Universidade Estadual de Maringá
Programa de Pós-graduação em Bioestatística

BEATRIZ REGINA BRUM

Modelo Conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV

Dissertação de Mestrado

Maringá - Paraná
2018

BEATRIZ REGINA BRUM

Modelo Conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá como requisito parcial para obtenção do título de Mestre em Bioestatística.
Orientadora: Profª. Drª. Isolde Previdelli

Maringá - Paraná

2018

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

Brum, Beatriz Regina
B893m Modelo conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV / Beatriz Regina Brum. -- Maringá, PR, 2018.
71 f.: il. color.

Orientadora: Profª. Drª. Isolde Previdelli.
Dissertação (mestrado) - Universidade Estadual de Maringá, Programa de Pós-Graduação em Bioestatística, 2018.

1. Biomarcadores. 2. HIV - Coinfecção. 3. HBV - Coinfecção. 4. HCV - Coinfecção. I. Previdelli, Isolde, orient. II. Universidade Estadual de Maringá. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 23.ed. 616.9792

Márcia Regina Paiva de Brito – CRB-9/1267

BEATRIZ REGINA BRUM

Modelo Conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV

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 - UEM

Prof. Dr. Giovani Loiola da Silva
Instituto Superior Técnico
Universidade de Lisboa

Profª. Dra. Rosangela Getirana Santana
Universidade Estadual de Maringá - UEM

Maringá, 21 de março de 2018.

Este trabalho é dedicado às pessoas que contribuíram de forma direta ou indireta com meu nível de conhecimento.

Agradecimentos

À Profª. Drª. Isolde T. S. Previdelli, pela coragem, determinação e iniciativa de trazer o curso de mestrado em Bioestatística para a Universidade Estadual de Maringá. Curso este, que me possibilitou ingressar e concluir esta etapa tão importante e recompensadora de minha vida. Agradeço a orientação, o incentivo ao longo do curso, a amizade, a confiança que depositou em minha pessoa, tornando-me mais confiante em minhas próprias percepções. Por fim, agradeço por ser essa pessoa que sempre acredita no próximo e no amanhã.

À Profª. Drª. Rosangela G. Santana, pelas verdades ditas, pela paciência em ensinar, por sempre se mostrar disposta a atender e tirar dúvidas, características essas que contribuiram muito com meu aprimoramento intelectual.

Ao colega e amigo, Prof. Dr. Omar Cleo Neves Pereira, por todo o auxílio computacional, todo o tempo empregado a me ajudar nestes dois anos de mestrado.

À CAPES, pelo financiamento cedido, a qual me possibilitou dedicação exclusiva à realização desta pesquisa.

À Universidade Estadual de Maringá e aos professores do Departamento de Bioestatística, os quais de forma magistral propiciaram-me conhecimento e condições para a conclusão deste trabalho.

Aos colegas e amigos, em especial a Aline Edilaine Medeiros, Felipe Emanoel Barletta Mendes, Tiago Peres da Silva Suguiura, Vinicius Garcia Tosta pela troca de conhecimento e o apoio.

À Claudinei Mesquita da Silva, por fornecer os dados para que o estudo fosse realizado e ao auxílio na área biológica.

Aos professores da UTFPR - Pato Branco, por me fazerem acreditar que o mestrado era algo possível.

"Em seu caminho para a ciência moderna, os
seres humanos descartaram o significado.
O conceito é substituído pela fórmula,
a causa por regras e probabilidade."
THEODOR W. ADORNO
- Dialética do Iluminismo -

Resumo

Estudos que têm em vista determinar associações entre respostas longitudinais e de sobrevivência frequentemente são analisados de forma separada. Com o avanço teórico estatístico e computacional permitiu-se analisar conjuntamente ambos os processos, sendo possível obter estimativas mais eficientes no processo longitudinal e de menor viés no modelo de sobrevivência. Propõem-se nesta pesquisa abordar a metodologia de modelos conjuntos, assumindo para a resposta longitudinal um modelo misto de efeitos aleatório normal multivariado e para a sobrevivência o modelo de riscos proporcionais. Neste estudo foram utilizados dados de uma coorte retrospectiva obtida no estado do Paraná/Brasil, entre os anos de 2002 a 2016, de indivíduos portadores da Imunodeficiência Humana (HIV), coinfectados pelos vírus HBV e HCV expostos à terapia antirretroviral combinada (CART). O planejamento de pesquisa foi composto por três grupos: controle (HIV), coinfecção por HBV e HCV. A distribuição conjunta do tempo de início da CART até o atingir a razão entre as células T CD4 (auxiliares) e CD8 (citotóxicas) com a contagem de células T CD4 em relação a essa razão foi modelada. As análises evidenciaram que há associação entre a contagem de CD4 e o risco de atingir a razão igual 0,9 e, sob o aspecto longitudinal, o grupo HBV não apresentou diferença na contagem de CD4 quando comparada ao controle, o que não lhe garantiu maior risco de atingir uma razão igual 0,9. O modelo conjunto mostrou-se satisfatório para o ajuste dos dados, indicando que do ponto de vista longitudinal a coinfecção por HCV apresentou maior severidade, enquanto no aspecto de risco a coinfecção por HBV mostrou pior índice prognóstico apontando uma maior vulnerabilidade de seu sistema imune.

Palavras-chave: Biomarcadores; Modelo conjunto; HIV; Coinfecção por HBV e HCV.

Abstract

Studies that aimed to determine associations between longitudinal and survival responses used to be analyzed separately. With the theoretical and computational advances of statistics, it allowed the joint analyzes of both process, making it possible to obtain more efficient and less biased estimates for both longitudinal and survival estimates. We propose in this study, to approach *Joint Models* assuming, for the longitudinal response, a multivariate mixed effects model and, for the survival response, the Cox proportional hazard model. In this study, were used data from a retrospective cohort research obtained in the state of Paraná/Brazil, between 2002 and 2014, of individuals bearing the human immunodeficiency vírus (HIV), coinfected with Hepatitis and exposed to combination antiretroviral therapy (cART). The research planning was composed by three groups: mono-infection (HIV), coinfection by Hepatitis B and by Hepatitis C. We modelled the joint distribution of the cART initial time, until it reaches the minimal lymphocytes ratio T (LT) CD4:CD8 and the LT CD4 count, regarding this ratio. The analysis has evidenced that there are associations between the LT CD4 and the risk of overcoming the minimal ratio considered and, under the longitudinal aspect, Hepatitis B does not present differences in the LT CD4 count, when compared with the mono-infection, which does not guarantee a risk increase of reaching a ratio considered healthy. The joint model has shown to be satisfactory for this data adjustment, which indicated the worst prognostics for the coinfection with Hepatitis B and higher vulnerability of its immune system.

Key words: Biomarkers; Joint Model; HIV; Coinfection by HBV and HCV.

Lista de figuras

Figura 1 – Representação do efeito aleatório no intercepto do i-ésimo indivíduo sendo representado por $b_{i0} = \beta_0 - \beta_{i0}$. 25
Figura 2 – Representação dos efeitos aleatórios no intercepto ($b_{i0} = \beta_0 - \beta_{i0}$) e na inclinação ($b_{i1} = \beta_1 - \beta_{i1}$) do i-ésimo indivíduo, sendo $\boldsymbol{b_i}$ o vetor de efeitos aleatórios. 26
Figura 3 – Representação gráfica de censura à esquerda. 32
Figura 4 – Representação gráfica de censura à direita. 32
Figura 5 – Representação gráfica de censura intervalar. 33
Figura 6 – Tempo de sobrevivência de seis indivíduos, os segmentos horizontais representam a trajetória dos indivíduos, $\times$ representa a ocorrência do evento e $\circ$ censura. 37
Figura 7 – Representação do método numérico por meio das tangentes em duas dimensões. 40
Figura 8 – Distribuição da variável CD4 54
Figura 9 – Gráfico de perfis da raiz quadrada da contagem de CD4 para os diferentes grupos de coinfecção 54
Figura 10 – Representação dos perfis médios dos dados observados. 55
Figura 11 – Boxplot da raiz quadrada da contagem de CD4 para os diferentes grupos de coinfecção 56
Figura 12 – (a) Curvas de sobrevivência estimadas por Kaplan-Meier segundo tipo de grupo 56
Figura 13 – Gráficos de resíduos 60

Lista de tabelas

Tabela 1 – Resumo dos dados contendo as 5 primeiras linhas. 53
Tabela 2 – Os valores do AIC, BIC e log Verossimilhança para os modelos Mod1 e Mod2. 57
Tabela 3 – Estimadores, erros padrão e valor p para o modelo proposto e, os D_{ij} denota o ij-elemento da matriz de covariância dos efeitos aleatórios. . . 59

