DAIANE DE OLIVEIRA GONÇALVES

MODELOS CONJUNTOS DE ANÁLISE DE DADOS LONGITUDINAIS CENSURADOS: ASPECTOS COMPUTACIONAIS E APLICAÇÃO NA RELAÇÃO DO SISTEMA DE IRRIGAÇÃO COM O SURGIMENTO DE MANCHAS FOLIARES DO CAFÉ

LAVRAS – MG

2024

DAIANE DE OLIVEIRA GONÇALVES

MODELOS CONJUNTOS DE ANÁLISE DE DADOS LONGITUDINAIS **CENSURADOS**:ASPECTOS COMPUTACIONAIS E APLICAÇÃO NA RELAÇÃO DO SISTEMA DE IRRIGAÇÃO COM O SURGIMENTO DE MANCHAS FOLIARES DO CAFÉ

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

Prof. Dr. Marcelo Ângelo Cirillo
Orientador

Profa. Dra. Natália da Silva Martins Fonseca
Coorientadora

LAVRAS – MG
2024

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Gonçalves, Daiane de Oliveira.

Modelos conjuntos de análise de dados longitudinais censurados : aspectos computacionais e aplicação na relação do sistema de irrigação com o surgimento de manchas foliares do caféGonçalves, Daiane de Oliveira. – 2024.

84 p. : il.

Orientador(a): Prof. Dr. Marcelo Ângelo Cirillo.

Coorientador(a): Profa. Dra. Natália da Silva Martins Fonseca.

Tese (doutorado) – Universidade Federal de Lavras, 2024.

Bibliografia.

1. Análise de Sobrevivência. 2. Censura. 3. Simulação. I. Cirillo, Marcelo Ângelo. II. Fonseca, Natália da Silva Martins. III. Título.

DAIANE DE OLIVEIRA GONÇALVES

MODELOS CONJUNTOS DE ANÁLISE DE DADOS LONGITUDINAIS CENSURADOS: ASPECTOS COMPUTACIONAIS E APLICAÇÃO NA RELAÇÃO DO SISTEMA DE IRRIGAÇÃO COM O SURGIMENTO DE MANCHAS FOLIARES DO CAFÉ

JOINT MODELS FOR THE ANALYSIS OF CENSORED LONGITUDINAL DATA: COMPUTATIONAL ASPECTS AND APPLICATION IN THE RELATIONSHIP BETWEEN THE IRRIGATION SYSTEM AND THE EMERGENCE OF COFFEE LEAF SPOTTING

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

APROVADA em 27 de junho de 2024.

Profa. Dr. Paulo Henrique Sales Guimaraes	UFLA
Prof. Dra. Cynthia Arantes Vieira Tojeiro	Ufg
Prof. Dr. Eduardo Yoshio Nakano	Usp
Profa. Dra. Gislene Araujo Pereira	UNIFAL

Prof. Dr. Marcelo Ângelo Cirillo
Orientador

Profa. Dra. Natália da Silva Martins Fonseca
Co-Orientadora

LAVRAS – MG
2024

AGRADECIMENTOS

Agradeço, em primeiro lugar, a Deus, cuja presença e orientação foram fundamentais em cada passo desta jornada. Aos meus amados pais, Maria Aparecida e Sebastião Waldevino, expresso minha gratidão pela constante dedicação, incentivo e apoio incondicional ao longo de toda esta trajetória acadêmica. À minha querida irmã Bruna Oliveira, minha eterna fonte de inspiração e suporte, agradeço por ser não apenas minha família, mas também um alicerce fundamental em minha vida. À minha sobrinha Helena, que mesmo antes de nascer me faz ter ainda mais determinação e amor à vida. À minha avó Marina, pelas inúmeras orações, carinho e apoio em todos os momentos. Aos meus avós falecidos Manuela, Oliveiros e Benedito, por terem sido minha força e inspiração de vida.

Ao meu namorado, Tobias, manifesto minha profunda gratidão por sua paciência, compreensão, amor e inabalável apoio, que foram imprescindíveis para enfrentar os desafios deste percurso. Aos amigos Bruna Silva e Luiz Otávio, reconheço a importância do companheirismo, da ajuda mútua e da amizade sincera, que tornaram esta jornada mais leve e significativa.

Expresso também minha gratidão ao meu orientador, Marcelo Ângelo Cirillo, e à minha coorientadora, Natália da Silva Martins Fonseca, por sua orientação especializada, incentivo constante e apoio acadêmico, que foram essenciais para o desenvolvimento deste trabalho.

Agradeço também aos respeitados docentes do programa de pós-graduação em Estatística e Experimentação Agropecuária da Universidade Federal de Lavras, cuja dedicação e vasto conhecimento foram essenciais para a minha formação acadêmica. Suas orientações e ensinamentos contribuíram significativamente para o desenvolvimento deste trabalho.

Por fim, agradeço a todos aqueles que, de alguma forma, contribuíram para a realização deste estudo e para o meu crescimento pessoal e acadêmico.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

Estudos relacionados a características de fenômenos/experimentos no tempo, como estudos longitudinais ou do tempo até a ocorrência de um evento de interesse, se fazem cada vez mais presentes em diversas áreas. Estes dois fenômenos podem ser tratados, respectivamente, por meio dos modelos lineares mistos e modelos de sobrevivência. No entanto, podem existir situações em que se objetiva investigar a relação entre uma ou mais respostas longitudinais e um evento de interesse, que pode ser realizada com o auxílio da modelagem conjunta de dados longitudinais e de sobrevivência. Porém, esses modelos podem apresentar problemas de convergência e serem computacionalmente exigentes, tornando inviável a utilização dos mesmos em muitos casos. Neste sentido, esta tese, foi separada em dois estágios, o primeiro compreende o objetivo de propor uso da medida da probabilidade de cobertura cruzada, como instrumento de diagnóstico da conexão dos modelos longitudinal e sobrevivência, para auxiliar na estimação de um modelo conjunto que envolva ambos os processos. O primeiro objetivo foi alcançado por meio de um estudo de simulação Monte Carlo, realizando uma comparação entre modelos longitudinais e de sobrevivência, em função de diferentes porcentagens de censura e estrutura de covariância das medidas repetidas. No segundo estágio aplicou-se a metodologia apresentada do modelo conjunto de dados longitudinais e de sobrevivência para compreender a influência do tipo de irrigação no tempo até a o aparecimento da mancha de phoma no café. Os resultados dessa tese mostraram que o procedimento utilizado para estimar probabilidades de cobertura cruzada, como instrumento de diagnostico da conexão dos modelos longitudinal e sobrevivência, se mostrou adequado, ao considerar o modelo Weibull. O aumento do percentual de censura ocasionou um impacto negativo em relação à convergência numérica para a obtenção das estimativas de máxima verossimilhança dos modelos conjuntos para dados longitudinais e de sobrevivência. Os resultados da aplicação proporcionaram uma análise ampla sobre a relação entre as variáveis e como o tempo está relacionado aos riscos de incidência da mancha de phoma em cafeeiro.

Palavras-chave: análise de sobrevivência; censura; modelos lineares mistos; simulação.

ABSTRACT

Studies related to the characteristics of phenomena or experiments over time, such as longitudinal studies or investigations into the time until the occurrence of an event of interest, are increasingly prevalent in various fields. These two phenomena can be addressed, respectively, through linear mixed models and survival models. However, there may be situations where the objective is to investigate the relationship between one or more longitudinal responses and an event of interest, which can be achieved with the aid of joint modeling of longitudinal and survival data. Nevertheless, these models may encounter convergence issues and be computationally demanding, making their use impractical in many cases. In this regard, this thesis was divided into two stages. The first stage aims to propose the use of the cross-coverage probability measure as a diagnostic tool for assessing the connection between longitudinal and survival models to aid in the estimation of a joint model involving both processes. This objective was achieved through a Monte Carlo simulation study, comparing longitudinal and survival models based on different censoring percentages and the covariance structure of repeated measures. In the second stage, the methodology presented for the joint model of longitudinal and survival data was applied to understand the influence of irrigation type on the time until the appearance of phoma leaf spot on coffee plants. The results of this thesis demonstrated that the procedure used to estimate cross-coverage probabilities, as a diagnostic tool for assessing the connection between longitudinal and survival models, proved to be adequate when considering the Weibull model. An increase in the censoring percentage resulted in a negative impact on numerical convergence for obtaining maximum likelihood estimates of joint models for longitudinal and survival data. The application results provided a comprehensive analysis of the relationship between the variables and how time is associated with the risks of phoma leaf spot incidence in coffee plants.

Keywords: survival analysis; censoring; linear mixed models; simulation.

INDICADORES DE IMPACTO

O estudo demonstra um potencial para impactar a sociedade e o setor cafeeiro, tanto no âmbito tecnológico quanto econômico. Neste, destaca-se a necessidade do aperfeiçoamento de modelos conjuntos, pois sua complexidade os torna inviáveis, apresentando problemas de convergência. Ao propor uso da medida da probabilidade de cobertura cruzada, o qual pode ser utilizado como instrumento de diagnóstico da conexão dos modelos longitudinal e sobrevivência, o estudo auxilia pesquisadores a proceder com a estimação de um modelo conjunto que envolva ambos os processos, buscando a minimização de problemas de convergência que, supostamente, poderão ocorrer em sua utilização. A aplicação da metodologia de modelagem conjunta de dados longitudinais e de sobrevivência permitiu uma análise mais robusta sobre a influência dos sistemas de irrigação no tempo até o surgimento da mancha de phoma no café, uma doença que pode comprometer seriamente a produção cafeeira. Socialmente, pode favorecer as práticas agrícolas, contribuindo para o desenvolvimento de técnicas de manejo mais eficazes, o que está diretamente relacionado ao desenvolvimento econômico. Portanto, o estudo tem o potencial de contribuir significativamente para o avanço e melhoria do setor cafeeiro, abrindo também à possibilidade da utilização dos modelos conjuntos em aplicações de outras doenças e cultivos, favorecendo o desenvolvimento e inovação no setor. Os impactos do presente estudo podem ser classificados dentro da temática Tecnologia e Produção da Política Nacional de extensão.

IMPACT INDICATORS

The study demonstrates the potential to impact society and the coffee sector, particularly in technological and economic aspects. It highlights the need for improving joint models, as their complexity often leads to infeasibility and convergence issues. By proposing using the cross-coverage probability measure, which can serve as a diagnostic tool for the connection between longitudinal and survival models, the study aids researchers in estimating a joint model that integrates both processes, aiming to minimize convergence problems that may arise during its application. Implementing this joint modeling methodology for longitudinal and survival data allowed for a more robust analysis of the influence of irrigation systems on the time until the appearance of Phoma leaf spot in coffee plants, a disease that can seriously jeopardize coffee production. Socially, the study may enhance agricultural practices by contributing to the development of more effective management techniques, which are directly linked to economic development. Therefore, the study has the potential to significantly contribute to the advancement and improvement of the coffee sector, while also opening the possibility of using joint

models for other diseases and crops, fostering development and innovation in the agricultural sector. The impacts of the present study can be classified under the Technology and Production theme of the National Extension Policy.

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

Muitos dados, atualmente, são coletadas ao longo do tempo. Estes são denominadas por dados longitudinais, tendo em vista que são obtidas nos mesmos elementos amostrais no decorrer de um período de tempo.

Os dados longitudinais equivalem a observações repetidas de uma variável aleatória de interesse, realizada em diferentes instantes de tempo para um mesmo indivíduo ou objeto (Hu; Szymczak, 2023). Na estatística, há diversas metodologias que podem ser aplicadas para a análise desses dados. Dentre as técnicas utilizadas para essas análises, destacam-se os modelos lineares de efeitos mistos e os modelos de sobrevivência, que são utilizados quando existe a presença de censura (observação incompleta da variável resposta).

Os modelos lineares de efeitos mistos são definidos como modelos que possuem efeitos fixos e efeitos aleatórios. Modelos de efeitos mistos são usados, principalmente, para descrever o relacionamento entre uma variável resposta e covariáveis em dados que são agrupados de acordo com um ou mais fatores de classificação (Pinheiro; Bates, 2006).

Os modelos lineares de efeitos mistos possibilitam predizer como as trajetórias de respostas individuais mudam ao longo do tempo, além da estimação dos parâmetros que descrevem como a resposta média muda na população de interesse. Estes modelos são capazes acomodar qualquer grau de desequilíbrio nos dados, ou seja, não é necessário o mesmo número de medições em cada indivíduo ou objeto. Tem-se ainda o fato de que os efeitos aleatórios são responsáveis pela correlação entre as medidas repetidas de forma relativamente parcimoniosa (Verbeke; Molenberghs, 1997).

Segundo Pinheiro e Bates (2006), quando os dados estão estruturados de maneira hierárquica em diversos níveis de efeitos aleatórios, estes são chamados de modelos de efeitos mistos multiníveis.

De acordo com Do Ha e Lee (2005), modelos lineares mistos têm sido propostos para a análise de dados de sobrevivência, nos quais os efeitos aleatórios atuam de forma linear no tempo de sobrevivência do indivíduo ou objeto.

Os modelos de sobrevivência são definidos para situações em que se objetiva avaliar o tempo decorrido até a ocorrência de um ou mais eventos de interesse. Este evento também pode ser denominado como falha. Porém, nem sempre o instante de ocorrência do evento de interesse tem seu tempo exato conhecido, ou até mesmo o evento de interesse não é observado, o que leva a definição de censura no modelo de sobrevivência. As observações censuradas

podem ser entendidas como aquelas observações parciais ou incompletas da variável resposta (Colosimo; Giolo, 2006).

Desse modo, os modelos de sobrevivência destacam-se justamente por considerarem essas observações incompletas (censuradas) na análise, permitindo que a análise estatística resulte em conclusões fidedignas, pois carregam informações sobre o tempo até a ocorrência do evento de interesse do elemento amostral em estudo.

Existe ainda a possibilidade de investigar a relação entre uma ou mais respostas longitudinais e um evento de interesse, com o auxílio dos modelos conjuntos. Estes modelos surgiram como uma ferramenta para a análise de dados com estas características, consequentemente sem a necessidade de serem separados. Entretanto, em muitos casos, a complexidade destes modelos podem torná-los inviáveis, pois a inclusão dos efeitos aleatórios torna-se computacionalmente exigente à medida que sua dimensionalidade aumenta (Murray; Philipson, 2022). Ademais, mesmo que existam na literatura abordagens para contornar os problemas de convergência, estas ainda são relativamente exigentes sendo a principal razão dos modelos conjuntos não serem uma opção direta para esse tipo de análise.

Dados os problemas de convergência na utilização dos modelos conjuntos, é de interesse avaliar uma possível conexão entre os modelos lineares mistos e os modelos de sobrevivência, para verificar a possibilidade da sua utilização na análise de dados com respostas longitudinais e evento de interesse. E, com isso constatar qual o efeito e influência de diferentes porcentagens de censura sobre a análise.

Com esta perspectiva, justifica-se a contribuição deste estudo, no uso de uma metodologia que propõe a obtenção das probabilidades de cobertura cruzadas, no sentido de que, as estimativas dos parâmetros, referente ao modelo longitudinal foram computadas com base no intervalo de confiança dos parâmetros do modelo de sobrevivência. Da mesma forma, gerou-se as probabilidades de cobertura para o modelo de sobrevivência, de tal forma que as estimativas dos parâmetros, referente à este modelo, foram obtidas com base no intervalo de confiança dos parâmetros do modelo longitudinal.

Em virtude do que foi mencionado, o objetivo deste trabalho é propor uso da medida da probabilidade de cobertura cruzada, como instrumento de diagnóstico da conexão dos modelos longitudinal e sobrevivência, de modo que, possa auxiliar o pesquisador a proceder com a estimação de um modelo conjunto que envolva ambos os processos e minimizar problemas numéricos de convergência que poderão supostamente ocorrer.

E, por fim, aplicar a metodologia apresentada do modelo conjunto de dados longitudinais e de sobrevivência em um estudo para compreender a influência do tipo de irrigação no tempo até a o aparecimento da mancha de phoma no café. Dentre as principais doenças observadas nos cafeeiros, a mancha de phoma se destaca, pois a sua ocorrência pode causar a queda de folhas e a seca e morte de ramos produtivos, resultando em perdas significativas na produção a ponto de inviabilizar a sua continuidade, representando um risco para o setor e sérios prejuízos financeiros. Portanto, objetiva-se avaliar qual sistema de irrigação faz com que o tempo até o aparecimento da mancha de phoma no café seja maior possível, visando a melhoria da qualidade e impulsionar o crescimento econômico do setor cafeeiro.

2 REFERENCIAL TEÓRICO

Nesta seção é apresentada a fundamentação teórica deste trabalho, abrangendo modelos para dados longitudinais, modelos de sobrevivência e os modelos conjuntos.

2.1 Estudos longitudinais

Estudos longitudinais podem ser caracterizados por aqueles que fornecem informações sobre indivíduos/objetos ao longo do tempo, contendo medidas repetidas para cada unidade de análise (Diggle *et al*., 2002).

Os estudos longitudinais apresentam diversas vantagens em relação aos estudos transversais. Entre estas, destacam-se a capacidade de avaliar o comportamento da resposta ao longo do tempo e a variação intra-indivíduo. Uma característica crucial dos estudos longitudinais é a possibilidade de utilizar cada indivíduo como seu próprio controle, proporcionando uma vantagem significativa na compreensão das mudanças ao longo do tempo (Hedeker; Gibbons, 2006).

De acordo com Singer e Andrade (1986), a principal desvantagem dos estudos longitudinais refere-se ao custo. Pois para a obtenção dos dados longitudinais existe um esforço para garantir a observação das unidades amostrais em diferentes instantes de tempo, e este período de observação pode ser longo. Uma outra desvantagem pode ser apontada em relação a análise dos dados, a qual exige mais cuidado que os estudos de dados transversais.

Para a análise de dados longitudinais, diversas metodologias podem ser adotadas, tais como a utilização de modelos de sobrevivência e modelos lineares mistos, os quais serão apresentados nas Seções 2.2 e 2.3. Existe, também, a possibilidade de considerar a dependência e associação entre dados longitudinais e dados de tempo até o evento de interesse, através dos modelos conjuntos, o qual será apresentado na Seção 2.4.

2.2 Análise de sobrevivência

De acordo com Liu (2012), a prática da análise de sobrevivência se refere ao ato de descrever, medir e analisar características de eventos para fazer previsões sobre processos de tempo até determinado evento de interesse, ou seja, o período de tempo até a ocorrência de um evento de interesse. Estudos em análise de sobrevivência envolvem fazer comparações entre grupos ou categorias de uma população, ou examinar as variáveis que poderão influenciar na sobrevivência de um indivíduo/objeto.

Segundo Moore (2016), análise de sobrevivência é o estudo dos tempos de sobrevivência e dos fatores que os influenciam. Os estudos envolvendo análise de sobrevivência podem ser aplicados em diversas áreas, sendo comumente utilizados nas áreas médicas. No âmbito acadêmico, a análise de sobrevivência é amplamente aplicada, devido à disponibilidade de dados longitudinais que registram histórias de vários processos de sobrevivência e as ocorrências de eventos diversos.

O conceito de sobrevivência aos poucos foi deixando de se referir somente a um evento médico e começou a se expandir para um escopo muito mais amplo de fenômenos caracterizados por processos de tempo até o evento de interesse (Liu, 2012).

Segundo Machin *et al*. (2006), a variável resposta dos dados de sobrevivência é não negativa e representa o tempo desde uma origem bem definida até um evento bem definido, quando este evento de interesse ocorre, tem-se o tempo de falha. Quando os eventos iniciais ou finais não são observados com precisão, diz-se que o tempo observado é censurado. Estas observações censuradas, mesmo que sejam incompletas, irão carregar informações sobre o tempo de sobrevida em estudo e que se excluídos podem levar a conclusões equivocadas.

Existem algumas categorizações das censuras, as quais podem ser ditas censura à direita, à esquerda e intervalar. A censura a direita pode ser entendida como a observação em que o tempo de ocorrência do evento de interesse está à direita do tempo registrado. A censura à esquerda pode ser observada quando o tempo registrado é maior que o tempo de falha. Na censura intervalar, o evento de interesse só é conhecido por ocorrer entre dois determinados pontos no tempo (Hosmer; Lemeshow, 1999).

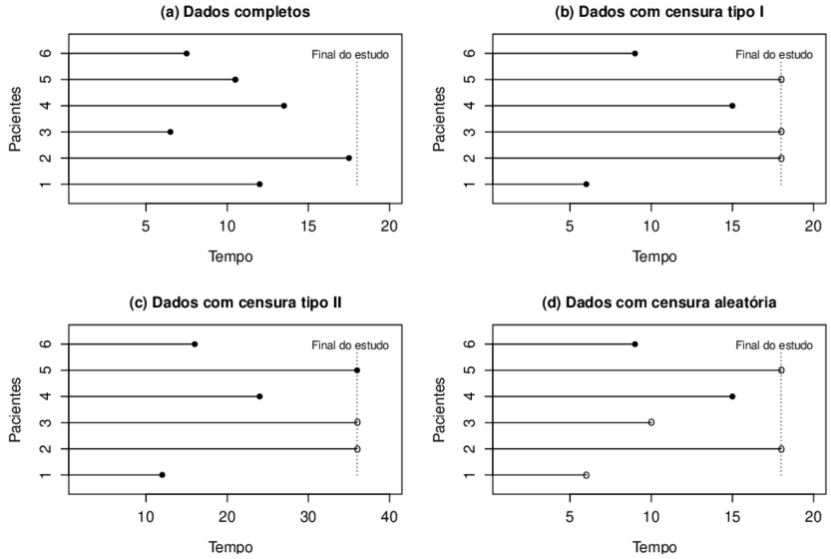
Para exemplificar os tipos de censura, suponhamos que haja um monitoramento de várias plantas de café para determinar quanto tempo leva até que apareça a mancha de phoma. Caso algumas plantas ainda não desenvolveram a mancha de phoma até o final do período de estudo, esses tempos de observação são censurados à direita.

