Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

Talita Evelin Nabarrete Tristão de Moraes

Análise de Sobrevivência com Fração de Cura: uma abordagem Bayesiana

Maringá - PR
2019

Talita Evelin Nabarrete Tristão de Moraes

Análise de Sobrevivência com Fração de Cura: uma abordagem Bayesiana

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito para o Exame de Defesa.

Orientadora: Prof. Dra. Isolde Previdelli
Coorientador: Prof. Dr. Giovani Loiola da Silva

Universidade Estadual de Maringá - UEM

Departamento de Estatística - DES

Programa de Pós-Graduação em Bioestatística

Maringá - PR

2019

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Moraes, Talita Evelin Nabarrete Tristão de

M827a Análise de sobrevivência com fração de cura : uma abordagem Bayesiana / Talita Evelin Nabarrete Tristão de Moraes. -- Maringá, PR, 2020.
76 f.: il. color.

Orientadora: Profa. Dra. Isolde Terezinha Santos Previdelli.
Coorientador: Prof. Dr. Giovani Loiola da Silva.
Dissertação (Mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2020.

1. Análise de sobrevivência (Bioestatística). 2. Estatística Bayesiana. 3. Análise longitudinal. 4. Modelos conjuntos mistos. 5. Fração de cura. I. Previdelli, Isolde Terezinha Santos, orient. II. Silva, Giovani Loiola da, coorient. III. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. IV. Título.

CDD 23.ed. 519.542

Márcia Regina Paiva de Brito - CRB-9/1267

TALITA EVELIN NABARRETE TRISTÃO DE MORAES

Análise de Sobrevivência com Fração de Cura: Uma Abordagem Bayesiana

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Profª. Dra. Isolde Previdelli
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Brian Alvarez Ribeiro de Melo
Universidade Estadual de Maringá – DES/UEM

Prof. Dr. Enrico Antônio Colosimo
Universidade Federal de Minas Gerais – ICEx/UFMG

Maringá, 31 de janeiro de 2020.

Scanned with CamScanner

“Dedico este trabalho primeiramente a mim, pelo meu esforço e dedicação no decorrer do estudo, dedicando ainda ao meu companheiro de todas as horas Cleber Silas Ferreia pela paciência e confiança que teve em mim”.

Agradecimentos

Agradeço a Deus pelos objetivos alcançados, durante todos os meus anos de estudos.

Aos familiares que de alguma forma contribuíram por essa conquista, principalmente ao meu irmão Paulo que me ajudou em todos os momentos difíceis.

Ao meu companheiro pelo apoio emocional que necessitei em vários momentos.

Aos amigos da turma de 2017 e aos novos da turma 2018, que sempre estiveram ao meu lado, pela amizade e pelo apoio demonstrado ao longo de todo o período de tempo em que me dediquei a este trabalho. E aos amigos em geral que sempre estiverem me apoiando em todos os momentos.

Um agradecimento especial a minha amiga Naiara Caroline, pelos dias e noites que passamos estudando e pelas ajudas no decorrer do mestrado. Principalmente por atravessar o oceano para explorarmos o mundo juntas, momentos que sempre serão lembrados.

Aos professores do mestrado pela ajuda na construção do conhecimento no decorrer das aulas.

A minha orientadora Isolde Previdelli pelas oportunidades oferecidas no mestrado, onde me oportunizou realizar um sonho de estudar em Portugal, tendo experiências indescritíveis.

Ao meu coorientador Giovani L. Silva pelos conhecimentos adquirido no período que estive em Portugal, pela paciência e acolhimento junto ao Instinto superior técnico de Lisboa.

À CAPES, pela bolsa que foi fundamental para minha permanência e dedicação ao curso.

“É muito melhor lançar-se em busca de conquistas grandiosas, mesmo expondo-se ao fracasso, do que alinhar-se com os pobres de espírito, que nem gozam muito nem sofrem muito, porque vivem numa penumbra cinzenta, onde não conhecem nem vitória, nem derrota.”

(Theodore Roosevelt)

Resumo

A utilização de modelos com fração de cura na análise de dados de sobrevivência provenientes da área da saúde tem crescido nos últimos anos devido às características de mudança na evolução de algumas doenças consideradas por vezes como crônicas atualmente e.g. câncer de mama ou AIDS. O presente trabalho visa analisar dados de sobrevivência com fração de cura em dois contextos atuais das Ciências da Saúde estando dividido em três partes. A primeira parte foca-se no estado da arte dos modelos de sobrevivência com ou sem fração de cura, incluindo a análise conjunta de dados longitudinais e de sobrevivência. As demais partes reportam-se aos resultados de dois artigos relacionados com esta temática e motivados por problemas reais. O primeiro artigo refere-se a análise de sobrevivência com fração de cura com aplicação a dados de pacientes com câncer de mama. Esses dados foram obtidos do site do Instituto Nacional de Câncer (INCA), durante o período de 2004 a 2008, levando em consideração 9 categorias do câncer de mama, segundo a Classificação Internacional de Doença (CID). No segundo artigo, faz-se uma análise conjunta de dados de sobrevivência de pacientes com AIDS, no período de 2002 a 2006, para os quais também se observa o número de células CD4 em exames dos pacientes ao longo do tempo (longitudinal). Em ambos os estudos, os pacientes são do estado do Paraná. Devido à complexidade dos modelos de sobrevivência abordados aqui, optou-se pela análise bayesiana dos mesmos via uma abordagem analítica por Aproximações de Laplace Encaixadas e Integradas (INLA).

Palavras-chaves: Fração de Cura. Análise Longitudinal. Modelos Conjuntos Mistos. INLA. Estatística Bayesiana. Análise de Sobrevivência.

Abstract

The use of cure fraction models in the analysis of survival data from the health area has been increased in recent years due to the characteristics of change in the evolution of some diseases, sometimes considered as chronic currently e.g. breast cancer or AIDS . The present work aims to analyze survival models with cure fraction in two current contexts of Health Sciences, being divided into three parts. The first part focuses on the state of the art of survival models with or without cure fraction, including joint analysis of longitudinal and survival data. The other parts report to the results of two articles. The first article refers to the survival analysis with a fraction of cure for application to data of patients with breast cancer. These data were obtained from the site of the National Cancer Institute (INCA), during the period from 2004 to 2008, taking into account 9 categories of breast cancer, according to the International Classification of Disease (ICD). In the second article, a joint analysis of survival data is made for AIDS patients in the period 2002 to 2006, for which the number of CD4 cells in patients examinations over time (longitudinal) is also observed. In both studies, the patients are from the state of Paraná. Due to the complexity of the survival models approached here, we opted for Bayesian analysis of them via an analytical approach by integrated nested Laplace approximations (INLA).

Keywords: Cure Fraction. Longitudinal analysis. Mixed Joint model. INLA. Bayesian Statistics. Survival Analysis.

Lista de ilustrações

Figura 3.3.1 Curvas de sobrevivência de Kaplan-Meier: (A) Todos as pacientes, com intervalos de confiança de 95%, (B) CID (2: C509, 1: C504, 0: Outros), (C) Raça (1: Branca, 0: Outras), (D) Morfologia (2: Carcinoma Ductal, 1: Carcinoma Lobular, 0: Outros), (E) Estado civil (1: Casada, 0: Outros) e (F) Faixa etária de mulheres diagnosticadas com câncer de mama no Estado do Paraná de 2004 a 2016. 42
Figura 3.3.2 Gráficos da função de sobrevivência das pacientes com câncer da mama, estimada pelos modelos Weibull com (1) e sem (2) fração de cura. . . . 45
Figura 3.3.3 Gráficos para diagnóstico do ajuste do modelo considerando a distribuição Weibull. 46
Figura 4.2.1 Curva de sobrevivência de Kaplan-Meier, com intervalos de confiança de 95%, para os dados de pacientes com HIV/AIDS. 52
Figura 4.2.2 Histogramas da contagem CD4 antes (esquerda) e após (direita) a transformação Box-Cox. 54
Figura 4.3.1 Curvas de sobrevivência de Kaplan-Meier: (A) Faixa etária, (B) PrevOI (0: Não, 1: Sim), (C) Sexo (0: Feminino, 1: Masculino). 60
Figura 4.3.2 Medidas de contagem $CD4$ ao longo do tempo (em dias). 60
Figura 4.3.3 Perfil médio de CD4 dos pacientes com HIV/AIDS por mulheres (esquerda) e homens (direita) ao longo do tempo (T em dias). 61
Figura 4.3.4 Distribuição das frequências relativas do número de exames por indivíduo ao longo do tempo (T em dias). 61
Figura 4.3.5 Gráficos Q-Q para os tempos de sobrevivência (T) com base nas distribuições Weibull (A), Exponencial (B) e Log-normal (C). 62
Figura 4.3.6 Trajetórias das medidas CD4 (esquerda) e densidades (direita) do tempo mediano de sobrevivência dos pacientes 3 e 90. 66

Figura 4.3.7 Gráficos de resíduos longitudinais bayesianos: resíduos *versus* índice (esquerdo) e resíduos *versus* tempos de recolha CD4 (direito). 67
Figura 4.3.8 Gráficos de resíduos de sobrevivência bayesianos: resíduos Cox-Snell (esquerdo), de desvio (centro) e de Martingale (direito). 67

Lista de tabelas

Tabela 1 – Características das pacientes com câncer da mama no Estado do Paraná no período de 2004 a 2016. 41
Tabela 2 – Valores de DIC e WAIC dos modelos de sobrevivência dos dados de câncer da mama, com e sem fração de cura e covariáveis selecionadas. 44
Tabela 3 – Quantidades *a posteriori:* média, desvio padrão (DP) e intervalos de credibilidade (IC) HPD a 95% dos parâmetros dos modelos de sobrevivência com fração de cura. 44
Tabela 4 – Estatísticas descritivas dos dados de pacientes com HIV/AIDS de algumas variáveis categorizadas. 59
Tabela 5 – Estatísticas descritivas dos dados de pacientes com HIV/AIDS das variáveis contínuas idade (anos), tempo de recolha de CD4 (TR em dias) e tempos de sobrevivência (TS em anos), com morte (M) ou censura (C). 60
Tabela 6 – Valores de DIC e WAIC dos modelos conjuntos dos dados HIV/AIDS, com termos longitudinal $w_1(t)$ e de sobrevivência $w_2(t)$. 64
Tabela 7 – Quantidades *a posteriori:* média, desvio padrão (DP) e intervalos de credibilidade (IC) HPD a 95% dos parâmetros do modelo V. 65

Lista de abreviaturas e siglas

AIDS	Síndrome da Imunodeficiência Adquirida
CD4	Grupamento de Diferenciação 4
CID	Classificação Internacional de Doenças
C500	Neoplasia maligna do mamilo e aréola
C501	Neoplasia maligna da porção central da mama
C502	Neoplasia maligna do quadrante superior interno da mama
C503	Neoplasia maligna do quadrante inferior interno da mama
C504	Neoplasia maligna do quadrante superior externo da mama
C505	Neoplasia maligna do quadrante inferior externo da mama
C506	Neoplasia maligna da porção axilar da mama
C508	Neoplasia maligna da mama com lesão invasiva
C509	Neoplasia maligna da mama, não especificada
HIV	*Human Immunodeficiency Virus* (vírus da imunodeficiência humana)
INCA	Instituto Nacional de Câncer
INLA	*Integrated Nested Laplace Approximation*
SUS	Sistema Único de Saúde
DIC	*Deviance Information Criterion* (O Critério de Informação deviance)

WAIC *Watanabe-Akaike Information Criterion* (Critério de Informação de Watanabe-Akaike)

Sumário

1 Introdução 16

2 Revisão da literatura 19

2.1 Análise de Sobrevivência 19

2.1.1 Modelos não paramétricos 20

2.1.2 Modelos paramétricos 21

2.1.3 Modelos semi-paramétricos 23

2.2 Modelo de sobrevivência com fração de cura 24

2.3 Modelos Conjuntos para dados Longitudinais e de Sobrevivência 25

2.4 Abordagem Bayesiana 26

2.4.1 Métodos computacionais 28

2.4.2 Seleção de Modelos 30

2.5 Análises de Resíduos 31

3 Modelos de sobrevivência com fração da cura para dados de câncer de mama 34

3.1 Introdução 34

3.2 Métodos 37

3.2.1 Material 37

3.2.2 Modelos com fração de cura 37

3.3 Resultados e discussão 40

3.3.1 Análise preliminar 40

3.3.2 Análise com fração de cura 43

3.4 Análise Final 46

4 Análise conjunta bayesiana de contagens longitudinais de CD4 e sobrevivência a longo prazo de pacientes com HIV no Paraná, Brasil 49

4.1 Introdução 49

4.2 Métodos 51

4.2.1 Material 51

4.2.2 Modelos conjuntos 53

4.2.2.1 Submodelo longitudinal 53

4.2.2.2 Submodelo de sobrevivência com fração de cura 55

4.2.2.3 Análise conjunta bayesiana 56

4.2.2.4 Avaliação de modelos 57
4.3 Resultados e discussão 59
4.3.1 Análise preliminar 59
4.3.2 Análise conjunta 62
4.4 Análise Final 67
5 Comentários finais 70

Referências 72

Capítulo 1

Introdução

Nas pesquisas experimentais ou observacionais, tem-se frequentemente por objetivo estudar a relação entre varáveis explicativas ou preditoras ou covariáveis e uma (ou mais) variável resposta. Estas relações em estatísticas são abordadas por meio de modelos de regressão, que podem ser lineares ou não-lineares, e são capazes de explicar o relacionamento entre as variáveis, realizar predições, estimar quantidades de interesse, entre outras.

Uma das metodologias bastante utilizadas na área da saúde são os modelos de regressão para dados de sobrevivência. Segundo Colosimo e Giolo (2006), na maioria das vezes, tem-se como variável resposta o tempo até a ocorrência de um evento de interesse, denominado tempo de falha ou tempo de sobrevivência. Uma das principais características dos dados de sobrevivência é a presença de censura, muito comum em estudos de coorte e ensaios–clínicos.

Coleman et al. (2011) utilizam-se dessa metodologia para investigar as disparidades de sobrevivência, com o objetivo de informar das políticas de saúde para elevar os seus padrões e reduzir as desigualdades na sobrevivência de pacientes com câncer. E ainda, usando da metodologia, Angelis et al. (2014) realizaram um estudo com registro de câncer em 29 países da Europa que mostrou diferenças significativas entre os países quanto à sobrevivência do câncer. Além disso, existem situações nos quais os indivíduos podem se tornar imunes ao evento de interesse e serem considerados como “curados”, ou seja, não suscetíveis ao evento de interesse. Para adequar essa nova característica pesquisadores tais como Berkson e Gage (1952) e Lambert et al. (2006), já relacionavam uma fração de cura a dados de sobrevivência, ou seja, a proporção de pacientes curados da doença, sendo essa uma medida útil para monitorar tendências na sobrevivência de

algumas doenças.

A presença da fração de cura em um conjunto de dados de sobrevivência é geralmente sugerida por um gráfico de curvas Kaplan-Meier, isto é, um método não paramétrico para estimar a função de sobrevivência. Quando existe fração de cura, o gráfico Kaplan-Meier apresenta um platô longo e estável com forte presença de dados censurados na parte extrema direita do gráfico sugerido por Corbière et al. (2009).

Além disso, tem-se o interesse em saber ainda o prosseguimento da doença nos pacientes ao longo do tempo, uma vez que os médicos e gestores da saúde necessitam destas informações para monitoramento da doença, tanto na evolução quanto na involução, para realização de algum prognóstico. Uma metodologia que fornece informações do ponto de vista quer da análise de sobrevivência quer da análise longitudinal é conhecida por modelos conjuntos mistos.

Segundo Rizopoulos (2012), esse tipo de modelagem vem crescendo nos últimos anos, e está sendo adotada nas áreas de pesquisa clínica tais como: i) análise do vírus da imunodeficiência humana (HIV) em que o interesse reside na associação entre contagem de células CD4 ou carga viral e o tempo até a ocorrência de Síndrome de imunodeficiência adquirida (AIDS), ii) estudos de cirrose hepática que investigam a associação entre bilirrubina sérica e o tempo até a morte de pacientes pela respectiva causa, iii) estudos de câncer de próstata nos quais o interesse reside na associação entre os níveis de antígeno prostático específico (PSA) e o tempo até o desenvolvimento do câncer de próstata.

Wulfsohn e Tsiatis (1997) demonstram que, ao ajustar um modelo conjunto para avaliar dados longitudinais e de sobrevivência simultaneamente, pode-se reduzir vieses e, como consequência, melhorar a sua precisão. Os pesquisadores Yu, Taylor e Sandler (2008) fazem uso desse modelo em dados de câncer de próstata, no qual estavam interessados em associar um biomarcador[1], ou seja, antígeno longitudinal específico da próstata ao tempo de sobrevivência de seus pacientes. O modelo de sobrevivência com fração de cura também pode ser utilizado na modelagem conjunta, como em Martins, Silva e Andreozzi (2017), que utilizaram de um modelo hierárquico bayesiano para a modelagem de dados longitudinais e de sobrevivência, considerando os efeitos do tempo num biomarcador e da fragilidade espacial no tempo de sobrevivência em pacientes com AIDS, com uma fração da população para descrever sobreviventes de longo prazo.

Esta metodologia é bastante informativa pois proporciona um entendimento da evolução e involução da doença no decorrer do tempo, i.e., o monitoramento da doença. Assim, um dos objetivos deste trabalho é aplicar a metodologia de modelos conjuntos para

[1] Em medicina, um biomarcador é um indicador mensurável da severidade ou da presença de algum estado de doença.

dados longitudinais e de Sobrevivência com fração de cura.

Este trabalho foi dividido em cinco capítulos. O primeiro se faz uma introdução geral do conteúdo abordado. O segundo uma revisão da literatura. O terceiro apresenta um modelo de sobrevivência com fração de cura para dados reais de pacientes com câncer de mama no estado do Paraná, entre os anos de 2004 a 2008, levando em consideração 9 topografias do câncer de mama segundo a Classificação Internacional de Doenças. No quarto capítulo, um modelo conjunto de sobrevivência com fração de cura e um biomarcador longitudinal foi adotado para dados de um estudo de coorte envolvendo 272 pacientes com HIV/AIDS no Estado do Paraná durante os anos de 2002 a 2006. No quinto Capítulo, um resumo geral dos capítulos anteriores.

Capítulo 2

Revisão da literatura

2.1 Análise de Sobrevivência

Segundo Colosimo e Giolo (2006), a análise de sobrevivência é uma metodologia utilizada para analisar dados em que na maioria das vezes tem como variável resposta, o tempo até a ocorrência de um evento de interesse, denominado por tempo de falha, tempo de vida ou tempo de sobrevivência. Uma das principais característica em relação aos dados de sobrevivência é a presença de censura, i.e., não obervação de alguns tempos de falha durante o estudo, com maior frequência em estudos de coorte e ensaios–clínicos.

Os dados de sobrevivência apresentam-se, na maioria das vezes, de forma incompleta. A censura e o trucamento são mecanismos essencial a estes dados. Contudo, havendo dados censurados, este tempo não é observado para todos os indivíduos. As censuras/truncamento são divididos em .

- *Censura a direita:* o evento de interesse não é observado até o término do estudo.
- *Censura a esquerda:* o evento de interesse já ocorre antes do indivíduo ser incluído no estudo.
- *Censura Intervalar:* o evento ocorreu em um intervalo de tempo.
- *Truncamento a Esquerda:* indivíduo que sofre desfecho antes do início do estudo é desconhecido.
- *Truncamento a Direita:* só são incluídos indivíduos que sofreram o evento.

Num conjunto de dados de sobrevivência com n indivíduos, o tempo de falhas ou censura do indivíduo $i - ésimo$ é representado, pelo par (t_i, δ_i) sendo t_i o tempo de falha, e δ_i

a variável indicadora de falha, isto é,

$$\delta_i = \begin{cases} 1, & \text{se} \quad t_i \text{ é um tempo de falha} \\ 0, & \text{se} \quad t_i \text{ é um tempo censurado.} \end{cases} \tag{2.1.1}$$

Existem alguns tipos de censura à direita: i) censura do tipo I, ocorre quando o tempo para o término do estudo é fixado antes do início e as observações que não apresentaram a falha até este instante são censuradas, ii) censura do tipo II, acontece quando uma quantidade de falhas a ser observada é previamente fixada, iii) censura do tipo III, quando um indivíduo é retirado do experimento após a ocorrência de uma falha que não está relacionada diretamente com o experimento de sobrevivência.

Outra particularidade na análise de sobrevivência é a distribuição do tempo de sobrevivência T (variável aleatória não negativa) que é caracterizado por três funções: função densidade de probabilidade, função de sobrevivência e função de risco (*hazard function*).

