UNIVERSIDADE FEDERAL DE MINAS GERAIS
Departamento de Estatística do Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística
Programa DINTER de Doutorado entre as Instituições UFMG e UFG

Márcio Augusto Ferreira Rodrigues

MODELOS DE SOBREVIVÊNCIA PARA DADOS CORRELACIONADOS NA PRESENÇA DE RISCOS COMPETITIVOS E CENSURA INTERVALAR

Belo Horizonte
2023

Márcio Augusto Ferreira Rodrigues

MODELOS DE SOBREVIVÊNCIA PARA DADOS CORRELACIONADOS NA PRESENÇA DE RISCOS COMPETITIVOS E CENSURA INTERVALAR

Versão Final

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Estatística.

Orientador: Enrico Antônio Colosimo - DEST/UFMG
Coorientadora: Juliana Vilela Bastos - FO/UFMG

Belo Horizonte
2023

Rodrigues, Márcio Augusto Ferreira.

R696m Modelos de sobrevivência para dados correlacionados ˆ na presença de riscos competitivos e censura intervalar [recurso eletrônico]/ Márcio Augusto Ferreira Rodrigues. – 2023.
1 recurso online (171 f. il, color.) : pdf.

Orientador: Enrico Antônio Colosimo
Coorientadora: Juliana Vilela Bastos
Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 135 -142

1. Estatística – Teses. 2. Análise de sobrevivência– Teses. 3. Riscos competitivos – Teses. 4. Censura Intervalar – Teses. I. Colosimo, Enrico Antônio. II. Bastos, Juliana Vilela. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.IV.Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA

FOLHA DE APROVAÇÃO

"Modelos de sobrevivência para dados correlacionados na presença de riscos competitivos e censura intervalar"

MÁRCIO AUGUSTO FERREIRA RODRIGUES

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Doutor em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 03 de outubro de 2023, pela banca constituída pelos membros:

Documento assinado digitalmente
gov.br ENRICO ANTONIO COLOSIMO
Data: 03/10/2023 15:11:00-0300
Verifique em https://validar.iti.gov.br

Prof(a). Enrico Antonio Colosimo - Orientador
DEST/UFMG

Documento assinado digitalmente
gov.br SILVANA SCHNEIDER
Data: 03/10/2023 16:29:42-0300
Verifique em https://validar.iti.gov.br

Prof(a). Silvana Shneider
IME/UFRGS

Documento assinado digitalmente
gov.br MAGDA CARVALHO PIRES
Data: 04/10/2023 10:47:17-0300
Verifique em https://validar.iti.gov.br

Prof(a). Magda Carvalho Pires
DEST/UFMG

Documento assinado digitalmente
gov.br VERA LUCIA DAMASCENO TOMAZELLA
Data: 03/10/2023 20:35:39-0300
Verifique em https://validar.iti.gov.br

Prof(a). Vera Lúcia Damasceno Tomazella
USFCar/ DEST

Documento assinado digitalmente
gov.br FABIO NOGUEIRA DEMARQUI
Data: 04/10/2023 11:08:33-0300
Verifique em https://validar.iti.gov.br

Prof(a). Fabio Nogueira Demarqui
DEST/UFMG

Belo Horizonte, 3 de outubro de 2023.

À minha família, em especial aos meus pais, Antônia e José,
e ao meu filho, Gustavo Henrique.

Agradecimentos

Agradeço primeiramente a Deus pelo dom da vida, por toda a energia que me foi dada, durante toda a caminhada para concluir esse trabalho, e pela oportunidade de realizar esse sonho.

À minha família pelo apoio e incentivo constante em toda minha trajetória.

Aos meus orientadores, prof. Enrico A. Colosimo e profa. Juliana V. Bastos pela paciência, ensinamentos, orientação presente e por todas a contribuições realizadas durante a elaboração deste trabalho.

Aos professores membros da banca: Profa. Silvana Schneider, Profa. Magda Carvalho Pires, Profa. Vera Lúcia Damasceno Tomazella e Prof. Fabio Nogueira Demarqui por participarem da banca e pelas contribuições apresentadas.

A todos os professores do Departamento de Estatística do ICEx/UFMG pelos conhecimentos e experiências compartilhadas em disciplinas e seminários do Programa.

A todos os integrantes do Estatrauma, nosso grupo de estudos de Estatística e Trauma liderado pela profa. Juliana V. Bastos, pela parceria e pelo conhecimento em traumatismo dentário compartilhado. Em especial, agradeço a Sylvia C. Coste, pelas reuniões sob o banco de dados e pela parceria durante todo o doutorado.

As professoras Marta e Renata, do IME/UFG, pelos momentos compartilhados e por terem aceitado participar dessa jornada.

A todos os servidores do Departamento de Estatística do ICEx/UFMG, em especial à Rogéria e ao Gideão, da secretaria do Programa de Pós-graduação em Estatística, pela atenção e ajuda incondicional que sempre me foi dada.

À Universidade Federal de Goiás - UFG, em especial a todos os professores do Instituto de Matemática e Estatística, pelo afastamento concedido para a realização deste trabalho.

Por fim, a todos que, de forma direta ou indereta, contribuíram para a realização deste trabalho.

"Ensinar não é transferir conhecimento, mas criar as possibilidades para a sua própria produção ou a sua construção".

(Paulo Freire)

Resumo

Na abordagem tradicional de análise de sobrevivência considera-se uma única causa para a ocorrência do evento de interesse, no entanto existem situações em que várias causas de falha são possíveis, porém somente a ocorrência da primeira delas pode ser observada. Quando um indivíduo está sob o risco de falhar por diferentes tipos de causas, esses diferentes tipos são denominados riscos competitivos ou concorrentes. Muitos estudos clínicos envolvendo dados de riscos competitivos são frequentemente sujeitos a censura intervalar. Isso significa que o tempo de falha não é observado com precisão, mas é conhecido apenas entre dois tempos de observação, como por exemplo visitas clínicas. Em diversos estudos, dados correlacionados estão presentes e uma análise estatística apropriada exige que estas correlações sejam consideradas por meio de modelagem de fragilidade ou ajustando a correlação intra conglomerado em um modelo marginal. Todavia, estudos investigando dados correlacionados na presença de riscos competitivos e censura intervalar não foram encontrados na literatura. Nesse contexto, propomos um modelo de regressão paramétrico em que a função de incidência acumulada é modelada por meio das funções taxa de falha causa específica. Nossa segunda contribuição consiste em um modelo semiparamétrico de regressão causa-específica utilizando expansão em série de Taylor para aproximar a função taxa de falha basal. Em ambos os modelos, para acomodar a presença de conglomerados, utilizamos um modelo tipo Generalized Estimation Equation (GEE) com matriz de trabalho independente e assim, um estimador de variância sanduíche é utilizado para ajustar a correlação dentro do agrupamento. Um estudo de simulação Monte Carlo foi conduzido e indicou um bom desempenho em termos de inferência em ambos os modelos propostos neste trabalho. Uma outra análise foi conduzida para um conjunto de dados reais sobre traumatismo dentário proveniente do Programa de Traumatismos Dentários da Faculdade de Odontologia da UFMG.

Palavras-chave: Censura intervalar; Conglomerado; Riscos competitivos; Variância sanduíche.

Abstract

In the classic survival analysis approach, a single cause is considered for the occurrence of the event of interest, however, there are situations in which several causes of failure are possible, but only the occurrence of the first of them can be observed. When an individual is at risk of failing due to different types of causes, these different types are called competitive or competing risks. Many clinical studies involving competing risks data are often subject to interval censoring. This means that the failure time is not observed precisely, but is only known between two observation times, such as clinical visits. In several studies correlated data are present and appropriate statistical analysis requires that these correlations be considered through frailty modeling or by adjusting the intra-cluster correlation in a marginal model. However, studies investigating correlated data in the presence of competing risks and interval censoring were not found in the literature. In this context, we propose a parametric regression model in which the accumulated incidence function is modeled using cause-specific failure rate functions. Our second contribution consists of a semiparametric cause-specific regression model using Taylor series expansion to approximate the basal failure rate function. In both models, to accommodate the presence of clusters, we use a Generalized Estimation Equation (GEE) model with an independent work matrix and thus, a sandwich variance estimator is used to adjust the correlation within the cluster. A Monte Carlo simulation study was conducted and indicated good performance in terms of inference in both models proposed in this work. Another analysis was conducted on a set of real data on dental trauma from the Dental Trauma Program of the UFMG School of Dentistry.

Keywords: Cluster; Competing risks; Interval-censored; Sandwich variance.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

CTD-FAO-UFMG Clínica de Traumatismos Dentários da Faculdade de Odontologia da Universidade Federal de Minas Gerais

GEE Equações de Estimação Generalizada, do inglês *Generalized Estimation Equation*

FDA Função de Distribuição Acumulada

FIA Função de Incidência Acumulada

ICEx-UFMG Instituto de Ciências Exatas da Universidade Federal de Minas Gerais

ICM Algoritmo Iterativo do Minorante Convexo, do inglês *Iterative Convex Minorant*

LTDA Lesões Traumáticas dento-alveolares

NPMLE Estimador não Paramétrico de Máxima Verossimilhança, do inglês *Nonparametric Maximum Likelihood Estimator*

UFMG Universidade Federal de Minas Gerais

Sumário

Capítulo 1

Introdução

Na pesquisa clínica o tempo de um ponto inicial, como por exemplo o diagnóstico de uma doença ou início de um tratamento, até a ocorrência de um evento crítico ou benéfico, como a morte ou a cura de uma doença, é frequentemente utilizado para avaliar a eficácia de um determinado procedimento ou avaliar possíveis fatores preditivos ou prognósticos. Na maioria dos casos, os tempos dos eventos não podem ser observados para todos os indivíduos devido à perda de acompanhamento dos indivíduos em estudo ou ao tempo delimitado de acompanhamento.

Com o objetivo de obter estimativas consistentes para a distribuição do tempo do evento ou para os efeitos das covariáveis nos tempos do evento, usando toda a informação disponível no estudo, métodos que consideram essas observações incompletas, chamadas censuras, foram desenvolvidos no âmbito do que é chamado análise de sobrevivência.

Na análise de sobrevivência tradicional, uma única causa para a ocorrência do evento é considerada, de modo que a informação obtida para cada indivíduo é o tempo até a ocorrência do evento de interesse ou até a última observação em que o indivíduo foi acompanhado, que é o caso da censura. No entanto, alguns estudos podem ter interesse na ocorrência de mais de um evento, como por exemplo, em estudo oncológicos destinados a comparar diferentes estratégias de tratamento, um evento pode ser o tempo até a cura do tumor e, outro evento pode ser o tempo até a morte relacionada ao tumor. Deste modo, cada sujeito pode falhar em um dos dois tipos de eventos possíveis e os tempos para diferentes tipos de eventos podem ter a mesma importância. Esses eventos

são conhecidos como riscos competitivos ou concorrentes e nesses casos a aplicação de métodos usuais desenvolvidos para a análise de sobrevivência tradicional, com um único tipo de evento possível, pode produzir resultados errôneos (Haller et al., 2013).

No contexto de riscos competitivos, os diferentes tipos de eventos são considerados mutuamente exclusivos e apenas o tempo para a ocorrência do primeiro evento é observado. Além disso, a distribuição conjunta dos tempos para os diferentes tipos de eventos não pode ser estimada a partir dos dados observados devido a problemas de identificabilidade (Pintilie, 2006).

Na modelagem de riscos competitivos duas abordagens baseadas no risco são as mais utilizadas. A primeira é baseada na modelagem das taxas de falha causa específica para cada tipo de falha e a partir delas obtem-se a função de incidência acumulada. A segunda abordagem é baseada na modelagem da função de incidência acumulada diretamente, através da função de taxa de falha de subdistribuição.

A modelagem da função de incidência acumulada pode ser realizada parametricamente de duas maneiras. A primeira delas assume modelos paramétricos separados para as funções taxa de falha causa específica, chamada de modelagem indireta e a segunda é modelando diretamente a função de incidência acumulada. Jeong e Fine (2006) discutem essas duas abordagens e modelam diretamente a função de incidencia acumulada por meio das distribuições Weibull e Gompertz. Jeong e Fine (2007) propõem um modelo de regressão paramétrico utilizando a distribuição Gompertz.

Em muitos estudos não é possível observar o tempo exato da ocorrência do evento, e sim um intervalo em que ele ocorreu, resultando em observações com censura intervalar. Vários métodos estão disponíveis na literatura para a análise de dados desta natureza. Finkelstein (1986) propôs um modelo de riscos proporcionais para dados censurados por intervalo. Sun (1996) propôs uma estatística de teste para dados censurados por intervalo semelhante ao teste log-rank. Zhao e Sun (2004) generalizaram o teste de log-rank de Sun (1996) para incluir tempos de falha exatos em dados censurados por intervalo.

Satten (1996) considerou uma abordagem de verossimilhança marginal para ajustar o modelo de riscos proporcionais. Heller (2010) propôs um método para estimar e

fazer inferências dos parâmetros do modelo de riscos proporcionais de Cox com censura intervalar baseada em equações de estimação. Lindsey (1998), Huang e Wellner (1993), Lawless e Babineau (2006) estudaram modelos paramétricos com censura intervalar.

Os resultados mencionados anteriormente são baseados na suposição de independencia entre as observações. No entanto, dados de sobrevivência em conglomerados são encontrados em muitos estudos clínicos. Uma suposição importante de dados agrupados é que as observações de diferentes conglomerados são independentes, enquanto que as observações dentro de um conglomerado podem ser correlacionadas. No contexto de dados dependentes pode-se considerar a abordagem dos modelos marginais. Entretanto, esses modelos são utilizados quando o interesse está apenas nos coeficientes do modelo, de modo que a correlação é considerada um parâmetro de pertubação. Essa análise é baseada na abordagem de Equações de Estimação Generalizada (GEE), proposta por Liang e Zeger (1986).

Vários experimentos dão origem a dados correlacionados e censurados por intervalos. Esse tipo de dado ocorre quando existem vários tempos de sobrevivência correlacionados de interesse e apenas observações censuradas por intervalo estão disponíveis para cada tempo de sobrevivência. Vários autores propuseram estratégias de análise para esse tipo de situação. Para a análise de regressão de dados mutivariados censurados por intervalo, Goggins e Finkelstein (2000) e Kim e Xue (2002) propuseram ajustar o modelo marginal de Cox aos dados. Considerando que o modelo de Cox pode não descrever bem o comportamento dos dados Chen et al. (2007) propuseram um modelo de odds proporcionais. Bogaerts et al. (2002) apresentam uma modelagem paramétrica e utilizam uma aborgadem do tipo Generalized Estimation Equation (GEE).

Outra situação que pode ocorrer no contexto de censura intervalar é a presença de riscos competivivos. Os dados de riscos competitivos censurados por intervalo surgem quando cada sujeito em estudo pode experimentar um evento ou falha entre várias causas possíveis e o tempo de falha não é observado diretamente, mas é conhecido por estar em um intervalo entre dois tempos.

Hudgens et al. (2014) apresentam uma modelagem paramétrica da função de incidência acumulada para dados de risco competitivos sujeitos a censura intervalar. Li

(2016b) utiliza o modelo de Fine-Gray para modelar a função de incidência acumulada com dados de riscos competitivos censurados por intervalo.

No contexto de risco competitivos é, também, comum o surgimento de dados correlacionados. Vários autores desenvolveram metodologias para tratar esse cenário. Por exemplo, Zhou et al. (2012) estenderam o modelo de Fine-Gray para dados em conglomerados. Logan et al. (2010) propuseram um método para modelar diretamente a função de incidência acumulada marginal e um estimador de variância sanduíche é derivado para ajustar a correlação dentro dos conglomerados.

1.1 Motivação e Objetivos

O presente trabalho foi motivado pela linha de pesquisa "Metodologia e estatística na pesquisa em traumatismos dentários" desenvolvida por meio da parceria estabelecida, desde 2015, entre o Programa Traumatismos Dentários da Faculdade de Odontologia da UFMG e o Departamento de Estatística do ICEx-UFMG. A principal motivação deste trabalho é proveniente do estudo realizado no CTD-FAO-UFMG relativo à cicatrização pulpar de dentes portadores de lesões traumáticas dento-alveolares (LTDA). As LTDA resultam de um impacto abrupto, principalmente sobre os dentes anteriores, causando danos tais como fraturas coronárias, fraturas radiculares, deslocamentos parciais (luxações) ou totais do dente (avulsão).

A avulsão dentária consiste no deslocamento total do dente de seu alvéolo causando a completa ruptura do feixe vásculo-nervoso apical assim como das fibras do ligamento periodontal que ligam a raiz dentária ao osso alveolar. A recolocação imediata do dente no alvéolo, manobra conhecida como reimplante dentário, é o tratamento de escolha mas seu prognóstico no longo prazo apresenta grande variabilidade, pois depende de uma série de fatores relacionados com o manejo do dente avulsionado imediatamente após ao trauma, o tratamento emergencial, características do dente avulsionado e a capacidade de resposta do paciente.

A cicatrização pulpar ideal após um reimplante consiste na revascularização total do tecido pulpar, observada somente após reimplantes imediatos de dentes ainda com

desenvolvimento radicular incompleto, que evolui para a obliteração da cavidade pulpar (OCP). Outra possibilidade de cicatrizaçao consiste na invaginação de tecido ósseo para dentro da cavidade pulpar (pulp bone). Entretanto, estes eventos são raros e o que se observa mais frequentemente é a necrose do tecido pulpar com grandes prejuízos estéticos, funcionais, emocionais e financeiros para o paciente, sua família e para a sociedade como um todo.

O conhecimento sobre o prognóstico pulpar pós-trauma, suas diferentes formas de cicatrização, bem como seus fatores determinantes é de grande relevância, uma vez que podem subsidiar condutas terapêuticas que impeçam a perda precoce destes dentes, o que justifica a investigação científica de dados desta natureza.

A análise de sobrevivência se presta muito bem para a avaliação do prognóstico de dentes traumatizados no longo prazo, pois permite incorporar na análise estatística informações contidas nos dados censurados, ou seja, a observação parcial da resposta tornando-se particularmente vantajosa tendo em vista os longos períodos necessários para se atingir tamanhos de amostra razoáveis. Do ponto de vista clínico, o tempo até a observação destas respostas, bem como seus fatores determinantes, representa uma informação tão importante quanto o próprio desfecho, pois tem influência direta na tomada de decisão sobre as condutas clínicas mais adequadas. Do ponto de vista estatístico, amostras constituídas por dentes com diferentes períodos de observação podem comprometer os resultados pois as diferentes probabilidades de apresentar o evento estão relacionadas ao período de observação. Por outro lado, ao se fixar períodos mínimos de acompanhamento necessários para o diagnóstico de eventos tardios, recaímos na dificuldade de se obter amostras de tamanho adequado.

No caso especifico desse estudo, existem características intrínsecas que devem ser consideradas na análise estatística. A primeira refere-se à existência de riscos competitivos, uma vez que existem mais de um possível desfecho para a resposta pulpar, a saber: necrose, revascularização com OCP e cicatrização com pulp bone. A segunda característica é a presença de conglomerados, pois em alguns casos o mesmo paciente pode apresentar mais de um dente reimplantado. Além disso, não é possível precisar uma data exata do ocorrência do evento, e sim um intervalo de tempo entre as consultas

de acompanhamento do paciente, configurando-se desta forma a terceira característica dos dados que, são observações com censura intervalar.

Por meio da revisão da literatura, evidenciou-se que existem metodologias que consideram estas características, tratadas duas a duas, ou seja : existem metodologias propostas para tratamento de riscos competitivos e conglomerados, riscos competitivos e censura intervalar e conglomerados com censura intervalar. Entretanto, não foram encontrados modelos para o tratamento de dados que apresentam estas três características simultaneamente. Motivados pela ausência de metodologia estatística que descreva essa estrutura de dados, o objetivo principal desse trabalho é propor modelos de regressão paramétricos e semiparamétricos para riscos competitivos correlacionados na presença de censura intervalar.

1.2 Organização do trabalho

O Capítulo 2 , apresentado a seguir, irá descrever os principais conceitos em análise de sobrevivência. São discutidas características dos dados de sobrevivência e apresentadas as principais funções que representam esses dados. Além disso, é feita uma introdução aos modelos de regressão e à censura intervalar.

O Capítulo 3 é dedicado à metodologia relacionada a riscos competitivos. É feita uma breve introdução expondo o seu cenário e características, assim como uma discussão a respeito da maneira de representar, matematicamente, os riscos competitivos. São apresentadas as principais medidas utilizadas nesse contexto e por último são apresentados os modelos de regressão mais utilizados no caso de riscos competitivos.

No Capítulo 4 apresentamos o modelo paramétrico proposto. É feita a especificação do modelo, a apresentação do modelo do tipo GEE com matriz de trabalho independente, que será utilizado na parte inferencial. São apresentadas a forma do modelo levando em consideração a distribuição exponencial e Weibull e e os seus respectivos modelos de regressão. Por último, apresentamos um estudo de simulação para o modelo.

No Capítulo 5 apresentamos o modelo semeparamétrico. Um modelo de regressão causa específica é apresentado em que um modelo de riscos proporcionais de Cox é ajus-

tado para cada causa de falha sob a suposição de independência. Uma expansão em série de Taylor é utlizada para aproximar a função taxa de falha linha de base e um estimador sanduíche é utilizado para corrigir a variância dos estimadores sob a suposição incorreta de independência. Finalizando o capítulo é apresentado um estudo de simulação.

No Capítulo 6 nos atemos a aplicação dos modelos propostos em um conjunto de dados reais sobre traumatismo dentário, proveniente do Programa de Traumatismos Dentários da Faculdade de Odontologia da UFMG, para exemplificar a metodologia desenvolvida.

Por fim, apresentamos as considerações finais e direcionamentos de possíveis trabalhos futuros no Capítulo 7.

Capítulo 2

Análise de sobrevivência

A análise de sobrevivência é um conjunto de técnicas estatísticas para o estudo dos tempos de sobrevivência e dos fatores que os influenciam. Os tipos de estudos com resultados de sobrevivência incluem ensaios clínicos, estudos observacionais prospectivos e retrospectivos e experimentos com animais. Exemplos de tempos de sobrevivência incluem tempo desde o nascimento até a morte, tempo desde a entrada em um ensaio clínico até a morte ou progressão da doença, ou tempo desde o nascimento até o desenvolvimento do câncer de mama (ou seja, idade de início). Existem também aplicações em ciências sociais ou financeiras, investigando, por exemplo, o tempo que um indivíduo está desempregado ou o tempo que uma empresa fica no mercado.

Usualmente, em análise de sobrevivência, as unidades em estudo são indivíduos. O instante de ocorrência do evento de interesse é denominado tempos de sobrevivência ou "tempo de falha". Além disso, a censura é um componente frequente dos dados de sobrevivência, pois resulta da observação parcial da resposta.

Em estudos ou experimentos em que o tempo de falha é a variável resposta, é comum haver dados censurados devido a fatores que interferem na observação do evento de interesse. Essa censura pode ocorrer por inúmeras razões, tais como a saída do indivíduo do estudo por motivos não relacionados ao evento de interesse ou a não ocorrência do evento de interesse ao fim do estudo, entre outros.

A análise de sobrevivência consiste então em usar um conjunto de métodos estatísticos para analisar o tempo até a falha. O tempo de falha se caracteriza pelo tempo

de início, escala de mensuração e o evento (falha).

2.1 Conceitos Básicos

Os dados de análise de sobrevivência estão sujeitos a uma variação aleatória, e como qualquer variável aleatória, formam uma distribuição. A distribuição dos tempos de sobrevivência é usualmente caracterizada por três funções: função de sobrevivência, $S(t)$, função de densidade de probabilidade (fdp), $f(t)$, e função de taxa de falha, $\lambda(t)$.

Seja T uma variável aleatória contínua positiva que representa o tempo de sobrevivência de um indivíduo, isto é, o tempo até a ocorrência do evento, com função de distribuição acumulada (FDA), $F(t)$.

Definição 2.1.1. *A função de distribuição acumulada (FDA) é dada por :*

$$F(t) = P(T \leq t) = \int_0^t f(u)du$$

em que $f(.)$ é a função densidade de probabilidade de T.

A função de sobrevivência, $S(t)$, é definida como a probabilidade de um indivíduo sobreviver por mais do que um determinado tempo t ou por no mínimo igual a t.

Definição 2.1.2. *A função de sobrevivência é dada por*

$$S(t) = P(T > t) = 1 - F(t).$$

A função de sobrevivência é também conhecida como taxa de sobrevivência acumulada e possui as seguintes propriedades:

i) $S(t)$ é uma função monótona não crescente em t;

ii) $S(t) = 1$ para $t = 0$;

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

O tempo de sobrevivência T tem uma função de densidade de probabilidade, $f(t)$, definida como a probabilidade de um indivíduo sofrer um evento em um intervalo instantâneo de tempo.

Definição 2.1.3. *A função densidade de probabilidade é dada por*

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

em que dt é um incremento de tempo infinitamente pequeno e $f(t) \geq 0$ para todo t, e tem a área abaixo da curva igual a 1.

A função de taxa de falha, $\lambda(t)$, é uma forma de quantificar o risco instantâneo de um acontecimento ocorrer por unidade de tempo, ou seja, é uma taxa instantânea de ocorrência de um evento no tempo $T = t$ condicionada a sobrevivência até o tempo t. Por meio de $\lambda(t)$ é possível descrever a forma com que a taxa instantânea muda com o tempo, ou seja, como o risco de um indivíduo falhar no tempo $t + dt$, com $dt \to 0$, dado que ele sobreviveu ao tempo t.

Definição 2.1.4. *A função de taxa de falha é definida como*

$$\lambda(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt},$$

também denominada de função de risco, taxa de incidência, taxa de falha instantânea, taxa de intensidade, força de mortalidade ou força de mortalidade condicional.

Da definição 2.1.4 é possível obter algumas relações entre as funções de sobrevivência, de taxa de falha e densidade, como segue

$$\begin{aligned} \lambda(t) &= \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt)}{dt} \frac{1}{P(T \geq t)} \\ &= \lim_{dt \to 0^+} \frac{F(t + dt) - F(t)}{dt} \frac{1}{S(t)} = \frac{f(t)}{S(t)}, \end{aligned} \tag{2.1}$$

$$\lambda(t) = -\frac{d}{dt} \ln S(t), \tag{2.2}$$

$$S(t) = \exp\left\{ -\int_0^t \lambda(u) du \right\}, \tag{2.3}$$

$$f(t) = \lambda(t) \exp\left\{ -\int_0^t \lambda(u) du \right\}. \tag{2.4}$$

Outra função importante em análise de sobrevivência é a função de taxa de falha acumulada, pois fornece a taxa de falha acumulada do indivíduo.

Definição 2.1.5. *A função de taxa de falha acumulada é definda por*

$$\Lambda(t) = \int_0^t \lambda(u)du,$$

que é finita para algum $t > 0$ *e* $\Lambda(t) = \int_0^\infty \lambda(u)du = \infty$.

Segundo Colosimo e Giolo (2006) esta função não tem interpretação direta, no entanto é bastante útil para avaliar a função de taxa de falha, principalmente no processo de estimação não paramétrica em que a função de taxa de falha acumulada apresenta propriedades ótimas e a função de taxa de falha é difícil de ser estimada.

Os dados de sobrevivência comportam-se de maneira diferente, o que gera problemas para a sua análise. Uma característica é a presença de censura, ou seja, apresentam observações incompletas, que por algum motivo não foi possível observar a ocorrência do evento de interesse. Em geral, a censura ocorre quando o indivíduo no estudo não sofre o evento de interesse durante o período que se encontra sob observação.

Para Colosimo e Giolo (2006) duas razões justificam considerar a censura na modelagem: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de sobrevivência; (ii) a omissão das censuras no cálculo das estimativas de interesse podem acarretar conclusões viciadas.

Existem três principais tipos de censuras, a saber:

i) Censura tipo I - Considere um estudo com n indivíduos, cada um com tempo de sobrevivência T_i associado. Seja T o tempo de seguimento pré-determinado do estudo. Se $T_i \leq T$, então o indivíduo sofreu o evento, caso contrário ($T_i > T$), significa que ocorreu censura do tipo I. T é chamado tempo de censura, que nesse caso é um valor fixo (não aleatório).

ii) Censura tipo II - A censura do tipo II ocorre quando em um estudo, um conjunto de n indivíduos ou objetos é observado até que um número pré-determinado de r falhas ocorra. Os $n - r$ indivíduos são censuras do tipo II. Nesse tipo de estudo tem-se o tempo para o término do estudo aleatório e o número de falhas fixo. Observa-se que a censura tipo II difere da censura tipo I em que o tempo para o término do estudo é fixo e o número de falhas que é aleatório.

iii) Censura aleatória - Na maioria dos estudos clínicos e epidemiológicos o período de estudo é fixo e os pacientes entram no estudo em momentos diferentes durante esse período. As censuras aleatórias podem ocorrer de uma das seguintes maneiras:

   a) *Término do estudo* (censura administrativa): o indivíduo é acompanhado até o final do estudo sem ter experimentado o evento de interesse;

   b) *Perda de observação*: o indivíduo abandona o estudo devido a fatores externo;

   c) *Retirado do estudo*: o indivíduo não se adapta ao tratamento e é preciso retirá-lo do estudo devido a, por exemplo, efeitos colaterais;

   d) *Risco competitivo*: o indivíduo falha por outra causa não relacionada ao evento de interesse em estudo.

Em todas as situações anteriores, apenas informações parciais são disponíveis. No primeiro caso, por exemplo, sabe-se apenas que o tempo entre o início do estudo e o evento de interesse é maior do que o tempo de fato observado.

Todos os tipos de censuras abordados anteriores são designadas por censura à direita. Existem também censura à esquerda e censura intervalar. A censura à esquerda ocorre quando o tempo registrado é maior do que o tempo de falha, isto é, o evento de interesse já aconteceu quando o indivíduo foi observado. A censura intervalar ocorre quando o tempo de falha não é conhecido exatamente, mas sabe-se que pertence a um intervalo. Como censura intervalar é um tópico deste trabalho, vamos explorá-la na Seção 2.4 deste capítulo.

Um pressuposto importante associado a vários métodos de análise de sobrevivência é que o tempo de sobrevivência T é independente do tempo de censura C, isto é, os dados censurados não fornecem nenhuma informação com respeito ao tempo até o evento principal.

No entanto, Putter et al. (2007) adverte que deve-se tomar cuidado em relação à suposição da independência entre T e C, pois a depender da causa da censura essa suposição pode ser violada. Por exemplo, se a censura é causada pelo término do estudo, pode-se seguramente assumir que o tempo de sobrevivência é independente do mecanismo

de censura, uma vez que o que ocorre após o final do estudo não é do conhecimento ou interesse do pesquisador. Esta censura é denominada administrativa.

2.2 Função de Verossimilhança e Modelos Paramétricos

Seja T a variável aleatória representando o tempo de sobrevivência e seja C a variável aleatória representando o tempo de censura à direita e considere que T e C são independentes. Defina

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

e seja $\delta = I(T \leq C)$ uma variável aleatória indicadora do *status*, que assume valor 1 se o tempo de sobrevivência foi observado e 0 se o indivíduo foi censurado, isto é

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

Note que, a variável aleatória Y tende a ser "menor" do que o tempo de falha de interesse, T. Isso fica claro ao observar que $Y = min\{T, C\}$. Partindo do pressuposto de que T e C são independentes, a função de sobrevivência de Y é

$$\begin{aligned} S_Y(y) &= P(T > y, C > y) = P(T > y)P(C > Y) \\ &= S_T(y)S_C(y) \leq S_T(y). \end{aligned}$$

Observe que na prática, para cada indivíduo, obtem-se um par (Y, δ), assim, considere $(Y_1, \delta_1), (Y_2, \delta_2), ..., (Y_n, \delta_n)$ cópias independentes e identicamente distribuidas (iid) de (Y, δ).

Para dados censurados à direita, a contribuição de cada elemento da amostra para a construção da função de verossimilhança é dada por :

$$\begin{cases} f(y_i; \boldsymbol{\theta}), & \text{se o } i\text{-ésimo tempo de sobrevivência for observado} \\ S(y_i; \boldsymbol{\theta}), & \text{se o } i\text{-ésimo tempo de sobrevivência for censurado} \end{cases}$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros de interesse associado à distribuição do tempo de falha T e $i = 1, ..., n$. Portanto, a contribuição de cada observação não censurada na

função de verossimilhança, $L(\boldsymbol{\theta})$, é dada pela função de densidade e como as observações censuradas informam apenas que o tempo até a ocorrência do evento de interesse é maior que o tempo de censura observado sua contribuição é dada pela função de sobrevivência.

A função de verossimilhança $L(\boldsymbol{\theta})$ é dada por :

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} [f(y_i;\boldsymbol{\theta})]^{\delta_i} [S(y_i;\boldsymbol{\theta})]^{1-\delta_i} \\ &= \prod_{i=1}^{n} [\lambda(y_i;\boldsymbol{\theta})]^{\delta_i} S(y_i;\boldsymbol{\theta}) \end{aligned} \tag{2.5}$$

Os estimadores de máxima verossimilhança são os valores de $\boldsymbol{\theta}$ que maximizam $L(\boldsymbol{\theta})$ ou, equivalentemente, o logaritmo de $L(\boldsymbol{\theta})$, ou seja, $\log L(\boldsymbol{\theta})$. Eles são obtidos resolvendo o sistema de equações,

$$\mathbf{U}(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}.$$

Uma propriedade importante do estimador de máxima verossimilhança, $\hat{\boldsymbol{\theta}}$, é em relação a sua distribuição assintótica. Para grandes amostras, sob certas condições de regularidade, a distribuição de $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ...\hat{\theta}_k)$ é descrita, assintoticamente, por uma distribuição normal multivariada com vetor de médias $\boldsymbol{\theta}$ e matriz de variâncias e covariâncias $Var(\hat{\boldsymbol{\theta}}) \approx \mathbf{I}^{-1}(\boldsymbol{\theta})$, ou seja

$$\hat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, \mathbf{I}^{-1}(\boldsymbol{\theta})),$$

em que $\mathbf{I}(\boldsymbol{\theta}) = -\dfrac{\partial^2 \log L(\boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta}}$ é a matriz de informação observada e k é a dimensão de $\hat{\boldsymbol{\theta}}$.

Esta propriedade é utilizada para construir testes de hipóteses para $\boldsymbol{\theta}$. Considere o interesse em testar as seguintes hipóteses:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \ \textit{verus} \ H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

em que $\boldsymbol{\theta}_0$ é conhecido. Temos:

Teste de Wald

A estatística de teste é dada por,

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

que, sob H_0 verdadeira, tem aproximadamente uma distribuição qui-quadrado com k grau de liberdade. O valor de W é então comparado ao valor encontrado na tabela da distribuição de probabilidade qui-quadrado com k graus de liberdade. Dado um nível de significância α, se $P(\chi^2 > W) \leq \alpha$ rejeita-se H_0, ou seja, valores altos de W indicam rejeição da hipótese nula.

Sendo θ escalar, a estatística do teste reduz a

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\widehat{Var}(\hat{\theta})},$$

em que $\widehat{Var}(\hat{\theta}) = I(\hat{\theta})^{-1}$.

Teste da Razão de Verossimilhanças

Seja $\log L(\hat{\boldsymbol{\theta}})$ o logaritmo da função de verossimilhança maximizada sem restrição e $\log L(\boldsymbol{\theta}_0)$ o logaritmo da função de verossimilhança maximizada sob H_0. A estatística para esse teste é

$$TRV = -2 \log \left[\frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})} \right],$$

que, sob H_0 verdadeira, é descrita aproximadamente por uma distribuição qui-quadrado com k graus de liberdade. A hipótese nula é rejeitada para valores altos de TRV.

2.2.1 Distribuição do tempo de sobrevivência

Nesta seção, algumas distribuições que são comumente consideradas para modelos paramétricos de sobrevivência são apresentadas resumidamente. Uma abordagem detalhada pode ser encontrada em Colosimo e Giolo (2006).

Distribuição exponencial

A distribuição exponencial é a distribuição de sobrevivência uniparamétrica mais simples. Ela possui uma característica única, a taxa de falha dessa distribuição não depende do tempo, isto é, a taxa de falha é constante, conforme ilustrado na Figura 2.1. Assim, a função de taxa de falha pode ser escrita como,

$$\lambda(t) = \alpha, \quad 0 \leq t < \infty.$$

A função de sobrevivência é

$$S(t) = \exp(-\alpha t),$$

e com isso, a função de desidade de probabilidade é dada por,

$$f(t) = \alpha \exp(-\alpha t), \quad 0 \leq t < \infty.$$

O parâmetro α é uma constante positiva que pode ser estimado ajustando o modelo de sobrevivência aos dados observados.

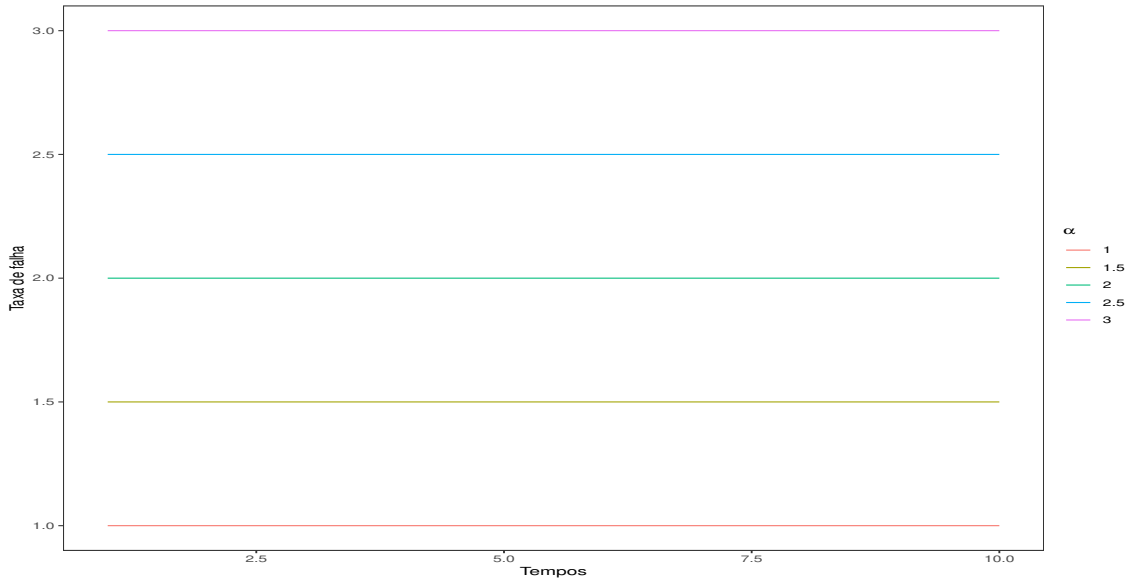


Figura 2.1 – Forma típica da função taxa de falha da distribuição exponencial para diferentes valores de α.

A distribuição exponencial é muito usada também em confiabilidade para descrever o tempo de vida de componentes ou sistemas. No entanto, na área médica, dificilmente uma doença não altera a taxa de falha ao longo do tempo, entretanto, em situações em que o estudo tem um acompanhamento muito curto é possível considerar que o risco do evento de interesse ocorrer é constante.

Distribuição Weibull

A distribuição Weibull foi proposta originalmente por Waloddi Weibull em 1951. Entre todas as famílias de distribuições paramétricas, o modelo Weibull é talvez o mo-

delo paramétrico mais amplamente aplicado em análise de sobrevivência devido à sua simplicidade e flexibilidade. Além disso, a distribuição Weibull pode ser formulada de várias maneiras, no entanto, duas abordagens são mais utilizadas: o modelo de taxa de risco proporcional e o modelo de tempo de vida acelerado.

A distribuição de probabilidade Weibull do tempo do evento T, uma função contínua, é caracterizada pelo uso de dois parâmetros: um parâmetro de escala, α, e outro de forma, τ. Com esses dois parâmetros, a distribuição Weibull, comparada com a distribuição exponencial, fornece um modelo paramétrico com maior flexibilidade. Esta é uma extensão da distribuição exponencial, que permite uma dependência da taxa de falha no tempo. Diferentes parametrizações da distribuição Weibull existem na literatura. Uma formulação possível é a que possui função de densidade dada por

$$f(t) = \tau\alpha^{\tau}t^{\tau-1}\exp(-(\alpha t)^{\tau}), \quad 0 \leq t < \infty. \tag{2.6}$$

A partir de (2.6) pode-se identificar facilmente que, quando quando $\tau = 1$, a função de densidade da Weibull se reduz a função de densidade exponencial. Portanto, a distribuição exponencial é um caso particular da distribuição Weibull quando $\tau = 1$.

Dadas as relações entre funções de sobrevivência, função de densidade e função taxa de falha, apresentadas na Seção 2.1, a função de sobrevivência para essa parametrização é derivada da Equação (2.6),

$$S(t) = \exp(-(\alpha t)^{\tau}), \quad 0 \leq t < \infty, \tag{2.7}$$

e a função de taxa de falha pode ser denotada por,

$$\lambda(t) = \tau\alpha^{\tau}t^{\tau-1}, \quad 0 \leq t < \infty, \tag{2.8}$$

em que α e τ são parâmetros positivos de escala e de forma, respectivamente.

A distribuição Weibull possui função taxa de falha que pode acomodar diferentes formas. Quando $\tau = 1$ a taxa de falha é constante; quando $\tau > 1$, a taxa de falha aumenta e tem forma não linear, para $\tau \neq 2$, ao longo de t; e, quando $\tau < 1$, a taxa de falha diminui de forma não linear com t. Portanto, quando $\tau \neq 1$, a taxa de falha é monotonicamente crescente ou monotonicamente decrescente ao longo do tempo. Dada

essa característica, uma das propriedades fundamentais para a distribuição Weibull é a monotonicidade. Essa propriedade é ilustrada na Figura 2.2, a seguir.

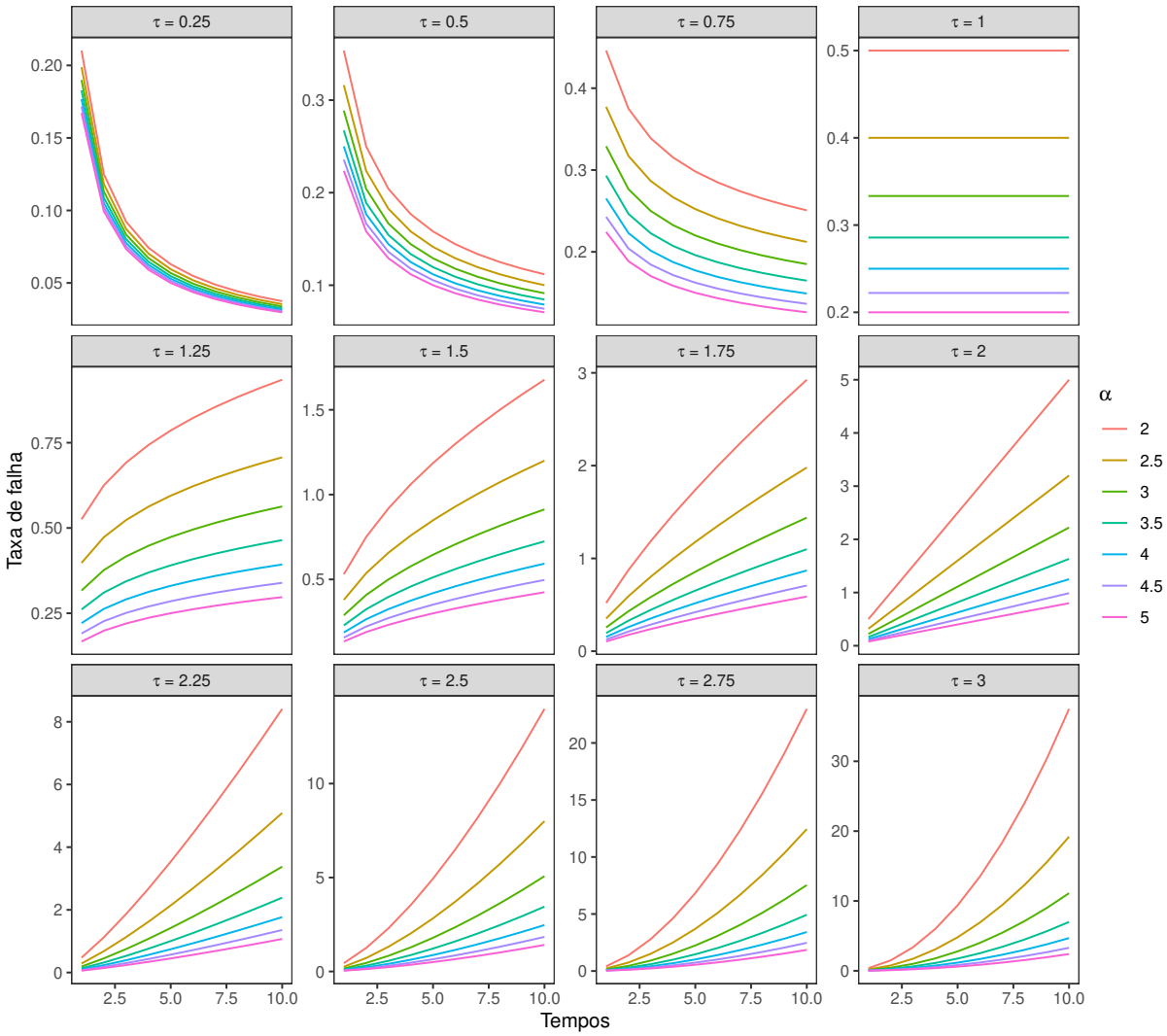


Figura 2.2 – Forma típica da função taxa de falha da distribuição Weibull para diferentes valores de α e τ.

Distribuição Gompertz

A distribuição Gompertz é uma distribuição de probabilidade contínua, com suporte positivo. Essa distribuição possui dois parâmetros, $\alpha > 0$ e $\tau \in \mathbb{R}$, em que α é o parâmetro de escala e τ é o parâmetro de forma. Ela foi proposta inicialmente em 1825 por Benjamin Gompertz, com o propósito de descrever a mortalidade humana. A

distribuição Gompertz é muito utilizada em estudos nas áreas de demografia, atuária e em estudos da área de sobrevivência.