No caso em que, no momento inicial do estudo, algumas plantas já apresentavam a doença, porém não sabe-se exatamente quando começou a se desenvolver, então o tempo desde o início do desenvolvimento da mancha de phoma até o início do estudo é censurado à esquerda.

Considerando que a mancha de phoma apareceu em algumas plantas entre duas datas de observação, mas não se sabe exatamente o momento específico a doença começou a se manifestar, então o tempo exato até o aparecimento da mancha de phoma para essas plantas será censurado intervalarmente.

O tipo de censura mais conhecido é censura à direita, a qual pode ser entendida através dos mecanismos de censura apresentados na Figura 2.1, em que o tempo de ocorrência do evento de interesse está à direita do tempo registrado, ou seja, o tempo de falha é maior que o tempo observado.

Figura 2.1 – Ilustração de alguns mecanismos de censura em que • representa a falha e ○ representa a censura.



Fonte: Colosimo e Giolo (2006)

Na Figura 2.1 (a), são apresentados os dados completos, ou seja, todos os indivíduos ou objetos passaram pelo evento de interesse antes do término do estudo. Já na Figura 2.1 (b), observa-se que alguns indivíduos ou objetos não vivenciaram o evento de interesse até o final do estudo, caracterizando dados com censura do tipo I. A Figura 2.1 (c) exibe dados com censura do tipo II, onde o estudo foi concluído após atingir um número previamente definido de falhas. Por fim, na Figura 2.1 (d), tem-se a ilustração dos dados com censura aleatória, em que o acompanhamento de certos indivíduos ou objetos foi interrompido por algum motivo, sendo que alguns deles não passaram pelo evento até o término do estudo.

Segundo Rizopoulos (2012), a função básica em análise de sobrevivência é a utilizada para descrever a distribuição de T^*, variável aleatória dos tempos de falha em estudo, sendo a função de sobrevivência, a qual é dada por:

$$S(t) = P(T^* > t) = \int_t^{\infty} f(s)ds, \tag{2.1}$$

assumindo a variável aleatória T^* contínua e positiva, e $f(\cdot)$ é a função densidade de probabilidade correspondente.

Outra função importante determinada em análise de sobrevivência é a função de risco, a qual irá retornar o risco instantâneo para um evento no intervalo de tempo $[t, t+\Delta t)$ desde a sobrevivência até t, e pode ser expressa por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t+\Delta t | T^* \geq t)}{\Delta t}, \tag{2.2}$$

considerando $t > 0$ ou, também,

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

em que $S(t)$ é a função de sobrevivência. Pode-se obter a seguinte relação:

$$h(t) = -\frac{d}{dt}\left[log\, S(t)\right]. \tag{2.4}$$

O desenvolvimento das expressões 2.3 e 2.4 é apresentado no Apêndice A.

A função de risco acumulada é definida por:

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

Pode-se observar uma relação importante entre a função de risco instantâneo e a função de sobrevivência determinada a seguir:

$$H(t) = \int_0^t h(u)du = -\log S(t), \tag{2.6}$$

ou seja,

$$S(t) = \exp\left\{-\mathcal{H}(t)\right\}. \tag{2.7}$$

2.2.1 Tipos de modelos para dados de sobrevivência

Segundo Rossello e González-del-Hoyo (2022), três abordagens básicas da análise de sobrevivência são: métodos não paramétricos, métodos paramétricos e métodos semiparamétricos.

Em relação ao método não paramétrico, para encontrar uma estimativa para a função de sobrevivência pode-se utilizar o estimador não paramétrico de Kaplan e Meier (1958). Este é um estimador muito conhecido e utilizado na análise de sobrevivência.

Métodos não paramétricos, como os de Kaplan-Meier, são relativamente simples e não fazem suposições sobre a distribuição dos tempos de sobrevivência. Eles são excelentes para análises univariadas, mas não são suficientes para lidar com problemas mais complexos (Rossello; González-del-Hoyo, 2022).

Kleinbaum e Klein (2010) denotam o estimador de Kaplan-Meier para a função de sobrevivência da seguinte forma:

$$\widehat{S}_{KM}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right). \tag{2.8}$$

onde $t_j,\ j = 1, \cdots, r$ são os r tempos distintos e ordenados de falha, d_j é o número de falhas em $t_j,\ j = 1, \cdots, r$, e n_j é o número de indivíduos sob risco em t_j.

De acordo com Atlan *et al*. (2021) quando um modelo é semi-paramétrico, apesar dos parâmetros de regressão serem conhecidos, a distribuição do tempo de sobrevivência ainda é desconhecida. Desse modo, não é totalmente paramétrico ou totalmente não paramétrico. Alguns métodos semiparamétricos são o modelo de Cox (Cox, 1972) e CoxBoost (Binder; Schumacher, 2008).

2.2.1.1 Modelo de riscos proporcionais de Cox

O modelo de riscos proporcionais de Cox tornou-se, de maneira significativa, o procedimento predominante para modelar a dados de sobrevivência (Therneau; Grambsch, 2000).

Segundo Rizopoulos (2012), em razão da popularidade do modelo de Cox (Cox, 1972), os modelos de riscos proporcionais prevaleceram até os estudos mais modernos. Estes modelos podem ser descritos da seguinte forma:

$$\begin{aligned} h_i(t|\boldsymbol{w}_i) &= \lim_{\Delta t \to 0} \frac{P(t \leq T^* < t + \Delta t | T^* \geq t, \boldsymbol{w}_i)}{\Delta t} \\ &= h_0(t) \exp(\boldsymbol{\gamma}^\top \boldsymbol{w}_i), \end{aligned} \tag{2.9}$$

em que $\boldsymbol{w}_i^\top = (w_{i1}, \cdots, w_{ip})$ é o vetor de covariáveis, as quais supõem-se estarem associadas ao risco de cada indivíduos e $\boldsymbol{\gamma}$ corresponde ao vetor de coeficientes de regressão. Na Expressão (2.9), $h_0(t)$ é a função denominada de risco basal, relacionada à função de risco associada a

$\boldsymbol{\gamma}^\top \boldsymbol{w}_i = 0$. Os modelos de riscos proporcionais de Cox assumem que as covariáveis têm um efeito multiplicativo sobre o risco de um evento.

Ainda de acordo com Rizopoulos (2012), a razão de riscos para um indivíduo i com vetor de covariáveis $\boldsymbol{w}_i$ em comparação ao indivíduo k com vetor de covariáveis $\boldsymbol{w}_k$ é:

$$\frac{h_i(t|\boldsymbol{w}_i)}{h_k(t|\boldsymbol{w}_k)} = \exp\left\{\boldsymbol{\gamma}^\top(\boldsymbol{w}_i - \boldsymbol{w}_k)\right\}, \tag{2.10}$$

que não depende de t.

A estimação dos parâmetros no Modelo de Riscos Proporcionais de Cox é, geralmente, realizada por meio do método de máxima verossimilhança parcial (Partial Likelihood Estimation) (Cox, 1972).

A verossimilhança parcial pode ser expressa da seguinte forma:

$$pl(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\gamma}^\top \boldsymbol{w}_i - \log\left\{\sum_{t_j \geq t_i} \exp(\boldsymbol{\gamma}^\top \boldsymbol{w}_j)\right\}\right], \tag{2.11}$$

em que $\boldsymbol{\gamma}$ é o vetor de coeficientes a serem estimados e δ_i é o indicador de falha ou censura, definido por

$$\delta_i = \begin{cases} 1, \text{ se } t_i \text{ é um tempo de falha} \\ 0, \text{ se } t_i \text{ é um tempo censurado.} \end{cases} \tag{2.12}$$

Segundo Rizopoulos (2012), não é exigido a especificação de $h_o(\cdot)$. Dessa forma, o modelo de risco proporcional com uma função de risco base não especificada é um modelo semiparamétrico que não faz suposições sobre a distribuição dos tempos de evento, mas presume que as covariáveis atuam multiplicativamente na taxa de risco. Os estimadores de máxima verossimilhança parcial são encontrados ao resolver as equações de escore associadas à verossimilhança parcial:

$$\frac{\partial pl(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^\top} = \sum_{i=1}^{n} \delta_i \left\{\boldsymbol{w}_i - \frac{\sum_{t_j \geq t_i} \boldsymbol{w}_j \exp(\boldsymbol{\gamma}^\top \boldsymbol{w}_j)}{\sum_{t_j \geq t_i} \exp(\boldsymbol{\gamma}^\top \boldsymbol{w}_j)}\right\} = 0. \tag{2.13}$$

2.2.1.2 Modelos paramétricos de sobrevivência

Segundo Maiorano (2018) na abordagem paramétrica assume-se que a função de sobrevivência tenha uma forma paramétrica específica, ou seja, que $S(t)$ seja modelada por alguma distribuição de probabilidade conhecida, tal como: Weibull, Exponencial, log-normal e Gama, dentre outras distribuições.

Segundo Rizopoulos (2012) a estimativa dos parâmetros da função de sobrevivência, quando esta assume a forma paramétrica, é, frequentemente, realizada por meio do método de máxima verossimilhança.

De acordo com Cox e Oakes (1984), a função de verossimilhança, considerando as informações das observações não censuradas e censuradas, pode ser representada por:

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [f(t_i;\boldsymbol{\theta})]^{\delta_i} [S(t_i;\boldsymbol{\theta})]^{1-\delta_i}, \tag{2.14}$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros, t_i é o tempo observado do i-ésimo indivíduo com o seu respectivo indicador de falha ou censura, δ_i, definido na Expressão (2.12).

Desse modo, o logaritmo da função de verossimilhança pode ser obtido por:

$$l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \{\delta_i \log[f(t_i;\boldsymbol{\theta})] + (1-\delta_i)[S(t_i;\boldsymbol{\theta})]\} + C, \tag{2.15}$$

em que C é uma constante real que não depende de $\boldsymbol{\theta}$ (o desenvolvimento da Expressão (2.15) é apresentado no Apêndice A).

Como os estimadores de máxima verossimilhança dos parâmetros nem sempre possuem solução analítica, é frequentemente necessário recorrer a métodos numéricos, como o método de Newton-Raphson, para obtê-los. (Verbeke; Cools, 1995).

2.2.1.2.1 Distribuição Weibull

A distribuição Weibull é uma das principais distribuições utilizadas na análise de sobrevivência, principalmente devido à sua notável flexibilidade em modelar diferentes padrões de falha e sua robustez em relação à censura à direita. A distribuição Weibull permite ajustar desde cenários com taxa de falha constante, quando o parâmetro de forma ($\sigma = 1$), até situações com taxa de falha crescente ($\sigma > 1$) ou decrescente ($\sigma < 1$), oferecendo grande versatilidade para representar comportamentos de risco variados ao longo do tempo (Lai; Murthy; Xie, 2006).

Além disso, o modelo Weibull se mostra robusto ao lidar com dados censurados à direita, uma característica comum em estudos de sobrevivência, garantindo estimativas consistentes e eficientes dos parâmetros mesmo quando o evento de interesse não é observado para todos os indivíduos ou sistemas até o final do estudo. Essas propriedades fazem da distribuição Weibull uma escolha adequada na presença de observações censuradas (Lai; Murthy; Xie, 2006).

A função densidade de probabilidade da distribuição Weibull para a variável aleatória T é dada por:

$$f(t;\alpha,\beta) = \frac{\beta}{\alpha^{\beta}} t^{\beta-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}, \tag{2.16}$$

em que $t \geq 0$, $\alpha > 0$ é o parâmetro de escala e $\beta > 0$ é o parâmetro de forma (Cox; Oakes, 1984). Tem-se, também, que a esperança e a variância da variável aleatória T são dadas, respectivamente por:

$$E[T] = \alpha\Gamma\left(1+\frac{1}{\beta}\right) \tag{2.17}$$

e

$$Var[T] = \alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right], \tag{2.18}$$

em que $\Gamma(r)$ é a função gama, definida por $\Gamma(r) = \int_0^{\infty} x^{r-1}e^{-x}dx$. Determina-se a função de sobrevivência $S(t)$ para a distribuição Weibull conforme a Expressão (2.19):

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}. \tag{2.19}$$

A partir da função de sobrevivência, pode-se definir a função de risco para a distribuição Weibull da seguinte forma:

$$h(t) = \frac{\beta}{\alpha^{\beta}} t^{\beta-1}, t \geq 0. \tag{2.20}$$

Com a finalidade de encontrar os estimadores dos parâmetros, considera-se uma amostra aleatória $t_1, \cdots, t_n$ de uma variável aleatória T com distribuição Weibull, em que $\boldsymbol{\theta}$ é o vetor de parâmetros, t_i é o tempo observado com o seu respectivo indicador de falha ou censura, δ_i, $i = 1, \cdots, n$. Desse modo, a função de verossimilhança pode ser escrita da seguinte forma:

$$L(\boldsymbol{\theta};\boldsymbol{t}) \propto \prod_{i=1}^{n}\left[\frac{\beta}{\alpha^{\beta}} {t_i}^{\beta-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}\right]. \tag{2.21}$$

Tomando o logaritmo de $L(\boldsymbol{\theta})$, tem-se:

$$l(\boldsymbol{\theta}) = k\left[\log(\beta)\right] - k\beta\left[\log(\alpha)\right] + \beta\sum_{i=1}^{n}\delta_i\log(t_i) - \sum_{i=1}^{n}\delta_i\log(t_i) - \alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta} + C. \tag{2.22}$$

em que k é o número de falhas, $\sum_{i=1}^{n}\delta_i = k$, e C é uma constante real que não depende de $\boldsymbol{\theta}$.

Derivando-se a Expressão (2.22) em relação a ambos os parâmetros α e β tem-se:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\alpha} = \frac{\beta}{\alpha}\left(-k + \alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta}\right) \tag{2.23}$$

e

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\beta} = \frac{k}{\beta} - k\left[\log(\alpha)\right] + \sum_{i=1}^{n}\delta_i\log(t_i) + \alpha^{-\beta}\left[\log(\alpha)\right]\sum_{i=1}^{n}{t_i}^{\beta} - \alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta}\log(t_i), \tag{2.24}$$

desse modo, igualando ambas as expressões (2.23) e (2.24) a zero, obtém-se:

$$\widehat{\alpha} = \left(\frac{\sum_{i=1}^{n}{t_i}^{\widehat{\beta}}}{k}\right)^{\frac{1}{\widehat{\beta}}} \tag{2.25}$$

e

$$\frac{1}{\widehat{\beta}} = \sum_{i=1}^{n}{t_i}^{\widehat{\beta}}\log(t_i) - \frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{k}. \tag{2.26}$$

O desenvolvimento das expressões 2.21, 2.22, 2.23, 2.25 e 2.26 é apresentado no Apêndice A. Observa-se que os estimadores de máxima verossimilhança não possuem solução analítica, logo a utilização de métodos numéricos, como o de Newton-Rapshon, se faz necessário para encontrar tais soluções. Para a obtenção das soluções o método de Newton-Raphson utiliza a matriz de derivadas segundas da função de log-verossimilhança.

Seja $U(\boldsymbol{\theta})$ a função escore dada pela derivação da log-verossimilhança em $\boldsymbol{\theta}$, como apresentado a seguir:

$$U(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \tag{2.27}$$

Para encontrar a estimativa de máxima verossimilhança basta igualar a função escore a 0 e isolar o $\boldsymbol{\theta}$. Portanto, para o estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$, tem-se:

$$U(\widehat{\boldsymbol{\theta}}) = 0, \tag{2.28}$$

expandindo $U(\widehat{\boldsymbol{\theta}})$ em série de Taylor em torno de um $\boldsymbol{\theta}_0$, tem-se que:

$$0 = U(\widehat{\boldsymbol{\theta}}) \cong U(\boldsymbol{\theta}_0) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)H(\boldsymbol{\theta}_0), \tag{2.29}$$

ou

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 - H(\boldsymbol{\theta}_0)^{-1}U(\boldsymbol{\theta}_0). \tag{2.30}$$

em que $H(\boldsymbol{\theta})$ é a matriz Hessiana, isto é, matriz de derivadas parciais de segunda ordem negativas de $l(\boldsymbol{\theta})$. O processo iterativo de Newton-Raphson pode ser obtido da Expressão (2.30), sendo representado por:

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - H(\boldsymbol{\theta}_j)^{-1}U(\boldsymbol{\theta}_j). \tag{2.31}$$

em que é iniciado com o valor $\boldsymbol{\theta}_0$ e então um novo valor de $\boldsymbol{\theta}_1$ é obtido a partir da Expressão (2.31) e assim consecutivamente, até que o processo se estabilize.

A matriz Hessiana é definida por:

$$H_{ij}(\alpha,\beta) = \begin{bmatrix} \frac{\partial^2}{\partial\alpha^2}l(\alpha,\beta) & \frac{\partial^2}{\partial\alpha\partial\beta}l(\alpha,\beta) \\ \frac{\partial^2}{\partial\beta\partial\alpha}l(\alpha,\beta) & \frac{\partial^2}{\partial\beta^2}l(\alpha,\beta) \end{bmatrix}$$

Considerando a distribuição Weibull, temos que os elementos (i,j) da matriz Hessiana são dados por:

$$\frac{\partial^2}{\partial\alpha^2}l(\alpha,\beta) = \frac{p\beta}{\alpha^2} + \beta(-\beta-1)\alpha^{-\beta-2}\sum_{i=1}^{n} {t_i}^{\beta},$$

$$\frac{\partial^2}{\partial\beta^2}l(\alpha,\beta) = -\frac{p}{\beta^2} - \sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta}\left[\log\left(\frac{t_i}{\alpha}\right)\right]^2,$$

e

$$\frac{\partial^2}{\partial\alpha\partial\beta}l(\alpha,\beta) = \frac{\partial^2}{\partial\beta\partial\alpha}l(\alpha,\beta) = -\frac{p}{\alpha} + \frac{1}{\alpha}\sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta} + \frac{\beta}{\alpha}\sum_{i=1}^{n}\left(\frac{t_i}{\alpha}\right)^{\beta}\log\left(\frac{t_i}{\alpha}\right).$$

Os elementos da matriz U foram apresentados nas expressões (2.23) e (2.24) considerando a distribuição Weibull. A matriz U é dada por:

$$U_{ij}(\alpha,\beta) = \begin{bmatrix} \frac{\partial}{\partial\alpha}l(\alpha,\beta) \\ \frac{\partial}{\partial\beta}l(\alpha,\beta) \end{bmatrix}$$

De acordo com Bolfarine e Sandoval (2010), o estimador de máxima verossimilhança é obtido quando $\mid \boldsymbol{\theta}_{j+1} - \boldsymbol{\theta}_j \mid < \varepsilon$, em que ε é o erro na estimação, ou seja, quando a diferença entre as iterações é menor que um erro ε.

2.3 Modelos lineares mistos

De acordo com Oliveira *et al*. (2021), os modelos lineares clássicos sofreram adaptações ao longo dos anos, passando a tratar, além dos efeitos fixos, também dos efeitos aleatórios, se tornando a classe de modelos lineares mistos, ou modelos lineares de efeitos mistos. Estes modelos são muito utilizados na análise de dados longitudinais.

A análise de dados longitudinais baseia-se no fato de que cada indivíduo da população tem seu próprio perfil de resposta médio específico ao longo do tempo. Formalmente, o modelo linear de efeitos mistos pode ser expresso da seguinte forma (Laird; Ware, 1982):

$$\begin{cases} y_i = \boldsymbol{X}_i\boldsymbol{\lambda} + \boldsymbol{Z}_i b_i + \boldsymbol{\varepsilon}_i, \\ b_i \sim N(0, \boldsymbol{D}), \\ \boldsymbol{\varepsilon}_i \sim N(0, \sigma^2 \boldsymbol{I}_{n_i}), \end{cases} \tag{2.32}$$

em que y_i representa a resposta ou variável dependente para a i-ésima observação, $i = 1, \cdots, n$. $\boldsymbol{X}_i$ e $\boldsymbol{Z}_i$ são matrizes de delineamento conhecidas, para os coeficientes regressão de efeitos fixos $\boldsymbol{\lambda}$, e os coeficientes de regressão de efeitos aleatórios b_i, respectivamente, e $\boldsymbol{I}_{n_i}$ denota a matriz de identidade n_i-dimensional, n_i é o número de observações para a i-ésima unidade (Rizopoulos, 2012). Os efeitos aleatórios são assumidos como normalmente distribuídos com média zero e $\boldsymbol{D}$ a matriz de variância-covariância, e são assumidos independentemente dos termos de erro $\boldsymbol{\varepsilon}_i$, ou seja, $cov(a_i, \boldsymbol{\varepsilon}_i) = 0$.