A função densidade de probabilidade *f(t)* é interpretada como o limite da probabilidade do evento de interesse ocorrer. No intervalo $(t, t + \triangle)$, dado por

$$f(t) = \lim_{\triangle t \to 0} \frac{P(t < T < t + \triangle t)}{\triangle t}. \tag{2.1.2}$$

De acordo Colosimo e Giolo (2006), a função de sobrevivência *S(t)* é definida como a probabilidade de uma observação não falhar até um certo tempo t,

$$S(t) = P(T \geq t), \tag{2.1.3}$$

Note-se que a função de distribuição acumulada é definida como a probabilidade de uma observação não sobreviver ao tempo t, isto é,

$$F(t) = 1 - S(t). \tag{2.1.4}$$

Tem-se ainda a função de risco instantânea *h(t)*, que é definida como a probabilidade do evento de interesse ocorrer entre os tempos t e $t + \triangle t$, dado que ele sobreviveu até o tempo t.

$$h(t) = \lim_{\triangle t \to 0} \frac{P(t \leq T < (t + \triangle t) | T \geq t)}{\triangle t}. \tag{2.1.5}$$

A definição ou não da distribuição do tempo de sobrevivência conduz a diferentes tipos de análises de sobrevivência i.e. não paramétrica, paramétrica e semi-paramétrica.

2.1.1 Modelos não paramétricos

O estimador de Kaplan-Meier é um instrumento fundamental nos modelos não paramétrico, ou seja, modelos nos quais não se assume uma distribuição de probabilidade para o

tempo de sobrevivência T. Estimativa da função de sobrevivência, dada por:

$$\hat{S(t)} = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right), \tag{2.1.6}$$

em que $t_1 < t_2 < \cdots < t_K$são os K tempos de falha distintos e ordenados, d_i é o número de falhas em t_i, $i = 1, \cdots, k$ e n_i é o número de indivíduos em risco em t_i, ou seja, indivíduos que não falharam e não foram censurados até o instante anterior a t_i.

As principais propriedades do estimador 2.1.6, são:

a) É não viciado para amostras grandes,

b) É fracamente consistente,

c) Converge assintótica para um processo gaussiano.

d) É o de máxima verossimilhança de *S(t)*.

Para mais detalhes, consulte por exemplo Colosimo e Giolo (2006).

2.1.2 Modelos paramétricos

Nos modelos de sobrevivência paramétricos há necessidade de definir uma família de distribuições de probabilidade para os tempos de sobrevivência, por exemplo, a distribuição exponencial, log-normal, Weibull. (vide e.g. (COLOSIMO; GIOLO, 2006)).

Confome a autora citada anteriormente, a distribuição exponencial foi a primeira distribuição a ser usada em modelos de sobrevivência. Em termos matemáticos, a distribuição exponencial é um dos modelos probabilísticos mais simples usados para descrever o tempo de falha, uma vez que se caracteriza por ter uma função de taxa de falha constante. Ela tem descrito adequadamente o tempo de vida de óleos isolantes e dielétricos, entre outros.

A função densidade de probabilidade do tempo de falha T com distribuição exponencial é dada por:

$$f(t) = \frac{1}{\alpha} \exp\left(-\frac{t}{\alpha}\right), \tag{2.1.7}$$

onde $t \geq 0$ e o parâmetro $\alpha > 0$ é o tempo médio de vida.

Além disso, as suas funções de sobrevivência $S(t)$ e de taxa de falha $h(t)$ são dadas, por

$$S(t) = \exp\left(-\frac{t}{\alpha}\right) \tag{2.1.8}$$

e

$$h(t) = \frac{1}{\alpha} \quad \text{para} \quad t > 0. \tag{2.1.9}$$

Outro modelo paramétrico basteante utilizada é a distribuição Weibull que atualmente a mais utilizada para modelar tempos de sobrevivência devido ao fato de sua função de risco apresentar uma grande variedade de formas, todas com uma propriedade básica: a sua taxa de falha é monótona, isto é, ela é crescente, decrescente ou constante. Segundo Colosimo e Giolo (2006), se T a sua função densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp^{\left(-\frac{t}{\alpha}\right)^{\gamma}} \quad , t > 0, \gamma > 0, \alpha > 0, \tag{2.1.10}$$

onde γ é o parâmetro de forma e α o de escala.

Para esta distribuição, as funções de sobrevivência e de risco são, respectivamente,

$$S(t) = \exp\left(-\frac{t}{\alpha}\right)^{\gamma} \tag{2.1.11}$$

e

$$h(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1}. \tag{2.1.12}$$

Tem-se ainda, que:

- Se $\gamma < 1$, a função de risco decresce;
- Se $\gamma > 1$, a função de risco cresce;
- Se $\gamma = 1$, tem-se a função de taxa de falha da distribuição exponencial, que é um caso particular da Weibull.

Assim como a distribuição Weibull, a distribuição log-normal é por vezes utilizada para caracterizar tempos de vida de produtos e indivíduos. Isto inclui fadiga de material, semicondutores, diodos e isolação elétrica. Ela também é bastante utilizada para descrever situações clínicas, tais como o tempo de vida de pacientes com leucemia. A função densidade de probabilidade de uma variável aleatória T com distribuição log-normal é dada por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\} \quad t > 0, \sigma >, \mu \in R \tag{2.1.13}$$

em que μ, é a média do logaritmo do tempo de falha, assim como σ é o desvio-padrão. Note-se que o logaritmo de uma variável com distribuição log-normal de parâmetros μ e σ tem uma distribuição normal com média μ e desvio-padrão σ.

As funções de sobrevivência e de taxa de falha de uma variável log-normal não apresentam uma forma analítica explícita e são, desse modo, representadas, respectivamente, por:

$$S(t) = \Phi\left(\frac{-log(t) + \mu}{\sigma}\right) \tag{2.1.14}$$

e

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.1.15}$$

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

2.1.3 Modelos semi-paramétricos

Outros modelos bastante usados em Análise de sobrevivência são os modelos semi-paramétricos, que possuem uma parte paramétrica e outra não paramétrica. Segundo Colosimo e Giolo (2006), o modelo semi-paramétrico de Cox é o mais utilizado em estudos clínicos por sua versatilidade.

De acordo com Lee e Wang (2003), o modelo de Cox possui a propriedade de que diferentes indivíduos têm funções de risco proporcionais, ou seja, a razão $\frac{h(t|\mathbf{x_1})}{h(t|\mathbf{x_2})}$ é constante ao longo do tempo t, onde x_1 e x_2 são valores do vetor de covariáveis de dois pacientes distintos. O modelo de regressão de Cox pode ser expresso como

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}), \tag{2.1.16}$$

onde $h_0(t)$ é a função de risco de base ou subjacente, i.e., a função de risco de um indivíduo quando o vetor de covariáveis $\mathbf{x}$ é nulo, e $g(x)$ representa o efeito das covariáveis. Pode-se, determinar a razão de risco de dois indivíduos com diferentes valores do vetor das covariáveis $\mathbf{x}_1$ e $\mathbf{x_2}$ por meio da seguinte equação

$$\frac{h(t|\mathbf{x_1})}{h(t|\mathbf{x_2})} = \frac{h_0(t)g(\mathbf{x_1})}{h_0(t)g(\mathbf{x_2})} = \frac{g(\mathbf{x_1})}{g(\mathbf{x_2})}. \tag{2.1.17}$$

Nota-se que a expressão (2.1.17) não depende do tempo, evidenciando a proporcionalidade dos riscos A função de risco 2.1.16 e a razão 2.1.17 são dadas respectivamente por

$$h(t|\mathbf{x}) = h_0(t)exp(x^T\beta) \qquad \text{e} \qquad \frac{h(t|\mathbf{x_1})}{h(t|\mathbf{x_2})} = \frac{exp({x_1}^T\beta)}{exp({x_2}^T\beta)}, \tag{2.1.18}$$

Para o uso do modelo de regressão de Cox, tem-se como suposição básica, que as taxas de falha sejam proporcionais ou, de forma equivalente para este modelo, que as taxas de falha acumulada sejam também proporcionais (COLOSIMO; GIOLO, 2006).

2.2 Modelo de sobrevivência com fração de cura

Segundo Maller e Zhou (1996), o modelo de sobrevivência com fração de cura comum supõe que a população em estudo é uma mistura de indivíduos que estão suscetíveis a experimentar o evento de interesse e de indivíduos que não estão suscetível a experimentar o evento. Existem várias abordagens para modelar a fração de indivíduos que não estão suscetíveis ao evento de interesse, desde abordagens paramétricas a não-paramétricas. Modelos de sobrevivência levando em consideração a fração de cura foram inicialmente propostos por Boag (1949), e mais tarde a metodologia de proporção de curados foi introduzida por Berkson e Gage (1952). Estes utilizaram uma função de sobrevivência em forma de mistura, onde a população é dividida em duas partes: uma que representa o tempo de sobrevivência para os indivíduos suscetíveis (não curados) e a outra com uma distribuição que permite calcular o tempo de sobrevivência dos imunes (curados).

De acordo com Ibrahim, Chen e Sinha (2014), o modelo de cura tem sido usado para modelar dados de sobrevivência para vários tipos de cânceres, incluindo câncer de mama, leucemia, câncer de próstata, câncer de cabeça e pescoço. Estas doenças possuem uma proporção significativa dos pacientes denominado "curados". Neste modelo, considera-se uma certa fração p da população curado, e os restantes $1-p$ para a população não curada. Diante disso, a função de sobrevivência para toda a população, que é denotada por $S(t)$, $t > 0$, é dada por,

$$S(t) = p + (1-p)S_0(t) = 1 - (1-p)F_0(t), \tag{2.2.1}$$

em que p é o parâmetro que representa a proporção de indivíduos que não estão sujeitos ao evento de interesse. $(0 < p < 1)$, $S_0(t)$ é a função de sobrevivência própria, ou seja, uma função para os indivíduos que estão sob risco de falha e uma função de sobrevivência para os indículos imunes ao evento, associada aos indivíduos não curados e $F_0(t) = 1 - S_0(t)$ que, se t tende ao infinito, então $S(t)$ tende a p, pois $\lim_{t\to\infty} S_0(t) = 0$. Dessa forma a função de sobrevivência do modelo com fração de cura padrão tem uma assíntota em p. A função densidade de probabilidade associada à 2.2.1 é dada por

$$f(t) = \frac{d(1 - S(t))}{dt} = (1-p)f_0(t), \tag{2.2.2}$$

onde $f_0(t)$ é a função densidade de probabilidade base para os indivíduos suscetíveis ao evento.

Considerando um conjunto de dados de sobrevivência com n indivíduos, com correspondentes distribuições densidade de probabilidade $f(t)$ e de sobrevivência $S(t)$ indexadas por um vetor paramétrico, a função de verossimilhança com base na amostra

$(t_i, \delta_i), i = 1, ...n$, é

$$\prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}. \tag{2.2.3}$$

Assim, a função de verossimilhança do modelo de sobrevivência com fração de cura para uma amostra (t_i, δ_i), $i = 1, \cdots, n$, é dada por

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} [(1-p) f_0(t_i|\boldsymbol{\theta})]^{\delta_i} [p + (1-p) S_0(t_i|\boldsymbol{\theta})]^{1-\delta_i}. \tag{2.2.4}$$

2.3 Modelos Conjuntos para dados Longitudinais e de Sobrevivência

Segundo Tsiatis, Degruttola e Wulfsohn (1995) , os modelos conjuntos para dados longitudinais e de sobrevivência são uma adição mais recente aos modelos de sobrevivência que podem lidar com variáveis dependentes do tempo conjuntamente. De acordo com Tsiatis e Davidian (2004) e Rizopoulos (2012), um modelo conjunto é composto de dois submodelos interligados, um para o processo longitudinal e outro para o tempo de falha, abordando questões de pesquisas que envolvem a estrutura de associação entre medidas repetidas e o tempo até a ocorrência do evento de interesse. Outros exemplos de modelos conjuntos são apresentados em Henderson, Diggle e Dobson (2000) e Rizopoulos (2012), que trabalham com modelos conjuntos, para dados longitudinal e de sobrevivência. Podem ser vistos em Brown e Ibrahim (2003), Yu, Taylor e Sandler (2008), outros exemplos de modelos conjuntos usando fração de cura. Há também estudos recentes com o enfoque Bayesiano como pode ser visto em Martins, Silva e Andreozzi (2017), Zhang et al. (2019) e Li e Luo (2019), entre outros.

A construção do modelo começa, habitualmente, por considerar submodelos separados que depois são ligados entre si. Daqui resulta que um modelo conjunto tem sempre três componentes essenciais: uma para o processo longitudinal, outra para o tempo até o evento e uma terceira que lida com as especificações e pressupostos que permitem ligar as duas primeiras componentes no modelo final.

Aqui a especificação do modelo conjunto será feita seguindo a nomenclatura de Martins, Silva e Andreozzi (2017), onde se realizou uma análise conjunta de dados longitudinais e de sobrevivência de AIDS com fração espacial para sobreviventes de longo prazo. Assim, para o modelo linear misto relativamente aos dados longitudinais será considerado a expressão

$$\begin{cases} \mathbf{Y}_i = \mathbf{x}_i^T \beta + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i \\ \mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}) \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \mathbf{R}_i), \end{cases} \tag{2.3.1}$$

em que

- $\mathbf{Y}_i$ é um vetor da variável resposta da i-ésima unidade amostral;
- $\mathbf{x}_i^T$ é o vetor da matriz de incidência dos efeitos fixos da i-ésima unidade amostral;
- β é o vetor de parâmetros de efeitos fixos;
- Z_i é a matriz associada aos efeitos aleatórios da i-ésima unidade amostral;
- $\mathbf{b_i}$ é um vector de efeitos aleatórios que descrevem o desvio do perfil longitudinal do indivíduo i relativamente á média populacional.
- $\mathbf{D}$ é a matriz das variâncias e covariâncias dos efeitos aleatórios;
- $\boldsymbol{\epsilon}_i$ é o vetor de erros aleatórios intra-indivíduos da i-ésima unidade amostral;
- $\mathcal{R}_i$ é a matriz de covariâncias dos erros aleatórios ϵ_i, cujos elementos da diagonal medem a variabilidade das medidas repetidas em cada indivíduo.

Seja $Y(t)$ a medida longitudinal no instante t, T o tempo de sobrevivência, e $Z = Z(t); t \geq 0$, o processo de covariáveis externas no tempo *t* usado para introduzir os efeitos aleatórios que explicam a correlação entre componentes longitudinais e de sobrevida. Assume-se que $Y(t)$ e T são independentes, condicionalmente a $\mathbf{Z_i}$ e $\mathbf{b_i}$ no $i - ésimo$ indivíduo. Então, a função de verossimilhança do modelo conjunto compreende em duas partes uma para os dados longitudinais, denotada por L_{1i} e outra para os dados de sobrevivência com fração de cura, denotada por L_{2i}.

$$\pi(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{i=1}^{n} \left[\prod_{j=1}^{n_i} f_L(y_{ij}|\mathbf{b}_i, \boldsymbol{\theta}_L) \right] f_S(t_i|\mathbf{b}_i, \boldsymbol{\theta}_S) \times \pi(\boldsymbol{\theta}). \tag{2.3.2}$$

2.4 Abordagem Bayesiana

A Inferência Bayesiana vem crescendo nos últimos anos devido ao importante avanço computacional. De acordo com Gelman, Hwang e Vehtari (2014), a Inferência Bayesiana pode ser entendida como o processo de ajustar um modelo de probabilidade para um conjunto de dados, resumindo o resultado por uma distribuição de probabilidade nos parâmetros do modelo e em quantidades não observáveis como predição para novas observações. Segundo Paulino et al. (2018) e Amaral-Turkman, Paulino e Müller (2019), na abordagem Bayesiana, é conveniente distinguir entre dois objetivos: a inferência sobre parâmetros desconhecidos e inferência sobre dados futuros. Na Inferência Bayesiana, a informação que

se tem sobre um parâmetro θ assume diferentes graus de incerteza e estes são representados por meio de modelos probabilísticos para θ. Assim, é possível que os pesquisadores possuam diferentes graus de incerteza sobre θ, considerando modelos distintos para θ.

Tem-se ainda, que na Inferência Bayesiana o parâmetro de interesse é tomado como uma quantidade aleatória, tendo uma distribuição de probabilidade associada. Isso distingue a Inferência Bayesiana da Inferência Frequentista. A combinação de informação *a priori* com a informação proveniente dos dados amostrais, através do Teorema de Bayes, origina a distribuição *a posteriori*.

O Teorema de Bayes fornece uma descrição para encontrar a probabilidade de um evento por meio do conhecimento prévio de condições associadas à ocorrência desse evento, sendo descrito como

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}, \tag{2.4.1}$$

sendo que os eventos A_i, $i = 1, ..., m$ representam os "antecedentes" a que o investigador atribui graus de credibilidade ou probabilidades *a priori* $P(A_i)$, de natureza subjetiva. Depois da informação adicional que consiste em saber que o acontecimento B se realizou, o investigador atualiza as suas probabilidades a priori através da fórmula Bayes e passa atribuir aos A_i as probabilidades *a posteriori* $P(A_i|B)$ (vide Paulino et al. (2018)).

Assim, tem-se como ideia básica, substituir $P(B|A_i)$ pela função de verossimilhança gerada por uma amostra $\mathbf{x} = (x_1, ..., x_n)$ através de uma distribuição de probabilidade indexada por um parâmetro θ. Qualquer informação subjetiva atribuída a esse parâmetro é chamada de informação *a priori*, e essa informação é quantificada formalmente por uma distribuição de probabilidades $h(\theta)$ *a priori* para o parâmetro θ.

Por meio, do Teorema de Bayes a distribuição $h(\theta)$ é combinada com a verossimilhança, denotada por $f(x|\theta)$ e, desta forma, origina a distribuição *a posteriori* denotada por $\pi(\theta|x)$, o que permite realizar toda a inferência. Desse modo, a distribuição *a posteriori* é expressa por

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta))}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}, \theta \in \mathbf{\Theta}, \tag{2.4.2}$$

onde Θ é o espaço paramétrico. Como o denominador de (2.4.2), conhecido por distribuição preditiva *f(x)*, não depende de θ, a distribuição *a posteriori* pode ser reescrita como

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta), \theta \in \mathbf{\Theta}. \tag{2.4.3}$$

Segundo Gelman, Hwang e Vehtari (2014), a função de verossimilhança tem um importante papel na fórmula de Bayes pois representa o meio através do qual os dados, **x**, transformam o conhecimento a priori sobre θ. Normalmente, os parâmetros associados às distribuições a prior são chamados de hiper-parâmetros para não haver confusão com o conjunto dos parâmetros do modelo. Nesse trabalho a os parâmetos de interesse foram estimados utilizando o método INLA (*Integrated Nested Laplace Approximation*) (Rue, Martino e Chopin (2009)), devido à dificuldade em obter as distribuições marginais *a posteriori* analiticamente.

2.4.1 Métodos computacionais

A Inferência Bayesiana necessita de programação computacional intensiva e os métodos mais difundidos para a obtenção de resumos da distribuição conjunta *a posteriori* são os Métodos de Monte Carlo com Cadeias de Markov – MCMC, especialmente o algoritmo de Metropolis-Hasting e o amostrados de Gibbs. A execução do algoritmo de Metropolis-Hasting e do amostrador de Gibbs em problemas complexos pode ser demorada para a obtenção da convergência do parâmetro.

Uma alternativa proposta aos métodos é o uso da estimação baseada na abordagem INLA (Integrated Nested Laplace Approximation), proposta por Rue, Martino e Chopin (2009). Segundo Amaral-Turkman, Paulino e Müller (2019), uma das principais vantagens no INLA é a obtenção de resultados dentro de segundos ou minutos em modelos que levaria horas ou mesmo dias com o uso dos algoritmos MCMC.

O método INLA permite a aproximação das distribuições marginais *a posteriori* de modelos estruturados aditivos. Estes modelos formam classes de modelos onde tem-se suposições mais flexíveis do que, por exemplo, a classe dos Modelos Lineares Generalizados (MLG).

De acordo com Amaral-Turkman, Paulino e Müller (2019), um subconjunto específico de modelos aditivos Bayesianos é a classe de modelos Gaussianos latentes (LGM) que pode ser modelada eficientemente usando a metodologia INLA implementada no *package INLA* no software (R versão 3.4.4 R Core Team (2018)). Essa classe compreende modelos bem conhecidos, como modelos mistos, modelos temporais e espaciais. Nesses modelos, a variável dependente (Y_i) é suposta pertencer a família exponencial com média μ_i. Ao utilizarmos uma função de ligação $g(.)$ temos um preditor $\eta_i = g(\mu_i)$, onde η pode ser escrito como,

$$\eta_i = \beta_o + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} w_{ki} f^{(k)}(u_{ki}) \epsilon_i, \tag{2.4.4}$$

onde β_0 é o intercepto, β são os efeitos lineares das covariáveis $\mathbf{z}$, $f^{(k)}(u_{ki})$ são funções desconhecidas das covariáveis u do modelo. Essas funções podem ter um pesos w_{ki}, fixos e conhecidos por cada observação, e ϵ_i o termo não estruturado. Para a utilização do método INLA, os modelos Estruturados Aditivos são escritos como modelo hierárquico Bayesiano que são amplamente utilizados em várias aplicações. A distribuição *a posteriori* de interesse é

$$\pi(\theta, \psi|\mathbf{y}) \propto \pi(\theta|\psi)\pi(\psi) \prod_i f(y_i|\theta_i, \psi),$$

onde $\mathbf{y}$ é o vetor de valores da variável resposta observado na amostra; θ é um vetor que contém todos os parâmetros desconhecidos em η_i (campo latente gaussiano); ψ é um vetor de hiperparâmetros (parâmetro da distribuição *a priori* gaussiana do campo latente); $\pi(\theta, \psi|y)$ é a distribuição *a posteriori* conjunta para parâmetros θ e ψ, $\pi(\psi)$ é distribuição *a priori* atribuída ao vetor θ, $\pi(\theta|\psi)$ é a distribuição *a priori* de θ dado o vetor de hiperparâmetros ψ, e $\prod_i f(y_i|\theta_i)$ a função de verossimilhança.