Sumário

1 Modelo Conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV . . . 13

1.1 Introdução . . . 13

1.2 Delineamento e Metodologia . . . 13

1.2.1 Cenário Imunológico . . . 13

1.2.2 Cenário da coinfecção por HBV e HCV . . . 14

1.2.3 Delineamento Experimental . . . 15

1.2.4 Objetivos do Experimento . . . 15

1.2.5 Organização do Trabalho . . . 16

2 Análise de Dados Longitudinais . . . 17

2.1 Introdução . . . 17

2.2 Caracterização de Modelos Mistos . . . 18

2.2.1 Estrutura de Covariância . . . 21

2.3 Estimação de Parâmetros . . . 22

2.4 Inferência . . . 26

2.4.1 Teste da Razão de Verossimilhança . . . 26

2.4.2 Teste de Wald . . . 27

2.5 Critérios de Informação . . . 28

2.6 Análise de Resíduos e Diagnóstico . . . 28

2.6.1 Resíduos Marginais . . . 29

2.6.2 Resíduos Condicionais . . . 29

2.6.3 Resíduos de Efeitos Aleatórios (EBLUP) . . . 29

2.6.4 Análise de Influência . . . 30

3 Análise de Sobrevivência . . . 31

3.1 Introdução . . . 31

3.2 O Tempo . . . 31

3.3 Censura . . . 32

3.4 Função de Sobrevivência . . . 33

3.5 Função de Risco . . . 33

3.6 Modelo semiparamétrico de Cox . . . 34

3.6.1 Estimação de parâmetros . . . 36

3.6.2 Função Escore . . . 39

3.6.3 Intervalo de Confiança IC . . . 40

3.6.4 O Risco Relativo-HR . . . 40

3.6.5 Estimação da Função de Sobrevivência Basal ($\lambda_0(t)$) . . . 41

3.7 Análise de Resíduos e Diagnóstico . . . 41

3.7.1 Resíduos de Schoenfeld 41
3.7.2 Resíduos Martingale 42
3.7.3 Resíduos $Deviance$ 43
3.7.4 Resíduos escore ($Dfbetas$) 43
3.8 Extensão do modelo de Cox 44
4 Modelo Conjunto 45
4.1 Estimação 46
4.1.1 Problemas de convergencia 48
4.2 Inferência 49
4.2.1 Teste da razão de verossimilhança 49
4.2.2 Teste de Wald 49
4.2.3 Intervalos de confiança 50
4.3 Análise de Resíduos e Diagnóstico 51
4.3.1 Resíduos marginais 51
4.3.2 Resíduos condicionais 51
4.3.3 Resíduos martingale 51
5 Aplicação à coorte de HIV coinfectada pelos vírus HBV e HCV 53
5.1 Análise descritiva dos dados 53
5.2 Modelagem conjunta 57
5.2.1 Seleção do modelo 57
5.2.2 Submodelo longitudinal 57
5.2.3 Submodelo de sobrevivência 58
5.2.4 Análise de Resíduos 59
5.2.5 Discussões 61
5.3 Considerações Finais 62
6 Apêndice A 63
Referências 67

1 Modelo Conjunto para um estudo de coorte de HIV coinfectada pelos vírus HBV e HCV

1.1 Introdução

Uma das primeiras ideias de modelo conjunto da história foi proposto por Wulfsohn e Tsiatis (1997), em que o tempo de sobrevivência pode ser estimado utilizando informações longitudinais com covariáveis dependentes do tempo e função de risco basal constante. Na sequência, Henderson et al. (2000) propuseram um modelo normal multivariado para a resposta longitudinal, assumindo uma estrutura de covariância AR(1) e o modelo semi-paramétrico de riscos proporcionais para a sobrevivência. Nas últimas décadas, modelos que relacionam respostas de processos distintos vêm sendo amplamente utilizados, principalmente no que diz respeito a relacionar o processo longitudinal e de sobrevivência. Tais estudos foram motivados principalmente por ensaios clínicos de HIV/AIDS (IBRAHIM et al., 2010).

Trabalhos de auto impacto científico podem ser vistos na literatura, entre eles estão Rizopoulos (2010), ao implementar no `R` uma biblioteca que facilita a utilização dessa metodologia, Sweeting e Thompson (2011) ao avaliarem dados de aneurisma da aorta abdominal, Su e Hogan (2011) com estudos da dinâmica natural do HIV, Asar et al. (2015), com pesquisa voltada à nefrologia, a implementação da macro JMFit no `SAS` por Zhang et al. (2016), Crowther et al. (2016) que aplicaram a metodologia a dados de de câncer de mama, entre outros. Nesses casos especificos para o aspecto longitudinal são necessária medidas obtidas ao longo do tempo (longitudinais), os marcadores biológicos, que ao serem extraídas do indivíduos assumem função identificadora de alterações no processo imune e metabólico, conforme visto em Rizopoulos (2012).

1.2 Delineamento e Metodologia

1.2.1 Cenário Imunológico

Entre os biomarcadores longitudinais mais populares na literatura voltados para a área da saúde estão os marcadores linfócitos T CD4 e CD8 da Imunodeficiência Humana (HIV). A principal característica da infecção pelo vírus do HIV é a redução das células CD4, responsáveis por combater infecções oportunistas. A depleção dessas células na corrente sanguínea fragiliza as defesas do organismo deixando-o susceptível a infecções

oportunistas.

A contagem do CD4 é considerada um dos marcadores biológicos mais eficientes na avaliação da progressão do HIV/AIDS e um indicador para o início da terapia antirretroviral (LU et al., 2015). O risco no diagnóstico de indivíduos com HIV positivo pode assumir vários aspectos, como o tempo até a morte (RIZOPOULOS, 2010), a supressão da carga viral (SU; HOGAN, 2011) e a mortalidade relacionada a transplante (TERRIN et al., 2015).

De acordo com Lanier (1995), ao comparar grupos não-infectados e infectados assintomáticos HIV, a razão CD4:CD8 é um indicador muito mais sensível da infecção por HIV do que a própria contagem absoluta de CD4. Essa razão auxilia no acompanhamento terapêutico, contribuindo para a avaliação imunológica dos pacientes tratados em seguimento de longo prazo e pode ser aplicada no monitoramento da disfunção imune (LU et al., 2015). O marcador em indivíduos com sistema imune considerado normal é de aproximadamente 2:1 (intervalo 1:1 a 3:1) (BAIN, 2014).

De certa forma os trabalhos publicados não apresentam concordância em afirmar se a contagem de CD4 deve ou não ser considerada como preditora da razão CD4:CD8, mas a mesma pode ser usada no acompanhamento da ativação imune (KUNIHOLM et al., 2016). Um estudo recente mostrou que o aumento do nível de CD4 isoladamente não pode ser considerado fator de recuperação do sistema imune, mas que seus altos níveis combinados com razão CD4:CD8 $\geq$ 0,8 são indicativos de um sistema imune ideal (LEE et al., 2017).

1.2.2 Cenário da coinfecção por HBV e HCV

Com a AIDS deixando de assumir o status de doença fatal, devido ao sucesso das terapias antirretrovirais, novos eventos começaram a gerar preocupação por parte dos profissionais da saúde, no que diz respeito ao agravamento do quadro clínico de pacientes HIV positivo.

Nos últimos anos, complicações clínicas vêm sendo causa de hospitalizações e óbitos de indivíduos portadores do HIV, entre elas estão as hepatopatias (insuficiência hepática crônica, cirrose e hepatocarcinoma) (BONACINI; PUOTI, 2000; VIRAIS, 2002; SHERMAN, 2004; DEEKS, 2011). Esse fato, combinado a supressão da carga viral e, consequentemente, a inibição da AIDS, tem motivado pesquisas com foco na mortalidade por questões não relacionadas à AIDS (DEEKS et al., 2013).

Entre 36,7 milhões de pessoas portadoras do HIV estima-se que 2,7 milhões estejam coinfectados por HBV e 2,3 milhões por HCV (ORGANIZATION et al., 2017). A mortalidade relacionada às hepatites virais entre os anos de 2000 a 2015 aumentou de 1,10 milhão para 1,34 milhão, correspondendo a um aumento de 22% (ORGANIZATION et al., 2017).

No Brasil, as estimativas mostram que cerca de 70% a 90% dos pacientes com HIV ou AIDS possuem algum marcador sorológico da hepatite B. A hepatite C, também ocorre

em um grande número de pessoas infectadas pelo HIV, atingindo em média 40% (VIRAIS, 2002). Entre os períodos de 2007 a 2016, foram notificados 212.031 casos confirmados de hepatite B, 5,2% apresentavam coinfecção com HIV, no mesmo período foram confirmados 149.537 casos de hepatite C, entre os quais se estimou que 9,8% apresentavam coinfecção com o HIV (SAÚDE, 2017).