A função de densidade de probabilidade para a variável aleatória tempo até o evento de interesse T, com distribuição Gompertz, é dada por

$$f(t) = \alpha e^{\tau t} \exp\left\{\frac{\alpha}{\tau}(1 - e^{\tau t})\right\}, \quad 0 \leq t < \infty. \tag{2.9}$$

Novamente, utilizando as relações entre funções de sobrevivência, função de densidade e função taxa de falha, apresentadas na Seção 2.1, a função de sobrevivência para essa parametrização é derivada da Equação (2.9)

$$S(t) = \exp\left\{\frac{\alpha}{\tau}(1 - e^{\tau t})\right\}, \quad 0 \leq t < \infty, \tag{2.10}$$

e a função de taxa de falha pode ser denotada por,

$$\lambda(t) = \alpha e^{\tau t}, \quad 0 \leq t < \infty. \tag{2.11}$$

Na Figura 2.3, a seguir, tem-se o comportamento da função taxa de falha (2.11) para diferentes valores de α e τ. A Equação (2.11) e a Figura 2.3, tanto analítica quanto graficamente, demonstram que a função de taxa de falha da Gompertz tem uma forma exponencial. Quando $\tau > 0$ a taxa de falha é crescente, quando $\tau < 0$ a taxa de falha é decrescente e quando $\tau = 0$ a taxa de falha é constante, reduzindo-se a distribuição exponencial.

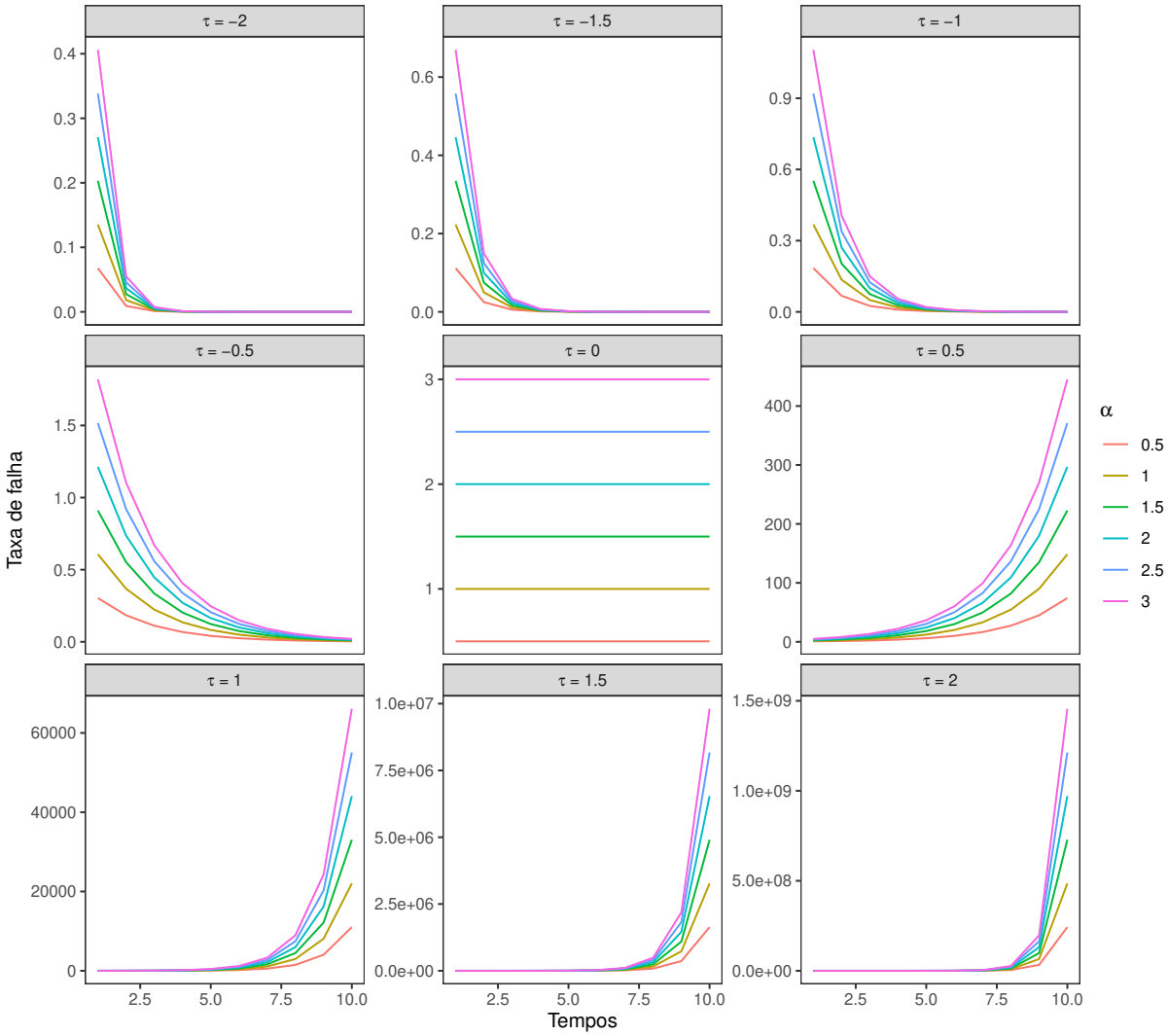


Figura 2.3 – Forma típica da função taxa de falha da distribuição Gompertz para diferentes valores de α e τ.

Outras ditribuições

Diversas distribuições de probabilidade podem ser utilizadas para modelar o tempo de sobrevivência T. Dentre elas, temos a gama, gama generalizada, lognormal, log-logística e normal inversa (Collett, 2015; Colosimo e Giolo, 2006; Cai e Prentice, 1997; Kalbfleisch e Prentice, 2011; Lawless, 2011).

2.3 Modelos de Regressão

Modelos de regressão são usados para modelar a dependência de covariáveis e estimar o impacto que elas podem ter na distribuição de sobrevivência. Para investigar a influência das covariáveis nos tempos dos eventos foram desenvolvidos modelos de regressão que podem ser aplicados na ocorrência de observações censuradas.

2.3.1 Modelo de Regressão de Cox

O modelo de regressão mais popular para dados de sobrevivência é o modelo de taxas de falhas proporcionais introduzido por Cox (1972) . Nesse modelo supõe-se que a razão das funções de taxa de falha são constantes ao longo do tempo e que cada uma das p covariáveis em consideração tem um efeito linear no logaritmo da taxa de falha, dadas as outras covariáveis. O modelo de Cox pode ser escrito como

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}), \tag{2.12}$$

com a função de taxa de falha basal $\lambda_0(t)$ não especificada para um indivíduo com vetor de covariáveis de zeros, o vetor p-dimensional de covariáveis $\mathbf{x}$ e o vetor de coeficientes da regressão $\boldsymbol{\beta}$. Os coeficientes $\boldsymbol{\beta}$ são parâmetros que quantificam o efeito de suas respectivas covariáveis no modelo de Cox, e $\exp(\beta)$ pode ser interpretada como a razão das funções de taxa de falha para a covariável x.

A razão das funções de taxa de falha entre dois indivíduos i e j pode ser calculada como

$$\frac{\lambda(t|\mathbf{x}_j)}{\lambda(t|\mathbf{x}_i)} = \frac{\lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_j)}{\lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_i)} = \exp(\boldsymbol{\beta}^T(\mathbf{x}_j - \mathbf{x}_i)). \tag{2.13}$$

Assim, a razão das funções taxa de falha entre dois indivíduos diferentes é constante no tempo, isto é, não depende do tempo. Devido a essa característica, o modelo de Cox é também chamado modelo de taxas de falhas proporcionais.

O coeficiente de regressão para a q-ésima covariável β_q pode ser interpretado como o logaritmo da razão das funções de taxa de falha entre dois indivíduos, diferindo em uma unidade na covariável x_q, e tendo valores iguais para todas as outras covariáveis

x_p, com $q < p$, isto é,

$$\beta_q = \log\left(\frac{\lambda(t|x_1, x_2, ..., x_{q-1}, x_q + 1, x_{q+1}, ..., x_p)}{\lambda(t|x_1, x_2, ..., x_{q-1}, x_q, x_{q+1}, ..., x_p)}\right). \tag{2.14}$$

Assumindo $k \leq n$ tempos de falha ordenados distintos, $(t_1, t_2, ..., t_k)$ uma estimativa para o vetor de coeficientes da regressão é obtida por maximização numérica da verossimilhança parcial introduzida por Cox, tratando a função de taxa de falha basal não especificado $\lambda_0(t)$ como um parâmetro de pertubação. Deste modo,

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{(i)})}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}^T \mathbf{x}_j)} \tag{2.15}$$

em que $\mathbf{x}_{(i)}$ é o vetor de covariáveis do indivíduo que falhou no instante t_i, e $R(t_i)$ representa o conjunto de indivíduos em risco no instante t_i, ou seja, todos os sujeitos que não falharam antes de t_i e ainda estão sob observação em t_i.

Usualmente um algorítimo do tipo Newton-Raphson é utilizado para encontrar o vetor de coeficientes da regressão, que maximiza a log-verossimilhança parcial

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \{\log(L(\boldsymbol{\beta}))\}.$$

Seja $\mathbf{I}(\boldsymbol{\beta})$ a matriz de informação observada, em que os elementos são dados por

$$-\frac{\partial^2 \log L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}.$$

O estimador, aproximado, da matriz de covariância do estimador de máxima verossimilhança é obtido pelo inverso da matriz de informação observada avaliada no estimador de máxima verossimilhança, $\hat{\boldsymbol{\beta}}$,

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1}.$$

O uso preciso do modelo de Cox requer que várias suposições sejam atendidas. Devemos ter mecanismo de censura não informativa, riscos proporcionais e linearidade nas covariáveis contínuas (Therneau e Grambsch, 2000). Uma descrição mais detalhada do modelo de regressão de Cox, incluindo adequação do modelo, testes estatísticos para os coeficientes da regressão, e extensões do modelo podem ser encontrados em vários livros (Colosimo e Giolo, 2006; Therneau e Grambsch, 2000; Kalbfleisch e Prentice, 2011; Collett, 2015).

2.3.2 Modelos de Regressão Paramétricos

Como alternativa ao modelo de regressão de Cox, modelos paramétricos de sobrevivência, assumindo que os tempos dos eventos tem uma distribuição pré-especificada e usando uma função de ligação adequada entre o preditor linear $\boldsymbol{\beta}^T\mathbf{x}$ e os parâmetros da distribuição do tempo do evento, podem ser usados.

Suponha que os dados sejam considerados como n pares de observações, em que o par para o i-ésimo indivíduo é $(t_i, \delta_i), i = 1, 2, ..., n$ e $\mathbf{x}_i$ o vetor de covariáveis. A função de verossimilhança pode então ser escrita como

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i|\mathbf{x}_i)\right]^{\delta_i} \left[S(t_i|\mathbf{x}_i)\right]^{1-\delta_i}, \tag{2.16}$$

em que $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\beta})$, sendo $\boldsymbol{\psi}$ o vetor de parâmetros do modelo paramétrico e $\boldsymbol{\beta}$ o vetor de parâmetros que quantificam o efeito de suas respectivas covariáveis no modelo.

Alternativamente,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[\frac{f(t_i|\mathbf{x}_i)}{S(t_i|\mathbf{x}_i)}\right]^{\delta_i} S(t_i|\mathbf{x}_i) = \prod_{i=1}^{n} \left[\lambda(t_i|\mathbf{x}_i)\right]^{\delta_i} S(t_i|\mathbf{x}_i). \tag{2.17}$$

Essa versão da função de verossimilhança é particularmente útil quando a função de densidade de probabilidade tem uma forma complicada, como costuma acontecer. Consequentimente, a função log-verossimilhança é

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left\{\delta_i \log[\lambda(t_i|\mathbf{x}_i)] + \log[S(t_i|\mathbf{x}_i]\right\}. \tag{2.18}$$

Para obter os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ é preciso substituir as funções de taxa de falha e sobrevivência por aqueles da distribuição em questão. As estimativas para os parâmetros são obtidas maximizando a função de verossimilhança ou a função log-verossimilhança, analítica ou numericamente. As variâncias das estimativas de máxima verossimilhança podem ser obtidas pelo inverso da matriz de informação observada.

Detalhes sobre os modelos de regressão paramétricos podem ser encontrados em Colosimo e Giolo (2006), Collett (2015), Lawless (2011) e Kalbfleisch e Prentice (2011).

2.4 Censura Intervalar

Uma razão, que também é uma característica importante que distingue a análise de dados de tempo de falha de outros campos estatísticos, é a existência de censura, como a censura intervalar. O tempo exato de ocorrência de um evento não é observado, mas o que se sabe é o intervalo em que o evento ocorre. Esses tipos de dados geralmente ocorrem em estudos biológicos, demográficos, econômicos e financeiros, epidemiológicos, experimentos psicológicos, confiabilidade e sociológicos. Por exemplo, em estudos médicos, os indivíduos são frequentemente acompanhados por exames periódicos e nesta situação, algumas observações podem ter seu estado alterado entre um exame e outro. Consequentemente, sabe-se apenas que o tempo real da ocorrência do evento é maior que o último tempo de observação no qual a mudança não ocorreu e menor ou igual ao primeiro tempo de observação em que a mudança foi observada. Assim, a ocorrência de um evento de interesse é observada apenas entre dois tempos de exame consecutivos, em vez da data exata.

Seja T uma variável aleatória não negativa representando o tempo de falha de um indivíduo em um estudo. Uma observação de T é censurada por intervalo se em vez de observar T exatamente, observa-se um intervalo $(L, R]$ tal que

$$T \in (L, R], \tag{2.19}$$

em que $L \leq R$, (Sun, 2006).

Os principais tipos de censura intervalar que frequentimente ocorrem na prática, conforme discutido por Sun (2006), Bogaerts et al. (2017), Gómez et al. (2009), são:

i) *Dados censurados por intervalo caso I ou dados de estado atual*: Nessa situação a variável aleatória T é conhecida apenas por ser maior ou menor que um tempo de monitoramento observado, ϵ. Nesse caso, o sujeito em estudo é observado apenas uma vez, assim todas as observações são da forma $(0, \epsilon]$ ou $[\epsilon, \infty]$. Ou seja, só é registrado se o evento já aconteceu ou ainda tem que acontecer no tempo ϵ e, portanto, todas as observações são censuradas à esquerda ou à direita.

ii) *Dados censurados por intervalo caso II*: Nessa situação, cada sujeito é observado

duas vezes, com dois tempos de observação, U e V, em que U e V são duas variáveis aleatórias tal que $U \leq V$ com probabilidade 1. Sabe-se assim, que o evento de interesse ocorreu antes do primeiro instante de monitoramento, $T \leq U$, entre os dois tempos de monitoramento, $U < T \leq V$ ou após o segundo tempo de monitoramento, $T > V$.

iii) *Dados censurados por intervalo caso K*: Em estudos longitudinais com acompanhamento periódico e tempos de monitoramento $U_1 \leq U_2 \leq ... \leq U_K$, o evento de interesse é observado apenas entre dois tempos de inspeção consecutivos $(U_i, U_{i+1}]$. Esse esquema de censura corresponde a uma extensão natural dos mecanismos do caso I e do caso II.

Em análise de sobrevivência, uma suposição básica e importante comumente usada é que o mecanismo de censura é independente ou não informativo sobre o tempo de falha de interesse. Para dados censurados por intervalo, caso II e caso K, a censura por intervalo independente significa que a distribuição conjunta de U e V ou os U_j's não contém parâmetros envolvidos na função de sobrevivência de T. A censura de intervalo independente assume que o intervalo $(L, R]$ não fornece nenhuma informação além do fato que T é simplesmente delimitado pelos dois valores observados, isto é,

$$P(T \leq t | L = l, R = r, L < T \leq R) = P(T \leq t | l < T \leq r) \tag{2.20}$$

e a distribuição conjunta de L e R é livre dos parâmetros envolvidos na função de sobrevivência de T.

2.4.1 Modelagem Paramétrica

Em algumas situações, a experiência do pesquisador pode sugerir uma distribuição particular para o tempo de sobrevivência T, nesse caso pode-se optar pela modelagem paramétrica. Especificando corretamente a distribuição de T, a função de sobrevivência estimada parametricamente fornecerá inferência mais precisa do que o estimador não paramétrico de máxima verossimilhança (NPMLE). No caso de dados censurados à direita, o NPMLE da função de sobrevivência é dado pelo estimador de Kaplan-Meier

(Kaplan e Meier, 1958). No caso de dados com censura intervalar, um dos métodos mais populares para obter um estimador não paramétrico de máxima verossimilhança (NPMLE) para a função de sobrevivência é o chamado algoritmo de Turnbull (Turnbull, 1976), que ao contrário do estimador de Kaplan-Meier, não tem uma forma analítica fechada e para ser obtido é preciso o uso de algoritimo iterativo (Colosimo e Giolo, 2006).

Modelos paramétricos, como os tratados no caso de censura a direita, são também de interesse na análise de dados de sobrevivência intervalar. A natureza intervalar dos dados deve ser levada em consideração na construção da função de verossimilhança. Cada indivíduo contribui para a função de verossimilhança com uma informação específica: um indivíduo que apresente um tempo exato de falha contribui com a probabilidade de ocorrência do evento nesse tempo, $f(t)$; se o indivíduo foi censurado à direita a contribuição é dada pela função de sobrevivência de T avaliada no último tempo observado; por outro lado, a contribuição de um indivíduo censurado à esquerda é dada pela função de distribuição acumulada (FDA) de T avaliada no tempo da primeira observação; enquanto que a contribuição de um individuo censurado em um intervalo é dada pela probabilidade de que o tempo de ocorrência do evento pertença a este intervalo, $[S(l) - S(r)]$.

Assim, considere um estudo com n indivíduos independentes e que as n observações consiste de tempo exato de falha, observações censuradas a esquerda, a direira e por intervalo. Seja T_i o tempo de sobrevivência de interesse para o i-ésimo indivíduo, $i = 1, ..., n$ e suponha que T_i seja descrito por um modelo paramétrico, temos

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ f(t_i|\mathbf{x}_i)^{\delta_1} S(t_i|\mathbf{x}_i)^{\delta_2} (1 - S(t_i|\mathbf{x}_i))^{\delta_3} [S(l_i|\mathbf{x}_i) - S(r_i|\mathbf{x}_i)]^{\delta_4} \right\} \tag{2.21}$$

em que $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ é o vetor de parâmetros descohecido do modelo, $\mathbf{x}_i$ é vetor de covariáveis para o individuo i, $(l_i, r_i]$ é o intervalo observado que contém o T_i e δ_j, $j = 1, ..., 4$, é a variável indicadora de tempo exato, censura à direita, censura à esquerda e censura intervalar, respectivamente. A função de verossimilhança fica determinada após a especificação do modelo paramétrico.

O estimador de máxima verossimilhança de $\boldsymbol{\theta}$ pode ser obtido como solução

$\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0}$ usando métodos numéricos, como o algoritmo de Newton-Raphson, em que

$$\mathbf{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \log L_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

2.4.2 Modelagem Semiparamétrica

O modelo de taxas de falhas proporcionais de Cox é um dos modelos de sobrevivência mais utilizados, em especial em aplicações médicas, desde a sua introdução em 1972 por Cox. Um grande atrativo desse modelo é a técnica de estimação da verossimilhança parcial que permite estimar os parâmetros da regressão ignorando a função de taxa de falha da linha de base. No entanto, esse recurso só se aplica a tempos de sobrevivência censurados à direita. Para dados censurados por intervalo, o modelo Cox ainda é bastante utilizado, no entanto a taxa de falha de linha de base deve ser estimada juntamente com os parâmetros da regressão. Várias abordagens têm sido sugeridas para um modelo Cox para dados censurados por intervalo, desde métodos puramente paramétricos até métodos semiparamétricos.

Seja $(l_i, r_i]$ o intervalo observado em que o evento, para o i-ésimo indivíduo, ocorreu e $\mathbf{x}_i$ o vetor de covariáveis do sujeito i, $i = 1, ..., n$. Considerando que somente censuras intervalares e censuras à direita tenham ocorrido, a função de verossimilhança é

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [S(l_i|\mathbf{x}_i) - S(r_i|\mathbf{x}_i)]^{\delta_i} [S(l_i|\mathbf{x}_i)]^{(1-\delta_i)},$$

em que δ_i é a variável indicadora de censura intervalar. Assumindo que $S(t|\mathbf{x})$ é especificada pelo modelo de Cox, a função de verossimilhança fica na forma,

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[[S_0(l_i)]^{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)} - [S_0(r_i)]^{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right]^{\delta_i} \left[[S_0(l_i)]^{\exp(\boldsymbol{\beta}^T \mathbf{x}_i)} \right]^{(1-\delta_i)} \tag{2.22}$$

em que $S_0(t)$ é a função de sobrevivência de base.

Pan (1999) propôs encontrar o estimador de máxima verossimilhança de $\boldsymbol{\theta}$ juntamente com a função de sobrevivência de base $S_0(t)$, assumindo que $S_0(t)$ é constante por partes e estendendo o algoritmo iterativo do minorante convexo(ICM) para dados de sobrevivência com censura intervalar. No algoritmo de Pan a função de verossimilhança

é parametrizada em termos da função taxa de falha acumulada $\Lambda_0(t)$, por ser mais conveniente porque $\Lambda_0(t)$ não possui limite superior, assim nenhuma restrição é necessária. Para maiores detalhes pode-se consultar Colosimo e Giolo (2006), Bogaerts et al. (2017) e Sun (2006).

2.5 Análise Multivariada

Um pressuposto fundamental na teoria apresentada até a última seção é que todos os tempos de sobrevivência são independentes. A análise de dados de sobrevivência multivariada é usada quando a suposição de independência não é satisfeita. Algumas situações em que isso pode ocorrer são:

a) os tempos de diferentes eventos são monitorados no mesmo sujeito;

b) medidas repetidas são tomadas sobre o mesmo sujeito ao longo do tempo;

c) os sujeitos dentro de um mesmo grupo ou conglomerado fornecem respostas correlacionadas decorrentes de relações genéticas ou efeitos ambientais ou sociais comuns.

Na situação c) existem duas fontes de variações nas observações. A primeira é a variabilidade entre sujeitos dentro de um conglomerado , e, a segunda é a variabilidade entre os conglomerados. Essas duas fontes de variação fazem com que a variância aumente e devem ser levadas em consideração na análise.

Existem duas classes principais de modelos que acomodam a questão da dependência dentro do conglomerado para dados de sobrevivência. Um deles são os modelos de fragilidade (Colosimo e Giolo (2006), Wienke (2010), Hanagal (2011)), que especificam explicitamente a dependência dentro do conglomerado por meio de efeitos aleatórios e fornecem inferência específica do conglomerado. Esses modelos normalmente impõem suposições sobre a estrutura da dependência dentro do conglomerado e a distribuição dos efeitos aleatórios e tendem a ser computacionalmente intensivos.

A outra classe de modelos são os modelos marginais (Wei et al. (1989), Liang e Zeger (1986),Liang e Zeger (1993), Cai e Prentice (1997), Spiekerman e Lin (1998)).

Esses modelos não dependem de suposições sobre a estrutura de dependência e têm uma interpretação de média populacional. Nesse sentido, uma abordagem é ignorar inicialmente as correlações entre os tempos de sobrevivência e ajustar os modelos assumindo que as respostas são independentes. Em seguida, a estimativa da covariância robusta é usada para obter erros padrão para os parâmetros estimados. Isso geralmente é chamado de matriz de trabalho de independência.

Diggle et al. (2013) recomendam o uso de modelos marginais quando o objetivo do estudo é fazer inferências baseadas na população e modelos condicionais quando se pretende fazer inferências sobre respostas individuais.

Quando o principal interesse de um estudo é estimar os efeitos marginais de variáveis explicativas dentro do conglomerado, as correlações entre os sujeitos tornam-se parâmetros de perturbação. Huster et al. (1989) propõem um modelo de matriz de trabalho de independência (IWM) para dados de sobrevivência correlacionados. Eles especificam modelos paramétricos para as distribuições marginais e usam a suposição incorreta de que os membros de cada conglomerado são independentes para obter um conjunto de equações de estimação. A verossimilhança de IWM resultante é o produto das verossimilhanças marginais de todos os conglomerados. Os parâmetros dos modelos marginais podem ser estimados maximizando essa verossimilhança. No entanto, o inverso da matriz de informação da verossimilhança IWM não fornece um estimador consistente para a matriz de covariância para a distribuição normal limitante das estimativas do parâmetro IWM. Huber et al. (1967) estudou a consistência e normalidade assintótica de estimadores obtidos da maximização da verossimilhança incorretas em condições mais gerais. Royall (1986) mostra como obter um estimador consistente da matriz de covariância, conhecida como robusto ou "estimador sanduíche", a partir dos resultados desenvolvidos por Huber et al. (1967).

Considere n conglomerados, e que cada um dos i conglomerados possui m_i indivíduos. Seja $T_i = (t_{1i}, t_{2i}, ..., t_{m_i i})$ o vetor de tempos de falhas para i-ésimo conglomerado e seja $C_i = (c_{1i}, c_{2i}, ..., c_{m_i i})$ o vetor de tempos de censura correspondente. Considere que T_j são independentes, C_j são independentes, t_{ij} é independente de c_{ij} para todo i e todo j, mas no entanto, os elementos de T_j podem ser correlacionados. Seja f_i e S_i a função

de densidade marginal e função de sobrevivência marginal, respectivamente. Seja $\boldsymbol{\theta}_i$ o i-ésimo vetor de parâmetros marginal e $\mathbf{x}_{ij}$ um vetor de covariáveis. Sob a hipótese do modelo de trabalho de independência, a função de verossimilhança conjunta é

$$L(\boldsymbol{\theta};\mathbf{x}) = \prod_{i=1}^{n}\prod_{j=1}^{m_i} f_i(y_{ij};\boldsymbol{\theta}_i,\mathbf{x}_{ij})^{\delta_{ij}} S_i(y_{ij};\boldsymbol{\theta}_i,\mathbf{x}_{ij})^{1-\delta_{ij}} \tag{2.23}$$

em que $y_{ij} = min(t_{ij}, c_{ij})$ e δ_{ij} é o indicador de censura, $i = 1, ..., n$ e $j = 1, ..., m_i$.

Huber et al. (1967) fornecem condições de regularidade que são suficientes para garantir uma distribuição normal limite para os estimadores de máxima verossimilhança sob o modelo IWM baseado na consistência das estimativas de parâmetros. A função log-verossimilhança tem a forma

$$\log L(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \delta_{ij}\log f_i(y_{ij};\boldsymbol{\theta}_i,\mathbf{x}_{ij}) + (1-\delta_{ij})\log S_i(y_{ij};\boldsymbol{\theta}_i,\mathbf{x}_{ij}). \tag{2.24}$$

Royall (1986) propôs uma abordagem para estimativa da variância que é robusta. Ele expressa a matriz de covariância da distribuição limite de $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ como

$$\mathbf{V}(\boldsymbol{\theta}) = \lim_{n\to\infty} \mathbf{I}(\boldsymbol{\theta})^{-1} E[\mathbf{U}(\boldsymbol{\theta})\mathbf{U}(\boldsymbol{\theta})^T]\mathbf{I}(\boldsymbol{\theta})^{-1}, \tag{2.25}$$

em que $\mathbf{U}(\boldsymbol{\theta}) = \dfrac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ é o vetor escore e $\mathbf{I}(\boldsymbol{\theta}) = -\dfrac{\partial^2 \log L(\boldsymbol{\theta})}{\partial^2 \boldsymbol{\theta}}$. Pode ser estimada por

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}\left\{\sum_{i=1}^{n}\mathbf{U}_i(\hat{\boldsymbol{\theta}})\mathbf{U}_i(\hat{\boldsymbol{\theta}})^T\right\}\mathbf{I}(\hat{\theta})^{-1}. \tag{2.26}$$

Se a suposição de independênicia é correta, isto é, os membros dos conglomerados são verdadeiramente independentes, então $E\left[\sum_{i=1}^{n}\mathbf{U}_i(\hat{\boldsymbol{\theta}})\mathbf{U}_i(\hat{\boldsymbol{\theta}})^T\right] = \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$ e $\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$, o que vai de acordo com a teoria assintótica para o caso de observações completamente indepenentes.

Em situações em que a esperança é difícil, ou até mesmo impossível, de ser calculada, $E[\mathbf{I}(\boldsymbol{\theta})]$ é estimado por $\mathbf{I}(\hat{\boldsymbol{\theta}})$.

Capítulo 3

Riscos Competitivos

Na abordagem tradicional da análise de sobrevivência considera-se uma única causa para a ocorrência do evento de interesse. No entanto existem situações em que várias causas de falha são possíveis, mas somente a ocorrência da primeira delas pode ser observada. Quando um indivíduo está sob risco de falhar por K diferentes tipos de causas, esses diferentes tipos são chamados de riscos competitivos ou concorrentes.

Na presença de riscos competitivos um indivíduo em estudo tem a possibilidade de sofrer um dos diferentes tipos de eventos possíveis, que impede a realização ou altera a probabilidade de ocorrência do evento de interesse. O termo riscos competitivos também inclui situações em que o intersse está na ocorrência do primeiro evento (Putter et al., 2007; Bakoyannis e Touloumi, 2011). Nessas situações, os métodos usuais de análise de sobrevivência são incorretamente utilizados, ao considerar apenas indivíduos que sofrem o evento de interesse, censurando todos os demais que falharam por qualquer evento competitivo. Neste caso, os dados censurados são assumidos não informativos. Assim, o tempo de sobrevivência por uma causa de interesse é considerado independente do tempo em que o indivíduo sofre um evento competitivo, considerado, neste caso, a causa da censura.

Muitos cuidados são necessários em relação à suposição de independência entre a censura e o evento de interesse. Dependendo da causa da censura, tal suposição pode ser violada, ou seja, as observações censuradas podem fornecer informações relevantes sobre o tempo de sobrevivência (Putter et al., 2007) . Além do que, como não se pode obter

informações sobre os valores dos tempos associados a cada uma das causas de falha para o mesmo indivíduo, é praticamente impossível testar a independência entre as distintas causas de falhas.

De outro modo, a censura devido à perda de seguimento é negativamente correlacionada ao tempo de sobrevivência quando os pacientes saudáveis desistem de continuar participando no estudo. A censura desses indivíduos causará um viés descendente da curva de sobrevivência estimada, superestimando assim a probabilidade de ocorrência do desfecho, uma vez que indivíduos com pior prognóstico são considerados representativos para os indivíduos já censurados. Agora, quando indivíduos com estado avançado da doença abandonam o estudo porque se tornam muito doentes para acompanhamento adicional acarreta uma correlação positiva entre o tempo da censura e o tempo de sobrevivência. A censura destes individuos causará um viés ascendente da curva de sobrevivência estimada. Assim, nesses casos, a ocorrência de censura contém uma informação sobre o tempo de sobrevivência, contestando, dessa forma, o pressuposto para mecanismo de censura aleatória, também conhecida por censura não informativa.

3.1 Representação de riscos competitivos

Duas abordagens diferentes para lidar com riscos competitivos podem ser encontradas na literatura. A primeira delas é por meio de uma abordagem de tempo de falha latente, implicando em uma distribuição conjunta dos tempos para os K tipos de eventos. Na outra abordagem, o processo de risco competitivo é representado por duas variáveis aleatórias, uma para o tempo do evento e outra para o tipo do evento.

Geralmente, a última abordadem é preferida devido a presença do problema de identificabilidade na formulação dos tempos de falha latentes. A seguir, uma introdução a essas representações é apresentada, discussões mais aprofundadas sobre as diferentes abordagens para dados de riscos competitivos podem ser encontradas em Pintilie (2006), Crowder (2001), e Beyersmann et al. (2011).

3.1.1 Riscos Competitivos como tempos de falha latentes

Uma maneira de abordar o problema de riscos competitivos é assumir que existem variáveis aleatórias $T_1, T_2, ..., T_K$ para o tempo de cada um dos K tipos de eventos possíveis. No contexto de risco competitivos somente o tempo para o primeiro evento, denotado por T, é observado, isto é,

$$T = \min\{T_1, T_2, ..., T_K\}.$$

Além disso, uma variável de censura, C, denotando o tipo de evento observado, é necessária. Assim, C é definida como $C = 0$ se a observação é censurada e $C = k$, $k = 1, 2, ..., K$, caso contrário.

Para estimar a função de sobrevivência da distribuição conjunta $S(t_1, t_2, ..., t_K)$, a correlação entre os tempos para diferentes tipos de eventos deve ser avaliada. No entanto, em uma configuração clássica de riscos competitivos, apenas um tipo de evento pode ser observado para cada indivíduo, então a estrutura de correlação não pode ser estimada a partir dos dados observados, sendo assim necessário assumir uma estrutura de correlação que não pode ser verificada.

Como a estrutura de correlação não pode ser estimada a partir dos dados observados, Prentice et al. (1978) questionou a plausibilidade da abordagem dos tempos de falha latentes e, assim, desestimula o seu uso. Tsiatis (1975) demonstrou que, para cada distribuição conjunta, assumindo independência entre os tempos de falha latentes, pode ser encontrada uma estrutura de dependência que fornece a mesma probabilidade. Beyersmann et al. (2011) discutiram e ilustram que não há ganho, mas problemas adicionais, quando a abordagem de tempo de falha latente é considerada.

3.1.2 Riscos Competitivos como variáveis aleatórias bivariadas

Conforme visto anteriormente, os dados de sobrevivência na abordagem tradicional são geralmente apresentados como um par (T, C). A variável aleatória C, indicadora de censuras, assume valor 1 se o evento de interesse foi observado e 0 se a observação foi censurada. Quando $C = 1$, T corresponde ao tempo até a ocorrênencia do evento, e quando $C = 0$, T é o tempo até a censura.

Esta definição pode ser estendida para situação de riscos competitivos, em que são possíveis $k \geq 2$ causas de falhas distintas e mutuamente exclusivas. Novamente, os dados são representados pelo par (T, C) e a variável aleatória indicadora de censura C será novamente definida como 0 se a observação for censurada, e no caso da observação não ser censurada, C assumirá k, em que k é o tipo da primeira falha (evento observado), $k = 1, 2, ..., K$. Assim, quando $C = k$, T representa o tempo até a ocorrência do evento pela causa k, e caso contrário, T é o tempo até a censura.

Os dados para um determinado indivíduo são valores observados de (T, C), assim, escrevemos (t_i, c_i) para os dados do i-ésimo indivíduo, $i = 1, 2, ..., n$, em que c_i é uma variável de estado que indica o tipo de evento $c_i \in \{1, 2, ..., K\}$ ou $c_i = 0$ quando o tempo do evento é censurado.

Como os dados de riscos competitivos não são representados por diferentes variáveis aleatórias para os tempos dos eventos possíveis, e sim por uma variável fornecendo o tempo do evento e uma variável indicando o tipo do evento, o conceito de dependência estatística entre os tempos para os diferentes tipos de eventos não se aplica a esta abordagem (Beyersmann et al., 2011, p.160).

3.2 Conceitos Básicos

Como no cenário de riscos competitivos os indivíduos podem falhar por diferentes tipos de eventos, as medidas usadas para a análise de sobrevivência tradicional com apenas um determinado tipo de evento devem ser adaptadas. Nesta seção, os conceitos e quantidades mais importantes e comumente usados são descritos.

Como na análise de sobrevivência tradicional, as funções de taxa de falha desempenham um papel importante para a análise de dados de riscos competitivos, uma vez que podem ser estimadas na presença de observações censuradas.

Na presença de riscos competitivos as funções de maior importância são:

- Função de taxa de falha causa específica, $\lambda_k(t)$ e
- Função de incidência acumulada associada a causa k, $F_k(t)$, $k = 1, ..., K$,

que serão definidas a seguir.

Segundo Porta Bleda et al. (2008), a distribuição conjunta do par (T,C) é completamente especificada por meio de qualquer um dos riscos causa específica, $\lambda_k(t)$, ou pelas funções de incidência acumulada, $F_k(t)$.

A função taxa de falha causa específica para o tipo de evento k é a adaptação natural da função taxa de falha mostrada na definição 2.1.4, e fornece a probabilidade de um indivíduo falhar por um evento do tipo k em um pequeno intervalo de tempo infinitesimal t até $t + \Delta t$ dado que não tenha falhado por qualquer evento até o instante t.

Definição 3.2.1. *A função taxa de falha causa específica, $\lambda_k(t)$, representa a taxa instantânea da ocorrência da k-ésima falha, para $k = 1, 2, ..., K$, no tempo t, condicionada a não ocorrência de qualquer outro tipo de falha entre os indivíduos sob risco até o tempo t. Sendo definida por :*

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, C = k | T > t)}{\Delta t}.$$

A taxa de falha geral ou total, $\lambda(t)$, de um evento de qualquer tipo pode ser encontrada somando todos as taxas de falha causa específica, isto é:

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k(t).$$

Segundo Putter et al. (2007), qualquer quantidade que pode ser determinada pela função de taxa de falha causa específica é estimável. Assim, a função taxa de falha acumulada causa específica é definida por

$$\Lambda_k(t) = \int_0^t \lambda_k(u) du.$$

Pintilie (2007) define a função subsobrevivência (*subsurvivor funtion*) ou função de sobrevivência causa específica como sendo a probabilidade de que um evento do tipo k não ocorra no tempo t, isto é, $S_k(t) = P(T > t, C = k)$, e assim

$$S_k(t) = \exp\{-\Lambda_k(t)\}.$$

Embora $S_k(t)$ possa ser estimada, ela não deve ser interpretada como uma função marginal de sobrevivência para a causa k. Ela só tem essa interpretação se as distribuições do tempo dos eventos concorrentes e a distribuição de censura forem independentes (Putter et al., 2007).

A função de sobrevivência global ou total, $S(t) = P(T \geq t)$, depende das funções taxa de falha causa específica acumulada para todos os K tipos de eventos,

$$S(t) = \exp\left(-\sum_{k=1}^{K} \Lambda_k(t)\right) = \exp\left(-\Lambda(t)\right),$$

em que $\Lambda(t)$ é chamada função taxa de falha causa específica global. Essa função de sobrevivência tem uma interpretação, representa a probabilidade de um indivíduo não sofrer falha por qualquer causa até o tempo t.

A função de sobrevivência global também pode ser expressa como,

$$S(t) = \prod_{k=1}^{K} S_k(t).$$

Da definição de função taxa de falha causa específica, $\lambda_k(t)$, temos que

$$\begin{aligned}\lambda_k(t) = &\lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, C = k)}{\Delta t P(T > t)} \\ = &\frac{1}{P(T > t)} \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, C = k)}{\Delta t} = \frac{f_k(t)}{S(t)},\end{aligned} \tag{3.1}$$

em que $f_k(t)$ é a função de densidade causa específica.

Segundo Collett (2015), a função de incidência acumulada para uma causa particular é um resumo mais útil dos dados do que uma função de sobrevivência. Esta é a probabilidade de sobreviver até t e falhar pela causa k, na presença de todas as outras causas.

Definição 3.2.2. *A função de incidência acumulada (FIA), $F_k(t)$, para a causa k, também conhecida como função de subdistribuição, é definida como a probabilidade de ocorrência da falha pela causa k antes do tempo t, isto é,*

$$F_k(t) = P(T \leq t, C = k), \quad k = 1, ..., K.$$

O valor máximo da função de incidência acumulada é

$$P(T < \infty, C = k) = P(C = k) = \pi_k,$$

em que π_k é a probabilidade de ocorrência da causa k.

Essa função também pode ser encontrada na literatura como, distribuição marginal, taxa de falha causa específico absoluta, curva de incidência bruta e probabilidade de falha causa específica (Pintilie, 2006). Ela pode ser expressa em termos das taxas de falha causa específica, $\lambda_k(t)$, e da função de sobrevivência global, $S(t)$,

$$F_k(t) = \int_0^t \lambda_k(u)S(u)du.$$

A função de distribuição global para T é a probabilidade de que um evento de qualquer tipo ocorra no ou antes do tempo t. Assim, ela é a soma das FIA's para todas as causas de falha, isto é

$$F(t) = P(T \leq t) = \sum_{k=1}^{K} P(T \leq t, C = k) = \sum_{k=1}^{K} F_k(t).$$

Portanto , temos que

$$S(t) = 1 - \sum_{k=1}^{K} F_k(t).$$

Note que a FIA para causa k não é propriamente uma função de distribuição, pois na presença de pelo menos duas causas de falhas distintas, F_k não converge para 1 quando t vai para o infinito, mas sim para a probabilidade global do evento do tipo k, isto é

$$\lim_{t\to\infty} F_k(t) = P(C = k) \leq 1.$$

Por essa razão ela também é conhecida como função de subdistribuição.

Portanto, para t tendendo ao infinito, as funções de incidência acumulada para todos os K tipos de eventos somam um, isto é,

$$\lim_{t\to\infty} \sum_{k=1}^{K} F_k(t) = 1.$$

Podemos relacionar a função de incidência acumulata para a causa k e as funções taxas de falha causa específica por meio da expressão

$$F_k(t) = \int_0^t \lambda_k(u)S(u)du = \int_0^t \lambda_k(u) \exp\left(-\sum_{l=1}^{K} \Lambda_l(u)\right) du.$$

Portanto, a função de incidência acumulada para a causa k depende das funções de taxa de falha causa específica de todos os K tipos de eventos, com isso os riscos para todos os tipos de eventos têm um efeito na probabilidade de um evento do tipo k. Como consequência desse fato, em comparações de grupos, uma maior taxa de falha causa específica para o evento do tipo k para um grupo não indica, necessariamente, uma maior incidência acumulada de evento (Putter et al., 2007).

A partir da função de incidência acumulada, $F_k(t)$, podemos obter quantidades que podem ser de interesse no contexto de riscos competitivos. Podemos obter a probabilidade de falha antes do instante t, quando a causa é k, por

$$P(T < t|C = k) = \pi_k^{-1} F_k(t),$$

em que π_k é a probabilidade da causa k ocorrer. Também podemos determinar a probabilidade de falha pela causa k, quando a falha ocorreu antes do tempo t, por

$$P(C = k|T < t) = \frac{F_k(t)}{F(t)}.$$

As estimativas dessas probabilidades podem ser obtidas a partir da função de incidência acumulada estimada.

Deve se tomar um certo cuidado ao interpretar a FIA e a função taxa de falha causa específica. Elas devem ser interpretadas considerando a presença das outras causas (Porta Bleda et al., 2008). Por exemplo, a taxa de falha causa específica da k-ésima causa não é o risco de falha pela causa k no tempo t, e sim o risco de falhar primeiro por causa k do que por outras causas. Portanto, essas funções não são interpretadas como se outras causas de falha estivessem ausentes, isto é, não devem ser interpretadas como funções marginais.

Para Porta Bleda et al. (2008), a escolha entre as funções taxa de falha causa específica e de incidência acumulada, em geral, é determinada pela questão de interesse

do pesquisador. Se o interesse é no número de indivíduos que falharam, no final, por uma causa específica, então usa-se as funções incidência acumulada. Por outro lado, se o interesse é verificar como o risco de falha devido às diferentes causas se comporta ao longo do tempo, as funções taxa de falha causa específica são mais indicadas.

Gray (1988) introduziu a chamada função de taxa de falha de subdistribuição que está diretamente ligado à função de incidência acumulada na presença de riscos competitivos.

Definição 3.2.3. *A função de taxa de falha de subdistribuição para a causa k, denotado por* $\gamma_k(t)$*, é*

$$\begin{aligned}\gamma_k(t) &= \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, C = k | T > t \cup \{T < t, C \neq k\})}{\Delta t} \\ &= -\frac{d}{dt} \log(1 - F_k(t)). \end{aligned} \tag{3.2}$$

A função de taxa de falha de subdistribuição é interpretada como a taxa instantânea de falha no tempo t da causa k entre os que ainda estão vivos ou aqueles que morreram por qualquer das outras $(K-1)$ causas concorrentes (Mozumder et al., 2017).

A função de taxa de falha de subdistribuição para a causa k difere da função taxa de falha causa específica pela definição de seu conjunto de risco. Para a função de taxa de falha de subdistribuição para a causa k no instante t, os indivíduos que falharam por causas diferentes de k antes de t permanecem no conjuto de risco.

A relação entre a função de incidência acumulada e a função de taxa de falha de subdistribuição é dada pela equação

$$F_k(t) = 1 - \exp(-\Gamma_k(t)), \tag{3.3}$$

em que $\Gamma_k(t)$ é a função taxa de falha de subdistribuição acumulada

$$\Gamma_k(t) = \int_0^t \gamma_k(s) ds. \tag{3.4}$$

Como $\gamma_k(t)$ possui as propriedades do risco para a função de incidencia acumulada $F_k(t)$, também chamada de função subdistribuição, é chamada função de taxa de falha de subdistribuição. Devido a sua relação direta com a função de incidência acumulada, a função de taxa de falha de subdistribuição tornou-se muito utilizados nos últimos anos.

3.3 Modelos de Regressão

Assim como na análise de sobrevivência tradicional, na análise de riscos competitivos, os pesquisadores, em muitas das vezes, têm interesse em identificar fatores prognósticos associados com a resposta, isto é, identificar a partir de p covariáveis, um subconjunto de covariáveis que afetam o risco de forma mais significativa, e com isso o tempo de sobrevivência do paciente. Para realizar tal objetivo, no contexto de risco competitivos, duas estratégias diferentes de modelagem de regressão são as mais utilizadas: modelar as taxa de falha causa específica ou modelar a taxa de falha de subdistribuição. A primeira é conhecida como modelo de regressão taxa de falha causa específca proposto por Prentice et al. (1978) e o segundo é o modelo de regressão taxa de falha de subdistribuição, introduzido por Fine e Gray (1999).

Outras abordagens são encontradas na literatura. Um resumo, com aplicação em um estudo clínico, de diferenes abordagens é apresentado por Haller et al. (2013).

3.3.1 Modelo de regresssão taxa de falha causa específica

Prentice et al. (1978) propuseram estimar o efeito de covariáveis nas taxas de falha causa específica. Para cada indivíduo i os dados $(t_i, c_i, \delta_i, \mathbf{x}_i)$ são observados, em que t_i é o tempo observado, c_i é a causa observada de falha, δ_i é um indicador de censura retornando o valor zero para uma obseração censurada e um valor um se algum evento foi observado, isto é, se $c_i = k, k = 1, ..., K$, e $\mathbf{x}_i$ é o vetor de covariáveis, que é considerado constante ao longo do tempo. A função de verossimilhança sob censura não informativa pode ser escrita como

$$\begin{aligned} L &= \prod_{i=1}^{n} \left(\left(\prod_{k=1}^{K} \lambda_k(t_i|\mathbf{x}_i)^{\delta_i} \right) S(t_i|\mathbf{x}_i) \right) \\ &= \prod_{i=1}^{n} \left(\left(\prod_{k=1}^{K} \lambda_k(t_i|\mathbf{x}_i)^{\delta_i} \right) \prod_{k=1}^{K} \exp\left(- \int_0^{t_i} \lambda_k(s|\mathbf{x}_i) ds \right) \right). \end{aligned} \tag{3.5}$$

Segundo Prentice et al. (1978), a forma da função de verossimilhança apresentada leva a algumas implicações: as funções taxa de falha e os coeficientes da regressão são identificáveis e podem ser estimados a partir dos dados observados, a função score para a

estimação dos coeficientes da regressão para o evento de interesse não muda quando todos os eventos concorrentes observados são tratados como observações censuradas. Assim, métodos padrões, usando um modelo paramétrico ou um modelo de Cox, podem ser aplicados considerando eventos concorrentes como observações censuradas. E os efeitos das covariáveis na taxa de falha causa específica para diferentes tipos de eventos podem ser estimados em modelos de regressão separados.