O principal objetivo do INLA é aproximar a distribuição *a posteriori* marginal dos parâmetros latentes e dos hiperparâmetros no modelo Gaussiano, respectivamente como

$$\pi(\theta_i|y) \qquad \text{e} \qquad \pi(\psi_k|y). \tag{2.4.5}$$

As distribuições *a posteriori* marginais 2.4.5 podem ser escritas como

$$\pi(\theta_i|y) = \int \pi(\theta_i, \psi|y)d\psi = \int \pi(\psi|y)\pi(\theta_i|\psi, y)d\psi \tag{2.4.6}$$

$$\pi(\psi_k|y) = \int \pi(\psi|y)d\psi_{-(k)}, \tag{2.4.7}$$

onde $\psi_{-(k)}$ denota o vetor sem o componente k. Estas integrais geralmente não são resolvíveis analiticamente, uma solução encontrada é aproximá-las. Estas distribuições requerem primeiras estimativas de $\pi(\theta_i|y)$ e $\pi(\psi_k|y)$ para em seguida, obter

$$\tilde{\pi}(\theta_i|y) \qquad \text{e} \qquad \tilde{\pi}(\psi_k|y), \tag{2.4.8}$$

onde $\tilde{\pi}(.|.)$ é a aproximação das equações 2.4.5, que é implementado em três etapas: (i) aproximação para $\pi(\psi_k|y)$; (ii) uma aproximação para $\pi(\psi|y)$ e (iii) uma integração numérica (AMARAL-TURKMAN; PAULINO; MÜLLER, 2019).

2.4.2 Seleção de Modelos

Para comparar os modelos é necessário utilizar critérios que permitam a seleção de um modelo que atenda às necessidades do pesquisador. No contexto Bayesiano várias medidas de seleção já foram propostas, sendo elas baseadas em medidas de ajustes de diferentes modelos. A seleção de modelos constitui uma importante ferramenta da modelagem estatística, por meio dela o pesquisador tem um embasamento adicional para justificar a escolha de um modelo adequado. A seguir, são apresentados os usuais critérios de seleção dos modelos.

- O Critério de Informação *deviance* (DIC) é um dos mais utilizados dentro da estatística Bayesiana, o *deviance* é definido por $D(\theta) = -2log(p(\mathbf{y}|\theta)) + C$, onde ***y*** representa os dados, θ os parâmetros desconhecidos do modelo e $p(y|\theta)$ a função de probabilidade e *C* é uma constante que cancela em todos os cálculos que comparam modelos diferentes, e que, portanto, não precisa ser conhecido. Spiegelhalter et al. (2002) expressa o Critério de Informação *deviance* como,

$$DIC = D(\overline{\theta}) + 2p_D, \tag{2.4.9}$$

  onde $D(\overline{\theta})$ é a média *a posteriori* do desvio do modelo e p_D é a penalização do modelo (número efetivo de parâmetros), dada pela diferença entre a média *a posteriori* da *deviance* ($\overline{D(\theta)}$) e a *deviance* das médias *a posteriori* ($D(\overline{\theta})$) dos parâmetros do modelo de interesse.sendo expresso por

$$DIC = 2(\overline{D\theta}) - D(\overline{\theta}). \tag{2.4.10}$$

- Critério de Informação de Watanabe-Akaike (WAIC) é definido por,

$$WAIC = DEV_W + p_W \tag{2.4.11}$$

  onde

$$DEV_W = \sum_{i=1}^{n} logE_{\boldsymbol{\theta}|\mathbf{y}}[f(y_i|\boldsymbol{\theta})] \qquad \text{e} \qquad p_w = \sum_{i=1}^{n} V_{y_i|\boldsymbol{\theta}}[logp(y_i\boldsymbol{\theta})].$$

  Por meio da distribuição *a posteriori* pode-se observar o comportamento dos parâmetros após a inclusão das novas informações. Dessa forma os valores pequenos destas medidas indicam o melhor modelo.

2.5 Análises de Resíduos

Para verificar a adequabilidade de um modelo usa-se frequentemente análise de resíduos. Dado os parâmetros desconhecidos e os preditores $\mathbf{x}_i$ para um ponto de dados y_i, resíduos bayesianos é dada por,

$$r_i = \epsilon_i(\beta) = y_i - x_i^T(\hat{\beta}), i = 1, ..., n. \tag{2.5.1}$$

Esses resíduos são aplicáveis ao submodelo longitudinal onde baseia-se em uma estimativa pontual do parâmetro $\hat{\beta}$. A priori desses resíduos são distribuídos como $N(0, \sigma^2)$, e *a posteriori* da distribuição pode ser obtida nas distribuição *a posteriori* de β e $\sigma^{\mathbf{2}}$.

Para obter um gráfico residual bayesiano, gera-se uma amostras da distribuição *a posteriori* de $\sigma^{\mathbf{2}}$ e depois e substituída no valor obtido pela densidade para produzir amostras da distribuição *a posteriori* dos resíduos. A média *a posteriori* ou a mediana do r_i pode ser calculada e examinada. Plotando esses resíduos versus os valores ajustados podemos detectar discrepâncias e revelar falhas nas suposições do modelo, como a suposição de normalidade ou homogeneidade.

No contexto da análise de sobrevivência, os métodos de avaliação do modelo têm de lidar com a existência de censura, o que os torna um pouco mais complicados do que os correspondentes métodos utilizados em outros modelos de regressão.De acordo com Rizopoulos (2012), tem-se algumas ferramentas para avaliar alguns pressupostos como heterocedasticidade, linearidade entre outros, para essas análise tem-se os resíduos de Cox-Snell, marginais e de Martingale, outro resíduo bastante utilizado na analise bayesiana é o resíduos de desvio.

O primeiro tipo de resíduos que será discutido aqui são os chamados resíduos Cox-Snell (COX; SNELL, 1968). Seja $\hat{\theta}$ o vector de parâmetros a estimar no modelo. Os resíduos de Cox-Snell é definido por,

$$r^*_{Ci} = H_i(t_i, \hat{\theta}|\mathbf{X}_i) = log(S(t_i, \hat{\theta}|\mathbf{X}_i)), \quad i = 1, ..., n, \tag{2.5.2}$$

onde $\hat{\theta}$ são as estimativas máxima de θ. Ou seja, os resíduos são as estimativas de probabilidade máxima dos riscos cumulativos para a sobrevivência observada no tempo dado as covariáveis.

Na versão bayesiana dos resíduos tem-se:

$$r_{C_i} = H_i(t_i, \theta|\mathbf{X}_i) = log(S(t_i, \theta|\mathbf{X}_i)), \quad i = 1, ..., n, \tag{2.5.3}$$

onde r_{C_i} é uma função de parâmetros desconhecidos, e a distribuição *a posteriori* é, portanto, simples de calcular, podendo extrair amostras do distribuição de θ e, em seguida, substituir essas amostras em (2.5.3) para produzir amostras a partir da distribuição *a posteriori* dos resíduos. A média posterior ou mediana dos $r'_{Ci}s$ podem ser calculados e avaliados (WANG; RYAN; FARAWAY, 2018).

Com base nos resíduos de Cox-Snell, dois outros tipos de resíduos podem ser calculados. A versão Bayesiana dos resíduos de Martingale é dado por

$$r_{M_i} = \delta_i - r_{C_i} \tag{2.5.4}$$

r_{M_i} nos fornece uma medida da diferença entre o indicador (δ_i) de falha e o número esperado de eventos que o indivíduo experimentaria.

Os resíduos de Martingale são uma realocação dos resíduos de Cox-Snell para uma média de zero para observações sem censura, de modo que eles seguem propriedades semelhantes aos resíduos na análise de regressão linear. De acordo com Wang, Ryan e Faraway (2018), na prática, esses resíduos não são muitas vezes distribuído simetricamente em torno de zero, o que os torna difíceis em relação a interpretação.

Já os resíduos de desvio (*Deviance*) são usados para obter resíduos que têm uma distribuição normal, e são dados por,

$$r_{D_i} = sgn(r_{M_i})[-2r_{M_i} + \boldsymbol{\delta}_i - r_{M_i}]^{\frac{1}{2}}, \quad i = 1, ..., n, \tag{2.5.5}$$

-

onde sgn(.) é o sinal da função, r_{M_i} é o resíduo de martingale (4.2.9) . Note-se ainda que r_{D_i} tem valor 0 quando r_{M_i} é zero.

Tais gráficos são facilmente e diretamente disponibilizados com o pacote do R (*brinla*).

Capítulo 3

Modelos de sobrevivência com fração da cura para dados de câncer de mama

Resumo

O câncer da mama é uma das doenças mais comum entre as mulheres em todo o mundo com cerca de 25% de casos novos a cada ano. No Brasil, esse percentual é de 29,7%. Em 2020, espera-se 66.280 novos casos de câncer de mama no Brasil, sendo este o câncer mais frequente nas mulheres das regiões Sul, Sudeste, Centro-Oeste e Nordeste, segundo o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Neste sentido, a análise de sobrevivência tem sido útil para a identificação e compreensão de fatores de risco e de prognóstico para pacientes com câncer em geral, bem como na comparação de tratamentos da doença realizados em vários ensaios clínicos. Com o objetivo de investigar o valor prognóstico de variáveis demográficas, clínico-patológicas e relacionadas com o tipo de tratamento em relação ao tempo de sobrevivência de doentes com câncer da mama, utilizaram-se dados de um estudo transversal obtido através do INCA, com 2092 pacientes diagnosticados com câncer de mama no estado do Paraná, no período de 2004 a 2006, sendo estes observados até 2016. Com base neste estudo, ajustaram-se modelos de mistura com distribuição de sobrevivência Weibull na presença de fração de cura, sob o enfoque bayesiano por meio de métodos de aproximações de Laplace encaixadas e integradas (INLA). Estes modelos de sobrevivência mostram-se mais convenientes e condizentes com a realidade e os seus resultados apontam para uma proporção de doentes curados de aproximadamente 57.6% na população em estudo. Na avaliação de potencias fatores de risco, conclui-se que a classificação C509, o tipo Carcinoma Ductal e a raça branca tiveram um efeito positivo no risco de morte devido ao câncer da mama, em oposição ao efeito negativo das faixas etárias mais velhas, a classificação C504, o tipo Carcinoma Lobular e pacientes casados. Notou-se ainda que a introdução desses fatores produziu um melhor ajustamento nos modelos de sobrevivência com fração de cura.

Palavras-chave: Câncer de Mama, fração de cura, INLA, estatística bayesiana, análise de sobrevivência.

3.1 Introdução

O tratamento do câncer tem tido muita evolução nas últimas décadas e continua a ter um crescente acompanhamento por parte dos pesquisadores na Medicina e não só. Segundo o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA) (INCA, 2020), a estimativa de incidência da câncer de mama no ano de 2020 no Brasil será 66.280 novos

casos, representando 29.7% dos vários tipos de câncer em mulheres. Araújo e Fernandes (2008) afirmam que o câncer de mama é mais preocupante devido à sua maior frequência e aos efeitos psicológicos causados na imagem pessoal, bem como que o impacto do diagnóstico do câncer de mama, quando o mesmo leva a sentimentos de medo da morte e de o câncer ser uma doença irremediável. Esse tipo de câncer ocorre na maior parte em mulheres, não isentando o sexo masculino mas com uma menor incidência, representando aproximadamente 1% do total de casos da doença (APARÍCIO, 2013).

De acordo com o INCA (INCA, 2020), o câncer de mama é uma doença causada pela multiplicação desordenada de células da mama, podendo evoluir de diferentes formas, e ainda se diferenciar em relação ao seu desenvolvimento. Esses comportamentos devem-se a característica própria de cada tumor. Observe-se ainda que há vários tipos de câncer de mama, que são discriminados pela Classificação Internacional de Doenças (CID) para Oncologia. Essa é uma ferramenta que permite obter mais informação sobre as patologias catalogadas pela Organização Mundial de Saúde (OMS), variando de C500 à C509. Assim há necessidade de estudar as classificações oncológicas CID separadamente, podendo então averiguar o comportamento individual referente à localização de cada tumor.

Vários modelos de sobrevivência têm sido usados para analisar dados de câncer tais como modelos paramétricos e.g. regressão Weibull (vide referências em (LEE; WANG, 2003; LAWLESS, 2003; COLOSIMO; GIOLO, 2006)) ou modelos semiparamétricos e.g. regressão Cox (COX, 1972). Tais exemplos de modelos de sobrevivência são conhecidos por modelos de riscos proporcionais, sendo bastante populares na análise de tempos de sobrevivência contínuos devido à sua fácil implementação e interpretação. Alternativamente, pode-se estudar dados de sobrevivência discretos utilizando modelos multiníveis. Suguiura (2017) apresentou um estudo de câncer de mama que levou em consideração o espaço geográfico das Regionais de Saúde no Estado do Paraná sob um modelo multinível Multinomial com efeitos aleatórios gaussianos. O seu principal objetivo foi verificar a frequência de ocorrência de determinado tipo de CID atendendo à localidade de residência do paciente, assim como outros fatores de risco da doença.

Em alguns ensaios clínicos de câncer, tem-se observado o aumento de uma proporção significativa de pacientes que não experimentam o evento de interesse durante o período em estudo, sendo comumente conhecidos por pacientes “curados” ou sobreviventes a longo prazo. Os modelos de sobrevivência que contemplam doentes “imunes” ao acontecimento de interesse são denominados por modelos de sobrevivência com fração de cura. Tais modelos têm tornado-se alvo de interesse por parte dos gestores de saúde, que querem acompanhar as tendências na sobrevivência de doenças curáveis.

Com base numa experiência de sobrevivência de pacientes após tratamento de câncer,

Berkson e Gage (1952) desenvolveram uma função simples, em termos de dois parâmetros fisicamente significativos, que se encaixa muito bem nesses dados de sobrevivência. Esses dois parâmetros podem ser usados para comparar sucintamente a mortalidade de dois grupos, diferentes em relação ao tratamento, tipo de câncer ou outras características. O primeiro parâmetro representa a proporção da população que está sujeita apenas a taxas de mortalidade "normais" ("curados"), enquanto o segundo é a taxa de mortalidade por câncer, para a qual o restante da população ("não curados") está sujeito. Lambert et al. (2006) relatam que em estudos de câncer, a cura ocorre quando a taxa de mortalidade (risco) retorna ao mesmo nível que o esperado na população geral.

Alguns estudos têm sido publicados em torno dessa metodologia, incorporando diferentes formas da fração de cura. Sposto (2002) analisou dados de crianças com linfomas e leucemias, com uma proporção significativa de pacientes curados após a terapia através de um modelo paramétrico de não-mistura para incorporar a mortalidade e assim obter as estimativas da fração de cura. Cucchetti et al. (2015) fizeram uma análise de sobrevivência com fração de cura para pacientes com fração de cura após hepatectomia de metástases hepáticas colorretais. Ramires et al. (2018) utilizaram de modelos aditivos generalizados, com o objetivo de investigar a proporção da taxa de cura em mulheres diagnosticadas com câncer de mama. Para mais detalhes sobre os modelos de sobrevivência com fração de cura, consulte e.g. os seguintes textos e as suas referências: Yakovlev e Tsodikov (1996) com inferências biologicamente significativas a partir de dados de câncer, Ibrahim, Chen e Sinha (2001) com esses modelos sob uma perspetiva bayesiana, e Amico e Keilegom (2018) que fazem uma revisão sobre modelos de cura.

Neste trabalho o objetivo principal é a análise de sobrevivência tendo em consideração a proporção de indivíduos não suscetíveis ao evento de interesse. Esta análise foi motivada por dados reais de pacientes com câncer da mama no Estado do Paraná, por meio de um estudo transversal e observacional, com entrada dos pacientes de 2004 a 2006, e fim do estudo em 2016. O capítulo está organizado da seguinte forma: a Seção 3.2 apresenta a metodologia utilizada na análise de sobrevivência com fração de cura; a Seção 3.3 expõe uma análise descritiva dos dados de pacientes câncer da mama e mostra os resultados inferenciais dos modelos de cura ajustados utilizando uma abordagem bayesiana via aproximações de Laplace encaixadas e integradas (INLA). Por fim, apresentam-se algumas conclusões sobre a análise de sobrevivência proposta.

3.2 Métodos

3.2.1 Material

Os dados utilizados neste estudo foram extraídos do INCA (INCA, 2020), referindo-se a pacientes diagnosticados com câncer de mama tratados num hospital de referência no Estado do Paraná no período de 2004 a 2006, sendo estes observados até 2016. Os dados originais possuem 2234 pacientes para os quais foram observadas 19 variáveis. Durante a análise exploratória dos dados, algumas dessas observações foram excluídas devido a informações duplicadas e/ou tinham tempo de sobrevivência igual a 0 dias. Assim, o conjunto de dados em estudo passou a ter 2103 pacientes.

O INCA é o órgão auxiliar do Ministério da Saúde no desenvolvimento e coordenação das ações integradas para a prevenção e o controle do câncer no Brasil. Tais ações compreendem a assistência médico-hospitalar, prestada direta e gratuitamente aos pacientes com câncer como parte dos serviços oferecidos pelo Sistema Único de Saúde, e a atuação em áreas estratégicas, como prevenção e detecção precoce, formação de profissionais especializados, desenvolvimento da pesquisa e geração de informação epidemiológica (INCA, 2020).

3.2.2 Modelos com fração de cura

A análise de sobrevivência lida com tempos de sobrevivência de indivíduos observados durante um determinado período, por exemplo, o tempo desde o diagnóstico de câncer da mama até a ocorrência do evento morte (falha) em cada indivíduo. Uma das principais características dos dados de sobrevivência é a presença de censura i.e. o evento de interesse não ocorre em alguns indivíduos em estudo, situação frequente em estudos de coorte e ensaios clínicos de câncer (vide Yakovlev e Tsodikov (1996)).

Um conjunto de dados de sobrevivência com n pacientes é usualmente representado pelo par (t_i, δ_i), onde t_i é o tempo de falha ou censura do paciente i e δ_i a sua variável indicadora de falha, isto é, $\delta_i = 1$, se houve o paciente i teve o evento de interesse, ou $\delta_i = 0$, no caso contrário, $i = 1, \ldots, n$. Frequentemente, adota-se um modelo paramétrico para os tempos de sobrevivência (T) que são caracterizados pela sua função de sobrevivência, $S(t)$ ou ou pela função de risco $h(t)$, $t \geq 0$. A primeira é a probabilidade de o paciente sobreviver pelo menos até ao instante t, enquanto a segunda é a taxa (instantânea) de ocorrência de morte no instante t, estando relacionadas em

$$S(t) \equiv P(T \geq t) = \frac{f(t)}{h(t)}, \tag{3.2.1}$$

onde $f(t)$ é a função densidade de probabilidade (f.d.p.) dos tempos de sobrevivência que são aqui considerados variáveis aleatórias absolutamente contínuas. Para mais detalhes sobre análise de sobrevivência, consulte e.g. Lawless (2003), Lee e Wang (2003) e Colosimo e Giolo (2006).

Os dados de sobrevivência têm sido alvo de muitas análises paramétricas, não paramétricas e semi-paramétricas devido às suas especificidades. Por exemplo, a observação de indivíduos que não são sucetíveis à ocorrência do evento de interesse durante o período de estudo. Neste caso, os modelos de sobrevivência resultantes são conhecidos por modelos com sobreviventes a longo-prazo ou modelos com fração de cura. Alguns trabalhos envolvendo modelos de cura têm sido publicados nos últimos anos, incluindo uma revisão desses modelos sob uma abordagem bayesiana no Capítulo 5 de Ibrahim, Chen e Sinha (2001).

Modelos de sobrevivência com fração de cura foram inicialmente propostos por Boag (1949), e mais tarde a metodologia de proporção de curados foi introduzida por Berkson e Gage (1952). Estes utilizaram a função de sobrevivência em forma de mistura, onde a população é dividida em duas partes: uma que representa o tempo de sobrevivência para os indivíduos suscetíveis (não curados) e a outra envolvendo uma distribuição para o tempo de sobrevivência dos indivíduos imunes (curados).

De acordo com Ibrahim, Chen e Sinha (2001), o modelo de cura tem sido usado para modelar dados de sobrevivência para vários tipos de câncer, incluindo câncer de mama, leucemia, câncer de próstata, câncer de cabeça e pescoço. Estas doenças possuem uma proporção significativa dos pacientes denominado “curados”. Nesse modelo, considera-se uma certa fração p da população curada, e o restante $1-p$ para a população não curada. Por conseguinte, a função de sobrevivência para toda a população é dada por

$$S(t) = p + (1-p)\, S_0(t) = \frac{(1-p)\, f_0(t)}{h(t)} \tag{3.2.2}$$

onde p representa a proporção de indivíduos que não estão sujeitos ao evento de interesse, $S_0(t)$ e $f_0(t)$ são, respectivamente, a função de sobrevivência (própria) e a f.d.p. dos indivíduos que estão sob risco de falha, ao passo que $h(t)$ é a função de risco de morte para toda a população, $t > 0$, $0 < p < 1$. Se t tende ao infinito, então (3.2.2) tende a p, possuindo a função de sobrevivência do modelo com fração de cura (padrão) uma assintota em p.

