EXTENSÃO PARA VÁRIAS COVARIÁVEIS DO MÉTODO DE ESTIMAÇÃO DA FUNÇÃO DE RISCO INSTANTÂNEO DO MODELO ADITIVO DE AALEN

LUCIANE TEIXEIRA PASSOS GIAROLA

2009

LUCIANE TEIXEIRA PASSOS GIAROLA

EXTENSÃO PARA VÁRIAS COVARIÁVEIS DO MÉTODO DE ESTIMAÇÃO DA FUNÇÃO DE RISCO INSTANTÂNEO DO MODELO ADITIVO DE AALEN

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Doutor".

Orientador

Prof. Mario Javier Ferrua Vivanco

LAVRAS
MINAS GERAIS - BRASIL
2009

Ficha Catalográfica Preparada pela Divisão de Processos Técnicos da Biblioteca Central da UFLA

Giarola, Luciane Teixeira Passos.
Extensão para várias covariáveis do método de estimação da função de risco instantâneo do modelo aditivo de Aalen / Luciane Teixeira Passos Giarola. - Lavras : UFLA, 2009.
119p. : il.

Tese (doutorado) - Universidade Federal de Lavras, 2009.
Orientador: Mario Javier Ferrua Vivanco.
Bibliografia.

1. Modelo de Aalen. 2. Suavização. 3. Distribuições paramétricas. 4. Bootstrap. I. Universidade Federal de Lavras. II. Título.

CDD-519.536

LUCIANE TEIXEIRA PASSOS GIAROLA

EXTENSÃO PARA VÁRIAS COVARIÁVEIS DO MÉTODO DE ESTIMAÇÃO DA FUNÇÃO DE RISCO INSTANTÂNEO DO MODELO ADITIVO DE AALEN

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, para obtenção do título de "Doutor".

APROVADA em 19 de Novembro de 2009

Profa. Sueli Ruiz Giolo	UFPR
Prof. João Domingos Scalon	UFLA
Prof. Fortunato Silva de Menezes	UFLA
Prof. Marcelo Ângelo Cirillo	UFLA
Prof. Carlos Artur Lopes Leite	UFLA

Prof. Mario Javier Ferrua Vivanco
UFLA
(Orientador)

LAVRAS
MINAS GERAIS-BRASIL

Ao meu filho Matheus, por iluminar minha vida com seu sorriso, sua alegria e vivacidade e ao meu marido Marcelo, por seu amor, companheirismo e dedicação,

OFEREÇO.

Aos meus pais, Arnaldo e Gilda, pela educação e formação que me proporcionaram chegar até aqui,

DEDICO.

AGRADECIMENTOS

Tenho muito o quê e a quem agradecer:

A Deus, pela oportunidade evolutiva e pelas conquistas da minha vida, afinal, "Não cai uma folha de uma árvore sem que Deus permita".

Ao meu filho Matheus, que me trouxe a alegria e a responsabilidade de ser mãe, por ter preenchido ainda mais minha vida com seu carinho, sua alegria e seu sorriso, que em diversos momentos me confortaram e me deram forças pra lutar, na confiança de que eu possa lhe proporcionar uma educação e um futuro melhor e também por sua tranquilidade que permitiu a continuidade deste trabalho.

Ao Marcelo, meu amado, por estar comigo nesta caminhada desde o ensino médio, amparando-me e possibilitando a realização de mais este sonho; por seu amor, carinho, compreensão, lealdade, dedicação e admiração; por acreditar em mim e por confortar-me e incentivar-me nos momentos difíceis. Esta vitória é nossa!

Aos meus pais, Arnaldo e Gilda, pelo amor, carinho, dedicação e apoio; por me ensinarem valores como honestidade, integridade, solidariedade, persistência, determinação e por me proporcionarem a formação que me permitiu chegar até aqui.

Aos meus irmãos e familiares, pelo carinho e amizade, pelas orações, pela torcida, pelo incentivo e por compreenderem minha ausência em momentos nos quais poderia desfrutar um pouco mais de suas companhias.

À Maria pelo carinho, zelo e capricho com que cuida do Matheus, pela dedicação, bondade e disponibilidade que me proporcionaram confiança, elemento fundamental para a realização deste trabalho.

À Viviane por cuidar da minha casa por mim e pelo carinho que tem

pelo meu filho.

À Clea, por sua simplicidade e disponibilidade; pelo apoio em momentos tão delicados e difíceis da minha vida.

À Angélica, pela amizade e confiança conquistadas; pela compreensão e paciência comigo e minha família, estando sempre pronta a nos ajudar.

À família Milton Teixeira pela certeza da amizade, do apoio, das orações e da torcida.

À Rô pelo apoio emocional, por me ajudar a tentar ser e fazer cada dia melhor, pela amizade.

À Rejane, pela amizade e afinidade, por compartilhar e lutar comigo pelos mesmos sonhos e objetivos, pelo incentivo e apoio, pelo companheirismo e parceria.

Aos amigos da pós graduação, pelo convívio, tanto nas horas de trabalho, como nas horas de lazer. Especialmente Verônica, Graziela e Fabrício pelo apoio.

Ao Prof Mario por me apresentar a Análise de Sobrevivência, pela orientação, pelos conhecimentos e pelo apoio, pela confiança, pela parceria e por me ensinar a ser mais independente.

Ao Prof Marcelo Cirillo pela oportunidade de convívio, aprendizado e amizade, pela confiança em mim depositada e pelo apoio computacional.

Ao professor Carlos Artur pela disponibilidade em me receber e em ceder os dados utilizados neste trabalho.

Aos professores do DEX-UFLA que contribuíram para minha formação e às funcionárias pela boa vontade e atenção, especialmente Josi (da pós), Maria, Selminha e Edila.

À Fapemig pelos dois anos de concessão de bolsa de estudos.

SUMÁRIO

LISTA DE TABELAS

LISTA DE FIGURAS

RESUMO

GIAROLA, Luciane Teixeira Passos. **Extensão para várias covariáveis do método de estimação da função de risco instantâneo do modelo aditivo de Aalen**. 2009. 119 p. Tese (Doutorado em Estatística e Experimentação Agropecuária) − Universidade Federal de Lavras, Lavras, MG.*

O modelo aditivo de Aalen avalia o risco de ocorrência de determinado evento ao longo do tempo. O modelo é não paramétrico e estima apenas os riscos acumulados. Portanto, não se tem nenhuma informação a respeito do risco de ocorrência do evento em determinado instante. Assim, neste trabalho, desenvolveu-se um método de estimação da função de risco instantâneo considerando diversas covariáveis, utilizando-se funções de risco acumuladas obtidas de distribuições paramétricas para suavizar as curvas de riscos acumulados do modelo, estratificando os dados em parcelas. Para verificar a adequacidade dos riscos acumulados paramétricos aos riscos acumulados não paramétricos, implementou-se um teste de significância por meio de métodos de computação intensiva. Concluiu-se que na presença de duas covariáveis dicotômicas, os riscos acumulados do modelo de Aalen podem ser estimados estratificando o conjunto de dados e que tal metodologia pode ser utilizada em um conjunto de dados com qualquer quantidade de covariáveis. Concluiu-se, ainda, que o teste de adequacidade proposto controla o erro tipo I, para diferentes proporções médias de censura ($p = 0,30; 0,20; 0,10$) e diferentes tamanhos amostrais ($n = 30, 50, 60, 90$), sendo mais conservativo para pequenas proporções médias de censura.

*Comitê Orientador: Mario Javier Ferrua Vivanco − UFLA (Orientador) e Marcelo Ângelo Cirillo − UFLA

ABSTRACT

GIAROLA, Luciane Teixeira Passos. **Extension to multiple covariates in the estimation of instantaneous risk function of the Aalen additive model**. 2009. 119 p. Thesis (Doctoral in Statistics and Agricultural Experimentation) - Universidade Federal de Lavras, Lavras, MG. *

The Aalen additive model assesses the risk of occurrence of an event over time. The model is non parametric and estimates only accumulated risks. So be has no information regarding the risk of occurrence of the event at a given time. Thus, this study developed a estimation of the instantaneous risk function considering several covariates, using accumulated risk functions obtained from parametric distributions to smooth cumulative risk by stratifying the data parcel. To check goodness of fit of the parametric curves cumulative risks to nonparametric cumulative risks, it was implemented a test through compute-intensive methods. It was concluded that in presence of two dichotomous covariates, the cumulative risk of model Aalen can be estimated stratifying the data set and that this methodology can be used in a data set with any number of covariates. It also was concluded that the test proposed of adequacy controls type I error for different average proportions censorship ($p = 0.30, 0.20, 0.10$) and different sample sizes ($n = 30, 50, 60, 90$), being more conservative for small average proportions of censorship.

*Guidance Committee: Mario Javier Ferrua Vivanco − UFLA (Adviser) and Marcelo Ângelo Cirillo − UFLA

1 INTRODUÇÃO

O principal foco da análise de sobrevivência está na função de risco. Para os pesquisadores é importante saber qual o risco de um indivíduo experimentar determinado evento. Há vários modelos que permitem estudar este risco e os fatores que nele influenciam. Os modelos dividem-se em três classes: paramétricos, semiparamétricos e não paramétricos.

Neste estudo o interesse está no modelo não paramétrico proposto por Aalen (1980). A principal característica deste modelo é que ele permite avaliar o efeito das covariáveis no risco de falha ao longo do tempo. Isto é feito por meio de funções de regressão. Assim, neste modelo não se estimam parâmetros e sim funções de regressão acumuladas, visto que o modelo é não paramétrico. O risco de um indivíduo experimentar o evento é modelado a partir dessas funções de regressão. Na verdade, cada função de regressão corresponde à diferença entre os riscos de dois grupos, isto é, cada função de regressão representa o risco adicional de falha de um grupo em relação a um outro grupo de referência. Logo, ao se estimar uma função de regressão, estima-se um risco de falha (adicional). Então, o modelo estima riscos acumulados, os quais não informam o risco de falha que um indivíduo possui em determinado instante de tempo.

Grunkemeier et al. (2006) desenvolveram um estudo para estimar a função de risco instantâneo adicional considerando apenas uma covariável. Os autores suavizaram os riscos acumulados do modelo de Aalen usando uma distribuição paramétrica conhecida: a distribuição Gompertz. Neste trabalho pretendeu-se estender tal estudo considerando diversas covariáveis.

Mais especificamente, pretendeu-se estimar a função de risco ins-

tantâneo do modelo de Aalen considerando inicialmente duas covariáveis e, posteriormente, generalizar a metodologia proposta para várias covariáveis. Para isto foram utilizadas distribuições paramétricas conhecidas para suavizar curvas de risco acumulado não paramétricas e, a partir da derivada da diferença entre elas, determinar a função de risco instantâneo. Pretendeu-se, ainda, desenvolver um teste de significância que permita avaliar quão bem a distribuição paramétrica utilizada suaviza a função de risco acumulado não paramétrico; obter uma expressão matemática para as funções de regressão instantâneas do modelo de Aalen para, por meio dela, poder predizer o risco adicional instantâneo de um grupo em relação a outro; validar a metodologia proposta utilizando dados simulados e aplicá-la a um conjunto de dados reais.

2 REFERENCIAL TEÓRICO

2.1 Análise de sobrevivência

2.1.1 Conceitos básicos

Em análise de sobrevivência, a variável resposta é, em geral, o tempo até a ocorrência de um evento de interesse (falha). Este tempo é dito tempo de falha, tempo de sobrevida ou de sobrevivência.

Uma das principais características de dados de sobrevivência é a presença de censuras. Censuras são observações parciais ou incompletas da resposta. São classificadas em à direita, à esquerda e intervalares. As censuras à direita podem ser do tipo I, do tipo II ou aleatória. A ocorrência de censura é, geralmente, representada por meio de uma variável indicadora de falha δ_i, a qual vale 1 se a observação falhou, isto é, se o evento ocorreu, e 0 se foi censurada. O índice i refere-se à observação (indivíduo). Para maiores detalhes os trabalhos de Allison (1995), Carvalho et al. (2005) e Colosimo & Giolo (2006) devem ser consultados.

Outra característica de dados de sobrevivência é a presença de covariáveis dependentes do tempo , isto é, variáveis medidas para cada observação ao longo do tempo e que podem interferir no tempo de falha.

O tempo de falha é usualmente representado por uma variável aleatória contínua não negativa T e apresenta forte assimetria, com uma grande cauda à esquerda. Isto ocorre porque grande parte destes tempos tem valores pequenos com poucas observações em tempos muito longos. Assim, em análise de sobrevivência, cada indivíduo é representado no banco de dados pelo par (T_i, δ_i), sendo T_i o tempo de falha ou censura do i-ésimo indivíduo e δ_i a variável indicadora de falha.

Esta variável T é, geralmente, especificada pela sua função de sobrevivência ou pela função de taxa de falha, também chamada função risco. A função de sobrevivência *S(t)* é definida como a probabilidade de uma observação não falhar até um certo tempo t. Matematicamente,

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

Em consequência, a função de distribuição acumulada $F(t)$ é definida como a probabilidade de uma observação falhar até um tempo t, isto é,

$$F(t) = 1 - S(t) \tag{2.1}$$

A função risco de T, denotada por $\lambda(t)$, descreve a forma com que a taxa instantânea de falha muda com o tempo. Ela é o limite do quociente entre a probabilidade de que o evento de interesse ocorra no pequeno intervalo de tempo $[t, t + \Delta(t)\big)$, condicionada à sobrevivência no início do intervalo, e o comprimento do intervalo, conforme a fórmula:

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

Segundo Fogo (2007), a função de risco é uma medida da propensão de falha como uma função da idade, no sentido de que a quantidade $\lambda(t)\Delta t$ é a proporção esperada de observações, com idade t, que poderão falhar em um pequeno intervalo de tempo $[t, t + \Delta t\big)$. Isto é, a quantidade $\lambda(t)\Delta t$ representa o risco de ocorrência do evento no intervalo $[t, t + \Delta t)$, sendo que este evento não ocorreu no intervalo $[0, t)$. Dessa forma, a função risco é uma taxa e não uma probabilidade. Tal função é bastante útil para se determinar a distribuição dos tempos de vida e para descrever o modo com

que a chance de ocorrência de um evento muda com o tempo. Ela também pode ser obtida pela relação

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{2.3}$$

sendo $f(t)$ a função de densidade de probabilidade de T.

Observam-se diferentes tipos de funções risco de acordo com a forma do gráfico de $\lambda(t)$. A Figura 1 apresenta três funções de taxa de falha: crescente, constante e decrescente. Modelos com função de risco crescente, por exemplo, surgem naturalmente em consequência do envelhecimento ou desgaste de material. É muito comum no final da vida das pessoas ou da vida útil de peças e equipamentos. Pode ser observada em pacientes em fase terminal e que não respondem mais ao tratamento. Modelos com função de risco constante não são comuns. São encontrados durante períodos isolados do tempo de vida de indivíduos ou equipamentos. Os modelos com função de risco decrescente também são menos comuns, podendo ser encontradas em situações nas quais há uma grande chance de falha nos períodos de tempo iniciais, como em alguns dispositivos eletrônicos, em pacientes que se recuperam de cirurgias simples ou em bebês nos primeiros anos de vida. Há ainda modelos que combinam as curvas apresentadas na Figura 1 em diferentes períodos de tempo como, por exemplo, a *curva da banheira*.

Outra função útil na análise de dados de sobrevivência é a função de taxa de falha acumulada, denotada por $\Lambda(t)$. Como o próprio nome sugere, ela representa o acúmulo de risco instantâneo ao longo do tempo; o risco de falha acumulado até o instante t, isto é, o risco de ocorrência do evento de

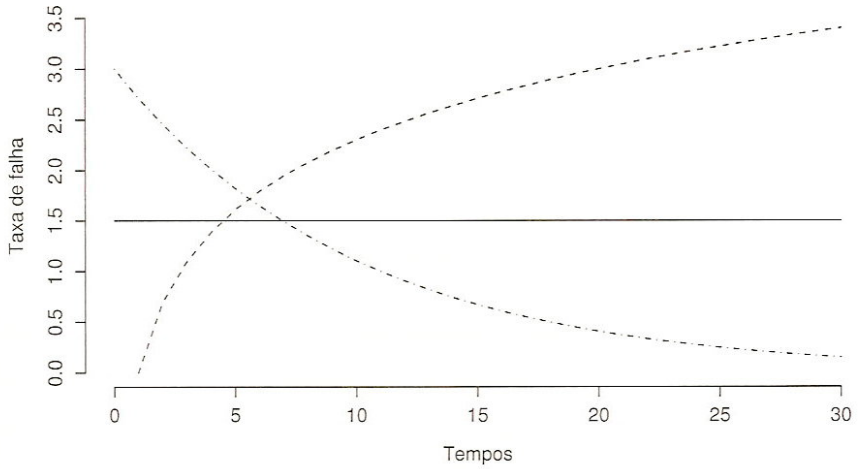

FIGURA 1 Funções de taxa de falha: - - - crescente, — constante e - · - · - decrescente.

interesse no intervalo $[0, t)$. Assim,

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

Tal função não apresenta uma interpretação direta como $\lambda(t)$, mas pode ser útil na avaliação de $\lambda(t)$. Isto ocorre na estimação não paramétrica, na qual $\lambda(t)$ é difícil de se estimar, como por exemplo no modelo aditivo de Aalen. A função de risco acumulado $\Lambda(t)$ é muito útil na seleção de modelos. O gráfico da estimativa desta função é utilizado para verificar a adequacidade de diversos modelos.

A função de taxa de falha acumulada relaciona-se com a função de sobrevivência através da expressão

$$\Lambda(t) = -\ln S(t) \tag{2.5}$$

São várias as formas de se modelar dados de sobrevivência. Há os modelos probabilísticos, como a distribuição Weibull, os modelos de regressão paramétricos, como os modelos de tempo de falha acelerado (AFT), os modelos semiparamétricos, como o modelo de riscos proporcionais proposto por Cox (1972) e os modelos não paramétricos, como o modelo proposto por Aalen (1980). Nesta tese,foram utilizados os modelos probabilísticos Weibull e Lognormal e o modelo de Aalen, os quais serão descritos posteriormente. O modelo de Aalen é construído utilizando-se a teoria de processos de contagem. Também serão necessários métodos não paramétricos para se estimar a função de risco acumulado.

2.1.2 Processos de contagem

A teoria de processos de contagem é muito importante em estudos de análise de sobrevivência, pois permite acomodar tanto censuras quanto situações mais complexas como, por exemplo, as que envolvem covariáveis dependentes do tempo , eventos recorrentes e eventos múltiplos. Neste texto será descrita tal teoria de forma simples para que se possa entender como os grupos de risco são formados a cada momento do tempo , o que é essencial para acomodar covariáveis tempo dependentes na organização dos bancos de dados.

Um processo de contagem é definido como um processo estocástico $N(t)$ com $t \geq 0$, de forma que $N(0) = 0$ e $N(t) < \infty$. A função $N(t)$ conta o número de eventos observados no intervalo $(0, t]$. Em um intervalo qualquer $(a, b]$, o número de eventos observados é definido por

$$N(a, b] = N(b) - N(a).$$

Para a i-ésima observação tem-se o processo de contagem univariado $N_i(t)$, o qual, segundo Fleming & Harrington (1991), é uma função escada com saltos de tamanho 1 e sua trajetória é contínua à direita. Um processo univariado está representado na Figura 2.

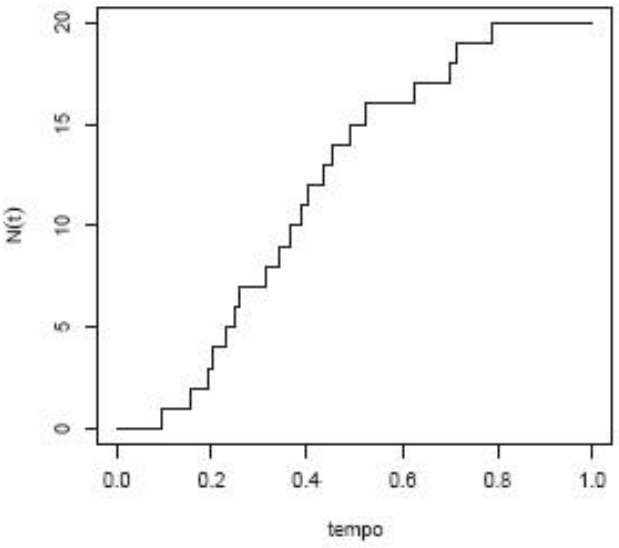

FIGURA 2 Representação gráfica de um processo de contagem univariado.

Segundo Aalen (1978),um processo de contagem multivariado dado por

$$N(t) = (N_1(t), N_2(t), ..., N_i(t), ..., N_n(t))',$$

é uma coleção de n processos de contagem univariados, os quais podem ser dependentes um do outro. Supõe-se, com probabilidade 1, que dois componentes $N_i(t)$ e $N_j(t)$ do processo de contagem não saltam ao mesmo tempo,

isto é, supõe-se que as observações apresentam tempos de falha diferentes, não há empates. Mas, em situações nas quais há um pequeno número de empates, pode-se somar um número aleatório entre 0 e 1 ao tempo de falha a fim de provocar o desempate (Mau, 1986; Aalen, 1989; Henderson & Milner, 1991).

A análise de sobrevivência pode ser pensada como um processo de contagem multivariado no qual $N(t)$ é o número de eventos observados até o tempo t e $\Delta N_i(t)$ é a diferença entre a contagem de eventos até o instante t e a contagem no momento imediatamente anterior a t, para o i-ésimo indivíduo. Em estudos que envolvem a ocorrência de um único evento uma única vez, $N_i(t)$ só pode assumir os valores 0 ou 1. Nestes casos,

$N_i(t) = 0$, se o indivíduo i não experimentou o evento até o tempo t e
$N_i(t) = 1$ se o indivíduo i sofreu o evento.

Assim, sob a óptica de processos de contagem, cada observação é representada no banco de dados pelo par de funções $(N_i(t), Y_i(t))$, sendo $N_i(t)$ o processo de contagem univariado (do i-ésimo indivíduo) do número de eventos observados no intervalo $[0, t]$ e $Y_i(t)$ indica se no tempo t o indivíduo estava sob risco de sofrer o evento. Logo,

$Y_i(t) = 1$, se o indivíduo i estiver sob risco no tempo t e
$Y_i(t) = 0$ se o indivíduo i sofreu o evento ou foi censurado.

Para exemplificar, considere a trajetória de dois indivíduos representada na Figura 3, sendo que um deles experimentou o evento e o outro foi

censurado.

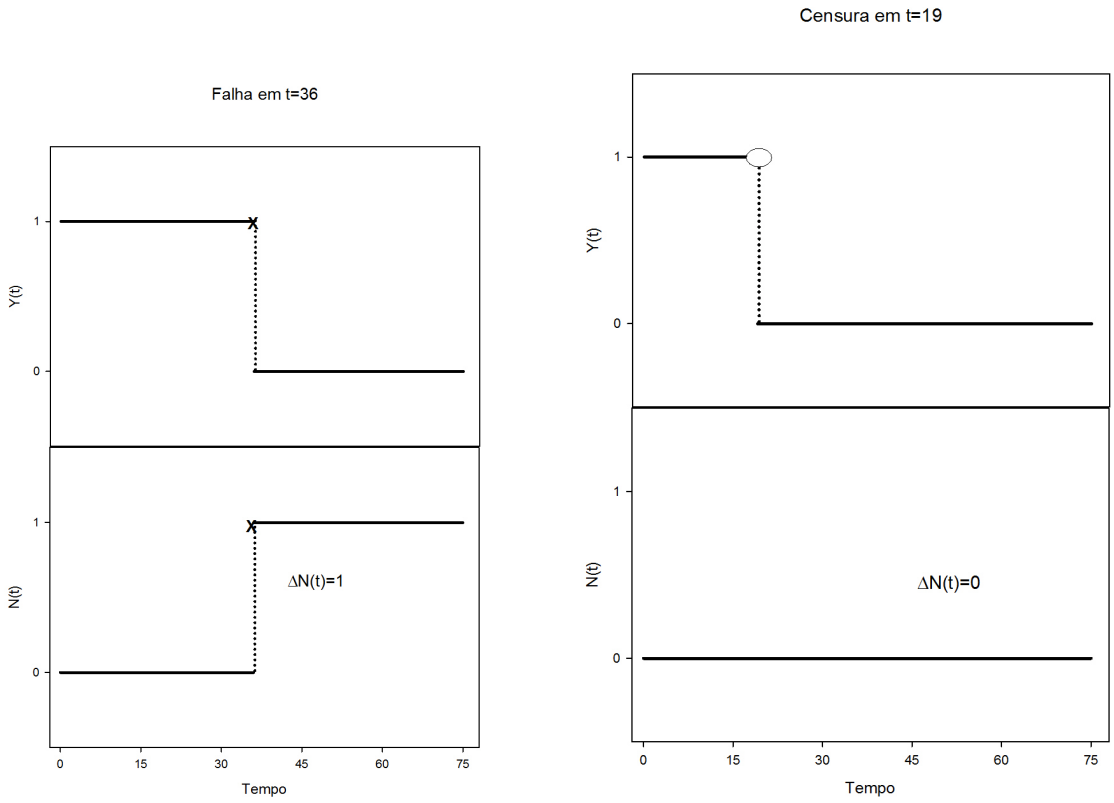

FIGURA 3 Trajetória de dois indivíduos: o primeiro experimentou o evento e o segundo foi censurado.

O primeiro indivíduo experimentou o evento em $t = 36$, logo, até este instante de tempo, encontrava-se em risco. Então, até $t = 36$, tem-se $N_1(t) = 0$ e $Y_1(t) = 1$. No instante em que ocorre o evento o processo de contagem salta para 1 e o indivíduo não mais se encontra em risco. Assim, $N_1(t) = 1$ e $Y_1(t) = 0$. Já o segundo indivíduo, pode ser representado pelo par $(0, 1)$ até o instante $t = 19$, quando então é censurado e $Y_2(t)$ passa a assumir o valor 0, pois o indivíduo não se encontra mais sob risco. Note que $N_2(t)$ permanece nulo por todo o tempo, pois o indivíduo não experimentou

o evento.