Quando os tempos dos indivíduos que falharam por eventos concorrentes são considerados como censura, esses são tratados como se houvesse a possibilidade do evento de interesse ocorrer no futuro e isso pode levar a uma taxa de falha causa específica superestimada.

Outra consideração, feita por Collett (2015), é que os modelos de causa específica são baseados na suposição de censura independente. Se os eventos concorrentes não ocorrem independentemente do evento de interesse esta suposição não é válida. No entanto, a suposição de riscos concorrentes independentes não pode ser testada usando os dados observados.

Apesar dessas desvantagens, essa abordagem pode ser justificada quando o interesse é em verificar como as variáveis influenciam o risco associado a uma determinada causa de falha, ignorando as falhas por outras causas (Collett, 2015).

Prentice et al. (1978) propuseram considerar um modelo de regressão do tipo Cox (Cox, 1972) para estimar os efeitos das covariáveis sobre a taxa de falha causa específica, assumindo taxas de falha causa específica proporcional.

$$\lambda_k(t|\mathbf{x}) = \lambda_{0k}(t)\exp(\boldsymbol{\beta}_k^T\mathbf{x}), \tag{3.6}$$

em que $\lambda_{0k}(t)$ descreve a taxa de falha causa específica linha de base para o tipo de evento k, $\mathbf{x}$ é o vetor de covariáveis $p-$dimensional e $\boldsymbol{\beta}_k$ é o vetor de coeficientes da regressão p-dimensional para o k-ésimo tipo de evento.

Seja $(t_{k1}, t_{k2}, ..., t_{kn_k})$ os tempos de falha distintos observados para a causa de falha k, e $n_k \leq n$, $k = 1, ..., K$. Quando o modelo de riscos proporcionais de Cox é usado para avaliar os efeitos das covariávies sobre as taxas de falha causa específicas, a

verossimilhança parcial pode ser denotada por

$$L = \prod_{k=1}^{K} \prod_{i=1}^{n_k} \frac{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_{(i)})}{\sum_{j \in R(t_{ki})} \exp(\boldsymbol{\beta}_k^T \mathbf{x}_j)}, \tag{3.7}$$

em que $\mathbf{x}_{(i)}$ o vetor de coeficiente da regressão do indivíduo que falhou pelo evento k no tempo t_{ki} e $R(t_{ki})$ é o conjunto sob risco em t_{ki}. Devido a fatoração da verossimilhança parcial, os coeficientes da regressão para os diferentes tipos de eventos podem ser estimados a partir de modelos separados, se não forem assumidos efeitos comuns ou a mesma taxa de falha linha de base. As estimativas dos coeficientes da regressão podem ser encontradas numericamente usando um algoritmo de Newton-Raphson.

Os coeficientes da regressão, $\beta_{k1}, ..., \beta_{kp}$, podem ser interpretados como log taxa de falha causa específca para o evento k.

Para estimar a função de incidência acumulada do modelo taxa de falha causa específica, as funções de taxa de falha linha de base $\lambda_{0k}(t)$ devem ser estimadas para todos os tipos de eventos $k = 1, ..., K$. Marubini e Valsecchi (2004) apresentaram uma versão do estimador de Breslow (Breslow, 1972) para a taxa de falha de linha de base em um modelo de riscos proporcionais

$$\hat{\lambda}_{0k}(t_{ki}) = \frac{d_k(t_{ki})}{\sum_{j \in R(t_{ki})} \exp(\hat{\boldsymbol{\beta}}_k^T \mathbf{x}_j)}, \tag{3.8}$$

em que $R(t_{ki})$ descreve o conjunto de risco no instante t_{ki}, $d_k(t_{ki})$ é o número de eventos do tipo k no instante t_{ki} e $\hat{\boldsymbol{\beta}}_k$ é a estimativa para o vetor de coeficientes da regressão para o tipo de evento k do modelo regressão taxa de falha causa especifico proporcional. Para cada instante t sem a ocorrência de um evento do tipo k, a estimativa para a taxa de falha linha de base causa específica é zero. Consequentemente, a taxa de falha linha de base causa específica acumulada para o evento k pode ser estimado como

$$\hat{\Lambda}_{0k}(t) = \sum_{i:t_{ki} \leq t} \hat{\lambda}_{0k}(t_{ki}). \tag{3.9}$$

Assumindo o vetor dos tempos observado e ordenados do evento do tipo k, $t_k = (t_{k1}, ..., t_{kn_k})$, o estimador para a função de incidência acumulada do evento evento k pode

ser escrito como

$$\begin{aligned}\hat{F}_k(t|\mathbf{x}) &= \sum_{i:t_{ki}\leq t} \hat{\lambda_k}(t_{ki}|\mathbf{x})\hat{S}(t_{k(i-1)}|\mathbf{x}) \\ &= \sum_{i:t_{ki}\leq t} \hat{\lambda}_{k0}(t_{ki})\exp(\boldsymbol{\beta}_k^T\mathbf{x})\exp\left(-\sum_{l=1}^{K}\hat{(\Lambda)}_l(t_{k(i-1)}|\mathbf{x})\right) \\ &= \sum_{i:t_{ki}\leq t} \hat{\lambda}_{k0}(t_{ki})\exp(\boldsymbol{\beta}_k^T\mathbf{x})\exp\left(-\sum_{l=1}^{K}\hat{(\Lambda)}_{l0}(t_{k(i-1)})\exp(\hat{\boldsymbol{\beta}}_l^T\mathbf{x})\right). \end{aligned} \tag{3.10}$$

Embora os eventos concorrentes possam ser tratados como observações censuradas para a estimativa da taxa de falha causa específica, os eventos concorrentes devem ser considerados adequadamente para a estimativa das funções de incidência acumulada. Como pode ser visto em (3.10), a função de incidência acumulada para o evento k depende das taxa de falha causa especifica de todos os tipos de eventos. Portanto, um efeito observado na taxa de falha causa específica não se traduz necessariamente em um efeito sobre a função de incidência acumulada.

3.3.2 Modelo de regressão taxa de falha de subdistribuição

Fine e Gray (1999) desenvolveram um modelo de regressão para dados de sobrevivência na presença de riscos competitivos, baseado na função de taxa de falha de subdistribuição (3.2), é conhecido como modelo de regressão Fine e Gray. Em seu artigo original, Fane e Gray propuseram o uso de um modelo de regressão para a função de taxa de falha de subdistribuição para um evento de interesse, assumindo a razão da função de taxa de falha de subdistribuição proporcional,

$$\gamma_k(t|\mathbf{x}) = \gamma_{k0}(t)\exp(\boldsymbol{\beta}_k^T\mathbf{x}) \tag{3.11}$$

em que $\gamma_k(t|\mathbf{x})$ denota a função de taxa de falha de subdistribuição para o evento de interesse dependendo do vetor de covariáveis $\mathbf{x}$, γ_{k0} é a função de taxa de falha de subdistribuição linha de base para um indivíduo com todas as covariáveis iguais a zero, e $\boldsymbol{\beta}_k$ é o vetor de coeficientes da regressão.

Em geral, a suposição de proporcionalidade não pode ser verdadeira para os modelos de regressão de taxa de falha de subdistribuição separados para diferentes tipos de

eventos (Beyersmann et al., 2011, p.139). Grambauer et al. (2010) investigaram o impacto da especificação incorreta do modelo e demostraram que o modelo de regressão de taxa de falha de subdistribuição tem uma interpretação adequada, mesmo quando as funções de taxa de falha de subdistribuição foram erroneamente considerados proporcionais.

Para estimar os coeficientes da regressão no modelo de regressão de taxa de falha de subdistribuição é necessário um conjunto de risco diferente do modelo de regressão taxa de falha causa específica. Embora a estimação dos coeficientes da regressão seja direta quando são observados dados completos para todos os indivíduos e sob censura administrativa, o procedimento de estimação torna-se complexo para dados incompletos com censura não administrativa.

Quando os dados completos estão disponíveis, isto é, o tempo do evento e o tipo do evento foram observados para cada indivíduo, a verossimilhança para o modelo de regressão pode ser escrita como mostrado em (3.7), mas com o conjuto de risco $R(t_{ki})$ definido como

$$R(t_{ki}) = \{j : (t_j \geq t_{ki}) \cup (t_j \leq t_{ki} \cap d_j \neq k)\}, \tag{3.12}$$

que inclui todos os indivíduos que ainda estão sob observação em t_{ki}, isto é, todos os indivíduos que não falharam por nenhuma causa antes de t_{ki} e todos os indivíduos que falharam por um evento diferente de k antes do tempo t_{ki}. A estimação dos coeficientes da regressão pode ser realizado conforme descrito para a regressão taxa de falha causa específica, mas usando o conjunto de risco definido em (3.12). A função de taxa de falha de subdistribuição linha de base pode ser estimado conforme (3.8), novamente utilizando conjunto de risco (3.12), incluindo assim indivíduos que falharam em um evento concorrente antes do ponto de tempo sob investigação.

Na presença de dados incompletos, ou seja, quando os indivíduos desistiram do estudo ou foram perdidos no acompanhamento, Fine e Gray propuseram usar uma função score ponderada para a estimativa dos parâmetros, a fim de obter estimativas imparciais para os coeficientes da regressão. Para construção da função score ponderada é utilizada a abordagem de probabilidade inversa de censura ponderada, introduzida por Robins e Rotnitzky (1992). A função score, que é maximizada para obter as estimativas de máxima verossimilhança parcial, é ponderada usando pesos dependentes do tempo

com base nas estimativas de Kaplan-Meier da função de sobrevivência. Cada indivíduo i contribui para função score com peso,

$$w_i(t) = r_i(t)\frac{\hat{G}(t)}{\hat{G}(min(t_i, t))}, \tag{3.13}$$

em que $r_i(t)$ indica o conhecimento do estado de vida do indivíduo i no instante t, ou seja, $r_i(t)$ é um se o individuo i estiver vivo no instante t ou se for conhecido que o indivíduo i falhou antes de t por qualquer causa de falha, e $r_i(t)$ é zero se o indivíduo i foi censurado antes do tempo t. $\hat{G}(t)$ é a estimativa de Kaplan-Meier da função de sobrevivência para os tempos censurados, obtido considerando todos os tempos de eventos, de qualquer tipo, no conjunto de dados como tempos censurados, e da mesma forma todos os tempos censurados como tempos de eventos, e calculando a estimativa de Kaplan-Meier a partir dos dados resultantes, Collett (2015).

A função de incidência acumulada pode ser estimada para um determinado vetor de covariáveis $\mathbf{x}$ a partir dos coeficientes da regressão estimados usando a relação entre a função de taxa de falha de subdistribuição e a função de incidência acumulada sem considerar os efeitos nos eventos concorrentes, da forma,

$$\begin{aligned}\hat{F}_1(t|\mathbf{x}) &= 1 - \exp\left(-\hat{\Gamma}_1(t|\mathbf{x})\right) \\ &= 1 - \exp\left(-\int_0^t \hat{\gamma}_1(s|\mathbf{x})ds\right) \\ &= 1 - \exp\left(-\int_0^t \hat{\gamma}_{01}(s)\exp(\hat{\boldsymbol{\beta}}_1\mathbf{x})ds\right).\end{aligned} \tag{3.14}$$

As duas abordagens de regressão baseada em risco, a regressão de risco específicos de causa e a regressão de taxa de falha de subdistribuição, são os métodos mais populares para análise de dados de riscos competitivos em ambientes médicos. Devido à semelhança das abordagens, os coeficientes de regressão obtidos a partir dos modelos de regressão são muitas vezes interpretados de forma igual, sem considerar que os métodos se concentram em quantidades diferentes, ou seja, a causa específica ou a função de taxa de falha de subdistribuição. Dependendo da quantidade de eventos concorrentes e dos efeitos das covariáveis nos eventos competitivos, as duas abordagens podem fornecer substancialmente coeficientes de regressão diferentes, pois a regressão taxa de falha causa específica visa

o risco instantâneo, enquanto o taxa de falha de subdistribuição está diretamente ligado ao função de incidência acumulada.

Capítulo 4

Modelo Paramétrico para dados correlacionados na presença de riscos competitivos e censura intervalar

Frequentemente, em dados de sobrevivência, o evento final é devido a uma dentre várias causas possíveis. Experimentar um determinado tipo de evento impede que o indivíduo experimente qualquer outro tipo de evento. Esses eventos são normalmente chamados de riscos concorrentes ou competitivos.

Essa estrutura de dados precisa ser considerada explicitamente para que a análise seja precisa. As análises que ignoram o aspecto de riscos competitivos dos dados, não diferenciando entre os tipos de eventos, podem perder informações sobre os efeitos da covariável e levar a uma interpretação imprecisa dos resultados.

Na presença de riscos competitivos há duas funções de maior importância. A primeira delas é a chamada função de taxa de falha de causa específica, que representa a taxa instantânea da ocorrência da k-ésima falha no tempo t, condicionada a não ocorrência de qualquer outro tipo de falha entre os indivíduos sob risco até o tempo t. A segunda função é chamada função de incidência acumulada (FIA), também conhecida como função de subdistribuição, devido ao fato de que a FIA para a causa k não é exatamente uma função de distribuição. A FIA para a causa k é definida como a probabilidade de ocorrência da falha pela causa k antes do tempo t, isto é, $F_k(t) = P(T \leq t, C = k)$.

Vários métodos foram propostos para a análise de dados de riscos competitivos. Trabalhos iniciais foram principalmente focados na estimativa e modelagem da taxa de falha de causa específica ou a taxa de falha instantânea de um evento (Prentice e Kalbfleisch, 1978). Para estudar os efeitos das covariáveis na probabilidade acumulada de uma causa particular, com indivíduos independentes, Fine e Gray (1999) propuseram um modelo de risco de subdistribuição proporcional, que tem uma correspondência direta com a função de incidência acumulada. Sun (2006) explorou um modelo de risco aditivo.

Em vários estudos o tempo do evento não é observado com precisão, mas é conhecido que ocorreu entre um intervalo de tempo, como por exemplo entre duas visitas clínicas, que dão origem a observações com censuras intervalar. Dados de riscos competitivos com censura intervalar podem surgir com frequência em diversas aplicações.

Jeong e Fine (2006) propuseram modelagem paramétrica para FIA para dados de riscos competitivos censurados à direita utilizando as distribuições Weibull e Gompertz, e Jeong e Fine (2007) apresentaram um modelo de regressão paramétrico utilizando a distribuição Gompertz no mesmo cenário. Hudgens et al. (2014) estenderam os modelos de Jeong e Fine (2006, 2007) para o caso de riscos concorrentes com censura intervalar.

Em muitas situações, os dados de riscos concorrentes não podem ser considerados independentes e métodos apropriados são necessários para explicar a correlação entre os indivíduos. A dependência desconhecida, dentro do conglomerado, entre as observações de falha necessita de métodos de regressão apropriados que levem em conta as correlações de uma maneira robusta para permitir inferência válida para os efeitos da covariáveis na função de incidência cumulativa do evento de interesse.

Modelos condicionais e marginais representam duas estratégias alternativas de modelagem de dados de riscos concorrentes em conglomerados. Os modelos condicionais especificam efeitos aleatórios para mensurar as correlações entre as observações de falha e tem uma interpretação dos parâmetros de efeitos fixos específica do conglomerado. Nesta abordagem, os modelos de fragilidades tem ganhado grande destaques (Hanagal, 2011; Wienke, 2010).

Embora os modelos de fragilidade sejam flexíveis, na medida em que modelam explicitamente a heterogeneidade entre os conglomerados, inferências válidas para os

parâmetros de efeitos fixos necessariamente depende da especificação correta da distribuição de fragilidade. Entretanto, o modelo marginal especifica os efeitos das covariáveis em toda a população de conglomerados, sem a necessidade de especificar as fragilidades não observadas (Liang e Zeger, 1986). O modelo marginal é computacionamente menos intensivo e, principalmente no contexto da nossa aplicação, tem interpretação populacional.

No cenário de risco competitivos, Zhou et al. (2012) usaram um modelo marginal de subdistribuição e uma estratégia de estimação assumindo uma estrutura de correlação de trabalho de independência. Um estimador de variância sanduíche foi desenvolvido para acomodar a dependência desconhecida dentro do conglomerado. Bogaerts et al. (2002) desenvolveram um estimador de variância sanduíche para o caso de observações em conglomerados censurados por intervalo, utilizando o modelo com matriz de trabalho independente. Com o estimador de variância sanduíche, inferência para os modelos de regressão marginal são geralmente robustos para suposições das correlações intra conglomerado.

Neste capítulo é apresentado a abordagem paramétrica de modelagem de dados de riscos competitivos em conglomerados na presença de censura intervalar. Na Seção 4.1, é apresentado o modelo para dados de riscos concorrentes censurados por intervalo proposto por Hudgens et al. (2014). Na Seção 4.2, propomos uma metodologia que estende o trabalho de Hudgens et al. (2014) para incluir dados correlacionados. Estudos de simulação são gerados para avaliar propriedades amostrais do modelo na Seção 4.3.

4.1 Dados de Riscos Competitivos Censurados por Intervalo

Os dados observados para cada indivíduo em um modelo de riscos competitivos podem ser representados por um par de variáveis aleatórias (T, C). A variável C assume valor zero se a observação do indivíduo for censurada à direita e assume valor k, caso contrário, em que k é o tipo de causa de falha observada $(k = 1, 2, \cdots, K)$. Se $C = k$

então T corresponde ao tempo até a falha pela causa k, caso contrário, T é o tempo até a censura. Cada indivíduo está sujeito a falhar devido a diferentes causas possíveis, mas a ocorrência de uma delas impede que todas as outras causas de falha ocorram. Seja, então, $\{1, ..., K\}$ um conjunto de eventos simultâneos mutuamente exclusivos e tempos de falha, $T_1, ..., T_K$, um para cada tipo de evento. Observa-se apenas o tempo mínimo de falha, $T = \min\{T_1, ..., T_K\}$. Deste modo, a distribuição conjunta de (T, C) é completamente especificada por meio das funções de incidência acumulada, $F_k(t)$, ou por meio das funções de taxa de falha de causa-específica, $\lambda_k(t)$ (Lawless, 2011).

A função de incidência acumulada para o k-ésimo evento é definida por

$$F_k(t) = P(T \leq t, C = k), \quad \text{para } k = 1, ..., K,$$

e corresponde à probabilidade de ocorrência da k-ésima causa de falha na presença das demais causas de falha. Também é conhecido que,

$$F_k(t) = \int_0^t \lambda_k(u) S(u) du, \tag{4.1}$$

em que $S(t) = P(T > t) = \exp\left(-\int_0^t \sum_{l=1}^{K} \lambda_l(u) du\right)$ é a função de sobrevivência global conforme apresentado na Seção 3.2 e

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, C = k | T > t)}{\Delta t}, \qquad k = 1, ..., K, \tag{4.2}$$

é a função de taxa de falha de causa específica para a k-ésima causa. A função $\lambda_k(t)$ determina a taxa de falha instantânea para um evento do tipo k no tempo t na presença de outras causas de falha. Assim, a FIA pode ser escrita como uma função de funções de taxa de falha de causa específica para todas as causas de falha K.

Existem diferentes maneiras de modelar parametricamente a função de incidência acumulada. As mais usadas são a modelagem direta da função de incidência acumulada, introduzida por Jeong e Fine (2006), e a abordagem chamada método de parametrização indireta, na qual modelos paramétricos separados são assumidos para funções de taxa de falha de causa específica (Prentice et al., 1978).

Os dados com censura intervalar ocorrem naturalmente em estudos de doenças em que os sintomas de interesse não são diretamente observáveis e exames laboratoriais

ou clínicos são necessários para detecção. O tempo exato para o evento de interesse T não é observado diretamente, mas é conhecido por pertencer a um intervalo de tempo $[L, R]$, tal que $0 \leq L < T \leq R \leq \infty$. No cenário envolvendo riscos competitivos, a contribuição na verossimilhança para k-ésima falha no intervalo é dada por $P(T \in [l, r], C = k) = F_k(r) - F_k(l)$ (Hudgens et al., 2014). Seja $T_1, ..., T_n$ o tempo de falha de n indivíduos, e para o i-ésimo indivíduo existe um intervalo aleatório $[L_i, R_i]$, em que L_i e R_i denotam o ponto aleatório esquerdo e direito, respectivamente, da censura intervalar. Defina $\Delta_{ik} = I(L_i < T_i < R_i, C_i = k)$ para $k = 1, ..., K$, de modo que se $\Delta_{ik} = 1$ indica que o i-ésimo indivíduo falhou devido a causa k durante o intervalo $[L_i, R_i]$. Para o caso de observações censuradas à direita, o tipo de falha é considerado desconhecido, definindo assim $\Delta_{i0} = 1$.

Note que não é possível observar (T, C) diretamente, mas sim $Y = (L, R, \Delta)$, e considere também que as observações (L_i, R_i) são independentes de (T, C), assim como, a distribuição de (L_i, R_i) não contém os parâmetros que definem a distribuição de (T, C). Neste cenário, seja $Y_1, ..., Y_n$ uma amostra aleatória de n cópias independentes e identicamente distribuídas de Y. Hudgens et al. (2014) demonstraram que a função de log-verossimilhança para dados de riscos competitivos censurados por intervalo pode ser escrita em termos das funções de incidência acumulada por

$$\log L(\Theta) = \sum_{i=1}^{n} \log l(Y_i; \Theta) \tag{4.3}$$

em que Θ é o vetor que consiste no vetor de parâmetros do modelo de $\Theta_1 \cup ... \cup \Theta_K$ e $l(Y_i; \Theta)$ é igual a

$$l(Y_i; \Theta) = \prod_{k=1}^{K} \{F_k(R_i; \Theta_k) - F_k(L_i; \Theta_k)\}^{\Delta_{ik}} \left\{1 - \sum_{k=1}^{K} F_k(L_i; \Theta_k)\right\}^{\Delta_{i0}} \tag{4.4}$$

O estimador de máxima verossimilhança (EMV), $\hat{\Theta}$, pode ser encontrado a partir da função escore, que é a derivada da Equação (4.3). Tomando o negativo das segundas derivadas da função log verossimilhança (4.3) tem-se a matriz de informação observada de Fisher. A variância de $\hat{\Theta}$ é o inverso da informação observada de Fisher. Expressões para as derivadas da função log verossimilhança (4.3) podem ser obtidas em função das funções de incidência acumulada, F_k.

4.2 Modelo Estendido para dados Correlacionados

Observe que a função de verossimilhança (4.3) foi construída considerando uma amostra aleatória simples da população. Em nosso contexto, além da presença de riscos competitivos e censura intervalar, também temos a presença de conglomerados, o que causa dependência entre indivíduos dentro de um mesmo conglomerado, mas a independência entre os conglomerados permanece válida. Assim, no contexto de dados dependentes, podemos utilizar a abordagem de modelos marginais. Esses modelos consideram apenas coeficientes de covariáveis, considerando a correlação dentro do conglomerado como um parâmetro de perturbação. A análise é baseada na abordagem Generalized Estimating Equations (GEE) proposta por Liang e Zeger (1986).

Para modelar dados de sobrevivência dependentes, censurados à direita, incluindo covariáveis, Huster et al. (1989) derivou uma abordagem GEE que permite fazer inferência estatística para as distribuições marginais enquanto trata a dependência entre os membros do conglomerado como parâmetros de perturbação. O método permite especificar as distribuições marginais independentemente da estrutura de associação e, além disso, deixa sem especificar a natureza da dependência entre os tempos de sobrevivência dos eventos dos membros do conglomerado. Os parâmetros do modelo marginal são estimados por máxima verossimilhança sob o pressuposto de independência. Este modelo foi chamado de modelo de matriz de trabalho independente por Huster et al. (1989). Na segunda etapa, os erros padrão dos estimadores são estimados por meio de uma correção pela variância sanduíche.

Considere n conglomerados, e que cada um dos i conglomerados tenha m_i indivíduos, $i = 1, ..., n$, de modo que $\sum_{i=1}^{n} m_i = m$ indivíduos em toda a amostra. Seja T_{ij} o tempo de falha para o j-ésimo indivíduo no i-ésimo conglomerado, $j = 1, ..., m_i$, de forma que se saiba apenas que o tempo de falha T_{ij} pertence a um intervalo, digamos $[L_{ij}, R_{ij}]$. Defina $\Delta_{ijk} = I(L_{ij} < T_{ij} < R_{ij}, C = k)$ para $k = 1, ..., K$, de modo que se $\Delta_{ijk} = 1$ indica que o j-ésimo indivíduo no i-ésimo conglomerado falhou pela causa k ao longo do intervalo $[L_{ij}, R_{ij}]$. Novamente, para o caso de observações censuradas à direita, o tipo de falha é considerado desconhecido, definindo assim $\Delta_{ij0} = 1$.

Assumindo que os tempos de falha T_{ij} são independentes e sob a hipótese do modelo de trabalho de independência, a função log-verossimilhança conjunta torna-se igual a

$$\log L(\Theta) = \sum_{i=1}^{n} \log l(Y_i; \Theta), \tag{4.5}$$

em que

$$l(Y_i; \Theta) = \prod_{j=1}^{m_i} \prod_{k=1}^{K} \{F_k(R_{ij}; \Theta_k) - F_k(L_{ij}; \Theta_k)\}^{\Delta_{ijk}} \left\{1 - \sum_{k=1}^{K} F_k(L_{ij}; \Theta_k)\right\}^{\Delta_{ij0}}. \tag{4.6}$$

Normalmente, uma estimativa de máxima verossimilhança $\hat{\Theta}$, que é obtida a partir da verossimilhança (4.5) que ignora a correlação entre indivíduos no mesmo conglomerado, é uma estimativa consistente de Θ para tempos de falha correlacionados e é robusta contra a especificação incorreta da estrutura de correlação. No entanto, a matriz Hessiana não fornece uma estimativa válida da matriz de covariância de $\hat{\Theta}$ devido à suposição de independência de trabalho incorreta (Peng et al., 2007).

Uma estimativa robusta e consistente da matriz de covariância é geralmente obtida pelo chamado estimador de variância sanduíche. Sob certas condições de regularidade, para um estimador consistente $\hat{\Theta}$, tem-se que $\sqrt{n}(\hat{\Theta} - \Theta)$ converge em distribuição para $N(0, \Lambda(\Theta))$ (Royall, 1986), em que,

$$\Lambda(\Theta) = \Upsilon(\Theta)^{-1} E[U(\Theta)U(\Theta)^T] \Upsilon(\Theta)^{-1}, \tag{4.7}$$

sendo que $U = \dfrac{\partial \log L(\Theta)}{\partial \Theta}$ é o vetor escore e $\Upsilon(\Theta) = E\left[\dfrac{-\partial^2 \log L(\Theta)}{\partial^2 \Theta}\right]$ é a matriz de informação esperada. Segundo Logan et al. (2010), um estimador consistente de $\Lambda(\Theta)$ é dado pela variância sanduíche,

$$\hat{\Lambda}(\hat{\Theta}) = I(\hat{\Theta})^{-1} \left\{\sum_i U_i(\hat{\Theta}) U_i(\hat{\Theta})^T\right\} I(\hat{\Theta})^{-1}, \tag{4.8}$$

em que U_i é a contribuição do i-ésimo conglomerado no vetor escore e I é menos a matriz de segundas derivadas de $\log L(\Theta)$. Expressões para a primeira e segunda derivadas de (4.5) podem ser obtidas em função das funções de incidência acumulada, F_k. Para isso, considerando $G_k(\Theta_k) = \{F_k(R_{ij}; \Theta_k) - F_k(L_{ij}; \Theta_k)\}$, em (4.6) temos que

$$\log l(y_i; \Theta) = \sum_{j=1}^{m_i} \sum_{k=1}^{K} \Delta_{ijk} \log G_k(\Theta_k) + \Delta_{ij0} \log \left(1 - \sum_{k=1}^{K} F_k(L_{ij}; \Theta_k)\right). \tag{4.9}$$

Então,

$$\frac{\partial \log l(y_i;\Theta)}{\partial \Theta_k} = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \frac{\Delta_{ijk}}{G_k(\Theta_k)}.\frac{\partial G_k(\Theta_k)}{\partial \theta_k} - \frac{\Delta_{ij0}}{1-\sum_{k=1}^{l} F_k(L_{ij};\Theta_k)}.\frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_k}, \quad (4.10)$$

em que,

$$\frac{\partial G_k(\Theta_k)}{\partial \Theta_k} = \frac{\partial F_k(R_{ij};\Theta_k)}{\partial \Theta_k} - \frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_k}.$$

De (4.10) temos que as derivadas de segunda ordem são:

$$\frac{\partial^2 \log l(y_i;\Theta)}{\partial \Theta_k^2} = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \Delta_{ijk}\left[\frac{\frac{\partial^2 G_k(\Theta_k)}{\partial \Theta_k^2}}{G_k(\Theta_k)} - \left(\frac{\frac{\partial^2 G_k(\Theta_k)}{\partial \Theta_k^2}}{G_k(\Theta_k)}\right)^2\right] - \Delta_{ij0}\left[\frac{\frac{\partial^2 F_k(L_{ij};\Theta_k)}{\partial \Theta_k^2}}{1-\sum_{k=1}^{K} F_k(L_{ij};\Theta_k)} - \left(\frac{\frac{\partial^2 F_k(L_{ij};\Theta_k)}{\partial \Theta_k^2}}{1-\sum_{k=1}^{K} F_k(L_{ij};\Theta_k)}\right)^2\right] \quad (4.11)$$

e

$$\frac{\partial^2 \log l(y_i;\Theta)}{\partial \Theta_k \Theta_{k'}} = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \Delta_{ijk}\left[\frac{\frac{\partial^2 G_k(\Theta_k)}{\partial \Theta_k \partial \Theta_{k'}}}{G_k(\Theta_k)} - \frac{\frac{\partial G_k(\Theta_k)}{\partial \Theta_k}\frac{\partial G_k(\Theta_k)}{\partial \Theta_{k'}}}{\{G_k(\Theta_k)\}^2}\right] - \Delta_{ij0}\left[\frac{\frac{\partial^2 F_k(L_{ij};\Theta_k)}{\partial \Theta_k \partial \Theta_{k'}}}{1-\sum_{k=1}^{K} F_k(L_{ij};\Theta_k)} - \frac{\frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_k}\frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_{k'}}}{\left\{1-\sum_{k=1}^{K} F_k(L_{ij};\Theta_k)\right\}^2}\right]. \quad (4.12)$$

Então

$$\frac{\partial \log L(\Theta)}{\partial \Theta_k} = \sum_{i=1}^{n} \frac{\partial \log l(y_i;\Theta)}{\partial \Theta_k}.$$

Na modelagem paramétrica, especificamos modelos paramétricos para as distribuições marginais, mas deixamos a natureza da dependência entre os membros do conglomerado completamente indefinida. Os parâmetros nos modelos marginais são então estimados usando a função de verossimilhança associada ao modelo que assume a independência dos membros (mesmo que esta suposição esteja incorreta), no caso presente é a Equação (4.5). Esta verossimilhança é o produto das verossimilhanças marginais de cada indivíduo no conjunto de dados, e então os erros padrão estimados são corrigidos usando a variância sanduíche (4.8).

Neste trabalho, modelamos a função de incidência acumulada usando a abordagem indireta, para a qual são considerados os modelos exponencial e Weibull para as distribuições marginais. Sem perda de generalidade, vamos considerar apenas duas causas de falha, $k = 1, 2$, a extensão para K causas de falha é imediata.

4.2.1 Modelo Exponencial

Considerando o modelo exponencial para a função de taxa de falha de causa-específica, temos que,

$$S_k(t;\alpha_k) = \exp(-\alpha_k t) \qquad \text{and} \qquad \lambda_k(t;\alpha_k) = \alpha_k \qquad \text{for } k = 1, 2.$$

A parametrização da função de incidência acumulada na Equação (4.1) então se reduz a

$$F_1(t;\Theta) = \int_0^t \alpha_1 \prod_{k=1}^{2} \exp(-\alpha_k t)du = \int_0^t \alpha_1 \exp\left(-\sum_{k=1}^{2}\alpha_k t\right) du = \frac{\alpha_1}{\alpha_1+\alpha_2}\left(1-\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right)$$

e

$$F_2(t;\Theta) = \int_0^t \alpha_2 \prod_{k=1}^{2} \exp(-\alpha_k t)du = \int_0^t \alpha_2 \exp\left(-\sum_{k=1}^{2}\alpha_k t\right) du = \frac{\alpha_2}{\alpha_1+\alpha_2}\left(1-\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right)$$

em que $\Theta = (\alpha_1, \alpha_2)$.

Na estimação dos parâmetros de interesse, propomos usar o log-verossimilhança (4.5) assumindo independência, tanto intra como entre conglomerados. Portanto, a log-verossimilhança para o modelo exponencial tem a forma,

$$\log L(\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\Delta_{ij1}\log\left[F_1(R_{ij};\Theta) - F_1(L_{ij};\Theta)\right] + \Delta_{ji2}\log\left[F_2(R_{ij};\Theta) - F_2(L_{ij};\Theta)\right] +$$
$$+ \Delta_{ij0}\log\left[1 - \sum_{k=1}^{2} F_k(L_{ij};\Theta)\right]\Bigg\}$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\Bigg\{\Delta_{ji1}\log\left[\frac{\alpha_1}{\alpha_1+\alpha_2}\left(\exp\left(-\sum_{k=1}^{2}\alpha_k L_{ij}\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_{ij}\right)\right)\right] +$$
$$+ \Delta_{ij2}\log\left[\frac{\alpha_2}{\alpha_1+\alpha_2}\left(\exp\left(-\sum_{k=1}^{2}\alpha_k L_{ij}\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_{ij}\right)\right)\right] +$$
$$+ \Delta_{ij0}\log\left[1 - \sum_{k=1}^{2} F_k(L_{ij};\Theta)\right]\Bigg\}. \tag{4.13}$$

Para maximizar $\log L(\Theta)$, pode-se usar, por exemplo, o algoritmo de Newton-Raphson. Para isso, precisamos da função score obtida calculando a derivada de $\log L(\Theta)$ em relação ao vetor de parâmetros Θ.

Expressões para as derivadas de primeira ordem de (4.13) são apresentadas a seguir. Observando a Equação 4.10 percebe-se que as derivadas da função log-verossimilhança dependem das derivadas de $F_1(t;\Theta)$ e $F_2(t;\Theta)$. Deste modo, as derivadas de primeira ordem de $F_1(t;\Theta)$ são,

$$\frac{\partial F_1(t;\Theta)}{\partial\alpha_1} = \frac{\alpha_2}{(\alpha_1+\alpha_2)^2}\left[1 - \exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right] + \frac{\alpha_1 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)$$

e

$$\frac{\partial F_1(t;\Theta)}{\partial\alpha_2} = -\frac{\alpha_1}{(\alpha_1+\alpha_2)^2}\left[1 + \exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right] + \frac{\alpha_1 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right).$$

As derivadas de segunda ordem de $F_1(t;\Theta)$ são,

$$\frac{\partial^2 F_1(t;\Theta)}{\partial\alpha_1^2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[2(\alpha_2^2+\alpha_1\alpha_2)t - (\alpha_1\alpha_2^2 + 2\alpha_1^2\alpha_2 + \alpha_1^3)t^2 + 2\alpha_2\right] - \frac{2\alpha_2}{(\alpha_1+\alpha_2)^3},$$

$$\frac{\partial^2 F_1(t;\Theta)}{\partial \alpha_2^2} = \frac{\alpha_1 \exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(-\alpha_1^2-\alpha_2^2-2\alpha_1\alpha_2)t^2-2(\alpha_1+\alpha_2)t-2\right]+\frac{2\alpha_1}{(\alpha_1+\alpha_2)^3}$$

e

$$\frac{\partial^2 F_1(t;\Theta)}{\partial \alpha_1 \partial \alpha_2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(\alpha_2^2-\alpha_1^2)t-(\alpha_1\alpha_2^2+2\alpha_1^2\alpha_2+\alpha_1^3)t^2-\alpha_1+\alpha_2\right]+\frac{\alpha_1-\alpha_2}{(\alpha_1+\alpha_2)^3}.$$

As derivadas de primeira ordem de $F_2(t;\Theta)$ são,

$$\frac{\partial F_2(t;\Theta)}{\partial \alpha_2} = \frac{\alpha_1}{(\alpha_1+\alpha_2)^2}\left[1-\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right]+\frac{\alpha_2 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)$$

e

$$\frac{\partial F_2(t;\Theta)}{\partial \alpha_1} = -\frac{\alpha_2}{(\alpha_1+\alpha_2)^2}\left[1+\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right]+\frac{\alpha_2 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right).$$

As derivadas de segunda ordem de $F_2(t;\Theta)$ são,

$$\frac{\partial^2 F_2(t;\Theta)}{\partial \alpha_2^2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[2(\alpha_1^2+\alpha_1\alpha_2)t-(\alpha_2\alpha_1^2+2\alpha_2^2\alpha_1+\alpha_2^3)t^2-2\alpha_1\right]-\frac{2\alpha_1}{(\alpha_1+\alpha_2)^3},$$

$$\frac{\partial^2 F_2(t;\Theta)}{\partial \alpha_1^2} = \frac{\alpha_2 \exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(\alpha_2^2-\alpha_1^2-2\alpha_1\alpha_2)t^2-2(\alpha_1+\alpha_2)t-2\right]+\frac{2\alpha_2}{(\alpha_1+\alpha_2)^3}$$

e

$$\frac{\partial^2 F_2(t;\Theta)}{\partial \alpha_1 \partial \alpha_2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(\alpha_1^2-\alpha_2^2)t+(-\alpha_2\alpha_1^2-2\alpha_2^2\alpha_1+\alpha_2^3)t^2-\alpha_2+\alpha_1\right]+\frac{\alpha_2-\alpha_1}{(\alpha_1+\alpha_2)^3}.$$

Portanto, as derivadas de primeira ordem da função de log-verossimilhança (4.13) são

$$\frac{\partial \log L(\Theta)}{\partial \alpha_1} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\Delta_{j1}\left[\frac{R_j \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right) - L_j \exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)} + \frac{\alpha_2}{\alpha_1(\alpha_1+\alpha_2)}\right] + \right.$$
$$\Delta_{j2}\left[\frac{R_j \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right) - L_j \exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)} + \frac{1}{(\alpha_1+\alpha_2)}\right] -$$
$$\left.\Delta_{j0}\left[\frac{1}{1-\sum_{k=1}^{2}F_k(L_j;\Theta)}\left(L_j \exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right) - \frac{2\alpha_2}{(\alpha_1+\alpha_2)^2}\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)\right)\right]\right\}$$

e

$$\frac{\partial \log L(\Theta)}{\partial \alpha_2} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\Delta_{j1}\left[\frac{R_j \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right) - L_j \exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)} + \frac{1}{(\alpha_1+\alpha_2)}\right] + \right.$$
$$\Delta_{j2}\left[\frac{R_j \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right) - L_j \exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)} + \frac{\alpha_1}{\alpha_2(\alpha_1+\alpha_2)}\right] -$$
$$\left.\Delta_{j0}\left[\frac{1}{1-\sum_{k=1}^{2}F_k(L_j;\Theta)}\left(L_j \exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right) - \frac{2\alpha_1}{(\alpha_1+\alpha_2)^2}\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)\right)\right]\right\}.$$

De maneira similar pode-se encontrar as expressões para as derivadas de segunda ordem da função de log-verossimilhança (4.13).

4.2.2 Modelo Weibull

Outra opção é modelar a função de incidência acumulada via distribuição Weibull para a função de taxa de falha de causa-específica. Nessa parametrização temos que

$$S_k(t;\alpha_k,\tau_k)=\exp[-(\alpha_k t)^{\tau_k}] \qquad \text{e} \qquad \lambda_k(t;\alpha_k,\tau_k)=\tau_k\alpha_k^{\tau_k}t^{\tau_k-1} \qquad k=1,2,$$

em que $\alpha_k>0$ é o parâmetro de escala e $\tau_k>0$ é o parâmetro de forma.

Neste caso, a versão parametrizada da função de incidência acumulada (4.1), com $k=1,2$, tem a forma

$$F_1(t;\Theta)=\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1}-(\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du$$

e

$$F_2(t;\Theta)=\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1}-(\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du,$$

em que $\Theta=(\alpha_1,\tau_1,\alpha_2,\tau_2)$. Essas integrais não possuem soluções analíticas, necessitando, portanto, de métodos numéricos.

Novamente, na estimativa dos parâmetros de interesse, propomos usar a log-verossimilhança (4.5) e assumindo tanto intra como entre a independência dos conglomerados e portanto, a log-verossimilhança para o modelo Weibull tem a forma,

$$\begin{aligned}
\log L(\Theta)=&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\Delta_{ij1}\log\left[F_1(R_{ij};\Theta)-F_1(L_{ij};\Theta)\right]+\Delta_{ij2}\log\left[F_2(R_{ij};\Theta)-F_2(L_{ij};\Theta)\right]+\\
&+\Delta_{ij0}\log\left[1-\sum_{k=1}^{2}F_k(L_{ij};\Theta)\right]\Bigg\}\\
=&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\Bigg\{\Delta_{ij1}\log\Bigg[\int_0^{R_{ij}}\exp\left[-(\alpha_1 u)^{\tau_1}-(\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du-\\
&-\int_0^{L_{ij}}\exp\left[-(\alpha_1 u)^{\tau_1}-(\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du\Bigg]+\\
&+\Delta_{ij2}\log\Bigg[\int_0^{R_{ij}}\exp\left[-(\alpha_1 u)^{\tau_1}-(\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du-\\
&-\int_0^{L_{ij}}\exp\left[-(\alpha_1 u)^{\tau_1}-(\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du\Bigg]+\\
&+\Delta_{ij0}\log\left[1-\sum_{k=1}^{2}F_k(L_{ij};\Theta)\right]\Bigg\}. \qquad (4.14)
\end{aligned}$$

Observe que as estimativas dos parâmetros desconhecidos não podem ser obtidas de forma fechada sendo necessária uma técnica numérica para computar essas estimativas. Pode-se usar os métodos de Newton-Raphson ou Gauss-Newton ou suas variantes para maximizar a Equação (4.14). Expressões para primeira e segunda derivadas de (4.14) dependem das derivadas de $F_1(t;\Theta)$ e $F_2(t;\Theta)$. Portanto, as derivadas de primeira ordem, em relação a α_k e τ_k são:

$$\frac{\partial F_1(t;\Theta)}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du,$$
$$\frac{\partial F_1(t;\Theta)}{\partial \tau_k} = \frac{\partial}{\partial \tau_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du,$$
$$\frac{\partial F_2(t;\Theta)}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du$$

e

$$\frac{\partial F_2(t;\Theta)}{\partial \tau_k} = \frac{\partial}{\partial \tau_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du$$

4.2.3 Modelo de Regressão

Para examinar os efeitos das covariáveis na resposta é necessário incluir uma estrutura de regressão. Na análise de sobrevivência, a heterogeneidade entre os indivíduos é explicada por covariáveis ou variáveis explicativas. Para construir o modelo de regressão introduzimos a covariável $\mathbf{x}$ assumindo que o parâmetro α_k depende das covariáveis por meio de $\alpha_k = \exp(\boldsymbol{\beta}_k^T\mathbf{x})$, na log-verossimilhança $\log L(\Theta)$, em que $\mathbf{x} = (X_1, X_2, ..., X_p)$ e $\boldsymbol{\beta}_k^T = (\beta_{k1}, \beta_{k2}, ..., \beta_{kp})$ são os vetores de coeficientes da regressão. Todos os parâmetros são incluídos simultaneamente no log-verossimilhança (4.13), para o modelo exponencial, e (4.14), para o modelo Weibull. Então, para obter o MLE de parâmetros desconhecidos, o sistema de equações normais foi derivado.

Para o modelo exponencial, o sistema de equações é

$$\frac{\partial \log L(\Theta)}{\partial \boldsymbol{\beta}_k} = 0,$$

e para o modelo de Weibull é

$$\frac{\partial \log L(\Theta)}{\partial \boldsymbol{\beta}_k} = 0 \qquad \text{e} \qquad \frac{\partial \log L(\Theta)}{\partial \tau_k} = 0.$$

As funções escore são subsequentemente definidas como zero e resolvidas usando um procedimento iterativo, como o algoritmo de Newton Raphson, para encontrar os estimadores de máxima verossimilhança.

4.3 Estudo de Simulação

Geralmente, o método da inversão pode ser aplicado para geração de números aleatórios considerando uma distribuição pré-definida. Para a aplicação do método da inversão é necessário um gerador de números aleatórios uniformemente distribuídos no intervalo de zero a um. Para dados de sobrevivência, a distribuição é definida principalmente por sua função taxa de falha ou função taxa de falha acumulada. Assim, os tempos de evento podem ser gerados resolvendo a equação

$$U = F(T|\mathbf{x}) \tag{4.15}$$

ou

$$U = S(T|\mathbf{x}), \tag{4.16}$$

em que $\mathbf{x}$ é um vetor de covariáveis e U é um número aleatório com distribuição $U[0,1]$.

Definindo a distribuição do tempo do evento pela função taxa de falha, a Equação (4.16) pode ser escrita como,

$$U = \exp[-\Lambda(T|\mathbf{x})] = \exp\left[-\int_0^T \lambda(s|\mathbf{x})ds\right], \tag{4.17}$$

em que $\lambda(T|\mathbf{x})$ é a função de taxa de falha e $\Lambda(T|\mathbf{x})$ é a função de taxa de falha acumulada.

