práticas.

3.2.6 Distribuição Log-Logística Discreta

A distribuição Log-Logística discreta (LLD) é caracterizada pelos parâmetros μ e β, sendo $\mu > 0$ e $\beta > 0$ com $t = 0, 1, 2, \ldots$ e apresenta as seguintes funções de probabilidade, sobrevivência e de risco, respectivamente:

$$p(t|\mu,\beta) = \frac{1}{1+\left(\frac{t}{\mu}\right)^{\beta}} - \frac{1}{1+\left(\frac{t+1}{\mu}\right)^{\beta}}, \tag{3.13}$$

$$S(t|\mu,\beta) = \frac{1}{1+\left(\frac{t+1}{\mu}\right)^{\beta}} \quad e \tag{3.14}$$

$$h(t|\mu,\beta) = \frac{\frac{1}{1+\left(\frac{t}{\mu}\right)^{\beta}} - \frac{1}{1+\left(\frac{t+1}{\mu}\right)^{\beta}}}{\frac{1}{1+\left(\frac{t}{\mu}\right)^{\beta}}}, \quad t = 0, 1, 2, \ldots. \tag{3.15}$$

A Figura 3.3 apresenta o comportamento da função de risco da distribuição Log-Logística discreta para diversos valores de μ e β

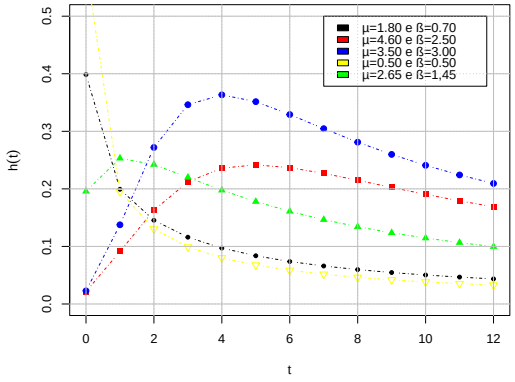


Figura 3.3: Riscos da distribuição LLD para diferentes valores dos parâmetros.
Fonte: Elaborado pelo autor

Assim como a Log-Normal, a distribuição Log-Logística também permite acomodar funções de risco unimodais, as Figuras 3.2 e 3.3 ilustram este fato.

Capítulo 4

Modelo de Regressão Log de Sobrevivências Proporcionais Para Dados Discretos

É apresentado nesse Capítulo o modelo de regressão log de sobrevivências proporcionais para dados discretos (MRLSPD). A formulação do modelo e as suas propriedades são apresentadas na Seção 4.1. A verificação da suposição do modelo é apresentada na Seção 4.2. O procedimento para estimação dos parâmetros (pontual e intervalar) são apresentados nas Seções 4.3 e 4.4, respectivamente. E na Seção 4.5 é apresentada uma abordagem geral dos testes de hipóteses.

4.1 Formulação do Modelo

O modelo de regressão log de sobrevivências proporcionais discreto (MRLSPD) proposto neste trabalho assume que a covariável $\boldsymbol{Z} = (z_1, z_2, ..., z_p)^T$ age multiplicativamente no loga-

ritmo da função de sobrevivência de uma variável aleatória discreta T, ou seja, assume que:

$$\log[S(t|\mathbf{Z})] = g(\boldsymbol{Z}^T\boldsymbol{\beta}) \log[S_0(t)]. \tag{4.1}$$

em que $S(.)$ é a função de sobrevivência de um indivíduo com vetor de covariável $\mathbf{Z}$, $g(.)$ uma função de ligação positiva que assume o valor 1 quando o seu argumento for igual a 0 e $S_0(t)$ é função de sobrevivência base (isto é, a função de sobrevivência de um indivíduo quando todas as covariáveis são nulas).

A partir da equação (4.1), tem-se que:

$$S(t|\mathbf{Z}) = [S_0(t)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})}. \tag{4.2}$$

Note que o intercepto β_0 não aparece no preditor linear $\boldsymbol{Z}^T\boldsymbol{\beta}$. Isso porque a função de sobrevivência base, $S_0(t)$, absorve este termo constante. Note também que, no caso em que T é uma variável aleatória contínua, o modelo de regressão log de sobrevivências proporcionais (4.2) é equivalente ao modelo de Riscos Proporcionais.

A partir das equação (2.20) deduz se a função de risco do MRLSPD que é expressa por:

$$h(t|\mathbf{Z}) = \begin{cases} 1 - [S_0(0)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})}, & \text{se} \quad t = 0 \\ \\ 1 - \frac{[S_0(t)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})}}{[S_0(t-1)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})}}, & \text{se} \quad t = 1, 2, ... \quad . \end{cases} \tag{4.3}$$

Da equação (2.21) obtém se a função de probabilidade do MRLSPD que é dada por:

$$p(t|\mathbf{Z}) = \begin{cases} 1 - [S_0(0)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})}, & \text{se} \quad t = 0 \\ \\ [S_0(t-1)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})} - [S_0(t)]^{g(\boldsymbol{Z}^T\boldsymbol{\beta})}, & \text{se} \quad t = 1, 2, \ldots \quad . \end{cases} \tag{4.4}$$

4.1.1 Comportamento da função de risco

Nesta seção apresenta-se o comportamento da função de risco do modelo, assim como a razão entre a função de risco e risco base. Essa ilustração será realizada considerando a distribuição Weibull discreta (apresentada na Seção 3.2.4 e considerando a função de ligação $g(.)$ como um parâmetro adicional da distribuição Weibull discreta.

As funções de risco e probabilidade do Modelo de Regressão Log de Sobrevivências Proporcional Weibull discreta (MRLSP-WD) podem ganhar outras formas equivalentes se substituirmos a função de ligação $g(\boldsymbol{Z}^T\boldsymbol{\beta})$ por uma constante λ, isto é, $g(\boldsymbol{Z}^T\boldsymbol{\beta}) = \lambda$. Assim, a partir de (3.8), (4.2), (4.3) e (4.4) e para t inteiro não negativo, temos a função de risco e a função de probabilidade do MRLSP-WD são dadas, respectivamente, por:

$$h(t|\mathbf{Z}) = 1 - b^{\lambda[(t+1)^\theta - t^\theta]} \quad \text{e} \tag{4.5}$$

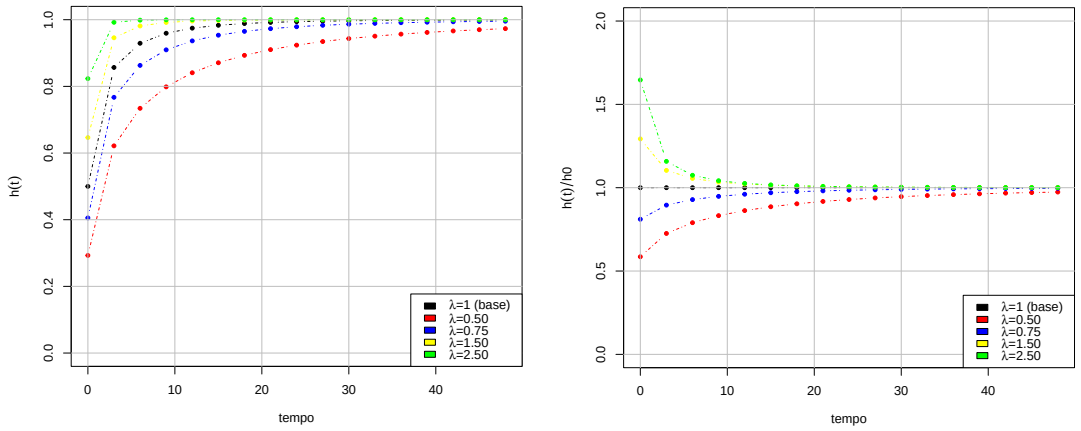
$$p(t|\mathbf{Z}) = b^{\lambda t^\theta} - b^{\lambda(t+1)^\theta}. \tag{4.6}$$

É interessante destacar que, se assumir se que a função base é da distribuição Weibul Discreta (WD) de risco constante quando $\theta = 1$, o MRLSP-WD também apresenta função de risco constante. De fato, neste caso o MRLSP-WD se reduz ao Modelo de Regressão Log de Sobrevivências Proporcionais Geométrico (MRLSP-G).

Para risco base crescente

As Figuras (4.1)(a) e (4.1)(b) apresentam, respectivamente, as funções de riscos e as razões de riscos para o MRLSP-WD com $\theta = 1.5$, $b = 0.50$ e diversos valores de λ: $\lambda = 1.0$, $\lambda = 0.5$, $\lambda = 0.75$, $\lambda = 1.50$ e $\lambda = 2.5$.

A função de risco do MRLSP-WD de base crescente é monótona crescente, para qualquer valor de λ e tende para 1 quando o tempo tende para infinito. Consequentemente, a razão entre riscos convergem para 1, como mostra a Figura 4.1(b). Segundo Murphy, Rossini e Vaart



(a) Riscos do MRLSP-WD

(b) Razão de riscos MRLSP-WD

Figura 4.1: Taxas de falhas e respectivas razões entre os riscos do MRLSP-WD para risco base crescente e diversos valores de λ ($\lambda = 1$ representa a distribuição base).