No período de 2007 a 2016, foram notificados 212.031 casos confirmados de hepatite B no Brasil, 5,2% apresentavam coinfecção com HIV. Entre os casos confirmados de hepatite C (149.537) no mesmo período, estima-se que 9,8% apresentam coinfecção com o HIV (SAÚDE, 2017).

1.2.3 Delineamento Experimental

Os dados do estudo são oriundos de uma coorte retrospectiva, de indivíduos diagnosticados com HIV, coinfectados por HBV e HCV, no período compreendido entre 01 janeiro de 2002 a 30 de junho de 2016, a partir de prontuários médicos do Centro Especializado de Doenças Infecciosas e Parasitárias (CEDIP), do Centro de Testagem em Cascavel/PR e Aconselhamento (CTA), em Maringá/PR, ambos vinculados ao Sistema Único de Saúde (SUS).

O número de indivíduos da amostra foi 147, todos submetidos a terapia antirretroviral (CART), sendo divididos em três grupos: i) grupo controle (controle), $n_{contr} = 108$, coinfecção por HBV, $n_B = 21$ e coinfecção por HCV com $n_C = 18$. O controle teve 74 casos censurados e 34 eventos, na coinfecção por HBV, 18 censuras e 3 eventos, e HCV, 10 censuras e 8 eventos.

Devido a coleta dos dados ter sido obtida diretamente dos prontuários, em que os registros informavam somente os intervalos de ocorrência (I, S) onde I < T < S. Como estratégia para contornar o problema do evento de interesse ocorrer entre duas observações consecutivas, foi utilizado o seguinte critério: Quando o evento não foi verificado no limite inferior e sim no limite superior assumiu-se como tempo de evento I. As censuras ocorridas no estudo são de caráter não informativo, mais detalhes podem ser vistos em (TABLEMAN; KIM, 2003).

1.2.4 Objetivos do Experimento

O foco principal desta pesquisa foi investigar a existência de associação entre a contagem de CD4 ao longo do tempo e o risco dos diferentes grupos em atingir uma razão CD4:CD8 considerada adequada. Buscou-se evidenciar entre os grupos coinfectados quais apresentaram redução no risco do evento. Neste estudo foi considerado como evento a razão CD4:CD8 $=$ 0,9, obtida através do ponto médio entre a razão dada pelo estudo realizado por Lee et al. (2017) e a razão preconizada pela maioria dos estudos (CD4:CD8

=1). Para a análise foi modelada a distribuição conjunta da contagem de CD4 do tempo de início do estudo, até o tempo de ocorrência da razão igual a 0,9. O modelo conjunto utilizou para o submodelo longitudinal um modelo misto normal multivariado, em que o biomarcador longitudinal é representado pela raiz quadrada da contagem de CD4, variável resposta do processo, sendo esta descrita pelos grupos (controle, HBV e HCV) e tempo (dias). As trajetórias não lineares da contagem de CD4 foram modeladas por meio de B-*splines*, devido alguns perfis terem apresentado comportamento não linear. Para o submodelo de sobrevivência foi utilizado o modelo de riscos proporcionais, considerando como resposta, a razão igual a 0,9, e como preditor linear os grupos.

1.2.5 Organização do Trabalho

O presente estudo está organizado no seguinte formato: O segundo e terceiro capítulos trazem uma revisão de literatura, que contempla uma breve abordagem de Modelos Mistos e Análise de Sobrevivência. O quarto capítulo aborda a metodologia de Modelo Conjunto. No quinto capítulo será apresentada a estrutura do modelo conjunto aplicado aos dados de infecção pelo HIV coinfecção por HBV e HCV, onde são apresentadas a análise descritiva dos dados, a seleção do modelo, o modelo proposto, os resultados, análise de resíduos, discussões e considerações finais acerca do problema.

2 Análise de Dados Longitudinais

2.1 Introdução

De acordo com Rizopoulos (2012), dados provenientes de estudos longitudinais podem ser definidos como dados resultantes de observações coletadas de indivíduos, as quais são repetidamente realizadas ao logo do tempo. Assim, cada indivíduo pode ser visto como um cluster, tendo em intervalos de tempo pré-determinados observações coletadas nos t períodos de tempo. Dessa forma, sempre que o tempo for um fator importante, as medidas repetidas recebem o nome de dados longitudinais, que têm por objetivo diagnosticar as diferenças que podem ocorrer ao longo do período de intervenção na variável resposta, sendo possível verificar o efeito dentro e entre indivíduos.

Estudo de ensaio clínico ou coorte, em que o objetivo do estudo é comparar efeito de tratamentos ou exposição a um fator, os n indivíduos são selecionados de uma população com o mesmo estado de interesse, uma doença ou condição médica. A coleta da primeira observação se refere ao tempo zero (linha de base), momento em que nenhum dos elementos da amostra sofreu intervenção em relação ao tratamento. Os indivíduos são considerados independentes entre si e as medidas repetidas são vistas como aninhadas dentro do indivíduo, evidenciando a estrutura de correlação desse delineamento.

Dentre as ferramentas utilizadas para resolver problemas de dados longitudinais destacam-se os Modelos Mistos. Sua abordagem permite que se façam suposições a respeito do padrão de correlação entre as medidas dentro do sujeito e entre sujeitos. Essas suposições podem assumir diversas formas de estrutura, das mais simples às mais complexas. Esses modelos têm como finalidade absorver maior variabilidade em relação a alguns modelos, aumentando a precisão das estimativas e, portanto, se aproximando mais do comportamento real dos dados.

Um amplo conjunto de pesquisas, com ênfase para a utilização dessa metodologia, pode ser encontrado na literatura. Entre elas podemos citar Waternaux et al. (1989), que abordaram efeitos aleatórios para modelar o termo de erro na análise de desenvolvimento cognitivo de crianças expostas a chumbo, no período pré e pós-natal, considerando o desenvolvimento de baixo a moderado nos primeiros 18 meses de vida. Foi verificado que crianças com alta concentração de chumbo nos 6 primeiros meses têm menor desempenho que crianças com alta concentração aos 18 meses. Pearson et al. (1994) ajustaram um modelo linear de efeitos mistos a dados longitudinais de casos controles em um estudo longitudinal de Baltimore, mostrando que casos de câncer podem ser caracterizados por uma taxa extremamente alta da mudança nos níveis de Antígeno Prostático Específico (PSA).

Foulley et al. (2000) mostraram como o algoritmo EM é uma ferramenta poderosa para obter as estimativas de máxima verossímilhança, permitindo calcular de forma separada os parâmetros da matriz R e G. Foram calculadas estimativas REML para componente de variância e covariância em um modelos mistos gaussianos para análise do crescimento facial de um conjunto de meninos e meninas com quatro diferentes idades. Fausto et al. (2008) ajustaram um modelo linear de efeitos mistos para estimar a velocidade de crescimento de crianças nascidas de mulheres infectadas pelo HIV, o acompanhamento ocorreu no ambulatório de AIDS pediátrico do Hospital das Clínicas da Universidade Federal de Minas Gerais, onde foi constatado a eficiência do modelo em descrever o comportamento longitudinal e a velocidade do crescimento. Ribeiro et al. (2016), por meio de um modelo misto, relatam a eficácia do tratamento à base de óleo de peixe (OP) na redução dos efeitos cognitivos causados pela isquemia cerebral. De acordo com Brown e Prescott (2014), os modelos mistos são capazes de lidar com dados em que as observações não são independentes, sendo possível modelar a estrutura de covariância dos mesmos, proporcionando grande flexibilidade na análise.

2.2 Caracterização de Modelos Mistos

Os Modelos Mistos recebem esse nome devido a sua estrutura ser composta de uma parte fixa e a outra aleatória, exceto o erro aleatório experimental. A ideia central sugere que a distribuição de probabilidade para as observações dentro de um cluster (indivíduo), tem a mesma forma geral para todos os clusters, ou seja, que o comportamento médio dos sujeitos é o mesmo (SCHABENBERGER; PIERCE, 2001). No entanto, o comportamento individual inerente a cada sujeito apresenta certa dependência, a qual pode ser modelada como um efeito aleatório em que a estrutura de correlação inerente ao indivíduo é captada ou modelada.

Para fins didáticos, será considerada primeiro a parte fixa do modelo de regressão múltipla multivariada como modelo marginal. Seja y_{ij} a sequência de observações realizadas para o i-ésimo sujeito no j-ésimo tempo de observação com $i = 1, ..., n$ e $j = 1, ..., n_i$, X_i $(n_i \times p)$ a matriz das n_i observações do i-ésimo indivíduo nas p variáveis explicativas , que podem ser tanto categóricas como numéricas, o modelo de regressão é denotado por,

$$y_{ij} \;=\; X_{1ij}\beta_1 + \cdots + X_{pij}\beta_p + \epsilon_{ij}, \tag{2.1}$$

onde ϵ_{ij}, são a sequência de erros aleatórios.