O caráter temporal de um processo de contagem exige o conhecimento dos indivíduos em cada instante $t \in [0, \infty)$. Esse conhecimento resume-se na informação a respeito dos instantes de ocorrência dos eventos e também, em alguns estudos, em informações adicionais sobre esses indivíduos no instante t. Isto é, consiste na informação a respeito dos pares $(N_i(t), Y_i(t))$ e em valores de um conjunto fixo de covariáveis, independentes ou não do tempo, em cada instante t. Esse conhecimento sobre o que aconteceu aos indivíduos até o instante t é chamado *filtragem* ou *história* do processo de contagem no instante t e será denotado por F_t. A história de um processo estocástico até um instante imediatamente anterior a t é denotada por F_t^- e representa os dados disponíveis, referentes ao conjunto de tempos de falhas e censuras, até um tempo $t^- < t$.

A estrutura probabilística de um processo de contagem pode ser descrita pelo processo de intensidade multivariado $\lambda(t) = (\lambda_1(t), \lambda_2(t), ..., \lambda_n(t))'$, sendo $\lambda_i(t)$ dado por

$$\lambda_i(t) = \lim_{\Delta(t) \to 0} \frac{P[N(t + \Delta(t)) \geq 1 | F_t^-]}{\Delta(t)}, \tag{2.6}$$

Assim, $\lambda_i(t)\Delta(t)$ representa a proporção esperada de ocorrência do evento de interesse (não necessariamente o primeiro) no pequeno intervalo de tempo $[t, t + \Delta(t))$, dado que no intervalo de tempo $[0, t)$, aconteceram F_t^- eventos. Logo, a predição do que acontecerá em tempos futuros depende do que aconteceu anteriormente.

É importante comparar a função intensidade (2.6) com a função risco (2.2) definida na seção 2.1.1. A função risco foca a ocorrência de um único evento. Já a função intensidade permite acomodar situações nas quais o

evento não é único ou ocorre várias vezes, sendo, portanto, mais ampla e mais geral que a outra. Nos casos em que o evento em questão é único e ocorre apenas uma vez, isto é, se não há eventos recorrentes, o produto entre a função intensidade $\lambda_i(t)$ e $\Delta(t)$ representa a proporção esperada de ocorrência do evento de interesse no pequeno intervalo de tempo $[t, t+\Delta(t))$, dado que o evento não ocorreu no intervalo de tempo $[0, t)$. Assim, apesar de terem expressões diferentes, as funções risco e intensidade possuem a mesma interpretação.

Como já mencionado anteriormente, dados de sobrevivência podem ser modelados de forma não paramétrica. A função risco acumulado, por exemplo, pode ser estimada por meio da estimativa Kaplan Meier da função de sobrevivência e da relação (2.5). Esta função de risco acumulado é de grande interesse para esta tese, pois será utilizada para medir o risco instantâneo de falha de um indivíduo.

2.1.3 Estimador de Kaplan-Meier

O estimador mais utilizado em estudos clínicos foi proposto por Kaplan & Meier (1958) e também ficou conhecido como estimador produto limite devido à forma como foi construído. Este estimador não paramétrico será utilizado para estimar a sobrevivência de indivíduos e, posteriormente, o risco acumulado de falha destes indivíduos.

Tal estimador é uma adaptação da função de sobrevivência empírica, a qual é uma função escada com saltos de tamanho $\frac{1}{n}$ nos tempos de falha, para uma amostra de tamanho n. Em sua construção, ele considera tantos intervalos de tempo quantos forem o número de falhas distintas. Os limites dos intervalos de tempo são os tempos de falha da amostra.

Considerando que em uma amostra de tamanho n existam $k(\leq n)$ falhas distintas nos tempos $t_1 < t_2 < t_3 < ... < t_k$, o estimador da função de sobrevivência proposto por Kaplan & Meier (1958) $\left(\widehat{S}_{KM}(t)\right)$ é o produto das probabilidades de sobrevivência nos intervalos $[t_{j-1}, t_j)$, $j = 1, ..., k$, sendo que o indivíduo sobreviveu até o tempo t_j, considerando $t_0 = 0$. Tais probabilidades de sobrevivência são estimadas pelo quociente entre o número de sobreviventes até o tempo t_j e o número de indivíduos em risco no tempo t_j. Segundo Carvalho et al. (2005), em linguagem de processos de contagem, o estimador de Kaplan-Meier pode ser escrito como

$$\widehat{S}_{KM}(t) = \left(\frac{R(t_1) - \Delta N(t_1)}{R(t_1)}\right)\left(\frac{R(t_2) - \Delta N(t_2)}{R(t_2)}\right)\ldots\left(\frac{R(t_k) - \Delta N(t_k)}{R(t_k)}\right),$$

ou seja,

$$\widehat{S}_{KM}(t) = \prod_{j:t_j \leq t}\left(\frac{R(t_j) - \Delta N(t_j)}{R(t_j)}\right), \tag{2.7}$$

sendo,

$R(t_j), j = 1, ..., k$, o número de indivíduos em risco no tempo t_j $\left(R(t_j) = \sum_{i=1}^{n} Y_i(t_j)\right)$,

$\Delta N(t_j) = \sum_{i=1}^{n} \Delta N_i(t_j)$, com $\Delta N_i(t_j)$ sendo a diferença entre a contagem de eventos até o instante t_j e a contagem no momento imediatamente anterior a t_j para o i-ésimo indivíduo.

No caso em que o evento é único e ocorre apenas uma vez, o termo $\Delta N(t_j)$ é equivalente a $N(t_j)$, com $N(t_j) = \sum_{i=1}^{n} N_i(t_j)$, isto é, o número de eventos observados no instante t_j.

A estimativa de Kaplan Meier da função de sobrevivência pode ser utilizada para estimar a função de risco acumulado $\Lambda_{KM}(t)$, obtendo-se, assim, uma estimativa não paramétrica para esta função.

Como a proposta desta tese é utilizar distribuições paramétricas para suavizar funções de risco acumulado não paramétricas, é necessário estudar as principais distribuições utilizadas em análise de sobrevivência e o método de estimação de seus parâmetros.

2.1.4 Risco acumulado paramétrico para dados de sobrevivência

Para a análise estatística de dados de sobrevivência podem ser utilizadas distribuições de probabilidade. Há vários modelos probabilísticos utilizados para descrever tempos de vida. Dentre eles destacam-se o modelo Exponencial, o modelo Weibull, o modelo Lognormal, o modelo Gama e o modelo Gama Generalizada, devido à comprovada adequação dos mesmos a diversas situações práticas.

Neste tipo de modelagem, considera-se que o tempo T até o evento segue uma distribuição conhecida de probabilidade e estimam-se os parâmetros correspondentes pelo método da máxima verossimilhança. A seguir serão descritas as distribuições mais utilizadas em análise de sobrevivência: Exponencial, Weibull, Lognormal, Gama e Gama Generalizada.

Distribuição Exponencial

Segundo Carvalho et al. (2005), a distribuição exponencial foi a mais utilizada historicamente na modelagem de tempo de sobrevivência, devido à sua simplicidade. Ela apresenta um único parâmetro e caracteriza-se por ter uma função de risco constante. Isto caracteriza a falta de memória desta distribuição, ou seja, dois indivíduos (um novo e um velho) que ainda não experimentaram o evento, têm o mesmo risco de falhar em um tempo futuro. Esta é uma situação rara na prática.

Considerando que a variável T possui uma distribuição exponencial, sua função densidade de probabilidade $f(t)$ é dada por

$$f(t) = \frac{1}{\alpha} exp\left[-\left(\frac{t}{\alpha}\right)\right], \quad t \geq 0, \tag{2.8}$$

sendo $\alpha > 0$ o tempo médio de vida. Sua função de distribuição acumulada é

$$F(t) = 1 - exp\left[-\left(\frac{t}{\alpha}\right)\right].$$

As funções de sobrevivência, taxa de falha e risco acumulado, as quais podem ser obtidas através de (2.1), (2.3) e (2.5), respectivamente, são dadas por

$$\begin{aligned} S(t) &= exp\left[-\left(\frac{t}{\alpha}\right)\right], \\ \lambda(t) &= \frac{1}{\alpha}, \\ \Lambda(t) &= \frac{t}{\alpha}. \end{aligned}$$

Distribuição Weibull

A distribuição Weibull é atualmente a mais utilizada para modelar dados de sobrevivência, talvez devido à flexibilidade de sua função risco. Ela possui dois parâmetros, um de forma e outro de escala.

Para uma variável aleatória T com distribuição de Weibull, tem-se:

$$\begin{aligned} f(t) &= \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} exp\left[-\left(\frac{t}{\alpha}\right)^\gamma\right], \quad t \geq 0, \qquad (2.9) \\ S(t) &= exp\left[-\left(\frac{t}{\alpha}\right)^\gamma\right], \\ \lambda(t) &= \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}, \\ \Lambda(t) &= \left(\frac{t}{\alpha}\right)^\gamma, \qquad (2.10) \end{aligned}$$

sendo γ o parâmetro de forma e α o de escala, ambos positivos.

A distribuição Weibull apresenta função risco monótona. Se $\gamma < 1$ a função risco é decrescente; se $\gamma > 1$ a função risco é crescente e se $\gamma = 1$ a função risco é constante. Neste último caso, a distribuição Weibull equivale à distribuição Exponencial.

Distribuição Lognormal

Esta distribuição também pode ser utilizada para caracterizar tempos de vida de produtos e indivíduos. Assumindo-se que a variável aleatória T segue uma distribuição Lognormal, isto é, que o logaritmo de T segue uma distribuição Normal com parâmetros μ e σ^2, tem-se:

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{2\pi} t\sigma} exp\left[-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right], \quad t > 0, \qquad (2.11) \\ S(t) &= \Phi\left(\frac{-log(t)+\mu}{\sigma}\right). \end{aligned}$$

Allison (1995) afirma que o comportamento da função de risco não é monótono. Ela cresce, atinge um valor máximo e depois decresce. Esse decréscimo

para valores grandes de T é pouco plausível na maioria das situações.

Distribuição Loglogística

A distribuição Loglogística se aplica a diversas situações práticas. Para uma variável aleatória T com esta distribuição, as funções de densidade, sobrevivência e risco instantâneo são expressas por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \left[1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right]^{-2} \tag{2.12}$$

$$S(t) = \frac{1}{1 + \left(\frac{t}{\alpha}\right)^{\gamma}}. \tag{2.13}$$

$$\lambda(t) = \frac{\gamma\left(\frac{t}{\alpha}\right)^{\gamma-1}}{\alpha\left[1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right]}.$$

De acordo com Colosimo & Giolo (2006), a função de risco desta distribuição apresenta um padrão similar ao da distribuição Lognormal para $\gamma > 1$.

Distribuição Gama

Por se ajustar a uma variedade de fenômenos, a distribuição Gama tem sido utilizada para descrever o tempo de vida de produtos e, mais recentemente, de pacientes. Também tem sido utilizada em situações que envolvem efeitos aleatórios como, por exemplo, em modelos de fragilidade.

Se a variável aleatória T segue uma distribuição Gama, então

$$\begin{aligned} f(t) &= \frac{1}{\Gamma(k)\alpha^k} t^{k-1} exp\left(-\frac{t}{\alpha}\right), \quad t > 0, \\ S(t) &= \int_t^\infty \frac{1}{\Gamma(k)\alpha^k} u^{k-1} exp\left(-\frac{u}{\alpha}\right) du, \end{aligned} \tag{2.14}$$

sendo k e α os parâmetros de forma e escala respectivamente, ambos positivos, e $\Gamma(k)$ a função gama definida por $\Gamma(k) = \int_0^\infty x^{k-1} exp(-x) dx$.

De acordo com Lawless (1982), a função taxa de falha para essa distribuição apresenta um padrão crescente ou decrescente convergindo para um valor constante quando t cresce de 0 a infinito.

Observe-se que tomando $k = 1$ em (2.14), obtém-se o modelo exponencial. Assim, a distribuição exponencial é um caso particular da distribuição gama.

Distribuição Gama Generalizada

A distribuição Gama Generalizada caracteriza-se por apresentar três parâmetros, um de escala e dois de forma, além de incluir, para determinados valores paramétricos, as distribuições Exponencial, Weibull, Lognormal e Gama, o que a torna muito útil.

Para uma variável aleatória T com distribuição Gama Generalizada, a densidade de probabilidade $f(t)$ é:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{\gamma k - 1} exp\left[-\left(\frac{t}{\alpha}\right)^\gamma\right], \quad t > 0, \tag{2.15}$$

sendo α o parâmetro de escala, k e γ parâmetros de forma e $\Gamma(k)$ a função gama.

De acordo com os valores dos parâmetros da distribuição Gama Generalizada, obtêm-se as distribuições Exponencial, Weibull, Lognormal e Gama. Para $\gamma = k = 1$, T segue a distribuição Exponencial com parâmetro α. Para $k = 1$, T segue a distribuição Weibull com parâmetros α e γ. Para $\gamma = 1$ tem-se que T apresenta distribuição Gama (k, α). Já a distribuição Lognormal surge como um caso limite da distribuição Gama Generalizada quando $k \to \infty$ (Lawless, 1982; Valença, 1994).

Há outros modelos, como por exemplo os modelos Gompertz, Binomial, Normal, etc. Mas os mais frequentes em dados de tempos de vida são os mencionados anteriormente.

Estimação dos Parâmetros dos Modelos

De acordo com o método de máxima verossimilhança, supondo-se que a variável aleatória T apresenta função densidade de probabilidade $f(t)$ e considerando-se que não há dados censurados, a função de verossimilhança $L(\theta)$ para os tempos observados é dada por

$$L(\theta) = \prod_{i \in O} f(t_i; \theta), \tag{2.16}$$

sendo O o conjunto de dados observados até a ocorrência do evento e θ o vetor de parâmetros a serem estimados. Então, a contribuição de cada indivíduo não censurado é a sua função densidade de probabilidade. Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$, ou de modo equivalente, que maximizam o logaritmo de $L(\theta)$. Para obtê-los resolve-se o sistema de equações $\frac{\partial log L(\theta)}{\partial \theta} = 0$.

Para um conjunto de dados com observações censuradas o método de

estimação de máxima verossimilhança é o mesmo, porém a função de verossimilhança é modificada, sendo então denominada função de verossimilhança parcial. Neste caso, para a construção da função de verossimilhança, supõe-se que os tempos de censura são independentes dos tempos de falha, isto é, as censuras são não informativas, pois não estão relacionadas ao evento em questão.

Quando os dados são censurados à direita, sabe-se apenas que o tempo de falha é maior que o tempo observado (t_+), ou seja, $S(t_+) = P(T > t_+)$. Assim, a contribuição de indivíduos censurados à direita (D) para $L(\theta)$ é a sua função de sobrevivência. Logo,

$$L(\theta) = \prod_{i \in O} f(t_i; \theta) \prod_{i \in D} S(t_{i+}; \theta),$$

que também pode ser escrita como

$$L(\theta) = \prod_{i=1}^{n} \big(f(t_i; \theta)\big)^{\delta_i} \big(S(t_i; \theta)\big)^{1-\delta_i},$$

sendo δ_i a variável indicadora de falha, definida na seção 2.1.1

Para dados censurados à esquerda, sabe-se que o evento ocorreu antes da data de início do estudo. Considerando t_- o tempo de sobrevida registrado para o indivíduo censurado à esquerda, $P(T \leq t_-) = F(t_-) = 1 - S(t_-)$. Desse modo, sendo E o conjunto dos indivíduos censurados à esquerda,

$$L(\theta) = \prod_{i \in O} f(t_i; \theta) \prod_{i \in E} (1 - S(t_{i-}; \theta)),$$

ou

$$L(\theta) = \prod_{i=1}^{n} \big(f(t_i; \theta)\big)^{\delta_i} \big(1 - S(t_i; \theta)\big)^{1-\delta_i},$$

No caso de censuras por intervalo, sabe-se apenas que a censura ocorreu dentro do referido intervalo; então $P(t_- \leq T \leq t_+) = F(t_+) - F(t_-) = S(t_-) - S(t_+)$. Logo,

$$L(\theta) = \prod_{i \in O} f(t_i; \theta) \prod_{i \in I} (S(t_{i-}; \theta) - S(t_{i+}; \theta)),$$

sendo I o conjunto de indivíduos censurados por intervalo. Ou

$$L(\theta) = \prod_{i=1}^{n} \big(f(t_i; \theta)\big)^{\delta_i} \big(S(t_{i-}; \theta) - S(t_{i+}; \theta)\big)^{1-\delta_i},$$

Em algumas situações, a solução do sistema de equações $\frac{\partial log L(\theta)}{\partial \theta} = 0$, para um conjunto de dados particular, deve ser obtida por meio de métodos numéricos, como por exemplo o método de Newton-Raphson. Para detalhes deste método os trabalhos de Lawless (1982) e Colosimo & Giolo (2006) devem ser consultados.

As distribuições paramétricas são úteis para descrever tempos de vida. Outro modelo utilizado é o modelo aditivo de Aalen.

2.1.5 O modelo de Aalen

Aalen (1980) propôs um modelo para dados de sobrevivência que identifica qual é o efeito das covariáveis sobre o risco de falha ao longo do tempo. Tal modelo é não paramétrico e não exige nenhuma pressuposição inicial, apenas assume que a ocorrência do evento é independente entre os indivíduos. Diversas aplicações do mesmo foram apresentadas. Mau (1986) investigou, por exemplo, o efeito da idade no risco de morte de pacientes com câncer de pulmão. Pereira (2004) avaliou o risco de desmame precoce

de crianças com até 18 meses de idade considerando algumas covariáveis, como conceito materno sobre o tempo ideal de amamentação. Grunkemeier et al. (2006) avaliaram o risco de hemorragias, tromboembolia, endocardites e o risco do maior escape paravalvular em relação ao implante de um dentre dois tipos de válvula cardíaca. Giarola et al. (2006) avaliaram o risco de óbito de pacientes com insuficiência renal crônica considerando fatores como diabetes, hipertensão, vasculite, entre outros. A seguir será descrito este modelo, o processo de estimação, testes para o efeito das covariáveis e um método gráfico para verificação da qualidade do ajuste.

No modelo de Aalen, o tempo de vida é descrito pela função intensidade. Considerando-se que o evento é único e ocorre apenas uma vez, funções intensidade e risco possuem a mesma interpretação, conforme descrito na seção 2.1.2. Assim, será utilizado nesta tese o termo risco.

Considerando-se r covariáveis na análise, a função risco, $\lambda_i(t)$, para o tempo de sobrevivência t do i-ésimo indivíduo no modelo linear de Aalen é uma combinação linear dos valores das covariáveis e é dado por:

$$\lambda_i(t) = \beta_0(t)Y_i(t) + \sum_{j=1}^{r} \beta_j(t)Y_i(t)Z_{ij}(t), \tag{2.17}$$

sendo $Y_i(t)$ a variável que indica se o indivíduo está ou não sob risco, conforme descrito na seção 2.1.2; $\beta_0(t)$ a função de risco de base, isto é, o risco para indivíduos com valor zero em todas as covariáveis; $Z_{ij}(t)$ é o valor observado no tempo t da j-ésima covariável para o i-ésimo indivíduo e $\beta_j(t)$ a função de regressão que mede o efeito sobre $\lambda_i(t)$ da j-ésima covariável no tempo t. Como essas funções de regressão são funções do tempo, a análise estatística das mesmas pode revelar mudanças no risco de falha com o passar do tempo devido à influência das covariáveis o e esta é uma das

vantagens deste modelo.

Considerando que n indivíduos são observados ao longo de um período de tempo para verificar a ocorrência de um determinado evento, o qual é independente entre os indivíduos, o modelo de Aalen para os riscos $\lambda_i(t), i = 1, 2, ..., n$, é escrito em sua forma matricial:

$$\lambda(t) = Z(t)\beta(t), \tag{2.18}$$

sendo $\beta(t) = (\beta_0(t), \beta_1(t), ..., \beta_r(t))'$ o vetor cujos elementos são as funções de regressão e $Z(t)$ uma matriz de dimensão $n \times (r+1)$, conforme a seguir:

$$Z(t) = \begin{bmatrix} Y_1(t) & Y_1(t)Z_{11}(t) & \cdots & Y_1(t)Z_{1r}(t) \\ Y_2(t) & Y_2(t)Z_{21}(t) & \cdots & Y_2(t)Z_{2r}(t) \\ \vdots & \vdots & \ddots & \vdots \\ Y_n(t) & Y_n(t)Z_{n1}(t) & \cdots & Y_n(t)Z_{nr}(t) \end{bmatrix}. \tag{2.19}$$

Deve-se observar que se o evento considerado ainda não ocorreu para o i-ésimo indivíduo e ele não é censurado, isto é, se ele está sob risco, então a i-ésima linha da matriz $Z(t)$ é o vetor $Z_i(t) = (1\ Z_{i1}(t)\ Z_{i2}(t)\ ...\ Z_{ir}(t))$, sendo $Z_{ij}(t)$ o valor da j-ésima covariável para o i-ésimo indivíduo. Caso contrário, se o indivíduo não está sob risco no tempo t, então a linha correspondente de $Z(t)$ contém apenas zeros.

Para exemplificar, suponha que em $t_k > 0$ apenas o indivíduo 1 já sofreu o evento ou foi censurado, então

$$Z(t_k) = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & Z_{21}(t_k) & \cdots & Z_{2r}(t_k) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1}(t_k) & \cdots & Z_{nr}(t_k) \end{bmatrix}.$$

Segundo Aalen (1989), o modelo de Aalen é consistente no sentido de que não é vulnerável a modificações no número de covariáveis modeladas. De acordo com Klein & Moeschberguer (1997), se uma covariável, independente das outras, é retirada do modelo, o novo modelo ainda é linear com as mesmas funções de regressão para as outras covariáveis; apenas a função de risco de base é modificada. Para melhor entendimento, pode-se supor um modelo com duas covariáveis, $Z_1(t)$ e $Z_2(t)$, dicotômicas, considerando-se $Y_i(t) = 1$, isto é,

$$\lambda_i(t) = \beta_0(t) + \beta_1(t)Z_{i1}(t) + \beta_2(t)Z_{i2}(t). \tag{2.20}$$

Para indivíduos que apresentam valor da covariável $Z_{i1}(t) = 1$, o modelo torna-se

$$\lambda_i(t) = \beta_0(t) + \beta_1(t) + \beta_2(t)Z_{i2}(t), \tag{2.21}$$

podendo ser escrito como

$$\lambda_i(t) = \beta_0^*(t) + \beta_2(t)Z_{i2}(t), \tag{2.22}$$

sendo $\beta_0^*(t) = \beta_0(t) + \beta_1(t)$. Assim, fixando-se a covariável $Z_1(t)$ em zero, o modelo ainda é linear e a função de regressão $\beta_2(t)$ que mede o efeito da covariável $Z_2(t)$ ao longo do tempo ainda é a mesma; esta é outra vantagem do modelo. Apenas a função de risco de base mudou de $\beta_0(t)$ para $\beta_0^*(t)$.

Aalen (1980) afirmou que o modelo de Aalen surgiu não para com-

petir com outros modelos já existentes, mas para complementar as análises, fornecendo informações mais detalhadas sobre o comportamento das covariáveis ao longo do tempo. Mau (1986) propôs utilizar os gráficos das funções de regressão acumuladas do modelo de Aalen como gráficos de diagnóstico, isto é, como uma ferramenta para detectar efeitos tempo dependentes das covariáveis, bem como os pontos em que a covariável muda sua influência. Segundo o autor, tais informações poderiam complementar, por exemplo, uma análise de Cox.

Porém, Aalen (1989) afirmou que os gráficos das funções de regressão acumuladas do modelo linear aditivo de Aalen não devem ser utilizados como uma ferramenta para a análise de Cox. O autor investigou, por meio de simulações, a eficácia de se utilizarem estes gráficos quando o verdadeiro modelo era de riscos proporcionais, não linear.

Modificando um pouco a proposta de Mau (1986), para uma análise exploratória dos dados, Henderson & Milner (1991) propõem sobrepor aos gráficos de regressão acumulada do modelo de Aalen uma estimativa da forma da curva esperada no modelo de riscos proporcionais. Assim, verdadeiras mudanças no efeito das covariáveis podem ser diferenciadas de mudanças aparentes.

O uso de repetições *Bootstrap* simples é então sugerido por Aalen (1993) para julgar quais características gráficas verdadeiramente refletem fenômenos reais e não são meramente variações aleatórias.

Desde que um modelo não é verdadeiro nem falso, ambos, o modelo de Cox e o modelo aditivo de Aalen, podem ser utilizados para obter informações de interesse a respeito dos dados, desde que os resultados sejam consistentes. O modelo linear aditivo de Aalen deve ser visto como uma

alternativa ao modelo de Cox e não ser fundamentalmente uma ferramenta de pesquisa para a análise de Cox. De acordo com Aalen (1993), o modelo a ser utilizado em casos práticos depende de qual deles melhor se ajusta aos dados.

Como o modelo de Aalen é não paramétrico, a estimação é feita sobre as funções de regressão acumulada.

2.1.5.1 Estimação das funções de regressão acumulada

No modelo de Aalen considera-se que as covariáveis atuam de modo aditivo à função de risco de base. Considera-se, ainda, que as funções de regressão são funções do tempo desconhecidas, e assim os efeitos das covariáveis podem variar durante o estudo. Logo, é necessário estimar estas funções de regressão. Mas nenhuma forma paramétrica é considerada para elas e é neste sentido que o modelo é considerado não paramétrico. Assim, estimam-se estas funções de regressão acumuladas, também conhecidas como coeficientes de risco, isto é, estima-se o vetor coluna $B(t)$ com elementos $B_j(t)$ dados por

$$B_j(t) = \int_0^t \beta_j(s)ds, \tag{2.23}$$

sendo $\beta_j(t)$ a função de regressão que representa o efeito sobre o risco $\lambda_i(t)$ da j-ésima covariável no tempo t.

Considerando-se, $t_1 < t_2 < \ldots < t_K < \ldots$, os tempos de falha ordenados, o estimador para a função (2.23) proposto por Aalen (1980) é

$$\widehat{B}_j(t) = \sum_{\forall t_K} X(t_K)I_K, \tag{2.24}$$

sendo $X(t)$ a inversa generalizada de $Z(t)$, dada em 2.19, e I_K um vetor de

zeros, exceto para o indivíduo que experimentou o evento no tempo t_K. Não são considerados tempos empatados, isto é, cada indivíduo tem um tempo de falha diferente. Isto não constitui em problema já que o tempo é uma variável aleatória contínua.