Fonte: Elaborado pelo autor

(1997), Bennett (1983) sugere que esse comportamento é útil para representar uma cura eficaz. A ideia que a taxa de mortalidade de um grupo de pessoas doentes se aproximaria da mortalidade de um grupo de controle conforme o tempo passa. Pode-se tomar como exemplo, a taxa de mortalidade para um grupo interrompendo um hábito, como fumar pode ser comparada com a taxa de um grupo de pessoas que não têm hábito de fumar.

Para risco base constante

A Figura 4.2 apresenta as funções de riscos para o MRLSP-WD com $\theta = 1.0$, $b = 0.50$ e diversos valores de lambda: $\lambda = 1.0$, $\lambda = 0.5$, $\lambda = 0.75$, $\lambda = 1.50$ e $\lambda = 2.5$.

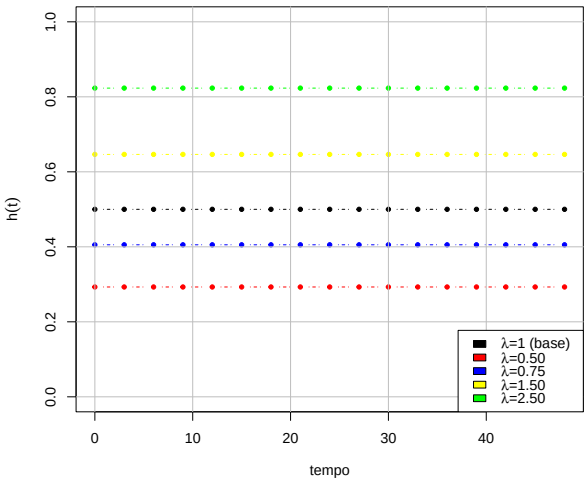


Figura 4.2: Taxas de falha do MRLSP-WD para risco base constante e diversos valores de λ ($\lambda = 1$ representa a distribuição base).

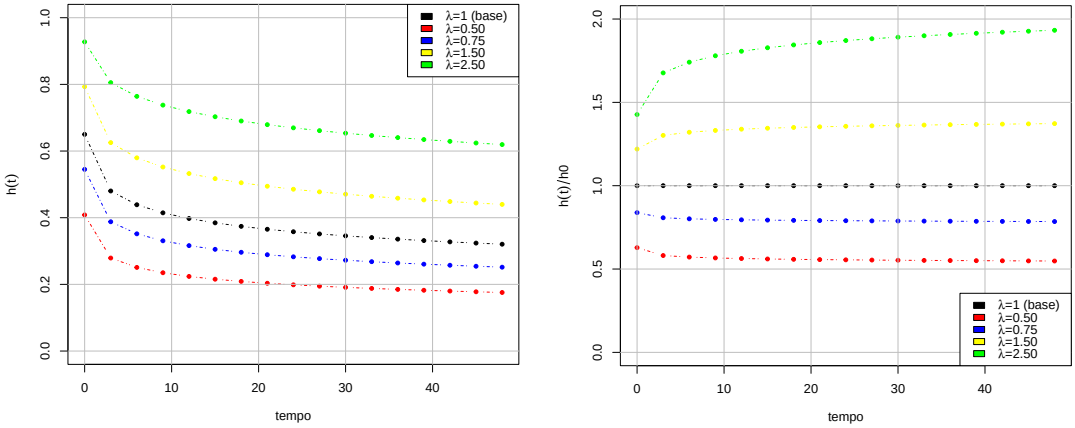
Fonte: Elaborado pelo autor

De acordo com a Figura 4.2, o risco do MRLSP de base constante é constante para qualquer valor de λ, este fato já foi referenciado na seção anterior. A partir deste fato pode-se concluir que as razões entre o risco do modelo e riscos bases são também constantes.

Para risco base decrescente

As Figuras (4.3)(a) e (4.3)(b) apresentam, respectivamente, as funções de riscos e as razões de riscos para a distribuição base Weibull discreta com $\theta = 0.8$, $b = 0.35$ e diversos valores de lambda: $\lambda = 1.0$, $\lambda = 0.5$, $\lambda = 0.75$, $\lambda = 1.50$ e $\lambda = 2.5$.

Com base na Figura 4.3(a), os riscos do MRLSP-WD decrescem com o tempo, este comportamento deve-se a função risco base ser monótona decrescente. Por outro, observa-se que os



(a) Riscos do MRLSP-WD (b) Razões de riscos do MRLSP-WD

Figura 4.3: Taxas de falhas e respectivas razões entre os riscos para risco base decrescente e diversos valores de λ ($\lambda = 1$ representa a distribuição base).

Fonte: Elaborado pelo autor

riscos tendem a se tornar paralelas entre eles quando o tempo tente ao infinito. Isto é, à medida que o tempo passa, os riscos do MRLSP-WD com risco base decrescente converge para um modelo de riscos proporcionais (Figura 4.3(b)).

É possível notar a partir das Figuras 4.1 a 4.3 que o MRLSP-WD apresenta: 1) a propriedade de riscos proporcionais quando a função de risco base é constante; 2) no limite (para t grande) a propriedade de riscos proporcionais quando a função base é decrescente e; 3) razão de riscos que tende a 1 no linite (fenômeno que representa uma "cura eficaz") quando o risco base é crescente.

4.2 Verificação da Suposição do MRLSP

Com o objetivo de investigar a suposição de log de sobrevivência proporcionais do modelo, na presente seção foram desenvolvidas ferramentas para este fim. Com o modelo proposto na Equação (4.1) é fácil verificar que a razão dos logaritmos das funções de sobrevivências é constante no tempo, isto é, pressupõe-se que a razão de logaritmos das funções de sobrevivências

para dois indivíduos são proporcionais.

Considerando, uma covariável **Z** dicotômica (que assume os valores 0 e 1), o MRLSP supõe que

$$\log\left[S(t|Z=1)\right] = \tau \log\left[S(t|Z=0)\right], \tag{4.7}$$

que resulta em

$$\log\left[-\log[S(t|Z=1)]\right] = \log(\tau) + \log\left[-\log[S(t|Z=0)]\right], \tag{4.8}$$

em que τ é a constante de proporcionalidade que não depende do tempo t.

Portanto, sob a suposição dos logaritmos das funções de sobrevivências serem proporcionais, a relação entre $\log\left[-\log[S(t|Z=1)]\right]$ e $\log\left[-\log[S(t|Z=0)]\right]$ é linear com coeficiente angular (declive) $m_0 = \log(\tau)$ e coeficiente linear (ordenada na origem) $m_1 = 1$, isto é,

$$y = m_0 + m_1 x, \tag{4.9}$$

em que, $y = \log\left[-\log[S(t|Z=1)]\right]$ e $x = \log\left[-\log[S(t|Z=0)]\right]$. Assim, a suposição de proporcionalidade do logaritmo da função de sobrevivência do MRLSP pode ser verificada graficamente, ajustando-se uma reta de regressão linear simples. A suposição de proporcionalidade será satisfeita se a reta de regressão apresentar coeficiente angular $m_1 = 1$.

É importante destacar que tal procedimento pode ser aplicado para covariáveis categóricas com três ou mais níveis, sendo necessário, nesse caso, comparar cada nível da covariável duas-a-duas. Outra circunstância em que é possível a utilização do método é quando há covariáveis numéricas, nessas situações é preciso categorizar os valores das covariáveis e compará-las duas-a-duas.

A análise gráfica é bastante informativa e para que determinada avaliação para tomada de decisão seja completa, é aconselhável que haja uma medida de evidência. Assim, ao considerar a Expressão (4.9), pode-se utilizar um teste de hipóteses, a fim de verificar se a razão de

logaritmo de taxa de falhas apresentam valores proporcionais entre si. Deste modo, seja $t_{(j)}$ o j-ésimo tempo distinto observado (censurado ou não censurado), $j = 1, 2, \ldots, J$. A verificação pode ser conduzida, ao testar a hipótese nula de que o coeficiente angular é igual a 1 contra a hipótese alternativa tal que o coeficiente angular é diferente de 1.

Assim, as hipóteses de interesse são descritas, em termos estatísticos, por:

$$H_0 : m_1 = 1 \quad vs \quad H_a : m_1 \neq 1. \tag{4.10}$$

A estatística do teste para a hipótese (4.10) é dada por:

$$M = \frac{\widehat{m_1} - 1}{\sqrt{\frac{\sum_{j=1}^{J}(x_j - \bar{x})^2}{(J-2)\sum_{j=1}^{J}(y_j - \bar{y})^2}}}, \tag{4.11}$$

em que, $\widehat{m_1} = \dfrac{J\sum_{j=1}^{J} x_j y_j - \sum_{j=1}^{J} x_j \sum_{j=1}^{J} y_j}{J\sum_{j=1}^{J} x_j^2 - \left(\sum_{j=1}^{J} x_j\right)^2}$, $\bar{x} = \dfrac{\sum_{j=1}^{J} x_j}{J}$ e $\bar{y} = \dfrac{\sum_{j=1}^{J} y_j}{J}$ com $y_j = \log\left[-\log[S(t_j|Z=1)]\right]$ e $x_j = \log\left[-\log[S(t_j|Z=0)]\right]$. Assumindo a normalidade de y, M segue uma distribuição t de Student com $J-2$ graus de liberdade.