Para a distribuição exponencial, considerando a função de taxa de falha $\lambda(t|\mathbf{x}) = \exp(\boldsymbol{\beta}^T\mathbf{x})$, os tempos podem ser gerados utilizando a Equação (4.17) com

$$T = -\frac{\ln(U)}{\exp(\boldsymbol{\beta}^T\mathbf{x})}, \tag{4.18}$$

e para o caso da distribuição Weibull, com função de taxa de falha $\lambda(t|\mathbf{x}) = \tau t^{\tau-1}[\exp(\boldsymbol{\beta}^T\mathbf{x})]^\tau$, os tempos são gerados com,

$$T = \frac{(-\ln(U))^{\frac{1}{\tau}}}{\exp(\boldsymbol{\beta}^T\mathbf{x})}. \tag{4.19}$$

Diferentes métodos para simulação de dados na presença de riscos competitivos foram usados na literatura para avaliar e comparar métodos estatísticos. Beyersmann et al. (2009) recomendam o uso da taxa de falha de causa-específica para a simulação de dados com riscos competitivos, pois a taxa de falha de causa-específica determina completamente o processo de riscos competitivos. Eles apresentam um algorítmo para a geração dos dados usando funções de taxa de falha de causa-específica predefinidas. Como para cada indivíduo a taxa de falha geral em cada ponto do tempo é a soma das taxas de falhas de causa-específica para todos os K eventos possíveis, o tempo do evento é gerado a partir de uma distribuição com taxa de falha $\lambda(t|\mathbf{x}) = \sum_{k=1}^{K} \lambda_k(t|\mathbf{x})$ no primeiro passo. Então, o tipo de evento é determinado por um experimento de Bernoulli com as probabilidades para cada tipo de evento k proporcional às taxas de falha de causa-específica $\lambda_k(t|\mathbf{x})$.

Portanto, dados de riscos competitivos com taxas de falha de causa-específica predefinidos, podendo depender de covariáveis, podem ser gerados conforme algoritmo a seguir. O algoritmo é apresentado para dois tipos de eventos, isto é, $K = 2$, podendo ser facilmente adaptado para $K > 2$:

1. Defina as taxas de falha causa-específica $\lambda_1(t|\mathbf{x})$ e $\lambda_2(t|\mathbf{x})$ para ambos os eventos;

2. Gere tempos de falha T com taxa de falha total $\lambda(t|\mathbf{x}) = \lambda_1(t|\mathbf{x}) + \lambda_2(t|\mathbf{x})$, utilizando a Equação (4.17);

3. Execute um experimento binomial para o tempo de falha simulado T com probabilidades $p_1 = \dfrac{\lambda_1(t|\mathbf{x})}{\lambda_1(t|\mathbf{x}) + \lambda_2(t|\mathbf{x})}$ para um evento do tipo $k = 1$ e $p_2 = \dfrac{\lambda_2(t|\mathbf{x})}{\lambda_1(t|\mathbf{x}) + \lambda_2(t|\mathbf{x})}$ para um evento do tipo $k = 2$;

4. Gere os tempos de censura C.

O método da inversão, apresentado anteriormente, pode ser aplicado para qualquer distribuição válida definida pela função de taxa de risco total, possivelmente usando métodos numéricos para o cálculo da função de taxa de falha total acumulada e para a solução da Equação (4.17), conforme é ilustrado em Beyersmann et al. (2011).

Um estudo de Monte Carlo foi realizado usando a linguagem R (R Core Team, 2022) para avaliar o desempenho de amostras finitas da metodologia proposta. A maximização da função log-verossimilhança (4.3) é realizada usando o algoritmo de otimização não linear quase-Newton BFGS implementado na função otimizada *optim* disponível em R. Diferentes cenários são considerados e para todos eles há $K = 2$ causas de falha, $n \in \{100, 500, 1000\}$, $m_i = 4$, $i = 1, ..., n$ e covariáveis X_1 e X_2 são gerados independentemente de Bernoulli (0.5) e distribuição normal padrão, respectivamente. O tempo de falha e os tipos de causa são simulados de acordo com os parâmetros $\Theta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0.3, 0.3, 0.3, 0.3)$, para o modelo exponencial

$$\lambda_k(t_i; \mathbf{x}_i) = \exp(\beta_{k1} X_{1i} + \beta_{k2} X_{2i}) \tag{4.20}$$

e $\Theta = (\beta_{11}, \beta_{12}, \tau_1, \beta_{21}, \beta_{22}, \tau_2) = (0.1, 0.1, 2.0, 0.1, 0.1, 2.0)$, para o modelo Weibull

$$\lambda_k(t_i; \mathbf{x}_i) = \tau_k t^{\tau_k - 1} [\exp(\beta_{k1} X_{1i} + \beta_{k2} X_{2i})]^{\tau_k}, \tag{4.21}$$

em que τ_k é o parâmetro de forma, β_{k1} e β_{k2} são os efeitos covariáveis, $i = 1, ..., n$ e $k = 1, 2$.

A geração de dados pode ser dividida nas seguintes etapas:

1 Um termo aleatório comum para cada conglomerado, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, é gerado a partir de uma distribuição gama, $\omega_{ik} \sim Gamma(\theta_k, \theta_k)$, de acordo com os passos de Brazauskas e Le-Rademacher (2016). θ_k é escolhido para obter a estrutura de dependência dentro do conglomerado em cada cenário.

2 Para cada indivíduo, covariáveis X_1 e X_2 são geradas independentemente a partir de uma distribuição Bernoulli $(0, 5)$ e normal padrão, respectivamente.

3 Geramos dados de riscos competitivos por meio do seguinte algoritmo, de acordo com Beyersmann et al. (2011):

   3.1 Especifique as funções de taxa de falha de causa-específica $\lambda_1(t; \mathbf{x})$ e $\lambda_2(t; \mathbf{x})$, para modelos exponenciais e Weibull, de acordo com as equações (4.13) e (4.14), respectivamente.

3.2 Simule tempos de falha T com taxa de falha de todas as causas $\lambda(t;\mathbf{x}) = \lambda_1(t;\mathbf{x}) + \lambda_2(t;\mathbf{x})$.

3.3 Determine o tipo de evento para cada indivíduo usando um experimento binomial para um tempo de falha simulado T, que decide com probabilidade $\dfrac{\lambda_k(t;\mathbf{x})}{\lambda_1(t;\mathbf{x}) + \lambda_2(t;\mathbf{x})}$ na causa k, $k = 1, 2$.

3.4 Gere tempos de censura à direita independentes $C \sim U[0, 1]$.

3.5 Os tempos observados são considerados o mínimo entre os tempos de falha T e os tempos censurados C, $t_{ij} = min(T_{ij}, C_{ij})$.

4 Os tempos de censura intervalar (L_{ij} e R_{ij}) são construídos conforme Santos Junior (2016), da seguinte forma:

4.1 Para todos os indivíduos censurados à direita, aplicamos os seguintes intervalos de censura

$$L_{ij} = t_{ij} \text{ and } R_{ij} = \infty;$$

4.2 Para indivíduos que falharam por uma das causas K, aplicamos os seguintes tempos de censura de intervalo:

$$L_{ij} = 0.0001 \text{ and } R_{ij} = c,$$

em que c é gerado a partir de uma distribuição uniforme $U[0.1; b]$. O valor b é escolhido de forma a aumentar ou diminuir o tamanho dos intervalos. Os tempos de vida t_{ij} são comparados com os intervalos criados, para verificar se t_{ij} pertence ao intervalo. Se t_{ij} não pertencer ao intervalo, novos intervalos são criados de modo que,

$$L_{ij} = R_{ij} \text{ and } R_{ij} = R_{ij} + c,$$

em que c é um novo valor gerado a partir de uma distribuição uniforme $U[0.1; b]$.

Os dados simulados foram gerados em três cenários. No cenário 1, foi considerada uma estrutura de correlação intra-conglomerado pequena, *tau de Kendall* igual a 0,20,

para isso os efeitos aleatórios específicos do conglomerado, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, são gerados a partir da distribuição gama, $\omega_{ik} \sim Gamma(2, 2)$, uma vez que *tau de Kendall* $= \frac{\theta}{\theta+2}$ para a distribuição $Gamma(\theta, \theta)$ em que $\theta > 0$ (Emura et al., 2019, p.22). Além disso, o comprimento dos intervalos censurados são gerados de uma distribuição uniforme $U(0.1, 0.3)$. No cenário 2, pretende-se verificar a influência do comprimento dos intervalos cencurados e, para isso, mantem-se a mesma extrutura do cenário 1 aumentando apenas o comprimento dos interalos censurados, que são gerados de uma distribuição uniforme $U(0.1, 0.5)$. No cenário 3, pretende-se verificar a influência da estrutura de correlação intra-conglomerados. Para isso mantem-se a mesma estrutura do cenário 1, aumentando apenas a correlação intra-conglomerado, *tau de Kendall* igual a 0,85, consequentemente os efeitos aleatórios específicos do conglomerado, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, são gerados a partir da distribuição gama, $\omega_{ik} \sim Gamma(1/11, 1/11)$.

Cada conjunto de dados é analisado usando a metodologia proposta na Seção 4.2.3. Cada resultado de simulação exibido nesta seção é baseado em conjuntos de dados simulados de tamanho $N = 1000$. Os desempenhos de todos os estimadores pontuais são comparados numericamente em termos de estimativa média, variância empírica, variância assintótica e valores de erro quadrático médio (MSE).

A média das estimativas dos parâmetros é calculada como $\hat{\theta} = \frac{\sum_{i=1}^{N} \hat{\theta}_i}{N}$, fornecendo uma estimativa do valor esperado das estimativas dos coeficientes para as N simulações. O viés relativo (rb) é estimado tomando a diferença entre $\hat{\theta}$ e θ, o verdadeiro valor do parâmetro populacional, dividido por e θ, tal que $rb = \frac{\hat{\theta} - \theta}{|\theta|}$.

Além disso, a probabilidade de cobertura (CP_1) de 95% é calculada para estimativas de intervalo considerando o erro padrão SE_1 e a probabilidade de cobertura (CP_2) de 95% para estimativas de intervalo considerando o estimador sanduíche do erro padrão SE_2.

A variância empírica é calculada por $Var(\hat{\theta}) = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \hat{\theta})^2}{N-1}$ e o erro quadrático médio é calculado por $MSE(\theta) = \frac{\sum_{i=1}^{N}(\hat{\theta} - \theta)^2}{N}$. Utilizando a variância empírica pode-se calcular o desvio-padrão empírico, SD.

O erro padrão (SE_1) foi calculado como menos o inverso da matriz de segundas

derivadas de $\log L(\Theta)$ e o estimador sanduíche de erro padrão (SE_2) foi calculado usando a variância sanduíche (4.8) com base na média dos resultados das N simulações.

4.3.1 Resultados

Os resultados do estudo de simulação são apresentados nas Tabelas 4.1 a 4.3 e Figuras 4.1 a 4.3, para o modelo exponencial, e nas Tabelas 4.4 a 4.6 e Figuras 4.4 a 4.6, para o modelo Weibull.

Tabela 4.1 – Cenário I para modelo exponencial - Neste cenário $\omega_{ik} \sim Gamma(2,2)$, causando assim uma fraca correlação entre os indivíduos dentro de um mesmo conglomerado (tau de Kendall = 0,20). O tamanho dos intervalos gerados a partir de $c \sim U(1, 0.1, 0.3)$.

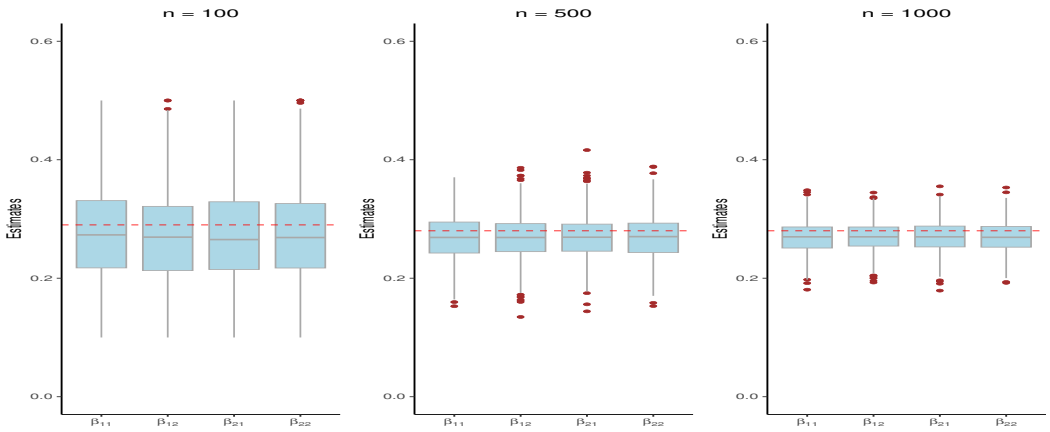
	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
100 conglomerados								
$\beta_{11} = 0.3$	0.276	0.008	-0.081	0.084	0.075	0.079	0.915	0.925
$\beta_{12} = 0.3$	0.271	0.008	-0.097	0.084	0.075	0.080	0.895	0.913
$\beta_{21} = 0.3$	0.271	0.008	-0.097	0.082	0.075	0.080	0.917	0.931
$\beta_{22} = 0.3$	0.270	0.007	-0.099	0.081	0.075	0.080	0.908	0.929
500 conglomerados								
$\beta_{11} = 0.3$	0.281	0.005	-0.064	0.058	0.053	0.056	0.921	0.926
$\beta_{12} = 0.3$	0.281	0.005	-0.064	0.056	0.053	0.056	0.915	0.937
$\beta_{21} = 0.3$	0.281	0.005	-0.064	0.057	0.053	0.056	0.913	0.939
$\beta_{22} = 0.3$	0.281	0.005	-0.064	0.057	0.053	0.056	0.911	0.934
1000 conglomerados								
$\beta_{11} = 0.3$	0.289	0.003	-0.036	0.037	0.033	0.036	0.921	0.946
$\beta_{12} = 0.3$	0.289	0.003	-0.036	0.036	0.033	0.036	0.925	0.940
$\beta_{21} = 0.3$	0.289	0.003	-0.036	0.037	0.033	0.036	0.923	0.949
$\beta_{22} = 0.3$	0.289	0.003	-0.036	0.037	0.033	0.036	0.921	0.946

Tabela 4.2 – Cenário II para modelo exponencial - mesma estrutura do Cenário I, exceto que aumentamos o tamanho dos intervalos, usando $c \sim U(1, 0.1, 0.5)$.

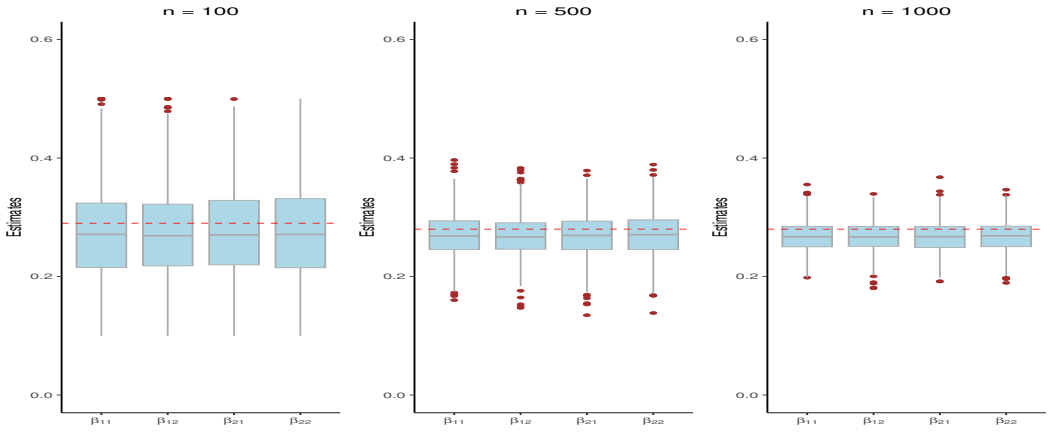
	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
100 conglomerados								
$\beta_{11} = 0.3$	0.269	0.009	-0.104	0.087	0.075	0.080	0.892	0.906
$\beta_{12} = 0.3$	0.268	0.009	-0.106	0.084	0.075	0.080	0.912	0.927
$\beta_{21} = 0.3$	0.268	0.009	-0.108	0.085	0.075	0.080	0.887	0.906
$\beta_{22} = 0.3$	0.268	0.009	-0.107	0.081	0.075	0.080	0.906	0.934
500 conglomerados								
$\beta_{11} = 0.3$	0.278	0.006	-0.073	0.066	0.063	0.066	0.906	0.922
$\beta_{12} = 0.3$	0.278	0.006	-0.073	0.065	0.063	0.064	0.902	0.919
$\beta_{21} = 0.3$	0.278	0.006	-0.073	0.065	0.063	0.064	0.918	0.929
$\beta_{22} = 0.3$	0.278	0.006	-0.073	0.064	0.063	0.063	0.907	0.908
1000 conglomerados								
$\beta_{11} = 0.3$	0.285	0.004	-0.049	0.046	0.043	0.046	0.926	0.942
$\beta_{12} = 0.3$	0.285	0.004	-0.049	0.045	0.043	0.044	0.922	0.949
$\beta_{21} = 0.3$	0.285	0.004	-0.049	0.045	0.043	0.044	0.928	0.949
$\beta_{22} = 0.3$	0.285	0.004	-0.049	0.044	0.043	0.044	0.927	0.948

Tabela 4.3 – Cenário III para modelo exponencial - mesma estrutura do Cenário I, exceto que aumentamos a correlação intra-conglomerado (tau de Kendall = 0,85) usando $\omega_{ik} \sim Gamma(1/11, 1/11)$.

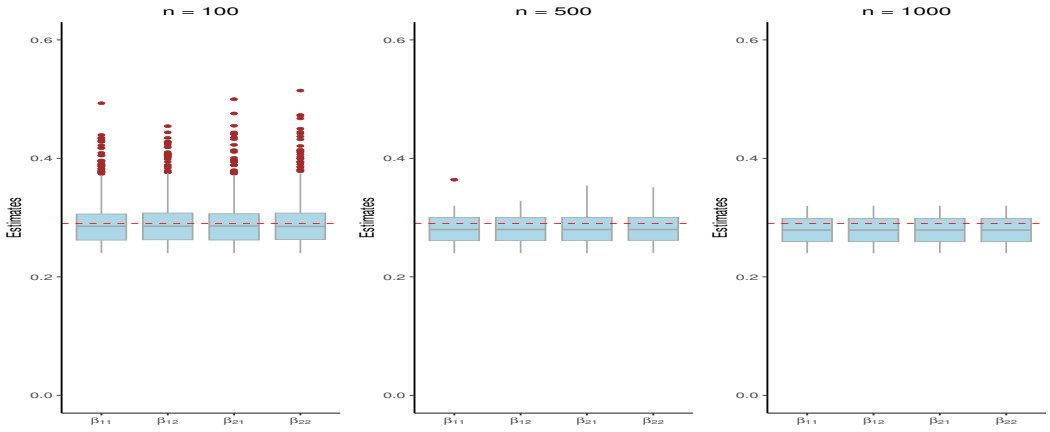
	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
100 conglomerados								
$\beta_{11} = 0.3$	0.269	0.010	-0.103	0.086	0.076	0.087	0.919	0.926
$\beta_{12} = 0.3$	0.269	0.010	-0.103	0.086	0.075	0.085	0.912	0.914
$\beta_{21} = 0.3$	0.269	0.010	-0.103	0.086	0.076	0.085	0.911	0.935
$\beta_{22} = 0.3$	0.269	0.010	-0.103	0.086	0.075	0.087	0.912	0.915
500 conglomerados								
$\beta_{11} = 0.3$	0.276	0.008	-0.081	0.058	0.050	0.056	0.901	0.925
$\beta_{12} = 0.3$	0.276	0.008	-0.081	0.058	0.051	0.056	0.901	0.929
$\beta_{21} = 0.3$	0.276	0.008	-0.081	0.058	0.051	0.056	0.911	0.937
$\beta_{22} = 0.3$	0.276	0.008	-0.081	0.058	0.050	0.056	0.911	0.931
1000 conglomerados								
$\beta_{11} = 0.3$	0.287	0.006	-0.043	0.038	0.031	0.037	0.931	0.944
$\beta_{12} = 0.3$	0.287	0.006	-0.043	0.038	0.032	0.037	0.931	0.948
$\beta_{21} = 0.3$	0.287	0.005	-0.043	0.038	0.032	0.037	0.931	0.949
$\beta_{22} = 0.3$	0.287	0.006	-0.043	0.038	0.031	0.037	0.931	0.947



(a) Cenário I

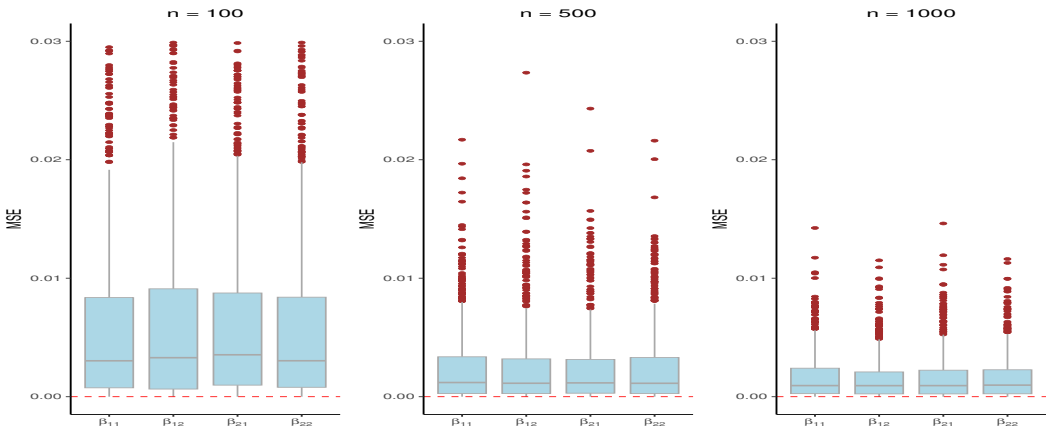


(b) Cenário II

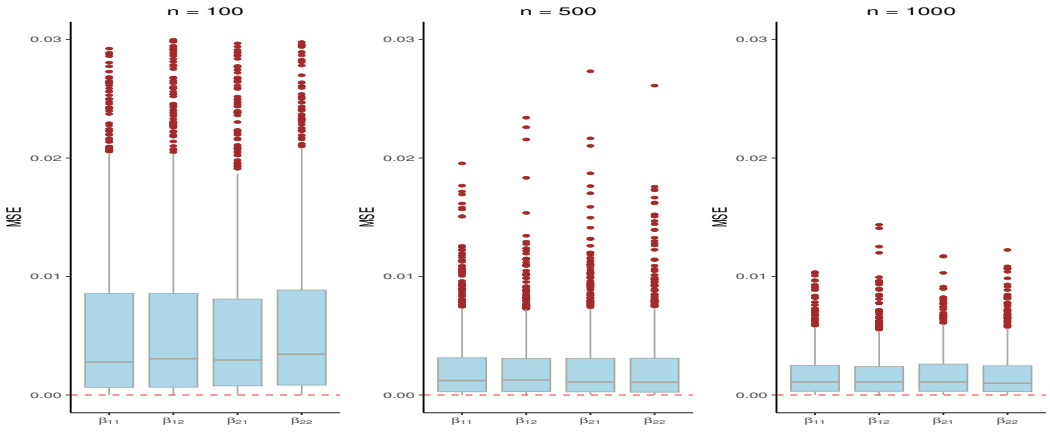


(c) Cenário III

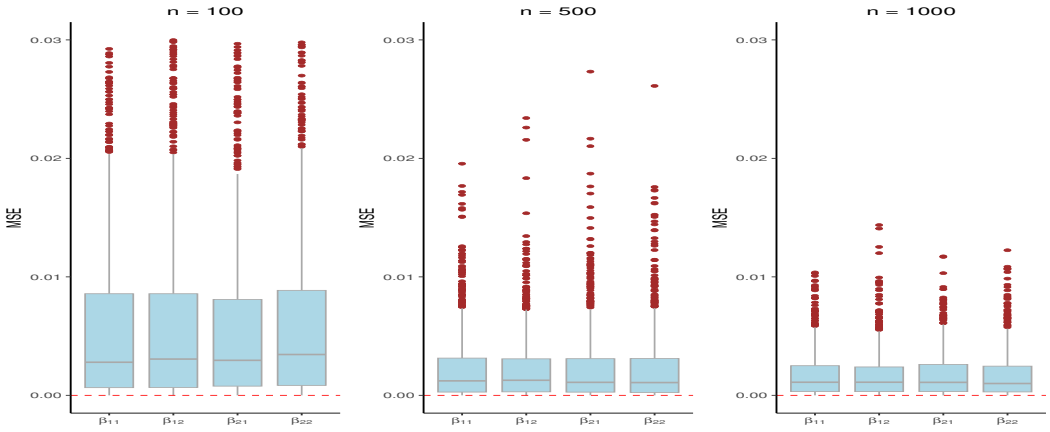
Figura 4.1 – Estimativas dos parâmetros do modelo Exponencial



(a) Cenário I

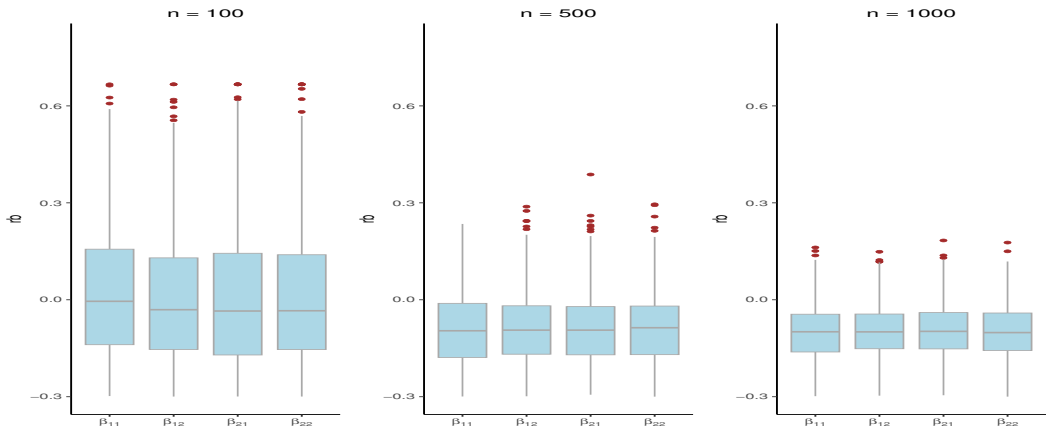


(b) Cenário II

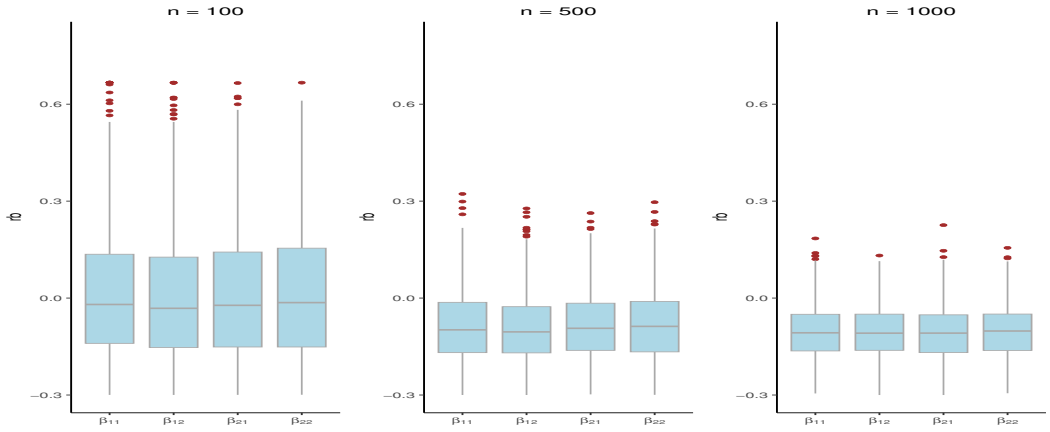


(c) Cenário III

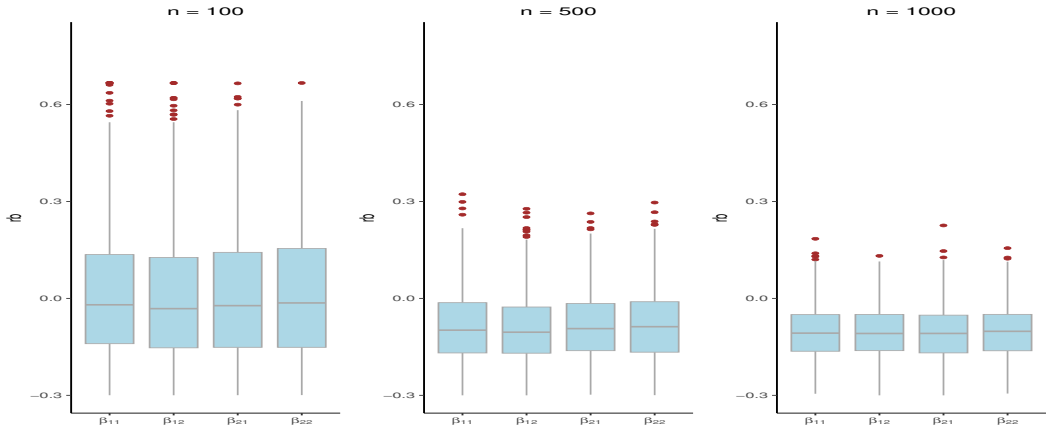
Figura 4.2 – Estimativas de erro quadrático médio (MSE) dos parâmetros do modelo Exponencial



(a) Cenário I



(b) Cenário II



(c) Cenário III

Figura 4.3 – Estimativas de viés relativo (rb) dos parâmetros do modelo Exponencial

De acordo com a Tabela 4.1 (modelo Exponencial) e a Tabela 4.4 (modelo Weibull), as estimativas médias estão próximas do valor real dos parâmetros e conforme o número de conglomerados aumenta, o MSE e rb diminuem. Por outro lado, as estimativas dos parâmetros de forma da Weibull são viesadas e apresentam taxa de convergência muito lenta. Além disso, o erro padrão sanduíche (SE_2) é próximo ao erro padrão (SD) empírico, indicando que a abordagem proposta efetivamente faz uma correção para a correlação do conglomerado. As probabilidades de cobertura (CP_2) estão próximas do nível nominal de 0,95, indicando uma boa aproximação das estimativas para a distribuição normal.

Tentamos avaliar o efeito do tamanho dos intervalos. Dessa forma, geramos diferentes cenários com tamanhos de intervalos maiores. Os resultados são apresentados na Tabela 4.2 e Tabela 4.5 , para o modelo exponencial e modelo Weibull, respectivamente. As estimativas médias estão próximas do valor real, mas o MSE, rb e SD são maiores quando comparados com o primeiro cenário. Além disso, o erro padrão sanduíche (SE_2) está muito próximo do erro padrão (SD) empírico e as probabilidades de cobertura estão próximas do nível nominal de 0,95. Observe que quando aumentamos a correlação dentro do conglomerado, as estimativas dos parâmetros têm um viés um pouco maior, como pode ser observado na Tabela 4.3 e na Tabela 4.6. Observa-se também que o erro padrão empírico é um pouco maior, mas o erro padrão sanduíche ainda é perto do erro padrão empírico. Em todos os cenários, o aumento do número de conglomerados leva a estimativas menos viesadas e intervalos de cobertura mais próximos do nível nominal.

Tabela 4.4 – Cenário I para o modelo Weibull - Neste cenário $\omega_{ik} \sim Gamma(2,2)$, causando assim uma correlação fraca entre os indivíduos do conglomerado (tau de Kendall = 0,20). O tamanho dos intervalos gerado a partir de $c \sim U(1, 0.1, 0.3)$.

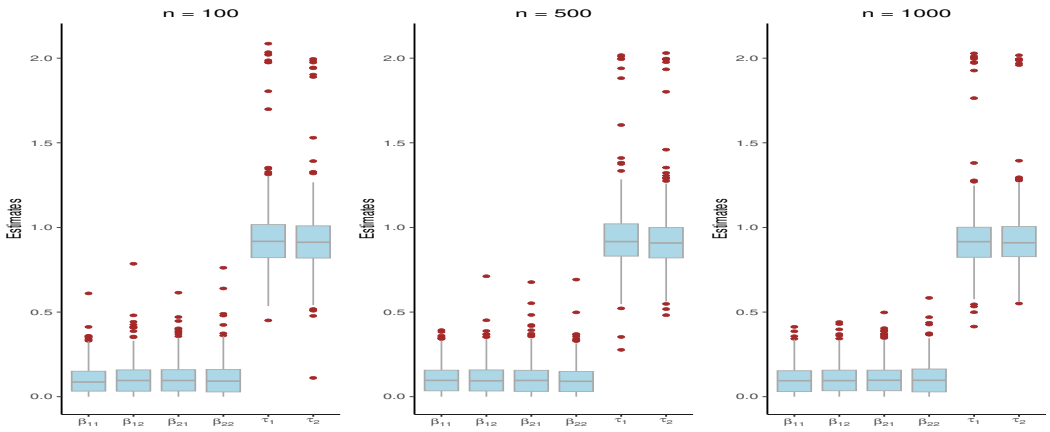
	Est	MSE	rb	SD	SE1	SE2	CP1	CP2
100 conglomerados								
$\beta_{11} = 0.1$	0.099	0.007	-0.008	0.092	0.097	0.095	0.976	0.997
$\beta_{12} = 0.1$	0.105	0.008	0.049	0.091	0.097	0.091	0.976	0.998
$\beta_{21} = 0.1$	0.103	0.007	0.034	0.094	0.098	0.096	0.976	0.994
$\beta_{22} = 0.1$	0.105	0.008	0.048	0.093	0.098	0.097	0.981	0.996
$\tau_1 = 2.0$	0.930	1.179	-0.535	0.181	0.165	0.168	0.032	0.032
$\tau_2 = 2.0$	0.924	1.189	-0.538	0.175	0.179	0.179	0.026	0.026
500 conglomerados								
$\beta_{11} = 0.1$	0.102	0.006	0.030	0.085	0.096	0.089	0.980	0.999
$\beta_{12} = 0.1$	0.102	0.006	0.023	0.088	0.097	0.095	0.982	0.999
$\beta_{21} = 0.1$	0.103	0.006	0.032	0.089	0.098	0.091	0.978	0.999
$\beta_{22} = 0.1$	0.100	0.006	-0.001	0.088	0.097	0.090	0.981	0.999
$\tau_1 = 2.0$	0.950	1.154	-0.525	0.167	0.126	0.157	0.455	0.468
$\tau_2 = 2.0$	0.941	1.162	-0.529	0.171	0.174	0.172	0.770	0.770
1000 conglomerados								
$\beta_{11} = 0.1$	0.101	0.005	0.030	0.082	0.096	0.084	0.980	0.999
$\beta_{12} = 0.1$	0.101	0.005	0.018	0.084	0.094	0.086	0.982	0.999
$\beta_{21} = 0.1$	0.102	0.005	0.028	0.087	0.095	0.086	0.978	0.999
$\beta_{22} = 0.1$	0.100	0.004	-0.001	0.083	0.092	0.085	0.981	0.999
$\tau_1 = 2.0$	0.980	1.141	-0.515	0.147	0.106	0.137	0.555	0.568
$\tau_2 = 2.0$	0.972	1.131	-0.519	0.141	0.148	0.142	0.770	0.770

Tabela 4.5 – Cenário II para o modelo Weibull - mesma estrutura do Cenário I, exceto que aumentamos o tamanho dos intervalos, usando $c \sim U(1, 0.1, 0.5)$.

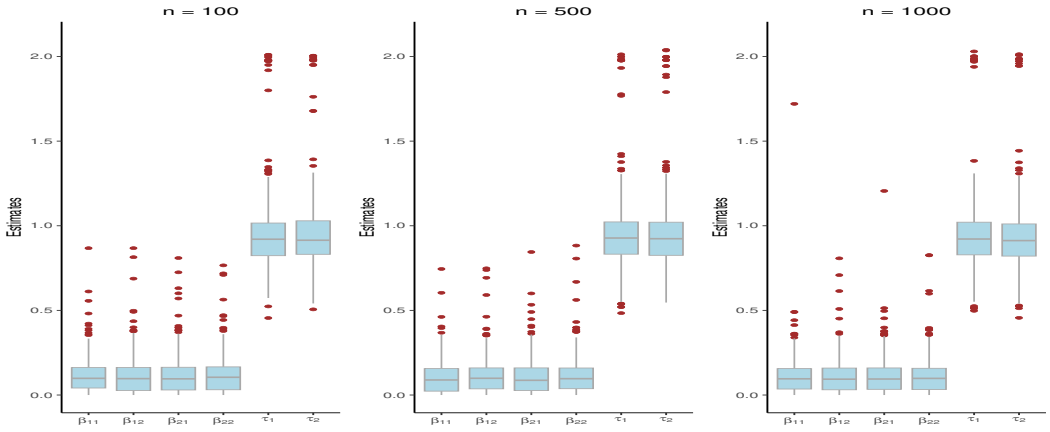
	Est	MSE	rb	SD	SE1	SE2	CP1	CP2
100 conglomerados								
$\beta_{11} = 0.1$	0.108	0.008	0.077	0.098	0.097	0.098	0.973	0.996
$\beta_{12} = 0.1$	0.107	0.009	0.068	0.097	0.096	0.097	0.965	0.992
$\beta_{21} = 0.1$	0.108	0.009	0.076	0.096	0.096	0.096	0.973	0.996
$\beta_{22} = 0.1$	0.109	0.008	0.087	0.094	0.096	0.094	0.974	0.995
$\tau_1 = 2.0$	0.940	1.165	-0.530	0.205	0.156	0.159	0.408	0.440
$\tau_2 = 2.0$	0.942	1.158	-0.529	0.197	1.115	1.015	0.433	0.433
500 conglomerados								
$\beta_{11} = 0.1$	0.101	0.008	0.007	0.088	0.096	0.089	0.966	0.996
$\beta_{12} = 0.1$	0.105	0.007	0.089	0.092	0.095	0.094	0.972	0.996
$\beta_{21} = 0.1$	0.101	0.008	0.014	0.093	0.095	0.094	0.977	0.998
$\beta_{22} = 0.1$	0.105	0.007	0.085	0.094	0.096	0.094	0.976	0.996
$\tau_1 = 2.0$	0.962	1.144	-0.519	0.188	0.125	0.153	0.404	0.470
$\tau_2 = 2.0$	0.959	1.131	-0.510	0.190	0.279	0.209	0.786	0.756
1000 conglomerados								
$\beta_{11} = 0.1$	0.101	0.006	0.007	0.085	0.093	0.087	0.966	0.996
$\beta_{12} = 0.1$	0.102	0.006	0.089	0.088	0.092	0.090	0.972	0.996
$\beta_{21} = 0.1$	0.101	0.006	0.014	0.089	0.091	0.089	0.977	0.998
$\beta_{22} = 0.1$	0.101	0.006	0.085	0.090	0.092	0.090	0.976	0.996
$\tau_1 = 2.0$	0.982	1.135	-0.509	0.178	0.112	0.145	0.404	0.470
$\tau_2 = 2.0$	0.979	1.127	-0.501	0.181	0.179	0.177	0.786	0.486

Tabela 4.6 – Cenário III para o modelo Weibull - mesma estrutura do Cenário I, exceto que aumentamos a correlação intra-conglomerado (tau de Kendall = 0,85) usando $\omega_{ik} \sim Gamma(1/11, 1/11)$.

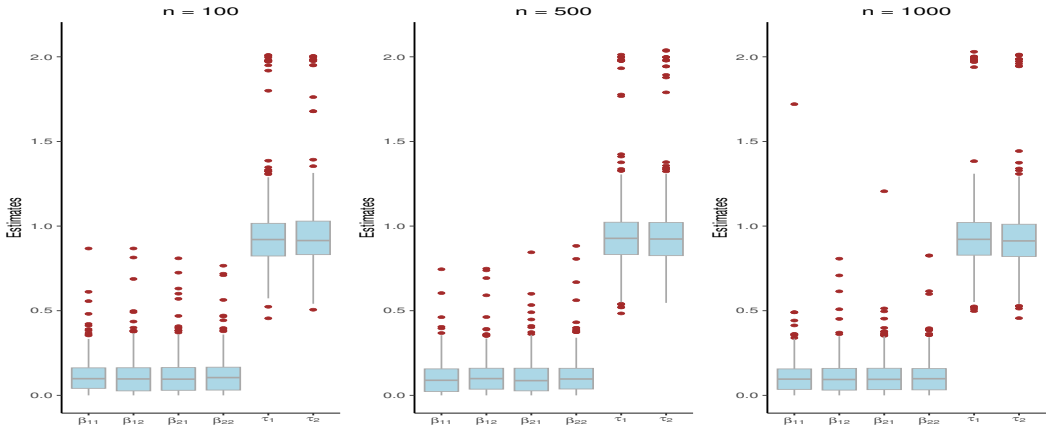
	Est	MSE	rb	SD	SE1	SE2	CP1	CP2
100 conglomerados								
$\beta_{11} = 0.1$	0.108	0.008	0.077	0.098	0.097	0.096	0.973	0.996
$\beta_{12} = 0.1$	0.107	0.009	0.068	0.095	0.096	0.096	0.965	0.992
$\beta_{21} = 0.1$	0.108	0.009	0.076	0.095	0.096	0.097	0.973	0.996
$\beta_{22} = 0.1$	0.109	0.008	0.087	0.092	0.096	0.094	0.974	0.995
$\tau_1 = 2.0$	0.940	1.165	-0.530	0.208	0.196	0.199	0.008	0.040
$\tau_2 = 2.0$	0.942	1.158	-0.529	0.199	1.115	1.115	0.033	0.033
500 conglomerados								
$\beta_{11} = 0.1$	0.101	0.007	0.007	0.096	0.096	0.090	0.966	0.996
$\beta_{12} = 0.1$	0.105	0.008	0.059	0.092	0.095	0.094	0.972	0.986
$\beta_{21} = 0.1$	0.101	0.008	0.014	0.091	0.095	0.093	0.977	0.988
$\beta_{22} = 0.1$	0.105	0.007	0.085	0.092	0.095	0.093	0.976	0.986
$\tau_1 = 2.0$	0.962	1.154	-0.521	0.198	0.155	0.171	0.004	0.770
$\tau_2 = 2.0$	0.959	1.151	-0.520	0.195	0.679	0.679	0.786	0.786
1000 conglomerados								
$\beta_{11} = 0.1$	0.101	0.007	0.007	0.088	0.086	0.088	0.966	0.996
$\beta_{12} = 0.1$	0.103	0.007	0.082	0.090	0.093	0.091	0.972	0.996
$\beta_{21} = 0.1$	0.101	0.006	0.013	0.090	0.093	0.090	0.977	0.998
$\beta_{22} = 0.1$	0.102	0.006	0.082	0.091	0.092	0.091	0.976	0.996
$\tau_1 = 2.0$	0.992	1.124	-0.520	0.188	0.150	0.165	0.004	0.770
$\tau_2 = 2.0$	0.989	1.123	-0.520	0.191	0.670	0.670	0.786	0.786



(a) Cenário I

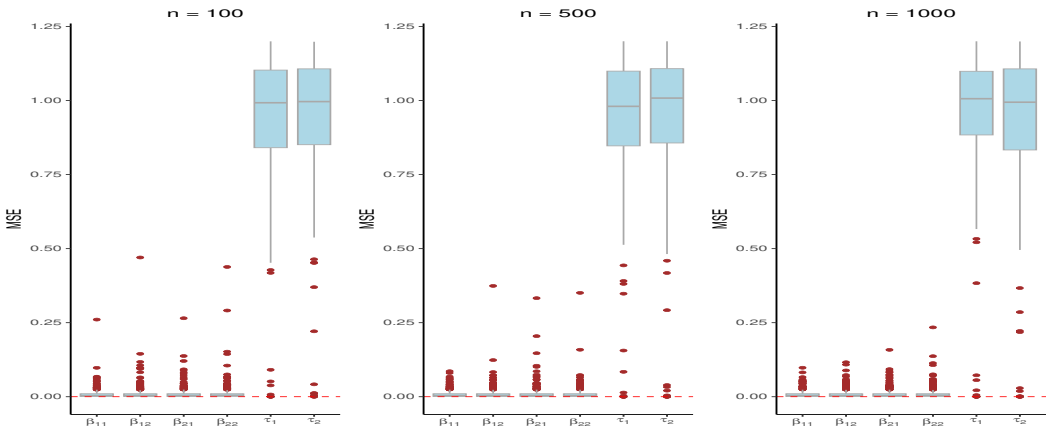


(b) Cenário II

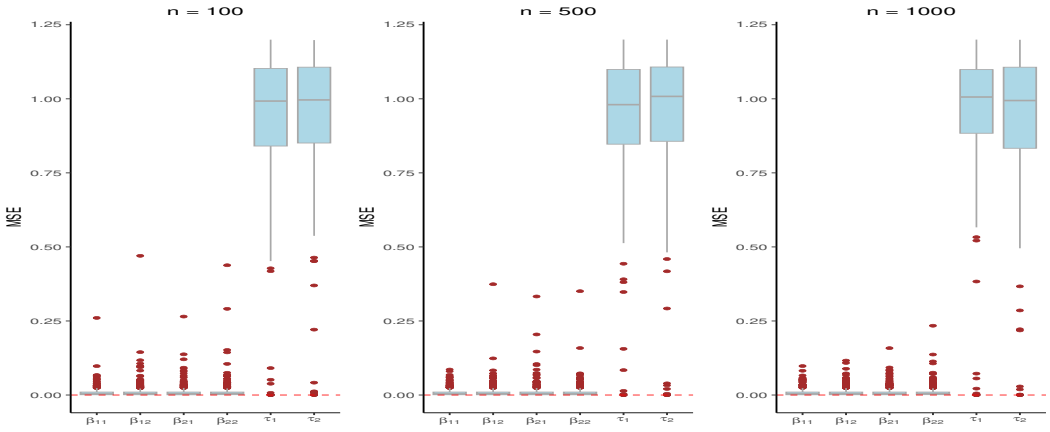


(c) Cenário III

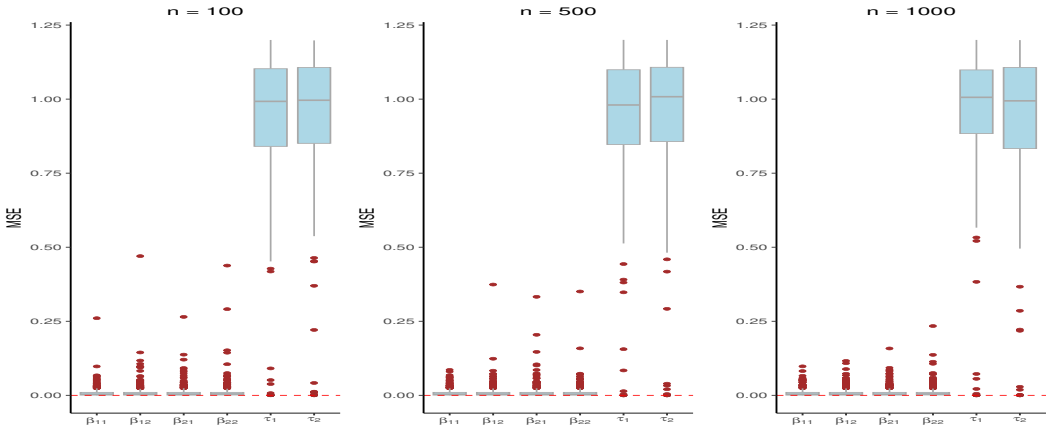
Figura 4.4 – Estimativas dos parâmetros do modelo Weibull



(a) Cenário I

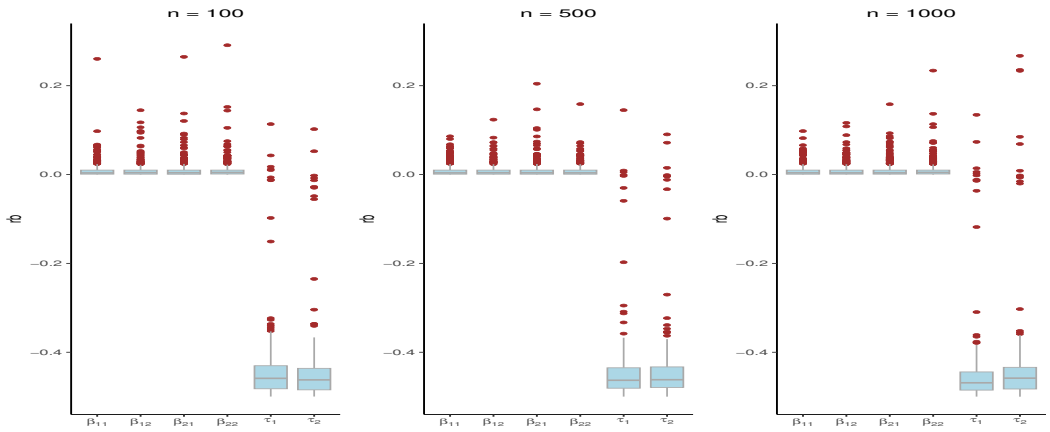


(b) Cenário II

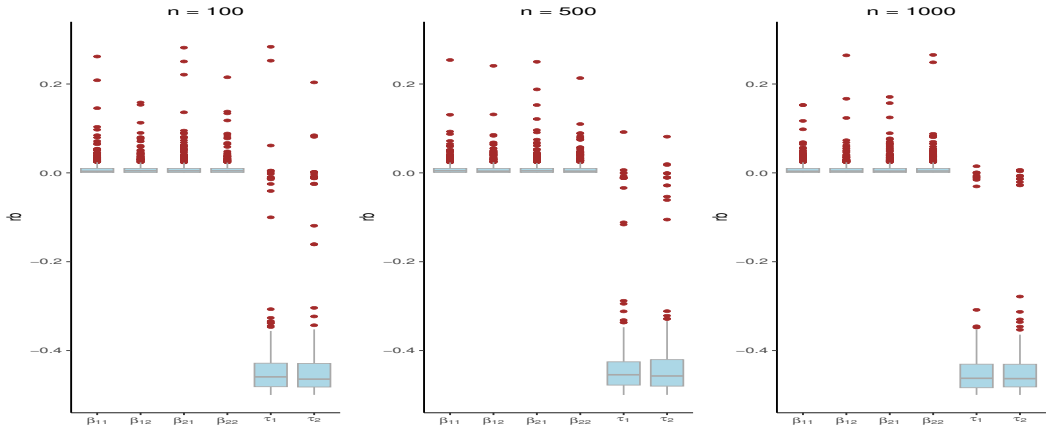


(c) Cenário III

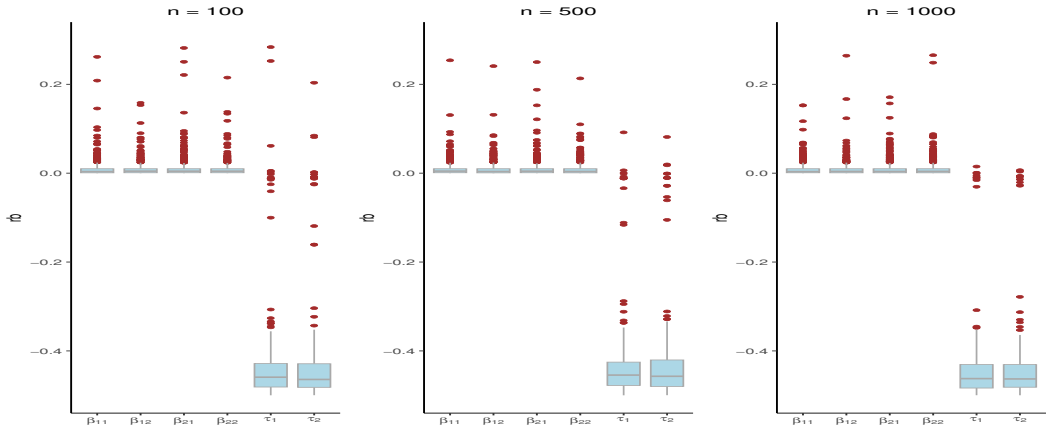
Figura 4.5 – Estimativas de erro quadrático médio (MSE) dos parâmetros do modelo Weibull



(a) Cenário I



(b) Cenário II



(c) Cenário III

Figura 4.6 – Estimativas de viés relativo (rb) dos parâmetros do modelo Weibull

Capítulo 5

Modelo de regressão Causa Específica para dados correlacionados na presença de riscos competitivos e censura intervalar

É comum estudos em coorte e ensaios clínicos os participantes do estudo estarem sob o risco de falha por múltiplos eventos mutuamente exclusivos e apenas é observado o tempo para a ocorrência do primeiro evento é de interesse (Kalbfleisch e Prentice, 2011; Collett, 2015). Os dados de riscos competitivos são censurados por intervalo quando o tempo exato da ocorrência do evento não é observado com precisão, é apenas conhecido por estar dentro de visitas ou exames periódocos de estudo, especialmente em ensaios clínicos ou estudos longitudinais (Sun, 2006) . Os dados obtitos são chamados por dados de riscos competitivos censurados por intervalo . No contexto de riscos competitivos as principais estimativas são a função de incidência acumulada (FIA) e a função de taxa de falha de causa específica.