Consequentemente, se os tempos de sobrevivência dos n pacientes com câncer da mama têm uma distribuição de probabilidade indexada por um vetor paramétrico $\boldsymbol{\theta}$, a função de verossimilhança de $\boldsymbol{\theta}$, com base no conjunto de dados dos pacientes $\mathcal{D} =$

$\{(t_i, \delta_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$, é expressa como

$$L(\boldsymbol{\theta}|\mathcal{D}) = \prod_{i=1}^{n} [(1-p) f_0(t_i|\boldsymbol{\theta})]^{\delta_i} [p + (1-p) S_0(t_i|\boldsymbol{\theta})]^{1-\delta_i}, \tag{3.2.3}$$

onde $\mathbf{x}_i$ é o vetor de covariáveis observadas para a $i - ésima$ observação. Por exemplo, adotando o modelo paramétrico Weibull $\mathcal{W}(\gamma, \lambda)$ para modelar o tempo dos indivíduos suscetíveis com parâmetro de forma γ e parâmetro de escala λ, a função de sobrevivência e a f.d.p. são iguais a $S_0(t|\gamma, \lambda) = \exp(-\lambda t^{\gamma})$ e $f_0(t|\gamma, \lambda) = \gamma \lambda t^{\gamma - 1} \exp(-\lambda t^{\gamma})$, respectivamente. Note-se que as covariáveis $\mathbf{x}$ são introduzidas no modelo Weibull através do parâmetro $\lambda \equiv \exp(\mathbf{x}^T \boldsymbol{\beta})$, sendo $\boldsymbol{\beta}$ o vetor de coeficientes de regressão associado ao vetor de covariáveis $\mathbf{x}$ do modelo.

Sob a perspectiva bayesiana, a inferência sobre os parâmetros de interesse no modelo de sobrevivência Weibull com fração de cura, será feita com base na função de verossimilhança (3.2.3) adicionando informação *a priori* sobre os parâmetros do modelo através de distribuições para os mesmos. Assim, a análise bayesiana permite inferir sobre os parâmetros (desconhecidos) do modelo avaliando a incerteza sobre eles através da sua distribuição *a posteriori*. De um modo mais simplificado, Paulino et al. (2018) expressa a distribuição *a posteriori* como

$$\text{distribuição} \textit{a posteriori} \propto \text{verossimilhança} \times \text{distribuição} \textit{a priori}. \tag{3.2.4}$$

Essa escolha deve-se ao fato de o modelo de cura de mistura proposto ser mais facilmente implementado sob essa abordagem inferencial.

Mesmo considerando distribuições bem conhecidas para os tempos de sobrevivência e os parâmetros do modelo, trabalhar com a distribuição *a posteriori* (3.2.4) pode não ser simples, depedendo fundamentalmente das distribuições *a priori* consideradas. Neste sentido, podem-se usar os métodos Monte Carlo via cadeias de Markov (MCMC), amplamente usados na análise bayesiana, ou aproximações de Laplace encaixadas e integradas (INLA) (veja-se um resumo desses métodos de computação bayesiana em Paulino et al. (2018) ou Amaral-Turkman, Paulino e Müller (2019)). Por conveniência, a obtenção de resultados inferenciais dos parâmetros de interesse foi feita via *software* R-INLA (RUE; MARTINO; CHOPIN, 2009) pelo seu melhor desempenho na implementação de modelos de sobrevivência Weibull com fração de cura. Além disso, o software R (R Core Team, 2018) foi também utilizado para a realização de outros cálculos e gráficos apresentados no trabalho.

Por fim, uma parte importante em qualquer análise estatística, incluindo na análise de sobrevivência, é a avaliação, seleção e comparação de modelos. Para comparar modelos é necessário definir critérios que permitam a seleção de um modelo parcimonioso. Ou seja,

um modelo capaz de explicar de forma satisfatória e simples os tempos de sobrevivência (variáveis resposta) com relação às variáveis explicativas (covariáveis). No contexto bayesiano várias medidas de seleção já foram propostas tais como medidas de desempenho preditivo ou resíduos de eliminação bayesianos preditivos. As medidas de desempenho preditivo aqui usadas foram os critérios de seleção DIC (*Deviance Information Criterion*) e WAIC (*Watanabe-Akaike Information Criterion*), propostos por Spiegelhalter et al. (2002) e Watanabe (2010), respectivamente. Valores pequenos dessas medidas indicam o melhor ajustamento do modelo. Para alguns detalhes de medidas de diagnóstico e adequação dos modelos bayesianos, veja-se e.g. Paulino et al. (2018) e Wang, Ryan e Faraway (2018)).

3.3 Resultados e discussão

3.3.1 Análise preliminar

O conjunto de dados aqui em estudo refere-se a pacientes diagnosticados com câncer de mama tratados num hospital de referência no Estado do Paraná. Numa fase preliminar, foram excluídas as covariáveis “Estado do endereço”, “Cidade endereço”, “Descrição topografia”, “Descrição doença” e “Código da doença” pela presença demasiada de valores omissos ou mesmo por não conterem informação relevante para o estudo. No decorrer do estudo foram observados 11 indivíduos do sexo masculino, que também foram excluídos do estudo visto que o objetivo aqui é analisar os indivíduos do sexo feminino com câncer de mama. Assim, o banco de dados passou a possuir 2092 observações e 14 variáveis.

A Tabela 1 mostra as estatísticas descritivas das covariáveis categorizadas em consideração, bem como a definição das suas categorias descrevendo assim as caraterísticas das pacientes com câncer da mama. Pode-se observar que a maioria das pacientes estão acima dos 50 anos e 73.42% delas são de raça branca. Quanto ao estado civil, tem-se que 46.27% são pacientes casadas.

Uma covariável relacionada com a doença é a Topografia, representada pelos códigos CID responsável pela classificação da identificação do tipo de câncer, que permite obter mais informações sobre as patologias catalogadas pela Organização Mundial de Saúde (OMS). Os CID são denominados pelos seguintes códigos e nomes: C500 - Neoplasia maligna do mamilo e aréola; C501 - Neoplasia maligna da porção central da mama; C502 - Neoplasia maligna do quadrante superior interno da mama; C503 - Neoplasia maligna do quadrante inferior interno da mama; C504 - Neoplasia maligna do quadrante superior externo da mama; C505 - Neoplasia maligna do quadrante inferior externo da mama; C506 - Neoplasia maligna da porção axilar da mama; C508 - Neoplasia maligna da mama com lesão invasiva; C509 - Neoplasia maligna da mama, não especificada. Devido à pouca

Tabela 1 – Características das pacientes com câncer da mama no Estado do Paraná no período de 2004 a 2016.

Variável	Descrição	Total	%	Censura(%)	Falha(%)
Idade	< 40	226	10.80	7.84	2.96
	$40 - 50$	551	26.34	21.46	4.88
	> 50	1315	62.86	45.46	17.40
Raça	Branco	1536	73.42	51.00	22.42
	Outra	556	26.58	23.76	2.82
Estado civil	Casado	968	46.27	36.42	9.85
	Outro	1124	53.73	38.34	15.39
Topografia	C509	1209	57.79	42.45	15.34
(CID)	C504	493	23.57	18.26	5.31
	Outros	390	18.64	14.05	4.59
Morfologia	Carcinoma Ductal	1372	65.58	48.61	16.97
	Carcinoma Lobular	206	9.85	7.41	2.44
	Outros	514	24.57	18.74	5.83
Status vital	Vivo	1468		70.17	0.00
	Morto	624		4.59	25.24
	Total	2.092			

frequência de algumas categorias, optou-se pelo agrupamento dos CID C500, C501, C502, C503, C505, C506 e C508 (categoria outras).

Observa-se ainda na Tabela 1 que 57.79% das pacientes foram diagnosticada com neoplasia maligna da mama, não especificada (C509), ao passo que 65.58% das pacientes em estudo foram diagnosticada com Carcinoma Ductal, corroborando com a literatura da área de que este tipo de câncer é mais comum em pacientes com câncer de mama. Um dos resultados que se tornou importante na decisão da metodologia para a análise dos dados em estudo foi perceber que 70.17% das pacientes estavam vivas no fim do estudo, considerado um percentual alto de sobreviventes. A censura foi aqui definida para as mulheres que não obtiveram óbitos por câncer de mama.

A Figura 3.3.1 apresenta as curvas de Kaplan-Meier para os tempos de sobrevivência das pacientes com câncer da mama, incluindo os seus intervalos de confiança a 95%, para todas as pacientes (Figura 3.3.1-A) e por cada covariável categorizada em estudo. Graficamente, há uma estabilização da estimativa Kaplan-Meier da função de sobrevivência de todas as pacientes. As mulheres com neoplasia maligna do quadrante superior externo da mama (C504) tendem a sobreviver mais que as mulheres com outros CID (Figura 3.3.1-B), e note-se que há um cruzamento entre as curvas das categorias C509 e outros. As mulheres que se declaram Branca têm menor sobrevivência do câncer de mama do que as mulheres não brancas (Figura 3.3.1-C). Não se vislumbra diferenças entre as curvas de sobrevivência das pacientes que possuem diferentes categorias de morfologia (Figura 3.3.1-D).

As pacientes casadas sobrevivem mais do que as que não são (Figura 3.3.1-E), enquanto as pacientes com idade entre 40 e 50 anos possuem maiores chances de sobreviver ao câncer da mama do que as mulheres com menos de 40 anos e acima dos 50 anos (Figura 3.3.1-F).

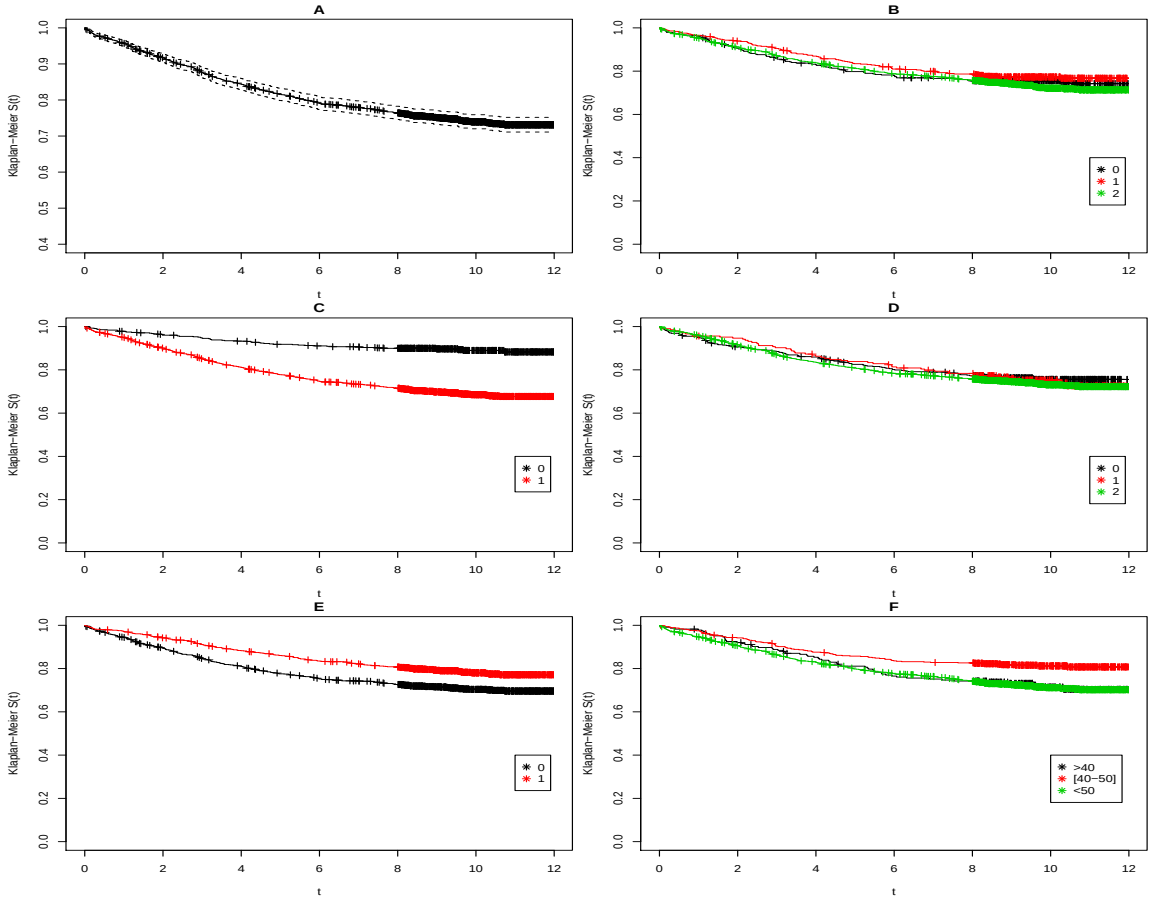

Figura 3.3.1 – Curvas de sobrevivência de Kaplan-Meier: (A) Todos as pacientes, com intervalos de confiança de 95%, (B) CID (2: C509, 1: C504, 0: Outros), (C) Raça (1: Branca, 0: Outras), (D) Morfologia (2: Carcinoma Ductal, 1: Carcinoma Lobular, 0: Outros), (E) Estado civil (1: Casada, 0: Outros) e (F) Faixa etária de mulheres diagnosticadas com câncer de mama no Estado do Paraná de 2004 a 2016.

Como pode ser visto na Figura 3.3.1, as funções de sobrevivência estimadas pelo estimador de Kaplan-Meier não tendem para zero à medida que se avança no tempo, indicando que os dados de pacientes com câncer da mama pode possuir uma fração de cura, ou seja, há suspeitas de uma proporção de pacientes não suscetíveis ao evento de interesse e consequentemente o modelo de sobrevivência com fração de cura (3.2.2) pode ser mais adequado.

Para concluir essa análise preliminar, algumas distribuições foram ajustadas aos tempos de sobrevivência das pacientes com câncer da mama, cujos resultados mostraram

evidências empíricas de que a distribuição Weibull foi a que melhor se adequou aos dados.

3.3.2 Análise com fração de cura

O modelo de sobrevivência Weibull com fração de cura (3.2.2) foi aqui utilizado na análise bayesiana dos dados das pacientes com câncer da mama do Estado do Paraná de 2004 a 2016. As covariáveis tidas em consideração no ajustamento do modelo proposto foram Faixa Etária (F), Morfologia (M), CID (Código da Topografia), Raça (R) e Estado Civil (EC). A inserção das covariáveis é feita através de uma fração de ligação log(.) no parâmetro de escala da distribuição Weibull, da seguinte forma:

$$\lambda \equiv \exp(\eta) = \exp\left[\beta_0 + \sum_{j=1}^{2}\left(\beta_j F_j + \beta_{j+2} M_j + \beta_{j+4} CID_j\right) + \beta_7 EC + \beta_8 R\right], \qquad (3.3.1)$$

onde as variáveis mudas/fictícias (*dummy variables*) são definidas por $F_1 = 1$ ($40 - 50$ anos), $F_2 = 1$ (> 50 anos), $M_1 = 1$ (Carcinoma Lobular), $M_2 = 1$ (Carcinoma Ductal), $CID_1 = 1$ (C504) e $CID_2 = 1$ (C509).

Na definição da distribuição *a posteriori* (3.2.4), a informação *a priori* considerada foi vaga no sentido que não obtêm elementos passados úteis para a caraterização prévia dos parâmetros do modelo com fração de cura (3.2.2). Enfim, as distribuições *a priori* não-informativas, vagas mas próprias, foram Normal $\mathcal{N}(0,1)$ para todos os parâmetros de regressão β_j, $j = 0, 1, \ldots, 8$ e Gama $\mathcal{G}(1,1)$ para o parâmetro de forma γ da Weibull. Por conveniência de implementação computacional dos modelos, foram consideradas transformações para o parâmetro γ e a probabilidade de cura p i.e. $\theta_1 = \log\gamma$ e $\theta_2 = \log\left[\frac{p}{(1-p)}\right]$.

Para avaliar os modelos usaram-se as medidas DIC e WAIC. Valores de DIC e WAIC dos modelos de sobrevivência dos dados de câncer da mama, com e sem fração de cura e covariáveis selecionadas estão na Tabela 6. Esses valores indicam que de fato o conjunto das covariáveis mostra-se necessário na análise das pacientes com câncer da mama, bem como que os modelos de cura são melhores do que os modelos sem fração de cura. Isso será discutido mais à frente. Dentre estes o modelo selecionado é proposto, i.e., o modelo de sobrevivência Weibull com fração de cura.

As inferências sobre os parâmetros quer do modelo de cura com covariáveis (modelo selecionado) quer dos demais modelos de sobrevivência foram também obtidas através do *software* R-INLA (RUE; MARTINO; CHOPIN, 2009) usando as famílias de distribuições Weibull e Weibull-cure. Alguns resultados inferenciais do modelo selecionado encontram-se na Tabela 3 tais como média, desvio padrão e intervalo de credibilidade HPD (*High Posterior Density*) a 95%. Nota-se que há evidência de influência das covariáveis Faixa

Tabela 2 – Valores de DIC e WAIC dos modelos de sobrevivência dos dados de câncer da mama, com e sem fração de cura e covariáveis selecionadas.

Modelo		DIC	WAIC
Fração de cura	Covariáveis		
Sem	Sem	4667.739	4667.574
	Com	4532.920	4533.068
Com	Sem	4646.372	4646.699
	Com	4521.263	4522.511

Etária, Estado Civil e Raça nos tempos de sobrevivência das pacientes com câncer da mama.

Tabela 3 – Quantidades *a posteriori:* média, desvio padrão (DP) e intervalos de credibilidade (IC) HPD a 95% dos parâmetros dos modelos de sobrevivência com fração de cura.

Parâmetro	Média	DP	IC HPD 95%
(Intercept)	-8.525	0.405	(-9.322, -7.739)
β_1 ($40-50$ anos)	-0.604	0.231	(-1.050, -0.145)
β_2 (>50 anos)	-0.157	0.198	(-0.537, 0.239)
β_3 (Carcinoma Lobular)	-0.041	0.222	(-0.481, 0.389)
β_4 (Carcinoma Ductal)	0.081	0.142	(-0.196, 0.360))
β_5 (C504)	-0.131	0.202	(-0.528, 0.266)
β_6 (C509)	0.124	0.171	(-0.212, 0.460)
β_7 (Casada)	-0.722	0.130	(-0.977, -0.468)
β_5 (Branca)	1.650	0.167	(1.329, 1.983)
γ	0.932	0.045	(0.844, 1.020)
p	0.576	0.028	(0.519, 0.629)

A Tabela 3 mostra ainda que a log-chance de curados (θ_2) na população das mulheres com câncer da mama foi estimada em aproximadamente 58%, quando se tem em conta as covariáveis Faixa Etária, Raça, Estado Civil, Código de topografia e a Morfologia da doença. No modelo selecionado pode-se observar que a Raça branca tem um efeito positivo na variável resposta i.e. as mulheres de raça branca com câncer de mama tem um maior risco de morte da doença do que as mulheres de outras raças. As pacientes diagnosticadas com câncer da mama entre 40 e 50 anos e as pacientes casadas têm um risco de morte menor do que as mulheres que estão com menos de 40 anos e acima de 50 anos e não são casadas, respectivamente. Outro fator importante é que as pacientes com CID C504 e C509 não são significativas no modelo selecionado, porém as pacientes com C504 têm um efeito “residual” negativo na variável resposta, contrariamente às pacientes com

C509. Isso foi confirmado em contacto informal com médico oncologista afirmando que as CID não indicam diferença da localização do câncer de mama.

A Figura 3.3.2 apresenta os gráficos da função de sobrevivência das pacientes com câncer da mama, estimada pelos modelos Weibull com (1) e sem (2) fração de cura. Tendo sido concluído que o primeiro modelo é melhor do que o segundo, pode-se induzir que o modelo sem fração de cura está a sobre-estimar essa função, bem como a curva de Kaplan-Meier na Figura 3.3.1-A, e que a fração de cura está num pantamar abaixo do que se pensava somente com o modelo Weibull.

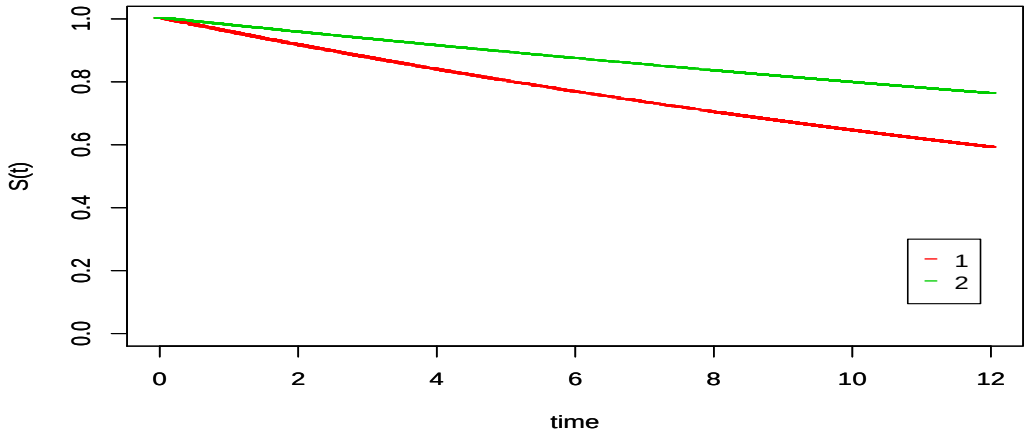

Figura 3.3.2 – Gráficos da função de sobrevivência das pacientes com câncer da mama, estimada pelos modelos Weibull com (1) e sem (2) fração de cura.