4.3 Procedimento para estimação pontual dos parâmetros

Seja uma amostra aleatória observada $\mathbf{t} = (t_1, t_2, ..., t_n)^T$ de uma variável aleatória discreta T com função de sobrevivência e probabilidade do MRLSP dadas pelas Equações (4.2) e (4.4) respetivamente. O vetor de parâmetros $\boldsymbol{\vartheta}$ é definido por $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)^T$ onde $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)^T$ o vetor de dimensão k de parâmetros da distribuição base e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ o vetor de parâmetros da regressão de dimensão p. Aqui, $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_n)^T$ é o vetor dos indicadores de censura, onde δ_i é variável indicadora da falha a direita que é igual a 1 se o tempo t_i for de falha ou igual a 0 se for censura a direita e $\boldsymbol{z_i} = (z_{i1}, z_{i2}, ..., z_{ip})^T$ é o vetor de covariáveis

do indivíduo $i, i = 1, 2, \ldots, n$. A função de verossimilhança do MRLSP é dada por:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{t}, \mathbf{Z}, \boldsymbol{\delta}) \propto \prod_{i=1}^{n} \left\{ \left[[S_0(t_i - 1)]^{\exp(\boldsymbol{Z^T\beta})} - [S_0(t_i)]^{\exp(\boldsymbol{Z^T\beta})} \right]^{\left(1-\mathbb{I}_{\{t_i=0\}}\right)\delta_i} \right.$$

$$\times \left[1 - [S_0(0)]^{\exp(\boldsymbol{Z^T\beta})}\right]^{\mathbb{I}_{\{t_i=0\}}\delta_i} \tag{4.12}$$

$$\left. \times \left[[S_0(t_i)]^{\exp(\boldsymbol{Z^T\beta})}\right]^{1-\delta_i} \right\}.$$

Ao aplicar o logaritmo na função de verossimilhança (4.12) obtém-se a função log-verossimilhança:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{t}, \mathbf{Z}, \boldsymbol{\delta}) = \sum_{i=1}^{n} \left\{ \left(1 - \mathbb{I}_{\{t_i=0\}}\right) \delta_i \log \left[[S_0(t_i - 1)]^{\exp(\boldsymbol{Z^T\beta})} - [S_0(t_i)]^{\exp(\boldsymbol{Z^T\beta})} \right] \right\}$$

$$+ \sum_{i=1}^{n} \left\{ \mathbb{I}_{\{t_i=0\}} \delta_i \log \left[1 - [S_0(0)]^{\exp(\boldsymbol{Z^T\beta})}\right] \right\} \tag{4.13}$$

$$+ \sum_{i=1}^{n} \left\{ (1 - \delta_i) \exp(\boldsymbol{Z^T\beta}) \log [S_0(t_i)] \right\} + \xi,$$

em que ξ é uma constante que não depende de $\boldsymbol{\theta}$ e $\boldsymbol{\beta}$.

Obtendo-se a primeira derivada da função de log-verossimilhança (função escore), e igualando a zero obtém-se o sistema de equações:

$$\boldsymbol{U(\vartheta)} = \frac{\boldsymbol{\partial \ell(\vartheta)}}{\boldsymbol{\partial \vartheta}} = \mathbf{0}, \tag{4.14}$$

em que $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)^T$.

O conjunto solução $\widehat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{\theta}}^T, \widehat{\boldsymbol{\beta}}^T)^T$ que satisfaz o sistema de equações (4.14) é o esti-

mador de máxima verossimilhança do MRLSP que, sob condições de regularidade adequadas, é consistente e possui distribuição assintótica normal multivariada com média $\boldsymbol{\vartheta}$ e matriz de variâncias e covariâncias dada por

$$\sum(\boldsymbol{\vartheta}) = \left[-\frac{\partial^2 \ell(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^T} \bigg|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}} \right]^{-1} = \left[-\boldsymbol{J}(\boldsymbol{\vartheta}) \bigg|_{\boldsymbol{\vartheta}=\widehat{\boldsymbol{\vartheta}}} \right]^{-1}. \tag{4.15}$$

Os elementos da matriz de informação observada de Fisher $\boldsymbol{J}(\boldsymbol{\vartheta})$ são obtidos numericamente, por meio de métodos computacionais de otimização utilizando algoritmo de Newton-Raphson, ou por métodos iterativos de Quase-Newton ou escore de Fisher que fornecem uma aproximação numérica precisa para essa matriz.

4.4 Procedimentos para estimação intervalar dos parâmetros

Para os parâmetros de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ o intervalo é obtido como descrito na Seção (2.4.2), uma vez que eles não apresentam nenhuma restrição. Assim, um intervalo assintótico com nível de $(1-\alpha) \times 100\%$ de confiança é dada por:

$$\left[\widehat{\beta}_j - Z_{(1-\alpha/2)} \sqrt{\widehat{Var}(\widehat{\beta}_j)}; \widehat{\beta}_j + Z_{(1-\alpha/2)} \sqrt{\widehat{Var}(\widehat{\beta}_j)} \right], \quad j = 1, 2, \ldots, p, \tag{4.16}$$

em que $\widehat{\beta}_j$ é o estimador de máxima verossimilhança que é obtido resolvendo o sistema (4.14), $Z_{(1-\alpha/2)}$ o quantil $1-\alpha/2$ de uma distribuição normal padrão e $\widehat{Var}(\widehat{\beta}_j)$ é a variância das estimativas dos parâmetros de regressão, dada por (4.15).

No que se refere aos parâmetros da distribuição indicada com a função de sobrevivência base, $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_k)^T$, estes podem apresentar limitações no espaço paramétrico. Os casos mais comuns ocorrem quando:

$$0 < \theta_j < 1, \quad j = 1, 2, \ldots, k \tag{4.17}$$

ou

$$\theta_j > 1, \quad j = 1, 2, \ldots, k. \tag{4.18}$$

Nestes casos, é interessante fazer uma transformação para deixá-los irrestritos e então construir intervalos de confiança cujos limites respeitem os seus respectivos espaços paramétricos. Assim, para o caso (4.17), pode-se fazer uma transformação log-log e para o caso (4.18) pode-se fazer uma transformação log como procede-se a seguir.

- **Intervalo de confiança para o caso** $0 < \theta_j < 1$

Consideremos a transformação log-log para θ_j. Seja $u_j = \log\left[-\log(\theta_j)\right]$, que resulta em $\theta_j = e^{-e^{u_j}}$. Pela propriedade de invariância dos estimadores de máxima verossimilhança temos que $\widehat{u}_j = \log\left[-\log(\widehat{\theta}_j)\right]$, que resulta em $\widehat{\theta}_j = e^{-e^{\widehat{u}_j}}$.

Assim, pelo método delta, temos que $\widehat{Var}(\widehat{u}_j) = \left[\frac{du_j}{d\theta_j}\Big|_{\theta_j=\widehat{\theta}_j}\right]^2 Var(\widehat{\theta}_j)$, que resulta em $\widehat{Var}(\widehat{u}_j) = \left[\frac{1}{\widehat{\theta}_j \log \widehat{\theta}_j}\right]^2 Var(\widehat{\theta}_j)$. Ademais, a função logarítmica é sobrejetiva ou seja, assume todos números reais, assim quer dizer que o parâmetro u_j é irrestrito, logo o seu intervalo é dado como descrito em (2.4.2):

$$\left[\widehat{u}_j - Z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\widehat{u}_j)}; \widehat{u}_j + Z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\widehat{u}_j)}\right]. \tag{4.19}$$

Considerando o intervalo de confiança para θ_j com a transformação feita, o intervalo de confiança assintótico ao nível de $(1-\alpha)\times 100\%$ de θ_j é dado por:

$$\left[\widehat{\theta}_j^{\exp\left\{\frac{Z_{(1-\alpha/2)}}{\widehat{\theta}_j \log \widehat{\theta}_j}\sqrt{\widehat{Var}(\widehat{\theta}_j)}\right\}}; \widehat{\theta}_j^{\exp\left\{-\frac{Z_{(1-\alpha/2)}}{\widehat{\theta}_j \log \widehat{\theta}_j}\sqrt{\widehat{Var}(\widehat{\theta}_j)}\right\}}\right]. \tag{4.20}$$

em que $\widehat{\theta}_j$ é o estimador de máxima verossimilhança que é obtido resolvendo o sistema (4.14), $Z_{(1-\alpha/2)}$ o quantil $1-\alpha/2$ de uma distribuição normal padrão e $\widehat{Var(\widehat{\theta}_j)}$ é a variância da estimativa do parâmetro da distribuição base, dada por (4.15).

- **Intervalo de confiança para o caso** $\theta_j > 1$

Consideremos a transformação log para θ_j. Seja $w_j = \log(\theta_j)$, que resulta em $\theta_j = e^{w_j}$. Pela propriedade de invariância dos estimadores de máxima verossimilhança temos que $\widehat{w}_j = \log(\widehat{\theta}_j)$, que resulta em $\widehat{\theta}_j = e^{\widehat{w}_j}$.