O modelo de riscos proporcionais de Cox é aplicável à função de taxa de falha de causa específica $\lambda_k(t)$. Uma vez que a função de taxa de falha de causa específica é uma generalização da função taxa de falha $\lambda(t)$ na análise de sobrevivência tradicional,

a interpretação das taxas de falha no modelo de Cox taxa de falha causa específica é semelhante a da taxa de falha padrão. Essencialmente, o modelo de regressão causa específica é uma extensão da regressão de Cox tradicional para cada tipo de evento, em que as falhas de eventos concorrentes são tratados como observações censuradas. A forma do modelo de regressão causa específica é praticamente idêntica à regressão de Cox tradicional.

Diferentes tipos de censura surgem dependendo da forma em que os dados do experimento são coletados. Em muitos estudos o tempo exato da ocorrência do evento de interesse não é exatamente observado, sabe-se apenas que pertence um intervalo observado, com limites conhecidos, como por exemplo duas consultas ou exames consecutivos. Esse tipo de censura é conhecido na literatura como censura intervalar, é o tipo de censura mais geral que tem como casos particulares a censura à direita e a censura à esquerda, conforme discutido na Seção 2.4.

A incorporação da censura intervalar no modelo de riscos proporcionais não permite o cancelamento da função de taxa de falha linha de base e, como resultado, a estimativa dos coeficientes de regressão e a derivação de suas propriedades assintóticas têm se mostrado desafiadoras. Finkelstein (1986) propôs um método semiparamétrico no qual a distribuição da linha de base e os parâmetros de regressão são ajustados simultaneamente, maximizando a verossimilhança total dos dados. O método proposto utiliza a partição do eixo do tempo com base nos pontos finais dos intervalos de tempo do evento. Para estender o modelo de Cox para dados com censura intervalar, Pan (1999) propôs uma abordagem semiparamétrica para aproximar a distribuição acumulada linha de base com constantes por partes, o que leva ao algoritmo iterativo do minorante convexo (ICM). Pan (2000) usa um procedimento de imputação múltipla para preencher os tempos de falha para os eventos censurados por intervalo e, em seguida, aplica a análise de verossimilhança parcial padrão. Vários autores sugeriram métodos de suavização para estimar a função linha de base. Kooperberg e Clarkson (1997) sugerem usar splines para modelar o logaritmo da função linha de base, enquanto Joly et al. (1998) utiliza splines monótonos diretamente para modelar a função linha de base. Cai e Betensky (2003) propõem o uso de spline linear penalizado para a função linha de base.

Dados com censura intervalar surgem com frequência em estudos longitudinais de modo que a ocorrência do evento só é observada em avalições periódicas. Em muitos desses estudos longitudinais, vários tipos de eventos podem ocorrer a um indivíduo. Se esses tipos de eventos se excluem mutuamente ou apenas o tempo para a ocorrência do primeiro evento é de interesse, tem se então a presença de riscos competitivos. Muitos autores tem apresentado metodologias para função de incidência acumulada para dados de riscos ompetitivos com censura intervalar, como Hudgens et al. (2014), Li (2016b), Li e Fine (2013), Groeneboom et al. (2008a,b), Jewell (2003) entre outros. Usando uma abordagem de verossimilhança penalizada Li (2016a) desenvolve um modelo de taxa de falha de causa-específica do tipo Cox para dados de riscos competitivos sob censura intervalar e truncamento à esquerda. No entanto, o modelo de taxa de falha de causa-específica para dados com censura intervalar tem sido pouco explorado.

Muitos estudos envolvendo censura intervalar ou riscos competitivos são realizados em conglomerados, gerando assim, uma correlação entre os tempos de sobrevivência de indivíduos pertencentes ao mesmo agrupamento ou conglomerado. A estrutura de dependência desconhecida entre as observações de falha dentro do conglomerado, geralmente complexa, necessita de métodos de regressão apropriados que levem em conta as correlações de maneira robusta para permitir inferência válida para os efeitos covariáveis. Muitos modelos para dados de sobrevivência multivariados têm surgido na literatura. Muitos desses estudos tem sido generalizar o modelo semiparametrico de riscos proporcionais de Cox para uma configuração multivariada. Lee et al. (1992) apresentam uma abordagem marginal por meio das equações de estimação generalizada (GEE) com base no modelo de Cox. Outros trabalhos consideraram um modelo marginal com suposição de matriz de trabalho de independência para analisar o modelo de Cox na presença de censura intervalar (Lin e Wei (1989), Goggins e Finkelstein (2000), Kim e Xue (2002)). No contexto de riscos competitivos Zhou et al. (2012) estenderam o modelo de riscos proporcionais de Fine-Gray (Fine e Gray, 1999) para situações em que os indivíduos dentro de um conglomerado podem ser correlacionados, utilizando a abordagem de modelos marginais. Logan et al. (2010) desenvolveram um estimador de variância sanduíche para ajustar a correlação dentro do conglomerados usando pseudovalores.

Neste capítulo é apresentada a abordagem semiparamétrica de modelagem de dados de riscos competitivos em conglomerados na presença de censura intervalar. Na Seção 5.1, é apresentado o modelo de regressão causa específica com censura intervalar e a abordagem proposta por Chen et al. (2013) para modelar a função basal por aproximação de série de Taylor. Na Seção 5.2, propomos uma metodologia para incluir dados correlacionados. Estudos de simulação são realizados para avaliar propriedades amostrais do modelo na Seção 5.3.

5.1 Modelo de Regressão Causa Específica com censura intervalar

Suponha que tenha K riscos competitivos ou causas de falha para o tempo de falha T e seja $C \in \{1, ..., K\}$ representa a causa da falha. Assim, seja (T_i, C_i) o par que representa o tempo de ocorrência do evento e tipo do evento para o i-ésimo indivíduo, $i = 1, ..., n$. Sob censura intervalar, o tempo de falha T_i do sujeito i é conhecido por estar no intervalo $[L_i; R_i]$. Estamos interessados em estimar os efeitos das covariávies X e para esse proposito assumimos, para a causa k, um modelo separado de riscos proporcionais pode ser assumido, conforme abaixo:

$$\lambda_k(t; \mathbf{x}) = \lambda_{0k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x}), \tag{5.1}$$

$$\Lambda_k(t; \mathbf{x}) = \Lambda_{0k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x}), \tag{5.2}$$

em que $\lambda_{0k}(t)$ é a função de base e $\Lambda_{0k}(t)$ é a função de base acumulada, não especificadas, para a k-ésima causa específica, $\mathbf{x}$ é o vetor de covariáveis de interesse e $\boldsymbol{\beta}_k$ é o vetor que representa o efeito das covariáveis na k-ésima causa de falha, $k = 1, ..., K$. Esse modelo é conhecido como um modelo semiparamétrico, pois a função de base é tratada de forma não paramétrica e a suposição paramétrica é feita no efeito das covariáveis na taxa de falha. Como os K modelos são mutuamente exclusivos, eles podem ser analisados independentemente. Sob o modelo de riscos proporcionais, temos que

$$S_k(t) = \exp[-\Lambda_{0k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x})] = [S_{0k}(t)]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x})}, \tag{5.3}$$

em que $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s)ds$ e $S_{0k}(t) = \exp\left[-\int_0^t \lambda_{0k}(s)ds\right]$.

Para o desenvolvimento da função de verossimilhança considera-se que os tempos de observação são independentes de $(T_i; C_i)$ condicionalmente em $\mathbf{x}_i$ (censura intervar independente) e que a distribuição de $[L_i; R_i]$ não contém os parâmetros de interesse (censura intervalar não informativa). Assim, assumindo que $L_i < R_i$ para todo $i = 1, ..., n$, a função de verossimilhança para a k-ésima causa de falha é escrita como

$$L = \prod_{i=1}^{n} [S_k(L_i|\mathbf{x}_i) - S_k(R_i|\mathbf{x}_i)]^{\delta_{i1}} [S_k(L_i|\mathbf{x}_i)]^{\delta_{i2}} [S_k(T_i^*|\mathbf{x}_i)]^{\delta_{i3}},$$

em que $T_i^* = \frac{L_i+R_i}{2}$ é o ponto médio do intevalo que contém uma causa de falha diferente da causa específica k. As variáveis δ_{i1}, δ_{i2} e δ_{i3} são variáveis indicadoras de censuras, denotando censura intervalar para a causa k, censura a direita e censura intervalar para as causas diferentes de k, respectivamente, de modo que $\delta_{i1} + \delta_{i2} + \delta_{i3} = 1$. Portanto, para k-ésima causa de falha, a função log-verossimilhança tem a forma

$$\begin{aligned} l(\boldsymbol{\beta}_k, S_{0k}) = \sum_{i=1}^{n} \Big\{ & \delta_{i1} \log \left[[S_{0k}(L_i)]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} - [S_{0k}(R_i)]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} \right] \\ & + \delta_{i2} \log [S_{0k}(L_i)]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} + \delta_{i3} \log [S_{0k}(T_i^*)]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} \Big\}. \end{aligned} \tag{5.4}$$

Os estimadores de máxima verossimilhança não possuem forma analítica, sendo necessário assim métodos númericos para a maximização da função (5.4). A maximização da função log verossimilhança depende da estimativa das funções linha de base $S_{0k}(t)$, $\lambda_{0k}(t)$ ou $\Lambda_{0k}(t)$, que pode ser especificada por uma distribuição adequada, como a distribuição Weibull ou Gompertz, no caso de uma abordagem paramética. Por outro lado, uma abordagem semiparamétrica alternativa pode ser usada, considerando que a função de linha de base seja constante por partes, o que leva à técnica do algorítmo iterativo do minorante convexo, que é discutida extensivamente por Pan (1999).

A abordagem semiparamétrica melhora o ajuste do modelo embora tenha a desvantagem de que a continuidade da função linha de base estimada não acontece, é apenas uma função degrau. Chen et al. (2013) utilizaram a série de Taylor para aproximar a função linha de base com o objetivo de obter uma função mais suave. Eles propuseram o uso de uma expansão em série de Taylor para o logarítmo da função taxa de falha linha

de base até uma ordem ótima para que possa ser aproximada com maior precisão possível a partir dos dados observados.

Aproximar funções usando um número finito de termos de suas séries de Taylor é uma prática comum. Esta técnica será utilizada aqui para obter a melhor aproximação da função de linha de base a partir dos dados observados e para evitar as desvantagens dos outros métodos propostos para a mesma situação. Na aproximação da série de Taylor, a ordem ótima pode ser determinada usando o teste da razão de verossimilhança (Chen et al., 2013).

Para assegurar que $\lambda_{0k}(t) \geq 0$, considere a expansão em série de Taylor do logaritmo da função taxa de falha linha de base dada por

$$\log[\lambda_{0k}(t;\boldsymbol{\alpha}_k)] = \alpha_{0k} + \alpha_{1k}t + \frac{\alpha_{2k}}{2!}t^2 + ... + \frac{\alpha_{qk}}{q!}t^q, \tag{5.5}$$

em que $\boldsymbol{\alpha}_k = (\alpha_{0k}, \alpha_{1k}, ..., \alpha_{qk})$ é o vetor de parâmetros de linha de base a ser estimado. Assim, a função taxa de falha acumulada linha de base pode ser obtida como

$$\begin{aligned}\Lambda_{0k}(t;\boldsymbol{\alpha}) &= \int_0^t \lambda_{0k}(s;\boldsymbol{\alpha})ds = \int_0^t \exp[\log(\lambda_{0k}(s;\boldsymbol{\alpha}))]ds \\ &= \int_0^t \exp[\alpha_{0k} + \alpha_{1k}s + \frac{\alpha_{2k}}{2!}s^2 + ... + \frac{\alpha_{qk}}{q!}s^q]ds. \end{aligned} \tag{5.6}$$

Como $S_{0k}(t) = \exp[-\Lambda_{0k}(t)]$, a função log verossimilhança (5.4) pode ser reescrita como

$$\begin{aligned} l(\boldsymbol{\beta}_k, S_{0k}) = &\sum_{i=1}^{n} \Big\{ \delta_{i1} \log \Big[[\exp[-\Lambda_{0k}(L_i)]]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} - [\exp[-\Lambda_{0k}(R_i)]]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} \Big] \\ &+ \delta_{i2} \log[\exp[-\Lambda_{0k}(L_i)]]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} + \delta_{i3} \log[\exp[-\Lambda_{0k}(T_i^*)]]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)} \Big\} \\ = &\sum_{i=1}^{n} \big\{ \delta_{i1} \log \big[\exp[-\Lambda_{0k}(L_i)\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)] - \exp[-\Lambda_{0k}(R_i)\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)] \big] \\ &+ \delta_{i2} \log[\exp[-\Lambda_{0k}(L_i)\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)] + \delta_{i3} \log[\exp[-\Lambda_{0k}(T_i^*)\exp(\boldsymbol{\beta}_k^T \mathbf{x}_i)] \big\}, \end{aligned} \tag{5.7}$$

em que $\Lambda_{0k}(L_i)$ e $\Lambda_{0k}(R_i)$ são a função de taxa de falha acumulada linha de base nos pontos limites esquerdo e direito dos intervalos observados, $k = 1, ..., K$.

Para a causa k, os coeficientes da regressão $\boldsymbol{\beta}_k$ são estimados utilizado métodos númericos, uma vez que nenhuma forma explícita dos estimadores de máxima verossimi-

lhança pode ser encontrada. O número ideal da série de Taylor pode ser determinado com base no seguinte procedimento, conforme apresentado por Chen et al. (2013):

1 Ajustar a função de verossmilhança apenas com o primeiro termo da série de Taylor, isto é, $q = 0$, e obter as estimativas de máxima verossimilhança dos parâmetros $\hat{\boldsymbol{\beta}}_k$ e $\hat{\boldsymbol{\alpha}}_k = \hat{\alpha}_{0k}$ e considere o valor ajustado da função de verossimilhança como $l_0 = \max[l(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\alpha}}_k]$. Observe que para $q = 0$ a função taxa de falha linha de base é função exponencial.

2 Ajustar a função de verossimilhança com mais uma ordem da série de Taylor, isto é, $q = 1$, em que os parâmetros são $\hat{\boldsymbol{\beta}}_k$ e $\hat{\boldsymbol{\alpha}}_k = (\hat{\alpha}_{0k}, \hat{\alpha}_{1k})$ e semelhante ao passo 1, o valor ajustado da função de verossimilhança é representado por $l_1 = \max[l(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\alpha}}_k]$. Observe que nesse caso, $q = 1$ a função linha de base é equivalente a função Gompertz.

3 Usando um nível de significância adequado, como $\alpha = 5\%$, e para a distribuição Chi quadrado com 1 grau de liberdade, então tem-se:

3.1 Se $-2(l_0 - l_1) < \chi^2_{1,(1-\alpha)}$ então a ordem selecionada da série de Taylor é $q = 0$ e portanto, as estimativas de máxima verossimilhaça dos parâmentros são $\hat{\boldsymbol{\beta}}_k$ e $\hat{\boldsymbol{\alpha}}_k = \hat{\alpha}_{0k}$.

3.2 Se a condição em (3.1) não for sastifeita então obteremos novas estimativas dos parâmetros em $q = 2$ e considere o novo valor ajustado da função de verossimilhança como $l_2 = \max[l(\hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\alpha}}_k]$ e, em seguida, repita o passo (3) novamente usando os valores l_1 e l_2.

Repita este procedimento até a condição de parada $-2(l_{q^*-1} - l_{q^*}) < \chi^2_{1,(1-\alpha)}$, em que a ordem da série de Taylor é $q = q^* - 1$ e, portanto, os estimadores de máxima verossimilhança são $\hat{\boldsymbol{\beta}}_k$ e $\hat{\boldsymbol{\alpha}}_k = (\hat{\alpha}_{0k}, \hat{\alpha}_{1k}, ..., \hat{\alpha}_{(q^*-1)k})$. Observe que a escolha do valor do nível de significância pode afetar o número de termos na aproximação de Taylor, em que menores valores para o nível de significância podem aumentar o número de termos ideal na aproximação de Taylor e vice-versa.

5.2 Modelo de Regressão Causa Específica com censura intervalar para dados correlacionados

Em muitas situações envolvendo riscos competitivos, os indivíduos no estudo podem ser correlacionados dentro do conglomerado, devido a fatores compartilhados não observados entre os indivíduos (Zhou et al., 2012). Portanto, o objetivo é desenvolver um modelo de regressão causa específica que considere a estrutura de correlação dos conglomerados e o mecanismo de censura intervalar.

Considere n conglomerados, e que cada um dos i conglomerados tenha m_i indivíduos, $i = 1, ..., n$ de modo que se tenha $\sum_{i=1}^{n} m_i = m$ indivíduos em toda a amostra. Suponha que tenha K riscos competitivos ou causas de falha para o tempo de falha T e $C \in \{1, ..., K\}$ representa a causa da falha. Seja T_{ij} o tempo de falha do j-ésimo indivíduo no i-ésimo conglomerado, $j = 1, ..., m_i$, de modo que é conhecido apenas que o tempo T_{ij} pertence a um intervalo, $[L_{ij}, R_{ij}]$ e seja $\mathbf{x}$ o vetor de covariávies de interesse.

Considere que que os tempos de observação são independentes de $(T_i; C_i)$ condicionalmente em $\mathbf{x}_i$ e que a distribuição de $[L_i; R_i]$ não contém os parâmetros de interesse. Defina a função taxa de falha de causa específica para a k causa dado as covariávies $\mathbf{x}$ por

$$\lambda_k(t; \mathbf{x}) = \lambda_{0k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x}), \tag{5.8}$$

e portanto temos,

$$S_k(t) = \exp[-\Lambda_{0k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x})] = [S_{0k}(t)]^{\exp(\boldsymbol{\beta}_k^T \mathbf{x})}. \tag{5.9}$$

Sob a hipótese de matriz de trabalho de independência, propomos a seguinte função de verossimilhança para a k-ésima causa de falha, considerando dados correlacionados sujeitos a censura intervalar:

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m_i} [S_k(L_{ij}|\mathbf{x}_{ij}) - S_k(R_{ij}|\mathbf{x}_{ij})]^{\delta_{ij1}} [S_k(L_{ij}|\mathbf{x}_{ij})]^{\delta_{ij2}} [S_k(T_{ij}^*|\mathbf{x}_{ij})]^{\delta_{ij3}},$$

em que $T_{ij}^* = \frac{L_{ij}+R_{ij}}{2}$ é o ponto médio do intervalo que contém uma causa de falha diferente da causa específica k. As variáveis δ_{ij1}, δ_{ij2} e δ_{ij3} são variáveis indicadoras de censuras,

denotando censura intervalar para a causa k, censura a direita e censura intervalar para as causas diferentes de k, respectivamente, de modo que $\delta_{ij1} + \delta_{ij2} + \delta_{ij3} = 1$.

Portanto, para k-ésima causa de falha, a função log verossimilhança tem a forma

$$l(\boldsymbol{\beta}_k, S_{0k}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\Big\{\delta_{ij1}\log\Big[[S_{0k}(L_{ij})]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})} - [S_{0k}(R_{ij})]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})}\Big] + \delta_{ij2}\log[S_{0k}(L_{ij})]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})} + \delta_{ij3}\log[S_{0k}(T_{ij}^*)]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})}\Big\}. \tag{5.10}$$

Usando a função taxa de falha acumulada (5.6) e o fato que $S_{0k}(t) = \exp[-\Lambda_{0k}(t)]$, a função log verossimilhança (5.10) pode ser reescrita como

$$\begin{aligned} l(\boldsymbol{\Theta}_k) =& \sum_{i=1}^{n}\sum_{j=1}^{m_i}\Big\{\delta_{ij1}\log\Big[[\exp[-\Lambda_{0k}(L_{ij})]]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})} - [\exp[-\Lambda_{0k}(R_{ij})]]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})}\Big] \\ &+ \delta_{ij2}\log[\exp[-\Lambda_{0k}(L_{ij})]]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})} + \delta_{ij3}\log[\exp[-\Lambda_{0k}(T_{ij}^*)]]^{\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})}\Big\} \\ =& \sum_{i=1}^{n}\sum_{j=1}^{m_i}\big\{\delta_{ij1}\log\big[\exp[-\Lambda_{0k}(L_{ij})\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})] - \exp[-\Lambda_{0k}(R_{ij})\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})]\big] \\ &+ \delta_{ij2}\log[\exp[-\Lambda_{0k}(L_{ij})\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})] + \delta_{ij3}\log[\exp[-\Lambda_{0k}(T_{ij}^*)\exp(\boldsymbol{\beta}_k^T\mathbf{x}_{ij})]\big\}, \end{aligned} \tag{5.11}$$

em que $\boldsymbol{\Theta}_k = (\boldsymbol{\beta}_k, \boldsymbol{\alpha}_k)$ e $\Lambda_{0k}(L_i)$ and $\Lambda_{0k}(R_i)$ são a função de taxa de falha acumulada linha de base nos pontos limites esquerdo e direito dos intervalos observados, $k = 1, ..., K$.

Como a log verossimilhança (5.10) não é a verdadeira log verossimilhança, não é possivel utilizar a inversa da matriz de informação de Fisher observada para $\boldsymbol{\Theta}_k$ para estimar a matriz de covariância dos estimadores, pois esta não fornece um estimador de covariância adequado por causa da suposição incorreta de independência. Embora geralmente dependente da correlação intraconglomerado, os estimadores são considerados assintoticamente normais conjuntamente com uma matriz de covariância que pode ser estimada consistentemente sem assumir uma estrutura de correlação específica (Wei et al., 1989). Utilizamos um estimador de variância sanduiche para estimar a matriz de covariância de $\hat{\boldsymbol{\Theta}}_k$.

Para obter o estimador de variância sanduíche tratamos a equação escore, utilizando a primeira derivada da função log verossimilhança (5.10), como a equação de estimação de forma semelhante ao GEE proposta por Liang e Zeger (1986). Para es-

timar a matriz de covariância das estimativas dos parâmetros $\boldsymbol{\beta}_k$ utilizamos o seguinte estimador de variância sanduíche como em Kor et al. (2012), expressa da seguinte forma,

$$nI^{-1}(\hat{\boldsymbol{\Theta}}_k)\left\{\sum_i U_i(\hat{\boldsymbol{\Theta}}_k)U_i(\hat{\boldsymbol{\Theta}}_k)^T\right\}I^{-1}(\hat{\boldsymbol{\Theta}}_k), \tag{5.12}$$

sendo que $U_i(\hat{\boldsymbol{\Theta}}_k) = \dfrac{\partial \log L_i(\hat{\boldsymbol{\Theta}}_k)}{\partial \hat{\boldsymbol{\Theta}}_k}|_{\boldsymbol{\Theta}_k=\hat{\boldsymbol{\Theta}}_k}$ é a contribuição do i-ésimo conglomerado no vetor de score e $I(\hat{\boldsymbol{\Theta}}_k) = -\dfrac{\partial^2 \log L(\boldsymbol{\Theta}_k)}{\partial^2 \boldsymbol{\Theta}_k}|_{\boldsymbol{\Theta}_k=\hat{\boldsymbol{\Theta}}_k}$.

Segundo Liu (2012) o método da variância sanduíche não especifica o padrão de dependência entre os tempos de falhas correlacionados, em vez disso, ele constrói um estimador robusto da matriz de variância e covariância para explicar a correlação intraconglomerados, da mesma forma que Zeger et al. (1988) utilizaram na análise de dados longitudinais. O estimador sanduíche resultará em estimativas de erros padrão assintoticamente corretas, independente da estrutura de correlação dos agrupamentos.

5.3 Estudo de Simulação

Dentro das especificações gerais do estudo de simulação, os tempos de vida podem ser gerados utilizando o método da função de distribuição inversa. A função de sobrevivência para o modelo de riscos proporcionais de Cox é dada por

$$S(t) = \exp[-\Lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x})].$$

Assim, a sua função de distribuição é

$$F(t) = 1 - \exp[-\Lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x})], \tag{5.13}$$

em que $\Lambda_0(t)$ é a função taxa de falha acumulada linha de base avaliado no tempo t, e $\mathbf{x}$ é um vetor de covariáveis com parâmetros associados $\boldsymbol{\beta}_k$.

Seja Z uma variável aleatória com distribuição F, então $U = F(Z)$ é descrita por uma distribuição uniforme no intervalo 0 até 1, isto é, $U \sim U[0,1]$. Além disso, se $U \sim U[0,1]$ então $(1-U) \sim U[0,1]$. Assim, seja T o tempo de sobrevivência do modelo de Cox, segue-se de (5.13) que

$$U = \exp[-\Lambda_0(t)\exp(\boldsymbol{\beta}_k^T\mathbf{x})] \sim U[0,1].$$

Se $\lambda_0(t) > 0$ para todo t, então Λ_0 pode ser invertida e T pode ser escrita como

$$T = \Lambda_0^{-1}[-log(U)\exp(\boldsymbol{\beta}_k^T\mathbf{x})]. \tag{5.14}$$

Portanto, simular tempos de sobrevivência a partir de um modelo de Cox com covariáveis requer a inversão da função de taxa de falha acumulada. Os parâmetros necessários para cada distribuição, a função de taxa de falha, a função de taxa de falha acumulada, o inverso da função de taxa de falha acumulada e a fórmula para simulação dos tempos de sobrevivência de cada distribuição são descritos na Tabela 5.1. Embora existam várias parametrizações diferentes da distribuição Weibull, utilizamos a parametrização de Bender et al. (2005).

Tabela 5.1 – Caracterização das distribuições exponencial, Weibull e Gompertz

Característica	Exponencial	Weibull	Gompertz
Parâmetro	$\alpha > 0$	$\alpha > 0$ e $\tau > 0$	$\alpha > 0$ e $-\infty < \tau < \infty$
Taxa de falha	$\lambda_0(t) = \alpha$	$\lambda_0(t) = \alpha\tau t^{\tau-1}$	$\lambda_0(t) = \alpha \exp(\tau t)$
Taxa de falha acumulada	$\Lambda_0(t) = \alpha t$	$\Lambda_0(t) = \alpha t^{\tau}$	$\Lambda_0(t) = \frac{\alpha}{\tau}[\exp(\tau t) - 1]$
Inversa da taxa de falha acumulada	$\Lambda_0^{-1}(t) = \alpha^{-1} t$	$\Lambda_0^{-1}(t) = (\alpha^{-1}t)^{1/\tau}$	$\Lambda_0^{-1}(t) = \frac{1}{\tau}\log(\frac{\tau}{\alpha}t + 1)$
Tempos simulados ($u \sim U(0,1)$)	$T = -\frac{log(u)}{\alpha \exp(\boldsymbol{\beta}_k^T \mathbf{x})}$	$T = \left(-\frac{log(u)}{\alpha \exp(\boldsymbol{\beta}_k^T \mathbf{x})}\right)^{1/\tau}$	$T = \frac{1}{\tau}\log\left(1 - \frac{\tau log(u)}{\alpha \exp(\boldsymbol{\beta}_k^T \mathbf{x})}\right)$

Estudos numéricos foram conduzidos para avaliar o desempenho da abordagem proposta. Para gerar os n conglomerados que satisfaçam modelo (5.8), assumimos que existem duas causas de falhas e consideramos que as taxas de falha causa específica para k-ésima causa de falha satisfaz,

$$\lambda_k(t; \mathbf{X}) = \omega_{ik}\lambda_{0k}(t) \exp(\boldsymbol{\beta}_k^T \mathbf{x}), \tag{5.15}$$

para cada conglomerado i, $i = 1, ..., n$, em que ω_{ik} são uma amostra aleatória de uma distribuição $Gamma(\theta_k, \theta_k)$. O parâmetro θ_k é escolhido de modo a obter a correlação desejada. Simulamos os tempos de sobrevivência do modelo na Equação (5.15), utilizado dois mecanismos diferentes para geração dos dados: o primeiro foi utilizando o modelo Gompertz e o segundo utilizando o modelo Weibull. Assim, de acordo com a Tabela 5.1 temos que,

1. $\lambda_{0k}(t) = \alpha \exp(\tau t)$, em que $\alpha > 0$ e $-\infty < \tau < \infty$, e
2. $\lambda_{0k}(t) = \alpha\tau t^{\tau-1}$, em que $\alpha > 0$ e $\tau > 0$

Diferentes cenários são considerados e para todos eles há $K = 2$ causas de falha, $n \in \{100, 500, 1000\}$, $m_i = 4$, $i = 1, ..., n$. São consideradas duas covariáveis independentes: $\mathbf{x} = (X_1, X_2)$ em que $X_1 \sim Ber(0, 5)$ e $X_2 \sim N(0, 1)$. O tempo de falha e os tipos de causa foram gerados a partir das duas funções de taxa de falha de causa-específica utilizando o algorítimo proposto por Beyersmann et al. (2009), apresentado na Seção 4.3, em que

$$\lambda_1(t; \mathbf{x}) = \omega_{i1}\lambda_{01}(t) \exp(0.3X_1 + 0.3X_2)$$

e

$$\lambda_2(t;\mathbf{x}) = \omega_{i2}\lambda_{02}(t)\exp(0.3X_1 + 0.3X_2).$$

A geração dos dados é feita da mesma forma que foi feito na secão 4.3, mas para facilitar o acompanhamento do leitor será repetido novamente a baixo.

1 Um termo aleatório comum para cada conglomerado, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, é gerado a partir de uma distribuição gama, $\omega_{ik} \sim Gamma(\theta_k, \theta_k)$, seguindo os passos de Brazauskas e Le-Rademacher (2016). θ_k é escolhido para obter a estrutura de dependência dentro do conglomerado em cada cenário.

2 Para cada individuo, covariáveis X_1 e X_2 são geradas independentemente;

3 Geramos dados de riscos competitivos por meio do seguinte algoritmo, de acordo com Beyersmann et al. (2011):

3.1 Especifique as funções de taxa de falha de causa-específica $\lambda_1(t;\mathbf{x})$ e $\lambda_2(t;\mathbf{x})$, para modelos exponenciais e Weibull, de acordo com as equações (4.13) e (4.14), respectivamente.

3.2 Simule tempos de falha T com taxa de falha de todas as causas $\lambda(t;\mathbf{x}) = \lambda_1(t;\mathbf{x}) + \lambda_2(t;\mathbf{x})$.

3.3 Determine o tipo de evento para cada indivíduo usando um experimento binomial para um tempo de falha simulado T, que decide com probabilidade $\dfrac{\lambda_k(t;\mathbf{x})}{\lambda_1(t;\mathbf{x}) + \lambda_2(t;\mathbf{x})}$ na causa k, $k = 1, 2$.

3.4 Além disso, gere tempos de censura à direita independentes $C \sim Exp(0, 8)$.

3.5 Os tempos observados são considerados o mínimo entre os tempos de falha T e os tempos censurados C, $t_{ij} = min(T_{ij}, C_{ij})$.

4 Os tempos de censura intervalar (L_{ij} e R_{ij}) são construídos da seguinte forma:

4.1 Para todos os indivíduos censurados à direita, aplicamos os seguintes intervalos de censura

$$L_{ij} = t_{ij} \text{ and } R_{ij} = \infty;$$

4.2 Para indivíduos que falharam por uma das causas K, aplicamos os seguintes tempos de censura de intervalo:

$$L_{ij} = 0.0001 \text{ and } R_{ij} = c,$$

em que c é gerado a partir de uma distribuição uniforme $U[0.1; b]$. O valor b é escolhido de forma a aumentar ou diminuir o tamanho dos intervalos. Os tempos de vida t_{ij} são comparados com os intervalos criados, para verificar se t_{ij} pertence ao intervalo. Se t_{ij} não pertencer ao intervalo, novos intervalos são criados de modo que

$$L_{ij} = R_{ij} \text{ and } R_{ij} = R_{ij} + c$$

em que c é um novo valor gerado a partir de uma distribuição uniforme $U[0.1; b]$.

Os dados simulados foram gerados em três cenário, usando a linguagem R (R Core Team, 2022) para avaliar o desempenho de amostras finitas da metodologia proposta. No cenário 1, foi considerada uma estrutura de correlação intra-conglomerado pequena, *tau de Kendall* igual a 0,20, e para isso, os efeitos aleatórios específico do conglomerado, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, são gerados a partir da distribuição gama, $\omega_{ik} \sim Gamma(2, 2)$. Além disso, o comprimento dos intervalos censurados são gerados de uma distribuição uniforme $U(0.1, 0.3)$. No cenário 2, pretende-se verificar a influência do comprimento dos intervalos censurados e para isso, mantem-se a mesma estrutura do cenário 1 aumentando apenas o comprimento dos intervalos censurados, que são gerados de uma distribuição uniforme $U(0.1, 0.5)$. No cenário 3, pretende-se verificar a influência da estrutura de correlação intra-conglomerados. Para isso mantem-se a mesma estrutura do cenário 1, aumentando apenas a correlação intra-conglomerado, *tau de Kendall* igual a 0,85, e para isso, os efeitos aleatórios específico do conglomerado, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, são gerados a partir da distribuição gama, $\omega_{ik} \sim Gamma(1/11, 1/11)$.

Cada conjunto de dados é analisado usando a metodologia proposta na Seção 5.2. Cada resultado de simulação exibido nesta seção é baseado em conjuntos de dados simulados de tamanho $N = 1000$. O desempenho de todos os estimadores pontuais são

comparado numericamente em termos de estimativa média, variância empírica, variância assintótica e valores de erro quadrático médio (MSE).

A média das estimativas dos parâmetros é calculada como $\hat{\theta} = \frac{\sum_{i=1}^{N} \hat{\theta}_i}{N}$, fornecendo uma estimativa do valor esperado do estimativas dos coeficientes para as N simulações. O viés relativo (rb) é estimado tomando a diferença entre $\hat{\theta}$ e θ, o verdadeiro valor do parâmetro populacional, dividido por e θ, tal que $rb = \frac{\hat{\theta} - \theta}{|\theta|}$.

A variância empírica é calculada por $Var(\hat{\theta}) = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \hat{\theta})^2}{N-1}$ e o erro quadrático médio é calculado por $MSE(\theta) = \frac{\sum_{i=1}^{N}(\hat{\theta} - \theta)^2}{N}$.

O erro padrão (SE_1) foi calculado como menos o inverso da matriz de segundas derivadas de $\log L(\Theta)$ e o estimador sanduíche de erro padrão (SE_2) foi calculado usando a variância sanduíche (4.8) com base na média dos resultados das N simulações.

Além disso, a probabilidade de cobertura (CP_1) de 95% é calculada para estimativas de intervalo considerando o erro padrão SE_1 e a probabilidade de cobertura (CP_2) de 95% para estimativas de intervalo considerando o estimador sanduíche do erro padrão SE_2.

5.3.1 Resultados

Os resultados do estudo de simulação são apresentados nas Tabelas 5.2 a 5.4 e Figuras 5.1 a 5.3, para o modelo de Cox com taxa de falha basal Gompertz, e nas Tabelas 5.5 a 5.7 e Figuras 5.4 a 5.6, para o modelo de Cox com taxa de falha basal Weibull.

Tabela 5.2 – Cenário I para modelo de Cox com taxa de falha basal Gompertz - Neste cenário $\omega_{ik} \sim Gamma(2,2)$, causando assim uma fraca correlação entre os indivíduos dentro de um mesmo conglomerado (tau de Kendall = 0,20). O tamanho dos intervalos gerados a partir de $c \sim U(1, 0.1, 0.3)$.

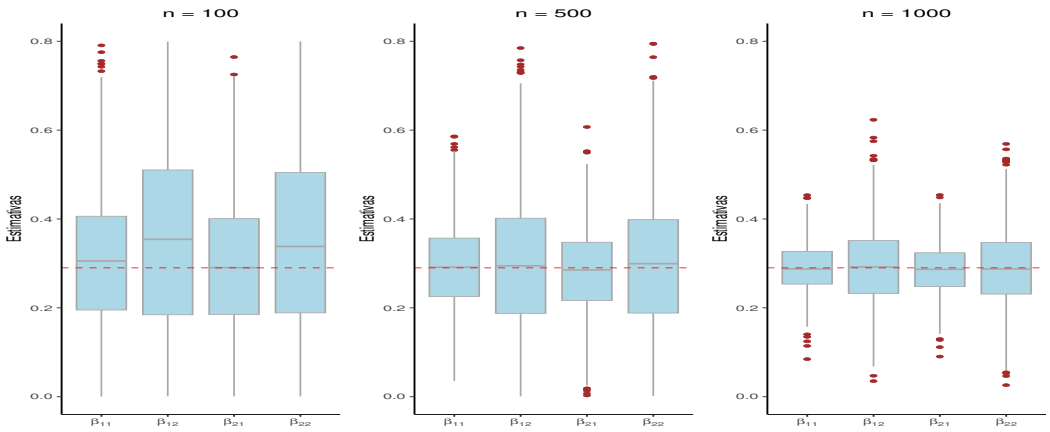
	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
100 conglomerados								
$\beta_{11} = 0.3$	0.295	0.030	-0.036	0.173	0.169	0.170	0.940	0.945
$\beta_{12} = 0.3$	0.297	0.099	-0.049	0.315	0.286	0.299	0.918	0.945
$\beta_{21} = 0.3$	0.282	0.029	-0.065	0.168	0.165	0.166	0.894	0.944
$\beta_{22} = 0.3$	0.290	0.087	-0.063	0.296	0.286	0.291	0.935	0.938
500 conglomerados								
$\beta_{11} = 0.3$	0.289	0.009	-0.035	0.097	0.097	0.097	0.947	0.948
$\beta_{12} = 0.3$	0.287	0.029	-0.042	0.170	0.163	0.166	0.932	0.942
$\beta_{21} = 0.3$	0.282	0.010	-0.061	0.097	0.094	0.095	0.926	0.946
$\beta_{22} = 0.3$	0.284	0.027	-0.053	0.164	0.163	0.164	0.932	0.947
1000 conglomerados								
$\beta_{11} = 0.3$	0.289	0.008	-0.033	0.095	0.096	0.096	0.947	0.948
$\beta_{12} = 0.3$	0.287	0.027	-0.041	0.165	0.163	0.166	0.932	0.942
$\beta_{21} = 0.3$	0.282	0.008	-0.059	0.095	0.094	0.095	0.946	0.956
$\beta_{22} = 0.3$	0.284	0.025	-0.050	0.161	0.162	0.161	0.932	0.946

Tabela 5.3 – Cenário II para modelo de Cox com taxa de falha basal Gompertz - mesma estrutura do Cenário I, exceto que aumentamos o tamanho dos intervalos usando $c \sim U(1, 0.1, 0.5)$.

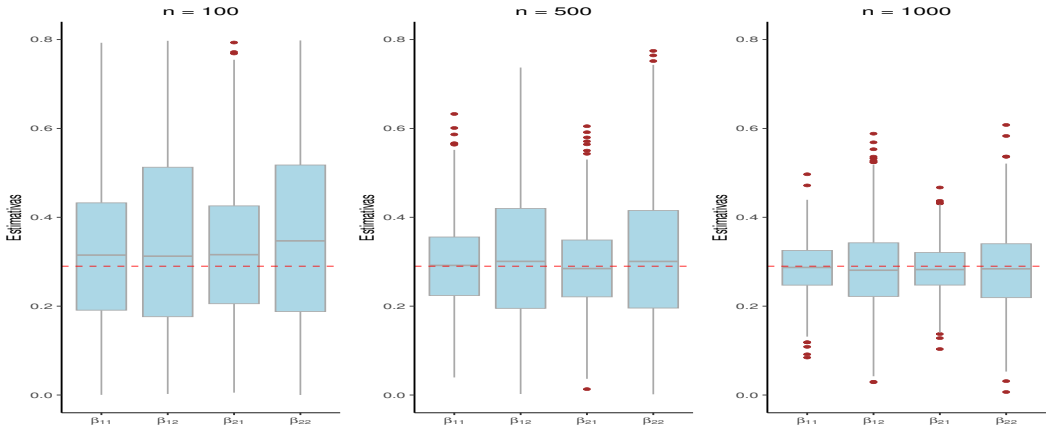
	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
100 conglomerados								
$\beta_{11} = 0.3$	0.291	0.033	-0.031	0.180	0.177	0.178	0.958	0.952
$\beta_{12} = 0.3$	0.296	0.097	-0.015	0.312	0.299	0.299	0.941	0.939
$\beta_{21} = 0.3$	0.287	0.033	-0.044	0.181	0.172	0.179	0.938	0.961
$\beta_{22} = 0.3$	0.288	0.099	-0.038	0.314	0.299	0.305	0.937	0.943
500 conglomerados								
$\beta_{11} = 0.3$	0.291	0.010	-0.031	0.099	0.100	0.100	0.954	0.952
$\beta_{12} = 0.3$	0.296	0.029	-0.015	0.174	0.170	0.173	0.949	0.954
$\beta_{21} = 0.3$	0.287	0.009	-0.044	0.099	0.098	0.097	0.925	0.951
$\beta_{22} = 0.3$	0.288	0.030	-0.038	0.174	0.169	0.172	0.911	0.953
1000 conglomerados								
$\beta_{11} = 0.3$	0.300	0.010	-0.001	0.099	0.100	0.100	0.954	0.952
$\beta_{12} = 0.3$	0.300	0.029	-0.001	0.172	0.169	0.172	0.949	0.954
$\beta_{21} = 0.3$	0.298	0.009	-0.006	0.098	0.098	0.098	0.935	0.961
$\beta_{22} = 0.3$	0.289	0.030	-0.036	0.170	0.168	0.170	0.911	0.935

Tabela 5.4 – Cenário III para modelo de Cox com taxa de falha basal Gompertz - mesma estrutura do Cenário I, exceto que aumentamos a correlação intra-conglomerado (tau de Kendall = 0,85) usando $\omega_{ik} \sim Gamma(1/11, 1/11)$.