Na Figura 3.3.3 é apresentado a análise dos resíduos do modelo selecionado sob uma perspetiva bayesiana (vide definições e detalhes em Wang, Ryan e Faraway (2018)). Os gráficos de resíduos bayesianos de Cox-Snell, de desvio e de Martingale sugerem que o ajustamento do modelo de sobrevivência Weibull com fração de cura não está essencialmente mal, uma vez que os resíduos de Cox-Snell estão em torno da reta de 45°, exceto na cauda, onde a variabilidade da estimativa de risco cumulativo de morte com câncer da mama é menor. Isso deve estar relacionado com a presença de dois grupos de valores de resíduos nos gráficos de resíduos de Martingale e de desvio. Todavia, essa questão deve ser pesquisada futuramente pois a própria definição de resíduos deve variar entre as pacientes “curadas” e “não curadas”.

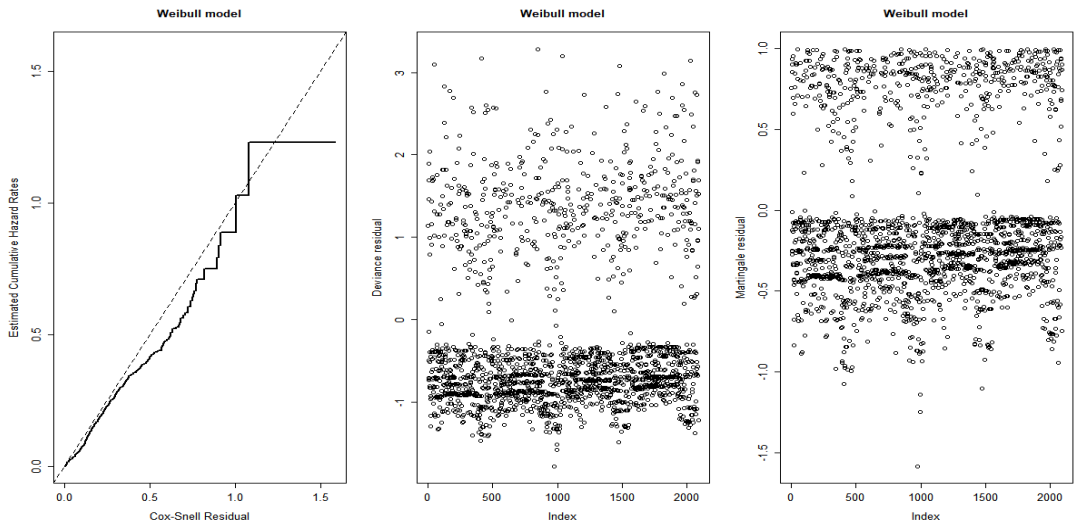

Figura 3.3.3 – Gráficos para diagnóstico do ajuste do modelo considerando a distribuição Weibull.

3.4 Análise Final

A análise de sobrevivência proposta para os dados de pacientes com câncer da mama no Estado do Paraná de 2004 a 2016 mostrou-se ser mais adequada por incluir uma fração de cura no modelo de sobrevivência Weibull, captando assim uma proporção de pacientes sobreviventes a longo prazo na população em estudo, que foi estimada pelo modelo em torno de 58%.

Com base nos resultados obtidos, as covariáveis Raça, Faixa Etária e Estado Civil tiveram efeitos significativos no risco de morte por câncer da mama. As mulheres de raça branca com câncer de mama tem um maior risco de morte da doença do que as mulheres de outras raças, enquanto que as pacientes diagnosticadas com câncer da mama entre 40 e 50 anos e as pacientes casadas têm um risco de morte menor do que as pacientes de outras faixas etárias e não casadas. Em estudos com base em registros oncológicos, deve-se encarar esses resultados com algum cuidado visto que há necessidade de confirmar a qualidade de recolha e tratamento dos dados que nem sempre é possível fazer por parte dos investigadores que analisam os dados já prontos (nosso caso).

Apesar de os métodos MCMC serem amplamente usados na análise bayesiana, foi também mais conveniente utilizar aproximações de Laplace encaixadas e integradas (INLA) na obtenção de resultados inferenciais, uma vez que o INLA já tem implementado a distribuição Weibull-cure. Isso foi crucial para se ter uma maior rapidez na execução dos programas computacionais (aqui executados no R-INLA). Todavia, faltou aqui desenvolver os modelos de cura para a predição de novos tempos de sobrevivência, como também fazer uma

análise de sensibilidade os valores fixados para os hiperparâmetros i.e. os parâmetros da distribuições *a priori*.

Os trabalhos recentes envolvendo metodologia mais inovadora no tratamento de dados de pacientes com câncer e/ou usando modelos de cura são estímulos para trabalhos futuros dos modelos de sobrevivência com fração de cura para dados de câncer da mama. Por exemplo, Wei e Wu (2019) propuseram um modelo de cura com riscos proporcionais por partes para incorporar o efeito do tratamento retardado e a fração de cura nos ensaios clínicos de imunoterapia contra o câncer. Seppä et al. (2019) usam modelos de excesso de mortalidade com efeitos aleatórios para estimar a variação na sobrevivência relativa ou *net survival* de pacientes com câncer. Este estudo avaliou o desempenho do INLA no monitoramento da variação regional para análises de dados de registro de câncer.

Capítulo 4

Análise conjunta bayesiana de contagens longitudinais de CD4 e sobrevivência a longo prazo de pacientes com HIV no Paraná, Brasil

Resumo

Recentemente, a análise de dados envolvendo pacientes com HIV/AIDS tem recebido atenção crescente com a modelagem conjunta de medidas repetidas de biomarcadores e os tempos de sobrevivência. Essa análise conjunta tem mostrado ser mais adequada por ter em conta a dependência entre os dois tipos de respostas longitudinal e de sobrevivência, que são frequentemente analisadas separadamente. Os modelos conjuntos de sobrevivência e longitudinal são aqui aplicados a um estudo de coorte de pacientes do estado do Paraná, Brasil, com HIV/AIDS durante os anos de 2002 a 2006. Neste caso, os tempos de sobrevivência de pacientes com HIV/AIDS são modelados conjuntamente com as medidas repetidas de contagens de linfócitos CD4, tendo em conta quer uma abordagem bayesiana quer uma potencial fração de cura dos pacientes. Isso deve-se ao fato de essas duas componentes do modelo conjunto partilharem efeitos aleatórios (modelos mistos) que são mais facilmente implementados via modelos gaussianos latentes. Todavia, apesar de os métodos Monte Carlo via cadeias de Markov (MCMC) serem amplamente usados na análise bayesiana, foi também mais conveniente utilizar aproximações de Laplace encaixadas e integradas (INLA) na obtenção de resultados inferenciais com base em efeitos aleatórios gaussianos. Os resultados da análise dos dados de HIV/AIDS no Paraná indicam que os modelos conjuntos apresentaram melhor desempenho comparativamente aos modelos mais tradicionais de sobrevivência, i.e., análise separada das contagens CD4 longitudinais e dos tempos de sobrevivência, incluindo uma fração de cura.

Palavras-chave: Análise de Sobrevivência. Fração de cura. Medidas Repetidas.Estatística Bayesiana. INLA. Modelos Conjuntos Mistos.

4.1 Introdução

O vírus HIV (Vírus da Imunodeficiência Humana), causador da doença contagiosa AIDS (Síndrome da Imunodeficiência Adquirida), ataca o sistema imunológico, que é o responsável por defender o nosso organismo de doenças. As células mais atingidas são os linfócitos CD4 que, quando atacados, alteram o seu DNA transformando-se em HIV. Indivíduos portadores do HIV (soropositivos) nem sempre tem AIDS, e muitos soropositivos vivem anos sem apresentar sintomas ou desenvolver a doença. Mas, eles podem transmitir o vírus a outros indivíduos pelas relações sexuais, sanguínea e até mesmo pelo leite materno (SSEP, 2019). Mangal et al. (2019) afirmam que o tratamento precoce pode contribuir para

a melhora da sobrevivência em pessoas vivendo com HIV em terapia anti-retroviral, com os maiores benefícios observados em mulheres e grupos etários mais jovens. Taquette e Souza (2019) reiteram que a qualificação e expansão das estratégias de comunicação sobre sexualidade nas escolas é urgente e essencial na prevenção do HIV e AIDS na adolescência.

Nomeadamente para pacientes com AIDS, há interesse em estudar o tempo de sobrevivência desses indivíduos, i.e., o tempo desde o diagnóstico da doença até a ocorrência de um evento de interesse (e.g. morte) que quando não observado define uma *censura*. Os dados de sobrevivência têm sido alvo de muitas analises devido às suas especificidades. Por exemplo, a observação de indivíduos que não são sucetíveis à ocorrência do evento durante o período de estudo (sobreviventes a longo-prazo ou fração de cura) (IBRAHIM; CHEN; SINHA, 2001). Para além da observação dos tempos de sobrevivência, os indivíduos em estudo frequentemente apresentam também medidas repetidas ao longo do tempo (e.g. medidas de biomarcadores) que em geral são analisadas separadamente.

Neste caso, o interesse da análise de dados poderia ser a análise do tempo até o evento num cenário longitudinal dependente, usando análise conjunta em vez de análise separada das componentes longitudinal e de sobrevivência (FAUCETT; THOMAS, 1996; WULFSOHN; TSIATIS, 1997; SOUSA, 2011; MARTINS; SILVA; ANDREOZZI, 2017). Segundo Rizopoulos (2012), uma das metodologias que consegue agregar as duas informações é a modelagem conjunta através de efeitos aleatórios (modelos mistos). Alguns estudos que utilizam essa metodologia estão relacionados à área da saúde, tais como estudos com dados de aneurisma da aorta abdominal (SWEETING; THOMPSON, 2011), de HIV (SU; HOGAN, 2011) e de câncer da mama (CROWTHER et al., 2016).

Vários modelos de sobrevivência e longitudinais podem ser adotados numa análise conjunta. Wulfsohn e Tsiatis (1997) aplicaram modelos conjuntos em que se pode estimar os parâmetros no modelo de Cox quando a covariável longitudinal é medida com pouca frequência e com erro de medição. Tournoud e Ecochard (2008) utilizaram modelos de fração de cura na modelagem conjunta para avaliar também a probabilidade de um bebê adquirir o HIV antes ou após o nascimento de acordo com a prática alimentar e as principais covariáveis maternas e infantis, bem como as ocasiões em que ocorre a transmissão do HIV de mãe para filho. Huang, Hu e Dagne (2014) exploraram modelos conjuntos bayesianos para os tempos de vida e as medidas longitudinais na tentativa de superar desvios de normalidade da resposta longitudinal, erros de medição e incerteza na especificação de um modelo de sobrevivência paramétrico, juntamente com a possibilidade do paciente sobreviver a longo prazo. Martins, Silva e Andreozzi (2017) também propuseram um modelo conjunto bayesiano, com pacientes a sobreviver a longo prazo e a inclusão de uma

fragilidade espacial na componente de sobrevivência.

Neste trabalho, o objetivo principal é propor uma análise conjunta de dados de sobrevivência de pacientes com HIV/AIDS residentes no Estado do Paraná, Brasil, num determinado período, levando em consideração as medidas ao longo do tempo do biomarcador contagem de células CD4 e a potencial presença de sobreviventes a longo prazo. Perante a complexidade de ajustar tais modelos conjuntos, optou-se pela análise bayesiana (vide e.g. Paulino et al. (PAULINO et al., 2018)) através de uma abordagem analítica por aproximações de Laplace encaixadas e integradas (INLA) (RUE; MARTINO; CHOPIN, 2009).

O restante deste trabalho evolui da seguinte forma: na Seção 4.2, apresentam-se uma descrição dos dados HIV/AIDS em estudo (Subseção 4.2.1) e os aspetos metodológicos dos modelos conjuntos propostos (Subseção 4.2.2), incluindo detalhes da sua abordagem bayesiana implementável no *software* INLA; na Seção 4.3 os modelos conjuntos são ajustados aos dados de HIV/AIDS, cujos resultados fornecem informação para os submodelos tanto longitudinal quanto de sobrevivência com fração de cura; as conclusões da análise conjunta proposta estão na Seção 4.4.

4.2 Métodos

4.2.1 Material

Os dados que motivaram este trabalho reportam-se a pacientes do Estado do Paraná, Brasil, diagnosticados com HIV/AIDS durante os anos de 2002 a 2006. Para cada paciente foi observado o seu tempo de sobrevivência i.e. o tempo (em anos) desde o diagnóstico de HIV/AIDS até a morte (disponível se a morte ocorreu antes de 31 de dezembro de 2006 e censurado no caso contrário) e as suas contagens de células CD4 (célula branca do sangue), medidas durante o acompanhamento do paciente em intervalos de tempo irregulares. Os dados incluem 272 pacientes e algumas covariáveis tais como sexo (1: homem; 0: mulher), idade (1: $\geq$ 50 anos; 0: $<$ 50 anos), infecção oportunista prévia na entrada do estudo (PrevOI=1: sem infecção prévia; PrevOI=0: no caso contrário), raça (1: branca; 0: outras) e estado civil (1: casado; 0: outros), bem como os tempos de recolha da medida CD4 (em dias) e o status (1: morte; 0: censurado).

Os dados de HIV/AIDS foram gerados pelo programa nacional de AIDS, que se utilizou de três principais bases eletrônicas: (i) SINAN-AIDS (sistema de informação de agravos de notificação de casos de AIDS), que é a mais importante base de dados de vigilância da AIDS; (ii) SISCEL (controle de testes laboratoriais), projetada para monitorar testes laboratoriais, tais como contagens de células CD4 e testes de carga viral para HIV/AIDS; (iii) SICLOM (sistema de controle logístico de drogas), desenvolvida para controlar a logís-

tica das entregas de tratamento da AIDS. Estas três bases de dados foram previamente combinadas em um único banco de dados com casos de HIV/AIDS usando um processo chamado vinculação de registro. Para mais detalhes, veja-se Fonseca et al. (2010) e Martins, Silva e Andreozzi (2016).

Entre algumas funções de interesse em análise de sobrevivência, a função de sobrevivência $S(t)$ tem um maior destaque por definir a probabilidade de um individuo sobreviver pelo menos até ao instante t, $t \geq 0$ (veja-se e.g. (LEE; WANG, 2003; LAWLESS, 2003; COLOSIMO; GIOLO, 2006)). A Figura 4.2.1 mostra a estimativa de Kaplan-Meier da função de sobrevivência $S(t)$ dos 272 pacientes com HIV/AIDS no período de 2002 a 2006, com intervalos de confiança de 95% e tempos de sobrevivência (t) em anos. Esta curva parece diminuir muito lentamente ao longo do tempo ocorrendo uma estabilidade na cauda direita da curva. Isso indica uma possível presença de sobreviventes de longo prazo. Assim, parece-nos sensato ajustar um modelo de sobrevivência que permita lidar com uma fração de cura para esses pacientes com HIV/AIDS.

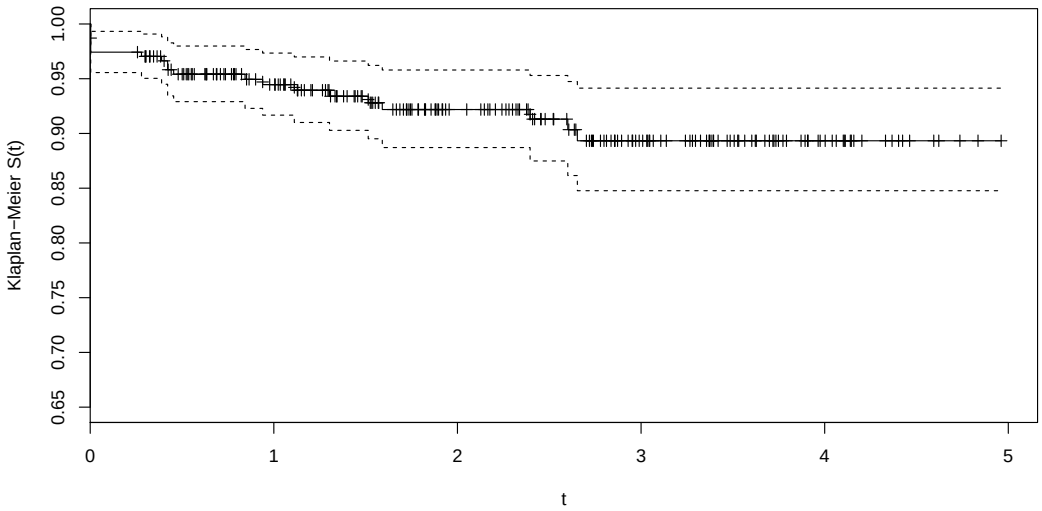

Figura 4.2.1 – Curva de sobrevivência de Kaplan-Meier, com intervalos de confiança de 95%, para os dados de pacientes com HIV/AIDS.

Além disso, mais de 85% das observações são censuradas nos dados de pacientes com HIV/AIDS. Segundo Maller e Zhou (1996) um alto percentual de censuras indica uma possível presença de sobreviventes a longo prazo, ou seja, pacientes com HIV/AIDS não estão suscetíveis ao evento de interesse. Assim, esse percentual de censura elevado é outro indicador da presença de pacientes “curados” no estudo e consequentemente da necessidade de introduzir no modelo estatístico componentes que permitam lidar com uma fração da população que não experimentará o evento de interesse. Já agora, todas as análises estatísticas apresentadas aqui foram feitas no *software* R (R Core Team, 2018).

4.2.2 Modelos conjuntos

Para a análise de dados com observação quer dos tempos de sobrevivência quer das medidas repetidas de uma segunda variável resposta, propõe-se modelos conjuntos mistos que lidam com a associação dos submodelos longitudinal e de sobrevivência através da introdução de efeitos aleatórios compartilhados entre os dois. Isto é, os efeitos aleatórios associados aos indivíduos do submodelo longitudinal entram no submodelo de sobrevivência através de várias formas e.g. uma combinação linear bem definida. De acordo com Guo e Carlin (2004), a associação entre os processos longitudinais e de sobrevivência pode surgir de duas maneiras: (i) através de variáveis explicativas comuns; (ii) por dependência estocástica entre variáveis comuns não observadas (efeitos aleatórios). Quando há associação entre os dois processos, deve-se diminuir as inferências tendenciosas e menos eficientes usando um modelo conjunto. As possíveis associações (lineares) entre os submodelos longitudinal e de sobrevivência utilizadas neste trabalho foram também usadas em Henderson, Diggle e Dobson (2000), Guo e Carlin (2004), Martins, Silva e Andreozzi (2016).

Os modelos conjuntos mistos têm recebido atenção crescente nos últimos anos, com extensões variadas, e.g., diferentes funções de risco e estruturas de associação, trajetórias longitudinais lineares e não lineares (veja-se, Ibrahim el al. (IBRAHIM; CHEN; SINHA, 2001), Rizopoulos (2012)). Neste sentido, Niekerk, Bakka e Rue (2019) mostram que esses modelos conjuntos podem ser implementados do ponto de vista do modelo gaussiano latente usando o *software* R-INLA (RUE; MARTINO; CHOPIN, 2009).

4.2.2.1 Submodelo longitudinal

Segundo Diggle et al. (2002), um estudo longitudinal tem como característica indivíduos para os quais são medidas uma ou mais variáveis repetidamente ao longo do tempo, de forma prospectiva ou retrospectiva. As medidas repetidas podem ser observadas em estudo de coorte possibilitando distinguir mudanças ao longo do tempo em cada individuo. Dados longitudinais induzem uma estrutura de covariância e, por meio de modelos lineares mistos, pode-se modelar tal peculiaridade com a inclusão de efeitos fixos e aleatórios no modelo (VERBEKE; MOLENBERGHS, 2009; WEST; WELCH; GALECKI, 2014).

Para a análise longitudinal dos dados de pacientes com HIV/AIDS foi observado a falta de simetria na variável resposta contagem de células $CD4$. O que põe em causa a suposição de distribuição Normal que é o modelo probabilístico usal em análise de medidas repetidas. Como meio de contornar essa limitação, realizou-se uma transformação na contagem $CD4$, resultando na variável resposta transformada $Y = \sqrt{CD4}$, afim de "normalizar" esta variáveis dos dados de HIV/AIDS (Figura 4.2.2). Para confirmar esta escolha foi

utilizado uma transformação Box-Cox, que consiste em encontrar um valor λ tal que a variável transformada Y^{λ} se aproximem de uma distribuição Normal, conforme algum critério de minimização da soma de quadrados residual (vide Box e Cox (BOX; COX, 1964)).