Assim, pelo método delta, temos que $\widehat{Var}(\widehat{w_j}) = \left[\frac{dw_j}{d\theta_j}\Big|_{\theta_j=\widehat{\theta}_j}\right]^2 Var(\widehat{\theta}_j)$, que resulta em $\widehat{Var}(\widehat{w_j}) = \left[\frac{1}{\widehat{\theta}_j}\right]^2 Var(\widehat{\theta}_j)$. Ademais, a função logarítmica é sobrejetiva ou seja, assume todos números reais, assim quer dizer que o parâmetro w_j é irrestrito, logo o seu intervalo é dado como descrito em (2.4.2):

$$\left[\widehat{w_j} - Z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\widehat{w_j})}; \widehat{w_j} + Z_{(1-\alpha/2)}\sqrt{\widehat{Var}(\widehat{w_j})}\right]. \tag{4.21}$$

Considerando o intervalo de confiança para θ_j com a transformação feita, o intervalo de confiança assintótico ao nível de $(1-\alpha)\times 100\%$ de θ_j é dado por:

$$\left[\widehat{\theta}_j e^{-\frac{Z_{(1-\alpha/2)}}{\widehat{\theta}_j}\sqrt{\widehat{Var}(\widehat{\theta}_j)}}; \widehat{\theta}_j e^{\frac{Z_{(1-\alpha/2)}}{\widehat{\theta}_j}\sqrt{\widehat{Var}(\widehat{\theta}_j)}}\right]. \tag{4.22}$$

em que $\widehat{\theta}_j$ é o estimador de máxima verossimilhança que é obtido resolvendo o sistema (4.14), $Z_{(1-\alpha/2)}$ o quantil $1-\alpha/2$ de uma distribuição normal padrão e $\widehat{Var}(\widehat{\theta}_j)$ é a variância da estimativa do parâmetro da distribuição base, dada por (4.15).

4.5 Procedimento para teste de hipóteses

Para testar a significância estatística dos parâmetros estimados do modelo, pode-se aplicar o Teste de Razão de Verossimilhança (TRV) ou teste de Wald. Os procedimentos para tal foram descrito na subseção (2.4.3).

Capítulo 5

Simulações Computacionais

Devido à complexidade do Modelo de Regressão Log de Sobrevivências Proporcionais Weibull Discreto (MRLSP-WD), as condições de regularidade não são fáceis de verificar analiticamente. Neste caso, estudos de simulação são necessários; veja, por exemplo, (Cardial, Cobre e Nakano, 2024), (Ha e Mackenzie, 2010), (Ortega et al., 2015) e (Barriga et al., 2019). Seguindo essa ideia, no presente capítulo, é descrito um estudo de simulação realizado para investigar se as propriedades assintóticas usuais dos estimadores de máxima verossimilhança se mantêm. Também foi avaliado o comportamento do modelo proposto na presença de dados censurados. O estudo do MRLSP-WD foi conduzido, considerando dados simulados no software R (R Core Team, 2024). Foram simulados tempos de sobrevivências do MRLSP-WD pelo método da transformação inversa, conforme o procedimento descrito a seguir. Inicialmente foram gerados tempos de sobrevivência contínuos por meio do Modelo Log de Sobrevivência Proporcionais Weibull (contínuo) que é dada pela Equação (4.2) com

$$S_0(t) = \left[\exp\left\{-\left(\frac{t}{\gamma}\right)^{\theta}\right\}\right], \quad \text{para} \quad t \geqslant 0, \tag{5.1}$$

em que $\theta > 0$ é o parâmetro de forma e $\gamma > 0$ é o parâmetro de escala.

Sejam os parâmetros θ e b da distribuição Weibull discreta, o parâmetro de escala, γ, da

distribuição Weibull contínua é obtida por

$$\gamma = \left\{ -\frac{1}{\log(b)} \right\}^{\frac{1}{\theta}}. \tag{5.2}$$

O parâmetro de forma, θ da distribuição Weibull contínua é o mesmo da distribuição Weibull discreta. Os tempos contínuos gerados foram discretizados considerando-se sua parte inteira, isto é, o maior inteiro menor ou igual ao valor gerado. Ademais, seguindo o procedimento descrito em Oliveira (2021), a censura foi incorporada independentemente do tempo de sobrevivência por meio de uma variável indicadora de censura gerada por uma distribuição Bernoulli, com os percentuais de censura que estão especificados na Seção 5.1.

Acrescenta-se ainda que, as covariáveis, foram incluídas no modelo levando em conta a função de ligação logarítmica, ou seja, $g(Z^T\beta) = \exp\{Z^T\beta\}$.

5.1 Tempo discretos sem presença de censura

As amostras dos tempos de sobrevivência foram simuladas, considerando duas covariáveis e diversos parâmetros do MRLSP-WD, tendo em vista os diferentes comportamentos das taxas de falha base conforme a Tabela 5.1.

Tabela 5.1: Casos das simulações.

Caso	b	θ	β_1	β_2	Taxa da Falha Base
C_1	0.95	1.50	2.0	1.0	Crescente
C_2	0.95	1.00	2.0	1.0	Constante
C_2	0.95	0.75	2.0	1.0	Decrescente

Fonte: Elaborado por autor

Os resultados das simulações apresentados neste trabalho consideraram no modelo uma covariável dicotômica, Z_1, simulada a partir de uma distribuição Bernoulli com probabilidade

de sucesso $p = 0.5$ e uma covariável Z_1, com distribuição normal padrão para cada cenário descrito.

As médias das estimativas, o erro quadrático médio (EQM) e a probabilidade de cobertura (PC) dos estimadores do MRLSP-WD foram calculadas a partir de M=10.000 réplicas de Monte Carlo, considerando amostras de tamanho $n = 30, 50, 100$ e 500. Para a construção dos intervalos de confiança e para o cálculo da PC foi considerado o nível de confiança de 95%.

A apresentação tabelada dos resultados está disposta por parâmetros, a fim de verificar o comportamento destes em relação ao tamanho amostral. Os parâmetros b e θ são provenientes da distribuição base WD e este segundo é responsável pelo comportamento da taxa de falha, e por este motivo são considerados diferentes valores do mesmo (Casos 1 a 3), frente a esses comportamentos. Ademais, os parâmetros β_1 e β_2 são os coeficientes de regressão associados as covariáveis Z_1 e Z_2, respectivamente.

Os resultados das simulações são apresentadas na Tabela 5.2, que contém as estimativas, erro quadrático médio e a probabilidade de cobertura dos parâmetros do modelo sem percentual de censura com os três (3) casos e diferentes tamanhos amostrais.

Tabela 5.2: Média das estimativas, EQM e PC dos parâmetros do MRLSP-WD considerando os casos da simulação e diversos tamanhos amostrais na ausência de censuras.

n	Caso	b			θ			β_1			β_2		
		Média	**EQM**	**PC**	**Média**	**EQM**	**PC**	**Média**	**EQM**	**PC**	**Média**	**EQM**	**PC**
	C1	0.9538	0.0009	0.9160	1.6658	0.1197	0.8830	2.2068	0.4082	0.9140	1.0872	0.0953	0.9230
30	C2	0.9538	0.0009	0.9160	1.1056	0.0490	0.8840	2.1919	0.3736	0.9160	1.0816	0.0887	0.9250
	C3	0.9538	0.0009	0.9210	0.8303	0.0281	0.8810	2.1961	0.3732	0.9170	1.0813	0.0896	0.9280
	C1	0.9542	0.0005	0.9390	1.6030	0.0538	0.9100	2.1365	0.1950	0.9420	1.0518	0.0515	0.9320
50	C2	0.9580	0.0005	0.9280	1.6479	0.0450	0.9070	2.2155	0.1179	0.9200	1.0518	0.0511	0.9300
	C3	0.9521	0.0005	0.9330	0.7914	0.0127	0.9140	2.1327	0.1877	0.9370	1.0528	0.0508	0.9360
	C1	0.9514	0.0003	0.9530	1.5456	0.0213	0.9360	2.0645	0.0838	0.9380	1.0392	0.0259	0.9320
100	C2	0.9515	0.0002	0.9450	1.5450	0.0214	0.9270	2.0623	0.0862	0.9380	1.0270	0.0260	0.9540
	C3	0.9515	0.0002	0.9460	0.7721	0.0052	0.9310	2.0611	0.0829	0.9400	1.0267	0.0256	0.9470
	C1	0.9504	0.0004	0.9530	1.5091	0.0035	0.9480	2.0133	0.0145	0.9600	1.0066	0.0038	0.9560
500	C2	0.9503	0.0001	0.9490	1.0057	0.0016	0.9470	2.0129	0.0144	0.9600	1.0061	0.0037	0.9540
	C3	0.9503	0.0001	0.9470	0.7543	0.0009	0.9470	2.0129	0.0143	0.9570	1.0062	0.0037	0.9540

Fonte: Elaborado pelo autor

Em relação ao parâmetro b, nota-se que as médias das estimativas são aproximadamente

iguais ao verdadeiro valor do parâmetro, independente do caso e tamanho amostral, e que à medida que o tamanho amostral aumenta há uma tendência das estimativas serem cada vez menos viesadas, isto é, o erro quadrático médio diminui significativamente e com probabilidade de cobertura em torno de 95%. Destaca-se ainda que, todas as estimativas de EQM são próximas de zero ($EQM(b) \leq 0,0009$). Além disso, as probabilidades de cobertura, não diferem mais de $0,034$ do nível de confiança adotado, mesmo para $n = 30$ (menor tamanho amostral), para todos os casos e tamanhos amostrais em estudo indicando uma ótima precisão do estimador mesmo para tamanhos amostrais menores.