Devido à forma como a matriz $Z(t)$ é construída, haverá um instante de tempo t a partir do qual ela se torna uma matriz singular. Seja τ o maior valor de t para o qual $Z(t)$ é uma matriz não singular. Para $t \leq \tau$, $X(t)$ pode ser inicialmente qualquer inversa generalizada de $Z(t)$. Aalen (1980) sugere a inversa generalizada de mínimos quadrados, dada por $X(t) = [Z(t)'Z(t)]^{-1}Z(t)'$. Deste modo, o estimador (2.24) somente está definido sobre o intervalo de tempo $t \leq \tau$.

$\widehat{B}_j(t)$ descreve a influência sobre o risco da j-ésima covariável ao longo do tempo para $j \geqslant 1$, o que pode ser analisado pela inclinação do gráfico de $\widehat{B}_j(t)$ *versus* o tempo t. Assim, é possível verificar se uma determinada covariável tem um efeito constante ou varia ao longo do tempo durante o período de estudo. Se $\beta_j(t)$ é constante, o gráfico deverá aproximar-se de uma linha reta. Inclinações positivas ocorrem em períodos nos quais aumentos nos valores das covariáveis estão associados com aumentos na função de risco. Já inclinações negativas ocorrem em períodos nos quais aumentos nos valores das covariáveis estão associados com decréscimos na função de risco. As funções de regressão acumulada têm inclinação aproximadamente nula em períodos nos quais a covariável não influencia o risco. Uma covariável pode ter um efeito constante em um intervalo inicial de tempo e, então, ter seu efeito modificado posteriormente. Neste caso, espera-se que o gráfico da função de regressão acumulada seja linear no início do intervalo de tempo quando o efeito é constante, como já mencionado, e, segundo Hosmer

& Lemeshow (1999), sua forma nos tempos seguintes dependerá de como o efeito da covariável muda. De acordo com Grunkemeier et al. (2006), a função de regressão pode ser pensada como o risco adicional de um grupo de indivíduos em relação a outro grupo devido a um fator de risco.

Dado o vetor $Z_i = (1, Z_{i1}, \ldots, Z_{ir})'$, cujos elementos são os valores das covariáveis fixadas no tempo $t = 0$ para o i-ésimo indivíduo, e $\widehat{B}(t)' = (\widehat{B}_0(t), \widehat{B}_1(t), ..., \widehat{B}_r(t))$, cujos elementos são as funções de regressão acumulada, o risco acumulado para cada indivíduo, $\Lambda_i(t)$, pode ser estimado por

$$\widehat{\Lambda}_i(t) = \widehat{B}(t)' Z_i, \tag{2.25}$$

e a função de sobrevivência, de acordo com (2.5), por $\widehat{S}_i(t) = e^{-\widehat{\Lambda}_i(t)}$.

Aalen (1989) afirmou que a função de sobrevivência pode, também, ser estimada como um produto, analogamente ao estimador Kaplan-Meier. Assim,

$$\widehat{S}(t) = \prod_{\forall t_K} [1 - (X(t_K) I_K)' Z] \tag{2.26}$$

As funções de sobrevivência não são monótonas; pode haver intervalos nos quais ela é crescente e intervalos nos quais é decrescente. Assim, a função risco pode assumir valores negativos. Esta é uma desvantagem deste modelo.

No modelo de Aalen, como o risco é medido em função da influência das covariáveis ao longo do tempo, é necessário testar quais covariáveis exercem influência no risco de falha.

2.1.5.2 Testes para os efeitos das covariáveis

Muitas vezes é de interesse testar se determinada covariável exerce influência no risco ou não. Isto corresponde a testar a hipótese de nulidade de que o efeito da covariável sobre a função risco é desprezível, ou seja,

$$H_{0j} : \beta_j(t) = 0, \quad \forall t \leq \tau \tag{2.27}$$

para algum $j \geqslant 1$ e τ o maior valor de t para o qual o estimador $\widehat{B}_j(t)$, dado por (2.24), está definido. Devido ao contexto não paramétrico, tal hipótese apenas pode ser testada em intervalos de tempo nos quais a matriz $Z(t)$ é não singular, isto é, apenas para $t \leq \tau$.

Aalen (1980) desenvolveu uma estatística de teste dada por $U_j V_{jj}^{-1/2}$, a qual segue uma distribuição assintótica normal padrão sob H_{0j}. U_j é o j-ésimo elemento do vetor

$$U = \sum_{t=t_1}^{t_K} K(t)X(t)I_K, \tag{2.28}$$

sendo $K(t)$ uma matriz diagonal de dimensão $(r+1)$ cujos elementos diagonais k_{ii} são funções peso não negativas (Aalen, 1989). Note que o somatório atinge todos os tempos de falha. V_{jj} é o elemento da j-ésima linha e j-ésima coluna do estimador V da matriz de covariância de U, dado por

$$V = \sum_{t_K} K(t_K)X(t_K)I_K^D X(t_K)'K(t_K) \tag{2.29}$$

A estatística de teste (2.28) é uma combinação ponderada do somatório do estimador (2.24). A escolha das funções peso k_{ii} depende das hipóteses alternativas de interesse e do peso relativo que se deseja atribuir

para intervalos de tempo diferentes.

Duas escolhas para a função peso são sugeridas por Aalen (1989). A primeira considera a função peso igual ao número de indivíduos que estão sob risco em dado tempo e a matriz $K(t_K)$ é então substituída pelo escalar $\sum_{i=1}^{n} Y_i(t)$. A segunda escolha considera $K(t) = \{diag[(Z(t)'Z(t))^{-1}]\}^{-1}$. Ambas as escolhas são apropriadas para hipóteses alternativas unilaterais.

Para testar simultaneamente que $s \leqslant r$ covariáveis não exercem influência na função risco, utiliza-se a estatística $U_A' V_A^{-1} U_A$, a qual tem uma distribuição assintótica qui-quadrado com s graus de liberdade, sendo U_A o subvetor correpondente de U e V_A a submatriz correpondente de V.

Em Estatística, após ajustar um modelo, é necessário verificar sua adequacidade, isto é, se o modelo ajustado explica bem o comportamento dos dados. No caso do modelo de Aalen, isto é feito utilizando-se os resíduos de Cox-Snell.

2.1.5.3 Adequacidade do modelo

A verificação da adequacidade do ajuste de um modelo é muito importante em diversas áreas da estatística. Geralmente, isto é feito por meio da análise de resíduos. Os resíduos de Cox-Snell são bastante utilizados para esta finalidade. Segundo Carvalho et al. (2005), estes resíduos foram originalmente propostos para a avaliação da qualidade de ajuste do modelo de Cox. Eles são quantidades determinadas por

$$\widehat{e}_i = \widehat{\Lambda}(t_i|Z_i),$$

sendo t_i o tempo de falha ou censura e Z_i o vetor cujos elementos são os valores das covariáveis, ambos para o i-ésimo indivíduo.

De acordo com Allison (1995), a definição dos resíduos de Cox-Snell é baseada no fato de que o risco acumulado de um indivíduo no tempo de falha segue uma distribuição exponencial de média 1. Estimando-se os resíduos para todos os indivíduos, podem-se utilizar técnicas gráficas para verificar a adequacidade do modelo. Assim, o gráfico de $\widehat{e}_i$ versus $\widehat{\Lambda}(\widehat{e}_i)$ deve ser aproximadamente uma reta com inclinação 1, se o modelo exponencial for adequado.

Segundo Colosimo & Giolo (2006), também pode-se construir o gráfico das curvas de sobrevivência desses resíduos, obtidas por Kaplan Meier e pelo modelo exponencial padrão, para verificar a adequacidade do modelo ajustado. Quanto mais próximas elas estiverem, melhor a adequacidade do modelo ajustado.

No caso do modelo de Aalen, o risco acumulado pode ser estimado apenas enquanto as funções de regressão acumuladas puderem ser estimadas, isto é, apenas enquanto $Z(t)$ for uma matriz não singular. Seja R o tempo a partir do qual o risco acumulado não pode mais ser estimado. Considerando R como o tempo de censura, todos os indivíduos que ainda se encontram em risco neste tempo são considerados censurados. Seja Q_i o tempo de falha ou censura do i-ésimo indivíduo. Fixados os valores das covariáveis no tempo zero, a estimativa do risco acumulado por indivíduo no tempo t é dada pela equação (2.25). Segundo Aalen (1989) o resíduo é definido como $\widehat{\Lambda}_i(Q_i)$, sendo considerado censurado se Q_i é tempo de censura. Então, um gráfico das estimativas do risco acumulado destes resíduos pode ser construído para verificar se eles seguem uma distribuição exponencial padrão. Outros métodos para verificar a qualidade do ajuste do modelo de Aalen são sugeridos por Aalen (1993) e Gandy & Jensen (2005).

Grunkemeier et al. (2006) utilizaram o modelo de Aalen para comparar o efeito de dois tipos de válvulas cardíacas em sere humanos no risco de hemorragias e outras doenças.

2.1.6 Estudos de Grunkemeier et al. (2006)

Grunkemeier et al. (2006) desenvolveram um estudo para comparar o efeito de dois tratamentos no risco de hemorragias, tromboembolia, endocardites e maior escape paravalvular. Os tratamentos consistem no implante de duas válvulas cardíacas: uma convencionalmente utilizada (testemunha) e outra chamada Silzone (tratamento). Desejava-se avaliar o efeito da válvula Silzone, em relação ao efeito da válvula convencional, no risco de falha. A análise foi baseada nas funções risco acumulado e risco instantâneo.

Nesse estudo havia um único fator de risco, o tipo de válvula, o qual foi dicotomizado. Utilizando-se o modelo de regressão aditivo proposto por Aalen (1980), descrito na seção 2.1.5, a expressão do risco acumulado $\Lambda_i(t)$ para o i-ésimo indivíduo neste estudo é:

$$\Lambda_i(t) = B_0(t) + B_1(t)Z_i(t), \tag{2.30}$$

sendo $Z_i(t)$ o valor da covariável tipo de válvula para o i-ésimo indivíduo, a qual assume o valor 1 para os indivíduos que possuem válvula Silzone e 0 para os indivíduos que possuem a válvula convencional; $B_0(t)$ uma função de risco de base acumulada, isto é, o risco acumulado para um indivíduo com a válvula convencional; $B_1(t)$ a função de regressão acumulada que representa a influência do tipo de válvula ao longo do tempo. Assim, o risco

para indivíduos que possuem a válvula convencional pode ser escrito como

$$\Lambda_{i0}(t) = B_0(t) \tag{2.31}$$

e o risco para indivíduos implantados com a válvula Silzone como

$$\Lambda_{i1}(t) = B_0(t) + B_1(t). \tag{2.32}$$

Então, é possível estimar a função de regressão acumulada para a válvula Silzone quando existe um único fator de risco dicotomizado. A função de regressão acumulada do modelo de Aalen para a covariável tipo de válvula é igual à diferença entre as duas funções de risco acumuladas, dadas por (2.31) e (2.32), isto é,

$$\Lambda_{i1}(t) - \Lambda_{i0}(t) = B_1(t) \tag{2.33}$$

Assim, $B_1(t)$ representa o risco acumulado adicional da válvula Silzone em relação à válvula convencional. Porém, mais interessante que a estimativa da função de regressão acumulada é a estimativa da função de regressão instantânea, pois ela representa o risco adicional da válvula Silzone em relação à válvula convencional no tempo t. Tal função é a derivada da função de regressão acumulada dada em (2.33). Logo,

$$\lambda_{i1}(t) - \lambda_{i0}(t) = \beta_1(t), \tag{2.34}$$

isto é, $\beta_1(t)$ é obtida tomando-se a diferença entre duas funções risco instantâneas.

Mas, como o modelo de Aalen é não paramétrico, não se conhece

uma expressão matemática para a função risco acumulado e, portanto, não se pode derivá-la. Grunkemeier et al. (2006) afirmaram que, para encontrar a função de regressão instantânea, pode ser utilizada a modelagem paramétrica, a fim de suavizar as quinas da função de regressão acumulada e, então, tomar a derivada matemática (inclinação) das curvas suavizadas resultantes para obter a função de regressão instantânea.

Assim, os autores utilizaram a distribuição Gompertz para suavizar as funções de risco acumulado, $\Lambda_{i0}(t)$ e $\Lambda_{i1}(t)$, e estimar, utilizando diferenciação e a equação (2.34), a função de regressão instantânea β_1, isto é, o risco instantâneo adicional para pacientes implantados com a válvula Silzone em relação aos pacientes implantados com a válvula convencional.

Mas como estimar a função risco instantâneo quando se têm diversas covariáveis? A resposta para esta questão encontra-se na metodologia desta tese, descrita posteriormente na seção 3.1.

Para verificar a adequacidade da suavização paramétrica, será proposto um teste estatístico implementado por meio da técnica *Bootstrap*.

2.2 Análise Bootstrap para dados de sobrevivência

O método *Bootstrap*, proposto por Efron (1979), é uma técnica estatística de reamostragem utilizada em diversos contextos. Foi proposto por Efron (1979). Em sua essência, fundamenta-se na idéia de que, na ausência de qualquer outro conhecimento da população, a distribuição amostral de uma estatística é a melhor "orientação" da população.

Considere-se o modelo de regressão $X_i = g_i(\beta) + \epsilon_i, i = 1, ..., n$, sendo os X_i's independentes e identicamente distribuídos, $g(\cdot)$ uma função conhecida com vetor de parâmetros β desconhecido e ϵ_i com distribuição de

probabilidade F. A partir de uma amostra observada, utilizam-se técnicas estatísticas, como mínimos quadrados ou máxima verossimilhança, para estimar β. Supondo-se que se deseja obter informação sobre a distribuição amostral de $\widehat{\beta}$, deve-se reamostrar N vezes, com reposição, considerando a amostra original. Cada reamostragem é dita uma amostra *Bootstrap* e produz uma estimativa *bootstrap* $\widehat{\beta}_j^*, j = 1, ..., N$, pelo mesmo método que se obteve $\widehat{\beta}$. Assim, a amostra aleatória $(\widehat{\beta}_1^*, \widehat{\beta}_2^*, ..., \widehat{\beta}_N^*)$ pode ser utilizada para estimar a distribuição *Bootstrap* de $\widehat{\beta}$. Segundo Efron (1981) o método *Bootstrap* fornece resultados assintoticamente corretos.

De acordo com Efron & Tibshirani (1993), o número ideal de reamostras *Bootstrap* é $N = \infty$. Na prática, N deve ser um número finito restrito ao poder computacional disponível.

Segundo Efron (1981), quando se têm dados censurados a estimativa *Bootstrap* será obtida reamostrando-se $(T_i, \delta_i, Z_{i1}, ..., Z_{ir})$, sendo T_i, δ_i, $i = 1, ..., n$, e $Z_{ij}, j = 1, ..., r$ o tempo de sobrevida, a variável indicadora de censura e o valor da j-ésima covariável para o i-ésimo indivíduo, respectivamente.

Em se tratando da aplicação da técnica *Bootstrap* no modelo de Aalen, Aalen (1993) utilizou a análise *Boostrap* para investigar o efeito de covariáveis tempo dependentes nos gráficos das funções de regressão acumulada do modelo de Aalen e afirmou que a técnica pode ser utilizada para julgar quaisquer características desses gráficos. Nesta Tese, o método *Bootstrap* foi utilizado para verificar a porcentagem de amostras com as quais se obtém uma boa suavização da função de risco acumulado do modelo de Aalen pela função de risco acumulado da distribuição Weibull. Também foi utilizado para verificar a adequacidade da curva de riscos acumulados da

distribuição Lognormal à curva de risco acumulado do modelo de Aalen.

3 MÉTODOS E MATERIAL

3.1 Métodos

3.1.1 Estimação da função de risco instantâneo do modelo aditivo de Aalen considerando diversas covariáveis.

Conforme descrito na seção 2.1.6, Grunkemeier et al. (2006) mostraram que é possível utilizar uma distribuição paramétrica para suavizar a função de risco acumulado não paramétrico quando se têm dados de sobrevivência na presença de uma única covariável e, desta forma, estimar a função de risco instantâneo do modelo de Aalen para tal covariável. Mas como estimar o risco instantâneo de falha do Modelo de Aalen para o caso de mais de uma covariável?

O modelo aditivo de Aalen fornece estimativas para a função de risco acumulado, por meio das estimativas das funções de regressão acumuladas, conforme a equação (2.25). Mas, é mais interessante estimar o risco instantâneo, pois este permite conhecer o risco de falha em cada instante do tempo, podendo-se fazer predições. Para isso, neste modelo, é necessário estimar as funções de regressão instantâneas. Isto pode ser feito utilizando-se métodos de suavização.

Uma função não paramétrica pode ser suavizada utilizando-se modelos paramétricos, os quais se ajustam aos dados dessa função. Dessa forma, conhecida a expressão matemática para a função, ela poderia ser derivada matematicamente. Neste trabalho, foram utilizados modelos paramétricos para suavizar as funções de risco acumulado não paramétrico do modelo de Aalen na presença de duas ou mais covariáveis. A partir da derivada da expressão matemática para estas funções, foi possível obter a função de risco

instantâneo deste modelo. Além da simplicidade do método, obtendo-se uma expressão matemática para a função risco, é possível predizer o risco.

De acordo com a teoria de processos de contagem aplicada a dados de sobrevivência, detalhada na seção 2.1.2, considerou-se um conjunto de dados de tempos de sobrevivência com duas covariáveis dicotomizadas, Z_1 e Z_2, no qual, para cada indivíduo, são conhecidos o tempo de sobrevida, se o indivíduo experimentou o evento ou foi censurado e os valores das covariáveis dicotomizadas Z_1 e Z_2. Assim, os indivíduos foram divididos em quatro parcelas, segundo os valores das covariáveis dicotomizadas, como se encontram na Tabela 1.

A expressão do risco acumulado obtido do modelo de Aalen para cada indivíduo deste conjunto de dados é:

$$\Lambda_i(t) = B_0(t) + B_1(t)z_{i1}(t) + B_2(t)z_{i2}(t), \tag{3.1}$$

sendo $B_0(t)$ o risco para um indivíduo que apresenta o valor 0 para as duas covariáveis, isto é, para um indivíduo da parcela $(0, 0)$, $B_1(t)$ e $B_2(t)$ as funções de regressão acumuladas que representam o efeito das covariáveis Z_1 e Z_2 no risco de falha, respectivamente.

TABELA 1 Conjunto de dados de tempos de sobrevivência, considerando-se duas covariáveis dicotomizadas Z_1 e Z_2, agrupado em parcelas.

Parcela	Tempo	Z_1	Z_2
	t_{11}	0	0
(0,0)	t_{12}	0	0
	$\vdots$	$\vdots$	$\vdots$
	t_{1n_1}	0	0
	t_{21}	0	1
(0,1)	t_{22}	0	1
	$\vdots$	$\vdots$	$\vdots$
	t_{2n_2}	0	1
	t_{31}	1	0
(1,0)	t_{32}	1	0
	$\vdots$	$\vdots$	$\vdots$
	t_{3n_3}	1	0
	t_{41}	1	1
(1,1)	t_{42}	1	1
	$\vdots$	$\vdots$	$\vdots$
	t_{4n_4}	1	1

Para observar o efeito de uma covariável sobre o risco acumulado, optou-se por formar Blocos, conforme mostrado na Tabela 2. Assim, fixando uma das covariáveis em um de seus dois possíveis valores (0 ou 1) os indivíduos foram agrupados em quatro Blocos de duas parcelas cada um.

Desta forma, no Bloco 1, foi avaliado o efeito da covariável Z_2 no risco de falha em indivíduos que assumem o valor 0 para a covariável Z_1. No Bloco 2, foi avaliado o efeito da covariável Z_2 no risco de falha em indivíduos que apresentam o valor 1 para a covariável Z_1. No Bloco 3, foi avaliado o efeito da covariável Z_1 no risco de falha em indivíduos que apresentam valor

TABELA 2 Blocos formados fixando-se uma das covariáveis em um de seus dois valores possíveis.

Bloco	Parcela	Tempo	Z_1	Z_2	Bloco	Parcela	Tempo	Z_1	Z_2
		t_{11}	0	0			t_{11}	0	0
	(0,0)	t_{12}	0	0		(0,0)	t_{12}	0	0
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
1		t_{1n_1}	0	0	3		t_{1n_1}	0	0
		t_{21}	0	1			t_{31}	1	0
	(0,1)	t_{22}	0	1		(1,0)	t_{32}	1	0
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
		t_{2n_2}	0	1			t_{3n_3}	1	0
		t_{31}	1	0			t_{21}	0	1
	(1,0)	t_{32}	1	0		(0,1)	t_{22}	0	1
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
2		t_{3n_3}	1	0	4		t_{2n_2}	0	1
		t_{41}	1	1			t_{41}	1	1
	(1,1)	t_{42}	1	1		(1,1)	t_{42}	1	1
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
		t_{4n_4}	1	1			t_{4n_4}	1	1

0 para a covariável Z_2. Analogamente, no Bloco 4 foi avaliado o efeito da covariável Z_1 no risco de falha em indivíduos que assumem o valor 1 para a covariável Z_2.

Considerando-se o Bloco 1, a expressão para o risco acumulado do modelo de Aalen para o i-ésimo indivíduo é:

$$\Lambda_{Bl1i}(t) = B_0(t) + B_2(t)z_{i2}(t), \tag{3.2}$$

sendo $B_2(t)$ a função de regressão acumulada que representa o efeito da covariável Z_2 no risco de falha quando esta passa do estágio 0 para o estágio

1. Assim, os riscos acumulados para os indivíduos referentes às parcelas $(0,0)$ e $(0,1)$ são, respectivamente:

$$\Lambda_{(0,0)}(t) = B_0(t) \tag{3.3}$$

e

$$\Lambda_{(0,1)}(t) = B_0(t) + B_2(t). \tag{3.4}$$

Tomando-se a diferença entre as equações (3.4) e (3.3), obteve-se a função de regressão acumulada $B_2(t)$, conforme a equação a seguir:

$$\Lambda_{(0,1)}(t) - \Lambda_{(0,0)}(t) = B_2(t). \tag{3.5}$$

A equação (3.5) revela que a função de regressão acumulada B_2 representa o risco acumulado adicional dos indivíduos da parcela $(0,1)$ em relação aos indivíduos da parcela $(0,0)$, isto é, o risco de falha acumulado adicional quando a covariável Z_2 muda do estágio 0 para o estágio 1, estando Z_1 fixa em zero.

De modo análogo, o procedimento é repetido para os outros blocos. Considerando-se o Bloco 2, o risco acumulado do modelo de Aalen para o indivíduo i pode ser matematicamente escrito como:

$$\Lambda_{Bl2i}(t) = B_0(t) + B_1(t) + B_2(t)z_{i2}(t), \tag{3.6}$$

isto é,

$$\Lambda_{Bl2i}(t) = B_0^*(t) + B_2(t)z_{i2}(t), \tag{3.7}$$

sendo $B_0^*(t) = B_0(t) + B_1(t)$ o risco acumulado para um indivíduo que assume o valor 1 para a covariável Z_1 e 0 para a covariável Z_2, isto é,

o risco acumulado para um indivíduo da parcela $(1,0)$; e $B_2(t)$ a função de regressão acumulada que representa o efeito da covariável Z_2 no risco de falha em relação à parcela $(1,0)$. Então, os riscos acumulados para os indivíduos das parcelas $(1,0)$ e $(1,1)$ são, respectivamente:

$$\Lambda_{(1,0)}(t) = B_0^*(t) \tag{3.8}$$

e

$$\Lambda_{(1,1)}(t) = B_0^*(t) + B_2(t), \tag{3.9}$$

A partir da diferença entre as funções risco dadas pelas equações (3.9) e (3.8), teve-se a função de regressão acumulada $B_2(t)$, que representa o risco acumulado adicional dos indivíduos da parcela $(1,1)$ em relação aos indivíduos da parcela $(1,0)$, isto é, o risco de falha acumulado adicional quando a covariável Z_2 muda do estágio 0 para o estágio 1, para Z_1 fixa em 1, conforme equação a seguir:

$$\Lambda_{(1,1)}(t) - \Lambda_{(1,0)}(t) = B_2(t). \tag{3.10}$$

Em relação ao Bloco 3, o risco acumulado do modelo de Aalen para o i-ésimo indivíduo é dado por:

$$\Lambda_{Bl3i}(t) = B_0(t) + B_1(t)z_{i1}(t), \tag{3.11}$$

sendo $B_0(t)$ o risco acumulado para um indivíduo da parcela $(0,0)$ e $B_1(t)$ a função de regressão acumulada que representa o efeito da covariável Z_1 no risco de falha em relação à parcela $(0,0)$. Então, os riscos acumulados

referentes às parcelas $(0,0)$ e $(1,0)$ são, respectivamente, dados por:

$$\Lambda_{(0,0)}(t) = B_0(t) \tag{3.12}$$

e

$$\Lambda_{(1,0)}(t) = B_0(t) + B_1(t). \tag{3.13}$$

Tomando-se a diferença entre as equações (3.13) e (3.12), encontrou-se a função de regressão acumulada $B_1(t)$ que, conforme a equação (3.14), representa o risco acumulado adicional de indivíduos da parcela $(1,0)$ em relação aos indivíduos da parcela $(0,0)$, isto é, o risco de falha acumulado adicional quando a covariável Z_1 muda do estágio 0 para o estágio 1, fixando Z_2 em zero.

$$\Lambda_{(1,0)}(t) - \Lambda_{(0,0)}(t) = B_1(t). \tag{3.14}$$

Analogamente, para o Bloco 4 tem-se a expressão do risco acumulado do modelo de Aalen para o indivíduo i:

$$\Lambda_{Bl4i}(t) = B_0(t) + B_1(t)z_{i1}(t) + B_2(t), \tag{3.15}$$

isto é,

$$\Lambda_{Bl4i}(t) = B_0^*(t) + B_1(t)z_{i1}(t), \tag{3.16}$$

sendo $B_0^*(t) = B_0(t) + B_2(t)$ o risco acumulado para o indivíduo da parcela $(0,1)$ e $B_1(t)$ a função de regressão acumulada que representa o efeito da covariável Z_1 no risco de falha em relação aos indivíduos da parcela $(0,1)$. Logo, os riscos acumulados para indivíduos das parcelas $(1,0)$ e $(1,1)$, respectivamente, são:

$$\Lambda_{(0,1)}(t) = B_0^*(t) \tag{3.17}$$

e

$$\Lambda_{(1,1)}(t) = B_0^*(t) + B_1(t). \tag{3.18}$$

Tomando-se a diferença entre as equações (3.18) e (3.17), encontrou-se a função de regressão acumulada $B_1(t)$:

$$\Lambda_{(1,1)}(t) - \Lambda_{(0,1)}(t) = B_1(t). \tag{3.19}$$

A equação (3.19) mostra que a função de regressão acumulada B_1 representa o risco acumulado adicional dos indivíduos da parcela $(1,1)$ em relação aos indivíduos da parcela $(0,1)$, isto é, o risco de falha acumulado adicional quando a covariável Z_1 muda do estágio 0 para o estágio 1, para Z_2 fixa em 1.