Segundo Pinheiro e Bates (2006), para realizar a estimação dos parâmetros do modelo linear misto, pode-se utilizar os princípios do método da máxima verossimilhança. A inferência relacionada à esses modelos é baseada na densidade marginal da variável resposta y_i, dada pela expressão:

$$p(y_i) = \int p(y_i|b_i)p(b_i)db_i. \tag{2.33}$$

Esta fundamentação se da pelo fato de que o comportamento probabilístico dos efeitos aleatórios do modelo linear misto não é conhecido (Costa, 2010). Supondo que a distribuição condicional das respostas longitudinais, os efeitos aleatórios $\{y_i|b_i\}$ e a distribuição dos efeitos aleatórios a_i são normais, a Expressão (2.33) tem uma solução fechada, tem-se que $\{y_i|b_i\} \sim N(\boldsymbol{X}_i\boldsymbol{\lambda}, \boldsymbol{V}_i)$, sendo $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top + \sigma^2\boldsymbol{I}_{n_i}$ a matriz de variância-covariância.

A função log-verossimilhança do modelo linear misto, admitindo-se a independência entre os sujeitos, pode ser dada por:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(y_i|b_i;\lambda,\sigma^2)p(b_i;\theta_b)db_i, \tag{2.34}$$

em que $\boldsymbol{\theta}^\top = (\alpha^\top, \sigma^2, \boldsymbol{\theta}_b^\top)$, sendo $\boldsymbol{\theta}_b = vech(\boldsymbol{D})$. Segundo Jinadasa (1988), para uma matriz $\boldsymbol{X}$, $vec(\boldsymbol{X})$ é definido como o vetor coluna obtido empilhando suas colunas, e $vech(\boldsymbol{X})$ é obtido de $vec(\boldsymbol{X})$ eliminando todos os elementos acima da diagonal de $\boldsymbol{X}$ (o desenvolvimento da Expressão (2.34) é apresentado no Apêndice A).

Supondo que $\boldsymbol{V}_i$ é conhecido, tem-se que o estimador de máxima verossimilhança do vetor de efeitos fixos $\boldsymbol{\lambda}$, pode ser obtido do seguinte modo:

$$l(\boldsymbol{\theta}) = \quad -\frac{1}{2}\sum_{i=1}^{n}\left(y_i^\top \boldsymbol{V}_i^{-1} y_i - y_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\boldsymbol{\lambda} - \boldsymbol{\lambda}\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} y_i + \boldsymbol{\lambda}\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\boldsymbol{\lambda}\right). \tag{2.35}$$

Derivando-se a Expressão (2.35) em relação à $\boldsymbol{\lambda}$ tem-se:

$$\frac{\partial}{\partial \boldsymbol{\lambda}} = \sum_{i=1}^{n}\left(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} y_i - \hat{\boldsymbol{\lambda}}\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right). \tag{2.36}$$

Igualando a zero e realizando o desenvolvimento necessário, obtêm-se:

$$\hat{\lambda} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right)^{-1}\sum_{i=1}^{n}\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} y_i. \tag{2.37}$$

A apresentação do desenvolvimento das expressões 2.35, 2.36 e 2.37 é apresentado no Apêndice A.

Ao maximizar a Expressão (2.34) condicional aos parâmetros em $\boldsymbol{V}_i$, $\hat{\lambda}$ tem uma forma fechada e corresponde ao estimador de mínimos quadrados generalizado.

A estimativa de $\boldsymbol{\lambda}$ também pode ser obtida quando $\boldsymbol{V}_i$ é desconhecido, neste caso utiliza-se a estimativa $\hat{\boldsymbol{V}}_i$ na Expressão (2.37). A obtenção da estimativa para $\boldsymbol{V}$ também pode ser a partir do método da máxima verossimilhança, considerando um dado valor de $\boldsymbol{\lambda}$. Porém, o estimador de máxima verossimilhança para os parâmetros em $\boldsymbol{V}_i$ nem sempre podem ser escritos de forma fechada. Deste modo, também, se faz necessário a utilização de métodos numéricos.

2.4 Modelagem conjunta de dados longitudinais e de sobrevivência

Em muitos casos é possível identificar a presença de dados longitudinais juntamente com o tempo até o evento de interesse. Segundo Wulfsohn e Tsiatis (1997), modelos clássicos, como o modelo linear misto para dados longitudinais e o modelo de riscos proporcionais de Cox para dados de tempo até o evento de interesse, não consideram dependências entre esses dois tipos de dados diferentes.

Um método que leva em consideração a dependência e associação entre dados longitudinais e dados de tempo até o evento de interesse são os modelos conjuntos para dados longitudinais e de sobrevivência (Rizopoulos, 2012). Estes são modelos que reúnem esses dois tipos de dados simultaneamente em um único modelo para que se possa inferir a dependência e associação entre o biomarcador longitudinal e o tempo até o evento de interesse para melhor avaliar o efeito de um tratamento.

Para apresentar a estrutura destes modelos, denota-se por T_i^* o tempo verdadeiro do evento para o i-ésimo sujeito, T_i o tempo do evento observado, definido como o mínimo do tempo potencial de censura C_i e T_i^* e $\delta_i = I(T_i^* \leq C_i)$ o indicador de falha ou censura. Para a covariável dependente do tempo, considera-se $y_i(t)$ o seu valor observado no tempo t para o i-ésimo objeto. Nota-se que não se observa $y_i(t)$ para qualquer tempo t, mas apenas em ocasiões muito específicas t_{ij} em que medidas foram tomadas. Assim, os dados longitudinais observados consistem na medições $y_{ij} = \{y_i(t_{ij}), j = 1, \cdots, n_i\}$ (Rizopoulos, 2012).

Como o objetivo é medir a associação entre o nível do marcador longitudinal e o risco de um evento, introduz-se o termo $m_i(t)$ que denota o verdadeiro valor do resultado longitudinal no tempo t. Para quantificar a associação entre $m_i(t)$ e o risco de um evento, o modelo de risco relativo pode ser expresso da seguinte forma:

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) &= \lim_{dt \to 0} \frac{Pr\{t \leq T_i^* < t + dt | T_i^* \geq t, \mathcal{M}_i(t), \boldsymbol{w}_i\}}{dt} \\ &= h_0(t) \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \varphi m_i(t)\}, \end{aligned} \tag{2.38}$$

em que $t > 0$, $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denota a trajetória das variáveis longitudinais até o ponto de tempo t, h_0 denota a função de risco basal, e $\boldsymbol{w}_i$ vetor de covariáveis fixas do i-ésimo indivíduo, relacionadas ao processo de sobrevivência, $\boldsymbol{\gamma}$ é o vetor de coeficientes de regressão associados à $\boldsymbol{w}_i$ e φ é um parâmetro desconhecido que quantifica o impacto do processo longitudinal na observação do evento.

Em particular, $\exp(\gamma_j)$ denota a razão de riscos para uma mudança de unidade em w_{ij} em um tempo qualquer t, enquanto $\exp(\varphi)$ denota o aumento do risco relativo de um evento no tempo t que resulta de uma unidade aumento em $m_i(t)$ no mesmo ponto de tempo.

Segundo Rizopoulos (2012), utilizando a relação conhecida entre a função de sobrevivência e a função de risco, obtém-se que:

$$\begin{aligned} S_i(t|\mathcal{M}_i(t),\boldsymbol{w}_i) &= Pr(T_i^* > t|\mathcal{M}_i(t),\boldsymbol{w}_i) \\ &= \exp\left(-\int_0^t h_0(s)\exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \varphi m_i(s)\}\, ds\right). \end{aligned} \tag{2.39}$$

Segundo Hsieh *et al*. (2006), a escolha de $h_0(\cdot)$ deve ser realizada de maneira diferente a considerada na análise de sobrevivência padrão, onde não se especifica $h_0(\cdot)$, porém no caso da modelagem conjunta de dados longitudinais e de sobrevivência, essa escolha pode levar a uma subestimação dos erros padrão das estimativas dos parâmetros.

As propostas encontradas na literatura para modelar de forma flexível a função de risco basal podem ser vistas nos trabalhos de Herndon e Jr (1990), Rosenberg (1995) e Whittemore e Keller (1986).

Com a finalidade de quantificar o efeito da covariável longitudinal para o risco de um evento, é necessário estimar $m_i(t)$ e reconstruir a trajetória das variáveis longitudinais $\mathcal{M}_i(t)$ para cada objeto. Deve-se considerar um modelo de efeitos mistos adequado para descrever a evolução do tempo específico do objeto. Inicialmente, considera-se os resultados longitudinais normalmente distribuídos e um modelo misto linear, do seguinte modo:

$$\begin{cases} y_i(t) = m_i(t) + \varepsilon_i(t) \\ m_i(t) = \boldsymbol{x}_i^\top(t)\boldsymbol{\lambda} + \boldsymbol{z}_i^\top(t)\boldsymbol{b}_i \\ b_i \sim \mathcal{N}(0,\boldsymbol{D}_i), \quad \varepsilon_i(t) \sim \mathcal{N}(0,\sigma^2), \end{cases} \tag{2.40}$$

em que $\boldsymbol{x}_i(t)$ é o vetor de variáveis explicativas associadas aos efeitos fixos para a i-ésima unidade no tempo t, $\boldsymbol{\lambda}$ é o vetor de coeficientes de regressão e $\boldsymbol{z}_i(t)$ é o vetor de variáveis explicativas associadas aos efeitos aleatórios $\boldsymbol{b}_i$, e o termo de erro $\varepsilon_i(t)$. $\boldsymbol{b}_i$ seguem uma distribuição normal multivariada com média zero e variância-covariância $\boldsymbol{D}_i$. Assume-se que o erro seja mutuamente independentes, independentes dos efeitos aleatórios e normalmente distribuído com média zero e variância σ^2.

2.4.1 Estimação dos parâmetros do modelo conjunto

Para se ajustar uma distribuição a um conjunto de dados é necessária a estimação dos seus parâmetros. Segundo Wulfsohn e Tsiatis (1997), um método de estimação apropriado para modelos conjuntos é o da máxima verossimilhança.

Para o entendimento do método da máxima verossimilhança, considera-se que o vetor de efeitos aleatórios $\boldsymbol{b}_i$ é subjacente aos processos longitudinal e de sobrevivência, e que, condicional a $\boldsymbol{b}_i$ estes processos são independentes. Formalmente, tem-se que:

$$p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}) \quad \text{e} \quad p(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}\}, \tag{2.41}$$

em que $i = 1, \cdots, n$, sendo n o número total de indivíduos, e $j = 1, \cdots, n_i$, sendo n_i o número de observações para a i-ésima unidade. $\boldsymbol{\theta} = (\boldsymbol{\theta}_t^\top, \boldsymbol{\theta}_y^\top, \boldsymbol{\theta}_b^\top)^\top$ denota o vetor de parâmetros completo, dividido em três subvetores que contém os parâmetros específicos para o respectivo componente do modelo: $\boldsymbol{\theta}_t^\top$ representa os parâmetros associados ao processo de sobrevivência (tempo de sobrevivência); $\boldsymbol{\theta}_y^\top$ refere-se aos parâmetros associados ao processo longitudinal; $\boldsymbol{\theta}_b^\top$ corresponde aos parâmetros associados aos efeitos aleatórios $\boldsymbol{b}_i$, esses parâmetros estão relacionados à variabilidade entre as unidades experimentais no contexto tanto do processo de sobrevivência quanto do processo longitudinal. $\boldsymbol{y}_i$ é o vetor de dimensão $n_i \times 1$ relacionado às observações para as respostas longitudinais do i-ésimo sujeito e $p(\cdot)$ uma função de densidade de probabilidade apropriada.

A contribuição do i-ésimo sujeito para a log-verossimilhança conjunta pode ser expressa da seguinte forma:

$$\begin{aligned} \log p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{\theta}) &= \log \int p(T_i, \delta_i, \boldsymbol{y}_i, \boldsymbol{b}_i|\boldsymbol{\theta}) db_i \\ &= \log \int p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\lambda}) \left[\prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i; \boldsymbol{\theta}_y\} \right] p(\boldsymbol{b}_i|\boldsymbol{\theta}_b) db_i, \end{aligned} \tag{2.42}$$

com a densidade condicional para a parte de sobrevivência $p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\lambda})$ assumindo a forma:

$$\begin{aligned} p(T_i, \delta_i|\boldsymbol{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\lambda}) &= h_i(T_i|\mathcal{M}(T_i); \boldsymbol{\theta}_i, \boldsymbol{\lambda})^{\delta_i} S_i(T_i|\mathcal{M}(T_i); \boldsymbol{\theta}_t, \boldsymbol{\lambda}) \\ &= \left[h_0(T_i) \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \varphi m_i(T_i)\} \right]^{\delta_i} \\ &\quad \times \exp\left(- \int_0^{T_i} h_0(s) \exp\{\boldsymbol{\gamma}^\top \boldsymbol{w}_i + \varphi m_i(T_i)\} ds \right), \end{aligned} \tag{2.43}$$

em que $h_0(\cdot)$ pode ser qualquer função positiva do tempo e a função de sobrevivência S_i é dada pela Expressão (2.39). A densidade conjunta para as respostas longitudinais juntamente com os efeitos aleatórios é dada por:

$$\begin{aligned} p(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta})p(\boldsymbol{b}_i;\boldsymbol{\theta}) &= \prod_j p\{y_i(t_{ij})|\boldsymbol{b}_i;\boldsymbol{\theta}_y\}p(\boldsymbol{b}_i;\boldsymbol{\theta}_b) \\ &= (2\pi\sigma^2)^{-n_i/2}\exp\{-\parallel \boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{Z}_i\boldsymbol{b}_i \parallel /2\sigma^2\} \\ &\quad \times (2\pi)^{-q_b/2}\det(\boldsymbol{D}_i)^{-1/2}\exp(\boldsymbol{b}_i^\top \boldsymbol{D}_i^{-1}\boldsymbol{b}_i/2), \end{aligned} \tag{2.44}$$

em que q_b denota a dimensionalidade do vetor de efeitos aleatórios e $\parallel \boldsymbol{x} \parallel = \left\{\sum_i x_i^2\right\}^{1/2}$ é a norma euclidiana do vetor $\boldsymbol{x}$.

A maximização do logarítimo da verossimilhança $\ell(\boldsymbol{\theta}) = \sum_i \log p(T_i, \delta_i, \boldsymbol{y}_i|\boldsymbol{\theta})$ em relação a $\boldsymbol{\theta}$ pode ser realizada utilizando algoritmos como o de *Expectation-Maximization* ou de Newton-Raphson (Rizopoulos, 2012).

De acordo com Mazzoleni (2020), com a finalidade de maximizar a função de verossimilhança, o vetor escore associado a log-verossimilhança é dado por:

$$\mathcal{S}(\boldsymbol{\theta}) = \sum_i \int \boldsymbol{A}(\boldsymbol{\theta}, \boldsymbol{b}_i)p(\boldsymbol{b}_i|T_i, \delta_i, \boldsymbol{y}_i;\boldsymbol{\theta})db_i \tag{2.45}$$

em que $\boldsymbol{A}(\cdot)$ denota o vetor escore completo dos dados (o desenvolvimento desta expressão é apresentado no Apêndice A), descrito por:

$$\boldsymbol{A}(\boldsymbol{\theta}, \boldsymbol{b}_i) = \frac{\partial}{\partial \boldsymbol{\theta}^\top}\left[\log p(T_i, \delta_i|\boldsymbol{b}_i;\boldsymbol{\theta}) + \log p(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta}) + \log p(\boldsymbol{b}_i;\boldsymbol{\theta})\right] \tag{2.46}$$

Comumente, ao ajustar modelos conjuntos para dados longitudinais e de sobrevivência, as integrais existentes na especificação do vetor escore não possuem uma solução analítica, ou seja, não apresentam forma fechada. Uma alternativa a ser utilizada é a a quadratura de Gauss-Hermite (Rizopoulos, 2012).

Segundo Murray e Philipson (2022), em muitos casos, a complexidade destes modelos podem tornar sua utilização inviável, principalmente na implementação de muitas abordagens clássicas, pois a inclusão dos efeitos aleatórios os tornam computacionalmente exigente. Também podem ser encontrados problemas de convergência, e, embora existam na literatura abor-

dagens utilizadas para a solução de tal problema, ainda são a principal razão destes modelos conjuntos não serem uma opção direta para esse tipo de análise.

2.4.2 Predição da probabilidade de sobrevivência

Além do modelo conjunto utilizar uma estrutura capaz de ligar os submodelos longitudinais e de sobrevivência, ele permite realizar previsões específicas de cada indivíduo, representando uma melhoria em relação aos modelos de sobrevivência tradicionais (Papageorgiou et al., 2019).

Conforme descrito por Rizopoulos (2012), considerando um modelo conjunto ajustado para uma amostra aleatória $\mathcal{D}_n = \{T_i, \delta_i, y_i\}$, em que $i = 1, \ldots, n$, tem-se o interesse em realizar predições das probabilidades de sobrevivência para um novo sujeito i, o qual possui um conjunto de medidas longitudinais $\mathcal{Y}_i(t) = \{y_i(s); 0 \leq s < t\}$ e possui um vetor de covariáveis fixas $\boldsymbol{w}_i$.

Considerando a probabilidade condicional de sobreviver ao tempo $u > t$ dado a sobrevivência até t, tem-se:

$$\pi_i(u|t) = Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{w}_i, \mathcal{D}_n; \boldsymbol{\theta}^*), \tag{2.47}$$

em que $t > 0$ e θ^* representa os verdadeiros valores dos parâmetros.

Esta probabilidade condicional pode ser reescrita como:

$$\Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) = \int \frac{\mathcal{S}_i\{u|\mathcal{M}_i(u, \boldsymbol{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{\mathcal{S}_i\{t|\mathcal{M}_i(t, \boldsymbol{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}} p(\boldsymbol{b}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{b}_i, \tag{2.48}$$

em que $\mathcal{S}_i(\cdot)$ é a função de sobrevivência, $\mathcal{M}_i(\cdot)$ representa o histórico longitudinal e $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_t^\top, \boldsymbol{\theta}_y^\top, \boldsymbol{\theta}_b^\top\right)$ é o vetor de parâmetros do modelo conjunto (o desenvolvimento desta expressão é apresentado no Apêndice A).

A partir deste resultado, Rizopoulos (2012) propôs o estimador de primeira ordem de $\pi_i(u|t)$, utilizando a estimativa Bayesiana empírica para $\boldsymbol{b}_i$, da seguinte forma:

$$\tilde{\pi}_i(u|t) = \frac{\mathcal{S}_i\left\{u|\mathcal{M}_i\left(u, \hat{\boldsymbol{b}}_i^{(t)}, \hat{\theta}\right); \hat{\theta}\right\}}{\mathcal{S}_i\left\{t|\mathcal{M}_i\left(t, \hat{\boldsymbol{b}}_i^{(t)}, \hat{\theta}\right); \hat{\theta}\right\}} + O([n_i(t)]^{-1}), \tag{2.49}$$

em que $\hat{\boldsymbol{\theta}}$ representa as estimativas de máxima verossimilhança, $\hat{\boldsymbol{b}}_i^{(t)}$ a moda da distribuição condicional $\log p(\boldsymbol{b}_i|T_i^* > t, \mathcal{Y}_i(t); \hat{\boldsymbol{\theta}})$, e $n_i(t)$ é o número de respostas longitudinais para o sujeito i até o tempo t.

Com a finalidade de produzir erros padrão válidos, Rizopoulos (2011) e Proust-Lima e Taylor (2009) propuseram a utilização de esquemas de simulação Monte Carlo.

REFERÊNCIAS

ATLAM, M.; TORKEY, H.; EL-FISHAWY, N.; SALEM, H. Coronavirus disease 2019 (covid-19): survival analysis using deep learning and cox regression model. **Pattern Analysis and Applications**, Springer, v. 24, n. 3, p. 993–1005, 2021.

BINDER, H.; SCHUMACHER, M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. **BMC Bioinformatics**, Springer, v. 9, n. 1, p. 1–10, 2008.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. 2. ed. São Paulo: Sociedade Brasileira de Matemática, 2010.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada.** Editora Blucher, 2006.

COSTA, T. R. da. **Modelos Lineares Mistos: Uma aplicação na produção de leite de vacas da raça Sindi**. 79 f. Dissertação (Mestrado em Biometria e Estatística Aplicada) — Universidade Federal Rural de Pernanbuco, 2010.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972.

COX, D. R.; OAKES, D. **Analysis of survival data**. New York: Chapman and Hall/CRC, 1984.

DIGGLE, P.; HEAGERTY, P.; LIANG, K. Y.; ZEGER, S. **Analysis of longitudinal data**. New York: Oxford university press, 2002.

DO HA, I.; LEE, Y. Multilevel mixed linear models for survival data. **Lifetime Data Analysis**, Springer, v. 11, n. 1, p. 131–142, 2005.

HEDEKER, D.; GIBBONS, R. D. **Longitudinal data analysis**. Hoboken: Wiley-Interscience, 2006.