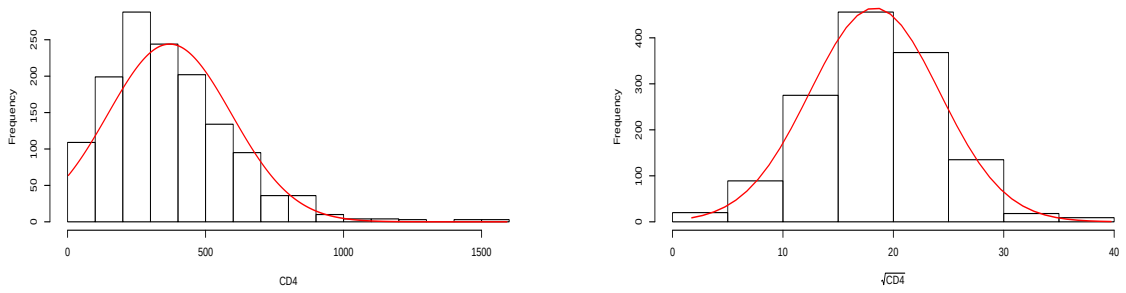

Figura 4.2.2 – Histogramas da contagem CD4 antes (esquerda) e após (direita) a transformação Box-Cox.

Tendo em conta o conjunto de dados HIV/AIDS com n indivíduos, a variável resposta longitudinal, i.e. a medida repetida Y_{ij} do indivíduo i observada no instante j, é suposta ter um modelo linear misto, definido por

$$\mathbf{Y}_i = \mathbf{x}_i^T \beta + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{4.2.1}$$

onde $\mathbf{Y}_i = \{Y_{ij}\}$, Y_{ij} denota a raiz quadrada do valor de contagem CD4 do indivíduo i no instante j, $\mathbf{x}_i^T$ o vetor de efeitos fixos do i-ésimo indivíduo, β o vetor de parâmetros de efeitos fixos, $\mathbf{Z}_i$ a matriz associada aos efeitos aleatórios do indivíduo i, $\mathbf{b}_i$ o vetor de efeitos aleatórios que descrevem o desvio do perfil longitudinal do indivíduo i relativamente á média populacional, e $\boldsymbol{\epsilon}_i$ é o vetor de erros aleatórios associados ao indivíduo i que tem n_i instantes de contagem CD4, $j = 1, \ldots, n_i$, $i = 1, \ldots, n$.

As suposições distribucionais do submodelo longitudinal (4.2.1) recaem sobre os efeitos aleatórios e os erros aleatórios gaussianamente distribuídos, ou seja,

$$\mathbf{b}_i \sim \mathcal{N}(0, \mathbf{D}), \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \mathbf{R}_i), \tag{4.2.2}$$

onde $\mathbf{D}$ é a matriz de covariâncias dos efeitos aleatórios e $\mathbf{R}_i$ é a matriz de covariâncias dos erros aleatórios $\boldsymbol{\epsilon}_i$ do i-ésimo indivíduo, cujos elementos da diagonal medem a variabilidade das medidas repetidas em cada indivíduo. Consequentemente, a construção da função de verossimilhança do submodelo longitudinal é feita simplesmente pelo produto das densidades gaussianas quer dos Y_{ij} condicionalmente aos efeitos aleatórios quer dos próprios efeitos aleatórios $\mathbf{b}_i$.

4.2.2.2 Submodelo de sobrevivência com fração de cura

Na construção do modelo conjunto proposto, o submodelo de sobrevivência pode ter diferentes abordagens tais como modelos semi-paramétricos (e.g. regressão Cox) ou modelos paramétricos (e.g. regressão Weibull). Por conveniência, optou-se pelo segundo que de acordo com alguns autores pode ser mais eficiente (vide e.g. Khan e Ababneh (KHAN; ABABNEH, 2016)).

Se os tempos de sobrevivência seguem uma distribuição Weibull ou seja $\mathcal{W}(\rho, \eta(t))$ com função de risco $h(t|\rho, \eta(t)) = \rho t^{\rho-1}$, então as covariáveis podem ser introduzidas no modelo através do seu parâmetro de escala i.e. $\eta(t) = \exp(\mathbf{x_i}^T \boldsymbol{\beta})$. De forma a permitir que o risco de morte seja uma função do processo gaussiano latente $\mathbf{b}(t)$, a função de risco associada ao i-ésimo indivíduo no instante t é aqui definida por

$$h_i(t) = h_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}_2 + \boldsymbol{\gamma}^T g(\mathbf{b}_i(t))) \tag{4.2.3}$$

onde $h_0(t) = \rho t^{\rho-1}$ é a função de risco de base, $\boldsymbol{\beta}_2$ é um vetor de efeitos fixos das covariáveis $\mathbf{x}_i^T$ que podem coincidir com as covariáveis utilizadas no submodelo longitudinal, $g(\cdot)$ é uma função de ligação que especifica quais componentes aleatórias do processo longitudinal estão relacionadas à função de risco $h_i(\cdot)$ e $\boldsymbol{\gamma}$ é o vetor de parâmetros que liga uma função da trajetória longitudinal (4.2.1) à função de risco (4.2.3). Para mais detalhes, consulte Martins, Silva e Andreozzi (2017).

No modelo conjunto com componentes longitudinal (4.2.1) e de sobrevivência (4.2.3), inclui-se uma fração de cura no submodelo de sobrevivência para ter em conta potenciais sobreviventes a longo prazo, como em Maller e Zhou (1996), Ibrahim, Chen e Sinha (2001), Lambert et al. (2006), entre outros. O modelo com fração de cura comum supõe que a população em estudo é uma mistura de indivíduos que estão suscetíveis a experimentar o evento de interesse e indivíduos que não estão suscetível a essa experimentação. Desta forma, a função de sobrevivência $S(t)$ para a população em estudo é dada por

$$S(t) = p + (1-p)S_0(t) = 1 - (1-p)F_0(t), \tag{4.2.4}$$

onde p representa a proporção de indivíduos curados e $S_0(t)$ e $F_0(t)$ são as funções de sobrevivência e de distribuição dos indivíduos não curados, respectivamente. Note-se que $\lim_{t\to\infty} S(t) = p$ é conhecido por fração de cura (ou fração de indivíduos não suscetíveis ao evento de interesse).

Consequentemente, dada uma amostra aleatória $\{(t_i, \delta_i), i = 1, \cdots, n\}$, onde t_i é o tempo de sobrevivência do indivíduo i e δ_i é o seu indicador de falha (morte), a função de verossimilhança do modelo de mistura (4.2.4) é expressa como

$$\prod_{i=1}^{n} [(1-p) f_0(t_i|\boldsymbol{\theta}^\star)]^{\delta_i} [p + (1-p) S_0(t_i|\boldsymbol{\theta}^\star)]^{1-\delta_i}, \tag{4.2.5}$$

onde $f_0(\cdot|\boldsymbol{\theta}^\star)$ representa a função de densidade de probabilidade dos indivíduos não curados dado o vetor de parâmetros $\boldsymbol{\theta}^\star$. A função de verossimilhança (4.2.5) é facilmente estendida aos modelos de sobrevivência paramétricos, condicionalmente ou não a efeitos aleatórios. Por exemplo, o modelo de regressão Weibull com função de risco (4.2.3) que em conjunto com a função de verossimilhança subjacente ao submodelo longitudinal (4.2.1) formam a função de verossimilhança do modelo conjunto proposto.

4.2.2.3 Análise conjunta bayesiana

Sob uma perspetiva bayesiana, a inferência sobre os parâmetros de interesse no modelo conjunto com componentes longitudinal (4.2.1) e de sobrevivência (4.2.3) será feita com base na função de verossimilhança desse modelo, com uma fração de cura (4.2.5), referida no fim da Subseção 4.2.2.2, mais informação *a priori* sobre os parâmetros do modelo, considerando distribuições para os mesmos. Assim, a análise bayesiana permite inferir sobre os parâmetros (desconhecidos) do modelo avaliando a incerteza sobre eles através da sua distribuição *a posteriori*. Essa escolha deve-se ao fato de essas duas componentes do modelo conjunto partilharem efeitos aleatórios (modelos mistos) que são mais facilmente implementados via modelos gaussianos latentes.

De um modo mais simplificado, Paulino et al. (2018) expressa a distribuição *a posteriori* como

$$\text{distribuição } \textit{a posteriori} \;\propto\; \text{verossimilhança} \times \text{distribuição } \textit{a priori}\ .$$

No modelo conjunto proposto, a verossimilhança é expressa pelo produto das duas funções de verossimilhança associadas aos submodelos longitudinal (4.2.1) e de sobrevivência (4.2.3) com fração de cura (4.2.5). Assim, a distribuição *a posteriori* de $\boldsymbol{\theta}$, que representa todos os parâmetros do modelo conjunto, incluindo os efeitos aleatórios $\mathbf{b}_i$, dado o conjunto $\mathcal{D}$ de dados observados, é definida por

$$\pi(\boldsymbol{\theta}|\mathcal{D}) \propto \prod_{i=1}^{n} \left[\prod_{j=1}^{n_i} f_L(y_{ij}|\mathbf{b}_i, \boldsymbol{\theta}_L) \right] f_S(t_i|\mathbf{b}_i, \boldsymbol{\theta}_S) \times \pi(\boldsymbol{\theta})\ , \tag{4.2.6}$$

onde $\pi(\boldsymbol{\theta})$ representa a distribuição conjunta *a priori* de $\boldsymbol{\theta} \equiv \{\boldsymbol{\theta}_L, \boldsymbol{\theta}_S, \mathbf{b}_i\}$, $f_L(y_{ij}|\mathbf{b}_i, \boldsymbol{\theta}_L)$ é a função densidade das medidas repetidas (Y_{ij}) e $\boldsymbol{\theta}_L$ o seu vetor de parâmetros, $f_S(t_i|\mathbf{b}_i, \boldsymbol{\theta}_S)$ representa a função densidade dos tempos de sobrevivência (T_i) e $\boldsymbol{\theta}_S$ o seu vetor de parâmetros. Note-se que $\pi(\boldsymbol{\theta})$ pode ser considerada um produto de distribuições *a priori* por grupos de parâmetros e.g. a distribuição dos efeitos aleatórios longitudinais $\boldsymbol{b}_i$ em (4.2.2). Em geral, decidiu-se por adotar distribuições *a priori* vagas mas próprias para os demais parâmetros e.g. a distribuição Normal com média zero e variância "grande" que é mais apropriada num cenário de pouco conhecimento *a priori* sobre os parâmetros.

Mesmo considerando distribuições bem conhecidas para as variáveis resposta e os parâmetros do modelo conjunto, trabalhar com a distribuição *a posteriori* (4.2.6) não é simples durante o processo inferencial dos parâmetros. Neste sentido, podem-se usar métodos numéricos, analíticos ou simulação estocástica e.g. os métodos Monte Carlo via cadeias de Markov (MCMC), amplamente usados na análise bayesiana, ou aproximações de Laplace encaixadas e integradas (INLA) (veja-se um resumo desses métodos de computação bayesiana em Paulino et al. (2018) ou Amaral-Turkman, Paulino e Müller (2019)). Por conveniência, a obtenção de resultados inferenciais dos parâmetros de interesse foi feita via *software* R-INLA (RUE; MARTINO; CHOPIN, 2009) pelo seu melhor desempenho com modelos gaussianos latentes.

4.2.2.4 Avaliação de modelos

A avaliação dos modelos conjuntos pode ser feita através de vários critérios e.g. a representação da qualidade de ajuste do modelo através da função de verossimilhança logaritmizada avaliada numa estimativa pontual dos parâmetros. Aqui optou-se por utilizar duas medidas de desempenho preditivo que generalizam esse critério de avaliação também penalizando a complexidade dos modelos mas tendo em conta a variabilidade *a posteriori* dos parâmetros do modelo. Ou seja, os modelos são aqui avaliados pelos critérios de seleção DIC (*Deviance Information Criterion*) e WAIC (*Watanabe-Akaike Information Criterion*), propostos por Spiegelhalter et al. (2002) e Watanabe (2010), respectivamente. Para maiores detalhes, consulte um resumo desses critérios de avaliação de modelos em Paulino et al. (2018) ou Amaral-Turkman, Paulino e Müller (2019).

Quanto à adequação dos modelos, a proposta aqui é fazer uma análise de resíduos quer para a componente longitudinal quer para a componente de sobrevivência mas sob uma perspetiva bayesiana. Para a primeira componente, apresenta-se uma construção de resíduos mais simples com base nos valores esperados ajustados das medidas repetidas, com os resíduos bayesianos representados genericamente por

$$r_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_1, \quad i = 1, \ldots, n. \tag{4.2.7}$$

Para obter um gráfico de resíduos bayesianos, gera-se uma amostra da distribuição *a posteriori* das componentes de variância do submodelo longitudinal, usada para substituir tais componentes na densidade *a posteriori* dos resíduos. Com base numa amostra da densidade dos resíduos, pode-se obter a sua média ou mediana *a posteriori*. O gráfico de resíduos é tradicionalmente formado pelos valores de resíduos *versus* os valores ajustados, servindo para detectar discrepâncias e revelar falhas nas suposições do modelo, como a suposição de normalidade, independência ou homogeneidade de variâncias.

No contexto da análise de sobrevivência, os métodos de avaliação de modelos têm de lidar com a existência de censura, o que os torna um pouco mais complicados do que os correspondentes métodos utilizados em outros modelos de regressão. Muitos procedimentos de avaliação são baseados em quantidades conhecidas como resíduos. Estes valores, calculados para cada indivíduo, têm a particularidade de, quando o ajuste do modelo é satisfatório, o seu comportamento ser conhecido, pelo menos de forma aproximada. De acordo com Rizopoulos (2012), tem-se algumas ferramentas para avaliar esses pressupostos, entre eles, os resíduos de Cox-Snell, marginais e de martingale, e os resíduos de desvio (*deviance*).

Os resíduos Cox-Snell, tendo em conta um submodelo de sobrevivência com função de sobrevivência $S(t)$ indexada por um vetor de parâmetros $\boldsymbol{\theta}_S$, são aqui vistos em termos dos parâmetros do modelo, ou seja,

$$r_{C_i} \equiv \log(S(t_i, \boldsymbol{\theta}_S|\mathbf{x}_i)), \quad i = 1, \ldots, n. \tag{4.2.8}$$

Sendo r_{C_i} uma função do parâmetro $\boldsymbol{\theta}_S$, amostras da sua distribuição *a posteriori* são facilmente obtidas, bem como a sua média e mediana *a posteriori* (vide detalhes em Wang et al. (WANG; RYAN; FARAWAY, 2018)).

Com base nos resíduos de Cox-Snell bayesianos (4.2.8), dois outros resíduos podem ser obtidos na avaliação do submodelo de sobrevivência mas aqui sob uma abordagem bayesiana. Os resíduos de Martingale bayesianos, definidos a partir de

$$r_{M_i} = \delta_i - r_{C_i}, \tag{4.2.9}$$

onde δ_i é o indicador do evento de interesse referido anteriormente. Os resíduos de Martingale são uma realocação dos resíduos de Cox-Snell para uma média de zero para observações sem censura, de modo que eles seguem propriedades semelhantes aos resíduos em regressão linear gaussiana. Já os resíduos de desvio são usadas na pespetiva de obter resíduos que tenham uma distribuição Normal, sendo definidos a partir de

$$r_{D_i} = sgn(r_{M_i})[-2(r_{M_i} + \delta_i - r_{M_i})]^{\frac{1}{2}}, \quad i = 1, \ldots, n, \tag{4.2.10}$$

onde $sgn(\cdot)$ é o sinal da função (+ ou -), r_{M_i} é o resíduo de Martingale (4.2.9). Note-se ainda que r_{D_i} tem valor 0 quando r_{M_i} é zero. Estes gráficos de resíduos estão facilmente disponibilizados no pacote do R *brinla* (WANG; RYAN; FARAWAY, 2018).

4.3 Resultados e discussão

4.3.1 Análise preliminar

Inicialmente, os dados de pacientes com HIV/AIDS do Estado do Paraná durante 2002 e 2006 são aqui analisados preliminarmente visando a descrição das características dos pacientes. A partir da Tabela 4, observa-se que (i) 55.15% são pacientes do sexo masculino, (ii) 86.76% deles têm idade ente 16 e 50 anos, (iii) a grande maioria teve infeção oportunista prévia (71.32%), (iv) há um número elevado (92.28%) de pacientes censurados.

Tabela 4 – Estatísticas descritivas dos dados de pacientes com HIV/AIDS de algumas variáveis categorizadas.

Variável	Descrição	Total	(%)	Censura(%)	Falha(%)
Sexo	Feminino	122	44.85	42.65	2.20
	Masculino	150	55.15	49.63	5.52
Idade	< 50	236	86.76	81.25	5.51
	≥ 50	36	13.24	11.03	2.21
PrevOI	Nâo	194	71.32	68.01	3.31
	Sim	78	28.68	24.26	4.41
Média de CD4	> 200	59	21.69	20.59	1.10
	≤ 200	213	78.31	71.69	6.62
Status	Censura	251	92.28	-	-
	Morte	21	7.72	-	-
	Total	272			

Quanto às variáveis contínuas idade (anos) e tempo de sobrevivência (dias), a idade média e mediana dos pacientes são, respectivamente, 36.71 e 36 anos, não havendo pacientes com mais de 65 anos e menos de 16 anos, enquanto o tempo máximo de recolha da contagem de células CD4 foi 1.812 dias com média de 719.7 superior à mediana de 665.5 dias, conforme se observa na Tabela 5. Os tempos médio e mediando de sobrevivência são, respectivamente, 2.316 e 2.327 anos, sendo esses valores semelhantes aos dos tempos com censura mas superiores aos de morte devido ao HIV/AIDS.

Como pode ser visto também na Figura 4.3.1, bem como na Figura 4.2.1, as curvas de sobrevivência de Kaplan-Meier não tendem para zero à medida que tempo passa, indicando que os dados de pacientes com HIV/AIDS podem possuir uma fração de cura. Ou seja, a probabilidade de um paciente sobreviver até um determinado momento se mantêm num patamar aproximadamente constante e isso aponta para a presença de pacientes não suscetíveis ao evento de interesse. Além disso, os gráficos evidenciam uma maior sobrevivência dos pacientes mais jovens, sem infeção oportunista prévia e mulheres.

Tabela 5 – Estatísticas descritivas dos dados de pacientes com HIV/AIDS das variáveis contínuas idade (anos), tempo de recolha de CD4 (TR em dias) e tempos de sobrevivência (TS em anos), com morte (M) ou censura (C).

	Mínimo	1º quartil	Mediana	Média	3º quartil	Máximo
Idade	16.00	29.00	36.00	36.71	42.25	65.00
TR	0.0	301.0	665.5	719.7	1093.8	1812.0
TS	0.131	1.186	2.316	2.327	3.425	5.051
TS (M)	0.167	0.583	0.977	1.321	2.141	3.691
TS (C)	0.131	1.328	2.415	2.400	3.515	5.051

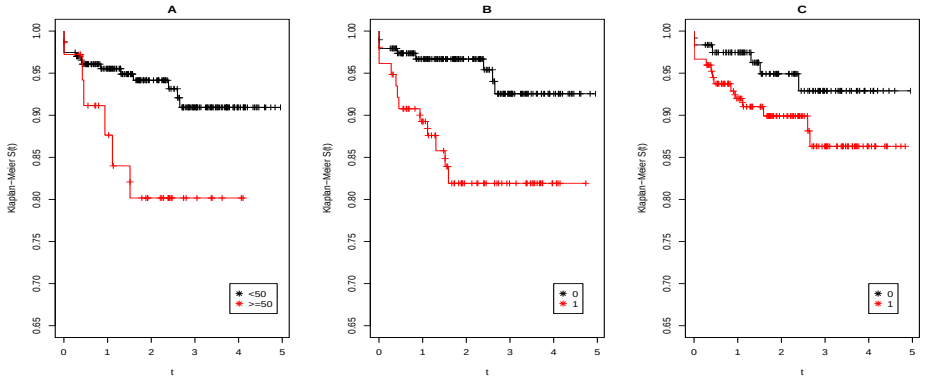

Figura 4.3.1 – Curvas de sobrevivência de Kaplan-Meier: (A) Faixa etária, (B) PrevOI (0: Não, 1: Sim), (C) Sexo (0: Feminino, 1: Masculino).

Na Figura 4.3.2, cada linha cinza representa a trajetória da medida de contagem CD4 para cada indivíduo e desta forma pode-se observar o comportamento de cada elemento da amostra no decorrer do tempo em relação à medida CD4. Note-se que há um comportamento distinto desses perfis no começo do estudo, indicando a necessidade de introduzir um efeito aleatório no intercepto da relação linear entre CD4 e o seu tempo de recolha pois cada perfil possui um intercepto diferente.

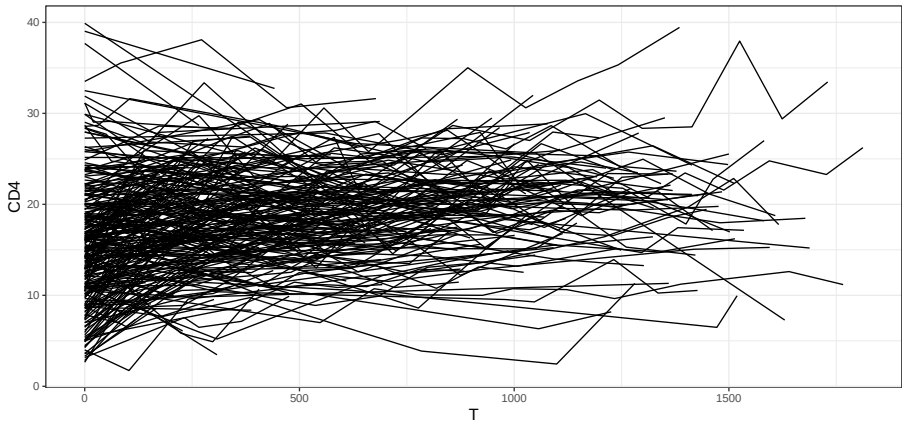

Figura 4.3.2 – Medidas de contagem $CD4$ ao longo do tempo (em dias).