As equações (3.5), (3.10), (3.14) e (3.19) mostraram que as respectivas funções de regressão acumuladas em cada bloco são obtidas pela diferença entre os riscos acumulados das parcelas deste bloco e representam o risco acumulado adicional de uma parcela em relação à outra, ou seja, o risco acumulado adicional quando a covariável muda de estágio. Assim, a função de regressão acumulada avalia quão maior é o risco acumulado de uma parcela em relação ao risco acumulado da outra parcela.

Derivando estas equações, (3.5), (3.10), (3.14) e (3.19), obteve-se a função de regressão instantânea em cada bloco. Por exemplo, derivando a equação (3.19), encontrou-se a função de regressão instantânea $\beta_1(t)$:

$$\lambda_{(1,1)}(t) - \lambda_{(0,1)}(t) = \beta_1(t), \tag{3.20}$$

sendo $\lambda_{(1,1)}(t)$ e $\lambda_{(0,1)}(t)$ os riscos de falha instantâneos para os indivíduos das parcelas $(1,1)$ e $(0,1)$, respectivamente. Assim, a função de regressão

instantânea $\beta_1(t)$ representa o risco de falha instantâneo adicional dos indivíduos da parcela $(1,1)$ em relação aos indivíduos da parcela $(0,1)$, isto é, o risco de falha instantâneo adicional quando a covariável Z_1 muda do estágio 0 para o estágio 1, fixando Z_2 em 1.

Então, estimar o risco de falha instantâneo de um grupo (parcela) em relação a outro é estimar a função de regressão instantânea que representa o efeito da covariável quando esta muda de estágio. Para isto, basta estimar o risco acumulado dos dois grupos e tomar a derivada matemática da diferença entre eles, caso se conheça uma expressão matemática para os riscos acumulados destes grupos. Mas no modelo de Aalen não há uma expressão analítica para as funções risco acumulado das parcelas, visto que o modelo é não paramétrico.

Assim, se para cada parcela puder ser ajustado um modelo paramétrico, como por exemplo os modelos descritos na seção 2.1.4, será possível estimar o risco instantâneo adicional para indivíduos de uma parcela em relação aos indivíduos da outra parcela. Supondo que para cada uma das parcelas de um bloco possa ser ajustado um modelo Weibull, o risco acumulado para os indivíduos de cada uma delas será expresso pela equação $\Lambda(t) = \left(\frac{t}{\alpha}\right)^{\gamma}$, dada em (2.10). Estes riscos acumulados Weibull constituem suavizações de riscos acumulados não paramétricos. Logo, a derivada matemática dos riscos acumulados Weibull, isto é, os riscos instantâneos Weibull, constituem suavizações dos riscos instantâneos não paramétricos. Tomando-se a diferença entre os riscos instantâneos Weibull das parcelas foi possível estimar a função de regressão instantânea, ou seja, o risco instantâneo adicional do modelo de Aalen de uma parcela em relação à outra. Neste caso, considerando, por exemplo, o Bloco 4, a expressão para a função de

regressão instantânea $\beta_1(t)$ é:

$$\beta_1(t) = \frac{\gamma_1}{\alpha_1^{\gamma_1}} t^{\gamma_1 - 1} - \frac{\gamma_0}{\alpha_0^{\gamma_0}} t^{\gamma_0 - 1}, \tag{3.21}$$

sendo γ_1, α_1 os parâmetros da distribuição Weibull para os indivíduos da parcela $(1,1)$ e γ_0, α_0 os parâmetros da distribuição Weibull para os indivíduos da parcela $(0,1)$, respectivamente.

TABELA 3 Conjunto de dados de tempos de sobrevivência, considerando três covariáveis dicotomizadas Z_1, Z_2 e Z_3, agrupado em parcelas.

Parcela	Tempo	Z_1	Z_2	Z_3	Parcela	Tempo	Z_1	Z_2	Z_3
	t_{11}	0	0	0		t_{51}	1	0	0
(0,0,0)	t_{12}	0	0	0	(1,0,0)	t_{52}	1	0	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	t_{1n_1}	0	0	0		t_{5n_5}	1	0	0
	t_{21}	0	0	1		t_{61}	1	0	1
(0,0,1)	t_{22}	0	0	1	(1,0,1)	t_{62}	0	0	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	t_{2n_2}	0	0	1		t_{6n_6}	1	0	1
	t_{31}	0	1	0		t_{71}	1	1	0
(0,1,0)	t_{32}	0	1	0	(1,1,0)	t_{72}	1	1	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	t_{3n_3}	0	1	0		t_{7n_7}	1	1	0
	t_{41}	0	1	1		t_{81}	1	1	1
(0,1,1)	t_{42}	0	1	1	(1,1,1)	t_{82}	1	1	1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\vdots$
	t_{4n_4}	0	1	1		t_{8n_8}	1	1	1

Se o conjunto de dados contiver três covariáveis dicotomizadas, Z_1, Z_2 e Z_3, os indivíduos poderão ser agrupados em $2^3 = 8$ parcelas de acordo com as triplas (z_{i1}, z_{i2}, z_{i3}), isto é, segundo os valores das três covariáveis dicotomizadas (Tabela 3).

A expressão para o risco acumulado do modelo de Aalen para cada indivíduo neste conjunto de dados é:

$$\Lambda_i(t) = B_0(t) + B_1(t)z_{i1}(t) + B_2(t)z_{i2}(t) + B_3(t)z_{i3}(t), \tag{3.22}$$

sendo B_0 o risco acumulado para um indivíduo da parcela $(0, 0, 0)$ e B_1, B_2, B_3 as funções de regressão acumuladas que avaliam o efeito das covariáveis Z_1, Z_2 e Z_3, respectivamente, no risco de falha.

O efeito de uma covariável sobre o risco acumulado poderá ser observado formando-se Blocos, conforme a Tabela 4. Assim, fixando duas das três covariáveis em um de seus dois valores possíveis (0 ou 1), os indivíduos poderão ser agrupados em $2^2.3 = 12$ blocos de duas parcelas cada um.

Os Blocos 7 a 12 são formados por indivíduos das seguintes parcelas:
Bloco 7: (1,0,0) e (1,1,0);
Bloco 8: (1,0,1) e (1,1,1);
Bloco 9: (0,0,0) e (1,0,0);
Bloco 10: (0,0,1) e (1,0,1);
Bloco 11: (0,1,0) e (1,1,0);
Bloco 12: (0,1,1) e (1,1,1).

Deste modo, em cada Bloco será avaliado o efeito de uma das três covariáveis no risco de falha para valores fixos das outras duas covariáveis. Por exemplo, no Bloco 1 será avaliado o efeito da covariável Z_3 no risco de falha de indivíduos que assumirão o valor 0 para as covariáveis Z_1 e Z_2; no

TABELA 4 Blocos 1 a 6 dentre os 12 Blocos formados fixando-se duas das covariáveis em um de seus dois valores possíveis.

Bloco	Parcela	Tempo	Z_1	Z_2	Z_3	Bloco	Parcela	Tempo	Z_1	Z_2	Z_3
		t_{11}	0	0	0			t_{71}	1	1	0
	(0,0,0)	t_{12}	0	0	0		(1,1,0)	t_{72}	1	1	0
		$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
1		t_{1n_1}	0	0	0	4		t_{7n_7}	1	1	0
		t_{21}	0	0	1			t_{81}	1	1	1
	(0,0,1)	t_{22}	0	0	1		(1,1,1)	t_{82}	1	1	1
		$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
		t_{2n_2}	0	0	1			t_{8n_8}	1	1	1
		t_{31}	0	1	0			t_{11}	0	0	0
	(0,1,0)	t_{32}	0	1	0		(0,0,0)	t_{12}	0	0	0
		$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
2		t_{3n_3}	0	1	0	5		t_{1n_1}	0	0	0
		t_{41}	0	1	1			t_{31}	0	1	0
	(0,1,1)	t_{42}	0	1	1		(0,1,0)	t_{32}	0	1	0
		$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
		t_{4n_4}	0	1	1			t_{3n_3}	0	1	0
		t_{51}	1	0	0			t_{21}	0	0	1
	(1,0,0)	t_{52}	1	0	0		(0,0,1)	t_{22}	0	0	1
		$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
3		t_{5n_5}	1	0	0	6		t_{2n_2}	0	0	1
		t_{61}	1	0	1			t_{41}	0	1	1
	(1,0,1)	t_{62}	1	0	1		(0,1,1)	t_{42}	0	1	1
		$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
		t_{6n_6}	1	0	1			t_{4n_4}	0	1	1

Bloco 2 será avaliado o efeito da covariável Z_3 no risco de falha de indivíduos que assumirão o valor 0 para a covariável Z_1 e o valor 1 para a covariável Z_2 e assim por diante.

Em cada um dos blocos, tomando-se a diferença entre os riscos acumulados das parcelas, obtém-se a função de regressão acumulada, a qual representará o risco acumulado adicional para indivíduos de uma parcela em relação a outra. A derivada matemática (inclinação) desta diferença fornecerá a estimativa da função de regressão instantânea.

Então, analogamente ao caso de duas covariáveis, obtendo-se um ajuste paramétrico para cada parcela, será possível estimar o risco instantâneo adicional para indivíduos de uma parcela em relação ao risco instantâneo para indivíduos da outra parcela.

Para efeito de ilustração considere-se, por exemplo, o Bloco 4. Neste Bloco será avaliado o efeito da covariável Z_3 no risco de falha de indivíduos que assumirão valores $z_{i1} = z_{i2} = 1$. O risco acumulado do modelo de Aalen para o i-ésimo indivíduo deste bloco é:

$$\Lambda_{BL4i}(t) = B_0(t) + B_1(t) + B_2(t) + B_3(t)z_{i3}(t), \tag{3.23}$$

ou seja,

$$\Lambda_{BL4i}(t) = B_0^*(t) + B_3(t), \tag{3.24}$$

sendo $B_0^*(t) = B_0(t) + B_1(t) + B_2(t)$ o risco acumulado para um indivíduo da parcela $(1, 1, 0)$ e $B_3(t)$ a função de regressão acumulada que representa o efeito da covariável Z_3 no risco de falha dos indivíduos do estágio zero em relação aos do estágio 1. Para os indivíduos das parcelas $(1, 1, 0)$ e $(1, 1, 1)$ deste bloco a expressão do modelo de Aalen é respectivamente,

$$\Lambda_{(1,1,0)}(t) = B_0(t) + B_1(t) + B_2(t) \tag{3.25}$$

e

$$\Lambda_{(1,1,1)}(t) = B_0(t) + B_1(t) + B_2(t) + B_3(t). \tag{3.26}$$

Tomando-se a diferença entre estas equações, ter-se-á a função de

regressão acumulada $B_3(t)$.

$$\Lambda_{(1,1,1)}(t) - \Lambda_{(1,1,0)}(t) = B_3(t). \tag{3.27}$$

Tal função de regressão representa o risco acumulado adicional de indivíduos da parcela $(1,1,1)$ em relação aos indivíduos da parcela $(1,1,0)$, isto é, o risco de falha acumulado adicional de indivíduos que assumem o valor 1 para as covariáveis Z_1 e Z_2, quando a covariável Z_3 muda do estágio 0 para o estágio 1. Derivando-se a equação (3.27) obter-se-á a função de regressão instantânea $\beta_3(t)$:

$$\lambda_{(1,1,1)}(t) - \lambda_{(1,1,0)}(t) = \beta_3(t), \tag{3.28}$$

sendo $\lambda_{(1,1,1)}(t)$ e $\lambda_{(1,1,0)}(t)$ os riscos de falha instantâneos para os indivíduos das parcelas $(1,1,1)$ e $(1,1,0)$, respectivamente. Logo, a função de regressão instantânea $\beta_3(t)$ representa o risco de falha instantâneo adicional de indivíduos da parcela $(1,1,1)$ em relação à parcela $(1,1,0)$.

Assim, para estimar o risco de falha instantâneo de indivíduos de uma parcela em relação a indivíduos de outra parcela, basta estimar a função de regressão instantânea. Logo, obtendo-se um ajuste paramétrico para o conjunto de dados das parcelas $(1,1,0)$ e $(1,1,1)$, é possível estimar o risco instantâneo do modelo de Aalen referente ao efeito da covariável Z_3 para indivíduos com $z_{i1} = z_{i2} = 1$. Supondo que para cada uma das parcelas ajusta-se um modelo Weibull, os riscos acumulados Weibull representam suavizações de riscos acumulados não paramétricos. Por consequência, os riscos instantâneos Weibull são suavizações dos riscos instantâneos não paramétricos. Tomando-se a diferença entre os dois primeiros obtém-se a função

de regressão instantânea:

$$\beta_3(t) = \frac{\gamma_1}{\alpha_1^{\gamma_1}} t^{\gamma_1 - 1} - \frac{\gamma_0}{\alpha_0^{\gamma_0}} t^{\gamma_0 - 1}, \tag{3.29}$$

sendo γ_1, α_1 os parâmetros da distribuição Weibull para os indivíduos da parcela $(1, 1, 1)$ e γ_0, α_0 os parâmetros da distribuição Weibull para os indivíduos da parcela $(1, 1, 0)$, respectivamente.

Quanto maior o número de covariáveis, maior o número de parcelas e mais blocos poderão ser formados. A metodologia acima pode ser utilizada em um banco de dados com n covariáveis dicotomizadas, $Z_1, Z_2, ..., Z_{n-1}, Z_n$. Neste caso, existirão 2^n parcelas e $n\ 2^{n-1}$ blocos. Os blocos serão formados fixando-se $n - 1$ covariáveis em um dos valores 0 ou 1. Para cada uma das duas parcelas de cada bloco, podem-se obter estimativas não paramétricas para o risco acumulado. Se for possível obter um ajuste paramétrico para cada uma das parcelas, ter-se-á uma expressão matemática para o risco acumulado não paramétrico que corresponderá a uma suavização deste risco. A partir desta suavização é possível estimar a função de regressão instantânea do modelo de Aalen. Para verificar isto neste caso de múltiplas covariáveis, considere-se o bloco no qual todas as covariáveis estão fixas em 1, exceto Z_1. Assim, este bloco é formado pelas parcelas $(0, 1, 1, ..., 1)$ e $(1, 1, 1, ..., 1)$. A expressão do modelo de Aalen para estas parcelas é, respectivamente:

$$\Lambda_{(0,1,1,...,1)}(t) = B_0(t) + B_2(t) + B_3(t) + ... + B_n(t) \tag{3.30}$$

e

$$\Lambda_{(1,1,1,...,1)}(t) = B_0(t) + B_1(t) + B_2(t) + B_3(t) + ... + B_n(t) \tag{3.31}$$

Tomando-se a diferença,

$$\Lambda_{(1,1,1,...,1)}(t) - \Lambda_{(0,1,1,...,1)}(t) = B_1(t), \tag{3.32}$$

obtém-se uma estimativa do risco acumulado de falha adicional para indivíduos da parcela $(1, 1, 1, ..., 1)$ em relação aos indivíduos da parcela $(0, 1, 1, ..., 1)$. Obtendo-se um ajuste paramétrico para cada uma das parcelas $(0, 1, 1, ..., 1)$ e $(1, 1, 1, ..., 1)$ será possível estimar o risco instantâneo adicional da parcela $(1, 1, 1, ..., 1)$ em relação à parcela $(0, 1, 1, ..., 1)$.

Como se pode observar, essa metodologia pode ser utilizada para dados com qualquer quantidade de covariáveis. Porém, é notório que quanto maior o número de covariáveis, maior o número de parcelas e menor o número de indivíduos por parcela. Assim, pode ocorrer de se terem poucas observações em determinadas parcelas, comprometendo o processo de estimação.

Para verificar-se quão bem os riscos acumulados obtidos a partir da distribuição paramétrica utilizada suavizam os riscos acumulados não paramétricos, contruiu-se um teste de significância.

3.1.2 Teste de adequacidade da suavização paramétrica por meio de simulação Monte Carlo.

Em concordância com os objetivos deste trabalho, foi proposto um teste de significância, construído através de técnicas de computação intensiva, com a finalidade de verificar a adequacidade da suavização paramétrica. Assim, a hipótese de nulidade é "H_0: O risco acumulado paramétrico obtido de uma distribuição paramétrica conhecida suaviza o risco acumulado não

paramétrico." Considerou-se a estatística definida por

$$D = \frac{max|\Lambda_P - \Lambda_{KM}|}{s_{\Lambda_{KM}}}, \tag{3.33}$$

sendo Λ_P e Λ_{KM}, respectivamente, os riscos acumulados obtidos da distribuição paramétrica conhecida e da estimativa Kaplan Meier da função de sobrevivência, doravante designados riscos Kaplan Meier, e $s_{\Lambda_{KM}}$ o desvio padrão dos riscos acumulados Kaplan Meier. O termo *max* refere-se ao maior valor, no caso, a maior diferença absoluta entre os riscos acumulados obtidos da distribuição paramétrica conhecida e os riscos acumulados Kaplan Meier.

A validação do teste proposto foi realizada por meio do cálculo da probabilidade empírica da ocorrência do controle do erro tipo I, assumindo-se o nível nominal de significância de 5%. Neste contexto, utilizaram-se 500 realizações Monte Carlo em cada parcela, considerando a geração das amostras sob H_0. Assumindo-se configurações entre diferentes tamanhos amostrais em todas as parcelas, fixados em $(n = 30, 50, 60, 90)$, e diferentes proporções médias de censura $(p = 30\%, 20\%, 10\%)$, justificou-se o uso do método *Bootstrap*, justamente pelo fato de que a distribuição amostral da estatística D dada em (3.33) não é conhecida.

Porém, vale ressaltar que as subamostras geradas, para uma dada proporção fixa de censuras, podem conter diferentes proporções de censuras, sendo algumas delas maiores que a proporção de censuras da amostra original. Além disso, as reamostras possuem uma quantidade de empates maior ou igual à quantidade de empates da amostra original.

Assim, visando a minimizar esse efeito e obter um valor-p mais preciso, utilizou-se o *Bootstrap* duplo, descrito em Davison & Hinkley (1997),

sendo que, para cada reamostragem feita no *Bootstrap* de 1º nível, uma outra reamostragem foi realizada, sendo esta definida como *Bootstrap* de 2º nível. A execução do algoritmo é ilustrada na Figura 4.

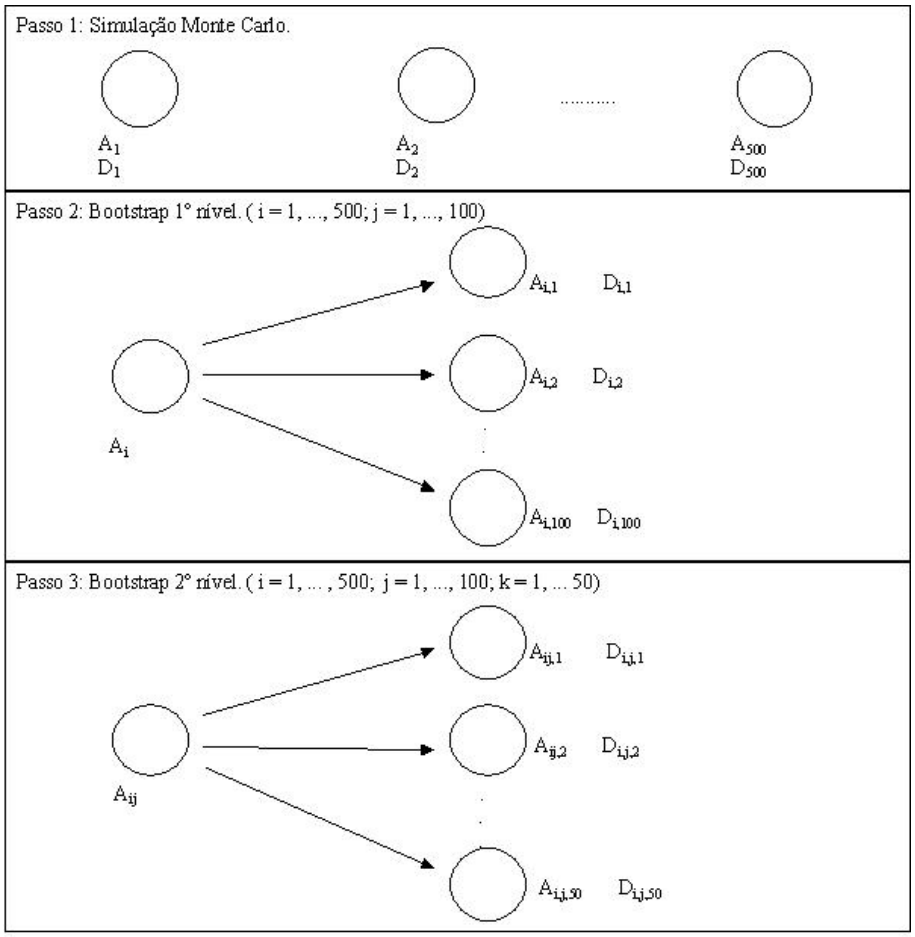

FIGURA 4 Passo 1: Esquema da geração das amostras por Simulação Monte Carlo (500 realizações) e obtenção da estatística $D_i (i = 1, ..., 500)$; Passo 2: Esquema do método *Bootstrap* de 1º nível aplicado a cada uma das amostras geradas no Passo 1 para obtenção das amostras A_{ij} e das estatísticas D_{ij}; Passo 3: Esquema do método *Bootstrap* de 2º nível aplicado a cada uma das subamostras *Bootstrap* do Passo 2 para obtenção das amostras A_{ijk} e das estatísticas D_{ijk}.

Em cada passo são obtidos os valores da estatística dada em (3.33), isto é,

$$D_i = \frac{max|\Lambda_{Pi} - \Lambda_{KMi}|}{s_{\Lambda_{KMi}}}, \qquad i = 1, ..., N, \tag{3.34}$$

no Passo 1,

$$D_{ij} = \frac{max|\Lambda_{Pij} - \Lambda_{KMij}|}{s_{\Lambda_{KMij}}}, \qquad j = 1, ..., n_{b1}, \tag{3.35}$$

no Passo 2 e

$$D_{ijk} = \frac{max|\Lambda_{Pijk} - \Lambda_{KMijk}|}{s_{\Lambda_{KMijk}}}, \qquad k = 1, ..., n_{b2}, \tag{3.36}$$

no Passo 3, sendo $N = 500$ o número de realizações Monte Carlo e $n_{b1} = 100$ e $n_{b2} = 50$ os números de amostras *Bootstrap* de 1º e 2º nível, respectivamente.

De acordo com esse algoritmo, as probabilidades de significância do teste proposto para verificar a proporção de suavizações adequadas são obtidas conforme as expressões (3.37), (3.38) e (3.39):

$$prob_1 = \frac{\#(D_{ij} \leq D_i)}{n_{b1}} \tag{3.37}$$

$$prob_2 = \frac{\#(D_{ijk} \leq D_{ij})}{n_{b2}} \tag{3.38}$$

$$prob_{cor} = \frac{\#(prob_2 \leq prob_1)}{n_{b1}} \tag{3.39}$$

Para cada reamostragem em cada parcela de cada tamanho amostral e cada proporção de censura, o valor-p, representado por $prob_1$, é obtido como a proporção do número de vezes em que o valor da estatística D_{ij}

dada em (3.35) for menor ou igual à estatística D_i, dada em (3.34). Assim, $prob_1$ é computado por (3.37).

Assim, é possível computar um valor auxiliar para o valor-p, dado em (3.38); porém, realizando-se $n_{b2} = 50$ reamostragens. O valor-p corrigido é estimado conforme a equação (3.39).

A probabilidade empírica da ocorrência do controle do erro tipo I, é calculada como sendo a proporção do número de vezes em que p-valor$<$ $0,05$. Logo, considerando-se $n = 500$ amostras Monte Carlo,

$$ptipo_1 = \frac{\#(prob_1 < 0,05)}{N} \tag{3.40}$$

e

$$ptipo_{1c} = \frac{\#(prob_{cor} < 0,05)}{N}, \tag{3.41}$$

sendo N o número de simulações.

3.2 Material

Foram utilizados neste trabalho dois conjuntos de dados, descritos a seguir. O primeiro corresponde a um conjunto de dados simulados. O outro são dados reais coletados no Hospital Veterinário da Universidade Federal de Lavras - UFLA.

3.2.1 Descrição do conjunto de dados simulados computacionalmente

Foram simulados computacionalmente dados de sobrevivência a partir de diferentes distribuições Weibull e agrupados em Blocos, conforme

descrito na seção 3.1.

Considerando-se duas covariáveis dicotomizadas denominadas Z_1 e Z_2, os indivíduos foram agrupados em quatro parcelas de acordo com os pares (z_{i1}, z_{i2}): $(0,0), (0,1), (1,0), (1,1)$. Para cada uma destas parcelas foram gerados tempos de falha e tempos de censura, segundo uma distribuição Weibull, considerando-se 30% de censuras, a partir de adaptações do programa desenvolvido por Rosa & Pedro Júnior (2008) que simularam dados de sobrevivência segundo uma distribuição exponencial.

A simulação foi feita por parcela porque desejava-se obter um ajuste paramétrico para os dados de cada parcela. A escolha da distribuição Weibull fundamenta-se no fato de que sua função risco tem um comportamento estritamente crescente para $\gamma > 1$ e também por ter inúmeras aplicações em análise de sobrevivência.

O programa para simular estes dados foi feito utilizando-se o software R Development Core Team (2009) e encontra-se no Anexo A. Tal programa foi construído a partir da transformação integral da probabilidade (Anexo B).