Assim a representação vetorial da resposta do i-ésimo indivíduo é escrita por,

$$\boldsymbol{Y_i} \;=\; \boldsymbol{X_i\beta} + \boldsymbol{\epsilon_i}, \tag{2.2}$$

em que $\boldsymbol{Y_i} = (y_{i1}, y_{i2}, \cdots, y_{in_i})^T$ corresponde as n_i observações do i-ésimo indivíduo e $\boldsymbol{X_i}$ a matriz de covariáveis fixas expressa por

$$\boldsymbol{X_i} = \begin{pmatrix} X_{11} & \dots & X_{1p} \\ X_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ X_{n_i1} & \dots & X_{n_ip} \end{pmatrix}_{n_i \times p} .$$

Sendo esta a matriz de p covariáveis explicativas das n_i observações do i-ésimo indivíduo.

O modelo marginal pode ser expresso em notação matricial como $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. A matriz $\boldsymbol{Y} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, \cdots, \boldsymbol{Y_n})^T$ de ordem ($\sum_{i=1}^{n} n_i$ x 1), determina o conjunto completo de respostas, sendo $\sum_{i=1}^{n} n_i$ o número total de observações dos n indivíduos descrito por

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \\ \vdots \\ \boldsymbol{Y_n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y_1} = \begin{pmatrix} y_{11} & \dots & y_{1n_1} \end{pmatrix}^T \\ \boldsymbol{Y_2} = \begin{pmatrix} y_{21} & \dots & y_{2n_2} \end{pmatrix}^T \\ \vdots \\ \boldsymbol{Y_n} = \begin{pmatrix} y_{n1} & \dots & y_{nn_n} \end{pmatrix}^T \end{pmatrix} .$$

Em modelos que possuem intercepto, a primeira coluna de $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ é composta pela unidade. Sendo essa uma matriz em bloco de ordem (n x p). O vetor de coeficientes da regressão é representado por $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ com dimensão p. A variância do modelo é dada por $\boldsymbol{V} = Var(\boldsymbol{Y}) = \sigma_e^2 I$, assumindo estrutura independente em que a correlação é igual a zero e I corresponde a matriz identidade.

Segundo Brown e Prescott (2014), o vetor de $\boldsymbol{\beta's}$, corresponde ao vetor com dimensão igual a dimensão das colunas da matriz $\boldsymbol{X}$, a qual traz todas as informações, como idade, sexo, por fim, medidas da linha de base. Variáveis categóricas representadas por zeros e uns também podem estar contidas na matriz, conhecida normalmente como matriz modelo. Ao incorporar efeitos aleatórios ($\boldsymbol{b}$) ao modelo marginal é possível obter o modelo condicional (SEARLE et al., 2009), a estrutura desse modelo para o i-ésimo indivíduo no j-ésimo tempo é expressa por

$$y_{ij} = \boldsymbol{X_{ij}\beta} + \boldsymbol{Z_{ij}b_i} + \varepsilon_{ij}, \tag{2.3}$$

onde a resposta, y_{ij}, corresponde ao vetor de dimensão n_i, $\boldsymbol{X_{ij}}$ e $\boldsymbol{Z_{ij}}$ são matrizes conhecidas, em que, $\boldsymbol{Z_{ij}}_{(n_i \times q)}$ pode ser uma submatriz de $\boldsymbol{X_{ij}}_{(n_i \times p)}$ e ε_{ij} de acordo com

Diggle (1988) é composto por três partes,

$$\varepsilon_{ij} = \sum_{k=1}^{k} z_{ij} b_{ik} + w_i(t_{ij}) + \epsilon_{ij}, \tag{2.4}$$

em que a primeira parcela corresponde ao efeito aditivo dos q efeitos aleatórios na regressão. O segundo é a contribuição do processo gaussiano, das observações que ocorrem no tempo, ou seja, cada ponto de tempo está associado a uma variável aleatória com distribuição normal. O terceiro é o erro de medição, assumindo que $\varepsilon_{ij} \sim N(0, \boldsymbol{V_i})$, onde a matriz de covariância $\boldsymbol{V_i} = \text{Var}(\boldsymbol{Y_i})$ é,

$$\boldsymbol{V_i} = \boldsymbol{Z_i}\text{Var}(\boldsymbol{b_i})\boldsymbol{Z_i}^T + \text{Var}(\boldsymbol{\epsilon_i}). \tag{2.5}$$

Sendo $\text{Var}(\boldsymbol{b_i}) = \boldsymbol{G}$, a matriz simétrica positiva que representa a variabilidade dos efeitos aleatórios, de ordem (q x q), no qual q, denota o número de efeito aleatório, supondo $\boldsymbol{b_i} \sim N(0, \boldsymbol{G})$, a matriz é escrita da seguinte forma,

$$\boldsymbol{G_i} = \begin{pmatrix} \text{Var}(b_{0i}) & \text{cov}(b_{0i}, b_{1i}) & \cdots & \text{cov}(b_{0i}, b_{qi}) \\ \text{cov}(b_{0i}, b_{1i}) & \text{Var}(b_{1i}) & \cdots & \text{cov}(b_{1i}, b_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(b_{0i}, b_{qi}) & \text{cov}(b_{1i}, b_{qi}) & \cdots & \text{Var}(b_{qi}) \end{pmatrix},$$

onde $\text{Var}(b_{ji})$, quando $j = 0, \cdots, q$, representam a variabilidade dos coeficientes da regressão linear com efeitos aleatórios e os demais são os termos de covariância.

De maneira análoga, $\text{Var}(\boldsymbol{\epsilon_i}) = \boldsymbol{R_i}$ de dimensão $(n_i \times n_i)$, onde cada elemento ϵ_i tem variância $\sigma^2 = Var(\epsilon_{n_i i})$, com $\boldsymbol{\epsilon_i} \sim N(0, \boldsymbol{R})$ e cada par de elementos têm covariância zero, uma vez que as respostas entre indivíduos são independentes, a matriz $\boldsymbol{R_i}$ é definida por,

$$\boldsymbol{R_i} = \begin{pmatrix} Var(\epsilon_{0i}) & cov(\epsilon_{0i}, \epsilon_{1i}) & \cdots & cov(\epsilon_{0i}, \epsilon_{n_i i}) \\ cov(\epsilon_{0i}, \epsilon_{1i}) & Var(\epsilon_{1i}) & \cdots & cov(\epsilon_{1i}, \epsilon_{n_i i}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\epsilon_{0i}, \epsilon_{n_i i}) & cov(\epsilon_{1i}, \epsilon_{n_i i}) & \cdots & Var(\epsilon_{n_i i}) \end{pmatrix}.$$

Dessa forma, é possível verificar que a variável resposta segue uma distribuição normal multivariada representada matricialmente por,

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zb_i}, \boldsymbol{V}). \tag{2.6}$$

Para modelar a estrutura de correlação $\boldsymbol{V}$, existente na variável resposta, diversas estruturas de correlação estão disponíveis nos aplicativos `R`, `SAS` entre outros.

2.2.1 Estrutura de Covariância

De acordo com Diggle (2002), a matriz de covariância deve ser suficientemente flexível para incluir diferentes variações aleatórias, como a variação devido a efeitos aleatórios, a variabilidade entre os tempos de medição e a variação dos erros de medida. Entre as estruturas de correlação de trabalho mais conhecidas temos:

- Simétrica Composta: Todas as observações dentro de um sujeito são igualmente correlacionadas. Um exemplo simples ocorre na piscicultura, considerando um tanque (cluster), em que os peixes são tratados com determinado tipo de ração espera-se que ao coletar quatro amostras do tanque para observação do peso dos peixes sejam observadas correlações iguais.

$$V_{ij} = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}_{n_i \text{x} n_i}$$

  O número de parâmetros a serem estimados se resume a dois.

- Auto-regressiva de Primeira Ordem $AR(1)$: A correlação entre um par de observações na mesma unidade decai quando o intervalo de tempo entre as medias aumenta.

$$V_{ij} = \sigma^2 \begin{pmatrix} 1 & \rho^1 & \dots & \rho^{n_i-1} \\ \rho^1 & 1 & \dots & \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-2} & 1 & \dots & \rho^1 \\ \rho^{n_i-1} & \rho^{n_i-2} & \dots & 1 \end{pmatrix}_{n_i \text{x} n_i}$$

  É natural que as covariâncias das medidas feitas entre tempos mais próximos sejam mais parecidas que as feitas em tempos mais distantes. Um exemplo desse tipo de estrutura pode ser visto em análises que procuram verificar o tempo de efeito de uma droga. Por exemplo, a dose injetada de medicamento nas primeiras observações apresenta uma alta concentração no organismo do paciente, na medida em que o tempo passa a dosagem no corpo diminui exponencialmente. Nessa estrutura apenas dois parâmetros devem ser estimados.