HERNDON, J. E.; HARRELL JUNIOR, F. E. The restricted cubic spline hazard model. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 19, n. 2, p. 639–663, 1990.

HOSMER, D. W.; LEMESHOW, S. **Applied survival analysis: regression modelling of time to event data**. Eur Orthodontic Soc, p. 386, 1999.

HSIEH, F.; TSENG, Y.-K.; WANG, J.-L. Joint modeling of survival and longitudinal data: likelihood approach revisited. **Biometrics**, Wiley Online Library, v. 62, n. 4, p. 1037–1043, 2006.

HU, J.; SZYMCZAK, S. A review on longitudinal data analysis with random forest. **Briefings in Bioinformatics**, v. 24, n. 2, p. bbad002, 2023.

IBRAHIM, J. G.; CHU, H.; CHEN, L. M. Basic concepts and methods for joint models of longitudinal and survival data. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 28, n. 16, p. 2796, 2010.

JINADASA, K. Applications of the matrix operators vech and vec. **Linear Algebra and its Applications**, Elsevier, v. 101, p. 73–79, 1988.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KLEINBAUM, D. G.; KLEIN, M. **Survival analysis**. New York: Springer, 2010.

LAI, C. D.; MURTHY, D. N.; XIE, M. Weibull distributions and their applications. In: Springer Handbooks. **Springer**, p. 63-78., 2006.

LAIRD, N. M.; WARE, J. H. Random-effects models for longitudinal data. **Biometrics**, JSTOR, p. 963–974, 1982.

LIU, X. **Survival analysis: models and applications**. Chichester: John Wiley & Sons, 2012.

MACHIN, D.; CHEUNG, Y. B.; PARMAR, M. **Survival analysis: a practical approach**. Chichester: John Wiley & Sons, 2006.

MAIORANO, A. C. **Modelagem conjunta de dados longitudinais e de sobrevivência para avaliação de desfechos clínicos do parto**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Departamento de Estatística, Universidade Federal de São Carlos, 2018.

MAZZOLENI, M. Joint models for time-to-event and multivariate longitudinal data: a likelihood approach. **Statistica Applicata-Italian Journal of Applied Statistics**, n. 2, p. 161–180, 2020.

MOORE, D. F. **Applied survival analysis using R**. New York: Springer, 2016.

MURRAY, J.; PHILIPSON, P. A fast approximate em algorithm for joint models of survival and multivariate longitudinal data. **Computational Statistics Data Analysis**, v. 170, p. 107438, 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167947322000184. Acesso em: 05 jan. 2023.

OLIVEIRA, R. A. de; BATISTELA, G. C.; SIMÕES, D.; PADOVANI, C. R. Aplicação de modelos lineares de efeitos mistos para avaliar a densidade básica da madeira de duas espécies e um híbrido de eucalyptus. **Scientia Forestalis**, v. 129, n. 49, p. e3201, 2021.

PAPAGEORGIOU, G.; MAUFF, K.; TOMER, A.; RIZOPOULOS, D. An overview of joint modeling of time-to-event and longitudinal outcomes. **Annual review of statistics and its application**, v. 6, p. 223–240, 2019.

PINHEIRO, J.; BATES, D. **Mixed-effects models in S and S-PLUS**. New York: Springer science & business media, 2006.

PROUST-LIMA, C.; TAYLOR, J. M. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment psa: a joint modeling approach. **Biostatistics**, v. 10, n. 3, p. 535–549, 2009.

RIZOPOULOS, D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. **Biometrics**. v. 67, n. 3, p. 819–829, 2011.

RIZOPOULOS, D. **Joint models for longitudinal and time-to-event data: With applications in R**. New York: CRC press, 2012.

ROSENBERG, P. S. Hazard function estimation using b-splines. **Biometrics**, JSTOR, p. 874–887, 1995.

ROSSELLO, X.; GONZÁLEZ-DEL-HOYO, M. Survival analyses in cardiovascular research, part i: the essentials. **Revista Española de Cardiología (English Edition)**, v. 75, n. 1, p. 67–76, 2022. Disponível em: https://www.sciencedirect.com/science/article/pii/S188558572100181X. Acesso em: 11 jun. 2023.

SINGER, J. M.; ANDRADE, D. d. Análise de dados longitudinais. **Simpósio Nacional de Probabilidade e Estatística**, Embrapa São Paulo, v. 7, 1986.

THERNEAU, T. M.; GRAMBSCH, P. M. The Cox Model. In: **Modeling Survival Data: Extending the Cox Model**. Statistics for Biology and Health. New York, NY: Springer, 2000.

VERBEKE, J.; COOLS, R. The newton-raphson method. **International Journal of Mathematical Education in Science and Technology**, Taylor & Francis, v. 26, n. 2, p. 177–193, 1995.

VERBEKE, G; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. Springer New York, 1997.

WHITTEMORE, A. S.; KELLER, J. B. Survival estimation using splines. **Biometrics**, v. 42, n. 3, p. 495–506, 1986. Disponível em: http://www.jstor.org/stable/2531200. Acesso em: 13 mar. 2023.

WULFSOHN, M. S.; TSIATIS, A. A. A joint model for survival and longitudinal data measured with error. **Biometrics**, v. 53, n. 1, p. 330–339, 1997. Disponível em: http://www.jstor.org/stable/2533118. Acesso em: 13 mar. 2023.

SEGUNDA PARTE - ARTIGOS

ARTIGO 1 - *Estudo de simulação de cenários viáveis de ocorrerem problemas de convergência numérica em ajuste de modelos conjuntos de dados de sobrevivência e longitudinais*

Redigido conforme as normas da Revista Brasileira de Computação Aplicada (versão em processo de submissão).

Revista Brasileira de Computação Aplicada, November, 2023

DOI: 10.5335/rbca.v11i3.9999
Vol. 11, Nº 3, pp. 1–9
Homepage: seer.upf.br/index.php/rbca/index

RBCA
ISSN 2176-6649

ORIGINAL PAPER

Simulation study of viable scenarios for potential numerical convergence issues in fitting joint models for longitudinal and survival data

Daiane de Oliveira Gonçalves, Natália da Silva Martins Fonseca, Marcelo Ângelo Cirillo

Received: . **Revised:** . **Accepted:** .

Abstract

Studies concerning the characteristics of phenomena/experiments over time, such as longitudinal studies or those focused on the time until an event of interest occurs, are increasingly essential in various fields. There may be instances where the investigation of the relationship between one or more longitudinal responses and an event of interest is warranted, a task achievable through the joint model of longitudinal and survival data. However, these models may have convergence problems and be computationally demanding, making their use unfeasible in many cases. In consideration of these factors, the objective of this study is to conduct a Monte Carlo simulation study involving various censoring percentages and correlation structures. The proposed cross-coverage probability will be employed as a diagnostic tool to identify circumstances conducive to numerical convergence, aiming to obtain maximum likelihood estimates for joint models applied to longitudinal and survival data. The results indicated similarity in terms of inference among the models, accounting for the impact of both the correlation structure and the censoring percentage. It was determined that the cross-coverage probability contributes to diagnosing the favorable behavior of the data, thereby facilitating the implementation of joint modeling.

Keywords: Censorship; longitudinal data; mixed linear models; simulation; survival analysis

Resumo

Estudos relacionados a características de fenômenos/experimentos no tempo, como estudos longitudinais ou do tempo até a ocorrência de um evento de interesse, se fazem cada vez mais necessários em diversas áreas. Podem existir situações em que se objetiva investigar a relação entre uma ou mais respostas longitudinais e um evento de interesse, que pode ser realizada com o auxilio da modelagem conjunta de dados longitudinais e de sobrevivência. Entretanto, esses modelos podem apresentar problemas de convergência e serem computacionalmente exigentes, tornando inviável a utilização dos mesmos em muitos casos. Tendo em vista esses fatores, o objetivo deste trabalho é realizar um estudo de simulação de Monte Carlo envolvendo diversas percentagens de censura e estruturas de correlação. A probabilidade de cobertura cruzada proposta será utilizada como ferramenta de diagnóstico para identificar circunstâncias favoráveis à convergência numérica, visando à obtenção de estimativas de máxima verossimilhança para modelos conjuntos aplicados a dados longitudinais e de sobrevivência. Como resultados, verificou-se a existência similaridade em termos de inferência entre os modelos, com efeito da estrutura de correlação e do percentual de censura. Constatou-se que a probabilidade de cobertura cruzada contribui com um diagnóstico sobre o bom comportamento dos dados, auxiliando para realização da modelagem conjunta.

Palavras-Chave: Censura; dados longitudinais; modelos lineares de efeitos mistos; simulação; análise de sobrevivência

1 Introduction

Many pieces of information are currently collected over time, known as longitudinal data, obtained from the same sample elements over an extended period. Longitudinal data represent repeated observations of a random variable of interest, collected at different time points for the same individual or object Hu and Szymczak (2023).

In statistics, numerous methodologies are available for analyzing such data. Among these techniques, mixed-effects linear and survival models stand out, with the latter being particularly useful when dealing with censored data (incomplete observations of the response variable).

Mixed-effects linear models are defined as models that include both fixed effects and random effects. They are primarily used to describe the relationship between a response variable and covariates in data grouped according to one or more classification factors Pinheiro and Bates (2006).

These models enable the prediction of how individual response trajectories change over time and the estimation of parameters describing how the mean response changes in the population of interest. They can accommodate any degree of imbalance in the data, meaning that the number of measurements does not need to be the same for each individual or object. Additionally, random effects account for the correlation between repeated measures in a relatively efficient manner Verbeke et al. (1997).

Survival models are designed for situations where the goal is to evaluate the time until the occurrence of one or more events of interest, often referred to as failures. However, the exact time of occurrence of the event of interest is not always known, or the event may not be observed at all, leading to censoring in survival models. Censored observations are partial or incomplete observations of the response variable Colosimo and Giolo (2006).

Thus, survival models are distinguished by their capacity to accommodate these incomplete (censored) observations in analysis, thereby enabling robust statistical conclusions by incorporating information about the time until the occurrence of the event of interest for the sampled elements.

There is also the possibility to investigate the relationship between one or more longitudinal responses and an event of interest. The statistical treatment of responses repeated over a period of time and observed in the same experimental unit can be applied in different situations involving specific models. In view of the above and given a longitudinal study considering n individuals, the use of a joint model (Viviani et al., 2014) allows the time until the occurrence of an event of interest to be modeled, including covariates that vary over time. In this case, Wu and Carroll (1988) suggest joint modeling using survival analysis techniques with random effects models.

The relationship between the mixed linear models with analysis of survival data such that random effects act linearly on the survival time of the individual or experimental unit is mentioned by Do Ha and Lee (2005). Rizopoulos (2012) includes random effects in survival data, allowing for the prediction of dynamic individual response trajectories over the observed period.

A joint model that simultaneously contemplates the longitudinal responses in the presence of censoring has been proposed. Zhang et al. (2014) recommend applying this in situations represented by survival models with measurement errors, missing data with time-dependent covariates and longitudinal models. However, in many cases, the numerical complexity of fitting these models can make them unfeasible since including random effects becomes computationally demanding as their dimensionality increases (Murray and Philipson, 2022).

Notably, the longitudinal process and the survival process are associated with latent variables. In this context, Rizopoulos and Lesaffre (2014) highlight that models with latent variables are defined based on the assumption of conditional independence. In practice, these models are difficult to implement since the specified integral with respect to the latent variable does not have a form. Therefore, numerical integration is needed, making these models very computationally demanding.

Another important issue is mentioned by Rizopoulos (2010): considering the accelerated time to failure, the specification of the joint model requires a complete longitudinal history for calculating the survival function and the risk function; in many applications, individuals and/or units may exhibit highly nonlinear longitudinal trajectories.

Given the previous description and considering the convergence problems that may occur, the use of latent variables and their implications in solving the integral with required computational demand, preliminarily evaluating the behavior of the data through individual process modeling of survival and longitudinal is worth investigating since similar parameter estimation results may otherwise occur. Therefore, the performance of a joint model can be better analyzed than that of other models.

This perspective justifies the contributions of this study, which presents a methodology that obtains the cross coverage probability. In the proposed methodology, the estimates of the longitudinal model parameters are computed based on the confidence interval of the parameters of the survival model. Thus, the coverage probabilities for the survival model are generated by inverting the intervals.

The main contribution of this work is the introduction and application of cross-coverage probability as a diagnostic tool. This tool is employed to identify circumstances conducive to numerical convergence in obtaining maximum likelihood estimates for joint models applied to longitudinal and survival data. This diagnostic significantly aids in overcoming convergence issues and the computational demands often associated with these models, thereby enhancing applicability in studies involving such data types and yielding more precise results while leveraging the advantages these models offer.

In view of the above, this study proposes using the measure of the probability of cross-coverage as a diagnostic tool for connecting longitudinal and survival models. This can help the researcher estimate a joint model that involves both processes and minimize possible numerical convergence problems.

2 Materials and methods

For a better compression of the construction of the panel of data with repeated measures in the absence and presence of censure, as well as the notation used in the subsequent sections, the layout described in Table 1 is followed.

The longitudinal process and simulated survival,

Table 1: Panel data layout with repeated measures ($m = 1, \cdots, M$), within each group ($g = 1, \cdots, G$) censored (δ).

Longitudinal Process			Survival Process	
Y	g	X	W	δ
y_{11}	1	x_{11}	w_{11}	δ_{11}
y_{21}	1	x_{21}	w_{21}	δ_{21}
⋮	⋮	⋮	⋮	⋮
y_{M1}	1	x_{m1}	w_{m1}	δ_{m1}
⋮	⋮	⋮	⋮	⋮
y_{1G}	G	x_{1g}	w_{1g}	δ_{1g}
y_{2G}	G	x_{2g}	w_{2g}	δ_{2g}
⋮	⋮	⋮	⋮	⋮
y_{MG}	G	x_{mg}	w_{mg}	δ_{mg}

including the categorical covariates based on this structure are described below in sections 2.1 – Monte Carlo simulation of the multilevel model for the longitudinal process; 2.2 – Monte Carlo simulation of the Weibull model for the survival process; 2.3 – Definition of the simulation scenarios and parametric values; and 2.4 – Adjustment of the models for the survival and longitudinal processes with inclusion of categorical covariates and estimation of the probabilities of cross coverage.

2.1 Monte Carlo simulation of the multilevel model for the longitudinal process

Y was assumed to be the dependent variable in the fit of the multilevel model with the distribution $Y_j \sim N_p\left(\underset{\sim}{\mu_{jg}}, \Sigma_a\right)$, for $j = 1, \cdots, M \cdot G$, where $g = 1, \cdots, G$ such that the dependence relationship with the regressor variable X was maintained by the Eq. (1) (Silva and Cirillo, 2018) and

$$\mu_{jg} = \beta_0\left(m_j - 1\right) + \beta_1 X_{jg}, \tag{1}$$

in which $X_{jg} \sim U(0,1) and \beta_0 = \beta_1 = 0.5$, fixed arbitrarily.

The autoregressive correlation structure of order 1, AR(1), was considered for the definition of the covariance matrix Σ_a, where $a = 1$. Its estimated correlations were given as a function of the α parameters used in the generalized estimation equations approach (Liang and Zeger, 1986) (2)

$$CORR\left(Y_{(g,j)}, Y_{(g,j+t)}\right) = \alpha^t, \text{where} \quad t = 1, \cdots, T. \tag{2}$$

For $\sum_2(a = 2)$, we proceeded by including the interchangeable correlation structure, according to the Eq. (3).

$$CORR\left(Y_{gj}, Y_{gj'}\right) = \begin{cases} 1, & \text{if } j = j' \\ \alpha, & \text{if } j \neq j' \end{cases}. \tag{3}$$

The inclusion of the degree of correlation ρ in the estimates of α in Eq. (2) and Eq. (3) was performed using the method for obtaining the limiting estimates of the covariance matrix proposed by Silva and Cirillo (2018). This method was applied to the GEE 2 models according to Eq. (4) and Eq. (5).

$$\alpha_0(1-\alpha_0)^{-1}\left\{\frac{t-(1-\alpha_0^t)}{1-\alpha_0}\right\} - t(t-1)\frac{\rho}{2} = 0, \tag{4}$$

where $-1/(t-1) \leq \rho \leq 1$, and

$$\alpha_0 = 2\rho\left\{\frac{t-(1-\rho^t)/(1-\rho)}{t(t-1)(1-\rho)}\right\}, \tag{5}$$

where $-1 \leq \rho \leq 1$.

The restriction presented in Eq. (4) is performed assuming that the exchangeable correlation matrix is true when considering it as a working correlation matrix; analogously, applying the restriction presented in Eq. (5) assumes the AR(1) structure to be true (Sutradhar and Das, 2000).

2.2 Monte Carlo simulation of the survival model

The survival process was simulated so that the percentage of censorship was controlled in the generated sample. To do so, the procedure used in Giarola et al. (2018) assumed two auxiliary random variables, $W_1 \sim Weibull(\alpha_1, \beta_1)$ and $W_2 \sim Weibull(\alpha_2, \beta_2)$. In this way $Z = W_2 - W_1$ was defined with the condition that $\alpha_1 = \alpha_2 = \alpha_c$. Substituting this into F_Z, we obtained in Eq. (6)

$$\begin{aligned} F(z) &= \int_0^\infty w_1^{\alpha_c - 1} - \exp\left\{-\left(\frac{w_1}{\beta_1}\right)^{\alpha_c} - \left(\frac{w_2}{\beta_2}\right)^{\alpha_c}\right\} dw_1 \\ &= \frac{1}{\alpha_c\left(\frac{1}{\beta_1^{\alpha_c}} + \frac{1}{\beta_2^{\alpha_c}}\right)}. \end{aligned} \tag{6}$$

Therefore,

$$F_z = \frac{\beta_1^{\alpha_c}}{\beta_1^{\alpha_c} + \beta_2^{\alpha_c}}. \tag{7}$$

Thus, given that W_2 considered the censoring time associated with the i-th observation and W_1 considered the failure time, the definition of the censoring percentage P was determined by Eq. (8)

$$P = \frac{\beta_{1g}^{\alpha_c}}{\beta_{1g}^{\alpha_c} + \beta_{2g}^{\alpha_c}}, \text{where} \quad g = 1, \cdots, G, \tag{8}$$

$$\beta_{2g}^* = \beta_{1g}\left(\frac{1-P}{P}\right)^{\frac{1}{\alpha_c}}. \tag{9}$$

Following these specifications, the censoring assignment was given by generating $F \sim Weibull(\alpha_g, \beta_{1g})$, where $g = 1, \cdots, G$ represents the time elapsed until failure, and $C \sim Weibull(\alpha_g, \beta^*_{2g})$ represents the censoring time. Therefore, $W = min(F, C)$ and δ is the censorship indicator, where $\delta = 1$ if $F < C$ and $\delta = 0$ otherwise.

2.3 Definition of simulation scenarios and parametric values

With the variables simulated in both processes as described in the previous sections, scenarios were used in the Monte Carlo simulation under the factor combinations described in Table 2.

The Monte Carlo simulation procedure is justified for simulating samples, which computationally control different scenarios (Table 2). These scenarios serve as instruments for investigating the performance of the joint model, generating empirical distributions of the parameters. From these distributions, it becomes possible to estimate coverage probabilities, providing a more robust and detailed view of the model's behavior under various scenarios.

Table 2: Scenarios considered for the simulation of data with different percentages of censorship (P), number of groups (Ng) and number of measurements (Nmed).

Scenario	P	Structure	Ng	Nmed
1	15	AR(1)	20	30
2	15	AR(1)	20	60
3	15	AR(1)	20	100
4	15	Uniform	20	30
5	15	Uniform	20	60
6	15	Uniform	20	100
7	15	AR(1)	50	30
8	15	AR(1)	50	60
9	15	AR(1)	50	100
10	15	Uniform	50	30
11	15	Uniform	50	60
12	15	Uniform	50	100
13	50	AR(1)	20	30
14	50	AR(1)	20	60
15	50	AR(1)	20	100
16	50	Uniform	20	30
17	50	Uniform	20	60
18	50	Uniform	20	100
19	50	AR(1)	50	30
20	50	AR(1)	50	60
21	50	AR(1)	50	100
22	50	Uniform	50	30
23	50	Uniform	50	60
24	50	Uniform	50	100

The values of the parameters of the Weibull model were defined arbitrarily, $\alpha = 12$ and $\beta = 4$. The correlation between repeated measures in the longitudinal process was determined to be $\rho = 0.5$.

2.4 Fit of the models for the survival and longitudinal processes with inclusion of categorical covariates and estimation of the probabilities of cross-coverage

Given the longitudinal process, the multilevel model was fitted with a random intercept (θ_0) and four categorical covariates ($\theta_1, \cdots, \theta_4$) whose systematic components were defined by the linear predictor Eq. (10).

$$\eta_k = \boldsymbol{\theta} \cdot X_k + \varepsilon, \tag{10}$$

in which $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2, \theta_3, \theta_4)$, $k = 1, \cdots, M \cdot G$ and $\varepsilon \sim N(0, 1)$.

For the survival process, the Weibull model was considered.

$$S(t) = \exp\left\{-\left(\frac{t}{\boldsymbol{\alpha}}\right)^{\beta}\right\}, \tag{11}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$.