Para observar o efeito de uma covariável sobre o risco do modelo de Aalen foram formados quatro Blocos de duas parcelas cada, fixando uma das covariáveis em um dos seus dois valores possíveis, conforme a Tabela 2 da seção 3.1, apresentada novamente a seguir (Tabela 5):

TABELA 5 Blocos formados fixando-se uma das covariáveis em um de seus dois valores possíveis.

Bloco	Parcela	Tempo	Z_1	Z_2	Bloco	Parcela	Tempo	Z_1	Z_2
		t_{11}	0	0			t_{11}	0	0
	(0,0)	t_{12}	0	0		(0,0)	t_{12}	0	0
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
1		t_{1n_1}	0	0	3		t_{1n_1}	0	0
		t_{21}	0	1			t_{31}	1	0
	(0,1)	t_{22}	0	1		(1,0)	t_{32}	1	0
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
		t_{2n_2}	0	1			t_{3n_3}	1	0
		t_{31}	1	0			t_{21}	0	1
	(1,0)	t_{32}	1	0		(0,1)	t_{22}	0	1
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
2		t_{3n_3}	1	0	4		t_{2n_2}	0	1
		t_{41}	1	1			t_{41}	1	1
	(1,1)	t_{42}	1	1		(1,1)	t_{42}	1	1
		$\vdots$	$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$
		t_{4n_4}	1	1			t_{4n_4}	1	1

Os parâmetros da distribuição Weibull em cada parcela foram tomados de modo a se obter significância para a covariável de interesse em cada Bloco no modelo de Aalen, de modo a possibilitar a validação da metodologia proposta. O tamanho amostral de cada parcela foi previamente fixado nos valores apresentados na Tabela 6. Nesta tabela também encontram-se os valores dos parâmetros α e γ da distribuição Weibull para cada uma das parcelas. A simulação foi feita considerando-se uma proporção média de 30% de censuras.

TABELA 6 Tamanhos amostrais e parâmetros da distribuição Weibull propostos para simular dados de tempos de vida, considerando-se 30% de censura.

Parcela	n	α	γ
(0,0)	90	8,5	1,5
(0,1)	42	4,0	2,2
(1,0)	82	4,5	2,0
(1,1)	64	2,0	3,5

Assim, para os dados da parcela $(0,0)$, por exemplo, foram gerados $n = 90$ dados da distribuição Weibull dada pela função de densidade de probabilidade

$$f(t) = \frac{1,5}{8,5^{1,5}} t^{1,5-1} exp\left[-\left(\frac{t}{8,5}\right)^{1,5}\right].$$

3.2.2 Descrição do conjunto de dados reais

Com o propósito de verificar a viabilidade em aplicar a metodologia proposta utilizou-se um conjunto de dados que foi fornecido pelo Hospital Veterinário da Universidade Federal de Lavras (UFLA) referente a 193 cães diagnosticados com otite externa causada pela síndrome do banho e tosa no período de 21 de Fevereiro de 2002 a 06 de Dezembro de 2006.

A otite externa é uma doença incurável em alguns casos, mas pode ser controlada de modo que os sinais clínicos sejam banidos. Há cães que curam e outros que se encontram na chamada "otite subclínica", isto é, o paciente tem a otite externa, mas ela só é diagnosticada por meio de testes especiais. Este quadro é aceitável sob o ponto de vista médico e considera-se que a doença está controlada.

Assim, a variável resposta foi o tempo até a cura ou controle da do-

ença. Este tempo é contado em dias a partir da primeira consulta clínica do paciente, na qual é diagnosticada a otite externa, até a cura completa ou o desaparecimento dos sinais clínicos. Ao final do estudo, os cães que mantinham pelo menos um sinal da afecção foram considerados dados censurados. Em alguns casos, o cão necessita de correção cirúrgica e o proprietário, por motivos financeiros, não autoriza o procedimento e decide não retornar mais, mesmo que o caso tenha grandes chances de ser resolvido. Estes cães também foram considerados como dados censurados, pois não atingiram a cura ou o controle da doença.

Neste estudo foram consideradas duas covariáveis referentes ao tipo de tratamento utilizado: tratamento por via tópica e tratamento por via sistêmica. Há dezenove cães tratados apenas pela via tópica, sendo nove censurados, dezesseis cães tratados apenas pela via sistêmica, dentre eles seis censurados e cento e cinquenta e oito cães tratados conjuntamente pelas duas vias, sendo vinte e nove censurados. Assim, os dados foram agrupados em três parcelas, de acordo com a via de tratamento, conforme a Tabela 7 a seguir. Nesta tabela, o valor 0 indica que o cão não foi tratado pela referida via e o valor 1 indica que o cão foi tratado pela referida via.

TABELA 7 Conjunto de dados de cães diagnosticados com otite externa tratados pelas vias Tópica (T) e Sistêmica (S) agrupado em parcelas.

Parcela	Tempo	T	S
(0,1)	$t_{2,1}$	0	1
	$t_{2,2}$	0	1
	$\vdots$	$\vdots$	$\vdots$
	$t_{2,16}$	0	1
(1,0)	$t_{3,1}$	1	0
	$t_{3,2}$	1	0
	$\vdots$	$\vdots$	$\vdots$
	$t_{3,19}$	1	0
(1,1)	$t_{4,1}$	1	1
	$t_{4,2}$	1	1
	$\vdots$	$\vdots$	$\vdots$
	$t_{4,158}$	1	1

4 RESULTADOS E DISCUSSÃO

4.1 Resultados para os dados simulados.

4.1.1 Estimação da função de risco instantâneo do modelo de Aalen considerando duas covariáveis.

Reunindo-se os dados simulados por parcela, obtiveram-se 132 observações para o Bloco 1, 146 para o Bloco 2, 172 para o Bloco 3 e 106 para o Bloco 4. As análises foram feitas por Blocos, conforme a metodologia descrita na seção 3.1. Em cada Bloco procedeu-se da seguinte forma: estimaram-se os riscos acumulados do modelo de Aalen, considerando-se que a distribuição dos dados é desconhecida; buscou-se uma função de riscos acumulados conhecida que melhor se ajustasse aos riscos acumulados do modelo de Aalen; determinou-se a diferença entre essas funções risco acumulado conhecidas; e utilizou-se a derivada matemática para obter uma expressão para o risco instantâneo do modelo de Aalen.

Em todos os Blocos a função risco acumulado que melhor se ajustou aos riscos acumulados do modelo de Aalen foi o risco acumulado Weibull, conforme esperado, pois os dados foram gerados a partir da distribuição Weibull. Para visualizar este resultado foram construídos gráficos, no mesmo sistema cartesiano, dos riscos acumulados não paramétricos, obtidos a partir da estimativa Kaplan Meier da função de sobrevivência e da equação (2.5), e dos riscos acumulados Weibull para cada parcela. Os resultados encontram-se descritos para cada Bloco individualmente.

No Bloco 1, os resultados do modelo de Aalen ajustado são dados na Tabela 8, para $t = 6,6$, aproximadamente. Nesta tabela, observou-se que a covariável Z_2 apresentou significância estatística, indicando que no

tempo $t = 6,6$ esta covariável influenciou no risco de ocorrência do evento, fixando-se em zero o valor da covariável Z_1.

TABELA 8 Estimativas obtidas em $t = 6,553301$ para o modelo de Aalen ajustado ao Bloco 1.

Covariável	Coeficiente	Erro padrão	IC(95%)	Valor p
Intercepto	0,797	0,122	[0,557;1,037]	0,000
Z_2	1,006	0,462	[0,101;1,911]	0,002

Foram construídos os gráficos das estimativas das funções de regressão acumuladas e seus respectivos intervalos de 95% de confiança para o modelo de Aalen ajustado ao Bloco 1, os quais se encontram na Figura 5. O primeiro gráfico desta figura (gráfico Intercepto) representa o risco acumulado para indivíduos da parcela $(0,0)$. Este gráfico possui uma inclinação positiva, o que indicou que indivíduos da parcela $(0,0)$ possuem um risco que se eleva gradativamene com o passar do tempo. O segundo gráfico corresponde ao efeito da covariável Z_2, ou seja, o risco adicional para indivíduos da parcela $(0,1)$ em relação ao risco para indivíduos da parcela $(0,0)$. Este gráfico indicou que indivíduos da parcela $(0,1)$ possuem um risco acumulado mais elevado que indivíduos da parcela $(0,0)$.

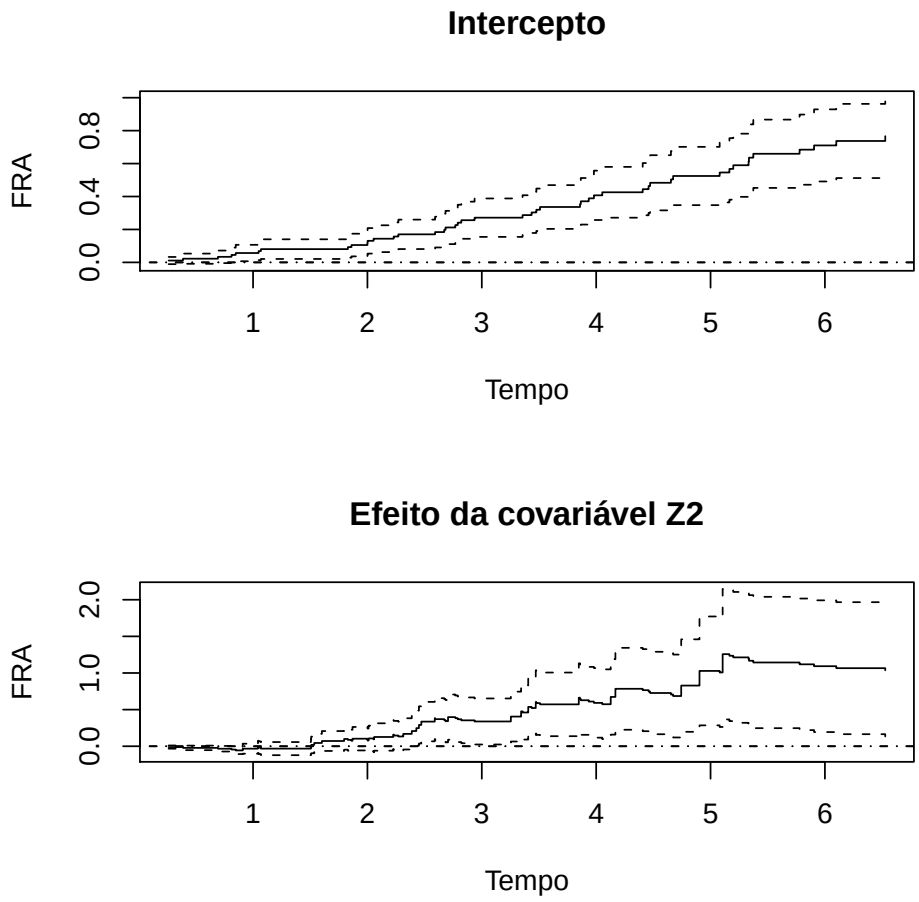

FIGURA 5 Funções de regressão acumulada (FRA) obtidas do modelo de Aalen ajustado para o Bloco 1 *versus* o tempo.

As estimativas $\widehat{\gamma}$ e $\widehat{\alpha}$ dos parâmetros da distribuição Weibull ajustada, obtidas para cada parcela pertencente ao Bloco 1, foram, respectivamente, $1,483782$ e $7,715627$ para a parcela $(0,0)$ e $2,153371$ e $3,965625$ para a parcela $(0,1)$. Assim, as expressões para as estimativas do risco

acumulado Weibull nas parcelas $(0,0)$ e $(0,1)$ são, respectivamente:

$$\widehat{\Lambda}_{(0,0)}(t) = \left(\frac{t}{7,72}\right)^{1,48} \tag{4.1}$$

e

$$\widehat{\Lambda}_{(0,1)}(t) = \left(\frac{t}{3,97}\right)^{2,15}. \tag{4.2}$$

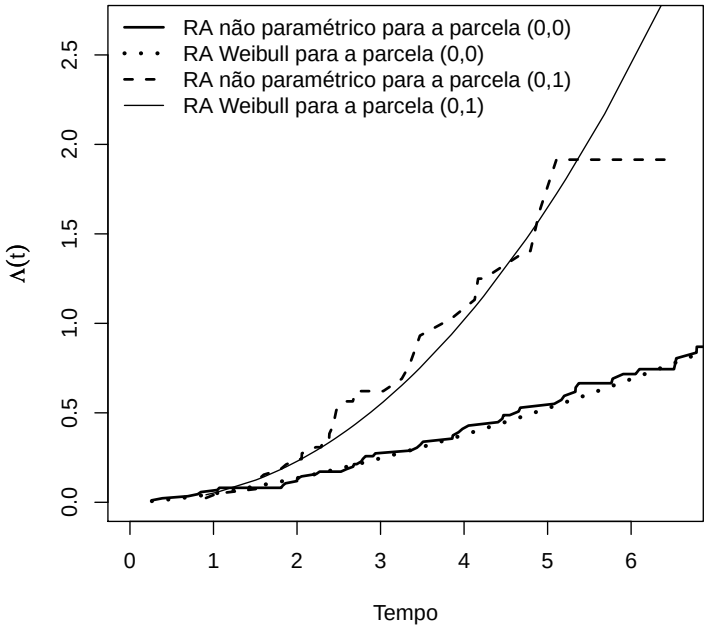

FIGURA 6 Estimativas dos riscos acumulados não paramétricos e Weibull para cada parcela do Bloco 1 em função do tempo.

Na Figura 6 encontram-se os gráficos dos riscos acumulados Weibull e dos riscos acumulados não paramétricos, obtidos das estimativas Kaplan Meier da função de sobrevivência, para as parcelas do Bloco 1. Observou-se nesta figura que o risco acumulado para indivíduos da parcela $(1,0)$ foi maior

que o risco acumulado para indivíduos da parcela $(0,0)$, conforme também se obteve no modelo de Aalen ajustado. Além disso, notou-se que as curvas de riscos acumulados Weibull suavizaram as curvas de riscos acumulados não paramétricos.

Tomando-se a diferença entre as equações (4.2) e (4.1) e derivando o resultado obteve-se uma expressão para a estimativa da função de regressão instantânea β_2 que corresponde ao efeito da covariável Z_2 sobre o risco ao fixar Z_1 em zero. Assim,

$$\widehat{\beta}_2(t) = \frac{2,15}{3,97^{2,15}} t^{1,15} - \frac{1,48}{7,72^{1,48}} t^{0,48} \tag{4.3}$$

O gráfico para as estimativas desta função de regressão instantânea $\beta_2(t)$ versus o tempo é apresentado na Figura 7. Este gráfico mostrou que o risco instantâneo para indivíduos da parcela $(0,1)$ é maior que o risco para indivíduos da parcela $(0,0)$ e esta diferença entre os riscos das parcelas é crescente. Além disso, em $t = 6$, por exemplo, obteve-se $\widehat{\beta}_2 = 0,71$, aproximadamente. Isto significa que em $t = 6$, em determinada unidade de tempo previamente definida (dias, meses, anos, ...), ter-se-iam, em média, $0,71$ falhas, o que corresponde a 1 falha a cada $1,42$ unidades de tempo ($\frac{1}{0,71} = 1,42$), em média. Logo, indivíduos da parcela $(0,0)$ precisam em média de $1,42$ unidades de tempo a mais que indivíduos da parcela $(0,1)$ para experimentar o evento.

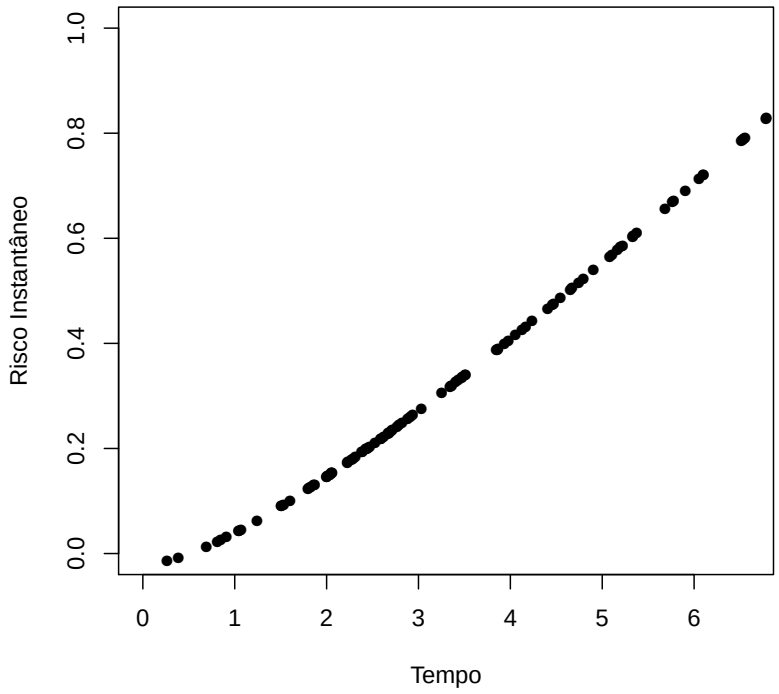

FIGURA 7 Estimativas $\widehat{\beta}_2$ da função de regressão instantânea do Bloco 1 em função do tempo.

Na Tabela 9 encontram-se as estimativas e intervalos de 95% de confiança resultantes do modelo de Aalen ajustado para o Bloco 2 no tempo $t = 2,6$, aproximadamente. A covariável Z_2 apresentou significância estatística, indicando que no tempo $t = 2,6$ esta covariável influenciou no risco de ocorrência do evento, fixando-se em 1 o valor da covariável Z_1.

TABELA 9 Estimativas obtidas em $t = 2,562483$ para o modelo de Aalen ajustado ao Bloco 2.

Covariável	Coeficiente	Erro padrão	IC(95%)	Valor p
Intercepto	0,345	0,074	[0,200; 0,491]	0
Z_2	1,331	0,563	[0,227; 2,436]	0

Os gráficos das funções de regressão acumuladas e seus respectivos intervalos de 95% de confiança para as parcelas deste Bloco 2 encontram-se na Figura 8. O gráfico Intercepto, correspondente à função de regressão acumulada para indivíduos da parcela $(1,0)$, indicou que o risco acumulado para indivíduos desta parcela cresce gradativamente com o passar do tempo, principalmente após $t = 1,8$, aproximadamente. O gráfico da função de regressão acumulada que representa o efeito da covariável Z_2 indicou que indivíduos da parcela $(1,1)$ possuem um risco maior de ocorrência do evento que indivíduos da parcela $(1,0)$ após o instante de tempo $t = 1,8$ aproximadamente. Antes deste instante de tempo a função de regressão acumulada é estável, indicando que o risco para indivíduos da parcela $(1,1)$ é similar ao risco para indivíduos da parcela $(1,0)$.

As estimativas $\widehat{\gamma}$ e $\widehat{\alpha}$ dos parâmetros da distribuição Weibull ajustada a cada parcela são, respectivamente, $1,938541$ e $4,695988$ para a parcela $(1,0)$ e $2,838148$ e $2,131648$ para a parcela $(1,1)$. Assim, para as parcelas $(1,0)$ e $(1,1)$, as expressões para as estimativas do risco acumulado Weibull são, respectivamente:

$$\widehat{\Lambda}_{(1,0)}(t) = \left(\frac{t}{4,69}\right)^{1,93} \tag{4.4}$$

e

$$\widehat{\Lambda}_{(1,1)}(t) = \left(\frac{t}{2,13}\right)^{2,84}. \tag{4.5}$$

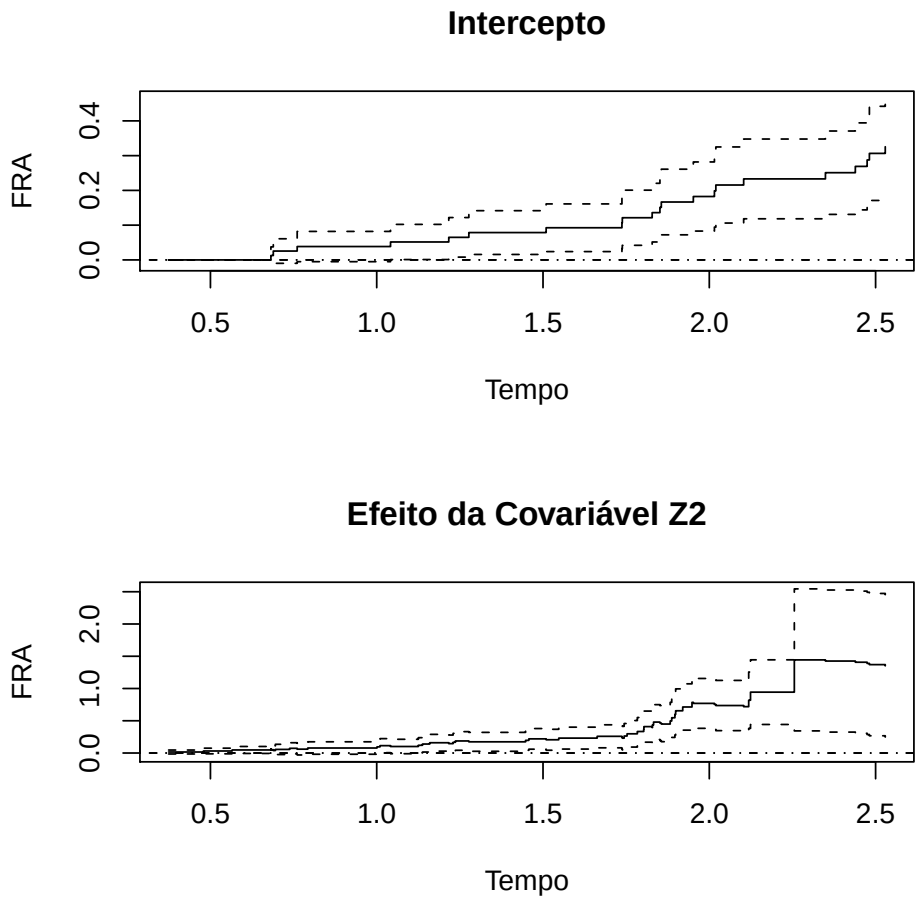

FIGURA 8 Funções de regressão acumulada (FRA) obtidas do modelo de Aalen ajustado para o Bloco 2 *versus* o tempo.

Os gráficos para as funções (4.4) e (4.5) e para os riscos acumulados não paramétricos obtidos da estimativa Kaplan-Meier da função de sobrevivência encontram-se na Figura 9. Esta figura indicou que o risco acumulado para indivíduos da parcela $(1, 1)$ é maior que o risco acumulado para indivíduos da parcela $(1, 0)$. Por simples inspeção notou-se que os ris-

cos acumulados Weibull foram boas suavizações para os riscos acumulados não paramétricos.

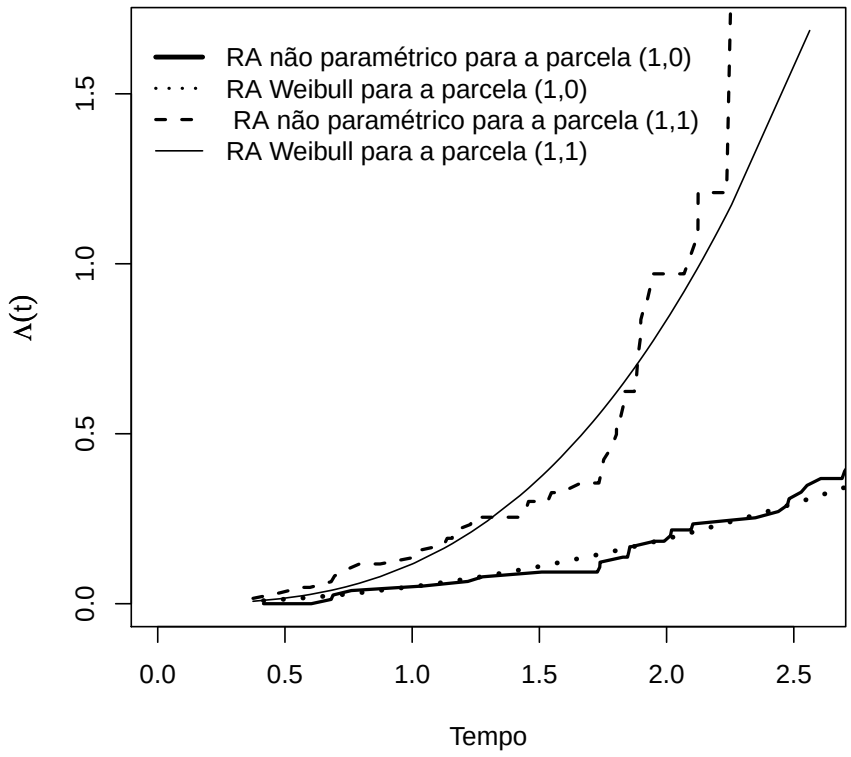

FIGURA 9 Estimativas dos riscos acumulados não paramétricos e Weibull para cada parcela do Bloco 2 em função do tempo.

A partir da derivada da diferença entre as equações (4.5) e (4.4) obteve-se uma expressão para a estimativa da função de regressão instantânea β_2 do Bloco 2 que representa o efeito da covariável Z_2 sobre o risco, fixando-se em 1 a covariável Z_1. Logo,

$$\widehat{\beta}_2(t) = \frac{2,84}{2,13^{2,84}} t^{1,84} - \frac{1,93}{4,69^{1,93}} t^{0,93} \tag{4.6}$$

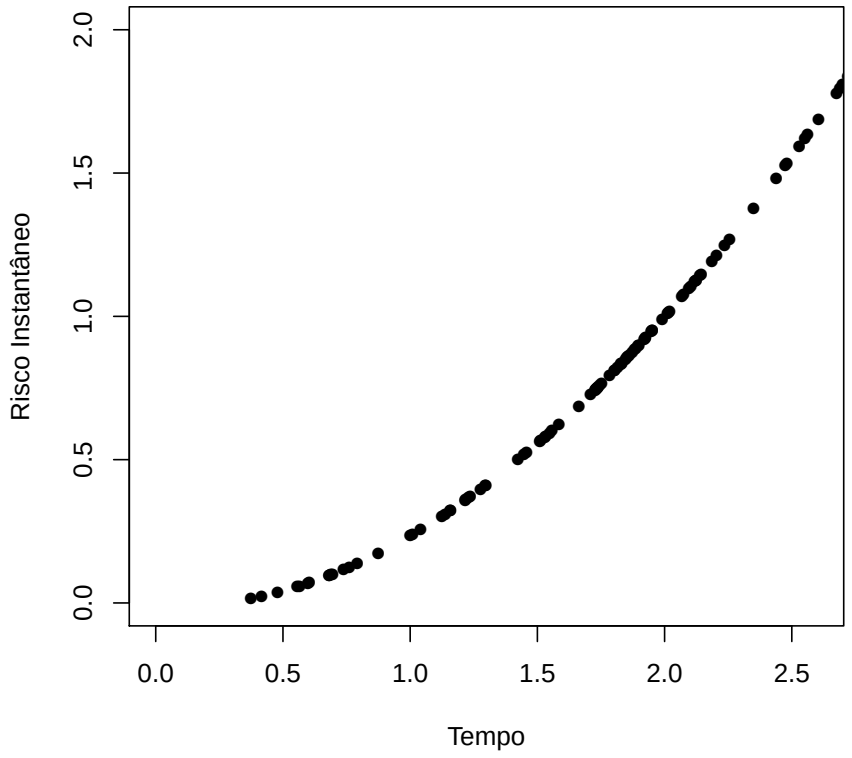

FIGURA 10 Estimativas $\widehat{\beta}_2$ da função de regressão instantânea do Bloco 2 em função do tempo.