- Não Estruturada: É muito comum não se saber nada a respeito dos dados, dessa forma torna-se conveniente estimar todos os parâmetros da matriz de covariância. Nos casos que existem poucas observações ao longo do tempo não se pode supor que as correlações tenham um comportamento regressivo. Esta não é uma estrutura parcimoniosa, dessa forma $n_i(n_i+1)/2$ parâmetros precisam ser estimados (SCHABENBERGER; PIERCE, 2001).

$$V_{ij} = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 & \dots & \sigma_{1n_i}^2 \\ \sigma_{21}^2 & \sigma_2^2 & \dots & \sigma_{2n_i}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i1}^2 & \sigma_{n_i2}^2 & \dots & \sigma_{n_i}^2 \end{pmatrix}_{n_i \text{x} n_i} .$$

Mais modelos de estrutura de covariância para a variabilidade de $\hat{\boldsymbol{Y}}$, podem ser vistas em Singer e Andrade (1986) e Verbeke e Molenberghs (1997).

2.3 Estimação de Parâmetros

O método de máxima verissimilhança usual faz uso da máxima verossimilhança perfilada. Nesse caso supõem-se conhecer $\boldsymbol{G}$ sendo possível por meio de mínimos quadrados determinar as estimativas de β, conhecendo $\hat{\beta}$ utiliza-se a verossimilhança perfilada para os parâmetros de variância $\boldsymbol{G}$. As estimativas que compõem $\boldsymbol{G}$ dependem de $\hat{\beta}$, a variabilidade de $\hat{\beta}$ é obtida pela inversa da informação de Fisher, que por sua vez está condicionada à $\boldsymbol{G}$. Para estimar $\hat{\boldsymbol{G}}$ são necessários processos iterativos extremamente trabalhosos e que subestimam a variabilidade dos erros das estimativas.

Para Diggle (2002), a estimação de parâmetros por meio de verossimilhança usual gera estimativas viesadas para as predições da variância. Uma alternativa a esse problema é a utilização do método de máxima verossimilhança restrita.

A máxima verissimilhança restrita ou residual é obtida por meio de uma soma entre logaritmo de verossimilhanças. Embora essa frase pareça soar de maneira estranha essa ideia surge de forma natural. O processo se dá primeiramente assumindo-se conhecer a estrutura de variância e estimando $\hat{\beta}$, por mínimos quadrados ordinais. Tal estimativa permite obter informações a respeito da variância dos parâmetros, pois os erros são obtidos calculando a diferença entre os valores observados e valores estimados a partir de $\boldsymbol{X}\hat{\beta}$, parte fixa do modelo. Os resíduos assumem uma distribuição de probabilidade normal multivariada dada por, $\boldsymbol{\epsilon} \sim$ N$(0, \boldsymbol{\sigma^2 I})$, em que $\boldsymbol{R_i} = \boldsymbol{\sigma^2 I_{n_i}}$. A distribuição de probabilidade de $\hat{\beta}$ é independente em relação aos resíduos onde $\hat{\beta} \sim$ N$(\hat{\beta}, (\boldsymbol{X_i'V_i^{-1}X_i})^{-1})$, portanto, a

verossimilhança da amostra pode ser escrita por,

$$L(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{y}) = L(\boldsymbol{\tau}; \boldsymbol{\epsilon}) \times L(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \boldsymbol{\tau}), \tag{2.7}$$

com vetor $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)$, vetor $\boldsymbol{\tau}$ representando os diferentes parâmetros da matriz $\boldsymbol{R}$ e $\boldsymbol{G}$. Nos casos em que não são verificados efeitos aleatórios $\boldsymbol{\tau} = \boldsymbol{\sigma^2 I}$. Então o logaritmo da função de verossimilhança residual é,

$$-\ell(\boldsymbol{\tau}; \boldsymbol{\epsilon}) = \ell(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{y}) - \ell(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \boldsymbol{\tau}), \tag{2.8}$$

onde

$$\ell(\boldsymbol{\beta}, \boldsymbol{\tau}; \boldsymbol{y}) = -1/2 \left\{ n \times n_i \ln(2\pi) + \sum_{i=1}^{n} \ln |\boldsymbol{V_i}| + \sum_{i=1}^{n} \left(\boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}} \right)' \boldsymbol{V_i}^{-1} \left(\boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}} \right) \right\}$$

e

$$\ell(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}, \boldsymbol{\tau}) = 1/2 \left\{ p \log(2\pi) + \ln |(\sum_{i=1}^{n} \boldsymbol{X_i'} \boldsymbol{V_i^{-1}} \boldsymbol{X_i})^{-1}| + \underbrace{\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)' (\boldsymbol{X_i'} \boldsymbol{V_i^{-1}} \boldsymbol{X_i})^{-1} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}} \right)}_{JJ} \right\}.$$

A parcela JJ pode ser excluída pois é apenas um termo a mais que em nada contribui, então a log verossimilhança restrita se resume a

$$\ell_{re}(\boldsymbol{\tau}; \boldsymbol{\epsilon}) = \frac{1}{2} \sum_{i=1}^{n} \ln |\boldsymbol{V_i}| + \frac{1}{2} \sum_{i=1}^{n} \left(\boldsymbol{Y_i} - \boldsymbol{X_i}\hat{\beta} \right) \boldsymbol{V_i}^{-1} \left(\boldsymbol{Y_i} + \boldsymbol{X_i}\hat{\beta} \right) + \frac{1}{2} \ln |\sum_{i=1}^{n} \boldsymbol{X_i'} \boldsymbol{V_i}^{-1} \boldsymbol{X_i}|. \tag{2.9}$$

Fitzmaurice et al. (2004) enfatizam que a principal ideia por trás da estimação via máxima verossimilhança restrita é utilizar uma parte dos dados para estimar a variância $\boldsymbol{\tau}$, onde a mesma é utilizada para a estimação de $\boldsymbol{\beta}$, de modo que a verossimilhança residual seja exclusivamente dependente de $\boldsymbol{\tau}$.

Dessa forma, devido ao determinante carregar um resumo da variabilidade de $\boldsymbol{\beta}$, ao maximizar a função log verossimilhança se obtem o estimador $\hat{\boldsymbol{\tau}}$, não viesado. A matriz $\hat{\boldsymbol{V}}_i$ composta por $\hat{\boldsymbol{\tau}}$ fica completamente especificada, permitindo determinar $\hat{\boldsymbol{\beta}}$ expresso por

$$\hat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{n} \boldsymbol{X_i'} \hat{\boldsymbol{V}}_i \boldsymbol{X_i} \right]^{-1} \sum_{i=1}^{n} \left(\boldsymbol{X_i'} \hat{\boldsymbol{V}}_i \boldsymbol{Y_i} \right). \tag{2.10}$$

Por meio da derivada de segunda ordem da função logarítmica de máxima verossimilhança é possível obter a variabilidade de $\hat{\boldsymbol{\beta}}$, dada por

$$Var(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X^T} \boldsymbol{V^{-1}} \boldsymbol{X})^{-1}. \tag{2.11}$$

Entretanto, os modelos mistos como próprio nome já diz não apresentam apenas parâmetros fixos, mas também de efeito aleatório, os quais podem ser preditos, mas não pelos métodos anteriormente citados. Para a predição dos efeitos aleatórios é utilizada a formulação da verossimilhança hierárquica proposta por Lee e Nelder (1996) e expressa por

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{b};\boldsymbol{\beta},\boldsymbol{\tau}) &= f_{Y|b}(\boldsymbol{y};\boldsymbol{\beta})f_b(\boldsymbol{b};\boldsymbol{\tau}) \\ &= \frac{1}{\sqrt{2\pi}^N det|\boldsymbol{R}|}exp\Big\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})\Big\} \\ &- \frac{1}{\sqrt{2\pi}^q det|\boldsymbol{G}|}exp\Big\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{Gb}\Big\}, \end{aligned} \tag{2.12}$$

a função densidade de probabilidade conjunta para variáveis observáveis e não observáveis, permite avaliar variáveis latentes ou aleatórias (HENRIK; POUL, 2011).