Com o intuito de verificar se há efeito da covariável sexo no decorrer do tempo de recolha do CD4, fez-se um gráfico de perfil médio de CD4 quer para homens quer para mulheres. A Figura 4.3.3 representa uma média geral das medidas repedidas, por meio de uma regressão ponderada localmente (*lowess*) para "suavizar" sobre toda a sua variabilidade e dar uma noção da tendência geral ou média (veja-se Cleveland (CLEVELAND, 1979)). Em ambos os painéis, esquerdo (feminino) e direito (masculino), percebe-se uma variação considerável da média da contagem CD4, e ainda alguma diferença entre homens e mulheres a partir de 100 dias.

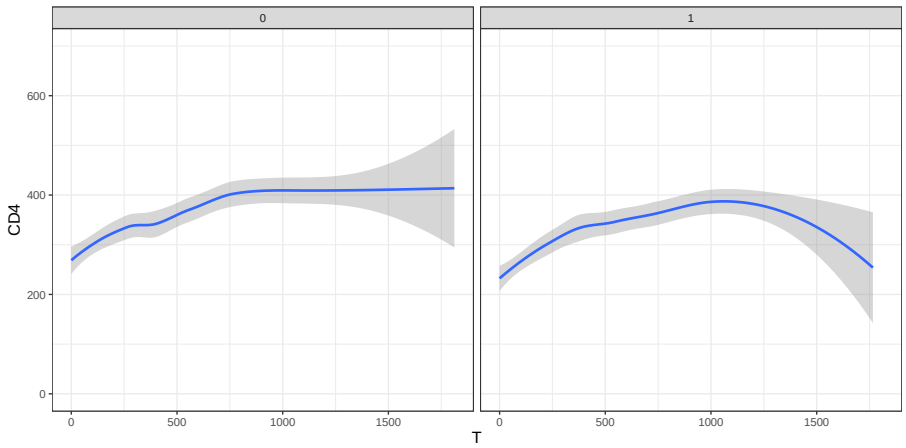

Figura 4.3.3 – Perfil médio de CD4 dos pacientes com HIV/AIDS por mulheres (esquerda) e homens (direita) ao longo do tempo (T em dias).

A Figura 4.3.4 representa graficamente a distribuição das frequências relativas do número de exames por indivíduo ao longo do tempo, sendo possível notar que a frequência de exames diminui com o passar do tempo e poucos são os pacientes que têm mais de 10 visitas aos laboratórios para medição da contagem de CD4 durante o período 2002-2006.

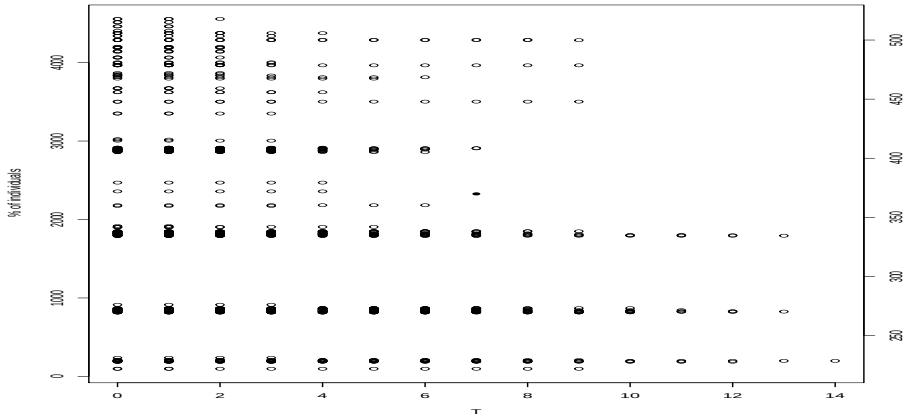

Figura 4.3.4 – Distribuição das frequências relativas do número de exames por indivíduo ao longo do tempo (T em dias).

Com o objetivo de investigar qual dos modelos de sobrevivência paramétricos tem um melhor ajustamento aos dados de pacientes com HIV/AIDS, produziram-se gráficos Q-Q comparando quantis empíricos e teóricos com base nas distribuições Weibull, Exponencial e Log-normal e apresentados na Figura 4.3.5. Note-se que a distribuição Weibull, incluindo o seu caso particular distribuição Exponencial ($\rho = 1$), se aproxima mais da distribuição empírica dos tempos de sobrevivência, sendo assim o modelo paramétrico selecionado para descrever os tempos de sobrevivência dos pacientes com HIV/AIDS.

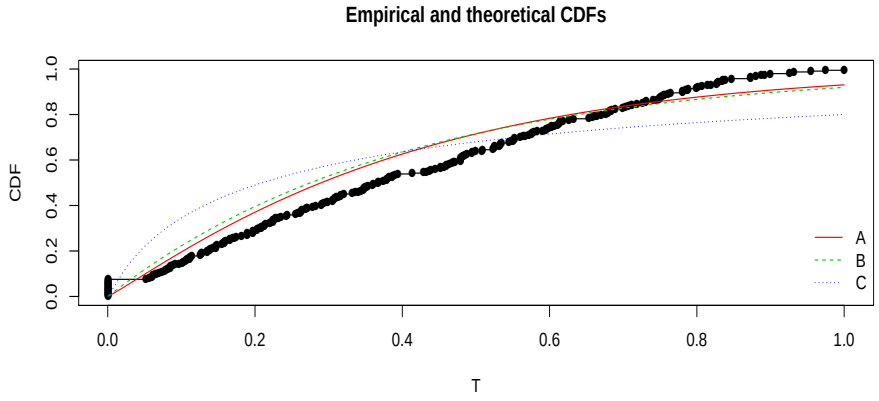

Figura 4.3.5 – Gráficos Q-Q para os tempos de sobrevivência (T) com base nas distribuições Weibull (A), Exponencial (B) e Log-normal (C).

4.3.2 Análise conjunta

Na análise conjunta dos dados de pacientes com HIV/AIDS do Estado do Paraná durante 2002 e 2006, o submodelo longitudinal é representado pela variável resposta Y_{ij} denotando a raiz quadrada da j-ésima medida de contagem de células CD4 observada no instante t_{ij} do paciente i, $i = 1, \ldots, 272$. Esta transformação fez com que a resposta longitudinal ficasse mais próxima da distribuição gaussiana, como foi visto na Figura 4.2.2.

Assim, a resposta Y do submodelo longitudinal segue uma distribuição gaussiana adicionando dois efeitos aleatórios, um no intercepto e outro na inclinação do tempo de recolha das medidas repetidas. As covariáveis selecionadas para os efeitos fixos são o tempo de recolha das medidas CD4 (t_{ij}), sexo (S), idade (I) e infecção oportunista prévia (PrevOI ou P), bem como a interação entre sexo e tempo de recolha CD4. Dada a expressão em 4.2.2.1, esse submodelo pode ser escrito como

$$y_i(t_{ij}) = (\beta_{11}+b_{1i})+(\beta_{12}+b_{2i})t_{ij}+\beta_{13}S_i+\beta_{14}I_i+\beta_{15}P_i+\beta_{16}S_it_{ij}+w_{1i}(t_{ij})+\epsilon_{ij} \qquad (4.3.1)$$

onde, por conveniência de notação, $w_{1i}(t_{ij}) = b_{1i} + b_{2i}\, t_{ij}$ representa os efeitos aleatórios longitudinais (correlacionados linearmente) para o intercepto e a inclinação ao longo dos

tempos de recolha CD4 e o vetor de efeitos aleatórios gaussianamente distribuídos i.e. $\mathbf{b}_i = (b_{1i}, b_{2i}) \sim \mathcal{N}(0, \mathbf{\Sigma})$.

Quanto ao submodelo de sobrevivência, a variável resposta é o tempo desde o diagnóstico até a ocorrência da morte do paciente com HIV/AIDS. O submodelo leva em consideração os dois efeitos aleatórios do submodelo longitudinal (4.3.1). As covariáveis selecionadas para os efeitos fixos são sexo (S), idade (I) e infecção oportunista prévia (PrevOI ou P). Para os tempos de sobrevivência desses pacientes, denotados genericamente aqui por T_i, considerou-se a distribuição Weibull com dois parâmetros, $T_i \sim \mathcal{W}(\rho, \eta_i(t))$, e com parâmetro de forma $\rho = 1$, i.e., distribuição Exponencial, $T_i \sim \mathcal{E}(\eta_i(t))$, onde a expressão do parâmetro de escala do paciente i é dada por

$$\eta_i(t) = \beta_{21} + \beta_{22}S_i + \beta_{23}I_i + \beta_{24}P_i + \gamma_i w_{2i}(t), \tag{4.3.2}$$

onde, por simplicidade, $w_{2i}(t) \equiv g(b_{1i}, b_{2i})$ representando b_{1i} e b_{2i} os efeitos aleatórios para o intercepto e a inclinação do tempo de recolha CD4 no submodelo longitudinal, respectivamente.

Devido à escassa informação *a priori*, utilizaram-se distribuições *a priori* não informativas para parâmetros do modelo conjunto (4.3.1)-(4.3.2). Por exemplo, distribuição Gama para a precisão dos erros aleatórios das medidas repetidas i.e. $\tau_e = 1/\sigma_e^2 \sim \mathcal{G}(0.001, 0.001)$ e distribuição gaussiana para os parâmetros dos efeitos fixos quer do submodelo longitudinal $\beta_{1k} \sim \mathcal{N}(0, 10^5)$ quer do submodelo de sobrevivência $\beta_{2k} \sim \mathcal{N}(0, 10^5)$, considerando a variância de β_{22}, β_{23} e β_{24} igual a 100 por conveniência computacional, e dos efeitos aleatórios $\mathbf{b}_i = (b_{1i}, b_{2i}) \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, como em (4.2.2), e $\gamma_i \sim \mathcal{N}(0, 100)$.

Na Tabela 6 estão os resultados de vários modelos conjuntos mistos considerados para avaliar a associação entre o submodelo longitudinal e o submodelo de sobrevivência dos dados de pacientes com HIV/AIDS. Os modelos conjuntos foram definidos pela forma diferenciada de introdução dos efeitos aleatórios individuais no intercepto e na inclinação do tempo de recolha CD4 formando uma variedade de modelos candidatos. Tais modelos foram avaliados pelos critérios de seleção de modelos DIC e WAIC, referidos na Subseção 4.2.2.4. Os valores dessas medidas e as inferências dos parâmetros do modelo conjunto foram obtidos através do *software* R-INLA Rue, Martino e Chopin (2009) usando as famílias de distribuições gaussiana e Weibull-cure.

O modelo I na Tabela 6 representa um modelo sem heterogeneidade não observada i.e. sem efeitos aleatórios. Os modelos II e IV possuem efeitos aleatórios somente no submodelo longitudinal sendo interpretados como uma análise separada em contraste à análise conjunta. Os modelos III e V permitem a partilhada somente dos efeitos aleatórios no intercepto, enquanto o modelo VI partilha somente os efeitos aleatórios da in-

Tabela 6 – Valores de DIC e WAIC dos modelos conjuntos dos dados HIV/AIDS, com termos longitudinal $w_1(t)$ e de sobrevivência $w_2(t)$.

Modelo	$w_1(t)$	$w_2(t)$	DIC	WAIC
Sem efeitos aleatórios				
I	0	0	8673.506	8674.140
Intercepto aleatório				
II	b_1	0	7446.220	7486.163
III	b_1	$\gamma_1(b_1)$	7431.265	7474.665
Interceptos aleatórios e inclinações aleatórias				
IV	$b_1 + b_2 t$	0	6839.904	6907.017
V	$b_1 + b_2 t$	$\gamma_1(b_1)$	6832.049	6900.460
VI	$b_1 + b_2 t$	$\gamma_1(b_2)$	6841.189	6990.238
VII	$b_1 + b_2 t$	$\gamma_1(b_1 + b_2)$	6891.477	6968.998
VIII	$b_1 + b_2 t$	$\gamma_1 b_1 + \gamma_2 b_2$	6891.477	9025.065

clinação. Os demais modelos levam em consideração o intercepto aleatório e inclinação aleatória do submodelo longitudinal partilhado com o submodelo de sobrevivência, tendo o modelo VII um parâmetro de associação comum tanto no intercepto como na inclinação e o Modelo VIII diferente parâmetro de associação em ambas componentes. Dentre os modelos conjuntos ajustados, o modelo V obteve o menor DIC e WAIC, sendo o modelo escolhido para descrever os dados de pacientes com HIV/AIDS.

A Tabela 7 resume algumas medidas sumárias *a posteriori* dos parâmetros de interesse do modelo *V* tais como a média, desvio padrão e intervalo de credibilidade HPD (*High Posterior Density*) a 95%. A covariável sexo tem um efeito negativo na média do biomarcador longitudinal, o que sugere que os homens têm contagens de CD4 inferiores ás das mulheres. O mesmo se passa com as covariáveis idade e PrevOI indicando que pacientes acima dos 50 anos e que já tiveram alguma doença oportunista prévia têm valores de CD4 menores que os pacientes no grupo com menos de 50 anos e que ainda não apresentavam qualquer doença oportunista antes da entrada no estudo, respectivamente. Esses resultados também permitem concluir que há interação entre sexo e o tempo de recolha das contagens CD4 (TempoR), mostrando que no decorrer do tempo os homens têm um aumento na medida CD4 superior ao das mulheres. Já no submodelo de sobrevivência nota-se que as covariáveis sexo, idade e infecção oportunista prévia (PrevOI) não foram significativas no risco de morte devido à HIV/AIDS.

Ainda sobre a Tabela 7, as quantidades *a posteriori* dos parâmetros de interesse fornecem fortes evidências de uma associação negativa (latente) entre os dois componentes

Tabela 7 – Quantidades *a posteriori:* média, desvio padrão (DP) e intervalos de credibilidade (IC) HPD a 95% dos parâmetros do modelo V.

Parâmetro	Média	DP	IC HPD 95%
β_{11}	18.057	0.642	(16.794, 19.315)
β_{12} (TempoR)	0.438	0.171	(0.103, 0.773)
β_{13} (Sexo)	-2.064	0.864	(-3.764, -0.371)
β_{14} (Idade)	-2.243	0.944	(-4.098, -0.389)
β_{15} (PrevOI)	-3.412	0.730	(-4.847, -1.979)
β_{16} (TempoR.Sexo)	0.602	0.234	(0.145, 1.063)
β_{21}	-2.226	0.859	(-3.897, -0.514)
β_{22} (Sexo)	1.337	0.775	(-0.047, 2.973)
β_{23} (Idade)	0.142	0.833	(-1.582, 1.661)
β_{24} (PrevOI)	0.904	0.697	(-0.511, 2.257)
ρ	0.722	0.150	(0.467, 1.056)
γ_1	-0.636	0.050	(-0.723, -0.528)
p	0.687	0.107	(0.444, 0.856)
σ_e^2	2.381	1.543	(2.336, 2.427)
σ_{b1}	6.412	2.532	(6.337, 6.486)
σ_{b2}	1.548	1.244	(1.511, 1.585)

submodelos, estimada em forma intervalar em $(-0.723, -0.528)$. A estimativa do parâmetro de forma da distribuição Weibull em cerca de 0.722 fornece algum indício que a distribuição Exponencial ($\rho = 1$) poderia ser uma boa alternativa para a distribuição dos tempos de sobrevivência. Além disso, há forte heterogeneidade não observada quer no intercepto quer na inclinação do tempo de recolha CD4 em volta de 6.412 e 1.548, respectivamente, bem como alguma dispersão presente nos erros aleatórios, cujo intervalo de credibilidade é $(2.336, 2.427)$. Por fim, há presença de fração de cura nos pacientes com HIV/AIDS, estimada aqui com média 0.687 e desvio padrão pequeno (0.107), confirmando a conjetura inicial que há pacientes com HIV/AIDS nesta amostra que sobrevivem a longo prazo.

A análise conjunta proposta para os submodelos longitudinal e de sobrevivência pode apresentar melhoria no tempo mediano de sobrevivência dos pacientes com HIV/AIDS quando comparada com aquela obtida separadamente. Para ilustrar esta situação vamos considerar 2 pacientes com HIV/AIDS: (i) Paciente 3 do sexo feminino, 33 anos, sem infecção oportunista prévia e com morte no 1348 $^\circ$ dia; (ii) Paciente 90 do sexo masculino, 47 anos, sem infecção oportunista prévia e censurado no 1044 $^\circ$ dia. A Figura 4.3.6 mostra a evolução diversificada da raiz quadrada da medida CD4 por exame ao longo do tempo dos 2 pacientes (gráficos da esquerda), como também as densidades *a posteriori* do tempo mediano de sobrevivência dos pacientes 90 e 186 para os modelos II (análise separada) e V (análise conjunta)(gráficos da direita). A análise separada aqui é entendida como a au-

sência de efeitos aleatórios longitudinais na função de risco de morte devido ao HIV/AIDS ($w_2(t) = 0$ no submodelo de sobrevivência (4.3.2)).

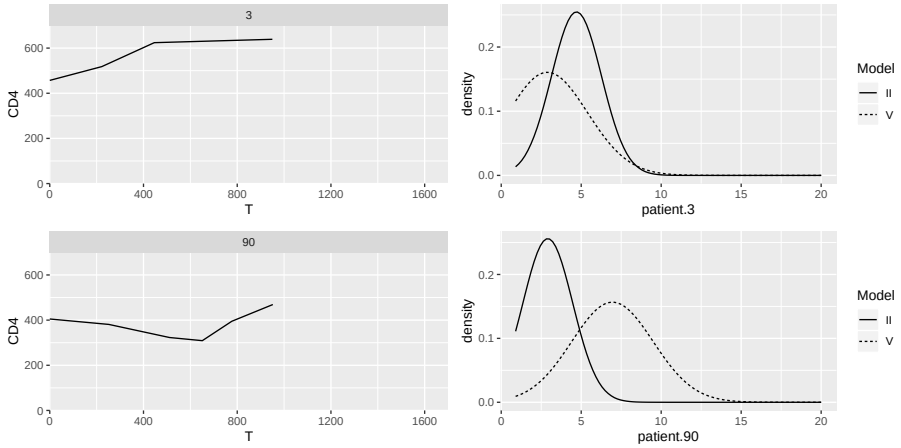

Figura 4.3.6 – Trajetórias das medidas CD4 (esquerda) e densidades (direita) do tempo mediano de sobrevivência dos pacientes 3 e 90.

Assim, a Figura 4.3.6 da indicação de que os pacientes 3 e 90 têm uma trajetória CD4 relativamente "boa" (começa relativamente baixo e depois aumenta). Mas para o paciente 90 os resultados do modelo separado (II) estão aquém dos do modelo conjunto (V) em termos da predição do tempo de sobrevivência mediano, enquanto que para o paciente 3 os resultados do modelo conjunto são mais "conservadores" do que os do modelo separado e isso pode fazer sentido visto que este paciente veio a morte de HIV/AIDS. Enfim, o modelo conjunto inverte, por vezes, os resultados do modelo separado no sentido de que agora o paciente 90 com "boa" trajetória CD4 sobrevive mais segundo o modelo conjunto. A tendência crescente das contagens CD4 pode ser enganadora pois a sua média é baseada no perfil longitudinal dos indivíduos que viveram mais tempo e que, por consequência, têm valores de CD4 mais elevados.

Para a avaliação do modelo conjunto V via análise de resíduos, produziram-se os gráficos de resíduos bayesianos para as componentes longitudinal e de sobrevivência. No primeiro caso, calcularam-se as distribuições *a posteriori* dos resíduos (4.2.7) para as medidas repetidas de contagem CD4. A Figura 4.3.7 apresenta os gráficos de resíduos longitudinais bayesianos: resíduos *versus* índice e resíduos *versus* tempos de recolha CD4. Em geral, os resíduos bayesianos mostram um padrão aleatório em torno de zero em ambos os gráficos mas com alguns valores grandes positivos e negativos. Entretanto, não há em geral indícios de falhas no ajustamento do modelo conjunto V em relação ao submodelo longitudinal para os dados de pacientes com HIV/AIDS.