Once the estimates $\hat{\theta}$ and $\hat{\alpha}$ were obtained, the asymptotic confidence intervals were computed with the nominal level $\gamma = 0.95$ considering the averages of the estimates of the parameters of the longitudinal and survival models, adjusted in 1000 Monte Carlo realizations Eq. (12).

$$IC(\theta_v, \gamma) = \hat{\theta}_v \pm 1.96\sqrt{\hat{Var}(\hat{\theta}_v)}, \tag{12}$$

$$IC(\alpha_v, \gamma) = \hat{\alpha}_v \pm 1.96\sqrt{\hat{Var}(\hat{\alpha}_v)}, \tag{13}$$

in which $v = 1, \cdots, 4$. v is the number of categorical variables.

As a function of the intervals, the estimate of the cross-coverage probability was denoted by $\widehat{CCP}_{\hat{\alpha}_i}\left(\hat{\theta}_i\right)$ as the frequency of the number of estimates $\hat{\alpha}_i$ obtained with the fit of the survival model, which is contained in the interval $IC(\theta_i, \gamma)$. Similarly, $\widehat{CCP}_{\hat{\theta}_i}(\hat{\alpha}_i)$ was estimated through the frequency of estimates $\hat{\theta}_i$ obtained with the fit of the multilevel model, which is contained in the interval $IC(\alpha_i, \gamma)$.

Following the recommendations of Bradley (1978) and Algina et al. (2005) and maintaining the nominal 95% confidence level, the confidence interval of this probability is [0.925; 0.975].

To obtain the results, a script was prepared in R software, version 4.0.4, for each scenario (Table 2) (R Core Team, 2022).

3 Results and discussion

Before discussing the simulated results, it should be noted that the first parameter of the survival model α_1 can be confused with the intercept (θ) of the longitudinal model. For didactic purposes and better clarification, let us assume the relative risk of the joint model, defined below:

$$h_i\left(t|\mathcal{M}_i(t), w_i\right) = h_0 \exp\left\{\gamma^\top w_i + \varphi m_i(t)\right\}. \quad (14)$$

Specifying h_0 by the distribution Weibull (α, β) results in the following expression.

$$\begin{aligned} h_i\left(t|\mathcal{M}_i(t), w_i\right) &= \frac{\beta}{\alpha^\beta} t^{\beta-1} \exp\left\{\gamma^\top w_i + \varphi m_i(t)\right\} \\ &= \frac{\beta}{\exp\left\{\beta \log(\alpha)\right\}} t^{\beta-1} \exp\left\{\gamma^\top w_i + \varphi m_i(t)\right\} \\ &= \beta t^{\beta-1} \exp\left\{-\beta \log(\alpha) + \gamma^\top w_i + \varphi m_i(t)\right\}. \end{aligned} \quad (15)$$

Therefore, the term $-\beta \log(\alpha)$ is conflated with the intercept effect of the linear predictor of the longitudinal model.

Given this understanding, although no inferences were made about relative risk, the results described in Table 3 correspond to the estimates of the cross-coverage probabilities. These are computed by the estimates of the first parameter of the survival model $\hat{\alpha}_i$ based on the confidence interval for the intercept of the longitudinal model.

Table 3: Probability of cross-coverage of the parameter α_1 of the survival model in relation to the intercept interval estimates θ_0 specified in the longitudinal model.

Scenario	P	Structure	$\widehat{CCP}_{\hat{\alpha}_1}\left(\hat{\theta}_0\right)$
1		AR1	0.9586
4	15%	Uniform	0.9513
9		AR1	0.9435
12		Uniform	0.9477
13		AR1	0.9660
16	50%	Uniform	0.9639
21		AR1	0.9568
24		Uniform	0.9439

The results described in Table 3 demonstrate that regardless of the correlation structure or whether the censorship percentage is specified as 15% or 50%, the estimates of the coverage probabilities approximately relate to the nominal confidence level defined in 95%.

The other scenarios involve the results of the other parameters of the Weibull and multilevel model in relation to the estimates of the cross-coverage probabilities. The extreme case, i.e., fewer groups (Ng=20) and fewer measurements (Nmed=30), follows the graphs illustrated in Fig. 1.

The results described in Fig. 1 demonstrate that, in general, the correlation structure to which the data are correlated has an impact. In this context, the estimates of the cross-coverage probabilities in both models showed discrepancies in at least one of the parameters in under the 95% nominal confidence level, with greater discrepancies when the percentage of censoring was high (50%).

This result did not occur under the uniform correlation structure, as shown in Figure 1(b)-1(d); that is, in both models, the longitudinal and survival processes were connected. In comparison with the results obtained by Villegas et al. (2013) and in relation to other survival models, the effect of correlation and percentages of censoring, different multiple failure approaches applying the Cox proportional hazards model are considered in different simulation scenarios.

Thus, assuming sample sizes fixed in n = (50,100,200,400); percentages of censorship p = 0%, 15%, 30% and 50%; number of recurring events $K = (3, 6, 9, 12)$; and the levels of correlation between the adjacent recurrence times fixed in $\rho = (0, 0.10, 0.45, 0.80)$, and without specifying the correlation structure, the authors concluded that the different approaches are stable against censorship and share a bias as the values increase For K recurrence levels, resulting in asymptotic confidence intervals that are imprecise relative to the specified nominal confidence level.

Fig. 2 illustrates the results of increasing the number of groups (Ng=50) and number of measurements (Nmeg=100). These results show that, given the correlation structure AR(1) (Fig. 2(a)-(c)) and for low censoring proportions, the estimates were reduced. This primarily caused an estimate in one of the parameters to be lower than the specification limit defined at 0.92 for the nominal confidence level. Thus, it has a limited ability to propose any recommendation regarding the existence of some connection between the longitudinal process and the survival process.

Under the uniform correlation, increasing the number of measurements resulted in reduced coverage probabilities in the presence of a high percentage of censorship (50), creating an estimate that is incoherent at the nominal confidence level in at least one of the model parameters of survival.

Štajduhar and Dalbelo-Bašić (2010) adapted the learning algorithms of Bayesian networks using a censorship weighting procedure proposed by Zupan et al. (2000), assuming nine different percentages of censoring p = 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80% and comparing the estimates of the Cox regression model. In this context, they concluded that the weighting procedure should be used with Bayesian networks only with intermediate data censorship (from 40% to 60%). If data censorship is light (up to 30%), the original algorithms should be used.

Lin et al. (2013) performed a simulation study comparing the performance of several maximum likelihood estimation (ML) methods, the log-probit regression method and the nonparametric Kaplan–Meier method (KM). Thus, samples were generated from the following distributions: log-normal, gamma, a mixture of two log-normals and log-normal with 30%

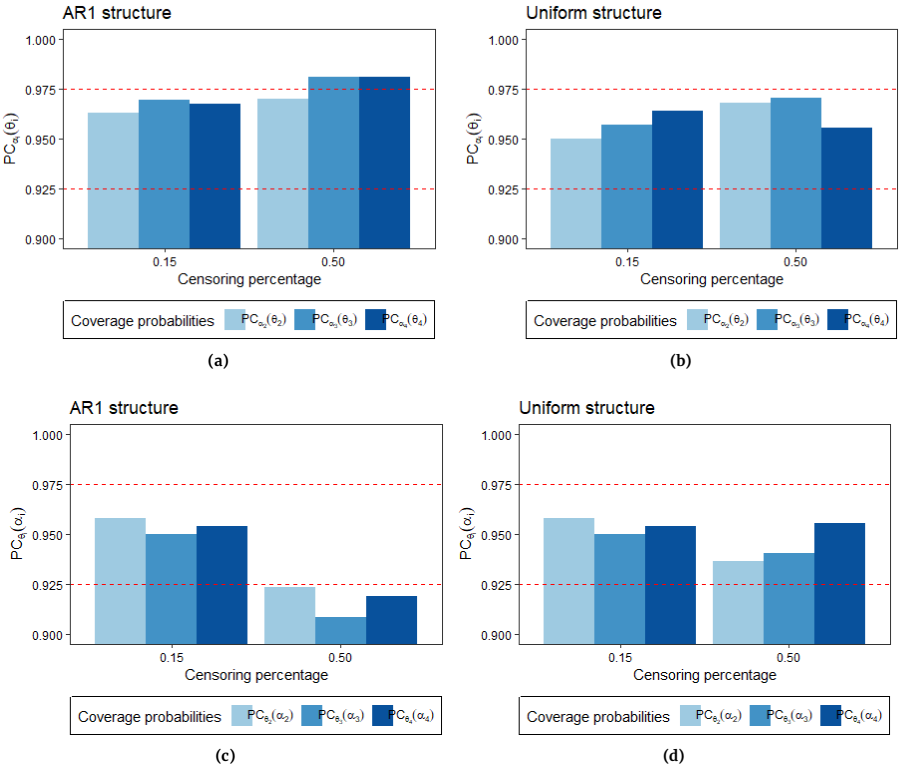


Figure 1: Estimates of the probabilities of cross-coverage, fixing the AR(1) and uniform correlation structures and censoring percentages of 0.15 and 0.50: (a) and (b), parameters of the longitudinal model in relation to the confidence interval of the parameters of the (c) and (d) parameters of the survival model in relation to the confidence interval of the parameters of the longitudinal model.

of observations at zero for different sample sizes. For each distribution evaluated, the percentage of censored observations was randomly generated from a uniform distribution ranging from 20% to 80%.

With these specifications, the results showed that the sample size had little impact on the accuracy of the estimates; however, the percentage of censored samples had the greatest impact, which is comparable to the results obtained by Antweiler and Taylor (2008) in the comparison of the maximum likelihood estimation methods, regression statistics by order and nonparameters for the analysis of left censored data; they concluded that with high percentages of censored data, the interval estimates were imprecise in relation to the nominal confidence level.

4 Conclusions

The proposed procedure used to estimate cross-coverage probabilities as a diagnostic tool for the connection of the longitudinal and survival models, was adequate when considering the Weibull model. Therefore, it can help researchers estimate a joint model that involves both processes and minimize possible numerical convergence problems.

In the context of scenario 13, which includes smaller numbers of groups and measurements involving the AR1 correlation structure, the estimates of the cross-coverage probabilities in both models showed discrepancies in at least one of the parameters when the percentage rate of censorship was 50%, given a specified 95% nominal confidence level. Thus, this rate of censorship presents more harmful results than other rates. Under the same

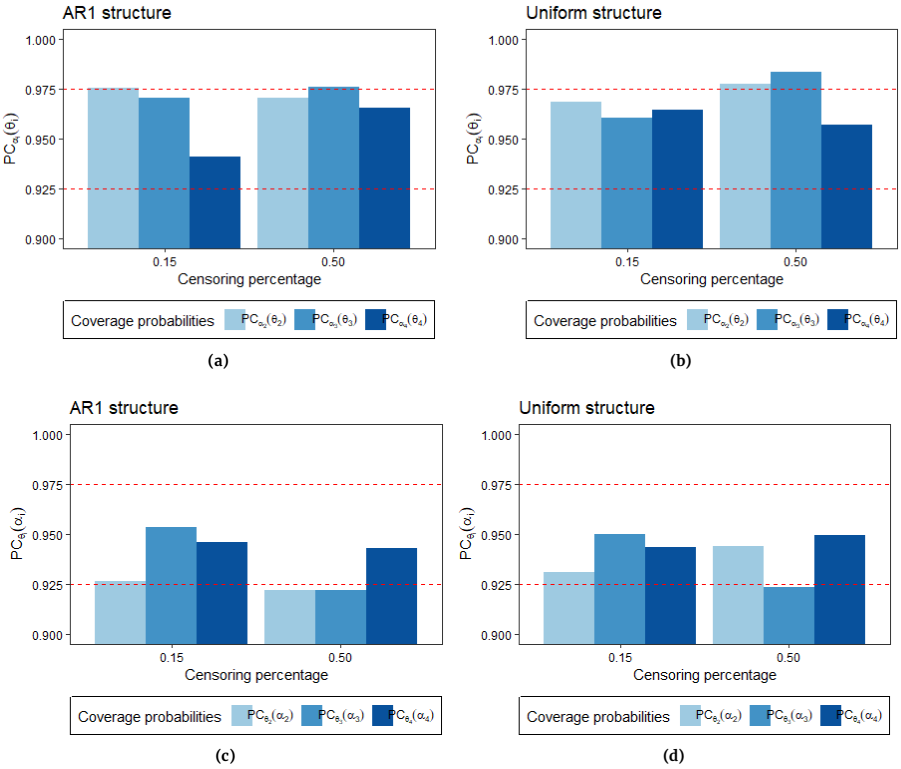


Figure 2: Estimates of the probabilities of cross-coverage, fixing the AR(1) and uniform correlation structures and censoring percentages of 0.15 and 0.50: (a) and (b), parameters of the longitudinal model in relation to the confidence interval of the parameters of the (c) and (d) parameters of the survival model in relation to the confidence interval of the parameters of the longitudinal model.

context but with a uniform correlation structure, as in scenarios 4 and 16, it is noted that a favorable condition exists for numerical convergence in obtaining maximum likelihood estimates for joint models of longitudinal and survival data.

Given the AR(1) correlation structure, increasing the number of groups and measurements and considering low censoring proportions leads to an estimate below the specification limit, which is defined as 0.92 for the nominal confidence level. This finding limits the ability to recommend the utilization of joint models for longitudinal and survival data.

For both correlation structures, increasing the number of measurements resulted in a reduction in the coverage probabilities in the presence of a high percentage of censorship, causing an estimate that was incoherent at the nominal confidence level in at least one of the parameters of the survival model. Thus, increasing the percentage of censorship negatively impacted the numerical convergence for obtaining maximum likelihood estimates of joint models for longitudinal and survival data.

The proposed methodology represents a significant advancement by offering a way to identify circumstances conducive to numerical convergence, which is essential for achieving results in joint modeling. The use of Monte Carlo simulation involving various levels of censoring and correlation structures provides a robust and comprehensive analysis, allowing for the evaluation of different scenarios and validation of the proposed methodology. Ultimately, this approach helps minimize computational challenges and convergence issues associated with joint models, thereby expanding their applicability across various fields.

In future studies, we aim to expand the methodology by considering additional parametric models. This could provide a broader view of scenarios where convergence issues may arise and evaluate the effectiveness of cross-coverage probability in these contexts. The findings of this study may contribute to the future development of computational tools incorporating the proposed methodology, assisting researchers in facilitating the estimation of joint models.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Algina, J., Keselman, H. and Penfield, R. D. (2005). An alternative to cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case., *Psychological methods* **10**(3): 317. https://psycnet.apa.org/doi/10.1037/1082-989X.10.3.317.

Antweiler, R. C. and Taylor, H. E. (2008). Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. summary statistics, *Environmental science & technology* **42**(10): 3732–3738. https://doi.org/10.1021/es071301c.

Bradley, J. V. (1978). Robustness?, *British Journal of Mathematical & Statistical Psychology* **31**(2): 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x.

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

Do Ha, I. and Lee, Y. (2005). Multilevel mixed linear models for survival data, *Lifetime Data Analysis* **11**: 131–142. https://doi.org/10.1007/s10985-004-5644-2.

Giarola, L. T. P., Vivanco, M. J. F., Cirillo, M. A. and Menezes, F. S. (2018). Extended method for several dichotomous covariates to estimate the instantaneous risk function of the aalen additive model, *Journal of Modern Applied Statistical Methods* **17**(1): 27. https://doi.org/10.22237/jmasm/1543852660.

Hu, J. and Szymczak, S. (2023). A review on longitudinal data analysis with random forest, *Briefings in Bioinformatics* **24**(2): bbad002. https://doi.org/10.1093/bib/bbad002.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22. https://doi.org/10.1093/biomet/73.1.13.

Lin, N. X., Logan, S. and Henley, W. E. (2013). Bias and sensitivity analysis when estimating treatment effects from the cox model with omitted covariates, *Biometrics* **69**(4): 850–860. https://doi.org/10.1111/biom.12096.

Murray, J. and Philipson, P. (2022). A fast approximate em algorithm for joint models of survival and multivariate longitudinal data, *Computational Statistics & Data Analysis* **170**: 107438. https://doi.org/10.1016/j.csda.2022.107438.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*, Springer science & business media.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data, *Journal of statistical software* **35**: 1–33. https://doi.org/10.18637/jss.v035.i09.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*, CRC press.

Rizopoulos, D. and Lesaffre, E. (2014). Introduction to the special issue on joint modelling techniques, *Statistical methods in medical research* **23**(1): 3–10. https://doi.org/10.1177/0962280212445800.

Silva, J. A. D. and Cirillo, M. A. (2018). Selection criterion of work matrix as a function of limiting estimates of the covariance matrix of correlated data in gee, *Biometrical Journal* **60**(5): 979–990. https://doi.org/10.1002/bimj.201800035.

Štajduhar, I. and Dalbelo-Bašić, B. (2010). Learning bayesian networks from survival data using weighting censored instances, *Journal of biomedical informatics* **43**(4): 613–622. https://doi.org/10.1016/j.jbi.2010.03.005.

Sutradhar, B. C. and Das, K. (2000). On the accuracy of efficiency of estimating equation approach, *Biometrics* **56**(2): 622–625. https://doi.org/10.1111/j.0006-341X.2000.00622.x.

Verbeke, G., Molenberghs, G. and Verbeke, G. (1997). *Linear mixed models for longitudinal data*, Springer.

Villegas, R., Julià, O. and Ocaña, J. (2013). Empirical study of correlated survival times for recurrent events with proportional hazards margins and the effect of correlation and censoring, *BMC medical research methodology* **13**: 1–10. https://doi.org/10.1186/1471-2288-13-95.

Viviani, S., Alfó, M. and Rizopoulos, D. (2014). Generalized linear mixed joint model for longitudinal and survival outcomes, *Statistics and Computing* **24**: 417–427. https://doi.org/10.1007/s11222-013-9378-4.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process, *Biometrics* pp. 175–188. https://doi.org/10.2307/2531905.

Zhang, D., Chen, M.-H., Ibrahim, J. G., Boye, M. E., Wang, P. and Shen, W. (2014). Assessing model fit in joint models of longitudinal and survival data with applications to cancer clinical trials, *Statistics in Medicine* **33**(27): 4715–4733. https://doi.org/10.1002/sim.6269.

Zupan, B., Demšar, J., Kattan, M. W., Beck, J. R. and Bratko, I. (2000). Machine learning for survival analysis: a case study on recurrence of prostate cancer, *Artificial intelligence in medicine* **20**(1): 59–75. https://doi.org/10.1016/S0933-3657(00)00053-1.

ARTIGO 2 - *Modelagem conjunta de dados longitudinais e de sobrevivência na avaliação do impacto do sistema de irrigação no tempo até a ocorrência da mancha de phoma em café arábica*

Redigido conforme as normas do Journal of the Science of Food and Agriculture (versão em processo de editoração).

Joint modeling of longitudinal and survival data for the evaluation of the impact of irrigation systems on the time until the occurrence of phoma leaf spot in Arabica coffee plants

Daiane de Oliveira Gonçalves[1], Natalia da Silva Martins Fonseca[2], Renata Cristina Martins Pereira[3], Edson Ampélio Pozza[3], Marcelo Ângelo Cirillo[4]

[1] Ph.D. student in Statistics and Agricultural Experimentation, Federal University of Lavras, E-mail: prof.daiane.oliveira@gmail.com
[2] Department of Statistics, Federal University of Alfenas
[3] Department of Plant Pathology, Federal University of Lavras
[4] Department of Statistics, Federal University of Lavras

The coffee crop is prominent in Brazilian agriculture, making the country a global power in this area. One of the main concerns in the coffee sector is disease, which can affect coffee productivity and quality. Thus, it is important to evaluate the factors that may affect coffee quality and thus enhance the development of strategies to reduce coffee losses and costs and optimize production. This study evaluated the influence of the type of irrigation (self-propelled, drip and center pivot) on the time until the occurrence of phoma leaf spot on Arabica coffee plants, considering the intensity of the disease. Additionally, the association between longitudinal incidence and the time until an event of interest was assessed, such as to investigate the optimal level of control, based on the joint modeling of longitudinal and survival data. The results of this study identify the effectiveness of drip irrigation system compared to other systems; the use of such systems was associated with an approximately 46.5% reduction in the risk of leaf spot disease compared to the use of self-propelled irrigation system. The use of a center pivot system increased the risk of disease progression compared to self-propelled. An association between the longitudinal and survival processes was also observed. An increase in the incidence of this disease is associated with a reduced risk of disease progression over time. These findings may contribute to establishing more efficient agricultural practices for coffee cultivation.

Keywords: Survival analysis. Coffee. Censorship. Failure. Incidence of diseases. Joint Model.