O logaritmo da verossimilhança conjunta é escrita como segue

$$\begin{aligned} \ell(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{b}) &= -\frac{1}{2}\log|\boldsymbol{R}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}) \\ &-\frac{1}{2}\log|\boldsymbol{G}| - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{G}^{-1}\boldsymbol{b}. \end{aligned} \tag{2.13}$$

Derivando o log da verossimilhança hierárquica (2.13) parcialmente em relação à $\boldsymbol{b}$, igualando a expressão 2.14 a zero, e resolvendo as equações com respeito à $\boldsymbol{b}$ tem-se,

$$\frac{\partial}{\partial \boldsymbol{b}}l(\boldsymbol{\beta},\boldsymbol{G},\boldsymbol{b}) = \boldsymbol{Z^T R^{-1}}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb}) - \boldsymbol{G}^{-1}\boldsymbol{b} \tag{2.14}$$

substituindo β por $\hat{\boldsymbol{\beta}}$ em 2.15, sendo conhecidos $\boldsymbol{G}$ e $\boldsymbol{R}$ obtemos $\hat{\boldsymbol{b}}$, definido como o melhor preditor linear imparcial ou não-enviesado ($BLUP$) de $\boldsymbol{b}$ e em relação a $\hat{\boldsymbol{\beta}}$ o melhor estimador não viesado ($BLUE$) de β (ROBINSON, 1991; SEARLE, 1971; MCLEAN et al., 1991).

$$\hat{\boldsymbol{b}} = \left[\boldsymbol{Z^T R^{-1} Z} - \boldsymbol{G}^{-1}\right]^{-1}\boldsymbol{Z^T R^{-1}}(\boldsymbol{y}-\boldsymbol{X}\beta), \tag{2.15}$$

sendo conhecidos $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{b}}$ é possível predizer $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{u}}$ (HENRIK; POUL, 2011).

De forma análoga à $\hat{\beta}$, a variabilidade de $\hat{\boldsymbol{b}}$ também é obtida a partir da segunda derivada do logaritmo da função de verossimilhança hierárquica, sendo escrita por

$$Var(\hat{\boldsymbol{b}}) = (\boldsymbol{Z^T R^{-1} Z} + \boldsymbol{G^{-1}})^{-1}. \tag{2.16}$$

Os efeitos aleatórios são uma característica específica dos sujeitos relacionados no estudo, cada um com suas particularidades específicas, vistas como variáveis latentes. Essa

aleatoriedade não pode ser explicada por covariáveis observáveis, as mesmas evidenciam a heterogeneidade entre e dentro dos sujeitos. Na sequência, são descritos dois modelos que servem de ilustração para compreender melhora essa latência.

Seja um estudo relacionado à evolução de uma covariável no tempo caracterizando um modelo misto gaussiano, que assume uma evolução média linear, com intercepto aleatório e mesma inclinação escrita por,

$$\begin{cases} y_{ij} = (\beta_0 + b_{i0}) + \beta_1 t_{ij} + \epsilon_{ij} \\ b_{i0} \sim N(0, \sigma_b^2) \\ \epsilon_{ij} \sim N(0, \sigma^2). \end{cases}$$

Por meio da Figura 1, é possível verificar o seguimento em negrito com intercepto β_0, correspondente à média longitudinal. O segmento cujo intercepto é β_{i0}, representa o intercepto do i-ésimo sujeito, de modo que, seja $b_{i0} = \beta_0 - \beta_{i0}$ o efeito aleatório do i-ésimo sujeito. Assim para amostras com n indivíduos se tem tantos efeitos aleatórios quanto sujeitos para este caso.

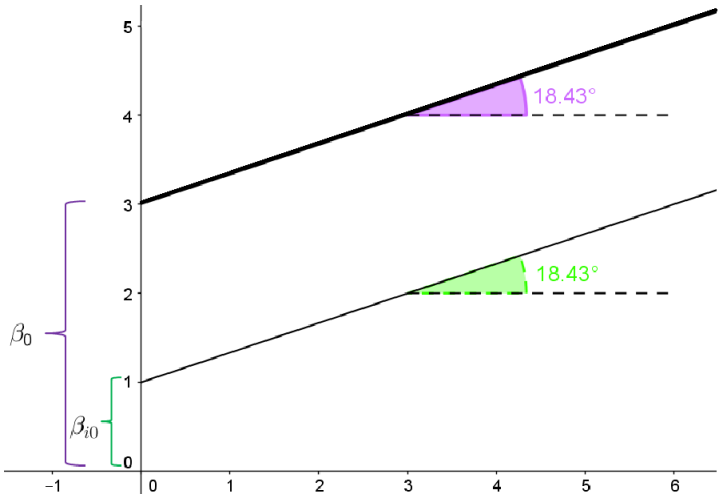

Figura 1 – Representação do efeito aleatório no intercepto do i-ésimo indivíduo sendo representado por $b_{i0} = \beta_0 - \beta_{i0}$.

O modelo anterior, embora de fácil entendimento, se apresenta em desvantagem devido ao fato de não levar em consideração as mudanças individuais das observações longitudinais, que podem alterar a inclinação de cada perfil. O modelo misto gaussiano com efeito aleatório tanto no intercepto quanto na inclinação, descreve indivíduos que além de apresentarem medidas distintas no início do estudo apresentam diferentes comportamentos

no passar do tempo, esse modelo é descrito por,

$$\begin{cases} y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_{ij} + \epsilon_{ij}. \\ \boldsymbol{b_i} \sim N(0, G) \\ \epsilon_{ij} \sim N(0, \sigma^2) \end{cases}$$

Veja que $\boldsymbol{b_i}$ representa um vetor de efeitos aleatórios, o modelo que apresenta esse vetor se torna muito mais flexível que o anterior, pois é natural supor que em medidas repetidas a correlação entre os pares de tempo assumam uma estrutura de correlação.

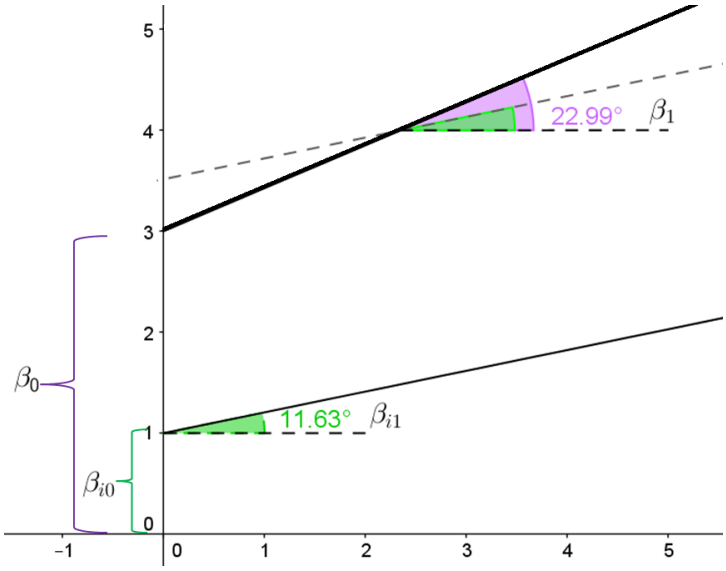

Figura 2 – Representação dos efeitos aleatórios no intercepto ($b_{i0} = \beta_0 - \beta_{i0}$) e na inclinação ($b_{i1} = \beta_1 - \beta_{i1}$) do i-ésimo indivíduo, sendo $\boldsymbol{b_i}$ o vetor de efeitos aleatórios.

Na Figura 2, o segmento em negrito representa a média longitudinal com efeito aleatório no intercepto dado por b_{i0} e na inclinação por $b_{i1} = \beta_1 - \beta_{i1}$. Isso é o mesmo que dizer que para cada n indivíduos é apresentado um vetor $\boldsymbol{b_i} = (b_{i0}, b_{i1})$ de efeitos aleatórios, o mesmo descreve as diferenças de indivíduo a indivíduo.

2.4 Inferência

2.4.1 Teste da Razão de Verossimilhança

Diversos são os métodos empregados para análise e escolha de um modelo. Considerando, $\boldsymbol{\tau} = (\sigma^2, \sigma^2_{b_{0i}}, \cdots, \sigma^2_{b_{qi}})$, o vetor de estimativas dos componentes de variância, a ideia da razão é descrever a distância entre o logaritmo da verossimilhança do modelo completo, sob H_0, $\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}})$ e o log da verossimilhança do modelo reduzido expresso por $\ell(\hat{\boldsymbol{\beta}}_{\mathbf{0}}, \hat{\boldsymbol{\tau}}_{\mathbf{0}})$.

O teste da razão de verossimilhança é usado basicamente na comparação de modelos encaixados. No caso dos modelos mistos, além dos parâmetros comuns da regressão, são comparados modelos de covariância, onde o modelo reduzido corresponde a um caso particular do modelo completo. Nesse caso, torna-se importante salientar que o interesse é testar uma determinada estrutura de correlação, o teste consiste em verificar qual dentre as estruturas melhor se adéqua ao comportamento dos dados.