Quanto ao submodelo de sobrevivência do modelo conjunto V, os gráficos de resíduos bayesianos de Cox-Snell (4.2.8), de desvio (4.2.10) e de Martingale (4.2.9) estão exibidos

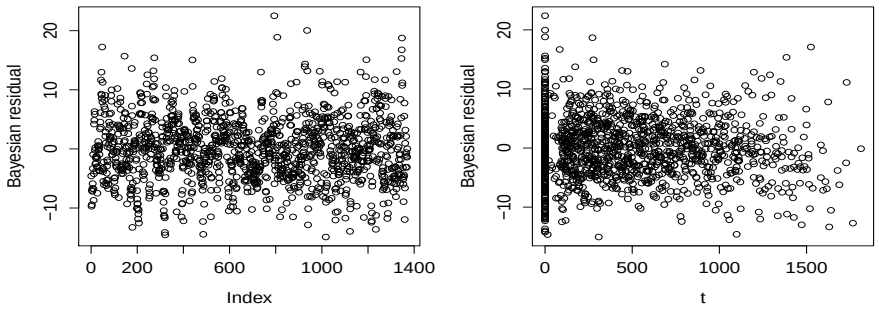

Figura 4.3.7 – Gráficos de resíduos longitudinais bayesianos: resíduos *versus* índice (esquerdo) e resíduos *versus* tempos de recolha CD4 (direito).

na Figura 4.3.8. Esses gráficos sugerem que o ajustamento desta componente do modelo V não está mal e.g. os resíduos de Cox-Snell estão em torno da reta de 45°, exceto na cauda, onde a variabilidade da estimativa de risco cumulativo de morte com HIV/AIDS é grande. Isso deve estar relacionado com alguns valores atípicos presentes nos resíduos de desvio e de Martingale. No gráfico de resíduos de Martingale há uma curva ajustada suavizada (linha sólida) com intervalos de credibilidade a 95% (linha tracejada) indicando que o uso de uma forma linear para a covariável idade na equação de regressão Weibull é suficiente.

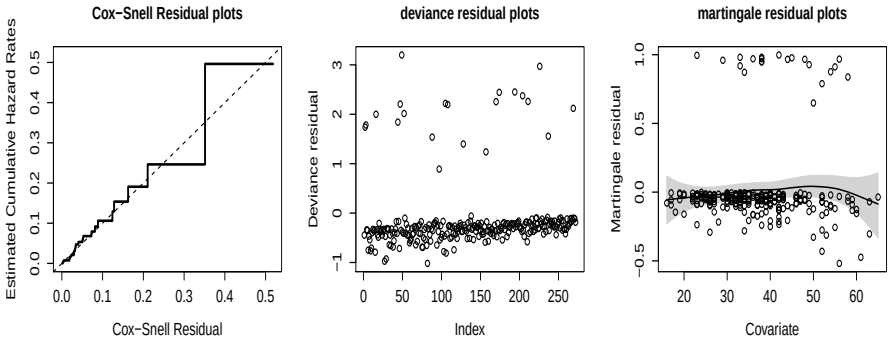

Figura 4.3.8 – Gráficos de resíduos de sobrevivência bayesianos: resíduos Cox-Snell (esquerdo), de desvio (centro) e de Martingale (direito).

4.4 Análise Final

A análise conjunta de dados envolvendo pacientes com HIV/AIDS no Estado do Paraná no período 2002 e 2006 mostrou-se ser mais adequada por ter em conta a dependência entre as medidas repetidas de contagem de células CD4 e os tempos de sobrevivência devido à doença, indicando também fração de cura para alguns pacientes. Por exemplo,

detetou-se essa melhor adequação da abordagem conjunta através da predição do tempo mediano de sobrevivência para alguns pacientes com boa ou má trajetórias CD4 que foram confirmas pelo modelo conjunto, contrariamente ao modelo separado i.e. que não partilha uma heterogeneidade não observada dos pacientes no submodelo longitudinal com o submodelo de sobrevivência.

Estes modelos conjuntos têm apresentado bons resultados em vários contextos, incluindo quando as variáveis podem ser de vários tipos com diferentes modelos probabilisticos. Por exemplo, Juarez-Colunga, Silva e Dean (2017) desenvolveram métodos conjuntos para situações em que, além da contagem de eventos, é registrada uma marca que indica uma medida de gravidade dos eventos. Neste caso, as variáveis respostas são discretas e os modelos Poisson e Binomial são ligados também via efeitos aleatórios e os resultados inferenciais foram obtidos através dos métodos MCMC.

Apesar de os métodos MCMC serem amplamente usados na análise bayesiana, foi também mais conveniente utilizar aproximações de Laplace encaixadas e integradas (INLA) na obtenção de resultados inferenciais com base em efeitos aleatórios gaussianos. Isso foi crucial para se ter uma maior rapidez na execução dos programas computacionais (aqui executados no R-INLA). Já agora, a opção por introduzir associação entre as duas componentes do modelo conjunto via efeitos aleatórios foi também vantajoso em termos computacionais e de interpretação dos resultados inferenciais. Uma alternativa aos modelos conjuntos mistos pode ser modelos mistos com cópulas a captar essa associação dos submodelos via a construção da função de sobrevivência bivariada usando a relação funcional entre as funções de sobrevivência marginais. Algumas propriedades assintóticas do estimador da função bivariada pode ser encontradas em Romeo et al. (2013).

Neste trabalho foi também destacado a presença de fração de cura no submodelo de sobrevivência utilizando uma abordagem abordagem paramétrica i.e. distribuição Weibull para os tempos de sobrevivência. Acredita-se que os pacientes de HIV/AIDS já podem ser considerados doentes crónicos e portanto faz todo o sentido adotar uma componente no modelo que contemple os sobreviventes a longo prazo.

Para trabalhos futuros espera-se fazer uma comparação mais ampla dos modelos conjuntos, comparando os modelos com regressão Weibull, Cox e Aalen para sobreviventes a longo prazo, sobretudo sob uma abordagem bayesiana. Já agora, faltou aqui desenvolver os modelos conjuntos mistos até a predição de tempos de sobrevivência, como também fazer uma análise de sensibilidade os valores fixados para os hiperparâmetros i.e. os parâmetros da distribuições *a priori*.

Por fim, os modelos conjuntos com componentes longitudinal e de sobrevivência com fração de cura, ajustados aqui aos dados de pacientes com HIV/AIDS no Estado do Pa-

raná, descreveram a população em causa de forma mais realista que os modelos separados. Dessa forma, os modelos conjuntos propostos podem ser mais condizente a analisar dados de pacientes com variáveis resposta longitudinal e de sobrevivência num contexto de diferentes enfermidades tais como câncer, doenças cardiovasculares, etc.

Capítulo 5

Comentários finais

Este trabalho apresentou resumidamente os principais conceitos de modelos de sobrevivência e modelos conjuntos , com fração de cura, onde os parâmetros foram estimados utilizando análise Bayesiana, por meio do método analítico INLA (Rue, Martino e Chopin (2009)). Os resultados expostos foram obtidos com a utilização do software livre R Core Team (2018).

No Capítulo 3 foi aplicado o modelo de sobrevivência com fração de cura a dados de pacientes com câncer da mama no estado do Paraná, descrevendo a população de forma mais realista que os modelos de sobrevivência tradicionais, visto que o indivíduo pode se curar e nunca apresentar o evento de interesse. Dessa forma, esse modelo mostra-se mais conveniente e condizente com a realidade. Após uma análise preliminar obteve-se as estimativas dos parâmetros do modelo selecionado, levando em consideração algumas discussões realizadas com um médico oncologista. Observou-se que a proporção de curados foi de aproximadamente 57.6%, quando se tem em conta as covariáveis Faixa etária, Raça, Estado civil, Código de topografia e a Morfologia da doença. Para esse modelo pode-se concluir que a a Topografia, a Morfologia e a Raça tiveram um efeito positivo no risco de morte, enquanto as covariáveis Faixa etária e Estado Civil tiveram um efeito negativo no risco de morte.

No Capítulo 4 teve-se como principal objetivo o uso de modelos conjuntos que exploram métodos de análise de sobrevida capazes de analisar variáveis dependentes do tempo conjuntamente. O modelo conjunto foi utilizado a dados reais de pacientes com HIV/AIDS no estado do Paraná, compreendo duas partes, uma componente longitudinal para capturar o comportamento dos diferentes perfis de CD4 ao longo do tempo e uma componente de sobrevivência com fração de cura para lidar com indivíduos sobreviventes a longo prazo.

Nesta aplicação podemos concluir que a covariável sexo tem um efeito negativo na média do biomarcador longitudinal, o que sugere que os homens têm contagens de CD4, durante o período de acompanhamento, inferiores às das mulheres. O mesmo se passa com as covariáveis Idade e PrevOI. Isto significa que pacientes acima dos 50 anos e que já tiveram alguma doença oportunista prévia têm valores de CD4 menores que os pacientes no grupo com menos de 50 anos e que ainda não apresentavam qualquer doença oportunista antes da entrada no estudo. Os resultados também permitem concluir que há interação entre sexo e o tempo, mostrando que no decorrer do tempo os homens têm um aumento na medida CD4 superior ao das mulheres. Já no modelo de sobrevivência nota-se que as covariáveis Sexo, Idade e PrevOI não foram significativas no risco de morte devido à HIV/AIDS.

Quanto aos resíduos utilizado neste trabalho os gráficos sugerem que o ajustamento tanto para o modelo de sobrevivência para pacientes com cancêr da mama quanto para os modelos de sobrevivência e longitudinal para os pacientes com HIV/AIDS não está mal e.g. os resíduos de Cox-Snell estão em torno da reta de 45°.

Como passos futuros temos ainda como desafios uma comparação mais ampla dos modelos conjuntos e.g. comparando os modelos com regressão Weibull, Cox e Aalen para sobreviventes a longo prazo, sobretudo sob uma abordagem bayesiana. Além disso, fazer uma análise de sensibilidade sobre os valores considerados para os parâmetros das distribuições *a priori* consideradas.

Agradecimentos

Um agradecimento ao médico oncologista Dr. Antonio Soares Safar pelas dúvidas sanadas no decorrer do trabalho.

Um agradecimento à Maria Goretti Fonseca, Cláudia Coeli, Valeska Andreozzi e Rui Martins por fornecer os dados para o estudo.

Referências

AMARAL-TURKMAN, M. A.; PAULINO, C. D.; MÜLLER, P. *Computational Bayesian Statistics: An Introduction*. [S.l.]: Cambridge University Press, 2019.

AMICO, M.; KEILEGOM, I. van. Cure models in survival analysis. *Annual Review of Statistics and Its Application*, v. 5, p. 311–342, 2018.

ANGELIS, R. D.; SANT, M.; COLEMAN, M. P.; FRANCISCI, S.; BAILI, P.; PIERANNUNZIO, D.; TRAMA, A.; VISSER, O.; BRENNER, H.; ARDANAZ, E. et al. Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE-5: A population-based study. *The Lancet Oncology*, Elsevier, v. 15, n. 1, p. 23–34, 2014.

APARÍCIO, M. M. E. F. *Análise de Sobrevivência do Cancro da Mama Masculina*. Dissertação (Mestrado) — Instituto Superior Técnico, Universidade de Lisboa, 2013.

ARAÚJO, I. M. de A.; FERNANDES, A. F. C. O significado do diagnóstico do câncer de mama para a mulher. *Escola Anna Nery Revista de Enfermagem*, Universidade Federal do Rio de Janeiro, v. 12, n. 4, p. 664–671, 2008.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, Taylor & Francis, v. 47, n. 259, p. 501–515, 1952.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B*, JSTOR, v. 11, n. 1, p. 15–53, 1949.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, Wiley Online Library, v. 26, n. 2, p. 211–243, 1964.

BROWN, E. R.; IBRAHIM, J. G. Bayesian approaches to joint cure-rate and longitudinal models with applications to cancer vaccine trials. *Biometrics*, Wiley Online Library, v. 59, n. 3, p. 686–693, 2003.

CLEVELAND, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, Taylor & Francis, v. 74, n. 368, p. 829–836, 1979.

COLEMAN, M.; FORMAN, D.; BRYANT, H.; BUTLER, J.; RACHET, B.; MARINGE, C.; NUR, U.; TRACEY, E.; COORY, M.; HATCHER, J. et al. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995–2007 (the International Cancer Benchmarking Partnership): An analysis of population-based cancer registry data. *The Lancet*, Elsevier, v. 377, n. 9760, p. 127–138, 2011.

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

CORBIÈRE, F.; COMMENGES, D.; TAYLOR, J. M.; JOLY, P. A penalized likelihood approach for mixture cure models. *Statistics in Medicine*, Wiley Online Library, v. 28, n. 3, p. 510–524, 2009.

COX, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, v. 34, p. 187–220, 1972.

COX, D. R.; SNELL, E. J. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 30, n. 2, p. 248–265, 1968.

CROWTHER, M. J.; ANDERSSON, T. M.-L.; LAMBERT, P. C.; ABRAMS, K. R.; HUMPHREYS, K. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine*, Wiley Online Library, v. 35, n. 7, p. 1193–1209, 2016.

CUCCHETTI, A.; FERRERO, A.; CESCON, M.; DONADON, M.; RUSSOLILLO, N.; ERCOLANI, G.; STACCHINI, G.; MAZZOTTI, F.; TORZILLI, G.; PINNA, A. D. Cure model survival analysis after hepatic resection for colorectal liver metastases. *Annals of Surgical Oncology*, Springer, v. 22, n. 6, p. 1908–1914, 2015.

DIGGLE, P.; HEAGERTY, P.; LIANG, K.-Y.; ZEGER, S. *Analysis of Longitudinal Data*. 2nd. ed. [S.l.]: Oxford University Press, 2002.

FAUCETT, C. L.; THOMAS, D. C. Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in Medicine*, v. 15, p. 1663–1685, 1996.

FONSECA, M. G. P.; COELI, C. M.; LUCENA, F. F. A.; VELOSO, V. G.; CARVALHO, M. S. Accuracy of a probabilistic record linkage strategy applied to identify deaths among cases reported to the Brazilian AIDS surveillance database. *Cadernos de Saúude Pública*, v. 26, n. 7, p. 1431–1438, 2010.

GELMAN, A.; HWANG, J.; VEHTARI, A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, Springer, v. 24, n. 6, p. 997–1016, 2014.

GUO, X.; CARLIN, B. P. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, Taylor & Francis, v. 58, n. 1, p. 16–24, 2004.

HENDERSON, R.; DIGGLE, P.; DOBSON, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, Oxford University Press, v. 1, n. 4, p. 465–480, 2000.

HUANG, Y.; HU, X. J.; DAGNE, G. A. Jointly modeling time-to-event and longitudinal data: a Bayesian approach. *Statistical Methods & Applications*, Springer, v. 23, n. 1, p. 95–121, 2014.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. *Bayesian Survival Analysis*. [S.l.]: Wiley, 2001.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. B ayesian survival analysis. *Wiley StatsRef: Statistics Reference Online*, Wiley Online Library, 2014.

INCA. *Câncer de mama*. 2020. Instituto Nacional de Câncer José Alencar Gomes da Silva. Disponível em: <https://www.inca.gov.br/tipos-de-cancer/cancer-de-mama>. Acesso em: 09/02/2020.

JUAREZ-COLUNGA, E.; SILVA, G. L.; DEAN, C. B. Joint modeling of zero-inflated panel count and severity outcomes. *Biometrics*, v. 73, p. 1413–1423, 2017.

KHAN, M. H. R.; ABABNEH, F. Efficiency of Weibull regression model over Cox regression model: A simulation study. *Journal of Biostatistics*, v. 12, p. 169–178, 2016.

LAMBERT, P. C.; THOMPSON, J. R.; WESTON, C. L.; DICKMAN, P. W. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, Oxford University Press, v. 8, n. 3, p. 576–594, 2006.

LAWLESS, J. F. *Statistical Models and Methods for Lifetime Data*. 2nd. ed. [S.l.]: Wiley, New York, 2003.

LEE, E. T.; WANG, J. *Statistical Methods for Survival Data Analysis*. [S.l.]: John Wiley & Sons, 2003.

LI, K.; LUO, S. Dynamic predictions in Bayesian functional joint models for longitudinal and time-to-event data: An application to Alzheimer's disease. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 28, n. 2, p. 327–342, 2019.

MALLER, R. A.; ZHOU, X. *Survival Analysis with Long-term Survivors*. [S.l.]: Wiley New York, 1996.

MANGAL, T. D.; MEIRELES, M. V.; PASCOM, A. R. P.; COELHO, R. A.; BENZAKEN, A. S.; HALLETT, T. B. Determinants of survival of people living with HIV/AIDS on antiretroviral therapy in Brazil 2006–2015. *BMC Infectious Diseases*, v. 19, p. 206, 2019.

MARTINS, R.; SILVA, G. L.; ANDREOZZI, V. Bayesian joint modeling of longitudinal and spatial survival aids data. *Statistics in Medicine*, v. 35, n. 19, p. 3368–3384, 2016.

MARTINS, R.; SILVA, G. L.; ANDREOZZI, V. Joint analysis of longitudinal and survival AIDS data with a spatial fraction of long-term survivors: A Bayesian approach. *Biometrical Journal*, Wiley Online Library, v. 59, n. 6, p. 1166–1183, 2017.

NIEKERK, J. van; BAKKA, H.; RUE, H. Joint models as latent Gaussian models-not reinventing the wheel. *arXiv preprint arXiv:1901.09365*, 2019.

PAULINO, C. D.; AMARAL-TURKMAN, M. A.; MURTEIRA, B.; SILVA, G. L. *Estatística Bayesiana*. [S.l.]: Fundação Calouste Gulbenkian, Lisboa, 2018.

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

RAMIRES, T. G.; CORDEIRO, G. M.; KATTAN, M. W.; HENS, N.; ORTEGA, E. M. Predicting the cure rate of breast cancer using a new regression model with four regression structures. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 27, n. 11, p. 3207–3223, 2018.

RIZOPOULOS, D. *Joint Models for Longitudinal and Time-to-event Data: With applications in R*. [S.l.]: Chapman and Hall/CRC, 2012.

ROMEO, J. S.; TANAKA, N. I.; LIMA, A. C. Pedroso-de; SALINAS-TORRES, V. H. Large sample properties for a class of copulas in bivariate survival analysis. *Metrika*, v. 76, p. 997–1015, 2013.

RUE, H.; MARTINO, S.; CHOPIN, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, Wiley Online Library, v. 71, n. 2, p. 319–392, 2009.

SEPPä, K.; RUE, H.; HAKULINEN, T.; LääRä, E.; SILLANPää, M. J.; PITKäNIEMI, J. Estimating multilevel regional variation in excess mortality of cancer patients using integrated nested Laplace approximation. *Statistics in Medicine*, v. 38, n. 4, p. 778–791, 2019.

SOUSA, I. A review on joint modelling of longitudinal measurements and time-to-event. *REVSTAT – Statistical Journal*, v. 9, n. 1, p. 57–81, 2011.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002.

SPOSTO, R. Cure model analysis in cancer: an application to data from the Children's Cancer Group. *Statistics in Medicine*, Wiley Online Library, v. 21, n. 2, p. 293–312, 2002.

SSEP. *Câncer de Mama*. 2019. Secretaria da Saúde do Estado do Paraná. Disponível em: <http://www.saude.pr.gov.br/modules/conteudo/conteudo.php?conteudo=3430>. Acesso em: 05/04/2019.

SU, L.; HOGAN, J. W. HIV dynamics and natural history studies: Joint modeling with doubly interval-censored event time and infrequent longitudinal data. *The Annals of Applied Statistics*, NIH Public Access, v. 5, n. 1, p. 400, 2011.

SUGUIURA, T. P. S. *Modelos multiníveis: gaussiano e multinomial*. Dissertação (Mestrado) — Universidade Estadual de Maringá, 2017.

SWEETING, M. J.; THOMPSON, S. G. Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, Wiley Online Library, v. 53, n. 5, p. 750–763, 2011.

TAQUETTE, S. R.; SOUZA, L. M. B. M. HIV-AIDS prevention in the conception of HIV-positive young people. *Revista de Saúde Pública*, v. 53, p. 80, 2019.

TOURNOUD, M.; ECOCHARD, R. Promotion time models with time-changing exposure and heterogeneity: Application to infectious diseases. *Biometrical Journal*, Wiley Online Library, v. 50, n. 3, p. 395–407, 2008.

TSIATIS, A. A.; DAVIDIAN, M. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, JSTOR, p. 809–834, 2004.

TSIATIS, A. A.; DEGRUTTOLA, V.; WULFSOHN, M. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association*, Taylor & Francis, v. 90, n. 429, p. 27–37, 1995.

VERBEKE, G.; MOLENBERGHS, G. *Linear Mixed Models for Longitudinal Data*. [S.l.]: Springer Science & Business Media, 2009.

WANG, X.; RYAN, Y. Y.; FARAWAY, J. J. *Bayesian Regression Modeling with INLA*. [S.l.]: Chapman and Hall/CRC, 2018.

WATANABE, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, v. 11, n. Dec, p. 3571–3594, 2010.

WEI, J.; WU, J. Cancer immunotherapy trial design with cure rate and delayed treatment effect. *Statistics in Medicine*, 2019. Early View.

WEST, B. T.; WELCH, K. B.; GALECKI, A. T. *Linear Mixed Models: A practical guide using statistical software*. [S.l.]: Chapman and Hall/CRC, 2014.

WULFSOHN, M. S.; TSIATIS, A. A. A joint model for survival and longitudinal data measured with error. *Biometrics*, JSTOR, p. 330–339, 1997.

YAKOVLEV, A. Y.; TSODIKOV, A. D. *Stochastic Models of Tumor Latency and Their Biostatistical Applications*. [S.l.]: World Scientific, Singapore, 1996.

YU, M.; TAYLOR, J. M. G.; SANDLER, H. M. Individual prediction in prostate cancer studies using a joint longitudinal survival–cure model. *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 481, p. 178–187, 2008.

ZHANG, H.; HUANG, Y.; WANG, W.; CHEN, H.; LANGLAND-ORBAN, B. Bayesian quantile regression-based partially linear mixed-effects joint models for longitudinal data with multiple features. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 28, n. 2, p. 569–588, 2019.