1. INTRODUCTION

Brazil is considered the main producer and exporter of coffee globally. According to data from the Instituto Brasileiro de Geografia e Estatística[1], coffee production in Brazil for the two species Arabica and Canephora reached 3.6 million tons in February 2024, representing an increase of 0.4% compared to that in the previous month and of 4.2% compared to that in the same period in 2023. The state of Minas Gerais is the main Brazilian producer of Arabica coffee, accounting for 70.3% of the country's total production estimated for 2024, indicating that the crop is promising this year. Minas Gerais production is projected to reach 1.7 million tons, representing an increase of 0.8% compared to that in the previous year.[2]

Coffee production in Brazil involves the creation of jobs, the generation of resources and the socioeconomic development of the country, comprising a complex

production chain that extends from rural producers to workers involved in various stages of processing, transport and marketing. Thus, it plays an important role in the economy of producing regions, also making a notable contribution to social sustainability.[2]

In addition, coffee plantations span diverse edaphoclimatic conditions, distinguished by their regional particularities. Notably, Brazil is characterized by its varied climates and soils, enabling the cultivation of coffees with unique characteristics and origins, including specific 'terroirs'. The country offers an extensive variety of coffee flavors and aromas that stand out among consumers worldwide.[2]

The microregions that specialize in the production of Arabica coffee are concentrated in the states of Minas Gerais and Bahia.[3] The state of Minas Gerais has the largest number of microregions specializing in the cultivation of Arabica coffee, representing 60% of all Arabica regions and contributing to 75.4% of the total coffee production in the state. These data highlight the fundamental role of Minas Gerais in the Brazilian coffee economy, which is driven by the extensive use of irrigation, mechanization, favorable landscapes and modern and highly efficient production methods in these territories.[4]

Due to the mountainous topography, the southern region of Minas Gerais depends heavily on manual labor, resulting in the constant presence of migrants during the months of May to August, the harvest period. In this scenario, thousands of workers from northern Minas Gerais and the state of Bahia move to procure income and employment on farms.[5] However, with the advent of mechanization in recent years, this is no longer the current reality. Minas Gerais now boasts a modern and highly productive coffee industry, characterized by favorable landscapes and extensive use of irrigation and mechanization.[4] It is worth noting that in economically disadvantaged communities, where access to advanced technology remains limited, conventional irrigation methods continue to be employed. Consequently, preventive studies on the effects of irrigation on coffee production are highly regarded and of significant value.

One of the main concerns in the coffee sector due the effect on coffee productivity and quality is the incidence of diseases, as they have the potential to cause significant financial losses.[6, 7] Among the main diseases observed in coffee plants, phoma leaf spot (Phoma tarda) is common because it can cause leaf drop, drought and the death of productive branches, resulting in significant losses in production.[8]

Symptoms of this disease occur from the stage of seedling cultivation in nurseries

to the production stage of the crop, which places it among the most relevant fungal diseases affecting coffee.[6, 9] In general, the fungus indiscriminately infects the leaves, fruits and branches of coffee plants, regardless of the species.[10] This infestation results in direct damage to final production, as it interferes with physiological processes, leading to the death of flower buds, inadequate development of new shoots, fruit drop and poor fruit formation.[11] This can lead to a considerable reduction in crop productivity.[12]

The intensity of the disease is directly related to the environment and the resistance of the cultivated variety.[13] In Brazil, in the entire area planted with *Coffea arabica*, the cultivars are susceptible to the phoma leaf spot. Among the environmental conditions, the temperature and hours of leaf wetness or the duration of water on the leaf are most important.[8] Therefore, the choice of irrigation system is highly important. Thus, the monitoring and analysis of coffee diseases can improve management practices, irrigation system choices and other measures, leading to reduced economic losses and production optimization.[14]

Therefore, understanding the influence of irrigation systems on the incidence of phoma leaf spot in coffee is necessary in view of the challenges faced in the sector. Notably, it is necessary to develop efficient management strategies to control the incidence of this disease, and irrigation systems play a key role by directly influencing environmental conditions favorable for disease progression. Variability in irrigation systems can result in different levels of leaf moisture, a factor responsible for increasing the infection of the pathogen, which causes phoma spot.

Thus, understanding how different irrigation systems affect the time until the appearance of phoma enables the targeting and implementation of management practices, positively impacting both production costs and environmental issues through the optimization of fungicide use. The influence of the irrigation system on disease incidence may lead to improvements in coffee quality, thereby enhancing competitiveness in the international market.

Some methods are used to analyze the factors that influence coffee productivity, as well as the impact of irrigation systems on coffee cultivation, especially regarding the progression of diseases such as rust. In this context, planning procedures and experimental designs, based on the technique of analysis of variance and linear mixed models, were used in studies on coffee plant productivity and disease incidence. [15,16,17,18] In addition to these methodologies, decision trees were also utilized to analyze the coffee rust epidemic. [19]

Although there are several methods available to monitor the incidence of phoma in coffee, some elements, such as the management methodology, fertilization, spraying, and pruning, in addition to the resistance and lifetime of coffee trees to particular pests or specific treatments, are still beyond the scope of analysis. These factors influence the profitability of the coffee industry.

Once the factors influencing survival are identified, survival analysis can be employed to investigate both the longevity of coffee plants and the development of specific diseases in the plants. However, survival models are rarely used as instruments for the analysis and management of issues involving these variables, especially plant diseases.

Studies related to the lifetime of individuals/objects, represented in this work by each plant, stand out in the literature because they consider data related to the time of occurrence of a particular event of interest, referred to as the time to failure. However, the exact time of occurrence of the event of interest is not always known, or the event may not be observed at all, leading to the concept of censorship in survival models. Censored observations can be understood as partial or incomplete observations of the response variable.[20]

Therefore, it is of interest to evaluate the influence of the type of irrigation system used on the time until the designed control level is reached or an epidemic, such as phoma leaf spot in coffee trees, occurs. The observations are censored because they provide information about the time until the occurrence of the event of interest. Additionally, it is necessary to evaluate the incidence of the disease over time, which provides valuable information about various trends, which is particularly important in the agricultural sector.

The joint modeling of longitudinal and survival data is appropriate when one wants to predict the time to an event with covariates measured longitudinally and related to the event.[21] This approach is based on a structure capable of linking the longitudinal and survival submodels and supports predictions specific to each individual, representing an improvement over traditional survival models.[22] Moreover, the predictions of joint models are often more accurate than those of individual models. These predictions can be used to characterize the progression of a disease, thus supporting decision-making, such as the most appropriate time to perform interventions.

Thus, in this study, the influence of the type of irrigation on the time until the appearance of phoma leaf spot on coffee plants is evaluated. Specifically, the objective is to determine which irrigation system delays the onset of phoma leaf spot, promoting

improvements in productivity and quality and boosting economic growth in the coffee sector.

2. METHODOLOGY

2.1 Data

A sample of 1750 records of the incidence of this disease in coffee plants was used to evaluate the influence of the type of irrigation on the time until the appearance of phoma spot. Data were collected in three coffee-growing areas located in Carmo do Rio Claro, Minas Gerais, Brazil, from August 2012 to December 2014.[14]

The coffee grown in the study region is of the species *Coffea arabica* L., variety Acaiá 474/19, and is watered with different irrigation systems, such as self-propelled, drip and center pivot systems. Weeds were controlled by mowing. The control of leaf miner and berry borer were performed with chemical methods.[14] Regarding fertilization, the managers of coffee plantations adhere to the relevant recommendations for soil fertility management.[23]

The sampling points were georeferenced using the GPS TRIMBLE 4600 LS® and the Total Station TC600®. In each of the areas irrigated with the self-propelled, drip and center pivot systems, 50 sampling points were established, with a total of 150 observations of the incidence of phoma in the coffee samples collected on each evaluation date. The analysis involved the evaluation of five plants per sampling point. The distance between the sampling points in the irrigated areas varied between 30 and 40 meters, accounting for the slope and shape of the terrain. Observations were recorded on days 0, 66, 104, 171, 245, 297, 374, 428, 475, 544, 598 and 658 after the start of follow-up. In the area irrigated in the self-propelled system, no observations were performed on day 658.[14]

An evaluation of the incidence of the disease, expressed as a percentage, was conducted by sampling 12 leaves from each plant using a nondestructive method. Leaves were selected from the middle third of the canopy from the third and fourth randomly chosen branches (Figure 1).[24]

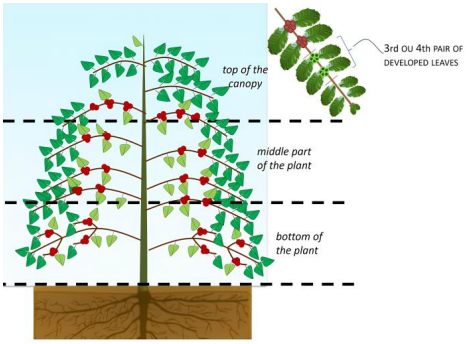


Figure 1 - Schematic representation of the division of the coffee plants used for sampling. Source: Adapted [25]

In summary, 60 leaves were evaluated per sampling point, or observation point, in each non-destructively disease assessment. These previously georeferenced plants were consistently used in all subsequent samplings, which were performed bimonthly throughout the period from August 2012 to December 2014.

Disease incidence was expressed as the percentage of diseased leaves in relation to the total sampled, as defined by Equation 1.[26]

$$I(\%) = (NFD/NFT) \cdot 100, \quad (1)$$

where I (%) indicates the percentage of disease incidence, NFD is the number of diseased leaves and NFT is the total number of leaves. In other words, at each sampling point, the number of diseased leaves among the 60 collected was recorded.

2.2 Statistical analysis

In the data analysis, a joint model was used to evaluate the time until disease onset. The longitudinal model was used to evaluate the progression of phoma leaf spot incidence, considering an explained linear trend over time, and in the survival model, which considered information on the irrigation system, a variable response time was assumed for the onset of disease detection for the plants studied. Self-propelled, drip and center pivot irrigation systems were considered covariates.

Regarding the survival model, the data used were characterized by the presence of censorship, where failure was represented by 1 and 0 represented censorship (absence of phoma disease).

At each sampling point, the time to development of Phoma leaf spot on coffee plants was recorded. If the disease was detected, it was classified as a failure. Conversely, if the disease had not manifested by the end of the study at a given sampling point, the observation was considered censored.

2.2.1 Joint model for longitudinal and survival data

A joint model is capable of considering the dependence and association between longitudinal data and time-to-event data.[27]

The joint model in this study comprises two submodels, namely, a mixed-effects model for longitudinal data and a model related to the time until the occurrence of an event, referring to the survival data.

A longitudinal data analysis was performed because each individual in the population has a specific mean response profile over time. The linear mixed-effects model can be expressed as follows:

$$\begin{cases} y_i = \boldsymbol{X}_i\boldsymbol{\lambda} + \boldsymbol{Z}_i b_i + \varepsilon_i, \\ b_i \sim N(0, \boldsymbol{D}), \\ \varepsilon_i \sim N\left(0, \sigma^2 \boldsymbol{I}_{n_i}\right), \end{cases} \tag{2}$$

in which y_i represents the response or dependent variable for the i-th observation. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known design matrices; for the fixed-effects regression coefficient α and the random-effects regression coefficient b_i, $\boldsymbol{I}_{n_i}$ denotes the n_i−dimensional identity matrix, and n_i is the number of observations for the i-th unit.[28] The random effects are assumed to be normally distributed with a mean of zero and covariance matrix $\boldsymbol{D_i}$, and are assumed independent of the error terms ε_i, that is, $cov(b_i, \varepsilon_i) = 0$.

Regarding the time-to-event submodel, the basic function from survival analysis is used to describe the distribution of T^*, the random variable associated with the studied failure times, and is called the survival function, represented by:

$$S(t) = P(T^* > t) = \int_t^{\infty} f(s)ds, \tag{3}$$

assuming that the random variable T^* is continuous and $f(\cdot)$ is the corresponding probability density function.[28]

The structure of the joint models is represented by T_i, the true time of the event for

the ith individual/object. In other words, T_i is the time of the observed event, defined as the minimum between the potential censoring time C_i and T_i, and $\delta_i = I(T_i^* \leq C_i)$, where δ_i is the failure or censor function. Regarding the time-dependent covariate, $y_i(t)$ is the observed value over time t for the ith individual/object.[29] Importantly, there is not a $y_i(t)$ for all times t, and t_{ij} is only available at times when measurements were obtained. Therefore, the observed longitudinal data consisted of measurements $y_{ij} = y_i(t_{ij})$, where $j = 1, \cdots, n_i$.[28]

$m_i(t)$ denotes the true value of the longitudinal result at time t. To quantify the association between $m_i(t)$ and the risk of an event, the relative risk model can be expressed as follows:

$$\begin{aligned} h_i(t|\mathcal{M}_i(t), \mathbf{w}_i) &= \lim_{dt\to 0} \frac{(Pr\{t \leq T_i^* < t + dt \,|T_i^* \geq t, \mathcal{M}_i(t), \mathbf{w}_i\})}{dt} \quad (4) \\ &= h_0 \exp\{\boldsymbol{\gamma}^T \mathbf{w}_i + \varphi m_i(t)\}, \end{aligned}$$

where $t > 0$; $\mathcal{M}_i(t) = \{m_i(s), 0 \leq s < t\}$ denotes the trajectories of the longitudinal variables at time point t; h_0 is the baseline risk function; $\boldsymbol{w}_i$ is a vector of fixed covariates for the ith individual related to the survival process; $\boldsymbol{\gamma}$ is the vector of regression coefficients associated with $\boldsymbol{w}_i$; and φ is an unknown parameter used to quantify the impact of the longitudinal process on the observation of an event.

In particular, $\exp(\gamma_j)$ denotes the hazard ratio per unit change in $w_{\{ij\}}$ at any time t, and $\exp(\varphi)$ denotes the increase in the relative risk of an event over time t resulting from a unit increase in $m_i(t)$ at the same point in time.

Using the known relationship between the survival function and the risk function, one can obtain the following expression:[28]

$$\begin{aligned} S_i(t|\mathcal{M}_i(t), \boldsymbol{w}_i) &= Pr(T_i^* > t \,|\mathcal{M}_i(t), \boldsymbol{w}_i) \quad (5) \\ &= exp\left(-\int_0^t h_0(s)\, exp\{\,(\,\boldsymbol{\gamma}^T \boldsymbol{w}_i + \varphi m_i(s)\,)\,\}\, ds\right). \end{aligned}$$

To quantify the effect of the longitudinal covariate on the risk of an event, it is necessary to estimate $m_i(t)$ and reconstruct the trajectory of the longitudinal variables $\mathcal{M}_i(t)$ for each object. An adequate mixed-effects model should be used to describe the evolution of an object over time. Initially, longitudinal normally distributed results and a linear mixed model are considered as follows:

$$\begin{cases} y_i = m_i(t) + \varepsilon_i \\ m_i(t) = \boldsymbol{x}_i^{\intercal}(t)\boldsymbol{\lambda} + \boldsymbol{z}_i^{\intercal}(t)\boldsymbol{b}_i \\ b_i \sim N(0, \boldsymbol{D_i}), \varepsilon_i(t) \sim N(0, \sigma^2), \end{cases} \quad (6)$$

where $\boldsymbol{x}_i(t)$is the vector of explanatory variables associated with the fixed effects for the ith unit at time t, $\boldsymbol{\lambda}$ is the vector of regression coefficients, $\boldsymbol{z}_i(t)$ is the vector of explanatory variables associated with random effects $\boldsymbol{b}_i$, and the error is given by $\varepsilon_i(t)$. $\boldsymbol{b}_i$ follows a multivariate normal distribution with a mean of zero and covariance matrix $\boldsymbol{D_i}$. The error is assumed to be mutually independent, independent of random effects and normally distributed with a mean of zero and variance σ^2.

2.2.2 Estimation of the parameters of the joint model

One estimation method appropriate for joint models is the maximum likelihood estimation method. The maximum likelihood estimates are derived as the modes of the log-likelihood function corresponding to the joint distribution of the observed results $\{T_i, \delta_i, y_i\}$.[30]

In the maximum likelihood method, it is assumed that the vector of parameters associated with random effects $\boldsymbol{b}_i$ underlies the longitudinal and survival processes. Formally, defined:[28]

$$p(T_i, \delta_i, \boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) = p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta})\, p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) \text{ e } p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) = \prod_j p\{ y_i(t_{ij}) | \boldsymbol{b}_i; \boldsymbol{\theta}\},$$

in which $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_t^{\intercal}, \boldsymbol{\theta}_y^{\intercal}, \boldsymbol{\theta}_b^{\intercal}\right)^{\intercal}$ denotes the complete parameter vector, divided into three subvectors that contain the specific parameters for the respective model components: $\boldsymbol{\theta}_t^{\intercal}$ represents the parameters associated with the survival process (survival time); $\boldsymbol{\theta}_y^{\intercal}$ denotes the parameters associated with the longitudinal process; $\boldsymbol{\theta}_b^{\intercal}$ corresponds to the parameters associated with random effects $\boldsymbol{b}_i$; and $\mathbf{y}_i$ is the vector of the longitudinal responses of object i. The parameters presented in $\boldsymbol{\theta}$ are related to the variability among the experimental units in the context of both the survival process and the longitudinal process.

The log-likelihood contribution for the i-th subject can be formulated as follows:

$$\log p(T_i, \delta_i, \mathbf{y}_i | \boldsymbol{\theta}) = \log \int p(T_i, \delta_i | \mathbf{b}_i; \boldsymbol{\theta}_t, \boldsymbol{\beta}) \left[\prod_j p\{y_i(t_{ij}) | \mathbf{b}_i ; \boldsymbol{\theta}_y\} \right] p(\mathbf{b}_i; \theta_b) d\mathbf{b}_i \ . \quad (7)$$

Algorithms such as expectation maximization (EM) are used to maximize $\ell(\boldsymbol{\theta}) = \sum_i \log p(T_i, \delta_i, \mathbf{y}_i | \theta)$ in relation to $\boldsymbol{\theta}$.[28] The optimization algorithm starts with EM iterations, and if convergence is not reached, quasi-Newton iterations are used.

To maximize the likelihood function, the score vector associated with the log likelihood is given by:[31]

$$\mathcal{S}(\boldsymbol{\theta}) = \Sigma_i \int \mathbf{A}(\boldsymbol{\theta}, \mathbf{b}_i) \, p(\mathbf{b}_i \mid T_i, \delta_i, \mathbf{y}_i; \boldsymbol{\theta}) \, d\mathbf{b}_i$$

in which $\mathbf{A}(\cdot)$ denotes the complete score vector for the data, described by:

$$\frac{\partial}{\partial \boldsymbol{\theta}^\top} [\log p(T_i, \delta_i \mid \mathbf{b}_i; \boldsymbol{\theta}) + \log p(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\theta}) + \log p(\mathbf{b}_i; \boldsymbol{\theta})]$$

For the statistical analyses, R software, version 4.0.4, was used,[32] with the aid of the JM package.

3. RESULTS AND DISCUSSION

The first descriptive measure observed was an average of 1.47% incidence of phoma spot on coffee leaves, ranging from 0 to 25%. The first evaluation of the plants was performed, and from then on, follow-up measurements were obtained until the disease occurred, with the last evaluation performed 658 days after the first evaluation.

Regarding the systems, plants irrigated by center pivot systems showed, on average, the highest incidence of phoma leaf spot, and these plants also had the highest incidence at 25% (Table 1 and Figure 2).

Table 1 - Incidence of phoma for drip, self-propelled and center pivot irrigation systems

	Minimum	Median	Mean	Maximum
drip	0.00	0.00	0.53	13.33
self-propelled	0.00	0.00	1.31	11.67
center pivot	0.00	1.67	2.57	25.00

Figure 2 shows the observed trajectories and survival curves for each irrigation system. Initially, the objective was to examine the influence of irrigation on the incidence of phoma disease in Arabica coffee leaves.

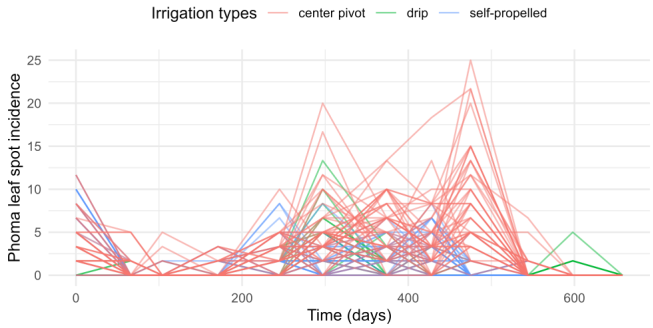


Figure 2 - Individual trajectories observed for each irrigation system: self-propelled, drip and center pivot systems.

When considering the time until the disease occurred, it was necessary to describe the process with censored observations. For the drip irrigation system, 44 of the 50 observations represented failure times, while six were censored. In the case of the self-propelled irrigation system, 47 observations were failures, and three were censored. Notably, 12% of the observations for the drip irrigation system and 6% for the self-propelled irrigation system were censored, as plants irrigated with these methods had not developed disease symptoms by the end of the study. In contrast, the center-pivot irrigation system had no censored data, indicating that Phoma leaf spot was present at all 50 sampling points.

For data with the presence of censoring, Kaplan-Meier estimates can be evaluated considering the different irrigation systems. These estimates include the median time until the occurrence of Phoma spots, along with the corresponding confidence intervals for each system (Table 2). This approach allows for a descriptive assessment of the time-to-event data while accounting for censored observations.