As estimativas desta função de regressão instantânea $\beta_2(t)$ em cada tempo foram ilustradas na Figura 10. Este gráfico permitiu dizer que o risco instantâneo para indivíduos da parcela $(1,1)$ foi maior que o risco para indivíduos da parcela $(1,0)$, enfatizando que esta diferença entre os riscos das parcelas é crescente. Além disso, em $t = 2,5$, por exemplo, tem-se $\widehat{\beta}_2 = 1,56$, aproximadamente. Em termos práticos, este resultado permitiu interpretar que, em $t = 2,5$, em determinada unidade de tempo previamente definida, ter-se-iam em média $1,56$ falhas, o que corresponde a 1 falha a cada $0,64$ unidade de tempo ($\frac{1}{1,56} = 0,64$) em média. Logo, indivíduos da parcela $(1,0)$ precisam em média de $0,64$ unidades de tempo a mais que indivíduos da parcela $(1,1)$ para experimentar o evento.

Os resultados descritos na Tabela 10 referem-se às estimativas e intervalos de 95% de confiança resultantes do modelo de Aalen ajustado para o Bloco 3 em $t = 8,43$, aproximadamente. Observou-se que a covariável Z_1 é significativa estatisticamente, o que indicou que o risco de ocorrência do evento é influenciado pela covariável Z_1 quando Z_2 é fixado em zero.

TABELA 10 Estimativas obtidas em $t = 8,429934$ para o modelo de Aalen ajustado ao Bloco 3.

Covariável	Coeficiente	Erro padrão	IC(95%)	Valor p
Intercepto	1,061	0,158	[0,750; 1,371]	0
Z_1	2,282	1,127	[0,074; 4,490]	0

A Figura 11 ilustra os gráficos das funções de regressão acumuladas para o Bloco 3. No primeiro gráfico (gráfico Intercepto) observou-se que o risco para indivíduos da parcela $(0,0)$ eleva-se gradativamente com o passar do tempo. No segundo gráfico (gráfico Efeito da covariável Z_1) observou-

se que o risco para indivíduos da parcela $(1,0)$ foi maior que o risco para indivíduos da parcela $(0,0)$ até $t = 6$, aproximadamente. Após este tempo a função de regressão acumulada se estabilizou, indicando que o risco para indivíduos de ambas as parcelas é similar.

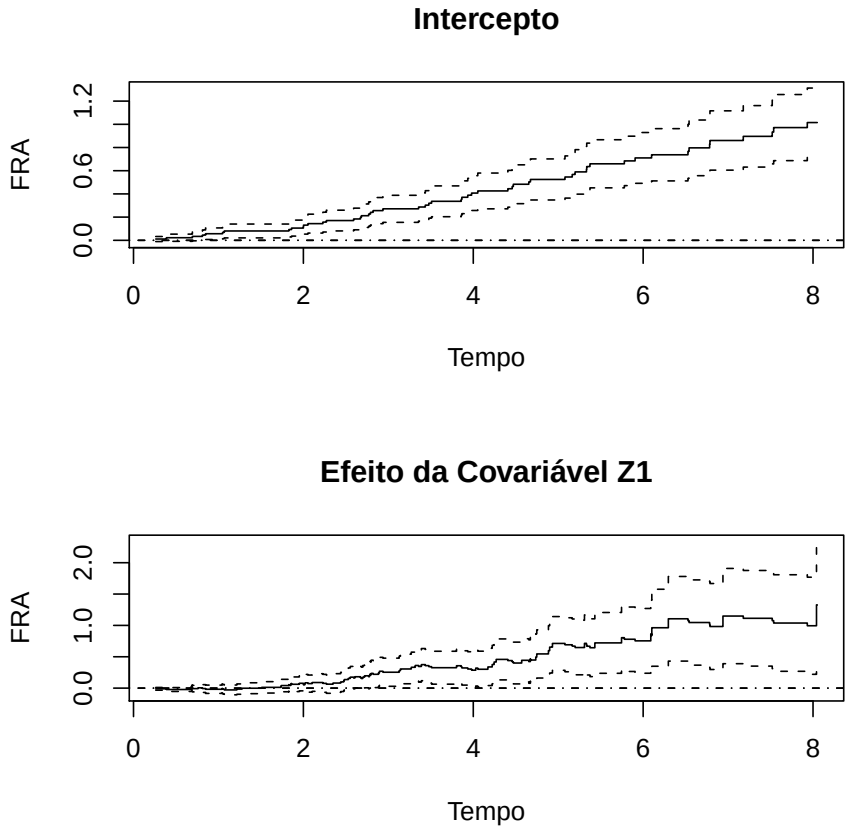

FIGURA 11 Funções de regressão acumulada (FRA) obtidas do modelo de Aalen ajustado para o Bloco 3 *versus* o tempo.

As expressões para as estimativas do risco acumulado Weibull para as parcelas $(0,0)$ e $(1,0)$ são, respectivamente, dadas pelas equações (4.1) e (4.4).

A Figura 12 apresenta os gráficos dos riscos acumulados não paramétricos, obtidos das estimativas Kaplan Meier da função de sobrevivência, e Weibull para cada uma das parcelas do Bloco 3. O risco acumulado para indivíduos da parcela $(1, 0)$ foi maior que o risco acumulado para indivíduos da parcela $(0, 0)$. Além disso, os riscos acumulados Weibull suavizaram os riscos acumulados não paramétricos.

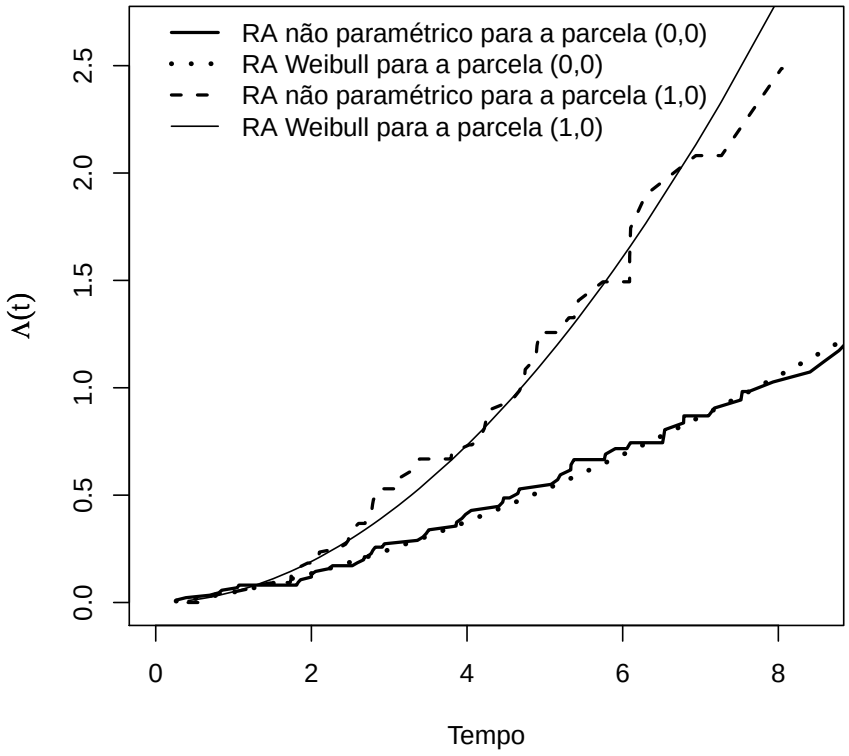

FIGURA 12 Estimativas dos riscos acumulados não paramétricos e Weibull para cada parcela do Bloco 3 em função do tempo.

Derivando a diferença entre as equações (4.4) e (4.1) obteve-se a estimativa da função de regressão instantânea β_1 do Bloco 3, que representa o efeito da covariável Z_1 sobre o risco de ocorrência do evento ao fixar Z_2 em zero. Portanto,

$$\widehat{\beta}_1(t) = \frac{1,93}{4,69^{1,93}} t^{0,93} - \frac{1,48}{7,72^{1,48}} t^{0,48}. \tag{4.7}$$

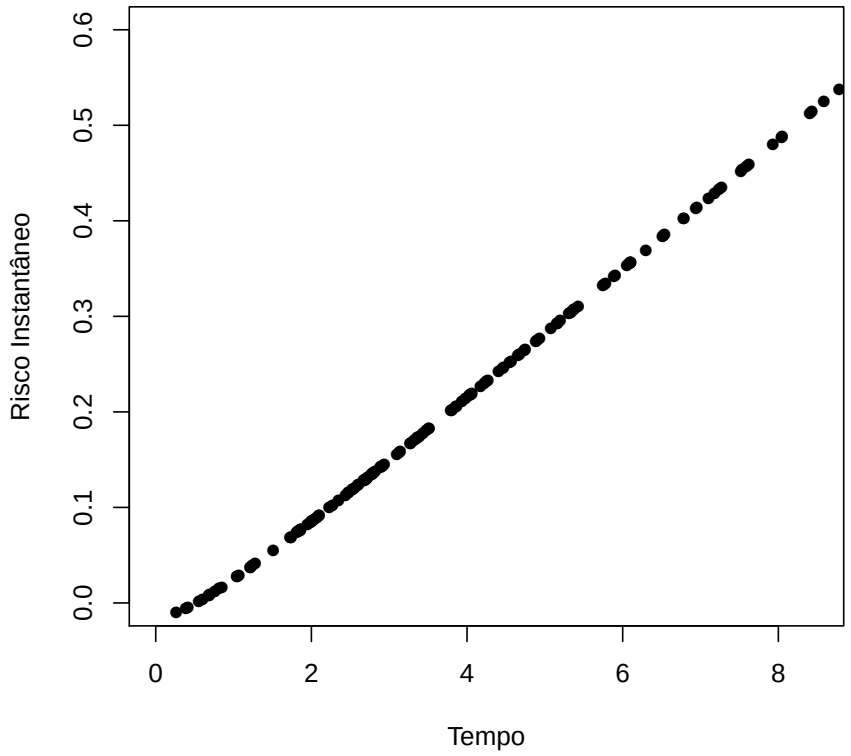

FIGURA 13 Estimativas $\widehat{\beta}_1$ da função de regressão instantânea do Bloco 3 em função do tempo.

O gráfico para a expressão desta estimativa de β_1 em função do tempo encontra-se na Figura 13 e indicou que o risco instantâneo para indivíduos da parcela $(1,0)$ é maior que o risco para indivíduos da parcela $(0,0)$ e esta diferença entre os riscos das parcelas é crescente. O gráfico também mostrou, por exemplo, que em $t = 8$, $\widehat{\beta}_1 = 0,48$, aproximadamente. Assim, em $t = 8$, em determinada unidade de tempo previamente definida, ter-se-iam em média $0,48$ falhas, o que corresponde em média a 1 falha a cada $2,06$ unidades de tempo ($\frac{1}{0,48} = 2,06$). Logo, em média, indivíduos da parcela $(0,0)$ precisam, em média, de $2,06$ unidades de tempo a mais que indivíduos da parcela $(1,0)$ para experimentar o evento.

Os resultados do modelo de Aalen ajustado para o Bloco 4 encontram-se descritos na Tabela 11 e os gráficos das funções de regressão acumuladas na Figura 14. Na Tabela 11 observou-se que no tempo $t = 2,6$ aproximadamente, a covariável Z_1 influenciou no risco de ocorrência do evento fixando-se Z_2 em 1.

TABELA 11 Estimativas obtidas em $t = 2,562483$ para o modelo de Aalen ajustado ao Bloco 4.

Covariável	Coeficiente	Erro padrão	IC(95%)	Valor p
Intercepto	0,507	0,130	[0,253; 0,761]	0
Z_1	1,170	0,573	[0,046; 2,294]	0

No primeiro gráfico da Figura 14 (gráfico Intercepto) constatou-se que, após $t = 1,5$, o risco para indivíduos da parcela $(0,1)$ aumentou gradativamente com o passar do tempo. Já no segundo gráfico (gráfico Efeito da covariável Z_1) verificou-se que o risco para indivíduos da parcela $(1,1)$ foi maior que o risco para indivíduos da parcela $(0,1)$ após, aproximada-

mente, $t = 1,8$, sendo que antes deste instante os riscos das duas parcelas são similares .

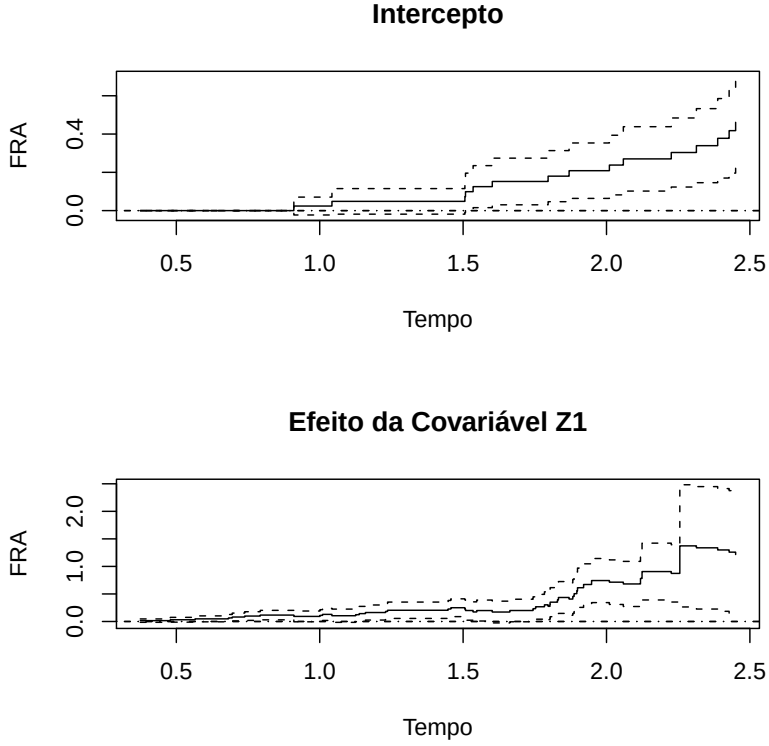

FIGURA 14 Funções de regressão acumulada (FRA) obtidas do modelo de Aalen ajustado para o Bloco 4 *versus* o tempo.

Para as parcelas $(0,1)$ e $(1,1)$, as expressões para as estimativas do risco acumulado Weibull são respectivamente dadas pelas equações (4.2) e (4.5).

Os gráficos dos riscos acumulados não paramétricos, obtidos das estimativas Kaplan meier da função de sobrevivência, e Weibull para cada uma das parcelas do Bloco 4 encontram-se na Figura 15, na qual verificou-se que

o risco acumulado para indivíduos da parcela $(1, 1)$ foi maior que o risco acumulado para indivíduos da parcela $(0, 1)$. Além disso, os riscos acumulados Weibull corresponderam a boas suavizações dos riscos acumulados não paramétricos.

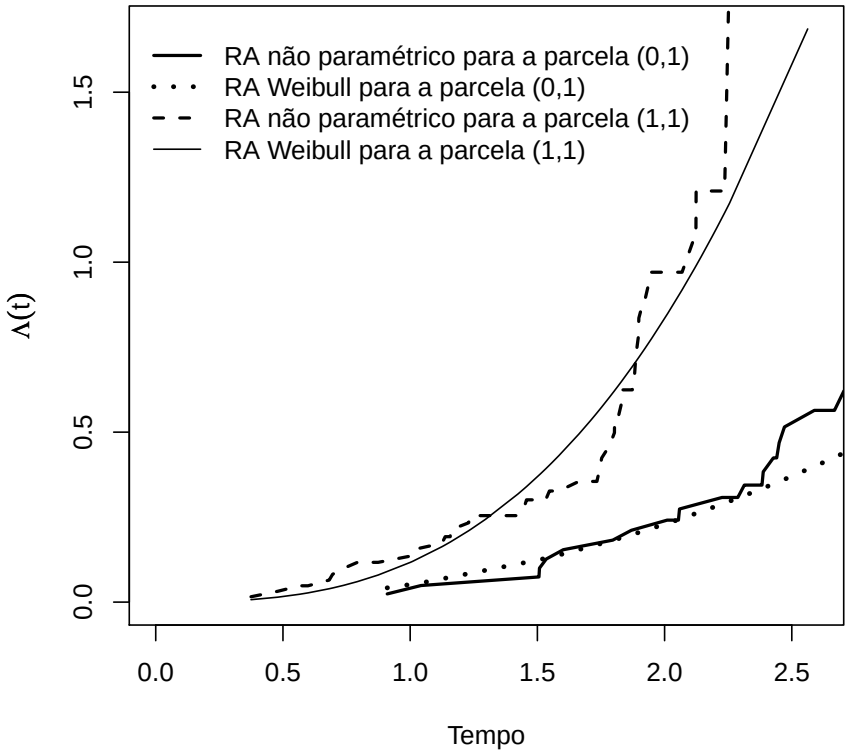

FIGURA 15 Estimativas dos riscos acumulados não paramétricos e Weibull para cada parcela do Bloco 4 em função do tempo.

Derivando a diferença entre as equações (4.5) e (4.2), estima-se a

função de regressão instantânea β_1 do Bloco 4:

$$\widehat{\beta}_1(t) = \frac{2,84}{2,13^{2,84}} t^{1,84} - \frac{2,15}{3,97^{2,15}} t^{1,15}. \tag{4.8}$$

O gráfico para esta expressão da estimativa $\widehat{\beta}_1$ em função do tempo, dado na Figura 16, indicou que o risco instantâneo para indivíduos da parcela $(1,1)$ é maior que o risco para indivíduos da parcela $(0,1)$ e esta diferença entre os riscos das parcelas é crescente. O gráfico também mostrou que em $t = 2.5$, por exemplo, $\widehat{\beta}_1 = 1,47$, aproximadamente. Então, em $t = 2.5$, em determinada unidade de tempo previamente definida, ter-se-iam em média $1,47$ falhas, isto é, 1 falha a cada $0,68$ unidades de tempo ($\frac{1}{1,47} = 0,68$) em média. Logo, indivíduos da parcela $(0,1)$ precisam de $0,68$ unidades de tempo a mais que indivíduos da parcela $(1,1)$ para experimentar o evento, em média.

A rotina utilizada para obter os resultados das Tabelas 8 a 11 e os gráficos das Figuras 5 a 16, encontra-se no Anexo C.

4.1.2 Teste de adequacidade da suavização do risco paramétrico ao risco acumulado não paramétrico.

Em consonância com os objetivos propostos, no que se refere aos desvios resultantes do ajuste do risco acumulado Weibull em relação aos riscos acumulados obtidos a partir da sobrevivência Kaplan Meier, avaliou-se a capacidade do teste de significância proposto na seção 3.1.2 em controlar o erro tipo I. Com este propósito, dada a hipótese de nulidade H_0 especificada na formalização do teste descrito na seção 3.1.2, as probabilidades empíricas referentes à ocorrência do erro tipo I foram obtidas por simulação

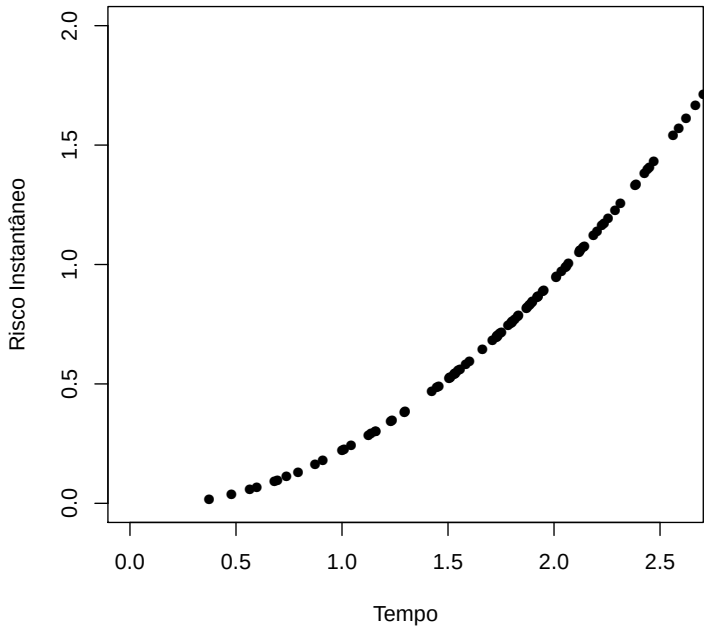

FIGURA 16 Estimativas $\widehat{\beta}_1$ da função de regressão instantânea do Bloco 4 em função do tempo.

Monte Carlo e pela implementação do método Bootstrap, considerando-se diferentes proporções de censura ($p = 0,30; 0,20; 0,10$) e diferentes tamanhos amostrais ($n = 30, 50, 60, 90$) em cada parcela. Os resultados obtidos encontram-se descritos na Tabela 12.

De acordo com os resultados encontrados na Tabela 12, mantendo-se o nível de significância fixado em 5%, pode-se observar que para uma porcentagem de censura definida em $p = 0,30$, em geral, o teste propiciou probabilidades empíricas mais condizentes com o nível de significância. Este fato foi notório, principalmente para tamanhos amostrais superiores a $n = 50$.

TABELA 12 Resultados obtidos para o cálculo da probabilidade empírica da ocorrência do controle do erro tipo I ($ptipo_1$).

		Parcela			
n	p	(0,0)	(0,1)	(1,0)	(1,1)
30	0,30	0,054	0,086	0,062	0,06
	0,20	0,106	0,07	0,062	0,066
	0,10	0,096	0,084	0,05	0,07
50	0,30	0,062	0,094	0,058	0,056
	0,20	0,058	0,05	0,074	0,078
	0,10	0,072	0,052	0,064	0,064
60	0,30	0,064	0,056	0,064	0,078
	0,20	0,056	0,044	0,058	0,048
	0,10	0,056	0,052	0,046	0,054
90	0,30	0,054	0,06	0,038	0,042
	0,20	0,056	0,068	0,058	0,052
	0,10	0,06	0,062	0,05	0,042

No contexto da análise de sobrevivência, de um modo geral, é esperado que a performance dos testes estatísticos seja prejudicada para altas porcentagens de censura e para grandes quantidades de empates. Reportando esta situação para os resultados (Tabela 12), percebeu-se uma certa "incoerência", pois à medida que a porcentagem de censura foi reduzida, em geral, o teste tende a não controlar o erro tipo I para as parcelas em estudo.

A ocorrência deste fato evidencia que a simples realização do *Bootstrap* de 1º nível poderá resultar em probabilidades imprecisas. Isto, supostamente, devido ao fato de no processo de reamostragem, para dada proporção de censuras fixa, as reamostras poderem conter diferentes proporções de censuras, sendo algumas delas maiores que a proporção de censuras da amostra original. Além disso, as reamostras possuem uma quantidade de empates maior ou igual à quantidade de empates da amostra original.

Mediante o problema detectado, uma alternativa viável para amenizá-lo foi a obtenção das probabilidades empíricas com a realização do *Bootstrap* de 2º nível. Nesse procedimento o p-valor obtido anteriormente foi corrigido de tal forma que foram consideradas apenas as subamostras obtidas no *Bootstrap* de 1º nível cujo p-valor era inferior ao nível nominal de significância. Dessa forma, acredita-se que as amostras selecionadas, de acordo com este critério, não foram afetadas pelo problema anteriormente mencionado. Os resultados são descritos na Tabela 13.

TABELA 13 Resultados obtidos para o cálculo da probabilidade empírica corrigida da ocorrência do controle do erro tipo I ($ptipo_{1c}$).

		Parcela			
n	p	(0,0)	(0,1)	(1,0)	(1,1)
30	0,30	0,056	0,05	0,066	0,03
	0,20	0,044	0,058	0,036	0,042
	0,10	0,04	0,026	0,026	0,022
50	0,30	0,046	0,066	0,044	0,042
	0,20	0,052	0,036	0,032	0,05
	0,10	0,026	0,032	0,026	0,05
60	0,30	0,042	0,06	0,044	0,064
	0,20	0,028	0,062	0,044	0,038
	0,10	0,026	0,024	0,03	0,022
90	0,30	0,034	0,054	0,042	0,038
	0,20	0,054	0,03	0,044	0,038
	0,10	0,036	0,046	0,044	0,028

Os resultados obtidos para as probabilidades empíricas corrigidas do controle do erro tipo I são, em geral, melhores, pois os p-valores obtidos estão menores e mais próximos do nível nominal de significância de 5%, indicando que o teste controlou o erro tipo I, incluindo as situações nas quais foram detectadas deficiências conforme discutido na Tabela 12. Além

disso, o controle do erro tipo I melhorou à medida que a porcentagem de censuras diminuiu, ao passo que para pequenas proporções de censuras o teste tornou-se conservativo.

A rotina utilizada para obter os resultados das tabelas 12 e 13 encontra-se no Anexo D.

4.2 Resultados para os dados reais.

De acordo com as especificações do conjunto de dados descrito na seção 3.2.2, supondo-se, por exemplo, que o pesquisador esteja interessado em avaliar o efeito do tratamento por via tópica quando se utiliza o tratamento por via sistêmica, foram analisadas apenas as observações refrentes às parcelas $(0,1)$ e $(1,1)$, cujos resultados serão aqui descritos. O procedimento foi o mesmo utilizado para os dados simulados (ver seção 4.1).