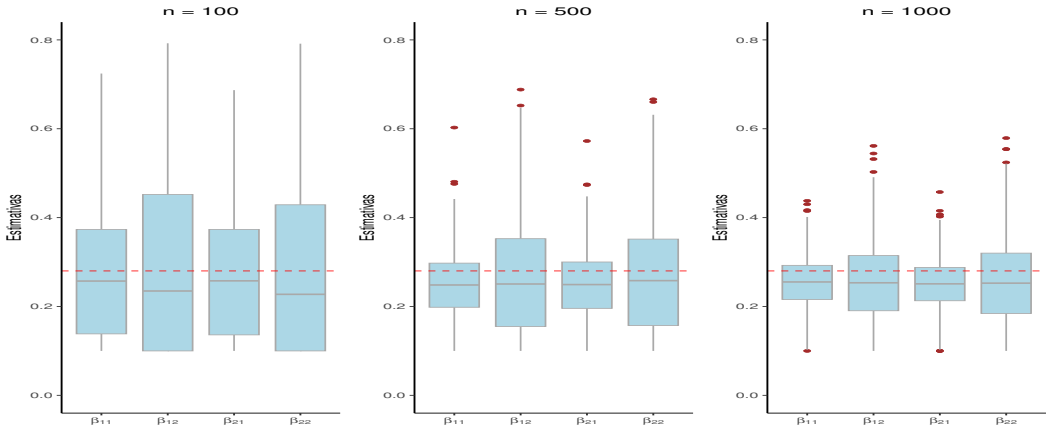
	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
100 conglomerados								
$\beta_{11} = 0.3$	0.250	0.023	-0.168	0.179	0.181	0.181	0.986	0.988
$\beta_{12} = 0.3$	0.262	0.056	-0.125	0.337	0.390	0.327	0.973	0.980
$\beta_{21} = 0.3$	0.250	0.022	-0.165	0.177	0.167	0.175	0.942	0.955
$\beta_{22} = 0.3$	0.264	0.054	-0.120	0.301	0.290	0.303	0.948	0.949
500 conglomerados								
$\beta_{11} = 0.3$	0.250	0.008	-0.168	0.102	0.135	0.102	0.891	0.914
$\beta_{12} = 0.3$	0.262	0.017	-0.125	0.186	0.188	0.187	0.982	0.988
$\beta_{21} = 0.3$	0.250	0.008	-0.165	0.106	0.109	0.105	0.875	0.959
$\beta_{22} = 0.3$	0.264	0.016	-0.120	0.193	0.198	0.191	0.920	0.954
1000 conglomerados								
$\beta_{11} = 0.3$	0.276	0.008	-0.081	0.101	0.105	0.101	0.891	0.914
$\beta_{12} = 0.3$	0.309	0.017	0.065	0.156	0.158	0.156	0.982	0.988
$\beta_{21} = 0.3$	0.275	0.008	-0.084	0.102	0.104	0.102	0.875	0.959
$\beta_{22} = 0.3$	0.303	0.016	0.044	0.183	0.188	0.183	0.920	0.954



(a) Cenário I

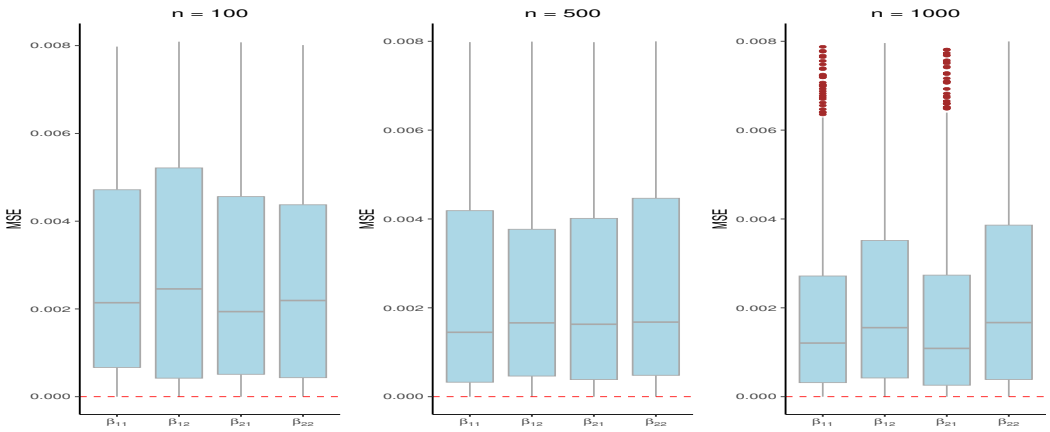


(b) Cenário II

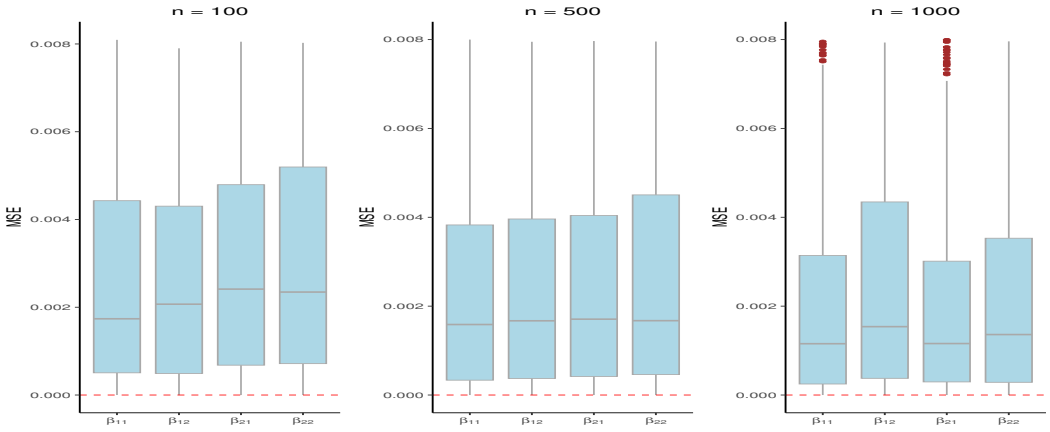


(c) Cenário III

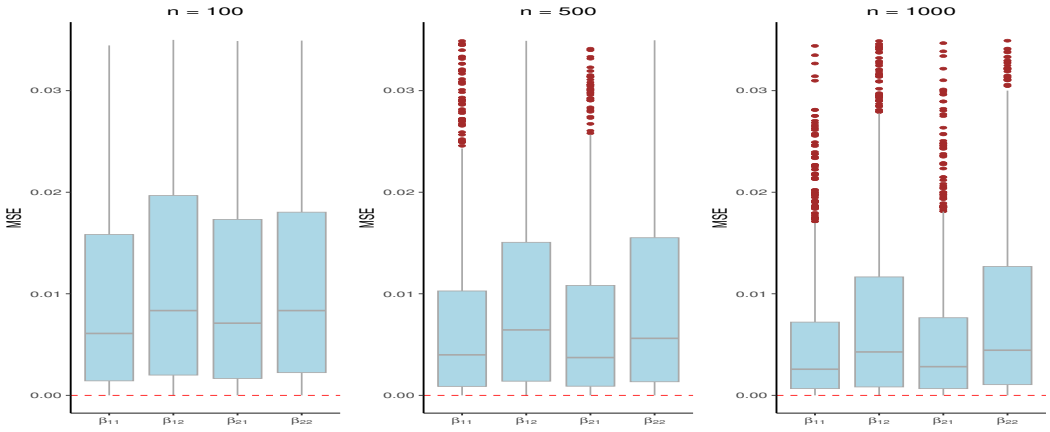
Figura 5.1 – Estimativas dos parâmetros do modelo de Cox com taxa de falha basal Gompertz



(a) Cenário I

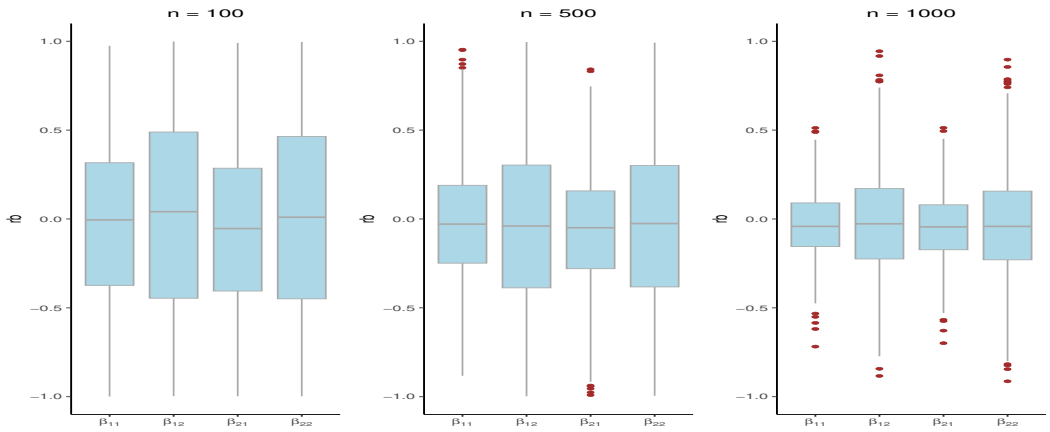


(b) Cenário II

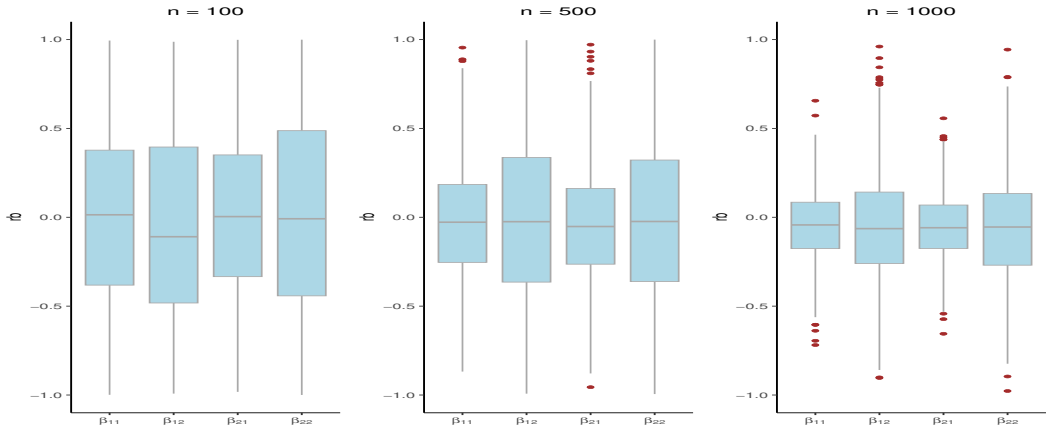


(c) Cenário III

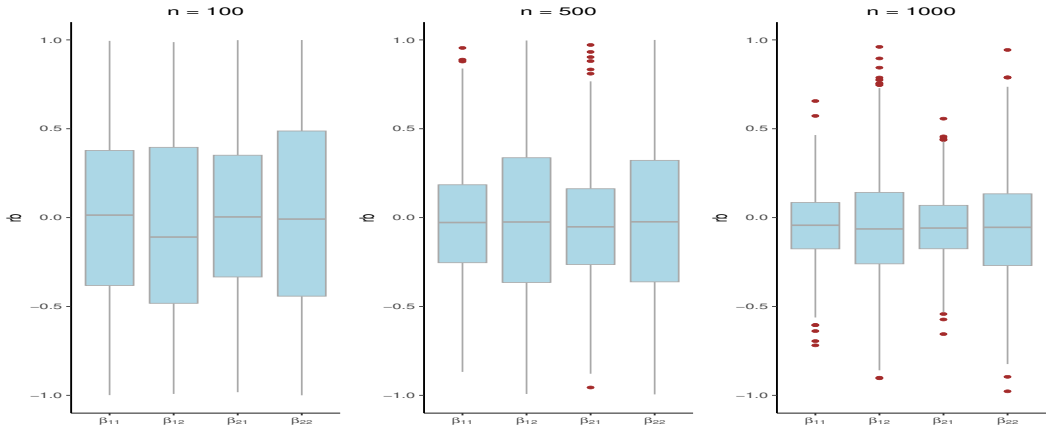
Figura 5.2 – Estimativas de erro quadrático médio (MSE) dos parâmetros modelo de Cox com taxa de falha basal Gompertz



(a) Cenário I



(b) Cenário II



(c) Cenário III

Figura 5.3 – Estimativas de viés relativo (rb) dos parâmetros do modelo de Cox com taxa de falha basal Gompertz

Conforme observado na Tabela 5.2, modelo de Cox com taxa de falha basal Gompertz, e na Tabela 5.5, modelo de Cox com taxa de falha basal Weibull, as estimativas médias estão próximas do valor real dos parâmetros e conforme o número de conglomerados aumenta, o MSE e rb diminuem. Além disso, o erro padrão sanduíche (SE_2) é próximo ao erro padrão (SD) empírico, indicando que a abordagem proposta efetivamente faz uma correção para a correlação do conglomerado. As probabilidades de cobertura (CP_2) estão próximas do nível nominal de 0,95, indicando uma boa aproximação das estimativas para a distribuição normal.

Tentamos avaliar o efeito do tamanho dos intervalos. Dessa forma, geramos diferentes cenários com tamanhos de intervalos maiores. Os resultados são apresentados na Tabela 5.3 e na Tabela 5.6, para o modelo de Cox com taxa de falha basal Gompertz e modelo de Cox com taxa de falha basal Weibull, respectivamente. As estimativas médias estão próximas do valor real, sendo que para o segundo modelo ficam mais afastadas. O MSE, rb e SD são maiores quando comparados com o primeiro cenário. Além disso, o erro padrão sanduíche (SE_2) está muito próximo do erro padrão (SD) empírico e as probabilidades de cobertura estão próximas do nível nominal de 0,95. Observe que quando aumentamos a correlação dentro do conglomerado, as estimativas dos parâmetros têm um viés um pouco maior, como pode ser observado na Tabela 5.4 e na Tabela 5.7. Observa-se também que o erro padrão empírico é um pouco maior, mas o erro padrão sanduíche ainda está próximo do erro padrão empírico. Em todos os cenários, com o aumento do número de conglomerados a média das estimativas se aproxima do valor real e leva a estimativas menos viesadas e intervalos de cobertura mais próximos do nível nominal.

Tabela 5.5 – Cenário I para modelo de Cox com taxa de falha basal Weibull - Neste cenário $\omega_{ik} \sim Gamma(2,2)$, causando assim uma fraca correlação entre os indivíduos dentro de um mesmo conglomerado (tau de Kendall = 0,20). Além disso, o tamanho dos intervalos foi gerado a partir de $c \sim U(1, 0.1, 0.3)$.

	Est	MSE	rb	SD	SE1	SE2	CP1	CP2
100 conglomerados								
$\beta_{11} = 0.3$	0.270	0.018	-0.099	0.132	0.157	0.128	0.977	0.936
$\beta_{12} = 0.3$	0.276	0.050	-0.088	0.222	0.299	0.223	0.991	0.944
$\beta_{21} = 0.3$	0.245	0.023	-0.183	0.140	0.172	0.137	0.909	0.947
$\beta_{22} = 0.3$	0.248	0.062	-0.173	0.244	0.299	0.237	0.986	0.935
500 conglomerados								
$\beta_{11} = 0.3$	0.271	0.006	-0.096	0.057	0.070	0.057	0.968	0.925
$\beta_{12} = 0.3$	0.276	0.013	-0.088	0.097	0.133	0.099	0.989	0.947
$\beta_{21} = 0.3$	0.248	0.009	-0.175	0.061	0.076	0.059	0.824	0.946
$\beta_{22} = 0.3$	0.249	0.018	-0.174	0.104	0.133	0.107	0.982	0.957
1000 conglomerados								
$\beta_{11} = 0.3$	0.275	0.004	-0.095	0.055	0.070	0.057	0.968	0.925
$\beta_{12} = 0.3$	0.278	0.010	-0.085	0.096	0.103	0.098	0.989	0.947
$\beta_{21} = 0.3$	0.258	0.006	-0.173	0.059	0.070	0.060	0.904	0.926
$\beta_{22} = 0.3$	0.257	0.014	-0.170	0.101	0.103	0.101	0.982	0.929

Tabela 5.6 – Cenário II para modelo de Cox com taxa de falha basal Weibull - Mesma estrutura do Cenário I, exceto que aumentamos o tamanho dos intervalos, usando $c \sim U(1, 0.1, 0.5)$.

	Est	MSE	rb	SD	SE1	SE2	CP1	CP2
100 conglomerados								
$\beta_{11} = 0.3$	0.256	0.018	-0.147	0.129	0.155	0.126	0.976	0.936
$\beta_{12} = 0.3$	0.264	0.049	-0.188	0.219	0.297	0.219	0.990	0.952
$\beta_{21} = 0.3$	0.233	0.025	-0.255	0.138	0.171	0.136	0.942	0.932
$\beta_{22} = 0.3$	0.232	0.064	-0.262	0.241	0.297	0.238	0.983	0.948
500 conglomerados								
$\beta_{11} = 0.3$	0.258	0.005	-0.145	0.056	0.069	0.057	0.952	0.930
$\beta_{12} = 0.3$	0.269	0.013	-0.135	0.096	0.132	0.098	0.989	0.941
$\beta_{21} = 0.3$	0.245	0.012	-0.252	0.062	0.076	0.059	0.849	0.924
$\beta_{22} = 0.3$	0.240	0.017	-0.235	0.105	0.132	0.104	0.983	0.900
1000 conglomerados								
$\beta_{11} = 0.3$	0.260	0.004	-0.143	0.055	0.059	0.055	0.952	0.938
$\beta_{12} = 0.3$	0.269	0.011	-0.132	0.094	0.102	0.094	0.989	0.941
$\beta_{21} = 0.3$	0.247	0.010	-0.250	0.061	0.071	0.058	0.949	0.912
$\beta_{22} = 0.3$	0.245	0.015	-0.230	0.103	0.123	0.103	0.983	0.936

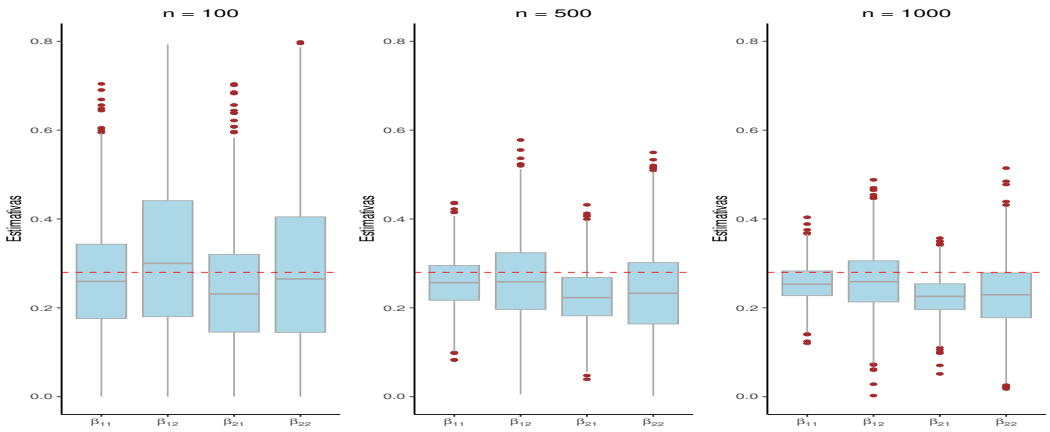
Tabela 5.7 – Cenário III para modelo de Cox com taxa de falha basal Weibull - Mesma estrutura do Cenário I, exceto que aumentamos o tamanho dos intervalos, usando $c \sim U(1, 0.1, 0.5)$.

	Est	MSE	rb	SD	SE1	SE2	CP1	CP2
100 conglomerados								
$\beta_{11} = 0.3$	0.251	0.019	-0.149	0.139	0.155	0.136	0.976	0.936
$\beta_{12} = 0.3$	0.262	0.051	-0.190	0.249	0.297	0.249	0.990	0.952
$\beta_{21} = 0.3$	0.230	0.026	-0.259	0.158	0.171	0.156	0.942	0.932
$\beta_{22} = 0.3$	0.231	0.065	-0.268	0.261	0.297	0.258	0.983	0.948
500 conglomerados								
$\beta_{11} = 0.3$	0.250	0.006	-0.145	0.068	0.069	0.069	0.952	0.930
$\beta_{12} = 0.3$	0.265	0.017	-0.139	0.109	0.132	0.107	0.989	0.941
$\beta_{21} = 0.3$	0.241	0.018	-0.255	0.075	0.076	0.079	0.849	0.924
$\beta_{22} = 0.3$	0.238	0.020	-0.239	0.123	0.132	0.121	0.983	0.900
1000 conglomerados								
$\beta_{11} = 0.3$	0.258	0.005	-0.144	0.066	0.068	0.067	0.952	0.940
$\beta_{12} = 0.3$	0.267	0.014	-0.135	0.106	0.123	0.106	0.989	0.941
$\beta_{21} = 0.3$	0.245	0.015	-0.253	0.072	0.076	0.072	0.949	0.912
$\beta_{22} = 0.3$	0.242	0.013	-0.234	0.113	0.125	0.114	0.983	0.936

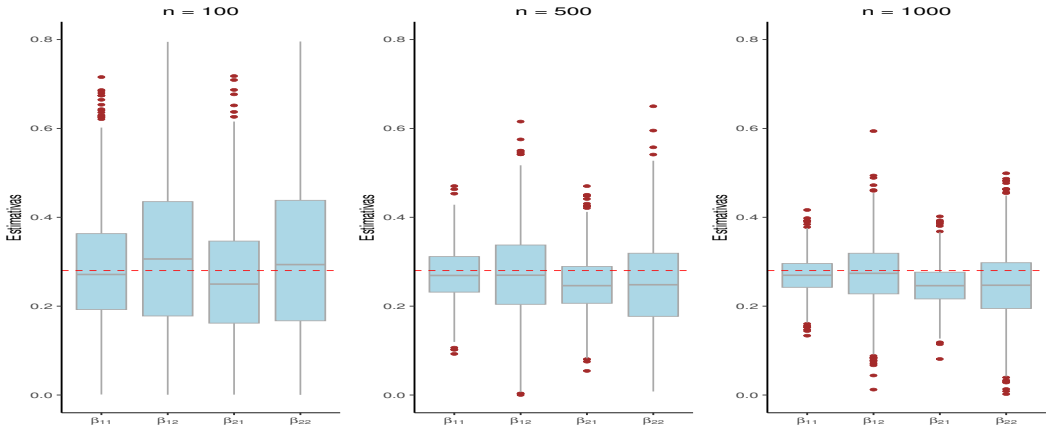
Na Tabela 5.8, tem-se as porcentagens, arredondadas, de vezes em que uma determinada ordem da série de Taylor foi selecionada, utilizando para isso o teste da razão de verossimilhança.

Tabela 5.8 – Ordem ótima (q^*) da série de Taylor em 1000 amostras simuladas

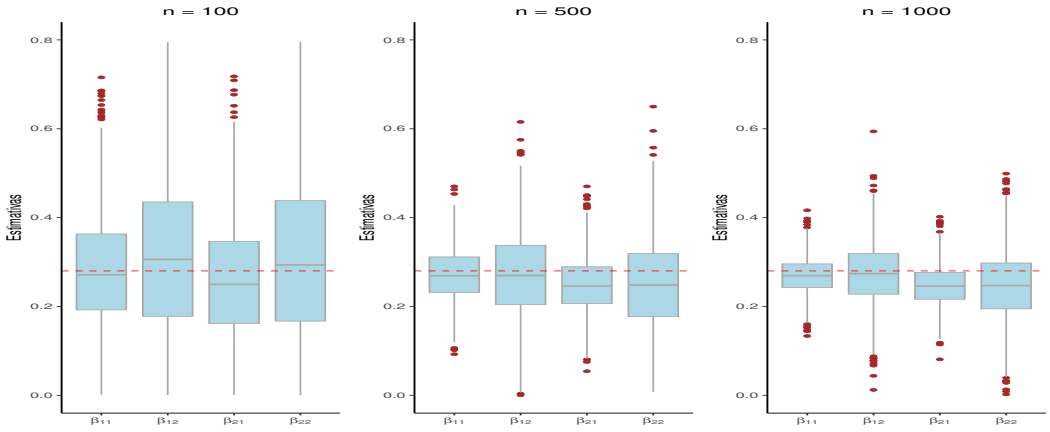
Taxa de Falha	Senários	$q = 0$	$q = 1$	$q = 2$	$q = 3$	$q = 4$
Gompertz	I	21%	51%	12%	10%	6%
	II	20%	53%	18%	4%	3%
	III	22%	58%	10%	6%	4%
Weibull	I	33%	24%	30%	7%	6%
	II	31%	25%	31%	8%	5%
	III	30%	20%	39%	9%	2%



(a) Cenário I

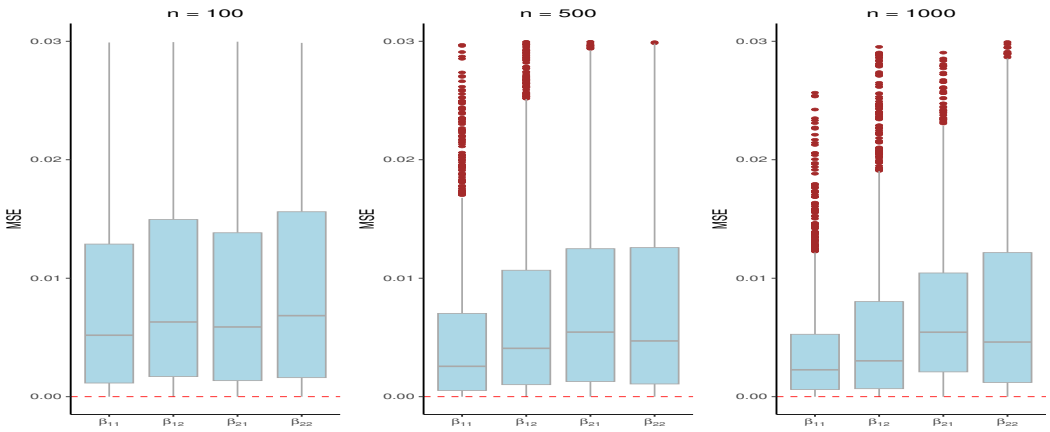


(b) Cenário II

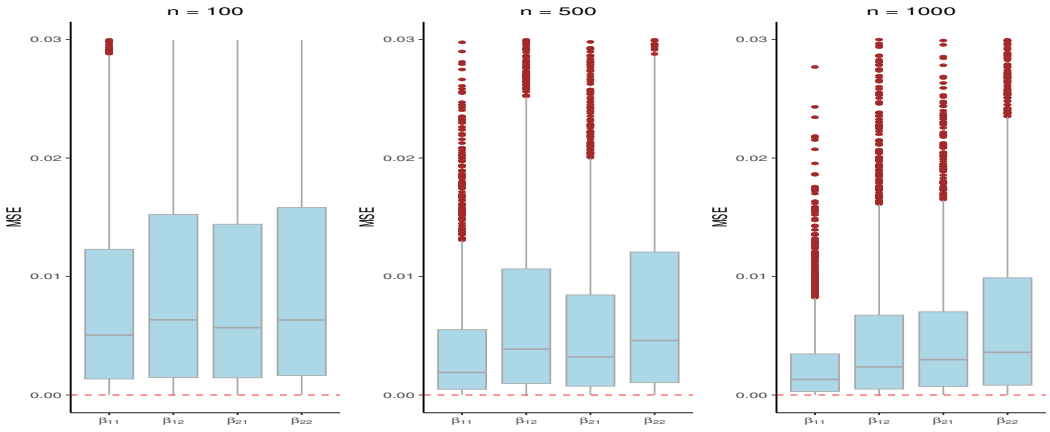


(c) Cenário III

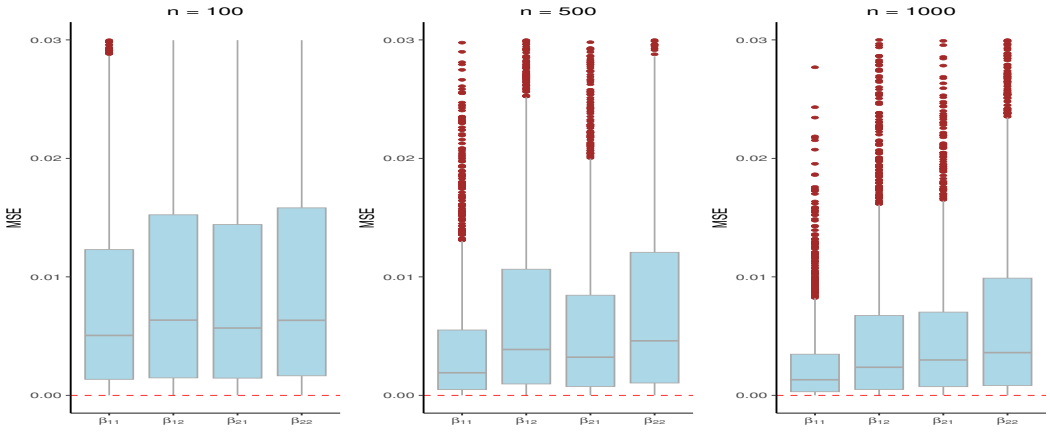
Figura 5.4 – Estimativas dos parâmetros do modelo de Cox com taxa de falha basal Weibull



(a) Cenário I

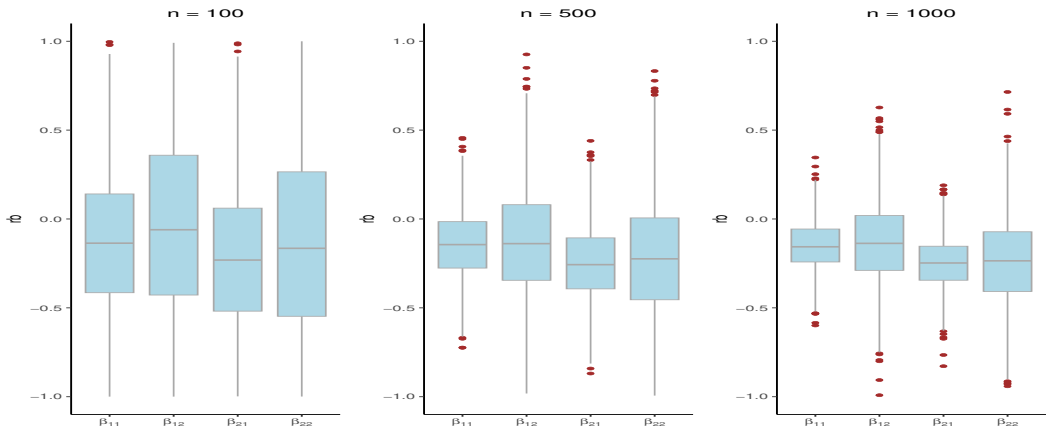


(b) Cenário II

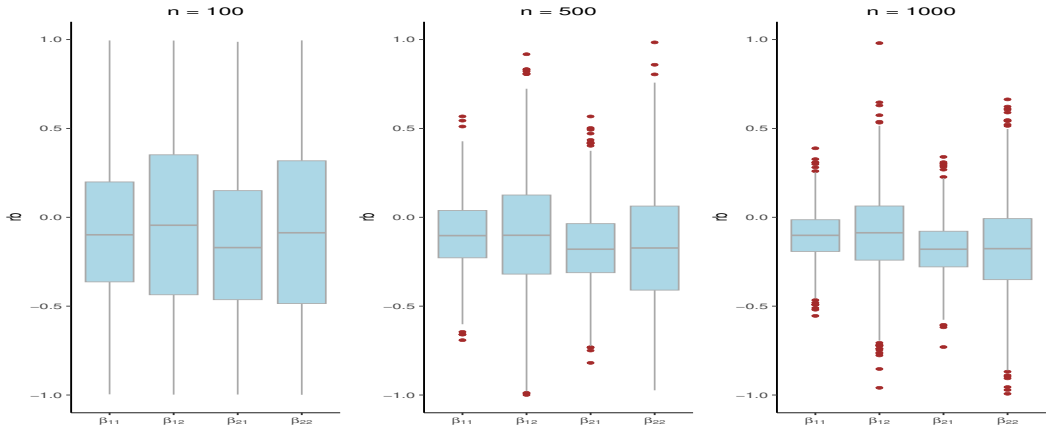


(c) Cenário III

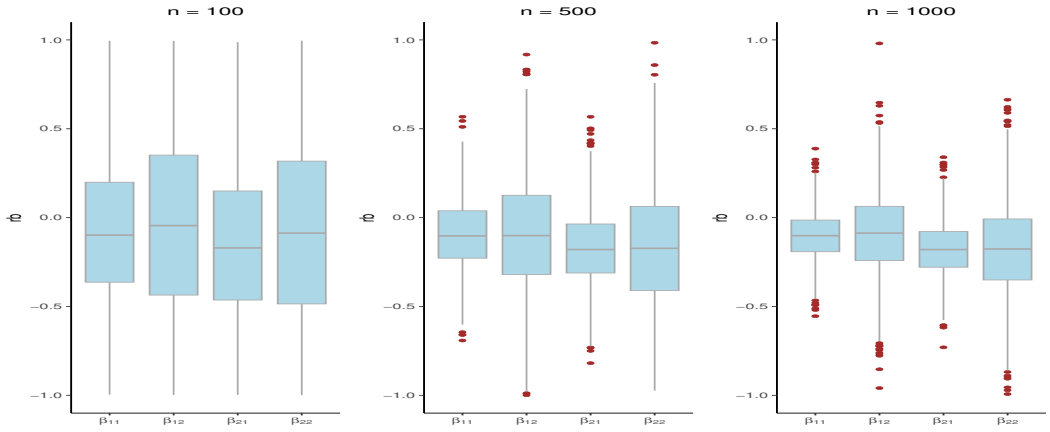
Figura 5.5 – Estimativas de erro quadrático médio (MSE) dos parâmetros modelo de Cox com taxa de falha basal Weibull



(a) Cenário I



(b) Cenário II



(c) Cenário III

Figura 5.6 – Estimativas de viés relativo (rb) dos parâmetros modelo de Cox com taxa de falha basal Weibull

Capítulo 6

Aplicação

Neste capítulo são aplicados os modelos estatísticos apresentados nos Capítulos 4 e 5 ao conjunto de dados de traumatismo dentário proveniente do Programa de Traumatismos Dentários da Faculdade de Odontologia da UFMG. Os dados são de um estudo retrospectivo para avaliar o prognóstico pulpar em dentes permanentes portadores de lesões por luxação. As luxações são classificadas em cinco categorias, de acordo com as estruturas envolvidas e seu grau de comprometimento, a saber: concussão, subluxação, luxação extrusiva, luxação lateral e luxação intrusiva. Conforme discutido na Seção 1.1, a recolocação imediata do dente é o tratamento de escolha após a ocorrência de luxações. Após um reimplante, a cicatrização pulpar pode ocorrer de três maneiras distintas e mutuamente excludentes: manutenção da vitalidade pulpar, obliteração da cavidade pulpar (OCP) e necrose pulpar.

O diagnóstico da cicatrização pulpar foi realizado a partir de dados clínicos associados aos dados radiográficos, não sendo assim possível determinar o instante correto da ocorrência do evento. Sabe-se apenas que o evento ocorreu entre dois exames consecutivos. Outra característica do estudo é que o mesmo paciente pode ter luxação em mais de um dente.

Para as análises e manipulação do banco de dados foi utilizado o software R Core Team (2022), versão 4.2.1.

6.1 Base de dados

O banco de dados desse estudo é composto por 224 pacientes e 427 dentes, distribuídos confome Figura 6.1 abaixo, em que pode-se observar que um pouco mais de 50% dos pacientes tiveram luxação em mais de 1 dente.

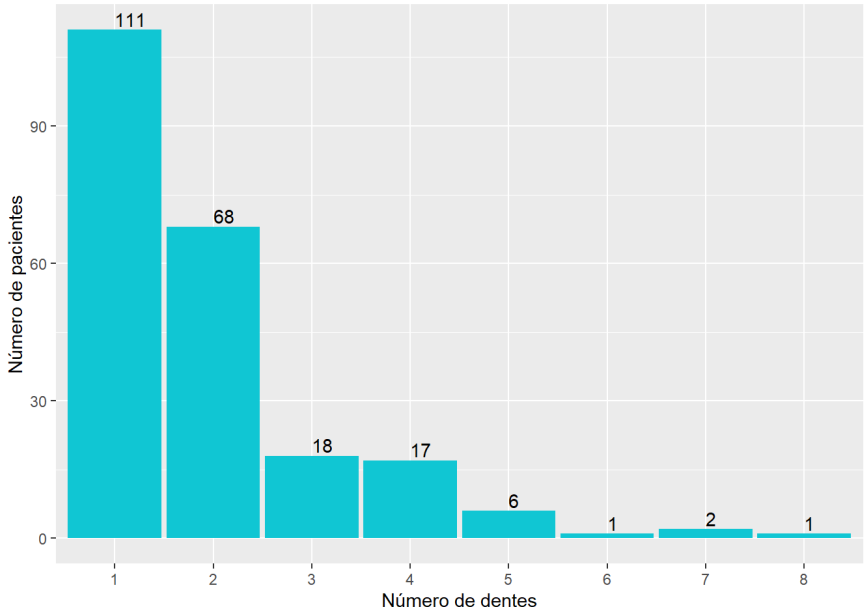


Figura 6.1 – Número de Dentes por pacientes.

A seguir, a Tabela 6.1 tem-se as medidas descritivas do tempo de acompanhamento, em dias, e idade, em anos, desses pacientes.

Tabela 6.1 – Descritivas das variáveis tempo de acompanhamento e idade.

Covariável	Min	1 Quartil	Mediana	Média	3 Quartil	Max
Tempo de acomp.(em dias)	11,7	199,3	509,5	748,9	1177,5	2522,3
Idade (em anos)	6,3	8,5	11	15,5	18,2	61,7

O banco de dados é composto por 26 covariáveis, das quais apenas um subconjunto com as principais variáveis de interesse é utilizado para fins de aplicação nesse trabalho. Elas são apresentadas na Tabela 6.2, a seguir.

Tabela 6.2 – Covariáveis do estudo.

Covariável	Categoria	N	%
Sexo	Masculino	136	60,71
	Feminino	88	39,29
Lesão Associada	Sim	66	15,46
	Não	361	84,54
Tratamento Emergencial	Sim	171	78,08
	Não	48	21,92
Diâmetro Apical	Aberto	138	32,32
	Fechado	289	67,68
Luxação	Lateral	157	36,77
	Subluxação	110	25,76
	Extrusiva	76	17,80
	Intrusiva	56	13,11
	Concussão	28	6,56
Diagnóstico Pulpar	Vitalidade	163	38,17
	Necrose	120	28,10
	PCO	55	12,88
	Censura à direita	89	20,85

Por meio das Tabelas 6.1 e 6.2 pode-se observar que a amostra é predominantemente constituida por crianças e adolecentes, uma vez que 75% deles tem idade igual ou inferior a 18,2 anos. Alem disso, $60,71\%$ são do sexo masculino e 78% tiveram algum tipo de tratamento emergencial. Essas são características a nivel de pacientes, quando olhamos para as características a nível de dente, observamos que $84,54\%$ dos dentes não tiveram nenhuma lesão associada, $67,68\%$ apresentaram diâmetro apical fechado, $36,77\%$ apresentaram luxação lateral e a vitalidade foi o diagnóstico pulpar para $38,17\%$ dos dentes. Observa-se também, a presença de censuras à direita, o que representa $20,85\%$ das observações.

Tabela 6.3 – Distribuição das covariáveis do estudo por Diagnóstico Pulpar.

Covariavel	Categoria	**Diagnóstico Pulpar**					
		Vitalidade		**Necrose**		**PCO**	
		N	%	N	%	N	%
Sexo	Masculino	52	38,24	42	30,88	11	8,09
	Feminino	31	35,23	25	28,41	8	9,09
Tratamento Emergencial	Sim	66	38,60	50	29,24	14	8,19
	Não	15	31,25	17	35,42	4	8,33
Lesão Associada	Sim	8	12,12	46	69,70	4	6,06
	Não	155	42,94	74	20,50	51	14,13
Diâmetro Apical	Aberto	55	39,86	35	25,36	27	19,57
	Fechado	108	37,37	85	29,41	28	9,69
Luxação	Lateral	63	40,13	41	26,11	22	14,01
	Subluxação	54	49,09	22	20,00	9	8,18
	Extrusiva	18	23,68	18	23,68	19	25,00
	Intrusiva	13	23,21	31	55,36	4	7,14
	Concussão	15	53,57	8	28,57	1	3,57

Esses dados apresentaram algumas características que devem ser consideradas na análise estatística. A primeira refere-se à existência de riscos competitivos uma vez que existe mais do que um resultado possível para a resposta pulpar, chamada: manutenção da vitalidade ($k = 1$), necrose ($k = 2$) e obliteração do canal pulpar - PCO ($k = 3$). A segunda característica é a presença de conglomerados, uma vez que a maioria dos pacientes apresentam mais de um dente luxado, conforme discutido anteriormente. Além disso, não foi possível especificar uma data exata de ocorrência do evento, mas um intervalo de tempo entre as consultas de acompanhamento do paciente.

A seguir são apresentados os ajustes dos modelos considerando essas covariáveis. A Seção 6.1.1 apresenta o ajuste dos modelos paramétricos e a Seção 6.1.2 apresenta o ajuste para o modelo semiparametrico.

6.1.1 Resultado dos Modelos Paramétricos

Nessa seção aplicamos a metodologia desenvolvida no Capítulo 4 aos dados de traumatísmo de dentário. Os resultados referentes às estimativas dos coeficientes juntamente com as estimativas do erro padrão sanduíche (SE) são relatados nas Tabelas 6.4 e 6.5 para exponencial e Weibull, respectivamente. O valor-p associado a cada covariável é calculado considerando o teste de Wald.

Tabela 6.4 – Estimativas dos coeficientes do modelo exponencial para três causas concorrentes no estudo de traumatismo dentário

Covariavel	**Vitalidade** ($k=1$)			**Necrose** ($k=2$)			**PCO** ($k=3$)		
	Estimativa	SE	p-value	Estimativa	SE	p-value	Estimativa	SE	p-value
Idade	0.0018	0.0009	0.0541	0.0023	0.0008	0.0044	-0.0076	0.0036	0.0364
Sexo (Mas)	0.1645	0.2356	0.4850	0.2914	0.2377	0.2200	0.3011	0.3613	0.4050
Luxação (Ext+Lat)	-0.1460	0.2384	0.5402	0.2664	0.2807	0.3426	0.9291	0.4384	0.0341
Luxação (Int)	-0.7403	0.4175	0.0762	1.0220	0.3112	0.0010	-0.0082	0.7028	0.9907
Lesão Assoc. (Sim)	-0.9781	0.3469	0.0048	1.4430	0.2064	2.71e-12	-0.4273	0.5140	0.4060
Diam. Apical (Aberto)	-0.0345	0.2181	0.8740	-0.2684	0.2437	0.2710	0.6228	0.3444	0.0706
Treat. Emerg. (sim)	-0.3077	0.2626	0.2410	-0.0335	0.2716	0.9020	-0.2561	0.4202	0.5420

Tabela 6.5 – Estimativas dos coeficientes do modelo Weibull para três causas concorrentes no estudo de traumatismo dentário

Covariavel	**Vitality** ($k=1$)			**Necrosis** ($k=2$)			**PCO** ($k=3$)		
	Estimativa	SE	p-value	Estimativa	SE	p-value	Estimativa	SE	p-value
Idade	0.0220	0.0114	0.0539	0.0281	0.0098	0.0044	-0.0915	0.0437	0.0364
Sexo (Mas)	0.1701	0.2355	0.4750	0.2857	0.2383	0.2310	0.3045	0.3609	0.3990
Luxação (Ext+Lat)	-0.3495	0.2221	0.1155	0.1747	0.2550	0.4930	0.9848	0.3994	0.0137
Luxação (Int)	-0.7467	0.4187	0.0745	1.0205	0.3102	0.0010	-0.0078	0.7028	0.9911
Lesão Assoc. (Sim)	-0.9822	0.3469	0.0046	1.4520	0.2073	2.49e-12	-0.4302	0.5140	0.4030
Diam. Apical (Aberto)	-0.0027	0.2837	0.9920	-0.4569	0.3597	0.2040	0.8663	0.3661	0.0180
Treat. Emerg. (sim)	-0.3454	0.2446	0.1580	-0.0355	0.2616	0.8020	-0.3366	0.4621	0.4660
τ	1.2250	0.1381	-	0.5050	0.2561	-	1.8370	0.2967	-

Tanto o modelo exponencial quanto o modelo Weibull indicam as mesmas covariáveis como significativas. A variável tipo de luxação apresenta efeito significativo para os eventos Necrose e PCO. A variável lesão associada apresenta efeito significativo para

a ocorrência dos eventos Vitalidade e Necrose e a variável diâmetro do forame apical tem efetio significativo para a ocorrência do evento PCO. Evidências experimentais sugerem que dentes com forame apical mais largo são mais propensos a sofrer crescimento vascular e regeneração nervosa, o que permite a preservação da vitalidade pulpar. Dados clínicos confirmaram essa premissa e também demonstraram que o risco de necrose pulpar aumentava com a gravidade da luxação e na presença de fraturas coronárias concomitantes. Além disso, estudos clínicos mostraram que a obliteração do canal pulpar (PCO) é mais frequentemente observada em lesões que envolvem deslocamento de dentes permanentes imaturos (Clark e Levin, 2019; Darley et al., 2020; Spinas et al., 2021). Pela Tabela 6.5, observa-se que as estimativas do parâmetro τ são próximas de 1 apenas para o evento vitalidade. Testando a hipótese de que τ é igual a 1, observa-se que não é significativo para os eventos necrose e PCO. Portanto, há uma indicação favorável para o modelo Weibull.