Os resultados que se referem ao estimador θ constata-se que as médias das estimativas se aproximam ao verdadeiro valor do parâmetro conforme o tamanho da amostra aumenta. Isso ocorre independente do cenário considerado (Tabela 5.2). Ademais, a probabilidade de cobertura distancia-se do nível de confiança adotado quando o tamanho amostral é pequeno e aproxima-se ao nível da confiança considerado com o tamanho amostral suficientemente grande.

Em relação ao parâmetro β_1 constata-se que as estimativas médias aproximam-se ao verdadeiro valor do parâmetro com aumento do tamanho amostral e com EQM não superior a 0.4082 para todos os casos e diminui com o aumento do tamanho da amostra. A probabilidade de cobertura se aproxima ao nível de confiança nominal quando a amostra cresce.

Para o parâmetro β_2 as estimativas médias estão mais próxima do verdadeiro valor do parâmetro com EQM não superior a 0.0953 para todos os casos. Com o aumento do tamanho da amostra, o EQM diminui e a probabilidade de cobertura se aproxima do nível de confiança adotado.

5.2 Tempos discretos com presença de censura

Do ponto de vista teórico, quando há a presença de censuras em determinado conjunto de dados é esperado que as estimativas obtidas se comportem de forma diferente do que foi apresentado na Seção 5.1 em relação as diferenças entre estimativas e verdadeiro valor do parâmetro,

devido ao fato do verdadeiro valor do parâmetro não ser mais o que foi proposto nos Casos da simulações, visto que as simulações são baseadas na Função Distribuição Acumulada (FDA) do modelo que por sua vez, não leva em conta a parte censurada. Visando avaliar o comportamento das estimativas do MRLSP-WD na presença de censura, nessa seção, serão avaliados os resultados (Médias das estimativas, EQM e PC) dispostos por parâmetros seguindo o mesmo raciocínio estabelecido na Seção 5.1 considerando percentuais de censura iguais a 5, 10 e 25%. Para a devida interpretação dos resultados, a respeito de cada um dos parâmetros, serão apresentados considerando os diversos Casos (Tabelas 5.3 e 5.4) e tamanhos amostrais (n = 30, 50, 100 e 500) separados pelos diversos percentuais de censura citados.

Tabela 5.3: Média das estimativas, EQM e PC dos parâmetros do MRLSP-WD considerando os casos da simulação e diversos tamanhos amostrais na presença de censuras para $n = 30$ e $n = 50$.

n	Perc. Cens.	Caso	b			θ			β_1			β_2		
			Média	**EQM**	**PC**	**Média**	**EQM**	**PC**	**Média**	**EQM**	**PC**	**Média**	**EQM**	**PC**
30	0.05	C1	0.9534	0.0008	0.9290	1.6014	0.1082	0.8830	2.2182	0.3460	0.9360	1.0256	0.0966	0.9230
		C2	0.9566	0.0008	0.9060	1.1036	0.0510	0.8760	2.2797	0.4358	0.9030	1.0930	0.1330	0.9290
		C3	0.9551	0.0009	0.9100	0.8152	0.0291	0.8690	2.2618	0.3950	0.9110	1.0708	0.0866	0.9190
	0.10	C1	0.9551	0.0009	0.8960	1.5991	1.1457	0.8210	2.3390	0.4735	0.9040	1.0281	0.1020	0.8850
		C2	0.9551	0.0008	0.9080	1.0343	0.0489	0.8590	2.2941	0.4065	0.9190	1.0915	0.0967	0.9090
		C3	0.9556	0.0009	0.9160	0.8027	0.0317	0.8640	2.3275	0.4669	0.8990	1.0485	0.0971	0,9070
	0.25	C1	0.9568	0.0008	0.9160	1.4129	0.1383	0.8910	2.4921	0.5732	0.9120	0.6755	0.2645	0.7070
		C2	0.9491	0.0008	0.9000	0.7546	0.1174	0.8552	2.5154	0.5264	0.9310	1.3088	1.7205	0.7140
		C3	0.9492	0.0008	0.9510	0.5674	0.0662	0.8460	2.5128	0.5273	0.9200	1.3149	0.7145	0.8180
50	0.05	C1	0.9489	0.0007	0.9120	1.4936	0.0744	0.8410	2.1528	0.1851	0.9500	0.9529	0.1734	0.8560
		C2	0.9487	0.0007	0.9130	0.9952	0.0324	0.8340	2.1485	0.1810	0.9470	1.9533	0.1707	0.8530
		C3	0.9487	0.0007	0.9190	0.7468	0.0181	0.8380	2.1261	0.1781	0.9480	1.9569	0.1703	0.8550
	0.10	C1	0.9468	0.0007	0.8950	0.7468	0.0181	0.8380	2.1461	0.1781	0.9480	1.9569	0.1703	0.8550
		C2	0.9467	0.0008	0.8960	0.9376	0.0455	0.7440	2.2006	0.2025	0.9440	1.8081	0.2453	0.7490
		C3	0.9465	0.0008	0.8980	0.7030	0.0255	0.7480	2.1972	0.2006	0.9420	1.8097	0.2430	0.7430
	0.25	C1	0.9460	0.0007	0.9360	1.1987	0.1873	0.7270	2.3640	0.2862	0.9310	1.4395	0.5133	0.8630
		C2	0.9458	0.0007	0.9360	0.8011	0.0824	0.7330	2.3573	0.2770	0.9310	1.4487	0.5062	0.8660
		C3	0.9456	0.0007	0.9420	0.6012	0.0464	0.7330	2.3572	0.2800	0.9320	1.4504	0.5034	0.8720

Fonte: Elaborado por autor

Observando atentamente as Tabelas 5.3 e Tabela 5.4 constata-se o seguinte em relação aos parâmetros do modelo:

Para todos os parâmetros do modelo, pode-se observar que quanto maior for o percentual

Tabela 5.4: Média das estimativas, EQM e PC dos parâmetros do MRLSP-WD considerando os casos da simulação e diversos tamanhos amostrais na presença de censuras para $n = 100$ e $n = 500$.

n	Perc. Cens.	Caso	b			θ			β_1			β_2		
			Média	**EQM**	**PC**	**Média**	**EQM**	**PC**	**Média**	**EQM**	**PC**	**Média**	**EQM**	**PC**
100	0.05	C1	0.9464	0.0005	0.8830	1.4467	0.0444	0.7920	2.0936	0.0968	0.9320	1.9077	0.0982	0.8120
		C2	0.9450	0.0005	0.9000	0.9559	0.0213	0.7650	2.0690	0.0851	0.9480	1.8907	0.1031	0.7950
		C3	0.9450	0.0005	0.9000	0.7171	0.0119	0.7680	2.0689	0.0842	0.9520	1.8918	0.1022	0.7970
	0.10	C1	0.9407	0.0006	0.8780	1.3319	0.0770	0.7860	2.0909	0.0855	0.9500	1.7330	0.1749	0.8190
		C2	0.9410	0.0006	0.8770	0.8912	0.0334	0.7950	2.0917	0.0850	0.9500	1.7428	0.1677	0.8310
		C3	0.9290	0.0015	0.8720	0.6446	0.0850	0.7260	2.0008	0.1023	0.9300	1.7385	0.1839	0.8780
	0.25	C1	0.9139	0.0022	0.8020	1.0344	0.2666	0.7670	2.0254	0.8760	0.9750	1.4012	0.4401	0.8110
		C2	0.9140	0.0021	0.8010	0.6929	0.1169	0.7750	0.0224	0.0748	0.9730	1.4103	0.4308	0.8220
		C3	0.9138	0.0022	0.8970	0.5202	0.0656	0.7800	2.0199	0.0748	0.9720	1.4127	0.4282	0.8250
500	0.05	C1	0.9344	0.0006	0.8810	1.3121	0.0581	0.7470	1.9488	0.0254	0.8610	1.7415	0.1046	0.7950
		C2	0.9346	0.0006	0.8870	0.8778	0.0251	0.7460	1.9509	0.0251	0.8500	1.7486	0.1005	0.7920
		C3	0.9347	0.0006	0.8950	0.6587	0.0139	0.7480	1.9511	0.0247	0.8550	1.7508	0.0991	0.7890
	0.10	C1	0.9270	0.0008	0.8780	1.1986	0.1122	0.7780	1.9474	0.0222	0.8910	1.5771	0.2119	0.7110
		C2	0.9270	0.0009	0.8840	0.8000	0.8490	0.7800	1.9480	0.0221	0.8810	1.5853	0.2051	0.7120
		C3	0.9273	0.0008	0.8890	0.6012	0.0271	0.7320	1.9394	0.0216	0.8900	1,5853	0.2011	0.7070
	0.25	C1	0.9223	0.0010	0.8060	1.0043	0.2574	0.7850	2.0523	0.0168	0.9580	1.2752	0.5455	0.7320
		C2	0.9223	0.0010	0.8030	0.6737	0.1118	0.7840	2.0501	0.0164	0.9560	1.2859	0.5302	0.7310
		C3	0.9220	0.0010	0.8040	0.5048	0.0631	0.7840	2.0480	0.0164	0.9580	1.2865	0.5297	0.7300

Fonte: Elaborado pelo autor

de censura nos dados maior são os desvios das estimativas em relação ao verdadeiro valor do parâmetro, que é comportamento natural para o estimador, visto que a maior quantidade de observações censuradas alteram o valor das estimativas a respeito do verdadeiro valor do parâmetro consideravelmente. As probabilidades de cobertura apresentadas nas Tabelas 5.3 e 5.4, que têm como grau de confiança 0,95, reforça essa afirmação. Note que, a maior quantidade de censuras (25%) conduzem aos piores resultados de probabilidade de cobertura e maiores diferenças a respeito do grau de confiança considerado para as simulações. É importante destacar que a probabilidade de cobertura estimada diferente do que o verdadeiro grau de confiança, neste caso não é sinônimo de o modelo produzir estimativas não condizentes e sim de que o verdadeiro valor do parâmetro não é aquele adotado para a gerar os dados (distribuição que não considera a parte censurada). Consequentemente o verdadeiro valor do parâmetro não é o

proposto e sim um valor que se aproxima das estimativas apresentadas nas Tabelas 5.3 e 5.4.