Para esse conjunto de dados, a função risco instantâneo do modelo de Aalen para o i-ésimo indivíduo é definida como $\lambda_i(t) = \beta_0(t) + \beta_1(t)T(t)$, sendo $\beta_0(t)$ o risco instantâneo para um animal submetido a tratamento apenas por via sistêmica, $\beta_1(t)$ a função de regressão referente ao efeito da covariável tratamento por via tópica e $T(t)$ a covariável Tratamento por via tópica. Os resultados obtidos para o modelo ajustado em $t = 317$ dias encontram-se descritos na Tabela 14. Nesta tabela observou-se um efeito significativo do tratamento por via tópica (T) no risco de cura ou controle da doença, quando os animais também foram tratados por via sistêmica.

TABELA 14 Estimativas obtidas em $t = 317$ dias para o modelo de Aalen ajustado aos dados de cães diagnosticados com otite externa.

Covariável	Coeficiente	Erro padrão	IC(95%)	Valor p
Intercepto	1,188	0,450	[0,305; 2,071]	0,002
T	1,050	0,538	[-0,004; 2,104]	0,009

A Figura 17 apresenta os gráficos das funções de regressão acumuladas e seus respectivos intervalos de 95% de confiança para o modelo de Aalen ajustado.

O primeiro gráfico da Figura 17 (gráfico Intercepto), permitiu interpretar o risco acumulado para os indivíduos da parcela $(0, 1)$, de modo que os cães tratados apenas por via sistêmica caracterizaram-se por um risco de cura ou controle da doença que se eleva gradativamente ao longo do tempo, principalmente nos primeiros dias. O segundo gráfico representa o efeito da covariável T, isto é, o risco adicional para cães tratados pelas duas vias em relação ao risco de cães tratados apenas pela via sistêmica. Este gráfico evidenciou que cães tratados pelas duas vias apresentaram um risco de cura/controle da doença maior que cães tratados apenas pela via sistêmica por aproximadamente 150 dias, sendo que após este período esta diferença se estabilizou e, portanto, os riscos tornaram-se similares.

A seguir, buscou-se a distribuição que melhor se ajustasse aos dados de cada uma das parcelas. Foram avaliados os modelos Exponencial, Weibull, Lognormal e Loglogístico.

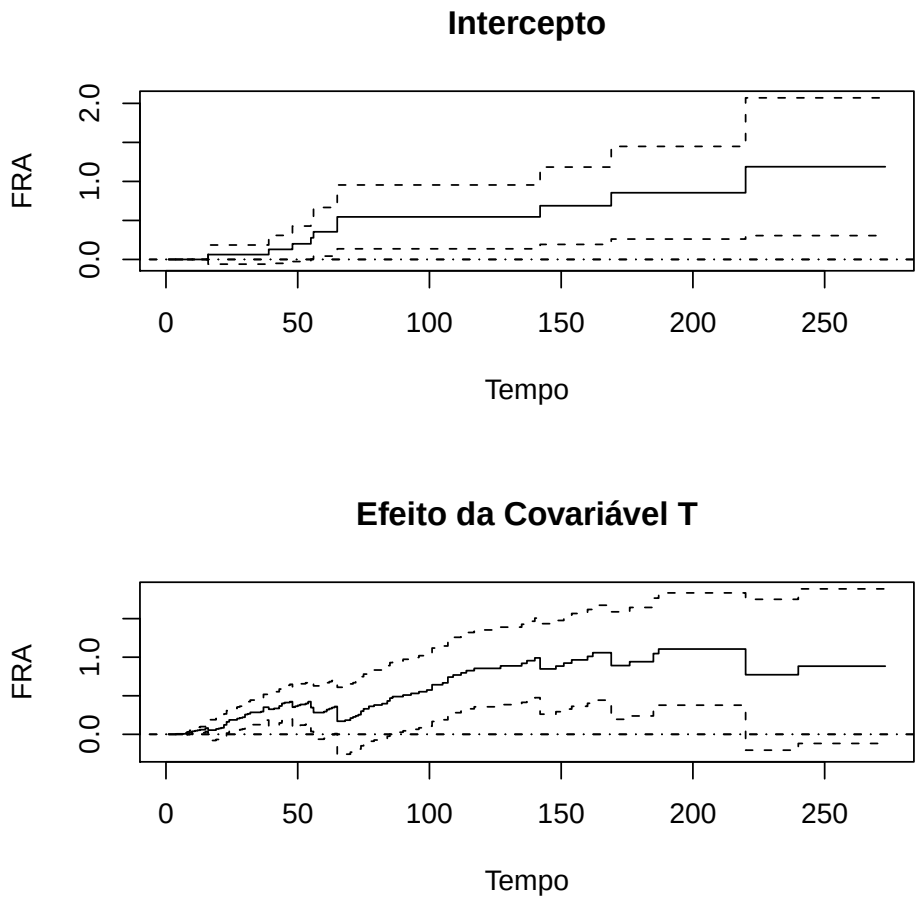

FIGURA 17 Funções de regressão acumulada (FRA) obtidas do modelo de Aalen ajustado para os dados de cães diagnosticados com otite externa *versus* o tempo em dias.

Considerando a natureza dos dados, percebeu-se que o risco de cura ou controle da doença é pequeno logo nos primeiros dias, elevando-se a medida em que o animal é tratado. Porém, há casos nos quais a otite externa é tratável, mas incurável e, assim, para tempos maiores, o risco de cura torna-se cada vez menor. Este comportamento nos remete, a princípio,

ao comportamento da função de risco de uma distribuição Lognormal.

Para ambas as parcelas, as estatísticas fornecidas pelos critérios de Akaike (AIC) e Bayesiano (BIC), descritas na Tabela 15, confirmam que o modelo Lognormal é adequado aos dados.

TABELA 15 Critérios de Akaike (AIC) e Bayesiano (BIC) para a seleção de modelos.

Modelo	AIC	BIC
Exponencial	1457,6	1460,6
Weibull	1459,4	1465,5
Lognormal	1429,6	1435,7
Loglogístico	1431,2	1437,3

A qualidade do modelo ajustado foi verificada por meio dos resíduos de Cox-Snell para cada uma das parcelas. Os gráficos dos mesmos encontram-se nas Figuras 18 e 19.

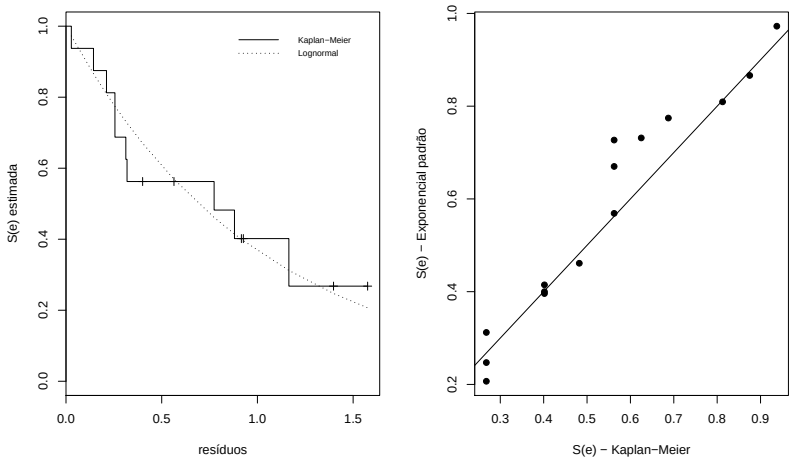

FIGURA 18 Gráficos dos resíduos de Cox-Snell para os dados da parcela (0,1).

No gráfico à direita da Figura 18, observaram-se alguns pontos distantes da reta. Isto ocorre devido à pequena quantidade de cães tratados apenas por via sistêmica. Apesar disto, pelo gráfico à esquerda da Figura 18 o modelo ajustado pode ser considerado satisfatório.

Na Figura 19, no gráfico à direita, observou-se que os resíduos aproximam-se da reta traçada, indicando que modelo ajustado foi adequado aos dados. Este resultado também pôde ser observado no gráfico à esquerda da Figura 19.

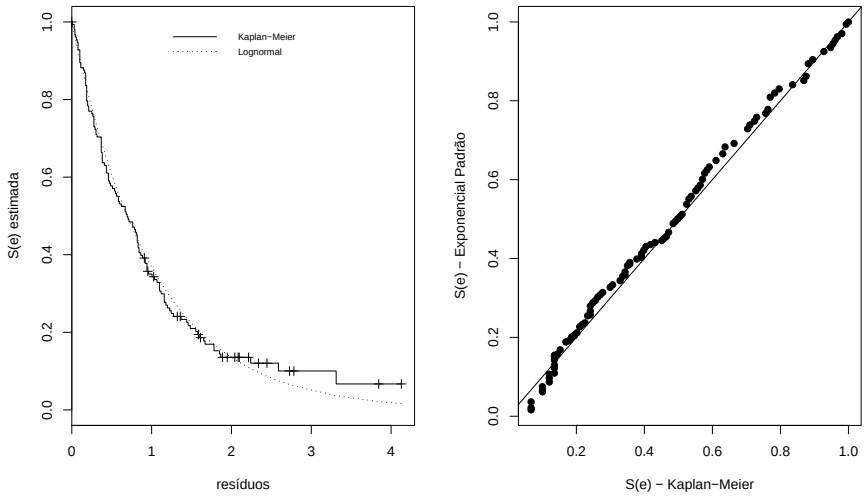

FIGURA 19 Gráficos dos resíduos de Cox-Snell para os dados da parcela (1,1).

As estimativas $\widehat{\mu}$ e $\widehat{\sigma}$ dos parâmetros da distribuição Lognormal ajustada foram respectivamente $4,127905$ e $1,085836$ para os cães tratados apenas por via sistêmica e $4,847469$ e $1,115060$ para os tratados pelas vias sistêmica e tópica simultaneamente. Assim, as expressões para as estimativas do

risco acumulado Lognormal nas referidas parcelas foram, respectivamente:

$$\widehat{\Lambda}_{(0,1)}(t) = -log\left[\Phi\left(\frac{-log(t) + 4,13}{1,09}\right)\right] \tag{4.9}$$

e

$$\widehat{\Lambda}_{(1,1)}(t) = -log\left[\Phi\left(\frac{-log(t) + 4,85}{1,12}\right)\right] \tag{4.10}$$

Os gráficos dos riscos acumulados Lognormal e dos riscos acumulados não paramétricos obtidos das estimativas Kaplan Meier da função de sobrevivência para cada uma das parcelas encontram-se na Figura 20. Observou-se nesta figura que o risco de cura ou controle da otite externa de cães submetidos ao tratamento por ambas as vias é maior que o risco de cães submetidos a tratamento apenas por via sistêmica, analogamente ao resultado obtido com o modelo de Aalen ajustado. Por inspeção, foi possível perceber que as curvas de riscos acumulados da distribuição Lognormal suavizaram as curvas de riscos acumulados não paramétricos. A distância observada em tempos maiores entre os riscos acumulados de animais tratados pelas vias tópica e sistêmica simultaneamente é devida aos indivíduos censurados.

Assim, tomando-se a diferença entre as equações (4.10) e (4.9) e derivando o resultado, obteve-se uma expressão para a estimativa da função de regressão instantânea $\beta_1(t)$, a qual representa o risco instantâneo adicional de cães tratados pelas vias tópica e sistêmica em relação aos cães tratados apenas pela via sistêmica.

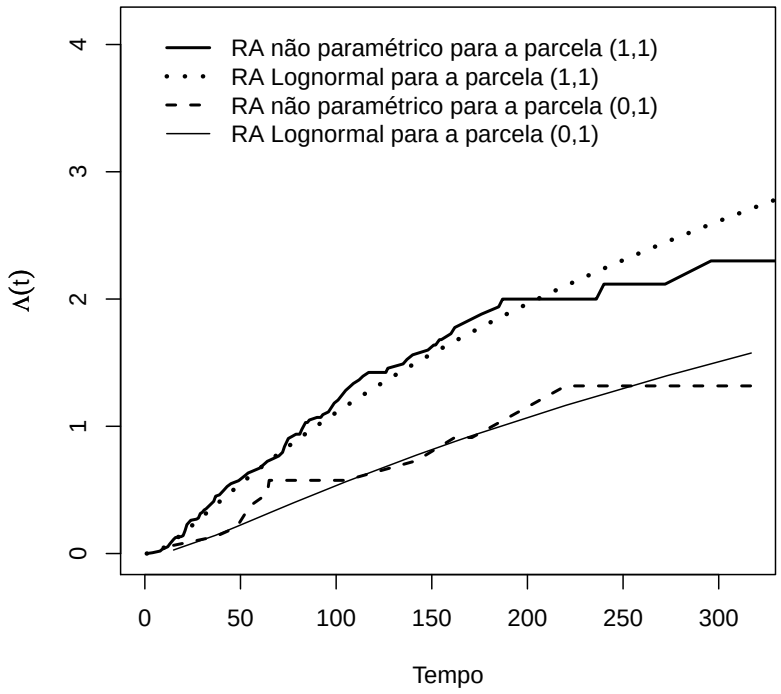

FIGURA 20 Estimativas dos riscos acumulados não paramétricos e Lognormal para cada parcela dos dados de otite externa em função do tempo.

Logo,

$$\widehat{\beta}_1(t) = \frac{\frac{1}{\sqrt{2\pi}t\sigma_1} exp\{\frac{-1}{2}\big(\frac{log(t)-\mu_1}{\sigma_1}\big)^2\}}{\Phi\big(\frac{-log(t)+\mu_1}{\sigma_1}\big)} - \frac{\frac{1}{\sqrt{2\pi}t\sigma_0} exp\{\frac{-1}{2}\big(\frac{log(t)-\mu_0}{\sigma_0}\big)^2\}}{\Phi\big(\frac{-log(t)+\mu_0}{\sigma_0}\big)}, \quad (4.11)$$

sendo μ_0, σ_0 e μ_1 σ_1 as estimativas dos parâmetros da distribuição Lognormal ajustada para os indivíduos submetidos ao tratamento sistêmico e ao tratamento conjunto, respectivamente.

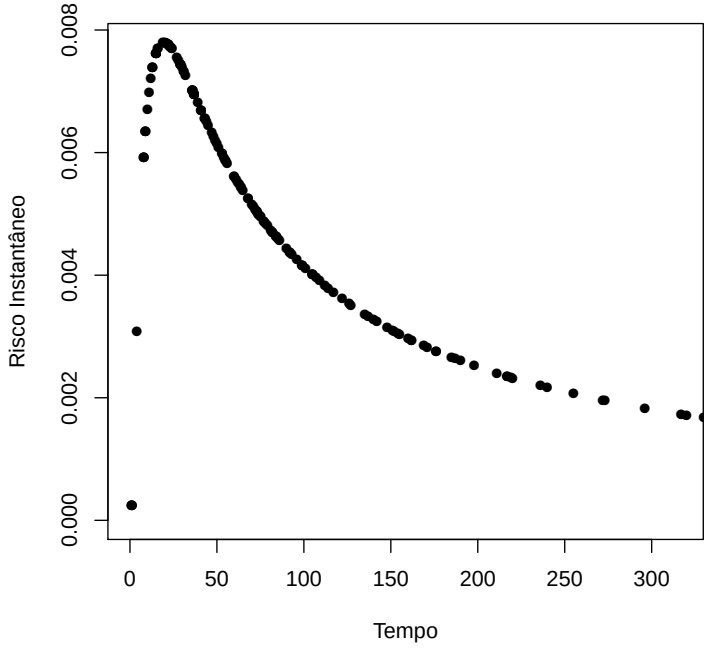

FIGURA 21 Estimativas da função de regressão instantânea β_1 em função do tempo.

O gráfico para as estimativas da função de regressão instantânea $\beta_1(t)$ *versus* o tempo é apresentado na Figura 21. Este gráfico mostrou que nos primeiros 20 dias o risco instantâneo de cura ou controle da doença é maior para cães tratados pelas duas vias que o risco para cães tratados apenas por via sistêmica. Após esse período essa diferença entre os riscos diminui e, com o passar do tempo, os riscos de cura ou controle tornam-se aproximadamente equivalentes. Além disso, em $t = 20$, $\widehat{\beta}_1 = 0,00779$, aproximadamente. Isto significa que em 1 dia ter-se-iam, em média, $0,00779$ curas, então tem-se 1 cura a cada 128 dias ($\frac{1}{0,00779} = 128,313$), em média. Logo, cães tratados apenas por via sistêmica precisam em média de 128

dias a mais que os cães tratados por ambas as vias (tópica e sistêmica) para serem curados ou terem seus quadros clínicos controlados.

Para verificar-se a adequacidade do ajuste da curva de riscos acumulados obtida da distribuição Lognormal ao risco acumulado não paramétrico obtido a partir das estimativas Kaplan-Meier, utilizou-se o teste proposto na seção 3.1.2. Foram feitas 2000 reamostragens no *Bootstrap* de 1º nível e 500 no de 2º nível. Os resultados obtidos para $prob_1$ e $prob_{cor}$ encontram-se na Tabela 16.

TABELA 16 Resultados das probabilidades de significância do teste proposto para verificar a proporção de suavizações adequadas.

	Parcela	
	(1,1)	(0,1)
$prob_1$	0,749	0,3435
$prob_{cor}$	0,8345	0,165

Os valores apresentados na Tabela 16 correspondem às probabilidades que permitem ao pesquisador tomar decisão a favor de "H_0: risco acumulado paramétrico obtido da distribuição Lognormal suaviza o risco acumulado Kaplan Meier." Exemplificando a interpretação desta probabilidade, pode-se verificar que, para a parcela (1,1), o nível de significância a ser utilizado poderia ser de no máximo 83%, aproximadamente, e, para a parcela $(0,1)$, de no máximo 16%, aproximadamente. Respeitando estes valores o pesquisador poderá tomar a decisão a favor de H_0.

As rotinas utilizadas para obter os resultados das Tabelas 14 e 16 e das Figuras 17 a 21 encontram-se nos Anexos E e F.

5 CONCLUSÕES

Na presença de covariáveis dicotômicas, as funções de risco acumulado do modelo de Aalen podem ser ajustadas por funções risco acumulado conhecidas. Assim, o risco instantâneo do modelo de Aalen adicional de um grupo em relação a outro pode ser estimado estratificando-se os dados em parcelas. A simplicidade e a facilidade de entendimento e interpretação são as grandes vantagens dessa técnica.

A metodologia pode ser utilizada para dados com qualquer quantidade de covariáveis categóricas. Porém, é notório que quanto maior o número de covariáveis, maior o número de parcelas e menor o número de indivíduos por parcela.

O Teste proposto para verificar a adequacidade do ajuste do risco acumulado paramétrico ao risco acumulado obtido a partir da estimativa Kaplan Meier da função de sobrevivência mostrou-se adequado no que se refere ao controle do erro tipo I.

Foi possível observar que a utilização do *Bootstrap* duplo proporcionou melhores resultados que o *Bootstrap* simples, tornando os p-valores em geral mais próximos do nível nominal de significância de 5%.

Em um estudo realizado com cães diagnosticados com otite externa, a metodologia proposta foi útil para investigar o efeito do tratamento por via tópica em relação ao tratamento por via sistêmica no risco de cura ou controle de otite externa causada por síndrome do banho-tosa em cães. Cães tratados apenas por via sistêmica apresentaram um risco de cura ou controle da doença que eleva gradativamente com o passar do tempo, principalmente nos primeiros dias. Além disso, animais tratados simultaneamente pelas

duas vias apresentaram um risco de cura ou controle da doença maior que animais tratados apenas por via sistêmica.

REFERÊNCIAS BIBLIOGRÁFICAS

AALEN, O. O. Further results on the non-parametric linear regression model in survival analysis. **Statistics in Medicine**, New York, v. 12, n. 1, p. 1569-1588, Mar. 1993.

AALEN, O. O. A linear regression model for the analysis of life times. **Statistics in Medicine**, New York, v. 8, n. 8, p. 907-925, Oct. 1989.

AALEN, O. O. A model for nonparametric regression analysis of counting processes. **Lecture Notes in Statistics**, New York, v.2, n. 1, p. 1-25, 1980.

AALEN, O. O. Nonparametric inference for a family of counting processes. **Annals of Statistics**, Hayward, v. 6, n. 4, p. 701-726, July 1978.

ALLISON, P. D. **Survival analysis using the SAS system**: a pratical guide. Cary: SAS Institute, 1995. 292 p.

CARVALHO, M. S.; ANDREOZZI V. L.; CODEÇO, C. T.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de sobrevida**: teoria e aplicações em saúde. 20. ed. Rio de Janeiro: FIOCRUZ, 2005. 396 p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada.** São Paulo: E. Blücher, 2006. 370 p.

COX, D. R. Regression models and life tables: with discussion. **Journal Royal Statistical Society, Series B, Methodological**, London, v. 34, n. 2, p. 187-220, 1972.

DAVISON, A. C.; HINKLEY, D. V. **Bootstrap methods and their application.** New York: Cambridge University, 1997. 592 p.

EFRON, B. Bootstrap method: another look at the jackknife. **Annals of Statistics**, Hayward, v. 7, n. 1, p. 1-26, 1979.

EFRON, B. Censored data and the bootstrap. **Journal of the American Statistical Association**, New York, v. 76, n. 374, p. 312-319, June 1981.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap.** New York: Chapman & Hall, 1993. 456 p.

FLEMING, T. R.; HARRINGTON D. P. **Counting processes and survival analysis.** New York: John Wiley, 1991. 448 p.

FOGO, J. C. **Modelo de regressão para um processo de renovação Weibull com termo de fragilidade.** 2007. 184 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba.

GANDY, A.; JENSEN, U. On goodness-of-fit for Aalen's additive risk model. **Scandinavian Journal of Statistics**, Stockholm, v. 32, n. 4, p. 425-445, Sept. 2005.

GIAROLA, L. T. P. ; KATAOKA, V. Y. ; VIVANCO, M. J. F. Aplicação do modelo aditivo de Aalen no estudo do tempo de tratamento até o óbito em pacientes com insuficiência renal crônica da cidade de Lavras, MG. In: REUNIÃO ANUAL DA REGIÃO BRASILEIRA DA SOCIEDADE INTERNACIONAL DE BIOMETRIA, 51., 2006, Botucatu. **Programas e Resumos**... Botucatu: UNESP, 2006. p. 129.

GRUNKEMEIER, G. L.; JIN, R.; IM, K.; HOLUBKOV, R.; KENNARD, E. D.; SCHAFF, H. V. Time-related risk of St. Jude Silzone heart valve. **European Journal of Cardiology Thoracic Surgery**, Oxford, v. 30, n. 1, p. 20-27, Jan. 2006.

HENDERSON, R.; MILNER, A. Aalen plots under proportional hazards. **Applied Statistic**, Oxford, v. 40, n. 3, p. 401-409, June 1991.

HOSMER, D. W.; LEMESHOW, S. **Applied survival analisys**: regression modeling of time to event data. New York: J. Wiley, 1999. 408 p.

KAPLAN, E.L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, New York, v. 53, n. 282, p. 457-481, June 1958.

KLEIN, J. P.; MOESCHBERGUER, M. L. **Survival analysis**: techinics for censored and truncated data. New York: Springer, 1997. 459 p.

LAWLESS, J. F. **Statistical models and methods for lifetime data.** New York: J. Wiley, 1982. 592 p.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias.** 2. ed. São Paulo: EDUSP, 2006. 428 p.

MAU, J. On a graphical method for the detection of time-dependent effects of covariates in survival data. **Journal of the Royal Statistical Society, Series C, Applied Statistics**, London, v. 35, n. 3, p. 245-255, June 1986.

MOOD, A. M., GRAYBILL, F. A., BOES, D. C. **Introduction to the theory of statistics.** 3.ed. New York: J. Wiley, 1974. 480 p.

PEREIRA, T. L. **Modelo de riscos proporcionais e aditivos para o tratamento de covariáveis dependentes do tempo.** 2004. 63 p. Dissertação (Mestrado em Estatística) - Universidade Federal de Pernambuco, Recife.

R DEVELOPMENT CORE TEAM. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2008. Disponível em: <http://www.R-project.org>. Acesso em: 10 ago. 2009.

ROSA, F. H. F. P.; PEDRO JÚNIOR, V. A. **MAE514** - Introdução à análise de Sobrevivência e aplicações. Disponível em: <http://feferraz.net/br/Listas/Analise_de_ Sobrevivência_-_Lista 1.> Acesso em: 19 de Maio de 2008.

VALENÇA, D. M. **O modelo de regressão gama generalizada para discriminar entre modelos paramétricos de tempo de vida.** 1994. 114 f. Dissertação (Mestrado em Estatística) - Universidade Estadual de Campinas, Campinas.

ANEXOS

ANEXO A Rotina para simulação de dados de sobrevivência segundo a distribuição Weibull.