Table 2 - Kaplan-Meier estimates of time (days) until the occurrence of phoma spots in drip, self-propelled, and center pivot irrigation systems

	time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
	66	50	1	0.98	0.0198	0.9420	1.000
	104	49	3	0.92	0.0384	0.8478	0.998
self-propelled	171	46	8	0.76	0.0604	0.6504	0.888
	245	38	25	0.26	0.0620	0.1629	0.415
	297	13	3	0.20	0.0566	0.1149	0.348
	374	10	6	0.08	0.0384	0.0313	0.205
	544	4	1	0.06	0.0336	0.0200	0.180

	104	50	1	0.98	0.0198	0.9420	1.000
	171	49	1	0.96	0.0277	0.9072	1.000
drip	245	48	17	0.62	0.0686	0.4991	0.770
	297	31	20	0.22	0.0586	0.1305	0.371
	428	11	3	0.16	0.0518	0.0848	0.302
	598	8	2	0.12	0.0460	0.0566	0.254
	66	50	12	0.76	0.0604	0.6504	0.888
	104	38	4	0.68	0.0660	0.5623	0.822
center pivot	171	34	8	0.52	0.0707	0.3984	0.679
	245	26	18	0.16	0.0518	0.0848	0.302
	297	8	6	0.04	0.0277	0.0103	0.156
	374	2	2	0.00	-	-	-

Based on the survival curves (Figure 3), the plants irrigated by center pivot systems displayed unfavorable results compared to those irrigated with the other systems, with a shorter time until the occurrence of the disease. On the other hand, the drip irrigation system notably delayed the development of the disease compared to the other systems.

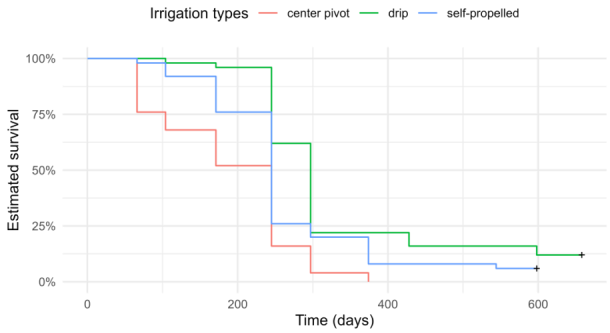


Figure 3 - Kaplan–Meier-estimated survival curves for self-propelled, drip and center pivot irrigation systems

To comprehensively assess the influence of the irrigation system on the time until the occurrence of the disease and on the incidence of the disease over time, we chose to use a joint model.

The baseline Weibull risk function was specified in the joint model to avoid underestimation of the standard errors of the model.

After fitting the submodels with longitudinal data and survival data, the following results were obtained (Table 3).

Table 3 - Estimates of maximum likelihood, standard errors and p values for the joint model with the Weibull risk function

	Estimates	Standard errors	p-values
Longitudinal process			
(Intercept)	1.3497	0.0979	<0.0001
Time	0.0114	0.0046	0.0126
Survival process			
(Intercept)	-16.5022	2.8908	<0.0001
Drip_Irrigation	-0.6248	0.2217	0.0048
CentralPivot_Irrigation	0.8952	0.2199	<0.0001
Assoct (φ)	-0.3605	0.1804	0.0457
log(shape)	1.1438	0.2078	<0.0001

For all processes involved in the joint model, an evaluation of the goodness of fit was performed, and in both cases, the results were confirmed be adequate.

According to the longitudinal results, each month, the incidence of phoma spot increased by an average of 0.01%. There was a significant difference in the risk of disease progression between the drip irrigation system and the center pivot irrigation system in relation to the self-propelled irrigation system. That is, for drip irrigation, the risk of disease occurrence was 46.5% lower than that for self-propelled irrigation. In other words, the use of drip irrigation is associated with an approximately 46.5% reduction in the risk of developing the disease compared to the use of self-propelled irrigation.

The drip or localized irrigation system provides a shorter duration of leaf wetness. In this way, a smaller number of spores germinate and the *Phoma spp* germ tube grows less, with a smaller number of infections or diseased leaves.[7]

In the area irrigated with the center pivot method, the risk of disease occurrence was approximately 2.448 times greater than that in the area irrigated with a self-propelled system.

The φ parameter was significant (Table 3); thus, there was an association between the longitudinal and survival processes. That is, changes in longitudinal variables over time are related to the risk of phoma occurrence.

A unit increase in the incidence of phoma spot was associated with a hazard ratio of 0.697. Therefore, an increase in the incidence of this disease is associated with a reduction in the risk of disease progression over time. That is, after peak or maximum disease occurrence is reached, the chance of the epidemic returning to an endemic process or the probability of incidence decreases is high.

A possible interpretation of this phenomenon is that the incidence of the disease

occurs only under certain weather conditions, namely, temperatures of 15 to 20°C and more than 4 hours of leaf wetness;[6,8] these conditions are common in winter. The water the pools on leaves may be provided by self-propelled and center pivot irrigation systems. As a result, after this period, there may fewer infections with the onset of autumn/summer, leading to a reduction in the risk of disease over time.

According to the predicted curves of survival probabilities and longitudinal trajectories, for three numbers of sampling points, namely, 50, 100 and 150, drip-irrigated Arabica coffee plants presented a higher probability of survival in relation to the progression of leaf spot than did plants irrigated with other systems.

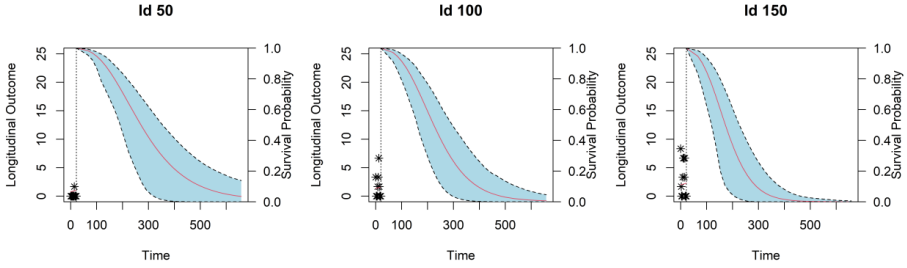


Figure 4 – Estimated survival probabilities for 50, 100 and 150 sampling points. The vertical dotted lines represent the time point of the last measurement of phoma incidence. To the left of this line is the adjusted longitudinal trajectory, and to the right, the curve represents the estimator of the probability of survival, with the respective 95% confidence intervals.

A comparison of the three curves (Figure 4) reveals that at 300 days, the estimated survival in cases with drip irrigation is approximately 50%; thus, half of the plants showed symptoms of the disease at that time. However, when considering self-propelled and center pivot irrigation systems, these estimated probabilities decrease to 30 and 10%, respectively. This implies that in 300 days, approximately 70% of the plants irrigated by self-propelled plants and 90% of the plants irrigated by center pivots will be infected. These results highlight not only the difference in the time to infection among irrigation systems but also the influence of these systems on plant health and disease resistance.

Figure 4 shows an increasing trend in the incidence of phoma spotting over time. This result highlights the importance of monitoring plants and implementing effective control measures, thus smoothing or flattening the progress curve and reducing the incidence of the disease and its harmful effects.

Based on analyses of three other georeferenced sampling points, with identification

numbers 2, 75 and 121, with drip and center pivot irrigation systems, it was evident that the predicted curves of the probabilities of survival and of the longitudinal trajectories (Figure 5) were similar to those shown above (Figure 4).

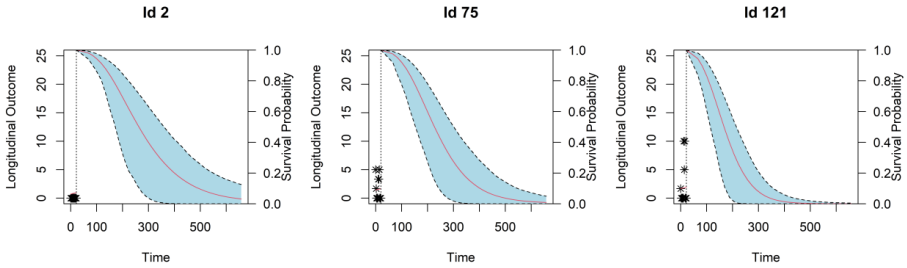


Figure 5 – Estimated survival probabilities for the plants at sampling points 2, 75 and 121. The vertical dotted line represents the time of the last measurement of phoma incidence. To the left of this line is the adjusted longitudinal trajectory, and to the right, the curve represents the probability of survival, with the respective 95% confidence intervals.

In addition, the advantage of the drip irrigation system is highlighted, as it shows more favorable results in relation to the time until the progress of the phoma stain in Arabica coffee plants than do the other irrigation methods.

Brazilian coffee producers are increasingly committed to meeting the ever-changing preferences of consumers by pursuing certifications, environmental preservation, geographical distinctions, and quality seals for their products.[3] The joint modeling of longitudinal and survival data offers several advantages and provides more comprehensive results, aiding producers in their decision-making processes.

When investigating the influence of drip irrigation management on the incidence of phoma leaf spot in coffee plants, it was concluded that plants subjected to greater water stress become more susceptible to infection. Identifying the relationships between the irrigation system and disease incidence can reduce the impact of the disease and lead to improved outcomes.[33]

4. CONCLUSIONS

The results of this study provide information on the incidence of phoma leaf spot in Arabica coffee plants, highlighting the influence of irrigation systems. There is a significant increase in the incidence of phoma over time.

There was also a significant difference among the irrigation systems in the time until disease progression in the Arabica coffee plantation, indicating that plants subjected to different irrigation systems experience changes in the time to infection. Drip irrigation stood out for providing a slower rate of disease progression, indicating its effectiveness in creating a less favorable environment for infection compared to self-propelled and center pivot systems. This result highlights the drip irrigation system's ability to reduce disease incidence, contributing to the overall health of the coffee plants.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

1. Instituto Brasileiro de Geografia e Estatística - IBGE. *Levantamento Sistemático da Produção Agrícola*: Estatística da Produção Agrícola (2024). Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/2415/epag_2024_fev. pdf [20 março 2024].

2. Araújo MDRP, Silva PL and Rocha APS. Cafeicultura: Evolução do café no Brasil, Minas Gerais e no município de João Pinheiro–MG. *Revista Contemporânea*. **3**: 21683-21706 (2023).

3. Volsi B, Telles TS, Caldarelli CE and Camara MRGD. The dynamics of coffee production in brazil. *PLoS ONE* **14**: e0219742 (2019).

4. Fernandes, A. L. T., Partelli, F. L., Bonomo, R., and Golynski, A. A moderna cafeicultura dos cerrados brasileiros. *Pesquisa Agropecuária Tropical* **42**: 231–240 (2012).

5. Cardoso HJM and Alves FD. Territorialização da mobilidade populacional entre os municípios de Carmo do Rio Claro-MG e Santa Luz-BA, in *Faces da agricultura familiar na diversidade do rural brasileiro*, ed. por Alves FD and Vale AR. Appris, Curitiba, 8–23 (2016).

6. Pozza EA. Diagnose e controle de doenças, in *Cafeicultura do Cerrado*, ed. por Carvalho RC, Ferreira AD, Andrade VT, Botelho CE and Felicori JP. EPAMIG, 347–430 (2021).

7. Ferrão RG, da Fonseca AFA, Ferrão MAG and Muner LH. *Café Conilon*. 2. ed. Vitória, ES: Incaper (2017). ISBN 978-85-89274-26-5.

8. Lorenzetti ER, Pozza EA, de Souza PE, Santos LA, Alves E, da Silva AC, Maia

FGM and Carvalho RRC. Effect of temperature and leaf wetness on Phoma tarda and Phoma leaf spot in coffee seedlings. *Coffee Science* **10**: 1–9 (2015).

9. Lima LMD, Pozza EA, Torres HN, Pozza AA, Salgado M and Pfenning LH. Relação nitrogênio/potássio com mancha de Phoma e nutrição de mudas de cafeeiro em solução nutritiva. *Tropical Plant Pathology* **35**: 223–228 (2010).

10. Deb D, Khan A and Dey N. Phoma diseases: Epidemiology and control. *Plant Pathology* **69**: 1203–1217 (2020).

11. Aveskamp M, Gruyter JD and Crous P. Biology and recent developments in the systematics of phoma, a complex genus of major quarantine significance. *Fungal diversity*, **31**:1–18 (2008).

12. Pozza E, Carvalho VD and Chalfoun S. Sintomas de injúrias causadas por doenças em cafeeiro, in *Semiologia do cafeeiro*: sintomas de desordens nutricionais, fitossanitárias e fisiológicas, ed por Guimarães RJ, Mendes ANG e Baliza DP. Editora Ufla, Lavras (2010).

13. Catarino ADM, Pozza EA, Pozza AAA, Santos LSD, Vasco GB and Souza PED. Calcium and Potassium Contents in Nutrient Solution on Phoma Leaf Spot Intensity in Coffee Seedlings. *Revista Ceres*. **63**: 486–491 (2016).

14. Pires MSO, Alves MC and Pozza EA. Multispectral radiometric characterization of coffee rust epidemic in different irrigation management systems. *International Journal of Applied Earth Observation and Geoinformation* **86**: 102016 (2020).

15. Nunes VDV. *Produtividade e incidência de doenças no cafeeiro sob diferentes lâminas de irrigação.* (Tese de Doutorado) Universidade Federal de Viçosa, Viçosa, (2006).

16. Custódio, AAP, Pozza, EA, Custódio, AAP, de Souza, PE, Lima, LA and da Silva, AM. Effect of center-pivot irrigation in the rust and brown eye spot of coffee. *Plant Disease*, **98**: 943-947 (2014).

17. Teixeira ADG. *Comportamento de cultivares de café arábica com e sem irrigação nas regiões das montanhas do estado do Espírito Santo*. (Dissertação de Mestrado) Universidade Federal do Espírito Santo, Alegre (2014).

18. Santin MR. *Caracterização agronômica de acessos de café Conilon irrigado no Cerrado do Planalto Central*. (Tese de Doutorado) Universidade de Brasília/Faculdade de Agronomia e Medicina Veterinária, Brasília (2016).

19. Meira CA, Rodrigues LH and Moraes SA. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. *Tropical Plant Pathology* **33**: 114–124 (2008).

20. Turkson AJ, Ayiah-Mensah F and Nimoh V. Handling censoring and censored data in survival analysis: a standalone systematic literature review. *International journal of mathematics and mathematical sciences*, **2021**: 9307475 (2021).

21. Long JD and Mills JA. Joint modeling of multivariate longitudinal data and survival data in several observational studies of huntington's disease. *BMC Medical Research Methodology* **18**: 138 (2018).

22. Papageorgiou G, Mauff K, Tomer A and Rizopoulos, D. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application* **6**: 223–240 (2019).

23. Guimarães PTG, Garcia AWR, Alvarez VH, Prezotti LC, Viana AS, Miguel AE, Malavolta E, Corrêa JB, Lopes AS, Nogueira FD, Monteiro AVC e Oliveira JA. Cafeeiro, in *Recomendações para o uso de corretivos e fertilizantes em Minas Gerais*: 5ª aproximação, ed. por Ribeiro AC, Guimarães PTG, Alvarez VHA. Viçosa: Comissão de Fertilidade do Solo do Estado de Minas Gerais (CFSEMG), pp. 289–302 (1999).

24. Huerta S. Par de hojas representativo del estado nutricional del cafeto. *Cenicafé Colombia* **14**: 111–127 (1963).

25. De Oliveira Aparecido LE, Rolim GS, Moraes JRSC, Costa CTS and Souza PS. Machine learning algorithms for forecasting the incidence of Coffea arabica pests and diseases. *International Journal of Biometeorology* **64**: 671-688 (2020).

26. Campbell CL and Madden LV. *Introduction to Plant Disease Epidemiology*. John Wiley & Sons (1990).

27. Wu L, Liu W, Yi GY and Huang Y. Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. *Journal of Probability and Statistics* (2012).

28. Rizopoulos D. *Joint models for longitudinal and time-to-event data:* With applications in R. New York: CRC press (2012).

29. Elashoff RM, Li G and Li N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. Biometrics, *Oxford University Press* **64**: 762–771 (2008).

30. Wulfsohn MS and Tsiatis AA. A joint model for survival and longitudinal data measured with error. *Biometrics* **53**: 330–339 (1997).

31. Mazzoleni M. Joint models for time-to-event and multivariate longitudinal data: a likelihood approach. *Statistica Applicata-Italian Journal of Applied Statistics* **2**:161–180 (2020).

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

33. Santos LSD, Pozza EA, Faria MD, Oliveira and Silva MDL, Custódio, ADP, Vasco GB e Melo e Castro BD. Incidence of the phoma leaf blight in drip irrigated coffee

trees under different irrigation managements. *Coffee Science* **9**: 77–89 (2014).

TERCEIRA PARTE

3 CONSIDERAÇÕES FINAIS

O principal ponto deste trabalho foi abordar modelos conjuntos de dados longitudinais e de sobrevivência, identificando algumas limitações e utilizando o mesmo para a análise de um conjunto de dados reais. O desenvolvimento do trabalho foi realizado em duas partes.

No primeiro momento, foi proposto uso da medida da probabilidade de cobertura cruzada, como instrumento de diagnótico da conexão dos modelos longitudinal e sobrevivência, auxiliando com a estimação de um modelo conjunto que envolva ambos os processos, podendo minimizar problemas numéricos de convergência que poderão supostamente ocorrer.

Na segunda parte, utilizou-se o modelo conjunto de dados longitudinais e de sobrevivência para compreender a influência do tipo de irrigação no tempo até a o aparecimento da mancha de phoma do cafeeiro. Com a finalidade de avaliar qual sistema de irrigação aumenta ou proporciona maior tempo até o aparecimento da mancha de phoma do cafeeiro, podendo promover a melhoria na produtividade e na qualidade da bebida, ajudar produtores de café de maneira prática, contribuir para melhorar as práticas agrícolas na lavoura cafeeira.

APÊNDICE A - Desenvolvimento das expressões algébricas

Neste apêndice, apresentamos o desenvolvimento matemático das expressões utilizadas no corpo da tese. O objetivo é fornecer uma explicação detalhada, auxiliando assim na compreensão do trabalho.

Expressão 2.3

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{S(t) - S(t+\Delta t)}{\Delta t S(t)} \\ &= \lim_{\Delta t \to 0} \frac{1 - F(t) - [1 - F(t+\Delta t)]}{\Delta t S(t)} \\ &= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{F(t+\Delta t) - F(t)}{\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Expressão 2.4

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{-f(t)}{1 - F(t)} \\ &= -\frac{d[log(1 - F(t))]}{dt} \\ &= -\frac{d}{dt}[log(1 - F(t))] \\ &= -\frac{d}{dt}[log\, S(t)]. \end{aligned}$$

Expressão 2.15

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \log(L(\theta)) \\
&= \log\left\{\prod_{i=1}^{n}[f(t_i;\boldsymbol{\theta})]^{\delta_i}[S(t_i;\boldsymbol{\theta})]^{1-\delta_i}\right\}+C \\
&= \sum_{i=1}^{n}\log\left\{[f(t_i;\boldsymbol{\theta})]^{\delta_i}[S(t_i;\boldsymbol{\theta})]^{1-\delta_i}\right\}+C \\
&= \sum_{i=1}^{n}\left\{\log[f(t_i;\boldsymbol{\theta})]^{\delta_i}+\log[S(t_i;\boldsymbol{\theta})]^{1-\delta_i}\right\}+C \\
&= \sum_{i=1}^{n}\{\delta_i\log[f(t_i;\boldsymbol{\theta})]+(1-\delta_i)[S(t_i;\boldsymbol{\theta})]\}+C.
\end{aligned}
$$

Expressão 2.21

$$
\begin{aligned}
L(\boldsymbol{\theta};\boldsymbol{t}) &\propto \prod_{i=1}^{n}[p\,(t_i \mid \boldsymbol{\theta})]^{\delta_i}[S\,(t_i \mid \boldsymbol{\theta})]^{(1-\delta_i)} \\
&\propto \prod_{i=1}^{n}\left[\frac{\beta}{\alpha^{\beta}}t_i^{\beta-1}\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right]^{\delta_i}\left[exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right]^{(1-\delta_i)} \\
&\propto \prod_{i=1}^{n}\left[\frac{\beta}{\alpha^{\beta}}t_i^{\beta-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right] \\
&\qquad \times\left[\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right]^{\delta_i} \\
&\propto \prod_{i=1}^{n}\left[\frac{\beta}{\alpha^{\beta}}t_i^{\beta-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}\right].
\end{aligned}
$$

Expressão 2.22

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \log(L(\boldsymbol{\theta})) \\
&= \log\left\{\prod_{i=1}^{n}\left[\frac{\beta}{\alpha^{\beta}}{t_i}^{\beta-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right]\right\}+C \\
&= \sum_{i=1}^{n}\log\left\{\left[\frac{\beta}{\alpha^{\beta}}{t_i}^{\beta-1}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t}{\alpha}\right)^{\beta}\right\}\right]\right\}+C \\
&= \sum_{i=1}^{n}\left\{\log\left[\frac{\beta}{\alpha^{\beta}}{t_i}^{\beta-1}\right]^{\delta_i}+\log\left[\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}\right]\right\}+C \\
&= \sum_{i=1}^{n}\left\{\delta_i\log\left[\frac{\beta}{\alpha^{\beta}}{t_i}^{\beta-1}\right]-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}+C \\
&= \sum_{i=1}^{n}\left\{\delta_i\log(\beta)-\delta_i\log(\alpha)^{\beta}+\delta_i\log(t_i)^{\beta-1}-\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}+C \\
&= \sum_{i=1}^{n}\left\{\delta_i\log(\beta)-\beta\,\delta_i\log(\alpha)+(\beta-1)\delta_i\log(t_i)-\alpha^{-\beta}{t_i}^{\beta}\right\}+C \\
&= \sum_{i=1}^{n}\left\{\delta_i\log(\beta)-\beta\,\delta_i\log(\alpha)+\beta\,\delta_i\log(t_i)-\delta_i\log(t_i)-\alpha^{-\beta}{t_i}^{\beta}\right\}+C \\
&= \log(\beta)\sum_{i=1}^{n}\delta_i-\beta\,[\log(\alpha)]\sum_{i=1}^{n}\delta_i+\beta\sum_{i=1}^{n}\delta_i\log(t_i)-\sum_{i=1}^{n}\delta_i\log(t_i)-\alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta}+C \\
&= k\,[\log(\beta)]-k\beta\,[\log(\alpha)]+\beta\sum_{i=1}^{n}\delta_i\log(t_i)-\sum_{i=1}^{n}\delta_i\log(t_i)-\alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta}+C,
\end{aligned}
$$

em que C é uma constante real que não depende de $\boldsymbol{\theta}$ e k é o número de falhas, $\sum_{i=1}^{n}\delta_i = k$.