De modo geral, fundamentado nos resultados das simulações da presente seção, nota-se que o comportamento dos estimadores na presença de observações censuradas conduz a desvios das estimativas em relação ao verdadeiro valor do parâmetro que são maiores à medida que o percentual de censura aumenta. Consequentemente a probabilidade de cobertura se torna cada vez mais distante a respeito do verdadeiro grau de confiança.

Capítulo 6

Aplicação em Dados Reais

Neste Capítulo é apresentada uma aplicação do MRLSP-WD em dados reais especificamente num conjunto de dados de tempo de sobrevivência de pacientes que sofrem uma doença chamada de Leucemia Mieloide Aguda (AML). Na seção 6.1 faz-se apresentação de conjunto de dados, o gráfico da função de sobrevivência estimada via Kaplan-Meier, ajuste do MRLSP-WD e o respectivo gráfico da função de sobrevivência. Na subseção 6.1.1 faz a verificação da suposição de log de sobrevivência proporcionais e na subseção 6.1.2 faz-se a comparação do MRLSP em relação aos MRCSP-WD e MRWD através dos seus erros máximos.

6.1 Dados de Leucemia

A Leucemia Mieloide Aguda [*Acute Myeloid Leukemia* (AML)] é um tipo de câncer no qual a medula óssea produz um grande número de células sanguíneas anormais. É o tipo mais comum de leucemia aguda em adultos e piora rapidamente se não for tratado. O conjunto de dados da presente aplicação é proveniente do livro de Lee e Wang (2003) (e foi posteriormente trabalhado por Cardial, Cobre e Nakano (2024)) e se refere ao tempo de sobrevivência em semanas de 30 pacientes com AML. Este trabalho considerou a idade como possível prognóstico e a mesma foi consideradas de forma dicotômica:

$$Z_1 = \begin{cases} 1, & \text{se o paciente tem 50 anos ou mais.} \\ 0, & \text{caso contrário.} \end{cases}$$

Os dados da aplicação são apresentados na Tabela 6.1 com 20% de observações censurada e o Gráfico 6.1 apresenta a estimativa da função de sobrevivência (para ambos níveis da covariável Idade) obtida pelo estimador de Kaplan-Meier.

Tabela 6.1: Tempo de sobrevivência e Idade de 30 pacientes com AML.

Tempo de sobrevivência	Z_1	Tempo de sobrevivência	Z_1
18	0	8	1
9	0	2	1
28+	0	26+	1
31	0	10	1
19+	0	4	1
45+	0	3	1
6	0	4	1
18	0	18	1
8	0	8	1
15	0	3	1
23	0	14	1
28+	0	3	1
7	1	13	1
12	1	13	1
9	1	35+	1

Nota: "+"indica observações censuradas
Fonte: Lee e Wang (2003)

A seguir passa-se apresentar o gráfico da função de Sobrevivência via Estimador de Kaplan-Meier do conjunto dos dados da Tabela 6.1 para a variável Idade (Z_1). Por meio da Figura 6.1, observa-se que pacientes com menos de 50 anos (Z_1=0) tem maior probabilidade de sobrevivência visto que o Estimador de Kaplan-Meier para a referida categoria é maior em relação

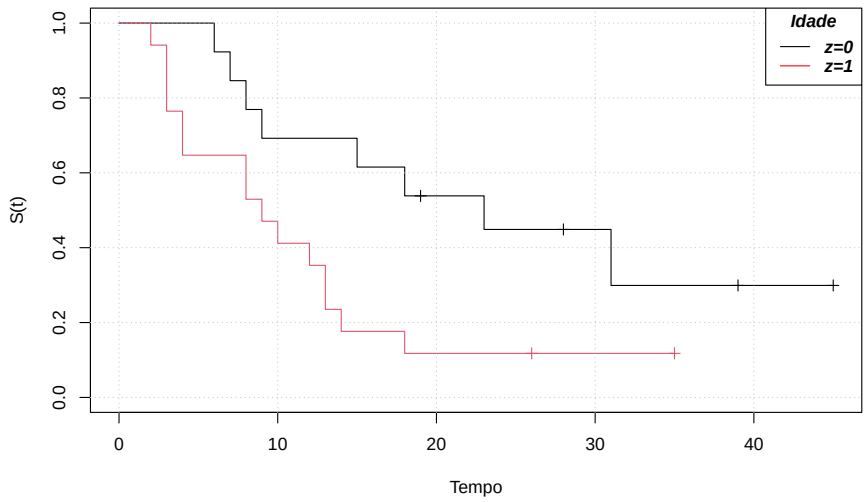


Figura 6.1: Função de sobrevivência via Estimador de Kaplan-Meier para o conjunto de dados da Tabela 6.1

Fonte: Elaborado pelo autor

á função de sobrevivência estimada dos pacientes com 50 anos ou mais (Z_1=1) para qualquer tempo t.

Dando prosseguimento ao estudo, foi realizado ajuste do Modelo de Regressão Log de Sobrevivências Proporcionais Weibull Discreto (MRLSP-WD). Inicialmente foi ajustado o modelo MRLSP-WD com propósito de verificar a significância de sua covariável ($H_0 : \beta_1 = 0$). Os resultados das estimativas, erro padrão e intervalo de confiança são apresentados na Tabela 6.2.

Os resultados da Tabela 6.2 comprovam que a covariável idade influencia a sobrevivência dos pacientes com AML (pois o intervalo de confiança do verdadeiro valor do parâmetro de regressão não inclui o valor zero).

A Figura 6.2 apresenta a estimativa da função de sobrevivência do MRLSP-WD para os dados da Tabela 6.1, indicando um bom ajuste do modelo a esse conjunto de dados.

Por meio das estimativas da Tabela 6.2, pode-se obter uma interpretação, quanto a log de sobrevivência para as diferentes categorias da covariável idade. Uma vez que $\exp(\beta_1)$ repre-

Tabela 6.2: Estimativas dos parâmetros do MRLSP-WD (dados da Tabela 6.1).

Variável	Parâmetro	Estimativa	Erro Padrão	I.C(95%)
	b	0.9881	0.0084	[0.9528, 0.9971]
	θ	1.2611	0.1962	[0.8545, 1.8611]
Idade	β_1	1.1484	0.4383	[0.2894, 2.0074]

Nível de referência: O paciente tem menos de 50 anos ($z_1 = 0$).

Fonte: Elaborado pelo autor

senta a razão dos log de sobrevivência dos diferentes grupos, constante ao longo do tempo, admitindo o grupo de pacientes que tenha 50 anos ou mais $(z_1 = 1)$. Nesse contexto, o log de sobrevivência do paciente que tenha 50 anos ou mais é $\exp(1.1484) = 3.153144$ vezes ao log de sobrevivência dos pacientes com menos de 50 anos.

6.1.1 Verificação da suposição de log de sobrevivência Proporcionais

A suposição de log de sobrevivências proporcionais foi verificada para os dados da Tabela 6.1, observando a referida proporcionalidade para cada um dos níveis da covariável idade, por meio do gráfico $\log(R_0(t)) \times \log(R_1(t))$ e o teste de hipótese proposto em na Seção 4.2. Aqui, $R_l(t) = -\log[S(t|Z = l)]$, l = 0, 1. Os resultados são apresentados na Figura 6.3 Com base dos resultados do teste apresentado na Figura 6.3 nota-se que, a suposição de log de sobrevivência proporcionais não foi rejeitada (tendo em vista um nível de significância de 5%), o gráfico apresentado entra em concordância para tal proposição, visto que os pontos formados pelas coordenadas estão próximos a reta de regressão ajustada cujo coeficiente angular não é significativamente diferente de 1 (p = 0,076).