Sejam n o número de observações e p a proporção de censuras.

```
> # ################## simulação de dados weibul ########################### #
>
>
> #########pacotes necessários######################
> require (survival)
> source("addreg.R")
>   p=0.3    # ##### proporção média de censuras ##### #
>   alfa=c(8.5,4.5,4,2) # parcelas (0,0); (1,0); (0,1); (1,1) respectivamente. #
>   gama=c(1.5,2,2.2,3.5)
>   n=c(90,82,42,64)
> ### função para gerar a weibull
>
>  rweib <- function(n,gama=1,alfa=1)
+   {
+      return((-(alfa^gama)*(log(runif(n,min=0,max=1))))^(1/gama))
+   }
> #  ################ gera uma amostra de tamanho n ################### #
>
>  est=0
>  auxdados=matrix(0,1,3)
> for (i in 1:4)
+ {
+    est=est+1
+    st=as.matrix(rep(est,n[i]))
+    cens=as.matrix(rep(0,n[i]))
+    falhas <- rweib(n[i],gama[i],alfa[i])
```

```
+    censuras <- rweib(n[i],gama[i],((alfa[i])*(((1-p)/p))^(1/gama[i]))) # veja Anexo B. #
+
+    # compara os valores e assume o menor; tempos de falha#
+
+     Y <- pmin(falhas,censuras)
+   for (j in 1:length(Y))
+   {
+              if (Y[j]==falhas[j]) cens[j,1]=1
+   }
+      dados=cbind(Y,cens,i)
+      auxdados=rbind(auxdados,dados)
+ }
>    dadest=auxdados[2:nrow(auxdados),1:3]
>  ### Formação dos blocos ####
>
>      pe1=dadest[1:n[1],1:2]   # (alfa=8.5, 1.5)
>      pe2=dadest[(n[1]+1):(n[1]+n[2]),1:2]
>      pe3=dadest[(n[1]+n[2]+1):(n[1]+n[2]+n[3]),1:2]
>      pe4=dadest[(n[1]+n[2]+n[3]+1):(n[1]+n[2]+n[3]+n[4]),1:2]
>      x1=as.matrix(rep(0,n[1]))
>      x2=x1
>      par1=cbind(pe1,x1,x2)
>      x3=as.matrix(rep(1,n[2]))
>      x4=as.matrix(rep(0,n[2]))
>      par2=cbind(pe2,x3,x4)
>      x5=as.matrix(rep(0,n[3]))
>      x6=as.matrix(rep(1,n[3]))
>      par3=cbind(pe3,x5,x6)
>      x7=as.matrix(rep(1,n[4]))
>      x8=x7
>      par4=cbind(pe4,x8,x7)
>      bloco1=rbind(par1,par3)
```

```
>      bloco2=rbind(par2,par4)
>      bloco3=rbind(par1,par2)
>      bloco4=rbind(par3,par4)
>           ####nomeando as colunas de cada bloco####
> tempo1=bloco1[,1];status1=bloco1[,2];x1b1=bloco1[,3];x2b1=bloco1[,4]
> tempo2=bloco2[,1];status2=bloco2[,2];x1b2=bloco2[,3];x2b2=bloco2[,4]
> tempo3=bloco3[,1];status3=bloco3[,2];x1b3=bloco3[,3];x2b3=bloco3[,4]
> tempo4=bloco4[,1];status4=bloco4[,2];x1b4=bloco4[,3];x2b4=bloco4[,4]
> bloco1=as.data.frame(bloco1)
> bloco2=as.data.frame(bloco2)
> bloco3=as.data.frame(bloco3)
> bloco4=as.data.frame(bloco4)
```

ANEXO B Desenvolvimento teórico para a construção das rotinas para simulação de dados de sobrevivência segundo a distribuição Weibull.

Deseja-se simular dados provenientes de uma distribuição Weibull na presença de censuras também provenientes de uma distribuição Weibull. Considerando-se disponível um gerador de números pseudoaleatórios uniformes em $(0, 1)$, pode-se utilizar a transformação integral da probabilidade (Mood et al., 1974; Magalhães, 2006).

Seja U uma variável aleatória uniforme e para qualquer função de distribuição contínua F define-se a variável aleatória X por $X = F^{-1}(U)$, então X tem função de distribuição F. Para o caso de uma variável aleatória Weibull com parâmetros α e γ, tem-se que

$$F(x) = 1 - exp\{-(\frac{x}{\alpha})^{\gamma}\}.$$

Então,

$$u = 1 - exp\Big\{-\Big(\frac{x}{\alpha}\Big)^{\gamma}\Big\},$$

donde se segue que

$$x = [-\alpha^{\gamma} log(1-u)]^{\frac{1}{\gamma}}$$

Assim, $F^{-1}(U) = [-\alpha^{\gamma} log(1-u)]^{\frac{1}{\gamma}}$ segue uma distribuição Weibull com parâmetros γ e α. Como $1-U$ também possui distribuição uniforme $(0,1)$, basta considerar a transformação $g(U) = [-\alpha^{\gamma} log(u)]^{\frac{1}{\gamma}}$. Desta forma geram-se números pseudoaleatórios com distribuição Weibull (γ, α).

Suponha-se que os dados estejam sujeitos a censuras aleatórias, também segundo uma distribuição Weibull e admita-se que estas censuras são independentes. Considere-se que p é, em média, a proporção de censuras $(0 < p < 1)$.

Sejam X_1 e X_2 variáveis aleatórias independentes com distribuição Weibull e vetores de parâmetros (γ_1, α_1) e (γ_2, α_2), respectivamente. Seja $Z = X_2 - X_1$. Logo, $P(Z \leqslant 0) = P((x_1, x_2) \in B_Z)$, sendo $B_Z = \{(x_1, x_2) \in R^2 | x_2 - x_1 \leqslant 0, x_1 \leqslant 0, x_2 \leqslant 0\}$. Então,

$$\begin{aligned} F_Z(z) &= \int\int_{B(Z)} f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 \\ &= \int_0^{\infty}\int_0^{x_1} \frac{\gamma_1}{\alpha_1}\Big(\frac{x_1}{\alpha_1}\Big)^{\gamma_1 - 1} exp\Big\{-\Big(\frac{x_1}{\alpha_1}\Big)^{\gamma_1}\Big\} \frac{\gamma_2}{\alpha_2}\Big(\frac{x_2}{\alpha_2}\Big)^{\gamma_2 - 1} exp\Big\{-\Big(\frac{x_2}{\alpha_2}\Big)^{\gamma_2}\Big\} dx_2 dx_1 \\ &= 1 - \frac{\gamma_1}{{\alpha_1}^{\gamma_1}} \int_0^{\infty} {x_1}^{\gamma_1 - 1} exp\Big\{-\Big(\frac{x_1}{\alpha_1}\Big)^{\gamma_1} - \Big(\frac{x_1}{\alpha_2}\Big)^{\gamma_2}\Big\} dx_1 \end{aligned}$$

A integral acima parece intratável. Mas, tomando-se $\gamma_1 = \gamma_2 = \gamma$, tem-se

que

$$\int_0^{\infty} {x_1}^{\gamma_1 - 1} exp\Big\{-\Big(\frac{x_1}{\alpha_1}\Big)^{\gamma_1} - \Big(\frac{x_1}{\alpha_2}\Big)^{\gamma_2}\Big\} dx_1 = \frac{1}{\gamma\Big(\frac{1}{{\alpha_1}^{\gamma}} + \frac{1}{{\alpha_2}^{\gamma}}\Big)}.$$

Logo,

$$F_Z(z) = \frac{{\alpha_1}^{\gamma}}{{\alpha_1}^{\gamma} + {\alpha_2}^{\gamma}}.$$

Considerando-se X_2 como o tempo de censura de determinado indivíduo e X_1 seu tempo de falha, deseja-se que

$$p = \frac{{\alpha_1}^{\gamma}}{{\alpha_1}^{\gamma} + {\alpha_2}^{\gamma}}.$$

Logo,

$$\alpha_2 = \alpha_1 \Big(\frac{1-p}{p}\Big)^{\frac{1}{\gamma}}$$

ANEXO C Rotina para a construção dos gráficos apresentados na seção 4.1.

Será escrita aqui apenas a rotina utilizada para o Bloco 1, pois para os outros blocos são necessárias pequenas adaptações.

```
>  ######Análise do Bloco1############
>
> ### Ajuste do modelo Aalen ###
>
> fitbl1<-addreg(Surv(tempo1,status1)~x2b1,bloco1)

> plot(fitbl1,xlab="Tempo",ylab="FRA",labelofvariable=c("Intercepto","Efeito da covariável Z2"))
> ###Ajuste não paramétrico (Kaplan Meier) ###
>
> ekmBL1<-survfit(Surv(tempo1,status1)~x2b1)
> ekm00<-survfit(Surv(tempo1[x2b1==0],status1[x2b1==0])~1) #KM parcela (0,0)#
```

```
> ekm01<-survfit(Surv(tempo1[x2b1==1],status1[x2b1==1])~1) #KM parcela (0,1)#
> t00<-ekm00$time
> t01<-ekm01$time
> sob00<-ekm00$surv
> sob01<-ekm01$surv
> racu00<--log(sob00)
> racu01<--log(sob01)
>   ###Ajuste Weibull###
>
> ##Parcela(0,0)##
>
>       ajust00<-survreg(Surv(tempo1[x2b1==0],status1[x2b1==0])~1,dist='weibull')
>       alpha00<-exp(ajust00$coefficients[1])
>       gama00<-1/ajust00$scale
>       s00<-exp(-(t00/alpha00)^gama00)
>       rac00<--log(s00)
> ##Parcela (0,1)##
>
>       ajust01<-survreg(Surv(tempo1[x2b1==1],status1[x2b1==1])~1,dist='weibull')
>       alpha01<-exp(ajust01$coefficients[1])
>       gama01<-1/ajust01$scale
>       s01<-exp(-(t01/alpha01)^gama01)
>       rac01<--log(s01)
> ## Curvas de riscos acumulados não paramétricos e Weibull ##
> plot(0, 0, type='n', xlim=range(0,t00),ylim=range(0,rac00),col=2,xlab="Tempo",ylab=expression(Lambda*(
> lines(t00,racu00,lwd=2)
> lines(t00,rac00,lty=3, lwd=3)
> lines(t01,racu01,lty=2,lwd=2)
> lines(t01,rac01)
> leg <- c('RA não paramétrico para a parcela (0,0) ',
+ 'RA Weibull para a parcela (0,0)',
+ 'RA não paramétrico para a parcela (0,1) ',
```

```
+ 'RA Weibull para a parcela (0,1) ')
> legend(3.65,lty=c(1,3,2,1),lwd=c(2,3,2,1),leg, bty="n")#
> ### Função de regressão instantânea ###
>
> r101<-(gama01/(alpha01^gama01))*tempo1^(gama01-1)#risco instantâneo para o estrato (0,1)#
> r100<-(gama00/(alpha00^gama00))*tempo1^(gama00-1)#derivada do risco acumulado##
> r1<-r101-r100
> plot(0, 0, type='n', xlim=range(0,tempo1), ylim=range(0,r1),xlab="Tempo",ylab="Risco Instantâneo", lty
> points(tempo1,r1, pch=16)
```

ANEXO D Programa para o cálculo da probabilidade do erro Tipo I para o teste proposto na seção 3.1.2

```
> #########pacote necessário######################
>
> require (survival)
> ### Valores paramétricos ############################## #
>
>   p=0.3    # ##### proporção média de censuras ##### #
>   alfa=c(8.5,4.5,4,2)
>   gama=c(1.5,2,2.2,3.5)
>   n=c(30,30,30,30)
>   taxa=0  ; taxacor=0; nsim=5
> ### função para gerar a weibull
>
>  rweib <- function(n,gama=1,alfa=1)
+   {
+      return((-(alfa^gama)*(log(runif(n,min=0,max=1))))^(1/gama))
+   }
> ##### gera uma amostra de tamanho n, controlando as censuras
>
> con_cens=function(n,alfa,gama,p)
```

```
+ {
+ for (i in 1:4)
+ {
+    est=est+1
+    st=as.matrix(rep(est,n[i]))
+    cens=as.matrix(rep(0,n[i]))
+    falhas <- rweib(n[i],gama[i],alfa[i])
+    censuras <- rweib(n[i],gama[i],((alfa[i])*(((1-p)/p))^(1/gama[i])))
+    # compara os valores e assume o menor; tempos de falha#
+     Y <- pmin(falhas,censuras)
+   for (j in 1:length(Y))
+   {
+        if (Y[j]==falhas[j]) cens[j,1]=1
+   }
+      dados=cbind(Y,cens,i)
+      auxdados=rbind(auxdados,dados)
+
+   # #### estimativas do parâmetro de cada estrato ###### #
+   }
+    return (auxdados)
+   }
> form_bl=function(dadest,n)
+ {
+
+  ### Formação dos blocos ####
+
+      pe1=dadest[1:n[1],1:2]   # (alfa=8.5, 1.5)
+      pe2=dadest[(n[1]+1):(n[1]+n[2]),1:2]
+      pe3=dadest[(n[1]+n[2]+1):(n[1]+n[2]+n[3]),1:2]
+      pe4=dadest[(n[1]+n[2]+n[3]+1):(n[1]+n[2]+n[3]+n[4]),1:2]
+      x1=as.matrix(rep(0,n[1]))   ;     x2=x1
+      par1=cbind(pe1,x1,x2)       ;     x3=as.matrix(rep(1,n[2]))
```

```
+      x4=as.matrix(rep(0,n[2]))   ;     par2=cbind(pe2,x3,x4)
+      x5=as.matrix(rep(0,n[3]))   ;     x6=as.matrix(rep(1,n[3]))
+      par3=cbind(pe3,x5,x6)       ;     x7=as.matrix(rep(1,n[4]))
+      x8=x7                       ;     par4=cbind(pe4,x8,x7)
+      blco1=rbind(par1,par3)   ;      blco2=rbind(par2,par4)
+      blco3=rbind(par1,par2)   ;      blco4=rbind(par3,par4)
+
+          #### nomeando as colunas de cada bloco ####
+
+      tempo1=blco1[,1];status1=blco1[,2];x1b1=blco1[,3];x2b1=blco1[,4]
+      tempo2=blco2[,1];status2=blco2[,2];x1b2=blco2[,3];x2b2=blco2[,4]
+      tempo3=blco3[,1];status3=blco3[,2];x1b3=blco3[,3];x2b3=blco3[,4]
+      tempo4=blco4[,1];status4=blco4[,2];x1b4=blco4[,3];x2b4=blco4[,4]
+
+      bl1=as.data.frame(blco1)
+      bl2=as.data.frame(blco2)
+      bl3=as.data.frame(blco3)
+      bl4=as.data.frame(blco4)
+
+    return (list(bloco1=bl1,bloco2=bl2,bloco3=bl3,bloco4=bl4,parcel1=par1,parcel2=par2,parcel3=par3,pa
+  }
>  ajweib=function(parcela,tempokm)
+   {
+      tempo=parcela[,1]
+      status=parcela[,2]
+
+     ###Ajuste Weibull###
+
+     ajust<-survreg(Surv(tempo,status)~1,dist='weibull')
+     alpha<-exp(ajust$coefficients[1])
+     gama<-1/ajust$scale
+     cbind(alpha,gama)
```

```
+     s<-exp(-(tempokm/alpha)^gama)
+     rac<--log(s)
+     return (riscoacw=rac)
+
+    }
> ajkm=function(parcela)
+  {
+
+   ### Ajuste não paramétrico (Kaplan Meier) ###
+
+     tempo=parcela[,1]
+     status=parcela[,2]
+
+     ekmp1<-survfit(coxph(Surv(tempo,status)~1,method="efron"))
+     tp1<-ekmp1$time    ; auxtp1=c(0,rep(length(tp1)))
+     sobp1<-ekmp1$surv  ;  auxsobp1=c(0,rep(length(sobp1)))
+
+   for (j in 1:length(sobp1))
+    {
+       if (sobp1[j]!=0)
+          {
+            auxsobp1[j]=sobp1[j]
+            auxtp1[j]=tp1[j]
+          }
+     }
+     racp1<--log(auxsobp1)
+     return (list(racup1=racp1,tempo1=auxtp1))
+   }
> for (f in 1:nsim)
+ {
+    est=0
+    auxdados=matrix(0,1,3)
```

```
+    auxdados=con_cens(n,alfa,gama,p)
+    dadest=auxdados[2:nrow(auxdados),1:3]
+    resbloco=form_bl(dadest,n)
+
+  ### dados para chamar o ajuste do modelo de Aalen ##########
+
+  ### Bloco 1 a cov=resbloco$bloco1[,4] ídem para bloco2
+
+  ### Bloco 3 a cov=resbloco$bloco1[,3] ídem para bloco4
+
+ ####### Dados a alterar no programa ########## #
+
+   nb2=5 ; nb1=10
+   par=resbloco$parcel4 ## Para outra parcela mude o indice
+   chama_km=ajkm(par)
+   ti=chama_km$tempo
+   chama_weib=ajweib(par,ti) ## Para outra parcela mude o indice
+
+
+ # ########### Estatística do teste de suavização ############## #
+
+    difpar=(max(abs(chama_weib-chama_km$racup1)))/sd(chama_km$racup1)
+
+ # ###################### Rotina Bootstrap ##################### #
+
+ contest=0  ; contest2=0  ; contest3=0
+
+ # #### Alterar numeros de reamostragem #### #
+
+    estb1=length(nb1)
+
+ # ################ fim de alteração #########
```

```
+
+ for (s in 1:nb1)
+ {
+  estb2=0 ; contest2=0 ; mpar=matrix(0,1,4)
+  for (k in 1:nrow(par))
+  {
+     u=round(runif(1,1,nrow(par)))
+     auxpar=par[u,]
+     mpar=rbind(mpar,auxpar)
+  }
+    mparaux=mpar[2:nrow(mpar),1:4]
+    chama_kmb1=ajkm(mparaux)
+    tib=chama_kmb1$tempo
+    chama_weibb=ajweib(mparaux,tib)
+    difparb=(max(abs(chama_weibb-chama_kmb1$racup1)))/sd(chama_kmb1$racup1)
+    #if (difparb<difpar) contest=contest+1
+
+  for (i in 1:nb2)
+ {
+   mpar2=matrix(0,1,4)
+  for (k1 in 1:nrow(mparaux))
+   {
+      u2=round(runif(1,1,nrow(mparaux)))
+      auxpar2=mparaux[u2,]
+      mpar2=rbind(mpar2,auxpar2)
+   }
+    mparaux2=mpar2[2:nrow(mpar2),1:4]
+    chama_kmb2=ajkm(mparaux2)
+    tib2=chama_kmb2$tempo
+    chama_weibb2=ajweib(mparaux2,tib2)
+    difparb2=(max(abs(chama_weibb2-chama_kmb2$racup1)))/sd(chama_kmb2$racup1)
+    estb2[i]=difparb2
```

```
+ }
+   estb1[s]=difparb
+ }
+
+ #resultados=function (est,p1,p2)
+
+ conta=0  ; prob2=length(nb1) ; conta3=0
+   ref=difpar
+  for (i in 1:length(estb1))
+  {
+   if (estb1[i]<=ref) conta=conta+1
+  }
+
+  ### bootstrap ###
+
+    prob1=conta/nb1
+ for (j in 1:length(estb1))
+ {
+    conta2=0
+    refb=estb1[j]
+ for (h in 1:length(estb2))
+  {
+      if (estb2[h]<=refb) conta2=conta2+1
+  }
+     prob2[j]=conta2/nb2  ### bootstrap 2 nivel
+  }
+
+ ## prob corrigido  ###
+
+ for (k in 1:length(prob2))
+ {
+       if (prob2[k]<=prob1) conta3=conta3+1
```

```
+ }
+     probcorr=conta3/nb1
+  if (probcorr<0.05) taxacor=taxacor+1
+  if (prob1<0.05) taxa=taxa+1
+ }
>   ptipo1c=taxacor/nsim
>   ptipo1=taxa/nsim
```

ANEXO E Programa utilizado com o conjunto de dados reais para estimação da função de regressão instantânea.

```
#########pacotes necessários######################

require(survival)
source("addreg.R")

################################################

otiteS<-read.table("S=1.txt", h=T) attach(otiteS)

## AALEN  ##

fitblS<-addreg(Surv(tempo,cens)~T,otiteS)
plot(fitblS,xlab="Tempo",ylab="FRA",labelofvariable=c("Intercepto","Efeito
da Covariável T"))

###Ajuste não paramétrico (Kaplan Meier) ###

ekm11<-survfit(Surv(tempo[T==1],cens[T==1])~1) #KM parcela (1,1)#
ekm01<-survfit(Surv(tempo[T==0],cens[T==0])~1) #KM parcela (0,1)#
t11<-ekm11$time t01<-ekm01$time sob11<-ekm11$surv sob01<-ekm01$surv
racu11<--log(sob11) racu01<--log(sob01)
```

```
##Parcela(1,1): ajuste Lognormal##

      ajust11<-survreg(Surv(tempo[T==1],cens[T==1])~1,dist='lognorm')
      alpha11<-ajust11$coefficients[1]
      gama11<-ajust11$scale
      s11<-pnorm((-log(t11)+alpha11)/gama11)
      rac11<--log(s11)

##Parcela (0,1): ajuste Lognormal##

      ajust01<-survreg(Surv(tempo[T==0],cens[T==0])~1,dist='lognorm')
      alpha01<-ajust01$coefficients[1]
      gama01<-ajust01$scale
      s01<-pnorm((-log(t01)+alpha01)/gama01)
      rac01<--log(s01)

## Curvas de riscos acumulados não paramétricos e Lognormal ##
plot(0, 0, type='n',
xlim=range(0,t11),ylim=range(0,rac11),col=2,xlab="Tempo",ylab=expression(Lambda*(t)),pch=16)
lines(t11,racu11,lwd=2)
lines(t11,rac11,lty=3, lwd=3)
lines(t01,racu01,lty=2,lwd=2)
lines(t01,rac01)
leg <- c('RA não
paramétrico para a parcela (1,1) ', 'RA Lognormal para a parcela
(1,1)', 'RA não paramétrico para a parcela (0,1) ', 'RA Lognormal
para a parcela (0,1) ')
legend(0.2,4.2,lty=c(1,3,2,1),lwd=c(2,3,2,1),leg, bty="n")#

### Função de regressão instantânea ###

f11<-(1/(sqrt(2*pi)*tempo*gama11))*exp((-1/2)*((log(tempo)-alpha11)/gama11)^2)
```

```
s11<-pnorm((-log(tempo)+alpha11)/gama11) r11<-f11/s11

f01<-(1/(sqrt(2*pi)*tempo*gama01))*exp((-1/2)*((log(tempo)-alpha01)/gama01)^2)
s01<-pnorm((-log(tempo)+alpha01)/gama01) r01<-f01/s01

r1<-r11-r01 plot(0, 0, type='n', xlim=range(0, tempo),
ylim=range(0,r1),xlab="Tempo",ylab="Risco Instantâneo", lty=3)
points(tempo,r1, pch=16)
```

ANEXO F Programa utilizado com o conjunto de dados reais para o cálculo da probabilidade do erro Tipo I.

```
> #########pacotes necessários######################
>
> require (survival)
> source("addreg.R")
> #  ################### Valores paramétricos ############################ #
>
>   taxa=0  ; taxacor=0;
> otite=read.table("S=1.txt",h=T)
> attach(otite)

>      par1=otite[1:158,1:4]
>      par2=otite[159:174,1:4]
>  ajlognorm=function(parcela,tempokm)
+   {
+      tempo=parcela[,1]
+      status=parcela[,2]
+
+     ###Ajuste Lognormal###
+       ajust<-survreg(Surv(tempo,status)~1,dist='lognorm')
+       alpha<-ajust$coefficients[1]
```

```
+       gama<-ajust$scale
+       cbind(alpha,gama)
+       s<-pnorm((-log(tempokm)+alpha)/gama)
+       rac<--log(s)
+
+     return (riscoaclogn=rac)
+    }
> ajkm=function(parcela)
+  {
+
+   ### Ajuste não paramétrico (Kaplan Meier) ###
+
+     tempo=parcela[,1]
+     status=parcela[,2]
+
+     ekmp1<-survfit(coxph(Surv(tempo,status)~1,method="efron"))
+     tp1<-ekmp1$time    ; auxtp1=c(0,rep(length(tp1)))
+     sobp1<-ekmp1$surv  ;  auxsobp1=c(0,rep(length(sobp1)))
+   for (j in 1:length(sobp1))
+    {
+       if (sobp1[j]!=0)
+          {
+            auxsobp1[j]=sobp1[j]
+            auxtp1[j]=tp1[j]
+          }
+     }
+     racp1<--log(auxsobp1)
+     return (list(racup1=racp1,tempo1=auxtp1))
+   }
> ####### Dados a alterar no programa ########## #
>
>   nb2=5 ; nb1=10
```

```
>   par=par1                ## Para outra parcela mude o indice
>   chama_km=ajkm(par)
>   ti=chama_km$tempo
>   chama_lognorm=ajlognorm(par,ti) ## Para outra parcela mude o indice
> # ############ Estatística do teste de suavização ############## #
>
>    difpar=(max(abs(chama_lognorm-chama_km$racup1)))/sd(chama_km$racup1)
> # ###################### Rotina Bootstrap ##################### #
>
> contest=0  ; contest2=0  ; contest3=0
> estb1=matrix(0,nb1,1)
> estb2=matrix(0,nb2,1)
> # ############### fim de alteração #########
>
> for (s in 1:nb1)
+ {
+   contest2=0 ; mpar=matrix(0,1,4)
+  for (k in 1:nrow(par))
+    {
+        u=round(runif(1,1,nrow(par)))
+        auxpar=as.matrix(par[u,])
+        mpar=rbind(mpar,auxpar)
+    }
+    mparaux=mpar[2:nrow(mpar),1:4]
+    chama_kmb1=ajkm(mparaux)
+    tib=chama_kmb1$tempo
+    chama_lognormb=ajlognorm(mparaux,tib)
+    difparb=(max(abs(chama_lognormb - chama_kmb1$racup1)))/sd(chama_kmb1$racup1)
+
+  for (i in 1:nb2)
+ {
+   mpar2=matrix(0,1,4)
```

```
+  for (k1 in 1:nrow(mparaux))
+   {
+      u2=round(runif(1,1,nrow(mparaux)))
+      auxpar2=mparaux[u2,]
+      mpar2=rbind(mpar2,auxpar2)
+   }
+    mparaux2=mpar2[2:nrow(mpar2),1:4]
+    chama_kmb2=ajkm(mparaux2)
+    tib2=chama_kmb2$tempo
+    chama_lognormb2=ajlognorm(mparaux2,tib2)
+    difparb2=(max(abs(chama_lognormb2 - chama_kmb2$racup1)))/sd(chama_kmb2$racup1)
+    estb2[i,1]=difparb2
+  }
+    estb1[s,1]=difparb
+  }
> conta=0  ; prob2=matrix(0,nb1,1)  ; conta3=0
>   ref=difpar
>  for (i in 1:nrow(estb1))
+  {
+   if (estb1[i,1]<=ref) conta=conta+1
+  }
>  ### bootstrap ###
>
>    prob1=conta/nb1
> for (j in 1:nrow(estb1))
+ {
+    conta2=0
+    refb=estb1[j,1]
+ for (h in 1:nrow(estb2))
+  {
+      if (estb2[h,1]<=refb) conta2=conta2+1
+  }
```

```
+     prob2[j,1]=conta2/nb2  ### bootstrap 2 nivel
+  }
> ## prob corrigido  ###
>
> for (k in 1:nrow(prob2))
+ {
+       if (prob2[k,1]<=prob1) conta3=conta3+1
+ }
> probcorr=conta3/nb1
>   # ############ probabilidades de significância #################
>
>   prob1

>   probcorr
```