A estatística da razão de verossimilhança é dado por

$$RL = 2\left[\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) - \ell(\hat{\boldsymbol{\beta}}_{\mathbf{0}}, \hat{\boldsymbol{\tau}}_0)\right], \tag{2.17}$$

com distribuição assintótica qui-quadrado, χ^2_{gl}, cujos graus de liberdade correspondem a diferença entre o número de parâmetros a serem testados.

Entretanto, a distribuição nula assintótica qui-quadrado para o teste da razão de verossimilhança não é válido em todos os casos. Os testes de hipótese para a significância dos efeitos aleatórios, são baseados na normalidade assintótica dos estimadores de máxima verossimilhança restrita. Quando os valores da hipótese nula encontram-se na fronteira da hipótese alternativa, violam o pressuposto de normalidade assintótica dos estimadores (COX; HINKLEY, 1979).

Stram e Lee (1994) mostraram que a distribuição assintótica para testar a significância dos efeitos aleatórios, em muitos casos, ao invés de ser uma única qui-quadrado é uma mistura de qui-quadrados. Em West et al. (2014), podem ser vistos exemplos que apresentam uma alternativa a esse problema de fronteira, nos quais testes de razão de verossimilhança com base na verossimilhança restrita para parâmetros de covariância são realizados. Nesses casos não se testa especificamente os efeitos aleatórios, mas sim a variância e covariância.

2.4.2 Teste de Wald

Uma alternativa ao teste da razão de verossimilhança para a inferência dos parâmetros do modelo é o teste de Wald, que sob hipótese nula tem distribuição assintoticamente $\chi^2_{rank(\boldsymbol{L})}$. O $rank$ de $\boldsymbol{L}$ é dado pelo número de linhas ou colunas linearmente independentes (LI). A estatística corresponde à estimativa do vetor de parâmetros dividida pelo erro padrão assintótico e pode ser escrita como

$$W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{L}'[\boldsymbol{L}(\sum_{i=1}^{n} \boldsymbol{X_i'}\hat{\boldsymbol{V}}_{\boldsymbol{i}}\boldsymbol{X}_i)^{-1}\boldsymbol{L}']\boldsymbol{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{2.18}$$

considerando que $\boldsymbol{L}$ é uma matriz conhecida (matriz de contrastes) cujas linhas correspondem ao número de restrições sob H_0 e as colunas ao número de parâmetros do modelo.

Contudo essa estatística não é recomendada para testar os parâmetros de covariância para os casos em que os mesmos encontram-se no limite do espaço paramétrico, pois elas podem ou não render estimativas válidas, como visto em Verbeke (1997). É recomendado o teste da razão de verossimilhança para esses casos (WEST et al., 2014), ao menos que se use inferência Baysiana a qual não será abordada neste trabalho.

2.5 Critérios de Informação

Os critérios de AIC e BIC são os mais utilizados. O critério de informação Akaike (AIC) é baseado no log da verossimilhança, sendo penalizado pelo número de parâmetros e escrita por

$$AIC = -2.l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) + 2(p+q), \tag{2.19}$$

em que, $(p+q)$, é o número total de parâmetros estimados de efeitos fixos e aleatórios, a estatística pode ser calculada com base na verossimilhança usual como na restrita (AKAIKE, 1974).

O critério BIC aplica uma maior penalidade para modelos que apresentam um grande número de parâmetros de acordo com West et al. (2014). Sua expressão é como segue,

$$BIC = -2.l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}}) + (p+q)\log(N), \tag{2.20}$$

onde $(p+q)$ é o número de parâmetros a serem estimados e N, número total de observações.

Basicamente esses dois critérios avaliam a qualidade do ajuste do modelo, condição esta que é caracterizada quando o modelo possui menores valores de AIC e BIC.

2.6 Análise de Resíduos e Diagnóstico

O diagnóstico descreve o conjunto de técnicas utilizadas, com finalidade de determinar qual dentre os modelos analisados é o mais parcimonioso. Por meio da avaliação do ajuste, busca-se verificar a existência de pontos discrepantes, garantir que os pressupostos de homocedasticidade, independência e normalidade dos erros se cumpram. A análise de influência, por sua vez, analisa se a extração de pontos atípicos detectados pelo processo anterior afetam as estimativas dos parâmetros. Nos casos em que ocorrem alterações, um novo modelo deve ser sugerido ou devem ser feitas investigações acerca dos pontos atípicos para possíveis exclusões. Nessa seção são apresentados os resíduos marginais, condicionais e $EBLUP$ sugeridos por Nobre e Singer (2007).

2.6.1 Resíduos Marginais

A finalidade dos resíduos marginais é basicamente verificar a linearidade da variável resposta em relação ás covariáveis. Ao verificar graficamente esta relação, é esperado que os resíduos estejam aleatoriamente espalhados em torno do zero, sendo que a expressão que corresponde a esses resíduos é,

$$r_i^{ym} = \boldsymbol{y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}. \tag{2.21}$$

Por meio dos resíduos marginais é possível compreender o quanto cada perfil individual se desvia do perfil médio do modelo.

2.6.2 Resíduos Condicionais

Os resíduos condicionais consistem na diferença entre o valor observado $\boldsymbol{y_i}$ e o valor condicional previsto. Considerando o resíduo do i-ésimo indivíduo como,

$$r_i^{yc} = \boldsymbol{y_i} - \boldsymbol{X_i}\hat{\boldsymbol{\beta}} - \boldsymbol{Z_i}\hat{\boldsymbol{b_i}}. \tag{2.22}$$

Esse resíduo acaba não sendo recomendado para avaliar a adequação do modelo, pois tende a ser correlacionado possivelmente com as diferentes variâncias dos indivíduos como visto em Pinheiro e Bates (2000). Por outro lado, Nobre e Singer (2007) sugerem uma padronização dos mesmos garantindo média zero e variância um, tal como segue,

$$r_i^* = \frac{r_i^{yc}}{\hat{\sigma}\sqrt{p_{kk}}}, \tag{2.23}$$

sendo, p_{kk}, definida pelos autores como funções conjuntas de alavancagem dos efeitos fixos e aleatórios. Graficamente esses resíduos, quando plotados versus o índice das observações e valores estimados, auxiliam na identificação de valores discrepantes e na verificação de homocedasticidade dos erros.

2.6.3 Resíduos de Efeitos Aleatórios (EBLUP)

Em Pinheiro e Bates (2000), podem ser vistos exemplos em que os resíduos EBLUP são utilizados para análise gráfica. O plot de $\boldsymbol{b_i}$ *versus* o índice dos sujeitos e valores estimados auxiliam na identificação de *outliers*.

Para os casos em que os efeitos aleatórios apresentam distribuição normal q-dimensional e não há confundimento, a distância de Mahalanobis entre $\hat{\boldsymbol{b}_i}$ e $E(\boldsymbol{b_i}) = 0$, expressa por

$$Mah_i = \hat{b_i}^T \left\{ \hat{Var}[\hat{b_i}] \right\}^{-1} \hat{b_i}, \tag{2.24}$$

é utilizada, a qual segue uma distribuição χ^2_q (SINGER; ANDRADE, 1986). O gráfico QQ baseado na distribuição χ^2_q para Mah_i contribui para identificar a normalidade dos efeitos aleatórios.

2.6.4 Análise de Influência

A análise de influência, também conhecida como análise de sensibilidade, baseia-se em verificar possíveis alterações nas estimativas dos parâmetros ao se remover dos dados pontos considerados atípicos. Sem dúvida nenhuma, uma das formas mais utilizada para medir a influência das estimativas dos parâmetros é por meio da distância de CooK, que em Christensen et al. (1992) é estendida para os modelos mistos como

$$D(\boldsymbol{\beta}) = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{[i]}})' Var(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}_{[i]}})}{rank(\boldsymbol{X})}, \tag{2.25}$$

onde, $[i]$, denota o conjunto de observações removidas dos dados.

De forma análoga o diagnóstico para os componentes da variância tem estatística descrita por

$$D(\boldsymbol{\tau}) = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_{[i]})' Cov(\hat{\boldsymbol{\tau}})^{-1} (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_{[i]}), \tag{2.26}$$

onde $\boldsymbol{\tau} = (\sigma^2, \sigma_{b^2_{0i}}, \cdots, \sigma^2_{b_{qi}})$ é o vetor de estimativas dos componentes de variância dos efeitos aleatórios.

3 Análise de Sobrevivência

3.1 Introdução

A análise de sobrevivência é uma ferramenta amplamente utilizada, ganhando grande popularidade na biologia e medicina, uma vez que a metodologia tem um grande apelo de interpretação no cenário de acompanhamento até o evento, permitindo inferir sobre o tempo de vida, ou risco de morte dado prognóstico. Collett (2015) especifica aplicações como o tempo de sobrevivência de animais em estudos experimentais, o tempo de realização de uma tarefa em psicologia experimental, bem como testes de componentes eletrônicos na análise do tempo de vida de dispositivos, entre outros.