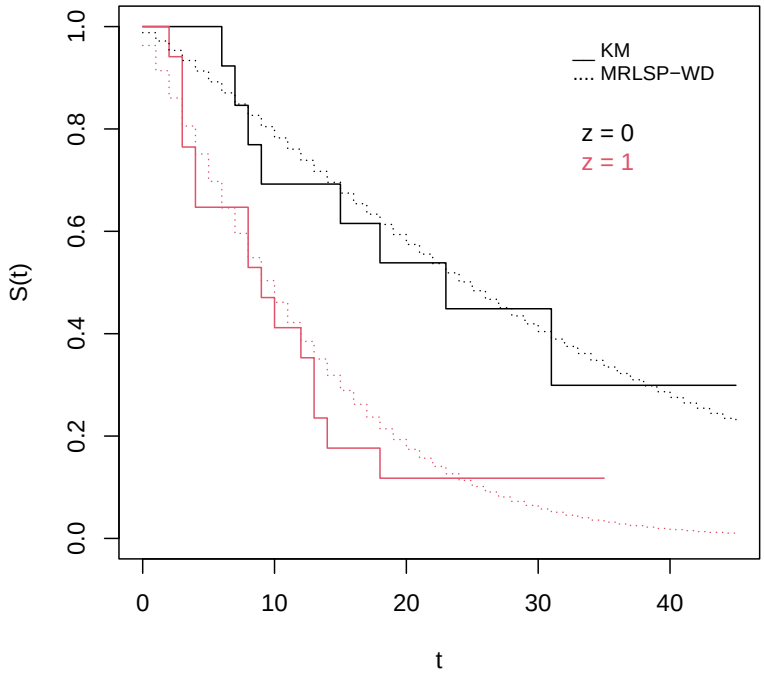


Figura 6.2: Função de sobrevivência estimada do MRLSP-WD para cada nível da covariável Idade do conjunto de dados da Tabela 6.1.

Fonte: Elaborado pelo autor

6.1.2 Comparação do MRLSP-WD com o Modelo de Regressão Chance de sobrevivência Proporcionais WD e Modelo de Regressão WD

A título de comparação, foram chamados dois modelos de regressão estudados por Cardial, Cobre e Nakano (2024) sendo: o Modelo de Regressão Chance de Sobrevivências Proporcionais Weibull discreto (MRCSP-WD) e o Modelo de Regressão Weibull discreto MRWD. A função de sobrevivência do MRCSP-WD é dada por Cardial, Cobre e Nakano (2024):

$$S(t|\boldsymbol{z}) = \frac{\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}S_0(t)}{1+(\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}-1)S_0(t)}, t = 0, 1, 2, \ldots, \tag{6.1}$$

em que $S_0(t)$ é função de sobrevivência base de uma distribuição Weibull discreta.

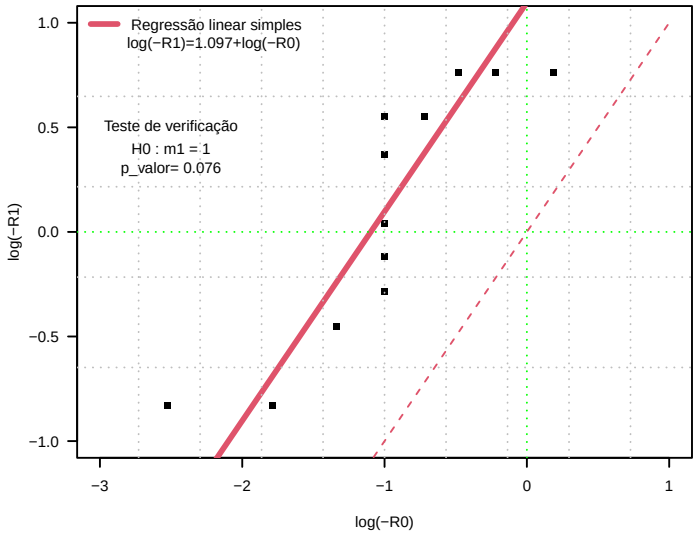


Figura 6.3: Verificação de suposição de log de sobrevivência proporcionais para a covariável Idade.

Fonte: Elaborado pelo autor

Já o modelo modelo de MRWD é um modelo de regressão mais simples levando em conta a distribuição WD (Equações 3.7, 3.8 e 3.9) com o parâmetro b recebendo o preditor linear (neste caso, com o intercepto β_0) por meio da função de ligação logito dada por:

$$b = \frac{\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{z}'\boldsymbol{\beta}\}}. \tag{6.2}$$

As estimativas dos parâmetros do MRCSP-WD e MRWD para os dados da Tabela 6.1 são apresentados nas Tabelas 6.3 e 6.4, respectivamente.

Para avaliação de ajuste dos modelos aos dados a título de comparação com o MRCSP-WD e o MRWD, foram calculados os erros máximos cometidos na estimação, denotados por ϵ, para cada um dos referidos níveis para os 3 modelos em estudo. Esse erro é apresentado na Tabela 6.5 e é baseado na diferença máxima entre as estimativas da função de sobrevivência dos modelos $\widehat{S}(t)$ e a estimativas empírica de Kaplan-Meier, $\widehat{S_{km}}(t)$. Outros procedimentos para verificar o

Tabela 6.3: Estimativas dos parâmetros do MRCSP-WD (dados da Tabela 6.1).

Variável	Parâmetro	Estimativa	Erro Padrão	I.C.(95%)
	b	0.9963	0.0017	[0.9928, 0.9981]
	θ	1.6038	0.1555	[1.2991, 1.9086]
Idade	β_1	-2.0152	0.5248	[-3,0483, -0.9867]

Nível de referência: O paciente tem menos de 50 anos ($z_1 = 0$).

Fonte: Cardial, Cobre e Nakano (2024)

Tabela 6.4: Estimativas dos parâmetros do MRWD (dados da Tabela 6.1).

Variável	Parâmetro	Estimativa	Erro Padrão	I.C.(95%)
	θ	1.2606	0.7864	[0.8479, 1.6733]
	β_0	4.2125	0.7864	[2.8798, 5.9626]
Idade	β_1	-1.1611	0.4531	[-2,0492, -0.2729]

Nível de referência: O paciente tem menos de 50 anos ($z_1 = 0$).

Fonte: Cardial, Cobre e Nakano (2024)

ajuste do modelo podem ser vistos em Brunello e Nakano (2024). Esse erro, denotado por ε, é definido por (Nakano e Carrasco, 2006):

$$\varepsilon = \text{máx}|\widehat{S}(t) - \widehat{S_{km}}(t)|. \tag{6.3}$$

A partir da Tabela 6.5, constata-se que os três modelos possuem um bom ajuste aos dados, estando as estimativas de sobrevivência dos referidos modelos sempre próximas as estimativas empíricas.

A Figura 6.4 apresenta a estimativa da função de sobrevivência do MRLSP-WD, MRCSP-WD e MRWD para os dados da Tabela 6.1, indicando um bom ajuste desses modelos à esse conjunto de dados. No entanto, como visto pela Tabela 6.5 o MRLSP-WD proposto neste trabalho foi o que apresentou o menor erro máximo na estimação da função de sobrevivência.

Tabela 6.5: Erros máximos provenientes da estimação dos modelos MRLSP-WD, MRCSP-WD e MRWD (dados da Tabela 6.1).

Categoria da Variável	$\varepsilon MRLSP-WD$	$\varepsilon MRCSP-WD$	$\varepsilon MRWD$
$z_1 = 0$	0.1204	0.1909	0.1477
$z_1 = 1$	0.1034	0.1341	0.1738

Nível de referência: O paciente tem menos de 50 anos ($z_1 = 0$).

Fonte: Elaborado pelo autor e por Cardial, Cobre e Nakano (2024)

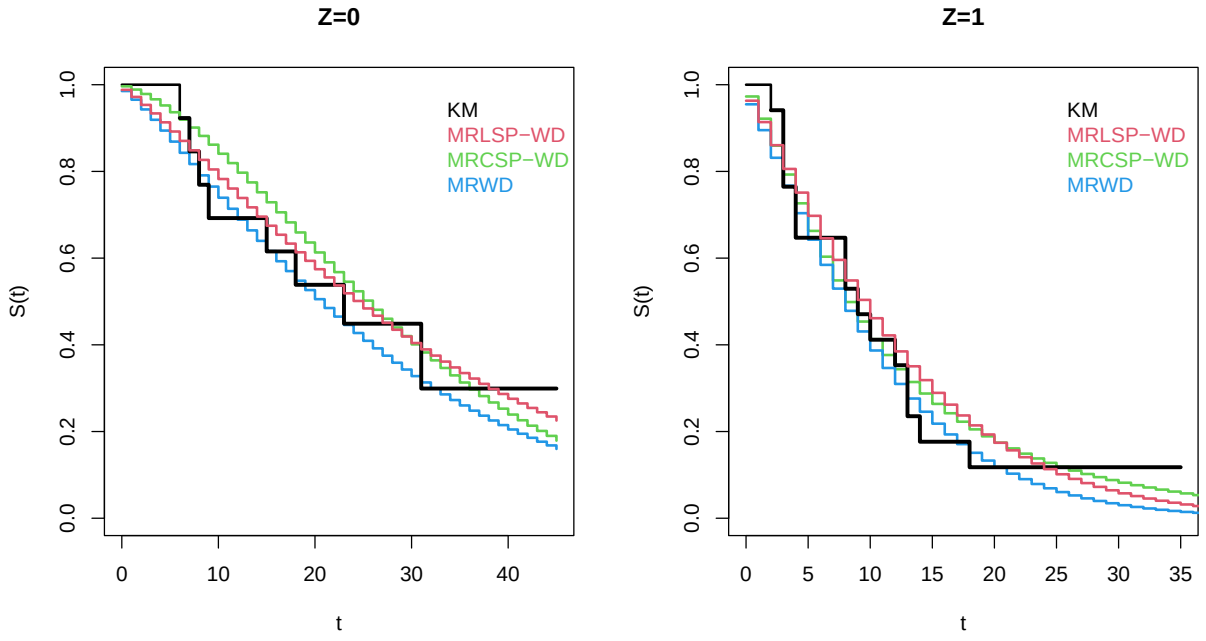


Figura 6.4: Função de sobrevivência estimada pelo MRLSP-WD, MRCSP-WD e MRWD para cada nível da covariável Idade do conjunto de dados da Tabela 6.1.