Expressão 2.23

$$
\begin{aligned}
\frac{\partial l(\boldsymbol{\theta})}{\partial\alpha} &= -\frac{k\beta}{\alpha}+\beta\,\alpha^{-\beta-1}\sum_{i=1}^{n}{t_i}^{\beta} \\
&= -\frac{k\beta}{\alpha}+\frac{\beta\,\alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta}}{\alpha} \\
&= \frac{\beta}{\alpha}\left(-k+\alpha^{-\beta}\sum_{i=1}^{n}{t_i}^{\beta}\right).
\end{aligned}
$$

Expressão 2.25

$$
\begin{aligned}
\frac{\widehat{\beta}}{\widehat{\alpha}}\left(-k+\widehat{\alpha}^{-\widehat{\beta}}\sum_{i=1}^{n}t_i^{\widehat{\beta}}\right)&=0\\
-k+\widehat{\alpha}^{-\widehat{\beta}}\sum_{i=1}^{n}t_i^{\widehat{\beta}}&=0\\
\widehat{\alpha}^{-\widehat{\beta}}\sum_{i=1}^{n}t_i^{\widehat{\beta}}&=k\\
\widehat{\alpha}^{-\widehat{\beta}}&=\frac{k}{\sum_{i=1}^{n}t_i^{\widehat{\beta}}}\\
\widehat{\alpha}^{\widehat{\beta}}&=\frac{\sum_{i=1}^{n}t_i^{\widehat{\beta}}}{k}\\
\widehat{\alpha}&=\left(\frac{\sum_{i=1}^{n}t_i^{\widehat{\beta}}}{k}\right)^{\frac{1}{\widehat{\beta}}}.
\end{aligned}
$$

Expressão 2.26

$$
\begin{aligned}
\frac{k}{\widehat{\beta}}-k\left[\log(\widehat{\alpha})\right]+\sum_{i=1}^{n}\delta_i\log(t_i)+\widehat{\alpha}^{-\widehat{\beta}}\left[\log(\widehat{\alpha})\right]\sum_{i=1}^{n}t_i^{\widehat{\beta}}-\widehat{\alpha}^{-\widehat{\beta}}\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)&=0\\
\frac{k}{\widehat{\beta}}-k\left[\log(\widehat{\alpha})\right]+\sum_{i=1}^{n}\delta_i\log(t_i)+\frac{k}{\sum_{i=1}^{n}t_i^{\widehat{\beta}}}\log(\widehat{\alpha})\sum_{i=1}^{n}t_i^{\widehat{\beta}}-\frac{k}{\sum_{i=1}^{n}t_i^{\widehat{\beta}}}\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)&=0\\
\frac{k}{\widehat{\beta}}-k\left[\log(\widehat{\alpha})\right]+\sum_{i=1}^{n}\delta_i\log(t_i)+k\left[\log(\widehat{\alpha})\right]-k\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)&=0\\
\frac{k}{\widehat{\beta}}+\sum_{i=1}^{n}\delta_i\log(t_i)-k\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)&=0\\
k\left[\frac{1}{\widehat{\beta}}+\frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{k}-\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)\right]&=0\\
\frac{1}{\widehat{\beta}}+\frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{k}-\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)&=0\\
\frac{1}{\widehat{\beta}}=\sum_{i=1}^{n}t_i^{\widehat{\beta}}\log(t_i)-\frac{\sum_{i=1}^{n}\delta_i\log(t_i)}{k}&.
\end{aligned}
$$

Expressão 2.34

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \log(L(\boldsymbol{\theta})) \\
&= \sum_{i=1}^{n} \log p(y_i, \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \int p(y_i|b_i;\lambda,\sigma^2) p(b_i;\theta_b) db_i. \qquad (3.1)
\end{aligned}
$$

Expressão 2.35

$$
\begin{aligned}
l(\boldsymbol{\theta}) &= \log(L(\boldsymbol{\theta})) \\
&= \sum_{i=1}^{n} \log p(y_i, \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log \left\{ (2\pi)^{-\frac{n}{2}} |V_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y_i - X_i\lambda)^\top V_i^{-1}(y_i - X_i\lambda) \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V_i| - \frac{1}{2}(y_i - X_i\lambda)^\top V_i^{-1}(y_i - X_i\lambda) \right\} \\
&= -\frac{n^2}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\sum_{i=1}^{n}(y_i - X_i\lambda)^\top V_i^{-1}(y_i - X_i\lambda) \\
&= -\frac{n^2}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\sum_{i=1}^{n}\left(y_i^\top - \lambda X_i^\top\right) V_i^{-1}(y_i - X_i\lambda) \\
&= -\frac{n^2}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|V_i| - \frac{1}{2}\sum_{i=1}^{n}\left(y_i^\top V_i^{-1} - \lambda X_i^\top V_i^{-1}\right)(y_i - X_i\lambda) \\
&= -\frac{n^2}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\log|V_i| \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\left(y_i^\top V_i^{-1} y_i - y_i^\top V_i^{-1} X_i\lambda - \lambda X_i^\top V_i^{-1} y_i + \lambda X_i^\top V_i^{-1} X_i\lambda\right).
\end{aligned}
$$

Expressão 2.36

$$
\begin{aligned}
\frac{\partial}{\partial\lambda} &= -\frac{1}{2}\sum_{i=1}^{n}\left(-y_i^\top V_i^{-1}X_i - X_i^\top V_i^{-1}y_i + 2\hat{\lambda}X_i^\top V_i^{-1}X_i\right)\\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(-y_i^\top \left(V_i^\top\right)^{-1}X_i - X_i^\top V_i^{-1}y_i + 2\hat{\lambda}X_i^\top V_i^{-1}X_i\right)\\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(-y_i^\top \left(V_i^{-1}\right)^\top X_i - X_i^\top V_i^{-1}y_i + 2\hat{\lambda}X_i^\top V_i^{-1}X_i\right)\\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(-\left[\left(V_i^{-1}\right)^\top X_i\right]^\top y_i - X_i^\top V_i^{-1}y_i + 2\hat{\lambda}X_i^\top V_i^{-1}X_i\right)\\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(-X_i^\top V_i^{-1}y_i - X_i^\top V_i^{-1}y_i + 2\hat{\lambda}X_i^\top V_i^{-1}X_i\right)\\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(-2X_i^\top V_i^{-1}y_i + 2\hat{\lambda}X_i^\top V_i^{-1}X_i\right)\\
&= \sum_{i=1}^{n}\left(X_i^\top V_i^{-1}y_i - \hat{\lambda}X_i^\top V_i^{-1}X_i\right).
\end{aligned}
$$

Expressão 2.37

$$
\begin{aligned}
\frac{\partial}{\partial\hat{\lambda}} = \sum_{i=1}^{n}\left(X_i^\top V_i^{-1}y_i - \hat{\lambda}X_i^\top V_i^{-1}X_i\right) &= 0\\
\sum_{i=1}^{n}X_i^\top V_i^{-1}y_i - \sum_{i=1}^{n}\hat{\lambda}X_i^\top V_i^{-1}X_i &= 0\\
\sum_{i=1}^{n}\hat{\lambda}X_i^\top V_i^{-1}X_i &= \sum_{i=1}^{n}X_i^\top V_i^{-1}y_i\\
\hat{\lambda}\sum_{i=1}^{n}X_i^\top V_i^{-1}X_i &= \sum_{i=1}^{n}X_i^\top V_i^{-1}y_i\\
\hat{\lambda} = \left(\sum_{i=1}^{n}X_i^\top V_i^{-1}X_i\right)^{-1}\sum_{i=1}^{n}X_i^\top V_i^{-1}y_i.&
\end{aligned}
$$

Expressão 2.45

$$
\begin{aligned}
\mathcal{S}(\boldsymbol{\theta}) &= \sum_i \frac{\partial}{\partial \boldsymbol{\theta}^\top} \log \int p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \int p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i \\
&= \sum_i \frac{1}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} \int \frac{\partial}{\partial \boldsymbol{\theta}^\top} \left\{ p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \boldsymbol{\theta}) \right\} d\boldsymbol{b}_i \\
&= \sum_i \int \left[\frac{\partial}{\partial \boldsymbol{\theta}^\top} \log \left\{ p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \boldsymbol{\theta}) \right\} \right] \\
&\times \frac{p(T_i, \delta_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{y}_i | \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i; \boldsymbol{\theta})}{p(T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta})} d\boldsymbol{b}_i \\
&= \sum_i \int A(\boldsymbol{\theta}, \boldsymbol{b}_i) p(\boldsymbol{b}_i | T_i, \delta_i, \boldsymbol{y}_i; \boldsymbol{\theta}) d\boldsymbol{b}_i.
\end{aligned}
$$

Equação 2.48

$$
\begin{aligned}
\Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) &= \int \Pr(T_i^* \geq u | T_i^* > t, \mathcal{Y}_i(t), \boldsymbol{b}_i; \theta) \times p(\boldsymbol{b}_i | T_i^* > t, \mathcal{Y}_i(t); \theta) d\boldsymbol{b}_i \\
&= \int \Pr(T_i^* \geq u | T_i^* > t, \boldsymbol{b}_i; \boldsymbol{\theta}) p(\boldsymbol{b}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{b}_i \\
&= \int \frac{\mathcal{S}_i\{u | \mathcal{M}_i(u, \boldsymbol{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}}{\mathcal{S}_i\{t | \mathcal{M}_i(t, \boldsymbol{b}_i, \boldsymbol{\theta}); \boldsymbol{\theta}\}} p(\boldsymbol{b}_i | T_i^* > t, \mathcal{Y}_i(t); \boldsymbol{\theta}) d\boldsymbol{b}_i.
\end{aligned}
$$

APÊNDICE B - Códigos

Neste apêndice, estão apresentados os códigos em R utilizados neste trabalho. Estes são apenas um exemplo, portanto, os parâmetros e configurações podem ser alterados.

Artigo 1

```
#-------------------------------------------------
## Pacotes necessários
require(MASS)
require(geepack)
library(geesmv)
library(gee)
library(Matrix)
library(JM)
library(xlsx)
require(mvtnorm)

#-------------------------------------------------
## gerando censura e falha (pela weibull)
con_cens=function(n,ng,alfa,gama,p)

{   est = 0
    auxdados=matrix(0,1,3)
    for (i in 1:ng)
  {
    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)
}

#-------------------------------------------------
## gerando a resposta weibull

rweib <- function(n,gama=1,alfa=1)
{
  return((-(alfa^gama)*(log(runif(n,min=0,max=1))))^(1/gama))
}
# calular o alpha a partir de rho uniforme
Alpha <- function(t, rho, xx)
  {
   ##### Para estrutura uniforme #### #
    for(i in 1:t)
    { ### Para estrutura Exchangeable (-1 < rho < 1) ####
    alpha_exc <- round(2*rho*(t - (1 - rho^t)/(1- rho))/(t*(t-1)*(1-rho)),4)
     xx <- alpha_exc
      }
  }

# calular o alpha a partir de rho AR1
Alpha <- function(t, rho, xx)
```

```
  {##### Para estrutura AR1 (-1/(t-1) < rho < 1) #### #
    for(i in 1:t)
      {
        fx <-  (xx/(1 - xx))*(t - ((1 - xx^t)/(1 - xx))) - t*(t-1)*rho/2
        fxl_1 <- (1/((1-xx)^2)) * (t - ((1 - xx^t)/(1 -xx)))
        fxl_2 <- (xx/(1 - xx))*(((t*(xx^(t-1))*(1 - xx)) - (1 - xx^t))/
                 ((1 - xx)^2))
        fxl <- (fxl_1 + fxl_2)
        razao <- fx/fxl
        alpha.ar <- xx - razao
        xx <- alpha.ar
      }
  }


#-------------------------------------------------
## Simulando os dados

Simula <- function(rho, n, t, nsim, sh, sc)
  {    for (i in 1:nsim)
      {
        Sigma <- cormax.exch(t,rho)
        Alp <- Alpha(t, rho, 0.2)  ##calcula os valores de alpha dado rho
        t1 <- 1:t
        #x <- matrix(rbinom(t,1, 0.5), ncol=1)
        tt <- t1-1
        x <- matrix(runif(t*2), ncol=2)
        x1 <- as.matrix(data.frame(tt,x))
        aux=c(5,5)
        mu <- x1[,2:3] %*% aux
        grupo <- rmvnorm(n, mean=mu, sigma=Sigma) #indiv são as linhas
        X <- matrix(rweibull(n*t, shape = sh, scale = sc), ncol=1)
        trat <- matrix(round(runif(n*t,1,4)),n*t,1)
```

```
        y <- matrix(as.vector(t(grupo)), ncol=1)
        ind <- rep(1:n, each=t)
        tempo <- rep(tt,n)
        resp <- data.frame(y,tempo, X,trat,ind)
        #resp1 <- resp[order(resp$tempo, decreasing = FALSE), ]
        dados <- data.frame(resp)
        colnames(dados) <- c("res", "T", "X1","as.factor(trat)","ind")
      }
    return (list(dados=dados))
  }

#-------------------------------------------------
## Gerando o painel

gera_painel <- function(n,ng,alfa,gama,p,rho,Grupo,med_rep,nsim,sh,sc)

{
## Simula dados de sobrevivência
    X = con_cens(n,ng,alfa,gama,p)
    amos <- X[2:nrow(X),]

## Simula dados Longitudinais
    ch_simula <- Simula (rho,Grupo,med_rep,nsim,sh,sc)
    dadosGEE <- as.matrix(ch_simula$dados)
    aux <-cbind(dadosGEE,amos)
    colnames(aux) <- c("cov_dep","T","cov_obs","as.factor(trat)",
                       "Grupo_long","Ywei","cens","Grupo_cens")


## Geração do Painel
  painel <- as.data.frame(aux)
    colnames(painel) <- c("cov_dep","T","cov_obs","trat","Grupo_long",
                          "Ywei","cens","Grupo_cens")
```

```
    painel.ind <- painel[!duplicated(painel$Grupo_cens),]

  return(list(dados=painel,ind=painel.ind))
}

#-------------------------------------------------
## valores parametricos a serem alterados

ng=50             ## Número de individuos
sh=4              ## Modelo Weibull - sh : shape
sc=12             ## Modelo Weibull - sc : escala
p=0.00            ## porcentagem de censura
rho=0.5           ## correlação entre medidas repetidas
nsim=1000         ## número de simulações

#-------------------------------------------------
## processo de sobrevivência

alfa= rep(sc,ng) ## alfa (Weibull) para cada individuo
gama=rep(sh,ng)  ## gama (Weibull) para cada individuo
nmed <- 100      ## Número de Medições por individuo

#-------------------------------------------------
## processo longitudinal

Grupo <- ng      ## Numero de individuos
med_rep <- nmed  ## Medidas repetidas por individuo
n=rep(nmed,ng)   ## tamanho amostral

#-------------------------------------------------
## Definições  gerar o painel

mests <- matrix(0,1,4)   ## Estimativas - sobrevivência
```

```
mestl <- matrix(0,1,5)    ## Estimativas - longitudinal

SOB <- matrix(0,1,4)      ## Distribuição das estimativas - sobre.
colnames(SOB) <- c("Tr1sob","Tr2sob","Tr3sob","Tr4sob")

LONG <- matrix(0,1,5)     ## Distribuição das estimativas - long.
colnames(LONG) <- c("inter","Tr1long","Tr2long","Tr3long","Tr4long")

#---------------------------------------------------
## Inicio do procedimento para gerar resultados

cont=1                    ## Inicio da contagem

while (cont <= nsim)
  {  ch_painel <- gera_painel(n,ng,alfa,gama,p,rho,Grupo,med_rep,
                              nsim,sh,sc)
    painel <- ch_painel$dados
    painel.ind <- ch_painel$ind

#---------------------------------------------------
## Modelo longitudinal
    fitlme <- lme(cov_dep ~ as.factor(trat):cov_obs, random = ~ 1
                  | Grupo_long, data=painel)
    mestl <- (summary(fitlme))$coefficients$fixed
    mestl <- t(as.matrix(mestl))

#---------------------------------------------------
## Modelo de Sobrevivência

   chute <- c(mestl[1,1],mestl[1,3],mestl[1,4],mestl[1,5])
    fitsurv <- survreg(Surv(Ywei, cens==0) ~ -1 + as.factor(trat),
                           init=chute,control=controle,dist="weibull",
                           data=painel.ind)
```

```
    mests <- (summary(fitsurv))$coefficients
    mests <- t(as.matrix(mests))

  LONG = rbind(LONG,mestl)
  SOB <- rbind(SOB,mests)

  cont=cont + 1
}

#-------------------------------------------------
## Fim do procedimento para geração de resultados

LONG
SOB
```

Artigo 2

```
# Pacotes necessários
library(lme4)
library(survival)
library(JM)

# dados de cafeeiro

# longitudinais (incidência de phoma ao longo do tempo)
dadosLong <- read.table("PhomaLongitudinal.txt",h=T, dec = ",")
# sobrevivencia (tempo até desenvolver a mancha de phoma)
dadosSobre <- read.table("PhomaSobrevivencia.txt",h=T, dec = ",")

## Modelagem conjunta

# Submodelo longitudinal
ModeloLme <- lme(Incidencia ~ meses, dadosLong, random = ~ 1 | id)
# Submodelo de Cox
ModeloCox <- coxph(Surv(dias,censura) ~ Tratamento, data = dadosSobre,
                   x = TRUE)
# Modelo conjunto
jmfit <- jointModel(ModeloLme, ModeloCox, timeVar = "meses",
                    method = "weibull-PH-GH")
summary(jmfit)
plot(jmfit)


# Produzindo previsões de probabilidades de sobrevivência
# considerando as unidades amostrais com identificações:
# 50, 100 e 150

dataID50 <- dadosLong[dadosLong$id == 50, ]
```

```
len_id <- nrow(dataID50)

dataID100 <- dadosLong[dadosLong$id == 100, ]
len_id <- nrow(dataID100)

dataID150 <- dadosLong[dadosLong$id == 150, ]
len_id <- nrow(dataID150)

# Plotando as informações
sfit1 <- survfitJM(jmfit, newdata = dataID50)
sfit2 <- survfitJM(jmfit, newdata = dataID100)
sfit3 <- survfitJM(jmfit, newdata = dataID150)

par(mfrow=c(1,3))
plotfit1 <- plot(sfit1, estimator="mean", include.y = TRUE, conf.int=0.95,
                 fill.area=TRUE, col.area="lightblue", main="Id 50")
plotfit2 <- plot(sfit2, estimator="mean", include.y = TRUE, conf.int=0.95,
                 fill.area=TRUE, col.area="lightblue", main="Id 100")
plotfit3 <- plot(sfit3, estimator="mean", include.y = TRUE, conf.int=0.95,
                 fill.area=TRUE, col.area="lightblue", main="Id 150")

# Produzindo previsões de probabilidades de sobrevivência
# considerando as unidades amostrais com identificações:
# 2, 75 e 121

dataID2 <- dadosLong[dadosLong$id == 2, ]
len_id <- nrow(dataID50)

dataID75 <- dadosLong[dadosLong$id == 75, ]
len_id <- nrow(dataID100)

dataID121 <- dadosLong[dadosLong$id == 121, ]
len_id <- nrow(dataID150)
```

```
# Plotando as informações
sfit1 <- survfitJM(jmfit, newdata = dataID2)
sfit2 <- survfitJM(jmfit, newdata = dataID75)
sfit3 <- survfitJM(jmfit, newdata = dataID121)

par(mfrow=c(1,3))
plotfit1 <- plot(sfit1, estimator="mean", include.y = TRUE, conf.int=0.95,
                 fill.area=TRUE, col.area="lightblue", main="Id 2")
plotfit2 <- plot(sfit2, estimator="mean", include.y = TRUE, conf.int=0.95,
                 fill.area=TRUE, col.area="lightblue", main="Id 75")
plotfit3 <- plot(sfit3, estimator="mean", include.y = TRUE, conf.int=0.95,
                 fill.area=TRUE, col.area="lightblue", main="Id 121")
```