6.1.2 Resultado do Modelo Semiparamétrico

Nessa seção aplicamos a metodologia desenvolvida no capítulo 5 aos dados de traumatímo dentário. Ajustamos a abordagem de expansão da série de Taylor com diferentes ordens q. Utilizando o teste da razão de verossimilhança encontramos que o q ótimo é igual a 1 para os eventos Vitalidade e Necrose, e igual a zero para o evento PCO. Os resultados referentes às estimativas dos coeficientes juntamente com as estimativas do erro padrão sanduíche (SE) são relatados na Tabela 6.6, a seguir.

Tabela 6.6 – Estimativas dos parâmetros do modelo semiparamétrico para três causas concorrentes no estudo de traumatismo dentário

Covariavel	**Vitalidade** ($k = 1$)			**Necrose** ($k = 2$)			**PCO** ($k = 3$)		
	Estimativa	SE	p-value	Estimativa	SE	p-value	Estimativa	SE	p-value
Idade	0.0250	0.0134	0.0542	0.0295	0.0128	0.0044	-0.0819	0.0349	0.0364
Sexo (Mas)	0.2701	0.2358	0.4950	0.2937	0.2493	0.2110	0.3241	0.3589	0.3990
Luxação (Ext+Lat)	-0.3095	0.2232	0.1215	0.1947	0.2652	0.4930	0.8948	0.3784	0.0137
Luxação (Int)	-0.7167	0.4287	0.0735	1.0301	0.3702	0.0012	-0.0101	0.7121	0.9911
Lesão Assoc. (Sim)	-0.9321	0.3489	0.0043	1.5110	0.2163	2.49e-10	-0.4501	0.5148	0.4030
Diam. Apical (Aberto)	-0.0019	0.2847	0.9810	-0.4668	0.3696	0.2140	0.8861	0.3551	0.0180
Treat. Emerg. (sim)	-0.3152	0.2446	0.1490	-0.0451	0.3613	0.7928	-0.3476	0.4728	0.4660

Observamos que o modelo semiparamétrico forneceu resultados semelhantes ao modelo paramétrico. A variável Luxação é significativa apenas para os eventos Necrose e PCO, sendo que a ocorrência de luxação intrusiva aumenta em 2,8 (exp[1, 0301]) vezes o risco de ocorrer necrose quando comparado com a ocorrência de subluxação e a ocorrência de luxação (Ext+Lat) aumenta em 2,45 (exp[0, 8948]) o risco de ocorrer PCO em relação a ocorrência de subluxação. A variável lesão associada é significativa apenas para os eventos Vitalidade e Necrose, de modo que um paciente com lesão associada tem um risco 0,39 (exp[−0, 9321]) vezes menor de ocorrer Vitalidade do que um paciente sem lesão associada e possui um risco 4,53 (exp[1.5110]) vezes maior de ocorrer Necrose quando comparado com aqueles que não possui lesão associada. A última variável a apresentar efeito significativo é a variável diâmetro apical, sendo significativa apenas para PCO. Assim, um paciente com diâmetro apical aberto possui risco 2,42 (exp[0.8861]) vezes maior de ocorrer PCO do que aqueles com diâmetro apical fechado.

Capítulo 7

Considerações Finais

Nessa tese são propostos modelos para análise de dados de sobrevivência correlacionados na presença de riscos competitivos e censura intervalar. Pesquisas iniciais na literatura da área apontaram a inexistência de metodologias para lidar com dados dessa natureza. Foi proposto inicialmente um modelo paramétrico em que foram considerados as distribuições exponencial e Weibull para a função taxa de falha. O segundo modelo proposto é um modelo semiparamétrico em que foi utilizado a abordagem de aproximação em série de Taylor para a função taxa de falha basal.

Embora façam suposições sobre a distribuição dos dados, os modelos paramétricos oferecem alguma flexibilidade na escolha das distribuições, permitindo assim adaptar o modelo às características específicas dos dados. Quando os dados são descritos por uma distribuição específica, os métodos paramétricos podem ser mais eficientes e pode resultar em estimativas mais precisas dos parâmetros.

Ao contrário dos modelos paramétricos, que assumem uma distribuição específica para os tempos dos eventos, os modelos semiparamétricos não dependem de suposições rígidas sobre uma distribuição subjacente. Isso torna esses modelos mais flexíveis e mais robustos, e particularmenente útil quando a distribuição subjacente dos tempos dos eventos não é conhecida com precisão. Devido à sua capacidade de combinar características paramétricas e não paramétricas, esses modelos podem ser adequados para uma ampla gama de cenários e tipos de dados. No entanto, é importante notar que a modelagem semiparamétrica também tem desafios, como a possibilidade de complexidade computa-

cional maior.

Hudgens et al. (2014) apresentaram um modelo paramétrico para dados de riscos competitivos com censura intervalar. Baseado no trabalho de Hudgens et al. (2014), propomos um modelo paramétrico que, além de lidar com riscos competitivos e censura intervar, pudesse acomodar a presença conglomerados. Para isso, modelos paramétricos são especificados para as funções taxa de falha causa-específicas e assim, a FIA é modelada indiretamente.

O segundo modelo proposto é um modelo de regressão causa específica para dados de sobrevivência correlacionados na presença de riscos competitivos e censura intervalar. Essencialmente, o modelo de regressão causa específica é uma extensão da regressão de Cox tradicional para cada tipo de evento, em que as falhas de eventos concorrentes são tratados como observações censuradas. A incorporação da censura intervalar no modelo de riscos proporcionais não permite o cancelamento da função de taxa de falha linha de base, assim como acontece no caso de censura a direita, e como resultado, distribuição da linha de base e os parâmetros de regressão são ajustados simultaneamente, maximizando a verossimilhança total dos dados. Para isso, utilizamos a abordagem proposta por Chen et al. (2013) que utilizaram a série de Taylor para aproximar a função linha de base. Essa abordagem tem a vantagem de produzir funções suaves enquanto que as outras abordagens produzem funções do tipo degrau.

Tanto para o modelo paramétrico quanto para o semiparamétrico, propusemos uma abordagem marginal para lidar com a presença de conglomerados. A abordagem marginal proposta apresenta vantagens: o método de estimação é de fácil execução e o modelo não depende da especificação de uma estrutura de correlação e, portanto, é robusto. Essa abordagem é particularmente útil se o efeito médio da população for de interesse principal e a estrutura de correlação não for de interesse ou não puder ser especificada adequadamente devido à falta de informações suficientes. Utilizamos um estimador de variância sanduíche para estimar a matriz de covariância de $\hat{\boldsymbol{\Theta}}_k$. Para obter o estimador de variância sanduíche tratamos a equação escore, utilizando a primeira derivada da função log verossimilhança, com equação de estimação de forma semelhante ao GEE proposta por Liang e Zeger (1986).

Neste trabalho foi realizado um exaustivo estudo de simulação considerando todos os modelos aqui propostos. Os dados simulados foram gerados em três cenários, de modo que no cenário 2, verificamos a influência do comprimento dos intervalos cencurados e para isso, mantem-se a mesma estrutura do cenário 1 aumentando apenas o comprimento dos intervalos censurados. Enquanto que no cenário 3, verificamos a influência da estrutura de correlação intra-conglomerados. Para isso mantivemos a mesma estrutura do cenário 1, aumentando apenas a correlação intra-conglomerado. No contexto dos cenários avaliados, notou-se que as estimativas ficaram mais próximas do verdadeiro valor a medida que o tamanho da amostra aumenta.

Sobre a análise com dados reais, em relação aos dados de traumatismo dentário, observamos que os modelos paramétricos e semiparamétricos tem resultados parecidos e indicam as covariáveis como significativas para o diagnóstico pulpar o que vai de encontro com a literatura da área de odontologia conforme pode ser observado em Clark e Levin (2019), Darley et al. (2020) e Spinas et al. (2021).

Como trabalhos futuros planejamos implementar outra forma para modelar a função taxa de falha basal, como por exemplo utilizando splines. Além disso, pretendemos desenvolver o modelo semiparamétrico de Fine-Gray com censura intervalar para dados em conglomerados.

Referências

Bakoyannis, G. e Touloumi, G. (2011), "Practical methods for competing risks data: A review," *Statistical Methods in Medical Research*, 21, 257–272.

Bender, R., Augustin, T., e Blettner, M. (2005), "Generating survival times to simulate Cox proportional hazards models," *Statistics in Medicine*, 24, 1713–1723.

Beyersmann, J., Latouche, A., Buchholz, A., e Schumacher, M. (2009), "Simulating competing risks data in survival analysis," *Statistics in Medicine*, 28, 956–971.

Beyersmann, J., Allignol, A., e Schumacher, M. (2011), *Competing risks and multistate models with R*, Springer Science & Business Media.

Bogaerts, K., Leroy, R., Lesaffre, E., e Declerck, D. (2002), "Modelling tooth emergence data based on multivariate interval-censored data," *Statistics in Medicine*, 21, 3775–3787.

Bogaerts, K., Komarek, A., e Lesaffre, E. (2017), *Survival Analysis with Interval-Censored Data : a Practical Approach with Examples in R, SAS, and BUGS*, Chapman and Hall/CRC, City.

Brazauskas, R. e Le-Rademacher, J. (2016), "Methods for generating paired competing risks data," *Computer methods and programs in biomedicine*, 135, 199–207.

Breslow, N. E. (1972), "Contribution to discussion of paper by DR Cox," *J. Roy. Statist. Soc., Ser. B*, 34, 216–217.

Cai, J. e Prentice, R. L. (1997)*Lifetime Data Analysis*, 3, 197–213.

Cai, T. e Betensky, R. A. (2003), "Hazard Regression for Interval-Censored Data with Penalized Spline," *Biometrics*, 59, 570–579.

Chen, D.-G., Yu, L., Peace, K. E., Lio, Y. L., e Wang, Y. (2013), "Approximating the Baseline Hazard Function by Taylor Series for Interval-Censored Time-to-Event Data," *Journal of Biopharmaceutical Statistics*, 23, 695–708.

Chen, M.-H., Tong, X., e Sun, J. (2007), "The proportional odds model for multivariate interval-censored failure time data," *Statistics in Medicine*, 26, 5147–5161.

Clark, D. e Levin, L. (2019), "Prognosis and complications of mature teeth after lateral luxation," *The Journal of the American Dental Association*, 150, 649–655.

Collett, D. (2015), *Modelling survival data in medical research*, CRC press.

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

Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–202.

Crowder, M. J. (2001), *Classical competing risks*, CRC Press.

Darley, R. M., e Silva, C. F., dos Santos Costa, F., Xavier, C. B., e Demarco, F. F. (2020), "Complications and sequelae of concussion and subluxation in permanent teeth: A systematic review and meta-analysis," *Dental Traumatology*, 36, 557–567.

Diggle, P., Heagerty, P., Liang, K.-Y., e Zeger, S. (2013), *Analysis of Longitudinal Data*, Oxford University Press.

Emura, T., Matsui, S., e Rondeau, V. (2019), *Survival Analysis with Correlated Endpoints: Joint Frailty-Copula Models*, Springer.

Fine, J. P. e Gray, R. J. (1999), "A proportional hazards model for the subdistribution of a competing risk," *Journal of the American statistical association*, 94, 496–509.

Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42, 845.

Goggins, W. B. e Finkelstein, D. M. (2000), "A Proportional Hazards Model for Multivariate Interval-Censored Failure Time Data," *Biometrics*, 56, 940–943.

Gómez, G., Calle, M. L., Oller, R., e Langohr, K. (2009), "Tutorial on methods for interval-censored data and their implementation in R," *Statistical Modelling*, 9, 259–297.

Grambauer, N., Schumacher, M., e Beyersmann, J. (2010), "Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified," *Statistics in Medicine*, 29, 875–884.

Gray, R. J. (1988), "A class of K-sample tests for comparing the cumulative incidence of a competing risk," *The Annals of statistics*, pp. 1141–1154.

Groeneboom, P., Maathuis, M. H., e Wellner, J. A. (2008a), "Current status data with competing risks: Consistency and rates of convergence of the MLE," *The Annals of Statistics*, 36, 1031–1063.

Groeneboom, P., Maathuis, M. H., e Wellner, J. A. (2008b), "Current status data with competing risks: Limiting distribution of the MLE," *The Annals of Statistics*, 36, 1064–1089.

Haller, B., Schmidt, G., e Ulm, K. (2013), "Applying competing risks regression models: an overview," *Lifetime data analysis*, 19, 33–58.

Hanagal, D. D. (2011), *Modeling survival data using frailty models*, Springer.

Heller, G. (2010), "Proportional hazards regression with interval censored data using an inverse probability weight," *Lifetime Data Analysis*, 17, 373–385.

Huang, J. e Wellner, J. A. (1993), "Regression models with interval censoring," *Probability theory and mathematical statistics (St. Petersburg, 1993)*, pp. 269–296.

Huber, P. J. et al. (1967), "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the fifth Berkeley symposium on mathematical*

statistics and probability, vol. 1, pp. 221–233, Berkeley, CA: University of California Press.

Hudgens, M. G., Li, C., e Fine, J. P. (2014), "Parametric likelihood inference for interval censored competing risks data," *Biometrics*, 70, 1–9.

Huster, W. J., Brookmeyer, R., e Self, S. G. (1989), "Modelling Paired Survival Data with Covariates," *Biometrics*, 45, 145.

Jeong, J.-H. e Fine, J. (2006), "Direct parametric inference for the cumulative incidence function," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55, 187–200.

Jeong, J.-H. e Fine, J. P. (2007), "Parametric regression on cumulative incidence function," *Biostatistics*, 8, 184–196.

Jewell, N. P. (2003), "Nonparametric estimation from current status data with competing risks," *Biometrika*, 90, 183–197.

Joly, P., Commenges, D., e Letenneur, L. (1998), "A Penalized Likelihood Approach for Arbitrarily Censored and Truncated Data: Application to Age-Specific Incidence of Dementia," *Biometrics*, 54, 185.

Kalbfleisch, J. D. e Prentice, R. L. (2011), *The statistical analysis of failure time data*, John Wiley & Sons.

Kaplan, E. L. e Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Kim, M. Y. e Xue, X. (2002), "The analysis of multivariate interval-censored survival data," *Statistics in Medicine*, 21, 3715–3726.

Kooperberg, C. e Clarkson, D. B. (1997), "Hazard Regression with Interval-Censored Data," *Biometrics*, 53, 1485.

Kor, C.-T., Cheng, K.-F., e Chen, Y.-H. (2012), "A method for analyzing clustered interval-censored data based on Cox's model," *Statistics in Medicine*, 32, 822–832.

Lawless, J. F. (2011), *Statistical models and methods for lifetime data*, vol. 362, John Wiley & Sons.

Lawless, J. F. e Babineau, D. (2006), "Models for interval censoring and simulation-based inference for lifetime distributions," *Biometrika*, 93, 671–686.

Lee, E. W., Wei, L. J., Amato, D. A., e Leurgans, S. (1992), "Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations," in *Survival Analysis: State of the Art*, pp. 237–247, Springer Netherlands.

Li, C. (2016a), "Cause-specific hazard regression for competing risks data under interval censoring and left truncation," *Computational Statistics & Data Analysis*, 104, 197–208.

Li, C. (2016b), "The Fine–Gray model under interval censored competing risks data," *Journal of Multivariate Analysis*, 143, 327–344.

Li, C. e Fine, J. P. (2013), "Smoothed nonparametric estimation for current status competing risks data," *Biometrika*, 100, 173–187.

Liang, K.-Y. e Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Liang, K. Y. e Zeger, S. L. (1993), "Regression Analysis for Correlated Data," *Annual Review of Public Health*, 14, 43–68.

Lin, D. Y. e Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association*, 84, 1074–1078.

Lindsey, J. (1998), "A study of interval censoring in parametric regression models," *Lifetime data analysis*, 4, 329–354.

Liu, X. (2012), *Survival Analysis: Models and Applications*, John Wiley & Sons.

Logan, B. R., Zhang, M.-J., e Klein, J. P. (2010), "Marginal Models for Clustered Time-to-Event Data with Competing Risks Using Pseudovalues," *Biometrics*, 67, 1–7.

Marubini, E. e Valsecchi, M. G. (2004), *Analysing survival data from clinical trials and observational studies*, vol. 15, John Wiley & Sons.

Mozumder, S. I., Rutherford, M. J., e Lambert, P. C. (2017), "A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function," *The Stata Journal*, 17, 462–489.

Pan, W. (1999), "Extending the iterative convex minorant algorithm to the Cox model for interval-censored data," *Journal of Computational and Graphical Statistics*, 8, 109–120.

Pan, W. (2000), "A Multiple Imputation Approach to Cox Regression with Interval-Censored Data," *Biometrics*, 56, 199–203.

Peng, Y., Taylor, J. M. G., e Yu, B. (2007), "A marginal regression model for multivariate failure time data with a surviving fraction," *Lifetime Data Analysis*, 13, 351–369.

Pintilie, M. (2006), *Competing risks: a practical perspective*, vol. 58, John Wiley & Sons.

Pintilie, M. (2007), "Analysing and interpreting competing risk data," *Statistics in Medicine*, 26, 1360–1367.

Porta Bleda, N., Gómez Melis, G., e Calle Rosingana, M. L. (2008), "The role of survival functions in competing risks," resreport, Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, http://hdl.handle.net/2117/2202, Relatório de Pesquisa.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., e Breslow, N. E. (1978), "The Analysis of Failure Times in the Presence of Competing Risks," *Biometrics*, 34, 541.

Putter, H., Fiocco, M., e Geskus, R. B. (2007), "Tutorial in biostatistics: competing risks and multi-state models," *Statistics in Medicine*, 26, 2389–2430.

R Core Team (2022), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Robins, J. M. e Rotnitzky, A. (1992), "Recovery of information and adjustment for dependent censoring using surrogate markers," in *AIDS epidemiology*, pp. 297–331, Springer.

Royall, R. M. (1986), "The Prediction Approach to Robust Variance Estimation in Two-Stage Cluster Sampling," *Journal of the American Statistical Association*, 81, 119–123.

Santos Junior, P. C. (2016), "Modelos semiparamétricos para dados de sobrevivência com censura intervalar," phdthesis, Departamento de Estatística, Universidade Federal de Minas Gerais, Tese de Doutorado em Estatística.

Satten, G. (1996), "Rank-based inference in the proportional hazards model for interval censored data," *Biometrika*, 83, 355–370.

Spiekerman, C. F. e Lin, D. Y. (1998), "Marginal Regression Models for Multivariate Failure Time Data," *Journal of the American Statistical Association*, 93, 1164–1175.

Spinas, E., Deias, M., Mameli, A., e Giannetti, L. (2021), "Pulp canal obliteration after extrusive and lateral luxation in young permanent teeth: A scoping review," *European Journal of Paediatric Dentistry*, 22, 55–60.

Sun, J. (1996), "A Non-Parametric Test For Interval-Censored Failure Time Data With Application To AIDS Studies," *Statistics in Medicine*, 15, 1387–1395.

Sun, J. (2006), *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer Nature.

Therneau, T. M. e Grambsch, P. M. (2000), "Expected survival," in *Modeling survival data: extending the Cox model*, pp. 261–287, Springer.

Tsiatis, A. (1975), "A nonidentifiability aspect of the problem of competing risks," *Proceedings of the National Academy of Sciences*, 72, 20–22.

Turnbull, B. W. (1976), "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 38, 290–295.

Wei, L. J., Lin, D. Y., e Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association*, 84, 1065–1073.

Wienke, A. (2010), *Frailty models in survival analysis*, CRC press.

Zeger, S. L., Liang, K.-Y., e Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049.

Zhao, Q. e Sun, J. (2004), "Generalized log-rank test for mixed interval-censored failure time data," *Statistics in Medicine*, 23, 1621–1629.

Zhou, B., Fine, J., Latouche, A., e Labopin, M. (2012), "Competing risks regression for clustered data," *Biostatistics*, 13, 371–383.

Apêndice

Apêndice A

Artigo submetido: Parametric Model for Correlated Survival Data in the Presence of Competing Risks and Interval Censored

Lifetime Data Analysis

Parametric Model for Correlated Survival Data in the Presence of Competing Risks and Interval Censored

--Manuscript Draft--

Manuscript Number:	
Full Title:	Parametric Model for Correlated Survival Data in the Presence of Competing Risks and Interval Censored
Article Type:	General Manuscript Submission
Keywords:	Clusters; GEE; Sandwich matrix; Weibull model; Work matrix.
Corresponding Author:	Marcio Ferreira Rodrigues, Ph.D. UFG: Universidade Federal de Goias Goiânia, Goias BRAZIL
Corresponding Author Secondary Information:	Augusto
Corresponding Author's Institution:	UFG: Universidade Federal de Goias
Corresponding Author's Secondary Institution:	
First Author:	Marcio Augusto Ferreira Rodrigues, Ph.D.
First Author Secondary Information:	Augusto
Order of Authors:	Marcio Augusto Ferreira Rodrigues, Ph.D.
	Enrico Antonio Colosimo, Ph.D.
	Juliana Vilela Bastos, Ph.D.
	Sylvia Cury Coste, Ph.D.
Order of Authors Secondary Information:	Augusto
Funding Information:	

Parametric Model for Correlated Survival Data in the Presence of Competing Risks and Interval Censored

Marcio A. F. Rodrigues[1], marcioaugusto@ufg.br,
[0000-0002-8746-8428],

Enrico A. Colosimo[2], enricoc57@gmail.com,
[0000-0001-8705-4674],

Juliana V. Bastos[3], julianavbtrauma@gmail.com
[0000-0002-2062-2566],

Sylvia C. Coste[3], sylviacury@hotmail.com
[0000-0003-3344-1585]

[1]Institue of Mathematics and Statistics, Universidade Federal de Goias, Goiania, Brazil
[2]Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[3]Restorative Dentistry Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

October 23, 2023

Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

Parametric Model for Correlated Survival Data in the Presence of Competing Risks and Interval Censored

Marcio A. F. Rodrigues · Enrico A. Colosimo · Juliana V. Bastos · Sylvia C. Coste

Received: date / Accepted: date

Abstract Interval censored in survival analysis occurs when the survival time of the individuals under study is known to only belong to a time interval. In classic survival analysis, a single cause for the occurrence of an event is considered. However, some studies may be interested in the occurrence of more than one event, which is called competing risks. Hudgens et al. (2014) considered the parametric modeling of the cumulative incidence function for competing risk data subject to interval censored. In this study, we extend the work of Hudgens et al. (2014) to accommodate the presence of clusters. The proposed methodology considers a Generalized Estimating Equations (GEE)-type model using an independent work matrix. The model performance was assessed using a simulated dataset. We demonstrated that the proposed methodology has good properties for small samples. The proposed method was applied to a real dental trauma dataset.

Keywords Clusters · GEE · Sandwich matrix · Weibull model · Work matrix

Institue of Mathematics and Statistics, Universidade Federal de Goias, Goiania, Brazil
first address
E-mail: marcioaugusto@ufg.br

Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
second address
E-mail: enricoc57@gmail.com

Restorative Dentistry Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
third address
E-mail: julianavbtrauma@gmail.com

Restorative Dentistry Department, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
forth address
E-mail: sylviacury@hotmail.com

1 Introduction

Traumatic dental injuries (TDIs) are unexpected and acute events caused by an abrupt impact, primarily to the anterior teeth. It can be considered a neglected public health concern due to their high prevalence worldwide; complex, prolonged, and expensive treatment; and great impact on the quality of life of patients and their families (Petti et al., 2018). TDI can simultaneously affect the mineralized tissues of the tooth, its supporting structures, and the dental pulp, with a multitude of possible trauma scenarios and complex healing patterns. The motivating data for this paper originated from a longitudinal study conducted at the dental trauma clinic at the Federal University of Minas Gerais, Brazil. That study aimed to evaluate pulp prognosis as well as its prognostic factors in permantent luxated teeth.

Pulp response after an acute TDI depends on the competition between the ingrowth of a neurovascular supply into the traumatized pulp tissue and bacterial invasion. The final pulp outcome can be pulp death (necrosis) or revascularization with the maintenance of pulp vitality, eventually followed by pulp canal obliteration (PCO). There are intermediate processes in pulpal response. Conclusive diagnosis of the pulp status can only be achieved in the long term. Therefore, survival analysis is used in clinical dental studies to account for the different probabilities of the event by considering variations in the observation periods (Andreasen and Andreasen, 1990). In this study, there are three mutually excluding possible outcomes/events: maintenance of vitality, necrosis, and PCO. As it was not possible to specify an exact date of occurrence for the event, a time interval between the patient's follow-up appointments, gave rise to interval-censored observations. Finally, as the same patient may have more than one traumatized tooth, characterized as a cluster, the outcomes were measured for each tooth individually.

The first issue in the present dataset is the competitive risk aspect of observed outcomes. The lack of differentiation between event types might generate bias estimates related to the covariate effects and lead to inaccurate interpretations of results. Several methods have been proposed for analyzing competitive risk data. Early work primarily focused on estimating and modeling cause-specific risk or the instantaneous risk of an event (Prentice et al., 1978). To study covariates effects on the cumulative probability of a particular cause, with independent individuals, Fine and Gray (1999) proposed a proportional sub-distribution risk model. This model has a direct correspondence with the cumulative incidence function (CIF). Sun (2006) explored an additive risk model.

The second aspect to be considered is that the timing of the event was not accurately observed. However, it was known to occur within a time interval, that is between two clinical visits, configuring interval-censored observations. Interval-censored concurrent risk data can arise frequently in many applications. Jeong and Fine (2006) proposed a CIF parametric for right-censored competitive risk data using Weibull and Gompertz distributions. Jeong and Fine (2007) presented a parametric regression model using the

Gompertz distribution in the same scenario. Hudgens et al. (2014) extended Jeong and Fine's (2006; 2007) models to the case of concurrent risks with interval-censored.

Lastly, concurrent risk data cannot be considered independent. Hence, appropriate methods are needed to explain the correlation among subjects within the same cluster. The unknown dependence structure, within the cluster, requires appropriate regression methods. These consider correlations in a robust way to allow valid inference for the effects of the covariates on incidence function of the event of interest. Conditional and marginal models represent two alternative strategies for modeling the data of competing risks in clusters. Conditional models specify random effects to measure correlations between failure observations and have a cluster-specific interpretation of fixed effects parameters (Lee and Nelder, 2004). In this approach, frailty models have gained great prominence (Hanagal, 2011; Wienke, 2010). Although frailty models are flexible, insofar as they explicitly model heterogeneity among clusters, valid inferences for the fixed-effects parameters necessarily depend on the correct specification of the frailty distribution. However, the marginal model specifies the effects of covariates across the entire population of clusters, without the need to specify unobserved correlation structure (Liang and Zeger, 1986). This model is computationally less intensive and, especially in the context of our application, has population interpretation. In the competitive risk scenario, Zhou et al. (2012) proposed a marginal subdistribution model and an estimation strategy assuming an independent working correlation structure for right-censored data. A sandwich variance estimator was developed to accommodate the unknown dependence within the cluster. Bogaerts et al. (2002) developed a sandwich variance estimator for observations in interval-censored clusters. They used the model with an independent work matrix in a classical survival analysis, that is, without a competitive risks structure. With this estimator, inferences to the marginal regression models are generally robust to assumptions of intra-cluster correlations.

To the best of our knowledge statistical methodologies that simultaneously consider the structure of competing risks, clusters, and interval censoring, have not been investigated in the existing literature. Thus, we present a parametric model for correlated data in the presence of competing risks and interval censoring. The variance of the parameters estimates is corrected for clustering using a sandwich estimate.

The remaining paper is organized as follows. In Section 2, we present the model for interval-censored competing risks data proposed by Hudgens et al. (2014). In Section 3, we propose a methodology that extends the work of Hudgens et al. (2014) to include correlated data. Simulation studies are generated to evaluate small sample properties of the model in Section 4. In Section 5, we illustrate the proposed method using a dental trauma real data. Finally, in Section 6, we present final remarks.

2 Interval censored competing risks data

The observed data for each individual in a competitive risk model can be represented by a pair of random variables (T, C). The variable C assumes value zero if the individual's observation is right-censored and assumes the value k, otherwise, where k is the type of observed failure cause ($k = 1, 2, \cdots, K$). If $C = k$, T corresponds to the time until the failure due to the cause k; otherwise, T is the time until censorship. Each subject is at risk of failure due to different possible causes, but the occurrence of one cause of failure precludes all other causes of failure. Let $\{1, ..., K\}$ be a set of mutually exclusive concurrent events and K failure times, $T_1, ..., T_K$, one for each type of event. Only the minimum failure time is observed, $T = \min\{T_1, ..., T_K\}$. The joint distribution of (T, C) is completely specified through either the cumulative incidence functions, $F_k(t)$, or through the cause-specific hazards, $\lambda_k(t)$ (Lawless, 2011).

The cumulative incidence function for the kth event is defined by

$$F_k(t) = P(T \leq t, C = k), \quad \text{for } k = 1, ..., K,$$

this corresponds to the probability of occurrence of the kth cause of failure in the presence of the other causes of failure. It is also known that

$$F_k(t) = \int_0^t \lambda_k(u) S(u) du \tag{1}$$

where $S(t) = P(T > t) = \exp\left(-\int_0^t \sum_{l=1}^{K} \lambda_l(u) du\right)$ is the survival function of all causes and

$$\lambda_k(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, C = k | T > t)}{\Delta t}, \qquad k = 1, ..., K \tag{2}$$

is the cause-specific hazard function for the kth cause. The function $\lambda_k(t)$ determines the instantaneous failure rate for an event of type k at time t in the presence of the other failure causes. Thus, the CIF can be written as a function of specific cause functions for all K failure causes.

There are different means to parametrically model the cumulative incidence function. The most used is the direct modeling of the cumulative incidence function, introduced by Jeong and Fine (2006), and the approach is the so-called indirect parameterization method, in which separate parametric models are assumed for the specific cause risk functions (Prentice et al., 1978).

Interval-censored data occur naturally in studies of diseases where the symptoms of interest are not directly observable, and laboratory or clinical examinations are required for detection. The exact time to the event of interest T is not directly observed but is known to fall within a time interval $[L, R]$, such that $0 \leq L < T \leq R \leq \infty$. In the scenario involving competing risks, the likelihood contribution for kth failure in the interval is given by $P(T \in [l, r], C = k) = F_k(r) - F_k(l)$ (Hudgens et al., 2014).

Let $T_1, ..., T_n$ be the failure time of n individuals. For the ith individual, there is a random interval censoring $[L_i, R_i]$, where L_i and R_i denote the left and right random endpoint, respectively, of the censoring interval. We define $\Delta_{ik} = I(L_i < T_i < R_i, C_i = k)$ for $k = 1, ..., K$, so that if $\Delta_{ik} = 1$ indicates the ith individual failed for cause k during interval $[L_i, R_i]$. For right-censored observations, the failure type is considered unknown, thus defining $\Delta_{i0} = 1$.

As it is not possible to observe (T, C) directly, but $Y = (L, R, \Delta)$, the observations (L_i, R_i) are independent of (T, C) and the distribution of (L_i, R_i) does not contain the parameters that govern the distribution of (T, C). Let $Y_1, ..., Y_n$ be a random sample of n independently and identically distributed copies of Y. Hudgens et al. (2014) demonstrated that the log-likelihood function for interval-censored competing risk data can be written in terms of the cumulative incidence functions by

$$\log L(\Theta) = \sum_{i=1}^{n} \log l(Y_i; \Theta) \tag{3}$$

where Θ is the vector consisting of total model parameters of $\Theta_1 \cup ... \cup \Theta_K$ and $l(Y_i; \Theta)$, which equals

$$l(Y_i; \Theta) = \prod_{k=1}^{K} \{F_k(R_i; \Theta_k) - F_k(L_i; \Theta_k)\}^{\Delta_{ik}} \left\{1 - \sum_{k=1}^{K} F_k(L_i; \Theta_k)\right\}^{\Delta_{i0}} \tag{4}$$

3 Extended model including correlated data

The likelihood function (3) was built considering a simple random sample of the population. In our context, in addition to the presence of competitive risks and interval censoring, we have the presence of clusters. It causes dependence among individuals within the same cluster, but the independence among clusters remains valid. Thus, in the context of dependent data, we can use the approach of marginal models. These models only consider covariate coefficients, considering the correlation within the cluster as a nuisance parameter. This analysis is based on the Generalized Estimating Equations (GEE) approach proposed by Liang and Zeger (1986).

To model dependent, right-censored, survival data including covariates, Huster et al. (1989) derived a GEE approach that allows statistical inference to be made for the marginal distributions while treating dependence among cluster members as perturbation parameters. More specifically, the method specifies the marginal distributions independently of the association structure. In addition, it leaves unspecified the nature of the dependence among the survival times of the events of cluster members. The parameters of the marginal model are estimated using maximum likelihood under the assumption of independence. This model was called the independent working matrix model by Huster et al. (1989). In the second step, the standard errors of the estimated parameters are estimated by using a correction through the sandwich variance.

Consider there are n clusters, and each of the i clusters has m_i individuals $(i = 1, ..., n)$ so there are $\sum_{i=1}^{n} m_i = m$ individuals in the entire sample. Let T_{ij} be the failure time for the jth individual in the ith clusters, $j = 1, ..., m_i$, so that it is only known that the failure time T_{ij} belongs to an interval, say $[L_{ij}, R_{ij}]$. Define $\Delta_{ijk} = I(L_{ij} < T_{ij} < R_{ij}, C = k)$ for $k = 1, ..., K$, so that if $\Delta_{ijk} = 1$ indicates the jth individual in the ith cluster failed by cause k along to the interval $[L_{ij}, R_{ij}]$. For the case of right-censored observations, the failure type is considered unknown, thus defining $\Delta_{ij0} = 1$.

Assuming that the failure times T_{ij} are independent and assuming the independence working model, the joint log-likelihood function becomes equal to

$$\log L(\Theta) = \sum_{i=1}^{n} \log l(Y_i; \Theta) \tag{5}$$

where

$$l(Y_i; \Theta) = \prod_{j=1}^{m_i} \prod_{k=1}^{K} \{F_k(R_{ij}; \Theta_k) - F_k(L_{ij}; \Theta_k)\}^{\Delta_{ijk}} \left\{1 - \sum_{k=1}^{K} F_k(L_{ij}; \Theta_k)\right\}^{\Delta_{ij0}}. \tag{6}$$

Under certain regularity conditions (Royall, 1986) for a consistent estimator $\hat{\Theta}$, $\sqrt{n}(\hat{\Theta} - \Theta)$ converges in distribution to $N(0, \Lambda(\Theta))$, where

$$\Lambda(\Theta) = \Upsilon(\Theta)^{-1} E[U(\Theta)U(\Theta)^T]\Upsilon(\Theta)^{-1} \tag{7}$$

where $U = \frac{\partial \log L(\Theta)}{\partial \Theta}$ is the score vector and $\Upsilon(\Theta) = E\left[\frac{-\partial^2 \log L(\Theta)}{\partial^2 \Theta}\right]$ is expected information matrix. According to Logan et al. (2010), a consistent estimator of $\Lambda(\Theta)$ is given by the sandwich variance

$$\hat{\Lambda}(\hat{\Theta}) = I(\hat{\Theta})^{-1} \left\{\sum_{i} U_i(\hat{\Theta}) U_i(\hat{\Theta})^T\right\} I(\hat{\Theta})^{-1} \tag{8}$$

where U_i is the contribution of ith cluster in the score vector and I is minus the matrix of second derivatives of $\log L(\Theta)$. Expressions for the first and second derivatives of (5) are presented in the Supporting Information.

Therefore, in parametric modeling, we specify parametric models for the marginal distributions but leave the nature of the dependency between cluster members completely unspecified. The parameters in the marginal models are then estimated using the likelihood function associated with the model that assumes the independence of the members (even if this assumption is incorrect). In the present case, it is equation (5). This likelihood is the product of the marginal likelihoods of each individual in the dataset. Then, the estimated standard errors are corrected using the sandwich variance (8).

In this work, we model the cumulative incidence function using the indirect approach, for which the exponential and Weibull models are considered for the marginal distributions. Without loss of generality, let us consider only two failure causes, $k = 1, 2$. The extension for K failure causes is immediate.

3.1 Exponential model

Considering the exponential model for the cause specific hazard function, we have:

$$S_k(t;\alpha_k) = \exp(-\alpha_k t) \qquad \text{and} \qquad \lambda_k(t;\alpha_k) = \alpha_k \qquad \text{for } k = 1, 2.$$

The parameterization of the cumulative incidence function in equation (1) reduces to

$$F_1(t;\Theta) = \int_0^t \alpha_1 \prod_{k=1}^{2} \exp(-\alpha_k t) du = \int_0^t \alpha_1 \exp\left(-\sum_{k=1}^{2} \alpha_k t\right) du = \frac{\alpha_1}{\alpha_1+\alpha_2}\left(1 - \exp\left(-\sum_{k=1}^{2} \alpha_k t\right)\right)$$

and

$$F_2(t;\Theta) = \int_0^t \alpha_2 \prod_{k=1}^{2} \exp(-\alpha_k t) du = \int_0^t \alpha_2 \exp\left(-\sum_{k=1}^{2} \alpha_k t\right) du = \frac{\alpha_2}{\alpha_1+\alpha_2}\left(1 - \exp\left(-\sum_{k=1}^{2} \alpha_k t\right)\right)$$

where $\Theta = (\alpha_1, \alpha_2)$.

In the estimation of the parameters of interest, we propose to use the log-likelihood (5) assuming both within and between cluster independence. The log-likelihood for the exponential model has the form

$$\begin{aligned}
\log L(\Theta) = &\sum_{i=1}^{n}\sum_{j=1}^{m_i} \{\Delta_{ij1} \log\left[F_1(R_{ij};\Theta) - F_1(L_{ij};\Theta)\right] + \Delta_{ji2} \log\left[F_2(R_{ij};\Theta) - F_2(L_{ij};\Theta)\right] + \\
&+ \Delta_{ij0} \log\left[1 - \sum_{k=1}^{2} F_k(L_{ij};\Theta)\right]\Bigg\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m_i} \Bigg\{\Delta_{ji1} \log\left[\frac{\alpha_1}{\alpha_1+\alpha_2}\left(\exp\left(-\sum_{k=1}^{2}\alpha_k L_{ij}\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_{ij}\right)\right)\right] + \\
&+ \Delta_{ij2} \log\left[\frac{\alpha_2}{\alpha_1+\alpha_2}\left(\exp\left(-\sum_{k=1}^{2}\alpha_k L_{ij}\right) - \exp\left(-\sum_{k=1}^{2}\alpha_k R_{ij}\right)\right)\right] + \\
&+ \Delta_{ij0} \log\left[1 - \sum_{k=1}^{2} F_k(L_{ij};\Theta)\right]\Bigg\}. \qquad (9)
\end{aligned}$$

To maximize $\log L(\Theta)$, one can use, for example, the Newton-Raphson algorithm. For this, we need the score function obtained by calculating the derivative of $\log L(\Theta)$ concerning the parameter vector Θ. The expressions for first and second derivatives of (9) are presented in the Supporting Information.

3.2 Weibull model

There is also the option of modeling the cumulative incidence function via Weibull distribution for the cause specific hazard function

$$S_k(t;\alpha_k,\tau_k) = \exp[-(\alpha_k t)^{\tau_k}] \qquad \text{and} \qquad \lambda_k(t;\alpha_k,\tau_k) = \tau_k \alpha_k^{\tau_k} t^{\tau_k - 1} \qquad \text{for } k = 1,2$$

where $\alpha_k > 0$ is the scale parameter and $\tau_k > 0$ is the shape parameter.

In this case, parameterized version of the cumulative incidence function (1), with $k = 1, 2$, has the form

$$F_1(t;\Theta) = \int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right] \tau_1 \alpha_1^{\tau_1} u^{\tau_1 - 1} du \tag{10}$$

and

$$F_2(t;\Theta) = \int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right] \tau_2 \alpha_2^{\tau_2} u^{\tau_2 - 1} du \tag{11}$$

where $\Theta = (\alpha_1, \tau_1, \alpha_2, \tau_2)$. These integrals do not have analytical solutions, thus requiring numerical methods.

Again, in the estimation of the parameters of interest, we propose to use the log-likelihood (5). Assuming both within and between cluster independence, the log-likelihood for Weibull model has the form

$$\begin{aligned}
\log L(\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \{ & \Delta_{ij1} \log\left[F_1(R_{ij};\Theta) - F_1(L_{ij};\Theta)\right] + \Delta_{ij2} \log\left[F_2(R_{ij};\Theta) - F_2(L_{ij};\Theta)\right] + \\
& + \Delta_{ij0} \log\left[1 - \sum_{k=1}^{2} F_k(L_{ij};\Theta)\right] \Bigg\} \\
= \sum_{i=1}^{n}\sum_{j=1}^{m_i} \Bigg\{ & \Delta_{ij1} \log\left[\int_0^{R_{ij}} \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right] \tau_1 \alpha_1^{\tau_1} u^{\tau_1 - 1} du - \right. \\
& \left. - \int_0^{L_{ij}} \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right] \tau_1 \alpha_1^{\tau_1} u^{\tau_1 - 1} du\right] + \\
& + \Delta_{ij2} \log\left[\int_0^{R_{ij}} \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right] \tau_2 \alpha_2^{\tau_2} u^{\tau_2 - 1} du - \right. \\
& \left. - \int_0^{L_{ij}} \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right] \tau_2 \alpha_2^{\tau_2} u^{\tau_2 - 1} du\right] + \\
& + \Delta_{ij0} \log\left[1 - \sum_{k=1}^{2} F_k(L_{ij};\Theta)\right] \Bigg\}.
\end{aligned} \tag{12}$$

Estimates of unknown parameters cannot be obtained in closed forms and it is necessary a numerical technique to compute these estimates. One may use Newton-Raphson or Gauss-Newton methods or their variants to maximize equation (12). Expressions for first and second derivatives of (12) are presented in the Supporting Information.

3.3 Regression model

To examine covariates effects on the response, it is necessary to include a regression structure. In survival analysis, the heterogeneity among the individuals is explained by covariates or explanatory variables. To build the regression model, we introduce covariate $\mathbf{x}$ assuming that the parameter α_k depends on the covariates through $\alpha_k = \exp(\boldsymbol{\beta}_k^T \mathbf{x})$ in the log-likelihood $\log L(\boldsymbol{\Theta})$. Here, $\mathbf{x} = (X_1, X_2, ..., X_p)$ and $\boldsymbol{\beta}_k^T = (\beta_{k1}, \beta_{k2}, ..., \beta_{kp})$ are the vector of regression coefficients. All parameters are simultaneously included in log-likelihood (9) for the exponential model, and (12) for the Weibull model. To obtain the MLE of unknown parameters, a system of normal equations is derived.

For the exponential model, the system of normal equations is

$$\frac{\partial \log L(\boldsymbol{\Theta})}{\partial \boldsymbol{\beta}_k} = 0$$

and for Weibull model, it

$$\frac{\partial \log L(\boldsymbol{\Theta})}{\partial \boldsymbol{\beta}_k} = 0 \qquad \text{and} \qquad \frac{\partial \log L(\boldsymbol{\Theta})}{\partial \tau_k} = 0.$$

The score functions are subsequently set to zero and solved using an iterative procedure such as the Newton-Raphson algorithm to find the maximum likelihood estimates.

4 Simulation study

A Monte Carlo study is conducted using the R language (R Core Team, 2022) to assess the finite sample performance of the proposed methodology. The maximization of the log-likelihood function (3) is conducted using the BFGS quasi-Newton nonlinear optimization algorithm implemented at the optim function available in R. Different scenarios are considered. For all of them, there are $K = 2$ causes of failure, $n \in \{100, 500, 1000\}$, $m_i = 4$, $i = 1, ..., n$. The covariates X_1 and X_2 are independently generated from Bernoulli (0.5) and standard normal distributions, respectively. The failure time and cause types are simulated according to the parameters $\boldsymbol{\Theta} = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}) = (0.3, 0.3, 0.3, 0.3)$, for the exponential model

$$\lambda_k(t_1; \mathbf{x}_i) = \exp(\beta_{k1} X_{1i} + \beta_{k2} X_{2i}) \tag{13}$$

and $\boldsymbol{\Theta} = (\beta_{11}, \beta_{12}, \tau_1, \beta_{21}, \beta_{22}, \tau_2) = (0.1, 0.1, 2.0, 0.1, 0.1, 2.0)$, for the Weibull model

$$\lambda_k(t_i; \mathbf{x}_i) = \tau_k t^{\tau_k - 1} [\exp(\beta_{k1} X_{1i} + \beta_{k2} X_{2i})]^{\tau_k} \tag{14}$$

where τ_k is the shape parameter, β_{k1} e β_{k2} are the covariate effects, $i = 1, ..., n$ and $k = 1, 2$.

Generating the data involves the following steps.

1 A common frailty term for each cluster, ω_{ik}, $i = 1, ..., n$ e $k = 1, 2$, is generated from a gamma distribution, $\omega_{ik} \sim Gamma(\theta_k, \theta_k)$, following the steps of Brazauskas and Le-Rademacher (2016). θ_k is chosen to obtain the within cluster dependence structure in each scenario.

2 For each individual, covariates X_1 and X_2 are independently generated from a Bernoulli (0.5) and standard normal distribution, respectively.

3 We generate competing risks data via the following algorithm, according to Beyersmann et al. (2011):

3.1 Specify the cause-specific hazards $\lambda_1(t; \mathbf{x})$ and $\lambda_2(t; \mathbf{x})$ for exponential and Weibull models according to the equations (9) and (12), respectively.

3.2 Simulate failure times T with all-cause hazard $\lambda(t; \mathbf{x}) = \lambda_1(t; \mathbf{x}) + \lambda_2(t; \mathbf{x})$.

3.3 Determine the type of event for each individual using a binomial experiment for a simulated failure time T, which decides with probability $\frac{\lambda_1(t;\mathbf{X})}{\lambda_1(t;\mathbf{X})+\lambda_2(t;\mathbf{X})}$ on cause k, $k = 1, 2$.

3.4 Additionally, generate independent right-censoring times $C \sim U[0, 1]$.

3.5 Observed times are taken as the minimum between the failure times T and the censored times C, $t_{ij} = min(T_{ij}, C_{ij})$.