Fonte: Elaborado pelo autor

Capítulo 7

Considerações Finais

O objetivo deste trabalho foi propor um Modelo de Regressão Log de Sobrevivências Proporcionais (MRLSP) como uma alternativa aos modelos de regressão Chances de Riscos Proporcionais (Vieira et al., 2023) e Chances de Sobrevivência Proporcionais (Cardial, Cobre e Nakano, 2024) discretos. Quando a variável aleatória é contínua, o MRLSP é equivalente ao Modelo de Riscos Proporcionais, mas essa equivalência não vale para nos casos em que a variável é discreta. Assim, MRLSP pode ser visto como uma alternativa discreta aos modelos de Riscos Proporcionais (Cox, 1972).

Este trabalho apresentou a formulação do modelo, assim como procedimentos para a verificação de sua suposição (de que o log das funções de sobrevivências são proporcionais). Apesar do MRLSP poder ser aplicado para qualquer distribuição discreta como base, este trabalho focou na distribuição Weibull discreta (WD), originando assim o Modelo de Regressão Log de Sobrevivências Proporcionais Weibull Discreto (MRLSP-WD). Inferências dos parâmetros do MRLSP-WD foram realizadas partir do método de máxima verossimilhança fazendo-se o uso de métodos computacionais de otimização para a obtenção das estimativas.

Simulações computacionais foram realizadas, via software R, avaliando o viés, o Erro Quadrático Médio (EQM) das estimativas e a Probabilidade de Cobertura (PC) dos intervalos de confiança assintóticos propostos. Essas simulações consideraram três cenários (casos) e di-

ferentes tamanhos de amostra e percentuais de censura. Na ausência de censuras atestou-se que os estimadores foram aproximadamente não viesados e consistentes. Já na presença de observações censuradas, à medida que a proporção de censura aumenta, há maiores desvios das estimativas em relação ao verdadeiro valor do parâmetro. Esse comportamento é uma característica intrínseca destas estimativas (quando há censura), visto que o verdadeiro valor do parâmetro não deve ser utilizado como referência.

O MRLSP-WD foi ilustrado a partir de dados sobre o tempo de sobrevivência de pacientes com leucemia. Nesta aplicação foi realizado teste de significância da covariável considerada. Aliado ao teste supracitado, foi realizada a verificação de suposição de log de sobrevivência proporcionais dessa covariável. O MRLSP-WD apresentou um bom ajuste aos dados da aplicação, demostrando que é uma alternativa viável para modelar dados de sobrevivência com covariáveis.

Referências Bibliográficas

Barriga, G.D.C. et al. (2019). “A new survival model with surviving fraction: An application to colorectal cancer data”. Em: *Statistical methods in medical research,* 28.9, pp. 2665–2680.

Bennett, S. (1983). “Analysis of survival data by the proportional odds model”. Em: *Statistics in medicine,* 2.2, pp. 273–277.

Berguer, M. e M. Schmid (2018). “Semiparametric regression for discrete time-to-event data”. Em: *Statistical Modelling,* 18.3-4, pp. 322–345.

Biazatti, E.C. e E.Y. Nakano (2020). “Uma proposta de orientação para o uso de modelos contínuos em dados de sobrevivência discretos”. Em: *REMAT: Revista Eletrônica da Matemática,* 6.2, e4002–e4002.

Brunello, G.H.V. e E.Y. Nakano (2015). “Inferência bayesiana no modelo weibull discreto em dados com presença de censura”. Em: *TEMA - Tend. Mat. Apl. Comput.,* 16.2, pp. 97–110.

— (2024). “A Bayesian Measure of Model Accuracy”. Em: *Entropy,* 26.6, p. 510.

Cardial, M.R.P., J. Cobre e E.Y. Nakano (2024). “A discrete Weibull proportional odds survival model”. Em: *Journal of Applied Statistics,* no plero, pp. 1–19.

Cardial, M.R.P., J.B. Fachini-Gomes e E.Y. Nakano (2020). “Exponentiated discrete Weibull distribution for censored data”. Em: *Brazilian Journal of Biometrics* 38.1, pp. 35–56.

Chakraborty, S. (2015). “Generating discrete analogues of continuous probability distributions A survey of methods and constructions”. Em: *Journal of Statistical Distributions and Applications,* 2, pp. 1–30.

Ciampi, A. e J. Etezadi-Amoli (1985). "A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates". Em: *Communications in Statistics-Theory and Methods,* 14.3, pp. 651–667.

Colosimo, E.A. e S.R. Giolo (2006). *Análise de sobrevivência aplicada*. Editora Blucher.

Cox, D.R. (1972). "Regression models and life-tables". Em: *Journal of the Royal Statistical Society: Series B (Methodological),* 34.2, pp. 187–202.

Cox, D.R. e D. Oakes (1984). *Analysis of survival data*. Vol. 21. CRC press.

Ha, I.D. e G. Mackenzie (2010). "Robust frailty modelling using non-proportional hazards models". Em: *Statistical modelling,* 10.3, pp. 315–332.

Jayakumar, K. e M.G. Babu (2018). "Discrete Weibull geometric distribution and its properties". Em: *Communications in Statistics-Theory and Methods,* 47.7, pp. 1767–1783.

Kalbfleisch, J.D. e R.L. Prentice (1980). *The statistical analysis of failure time data*. John Wiley & Sons.

Kaplan, E.L. e P. Meier (1958). "Nonparametric estimation from incomplete observations". Em: *Journal of the American statistical association,* 53.282, pp. 457–481.

Lee, E.T. e J.W. Wang (2003). *Statistical Methods for Survival Data Analysis*. John Wiley & Sons.

Louzada-Neto, F. (1997). "Extended hazard regression model for reliability and survival analysis". Em: *Lifetime Data Analysis,* 3, pp. 367–381.

Magalhães, M.N. (2006). *Probabilidade e variáveis aleatórias*. Edusp.

Migon, H.S., D. Gamerman e F. Louzada (2014). *Statistical inference: an integrated approach*. CRC press.

Murphy, S.A., A.J. Rossini e A.W. van der Vaart (1997). "Maximum likelihood estimation in the proportional odds model". Em: *Journal of the American Statistical Association,* 92.439, pp. 968–976.

Nakagawa, T. e S. Osaki (1975). "The discrete Weibull distribution". Em: *IEEE transactions on reliability,* 24.5, pp. 300–301.

Nakano, E.Y. (2017). *Um curso de análise de sobrevivência*. Departamento de Estatística, Universidade de Brasília, Brasília.

Nakano, E.Y. e C.G. Carrasco (2006). "Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência". Em: *TEMA - Tend. Mat. Apl. Comput.,* 7.1, pp. 91–100.

Nicolis, J.S., G. Meyer-Kress e G. Haubs (1983). "Non-uniform chaotic dynamics with implications to information processing". Em: *Zeitschrift für Naturforschung A,* 38.11, pp. 1157–1169.

Oliveira, F.A.P. (2021). *Procedimentos de geração de dados de sobrevivência com censura à direita*. Dissertação (Mestrado) — Universidade de Brasília.

Ortega, E.M.M. et al. (2015). "A power series beta Weibull regression model for predicting breast carcinoma". Em: *Statistics in medicine,* 34.8, p.1366–1388.

Padgett, W.J. e J.D. Spurrier (1985). "On discrete failure models". Em: *IEEE Transactions on Reliability,* 34.3, pp. 253–256.

Parreira, D.R.M. (2007). *Um modelo de risco proporcional dependente do tempo*. Dissertação (Mestrado) - Universidade Federal de São Carlos.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rohatgi, V.K. e A.K.E. Saleh (2015). *An introduction to probability and statistics*. John Wiley & Sons.

Royston, P. e M.K.B. Parmar (2002). "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects". Em: *Statistics in medicine,* 21.15, pp. 2175–2197.

Sarhan, A.M. (2017). "A two-parameter discrete distribution with a bathtub hazard shape". Em: *Communications for Statistical Applications and Methods,* 24.1, pp. 15–27.

Stein, W.E. e R. Dattero (1984). "A new discrete Weibull distribution". Em: *IEEE transactions on reliability,* 33.2, pp. 196–197.

Tutz, G. e M. Schmid (2016). *Modeling discrete time-to-event data*. Springer.

Vieira, M.G.F. et al. (2023). "Proportional odds hazard model for discrete time-to-event data". Em: *Axioms,* 12.12, p. 1102.

Vila, R., E.Y. Nakano e H. Saulo (2018). "Theoretical results on the discrete Weibull distribution of Nakagawa and Osaki". Em: *Statistics,* 53.2, pp. 339–363.

Wang, L. e L. Wang (2022). "An EM algorithm for analyzing right-censored survival data under the semiparametric proportional odds model". Em: *Communications in Statistics-Theory and Methods,* 51.15, pp. 5284–5297.

Yang, S. e R.L. Prentice (1999). "Semiparametric inference in the proportional odds regression model". Em: *Journal of the American Statistical Association,* 94.445, pp. 125–136.

Zhou, J., J. Zhang e W. Lu (2022). "TransModel: An R package for linear transformation model with censored data". Em: *Journal of Statistical Software,* 101, pp. 1–12.