4 The interval-censoring times (L_{ij} and R_{ij}) are constructed as follows:

4.1 For all right-censored individuals, we apply the following interval-censoring times:

$$L_{ij} = t_{ij} \text{ and } R_{ij} = \infty;$$

4.2 For individuals who failed for one of the K causes, we apply the following interval-censoring times:

$$L_{ij} = 0.0001 \text{ and } R_{ij} = c,$$

where c is generated from a uniform distribution $U[0.1; b]$ and b is chosen to increase or decrease the size of the intervals. Lifetimes t_{ij} are compared with created intervals, to verify if t_{ij} belongs to the interval. If t_{ij} does not belong to the interval, new intervals are created so that

$$L_{ij} = R_{ij} \text{ and } R_{ij} = R_{ij} + c$$

where c is a new generated value from a uniform distribution $U[0.1; b]$.

Each dataset is analyzed using the methodology proposed in Section 3. Each simulation result displayed in this section is based on $N = 1000$ simulated datasets. The performance of all point estimators is compared numerically in terms of their average estimate, empirical variance, asymptotic variance, and mean square error (MSE) values. Moreover, coverage probability (CP) of 95% is computed for interval estimates.

The average of parameter estimates is computed as $\hat{\theta} = \frac{\sum_{i=1}^{N} \hat{\theta}_i}{N}$, providing an estimate of the expected value of the coefficient estimates for N simulations. The relative bias (rb) is estimated by taking the difference between $\hat{\theta}$ and θ, the true population parameter value, divided by and θ, such that $rb = \frac{\hat{\theta}-\theta}{\theta}$.

Table 1 Scenario I for exponential and Weibull model - In this scenario $\omega_{ik} \sim Gamma(2,2)$, thus causing a weak correlation among the individuals within the same cluster (Kendall's tau = 0.20). In addition, the size of the intervals was generated from $c \sim U(1, 0.1, 0.3)$.

	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
Exponetial model								
100 clusters								
$\beta_{11} = 0.3$	0.276	0.008	-0.081	0.084	0.075	0.079	0.915	0.925
$\beta_{12} = 0.3$	0.271	0.008	-0.097	0.084	0.075	0.080	0.895	0.913
$\beta_{21} = 0.3$	0.271	0.008	-0.097	0.082	0.075	0.080	0.917	0.931
$\beta_{22} = 0.3$	0.270	0.007	-0.099	0.081	0.075	0.080	0.908	0.929
500 clusters								
$\beta_{11} = 0.3$	0.281	0.005	-0.064	0.058	0.053	0.056	0.921	0.926
$\beta_{12} = 0.3$	0.281	0.005	-0.064	0.056	0.053	0.056	0.915	0.937
$\beta_{21} = 0.3$	0.281	0.005	-0.064	0.057	0.053	0.056	0.913	0.939
$\beta_{22} = 0.3$	0.281	0.005	-0.064	0.057	0.053	0.056	0.911	0.934
1000 clusters								
$\beta_{11} = 0.3$	0.289	0.003	-0.036	0.037	0.033	0.036	0.921	0.946
$\beta_{12} = 0.3$	0.289	0.003	-0.036	0.036	0.033	0.036	0.925	0.940
$\beta_{21} = 0.3$	0.289	0.003	-0.036	0.037	0.033	0.036	0.923	0.949
$\beta_{22} = 0.3$	0.289	0.003	-0.036	0.037	0.033	0.036	0.921	0.946
Weibull model								
100 clusters								
$\beta_{11} = 0.1$	0.099	0.007	-0.008	0.092	0.097	0.095	0.976	0.997
$\beta_{12} = 0.1$	0.105	0.008	0.049	0.091	0.097	0.091	0.976	0.998
$\beta_{21} = 0.1$	0.103	0.007	0.034	0.094	0.098	0.096	0.976	0.994
$\beta_{22} = 0.1$	0.105	0.008	0.048	0.093	0.098	0.097	0.981	0.996
$\tau_1 = 2.0$	0.930	1.179	-0.535	0.181	0.165	0.168	0.032	0.032
$\tau_2 = 2.0$	0.924	1.189	-0.538	0.175	0.179	0.179	0.026	0.026
500 clusters								
$\beta_{11} = 0.1$	0.102	0.006	0.030	0.085	0.096	0.089	0.980	0.999
$\beta_{12} = 0.1$	0.102	0.006	0.023	0.088	0.097	0.095	0.982	0.999
$\beta_{21} = 0.1$	0.103	0.006	0.032	0.089	0.098	0.091	0.978	0.999
$\beta_{22} = 0.1$	0.100	0.006	-0.001	0.088	0.097	0.090	0.981	0.999
$\tau_1 = 2.0$	0.950	1.154	-0.525	0.167	0.126	0.157	0.455	0.468
$\tau_2 = 2.0$	0.941	1.162	-0.529	0.171	0.174	0.172	0.770	0.770
1000 clusters								
$\beta_{11} = 0.1$	0.101	0.005	0.030	0.082	0.096	0.084	0.980	0.999
$\beta_{12} = 0.1$	0.101	0.005	0.018	0.084	0.094	0.086	0.982	0.999
$\beta_{21} = 0.1$	0.102	0.005	0.028	0.087	0.095	0.086	0.978	0.999
$\beta_{22} = 0.1$	0.100	0.004	-0.001	0.083	0.092	0.085	0.981	0.999
$\tau_1 = 2.0$	0.980	1.141	-0.515	0.147	0.106	0.137	0.555	0.568
$\tau_2 = 2.0$	0.972	1.131	-0.519	0.141	0.148	0.142	0.770	0.770

The empirical variance is computed by $Var(\hat{\theta}) = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \hat{\theta})^2}{N-1}$ and mean square error is calculated by $MSE(\theta) = \frac{\sum_{i=1}^{N}(\hat{\theta} - \theta)^2}{N}$.

Standard error (SE_1) is calculated by minus the inverse the matrix of second derivatives of $\log L(\Theta)$, and standard error sandwich estimator (SE_2) is calculated using sandwich variance (8) based on the average of the N simulations results.

The simulation study results are presented in Tables 1 to 3 and Figures 1 to 6 in the Appendix

Table 2 Scenario II for exponential and Weibull model - Same structure as in Scenario I except that we increased the size of the intervals, using $c \sim U(1, 0.1, 0.5)$.

	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
Exponetial model								
100 clusters								
$\beta_{11} = 0.3$	0.269	0.009	-0.104	0.087	0.075	0.080	0.892	0.906
$\beta_{12} = 0.3$	0.268	0.009	-0.106	0.084	0.075	0.080	0.912	0.927
$\beta_{21} = 0.3$	0.268	0.009	-0.108	0.085	0.075	0.080	0.887	0.906
$\beta_{22} = 0.3$	0.268	0.009	-0.107	0.081	0.075	0.080	0.906	0.934
500 clusters								
$\beta_{11} = 0.3$	0.278	0.006	-0.073	0.066	0.063	0.066	0.906	0.922
$\beta_{12} = 0.3$	0.278	0.006	-0.073	0.065	0.063	0.064	0.902	0.919
$\beta_{21} = 0.3$	0.278	0.006	-0.073	0.065	0.063	0.064	0.918	0.929
$\beta_{22} = 0.3$	0.278	0.006	-0.073	0.064	0.063	0.063	0.907	0.908
1000 clusters								
$\beta_{11} = 0.3$	0.285	0.004	-0.049	0.046	0.043	0.046	0.926	0.942
$\beta_{12} = 0.3$	0.285	0.004	-0.049	0.045	0.043	0.044	0.922	0.949
$\beta_{21} = 0.3$	0.285	0.004	-0.049	0.045	0.043	0.044	0.928	0.949
$\beta_{22} = 0.3$	0.285	0.004	-0.049	0.044	0.043	0.044	0.927	0.948
Weibull model								
100 clusters								
$\beta_{11} = 0.1$	0.108	0.008	0.077	0.098	0.097	0.098	0.973	0.996
$\beta_{12} = 0.1$	0.107	0.009	0.068	0.097	0.096	0.097	0.965	0.992
$\beta_{21} = 0.1$	0.108	0.009	0.076	0.096	0.096	0.096	0.973	0.996
$\beta_{22} = 0.1$	0.109	0.008	0.087	0.094	0.096	0.094	0.974	0.995
$\tau_1 = 2.0$	0.940	1.165	-0.530	0.205	0.156	0.159	0.408	0.440
$\tau_2 = 2.0$	0.942	1.158	-0.529	0.197	1.115	1.015	0.433	0.433
500 clusters								
$\beta_{11} = 0.1$	0.101	0.008	0.007	0.088	0.096	0.089	0.966	0.996
$\beta_{12} = 0.1$	0.105	0.007	0.089	0.092	0.095	0.094	0.972	0.996
$\beta_{21} = 0.1$	0.101	0.008	0.014	0.093	0.095	0.094	0.977	0.998
$\beta_{22} = 0.1$	0.105	0.007	0.085	0.094	0.096	0.094	0.976	0.996
$\tau_1 = 2.0$	0.962	1.144	-0.519	0.188	0.125	0.153	0.404	0.470
$\tau_2 = 2.0$	0.959	1.131	-0.510	0.190	0.279	0.209	0.786	0.756
1000 clusters								
$\beta_{11} = 0.1$	0.101	0.006	0.007	0.085	0.093	0.087	0.966	0.996
$\beta_{12} = 0.1$	0.102	0.006	0.089	0.088	0.092	0.090	0.972	0.996
$\beta_{21} = 0.1$	0.101	0.006	0.014	0.089	0.091	0.089	0.977	0.998
$\beta_{22} = 0.1$	0.101	0.006	0.085	0.090	0.092	0.090	0.976	0.996
$\tau_1 = 2.0$	0.982	1.135	-0.509	0.178	0.112	0.145	0.404	0.470
$\tau_2 = 2.0$	0.979	1.127	-0.501	0.181	0.179	0.177	0.786	0.486

As shown in Table 1, the mean estimates are close to the true value of the parameters; and as the number of clusters increases, the MSE and rb decreases. However, Weibull shape parameters estimates are biased and show very slow convergence rate. Furthermore, the sandwich standard error (SE_2) is close to the empirical standard error (SD), indicating that the proposed approach effectively makes a correction for the cluster correlation. Coverage probabilities (CP_2) are close to the nominal level of 0.95, indicating a nice approximation of the estimates for the normal distribution.

We attempted to evaluate the effect of intervals size. By generating different scenarios with larger intervals sizes. The results are presented in Table 2, for

Table 3 Scenario III for exponential and Weibull model - Same structure as in Scenario I except that we increased the intra-cluster correlation (Kendall's tau = 0.85) by using $\omega_{ik} \sim Gamma(1/11, 1/11)$.

	$\hat{\beta}$	MSE	rb	SD	SE_1	SE_2	CP_1	CP_2
Exponetial model								
100 clusters								
$\beta_{11} = 0.3$	0.269	0.010	-0.103	0.086	0.076	0.087	0.919	0.926
$\beta_{12} = 0.3$	0.269	0.010	-0.103	0.086	0.075	0.085	0.912	0.914
$\beta_{21} = 0.3$	0.269	0.010	-0.103	0.086	0.076	0.085	0.911	0.935
$\beta_{22} = 0.3$	0.269	0.010	-0.103	0.086	0.075	0.087	0.912	0.915
500 clusters								
$\beta_{11} = 0.3$	0.276	0.008	-0.081	0.058	0.050	0.056	0.901	0.925
$\beta_{12} = 0.3$	0.276	0.008	-0.081	0.058	0.051	0.056	0.901	0.929
$\beta_{21} = 0.3$	0.276	0.008	-0.081	0.058	0.051	0.056	0.911	0.937
$\beta_{22} = 0.3$	0.276	0.008	-0.081	0.058	0.050	0.056	0.911	0.931
1000 clusters								
$\beta_{11} = 0.3$	0.287	0.006	-0.043	0.038	0.031	0.037	0.931	0.944
$\beta_{12} = 0.3$	0.287	0.006	-0.043	0.038	0.032	0.037	0.931	0.948
$\beta_{21} = 0.3$	0.287	0.005	-0.043	0.038	0.032	0.037	0.931	0.949
$\beta_{22} = 0.3$	0.287	0.006	-0.043	0.038	0.031	0.037	0.931	0.947
Weibull model								
100 clusters								
$\beta_{11} = 0.1$	0.108	0.008	0.077	0.098	0.097	0.096	0.973	0.996
$\beta_{12} = 0.1$	0.107	0.009	0.068	0.095	0.096	0.096	0.965	0.992
$\beta_{21} = 0.1$	0.108	0.009	0.076	0.095	0.096	0.097	0.973	0.996
$\beta_{22} = 0.1$	0.109	0.008	0.087	0.092	0.096	0.094	0.974	0.995
$\tau_1 = 2.0$	0.940	1.165	-0.530	0.208	0.196	0.199	0.008	0.040
$\tau_2 = 2.0$	0.942	1.158	-0.529	0.199	1.115	1.115	0.033	0.033
500 clusters								
$\beta_{11} = 0.1$	0.101	0.007	0.007	0.096	0.096	0.090	0.966	0.996
$\beta_{12} = 0.1$	0.105	0.008	0.059	0.092	0.095	0.094	0.972	0.986
$\beta_{21} = 0.1$	0.101	0.008	0.014	0.091	0.095	0.093	0.977	0.988
$\beta_{22} = 0.1$	0.105	0.007	0.085	0.092	0.095	0.093	0.976	0.986
$\tau_1 = 2.0$	0.962	1.154	-0.521	0.198	0.155	0.171	0.004	0.770
$\tau_2 = 2.0$	0.959	1.151	-0.520	0.195	0.679	0.679	0.786	0.786
1000 clusters								
$\beta_{11} = 0.1$	0.101	0.007	0.007	0.088	0.086	0.088	0.966	0.996
$\beta_{12} = 0.1$	0.103	0.007	0.082	0.090	0.093	0.091	0.972	0.996
$\beta_{21} = 0.1$	0.101	0.006	0.013	0.090	0.093	0.090	0.977	0.998
$\beta_{22} = 0.1$	0.102	0.006	0.082	0.091	0.092	0.091	0.976	0.996
$\tau_1 = 2.0$	0.992	1.124	-0.520	0.188	0.150	0.165	0.004	0.770
$\tau_2 = 2.0$	0.989	1.123	-0.520	0.191	0.670	0.670	0.786	0.786

the exponential and Weibull models. The mean estimates are close to the true value but the MSE, rb, and SD are larger when compared with the first scenario. Furthermore, the sandwich standard error (SE_2) is very close to the empirical standard error (SD) and the coverage probabilities are close to the nominal level of 0.95. Note that when we increase the correlation within the cluster, the parameter estimates have a slightly greater bias, as can be seen in Table 3. The empirical standard error is slightly greater but the sandwich standard error is still close to the empirical standard error. In all scenarios,

Table 4 Parameter estimates from exponential and Weibull models of the CIF for three competing causes in the dental trauma study

	Exponetial model								
	Vitality ($k = 1$)			**Necrosis** ($k = 2$)			**PCO** ($k = 3$)		
Covarientes	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Age	0.0018	0.0009	0.0541	0.0023	0.0008	0.0044	-0.0076	0.0036	0.0364
Gender (Male)	0.1645	0.2356	0.4850	0.2914	0.2377	0.2200	0.3011	0.3613	0.4050
Luxation (Ext+Lat)	-0.1460	0.2384	0.5402	0.2664	0.2807	0.3426	0.9291	0.4384	0.0341
Luxation (Int)	-0.7403	0.4175	0.0762	1.0220	0.3112	0.0010	-0.0082	0.7028	0.9907
Assoc. injury (Yes)	-0.9781	0.3469	0.0048	1.4430	0.2064	2.71e-12	-0.4273	0.5140	0.4060
Apical Diam. (open)	-0.0345	0.2181	0.8740	-0.2684	0.2437	0.2710	0.6228	0.3444	0.0706
Emerg. Treat. (yes)	-0.3077	0.2626	0.2410	-0.0335	0.2716	0.9020	-0.2561	0.4202	0.5420
	Weibull model								
Age	0.0220	0.0114	0.0539	0.0281	0.0098	0.0044	-0.0915	0.0437	0.0364
Gender (Male)	0.1701	0.2355	0.4750	0.2857	0.2383	0.2310	0.3045	0.3609	0.3990
Luxation (Ext+Lat)	-0.3495	0.2221	0.1155	0.1747	0.2550	0.4930	0.9848	0.3994	0.0137
Luxation (Int)	-0.7467	0.4187	0.0745	1.0205	0.3102	0.0010	-0.0078	0.7028	0.9911
Assoc. injury (Yes)	-0.9822	0.3469	0.0046	1.4520	0.2073	2.49e-12	-0.4302	0.5140	0.4030
Apical Diam. (open)	-0.0027	0.2837	0.9920	-0.4569	0.3597	0.2040	0.8663	0.3661	0.0180
Emerg. Treat. (yes)	-0.3454	0.2446	0.1580	-0.0355	0.2616	0.8020	-0.3366	0.4621	0.4660
τ	1.2250	0.1381	-	0.5050	0.2561	-	1.8370	0.2967	-

increasing the number of clusters lead to less biased estimates and coverage intervals closer to the nominal level.

5 Application to the dental study

The proposed methodology is illustrated with a real dental trauma dataset from a study carried out at the Dental Trauma Program of the School of Dentistry of the Federal University of Minas Gerais, which aimed to evaluate pulp prognosis of 427 luxated permanent teeth from 224 patients.

These data presented some characteristics that must be considered in the statistical analysis. The first refers to the existence of competitive risks as there is more than one possible outcome for the pulpal response, namely: maintenance of vitality ($k = 1$), necrosis ($k = 2$), and pulp canal obliteration - PCO ($k = 3$). The second characteristic is the presence of clusters because in some cases, the same patient has more than one luxated tooth. In addition, it was not possible to specify an exact date of occurrence of the event; hence, a time interval between the patient's follow-up appointments was selected.

A total of 21.25% of the data were right-censored and 78.75% were interval-censored. Regarding the pulp diagnosis variable, 32.97% presented vitality ($k = 1$), 32.60% necrosis ($k = 2$), and 13.19% PCO ($k = 3$). Six covariates were included in the study: gender (44% female and 56% male), age (mean 15.5 years), type of luxation (33.33% subluxation+concussion, 37% lateral, 15.75% extrusive, and 13.92% intrusive), concomitant injuries other than luxations (17.22% yes and 82.78% no), diameter of the apical foramen (32.23% open and 67.77% closed), and emergencial treatment (60.9% yes and 39.81% no).

We fitted the exponential and Weibull regression models based on methodology presented in Section 3.

The results concerning the coefficient estimates along with the sandwich standard error estimates (SE) are reported in Table 4. The p-value associated with each covariate is calculated using the Wald test. Both the exponential and Weibull models indicate the same covariates as significant. The covariates for

the type of luxation and concomitant injury are significant for the occurrence of the three events, and the diameter of the apical foramen is significant for the occurrence of necrosis and PCO. Experimental evidence suggests that teeth with wider apical foramen are more likely to experience vascular ingrowth and nerve regeneration, which allows for the preservation of pulp vitality. Clinical data confirmed this premise and also demonstrated that the risk of pulp necrosis increased with the severity of luxation and the presence of concomitant crown fractures. Furthermore, clinical studies have shown that PCO is more frequently observed in injuries that involve the displacement of immature permanent teeth (Clark and Levin, 2019; Darley et al., 2020; Spinas et al., 2021). Table 4 shows that the estimates of the τ parameter are close to 1 for the event of vitality only. It is observed that the hypothesis that equals 1 is not significant for the events necrosis and PCO. Therefore, there is a favorable indication for the Weibull model.

6 Final remarks

We proposed a GEE-type procedure for cluster competitive risks models under interval-censored dataset structure. Our approach is based on a marginal parametric likelihood for independent data and uses the sandwich variance for the estimates. This approach is very appealing for its computational simplicity and parameters that have a populational interpretation.

The simulation study showed that our proposed methods perform well for finite data sets. Three scenarios were considered based on exponential and Weibull parametric structures. Whereas scale parameters presented nice finite properties. Weibull shape parameters estimates are biased and showed very slow convergence rate.

We illustrated the application of our methodology with a real study of dental trauma demonstrating its practical importance. It can also be used in many other practical situations.

This topic has numerous avenues for related future research. The most interesting one seems to be considering a semi-parametric model, that is, a Cox-type model approach. Another promising research direction is using splines instead of fixed coefficients for continuous covariates. We are presently working on the former point and wish to report the results in a future paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Andreasen, F. M. and Andreasen, J. O. (1990). Treatment of traumatic dental injuries: Shift in strategy. *International Journal of Technology Assessment in Health Care,* **6,** 588–602.

2. Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing risks and multistate models with R*. New York: Springer.
3. Bogaerts, K., Leroy, R., Lesaffre, E., and Declerck, D. (2002). Modelling tooth emergence data based on multivariate interval-censored data. *Statistics in Medicine*, **21(24)**, 3775–3787.
4. Brazauskas, R. and Le-Rademacher, J. (2016). Methods for generating paired competing risks data. *Computer Methods and Programs in Biomedicine*, **135**, 199–207.
5. Clark, D. and Levin, L. (2019). Prognosis and complications of mature teeth after lateral luxation. *The Journal of the American Dental Association*, **150(8)**, 649–655.
6. Darley, R. M., e Silva, C. F., dos Santos Costa, F., Xavier, C. B., and Demarco, F. F. (2020). Complications and sequelae of concussion and subluxation in permanent teeth: A systematic review and meta-analysis. *Dental Traumatology*, **36(6)**, 557–567.
7. Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94**, 496–509.
8. Hanagal, D. D. (2011). *Modeling survival data using frailty models*. New York: Springer.
9. Hudgens, M. G., Li, C., and Fine, J. P. (2014). Parametric likelihood inference for interval censored competing risks data. *Biometrics,* **70**, 1–9.
10. Huster, W. J., Brookmeyer, R., and Self, S. G. (1989). Modelling paired survival data with covariates. *Biometrics*, **45**, 145–156 .
11. Jeong, J. H. and Fine, J. (2006). Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* **55**, 187–200.
12. Jeong, J. H. and Fine, J. P. (2007). Parametric regression on cumulative incidence function. *Biostatistics,* **8**, 184–196.
13. Lawless, J. F. (2011). *Statistical models and methods for lifetime data*. 2nd edition. Hoboken, New Jersey: Wiley.
14. Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, **19**, 219–238.
15. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
16. Logan, B. R., Zhang, M.-J., and Klein, J. P. (2010). Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics*, **67**, 1–7.
17. Petti, S., Glendor, U., and Andersson, L. (2018). World traumatic dental injury prevalence and incidence, a meta-analysis–one billion living people have had traumatic dental injuries. *Dental Traumatology*, **34,** 71–86.
18. Prentice, R. L., Kalbeisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, **34**, 541–554.
19. R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
20. Royall, R. M. (1986). The prediction approach to robust variance estimation in two-stage cluster sampling. *Journal of the American Statistical Association*, **81**, 119–123.
21. Spinas, E., Deias, M., Mameli, A., and Giannetti, L. (2021). Pulp canal obliteration after extrusive and lateral luxation in young permanent teeth: A scoping review. *European Journal of Paediatric Dentistry*, **22**, 55–60.
22. Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. New York: Springer.
23. Wienke, A. (2010). *Frailty models in survival analysis*. CRC Press.
24. Zhou, B., Fine, J., Latouche, A., and Labopin, M. (2012). Competing risks regression for clustered data. *Biostatistics,* **13**, 371–383.

Appendix A: Derivatives

The parameter estimates involve calculating the first and second order derivatives of the equation (5) for the assumed parametric model. Hence, in general,

considering
$G_k(\Theta_k) = \{F_k(R_{ij};\Theta_k) - F_k(L_{ij};\Theta_k)\}$, in (5) we have that

$$\log l(y_i;\Theta) = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \Delta_{ijk} \log G_k(\Theta_k) + \Delta_{ij0} \log\left(1 - \sum_{k=1}^{K} F_k(L_{ij};\Theta_k)\right). \quad (15)$$

Therefore,

$$\frac{\partial \log l(y_i;\Theta)}{\partial \Theta_k} = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \frac{\Delta_{ijk}}{G_k(\Theta_k)} \cdot \frac{\partial G_k(\Theta_k)}{\partial \Theta_k} - \frac{\Delta_{ij0}}{1 - \sum_{k=1}^{l} F_k(L_{ij};\Theta_k)} \cdot \frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_k} \quad (16)$$

where

$$\frac{\partial G_k(\Theta_k)}{\partial \Theta_k} = \frac{\partial F_k(R_{ij};\Theta_k)}{\partial \Theta_k} - \frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_k}.$$

From (16) we have that the second order derivatives are:

$$\frac{\partial^2 \log l(y_i;\Theta)}{\partial \Theta_k^2} = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \Delta_{ijk} \left[\frac{\frac{\partial^2 G_k(\Theta_k)}{\partial \Theta_k^2}}{G_k(\Theta_k)} - \left(\frac{\frac{\partial^2 G_k(\Theta_k)}{\partial \Theta_k^2}}{G_k(\Theta_k)}\right)^2\right] - \Delta_{ij0} \left[\frac{\frac{\partial^2 F_k(L_{ij};\Theta_k)}{\partial \Theta_k^2}}{1 - \sum_{k=1}^{K} F_k(L_{ij};\Theta_k)} - \left(\frac{\frac{\partial^2 F_k(L_{ij};\Theta_k)}{\partial \Theta_k^2}}{1 - \sum_{k=1}^{K} F_k(L_{ij};\Theta_k)}\right)^2\right]$$

and

$$\frac{\partial^2 \log l(y_i;\Theta)}{\partial \Theta_k \Theta_{k'}} = \sum_{j=1}^{m_i}\sum_{k=1}^{K} \Delta_{ijk} \left[\frac{\frac{\partial^2 G_k(\Theta_k)}{\partial \Theta_k \partial \Theta_{k'}}}{G_k(\Theta_k)} - \frac{\frac{\partial G_k(\Theta_k)}{\partial \Theta_k}\frac{\partial G_k(\Theta_k)}{\partial \Theta_{k'}}}{\{G_k(\Theta_k)\}^2}\right] - \Delta_{ij0} \left[\frac{\frac{\partial^2 F_k(L_{ij};\Theta_k)}{\partial \Theta_k \partial \Theta_{k'}}}{1 - \sum_{k=1}^{K} F_k(L_{ij};\Theta_k)} - \frac{\frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_k}\frac{\partial F_k(L_{ij};\Theta_k)}{\partial \Theta_{k'}}}{\left\{1 - \sum_{k=1}^{K} F_k(L_{ij};\Theta_k)\right\}^2}\right].$$

Then

$$\frac{\partial \log L(\Theta)}{\partial \Theta_k} = \sum_{i=1}^{n} \frac{\partial \log l(y_i;\Theta)}{\partial \Theta_k}.$$

Derivatives for Exponential Model

Considering the presence of two competitive risks, that is, $k = 1$ e 2, we have for the exponential model, without covariates that

$$F_1(t;\Theta) = \int_0^t \alpha_1 \prod_{k'=1}^{2} \exp(-\alpha_k t)du = \int_0^t \alpha_1 \exp\left(-\sum_{k=1}^{2}\alpha_k t\right) du = \frac{\alpha_1}{\alpha_1+\alpha_2}\left(1-\exp\left(-\sum_{k=1}^{l}\alpha_k t\right)\right)$$

and

$$F_2(t;\Theta) = \int_0^t \alpha_2 \prod_{k'=1}^{2} \exp(-\alpha_k t)du = \int_0^t \alpha_2 \exp\left(-\sum_{k=1}^{2}\alpha_k t\right) du = \frac{\alpha_2}{\alpha_1+\alpha_2}\left(1-\exp\left(-\sum_{k=1}^{l}\alpha_k t\right)\right)$$

where $\Theta = (\alpha_1, \alpha_2)$.

The first order derivatives of $F_1(t;\Theta)$ are

$$\frac{\partial F_1(t;\Theta)}{\partial \alpha_1} = \frac{\alpha_2}{(\alpha_1+\alpha_2)^2}\left[1-\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right] + \frac{\alpha_1 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)$$

and

$$\frac{\partial F_1(t;\Theta)}{\partial \alpha_2} = -\frac{\alpha_1}{(\alpha_1+\alpha_2)^2}\left[1+\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right] + \frac{\alpha_1 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right).$$

The second order derivatives of $F_1(t;\Theta)$ are

$$\frac{\partial^2 F_1(t;\Theta)}{\partial \alpha_1^2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[2(\alpha_2^2+\alpha_1\alpha_2)t - (\alpha_1\alpha_2^2+2\alpha_1^2\alpha_2+\alpha_1^3)t^2+2\alpha_2\right] - \frac{2\alpha_2}{(\alpha_1+\alpha_2)^3},$$

$$\frac{\partial^2 F_1(t;\Theta)}{\partial \alpha_2^2} = \frac{\alpha_1\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(-\alpha_1^2-\alpha_2^2-2\alpha_1\alpha_2)t^2 - 2(\alpha_1+\alpha_2)t - 2\right] + \frac{2\alpha_1}{(\alpha_1+\alpha_2)^3}$$

and

$$\frac{\partial^2 F_1(t;\Theta)}{\partial \alpha_1\partial \alpha_2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(\alpha_2^2-\alpha_1^2)t - (\alpha_1\alpha_2^2+2\alpha_1^2\alpha_2+\alpha_1^3)t^2 - \alpha_1+\alpha_2\right] + \frac{\alpha_1-\alpha_2}{(\alpha_1+\alpha_2)^3}.$$

The first order derivatives of $F_2(t;\Theta)$ are

$$\frac{\partial F_2(t;\Theta)}{\partial \alpha_2} = \frac{\alpha_1}{(\alpha_1+\alpha_2)^2}\left[1-\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right] + \frac{\alpha_2 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)$$

and

$$\frac{\partial F_2(t;\Theta)}{\partial \alpha_1} = -\frac{\alpha_2}{(\alpha_1+\alpha_2)^2}\left[1+\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)\right]+\frac{\alpha_2 t}{\alpha_1+\alpha_2}\exp\left(-\sum_{k=1}^{2}\alpha_k t\right).$$

The second order derivatives of $F_2(t;\Theta)$ are

$$\frac{\partial^2 F_2(t;\Theta)}{\partial \alpha_2^2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[2(\alpha_1^2+\alpha_1\alpha_2)t-(\alpha_2\alpha_1^2+2\alpha_2^2\alpha_1+\alpha_2^3)t^2-2\alpha_1\right]-\frac{2\alpha_1}{(\alpha_1+\alpha_2)^3},$$

$$\frac{\partial^2 F_2(t;\Theta)}{\partial \alpha_1^2} = \frac{\alpha_2\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(\alpha_2^2-\alpha_1^2-2\alpha_1\alpha_2)t^2-2(\alpha_1+\alpha_2)t-2\right]+\frac{2\alpha_2}{(\alpha_1+\alpha_2)^3}$$

and

$$\frac{\partial^2 F_2(t;\Theta)}{\partial \alpha_1\partial \alpha_2} = \frac{\exp\left(-\sum_{k=1}^{2}\alpha_k t\right)}{(\alpha_1+\alpha_2)^3}\left[(\alpha_1^2-\alpha_2^2)t+(-\alpha_2\alpha_1^2-2\alpha_2^2\alpha_1+\alpha_2^3)t^2-\alpha_2+\alpha_1\right]+\frac{\alpha_2-\alpha_1}{(\alpha_1+\alpha_2)^3}.$$

From (5) we have that the log-likelihood function for the exponential model is

$$\begin{aligned}\log L(\Theta) = &\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\Delta_{ij1}\log\left[F_1(R_{ij};\Theta)-F_1(L_{ij};\Theta)\right]+\Delta_{ji2}\log\left[F_2(R_{ij};\Theta)-F_2(L_{ij};\Theta)\right]+\\ &+\ \Delta_{ij0}\log\left[1-\sum_{k=1}^{2}F_k(L_{ij};\Theta)\right]\Bigg\}\\ &=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\Delta_{ji1}\log\left[\frac{\alpha_1}{\alpha_1+\alpha_2}\left(\exp\left(-\sum_{k=1}^{2}\alpha_k L_{ij}\right)-\exp\left(-\sum_{k=1}^{2}\alpha_k R_{ij}\right)\right)\right]+\right.\\ &+\ \Delta_{ij2}\log\left[\frac{\alpha_2}{\alpha_1+\alpha_2}\left(\exp\left(-\sum_{k=1}^{2}\alpha_k L_{ij}\right)-\exp\left(-\sum_{k=1}^{2}\alpha_k R_{ij}\right)\right)\right]+\\ &\left.+\ \Delta_{ij0}\log\left[1-\sum_{k=1}^{2}F_k(L_{ij};\Theta)\right]\right\}. \qquad (17)\end{aligned}$$

The first-order derivatives of the log-likelihood function are

$$\frac{\partial \log L(\Theta)}{\partial \alpha_1} = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\Delta_{j1}\left[\frac{R_j\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)-L_j\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)-\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)}+\frac{\alpha_2}{\alpha_1(\alpha_1+\alpha_2)}\right]+\right.$$

$$\Delta_{j2}\left[\frac{R_j\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)-L_j\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)-\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)}+\frac{1}{(\alpha_1+\alpha_2)}\right]-$$

$$\Delta_{j0}\left[\frac{1}{1-\sum_{k=1}^{2}F_k(L_j;\Theta)}\left(L_j\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)-\frac{2\alpha_2}{(\alpha_1+\alpha_2)^2}\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)\right)\right]\Bigg\}$$

and

$$\frac{\partial\log L(\Theta)}{\partial\alpha_2}=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\Bigg\{\Delta_{j1}\left[\frac{R_j\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)-L_j\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)-\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)}+\frac{1}{(\alpha_1+\alpha_2)}\right]+$$

$$\Delta_{j2}\left[\frac{R_j\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)-L_j\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)}{\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)-\exp\left(-\sum_{k=1}^{2}\alpha_k R_j\right)}+\frac{\alpha_1}{\alpha_2(\alpha_1+\alpha_2)}\right]-$$

$$\Delta_{j0}\left[\frac{1}{1-\sum_{k=1}^{2}F_k(L_j;\Theta)}\left(L_j\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)-\frac{2\alpha_1}{(\alpha_1+\alpha_2)^2}\exp\left(-\sum_{k=1}^{2}\alpha_k L_j\right)\right)\right]\Bigg\}.$$

Derivatives for the Weibull Model

Considering the Weibull model and the presence of two competitive risks, that is, $k=1,2$, we have

$$F_1(t;\Theta)=\int_0^t\exp\left[-(\alpha_1u)^{\tau_1}-(\alpha_2u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du \tag{18}$$

and

$$F_2(t;\Theta)=\int_0^t\exp\left[-(\alpha_1u)^{\tau_1}-(\alpha_2u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du \tag{19}$$

where $\Theta=(\alpha_1,\tau_1,\alpha_2,\tau_2)$. These integrals do not have analytical solutions, thus requiring numerical methods.

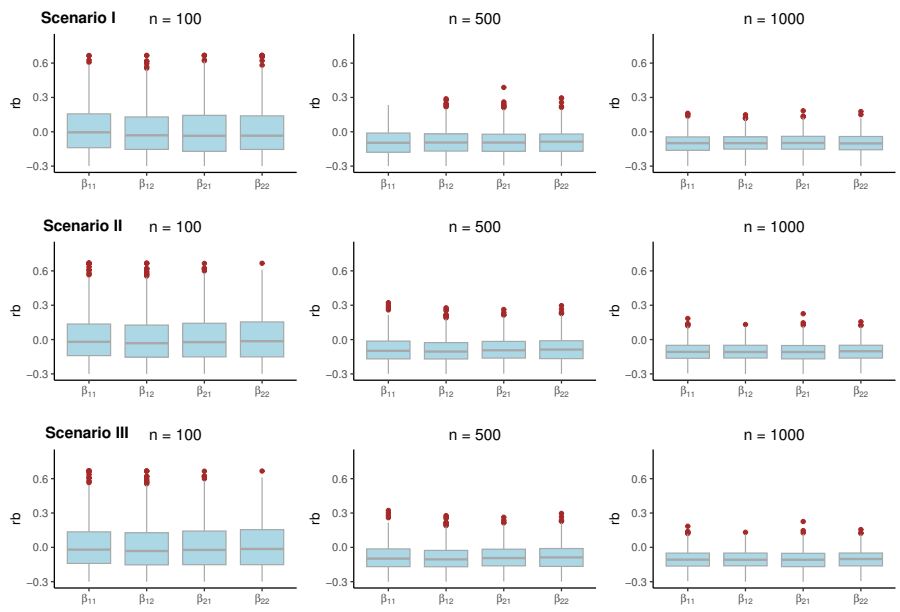


Fig. 1 Relative bias (rb) estimates of the parameters of the Exponential model

The first order derivatives of F_1 and F_2, about α_k and τ_k are:

$$\frac{\partial F_1(t;\Theta)}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du,$$

$$\frac{\partial F_1(t;\Theta)}{\partial \tau_k} = \frac{\partial}{\partial \tau_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_1\alpha_1^{\tau_1}u^{\tau_1-1}du,$$

$$\frac{\partial F_2(t;\Theta)}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du$$

and

$$\frac{\partial F_2(t;\Theta)}{\partial \tau_k} = \frac{\partial}{\partial \tau_k}\int_0^t \exp\left[-(\alpha_1 u)^{\tau_1} - (\alpha_2 u)^{\tau_2}\right]\tau_2\alpha_2^{\tau_2}u^{\tau_2-1}du$$

Appendix B: Additional Simulation Results

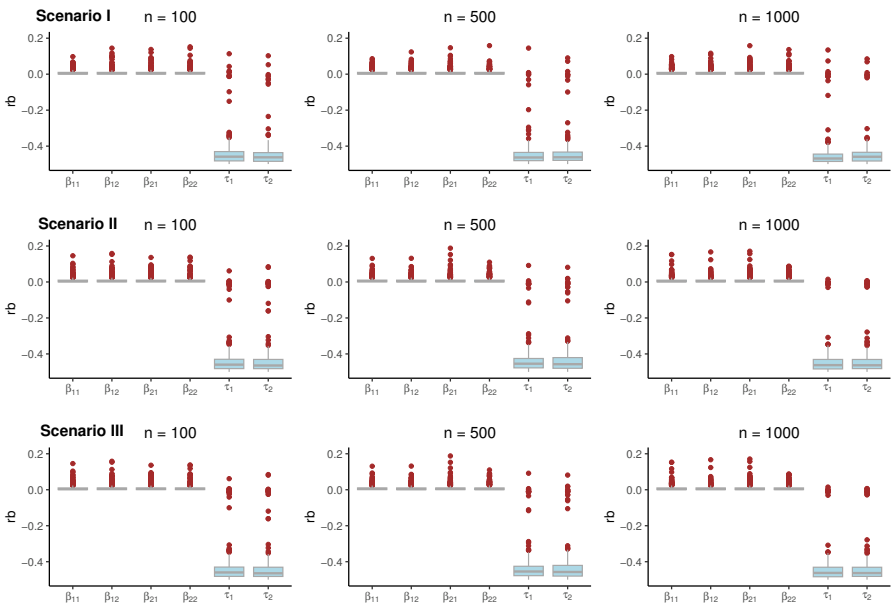


Fig. 2 Relative bias (rb) estimates of the parameters of the Weibull model

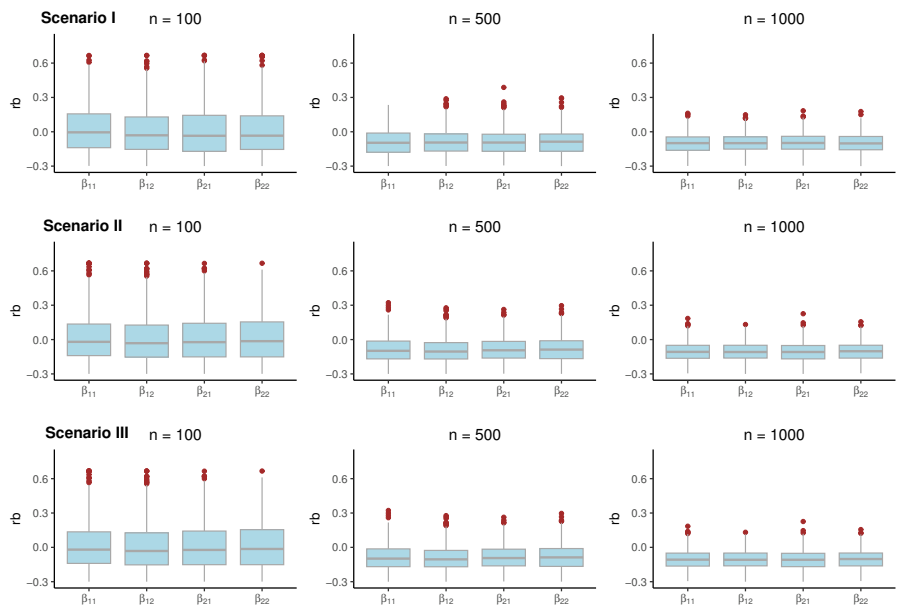


Fig. 3 Relative bias (rb) estimates of the parameters of the Exponential model

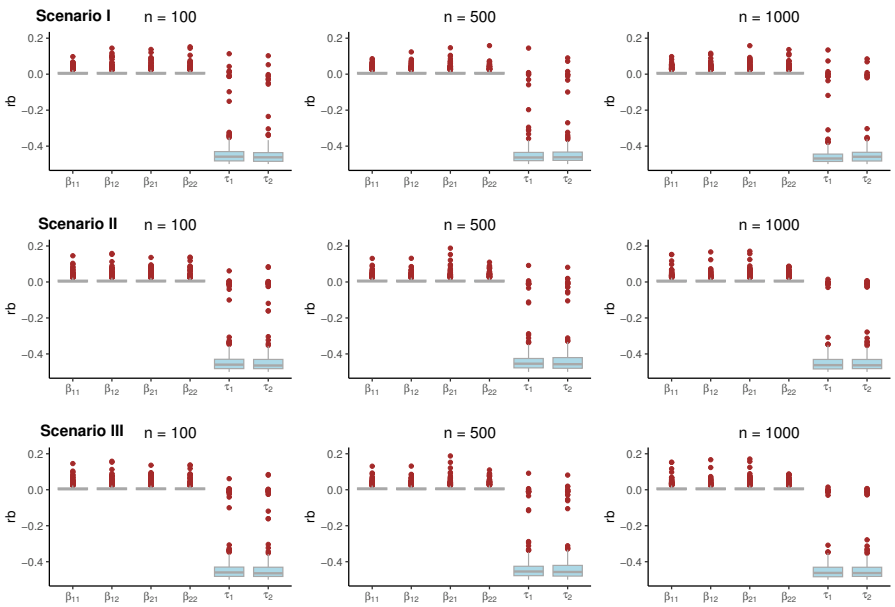


Fig. 4 Relative bias (rb) estimates of the parameters of the Weibull model

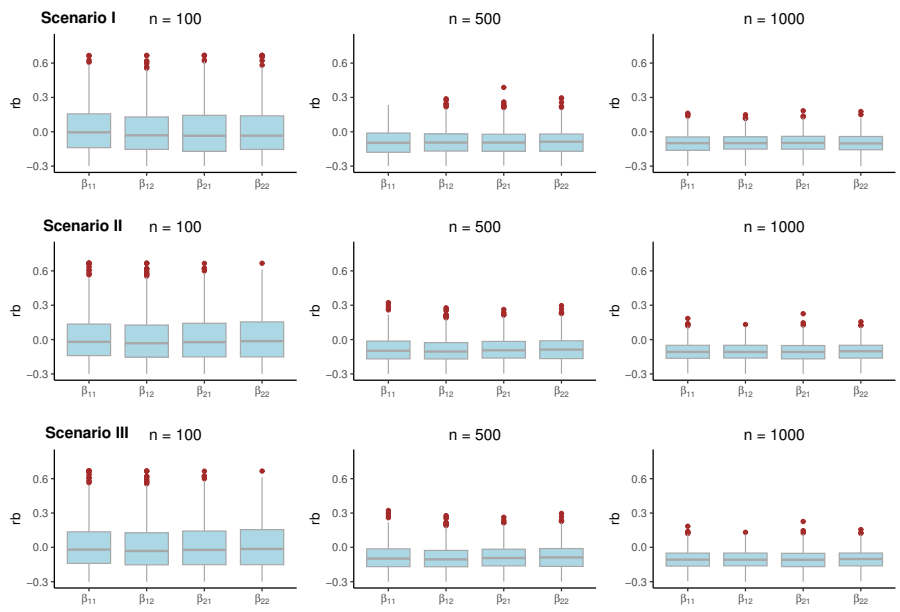


Fig. 5 Relative bias (rb) estimates of the parameters of the Exponential model

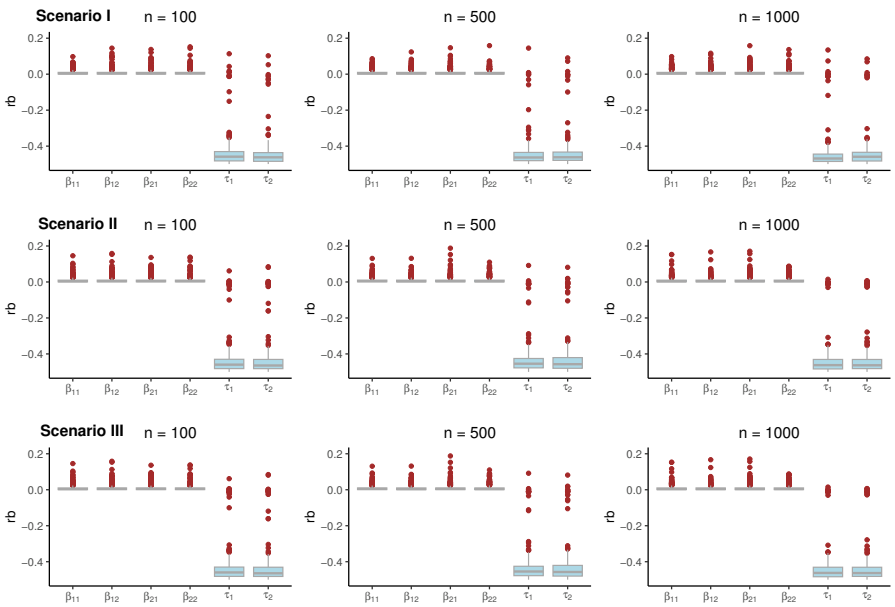


Fig. 6 Relative bias (rb) estimates of the parameters of the Weibull model