Segundo Carvalho et al. (2011), sempre que o tempo for objeto de interesse, seja esse, interpretado como o tempo até a ocorrência do evento ou o risco de ocorrência no tempo essa análise deve ser utilizada.

Os modelos de sobrevivência são modelos como qualquer outro, sendo compostos basicamente por uma variável resposta, covariáveis explicativas ou fatores de risco, função de ligação e estrutura de erro. Sua eficácia se dá devido a utilização parcial das informações do indivíduo (TABLEMAN; KIM, 2003). O que os distinguem dos demais modelos são as censuras, ou observações incompletas, presentes no conjunto de dados.

3.2 O Tempo

Dentre os elementos que compõem a análise de sobrevivência, o tempo sem dúvida nenhuma é o principal fator a ser analisado. Os n indivíduos envolvidos no estudo são observados durante um tempo de acompanhamento, dado em escala real (dias, meses e anos). Com base nessa métrica, t_i, representa o tempo de sobrevivência, que pode ser visto como o tempo decorrido da entrada do i-ésimo indivíduo no estudo até a ocorrência do evento. Em estudos de ensaios clínicos o recrutamento do indivíduo em um experimental ocorre no tempo inicial t_0 e o tempo final em t_f (COLLETT, 2015). O engajamento do paciente pode coincidir com o momento do diagnóstico de certa doença, com a realização de um procedimento cirúrgico entre outros.

No entanto, sabe-se, por exemplo, que em ensaios clínicos ou em coorte nem todos os pacientes são acompanhados até o evento de interesse. Muitos abandonam o estudo e outros não apresentam o desfecho de interesse antes do término da pesquisa. Quando isso ocorre, é necessário levar em consideração uma nova variável que possa indicar essa falta de informação dando a esta o nome de censura.

3.3 Censura

Quando há perda de informações o tempo de sobrevivência não é computado, esse comportamento recebe o nome de censura (CARVALHO et al., 2011). A censura é parte inerente dos dados de sobrevivência. A construção da verossimilhança do modelo leva em consideração essa ausência de dados, no que se refere tanto à verossimilhança de modelos paramétricos quanto semiparamétricos.

Na análise de dados de sobrevivência podem ser observados os seguintes tipos de censura: à esquerda, à direita, intervalar e não informativa.

- O primeiro caso de censura pode ser verificado quando o evento já ocorreu e não se sabe ao certo em que momento. Um exemplo desse tipo de censura pode ser visto em Carvalho et al. (2011), quando o objeto de estudo corresponde ao tempo entre a infecção pelo HIV e o diagnóstico de AIDS. Sabe-se que a infecção (evento) em T ocorreu, mas só foi detectada juntamente no exame que acusou AIDS, em t_0, assim o tempo de sobrevivência é menor que o instante observado ($T < t_0$).

Figura 3 – Representação gráfica de censura à esquerda.

- O segundo caso ocorre quando o tempo de acompanhamento chega ao fim e o evento de interesse não ocorreu. Nada se sabe sobre sua ocorrência, ou seja, o tempo de sobrevivência é maior que o tempo observado ($T > t_0$).

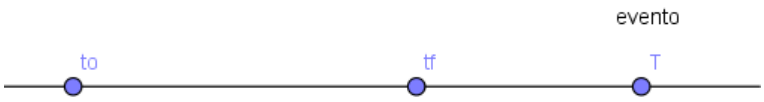

Figura 4 – Representação gráfica de censura à direita.

- No terceiro tipo a perda ocorre dentro do intervalo observado ($t_0 < T < t_f$). Esse tipo de censura decorre de monitoramentos feitos em períodos mais longos de tempo (semanas, meses e anos) muito comuns em estudos longitudinais.

- O último tipo considera independência entre o processo de risco e de censura. A censura não informativa pode ser identificada sobre perspectiva longitudinal. Nos casos que for identificado o abandono não aleatório, a censura não pode ser vista com base no risco do evento.

Figura 5 – Representação gráfica de censura intervalar.

Um indivíduo compreendido em um conjunto de dados em que há ocorrência de censura pode ser caracterizado por (t_i, δ_i), onde t_i é o tempo de observação do i-ésimo indivíduo e δ_i é a variável indicadora do status, sendo 1 se houve o desfecho e 0 caso contrário.

$$\begin{cases} \delta_i = 1; \text{o desfecho ocorreu para o } i\text{-ésimo indivíduo} \\ \delta_i = 0; \text{o desfecho não ocorreu para o } i\text{-ésimo indivíduo} \end{cases} \tag{3.1}$$

3.4 Função de Sobrevivência

A função de sobrevivência é a probabilidade de um indivíduo sobreviver por mais do que um determinado tempo t (COLLETT, 2003), sendo expressa por

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= \int_t^{\infty} f(t)\,\mathrm{d}t \\ &= 1 - F(t), \end{aligned} \tag{3.2}$$

onde $F(t)$, é a função de probabilidade acumulada de $f(t)$. Sendo $f(t)$ a função densidade de probabilidade, que descreve a probabilidade do i-ésimo indivíduo sofrer um determinado evento em um instante de tempo t. A função de sobrevivência é contínua e decrescente à direita, com $S(0) = 1$ e na medida em que o tempo cresce $S(t)$ converge para zero.

Por meio da função densidade de probabilidade, $S(t)$, é possível determinar a probabilidade de um paciente vir a sobreviver por mais de t unidade de tempo ou a probabilidade de um dispositivo vir a durar por mais que t unidades de tempo. Também é possível determinar os percentis do tempo de sobrevivência, sendo possível identificar a proporção de indivíduos que já sofreram o evento até o t-ésimo tempo.

3.5 Função de Risco

Embora seja chamada de função de risco, também pode ser vista como a taxa de incidência. O risco não é uma probabilidade e assume qualquer valor maior que zero. A

função de risco pode ser obtida através da função de sobrevivência como segue

$$\begin{aligned} \lambda(t) &= \lim_{\Delta \longrightarrow 0} \frac{P(t < T \leq t + \Delta | T > t)}{\Delta} \\ &= \frac{f(t)}{S(t)} \\ &= \frac{-d \ln(S(t))}{dt}, \end{aligned} \tag{3.3}$$

sendo $\lambda(t) \geq 0$ e $\int_0^\infty \lambda(t)dt = \infty$.

É importante notar a partir de 3.3 que as funções de risco e sobrevivência são ambas inversamente proporcionais, ou seja, quanto maior o risco, menor o tempo de sobrevivência e quanto maior o tempo de sobrevivência, menor é o risco (CARVALHO et al., 2011).

3.6 Modelo semiparamétrico de Cox

A principal motivação para a elaboração do modelo de riscos proporcionais foi a suposição de uma distribuição de probabilidade para o tempo de sobrevivência adotado pelos modelos paramétricos. Tal pressuposto, se inapropriado poderia levar a análise a conclusões inadequadas. Partindo desse ponto de vista, David et al. (1972) propuseram o modelo semiparamétrico de riscos proporcionais, que passou a ser um dos mais utilizados na análise de sobrevivência. Sua utilização abrange diversas áreas de conhecimento, principalmente no que diz respeito a estudos clínicos (COLOSIMO; GIOLO, 2006).

O modelo proposto por Cox assume proporcionalidade de risco, sendo apropriado a utilização de técnicas não paramétricas como o estimador de Kaplan-Meier, o qual possibilita acomodar os dados obtendo um primeiro contato com as curvas de sobrevivência, permitindo observar se os pressupostos de proporcionalidade não foram violados.

Em Colosimo e Giolo (2006), o modelo é apresentado de forma intuitiva, supõe-se que o interesse da pesquisa seja verificar o risco de morte em relação ao sexo (grupo).

Considerando $\lambda_0(t)$ como o risco de morte para o sexo feminino e $\lambda_1(t)$ o risco de morte para o sexo masculino, é assumindo proporcionalidade entre as funções de risco, de modo que, $\lambda_1(t)$ é proporcional a $\lambda_0(t)$, se e somente se, existe alguma constante C, $\forall\, t$, ao longo de T, em que

$$\frac{\lambda_1(t)}{\lambda_0(t)} = C, \tag{3.4}$$

então 3.4 representa o risco de morte de um paciente do sexo masculino em relação ao sexo feminino.

A variável binária x assume os valores 0 para o sexo feminino e 1 para o masculino. A constante 3.4 corresponde a $\exp\{x\gamma\}$. O modelo de riscos proporcionais para a covariável sexo fica completamente definida por

$$\lambda(t|\boldsymbol{X}) = \begin{cases} \lambda_0(t); \text{ se } x = 0 \\ \lambda_1(t) = \lambda_0(t)\exp\{\gamma\}; \text{ se } x = 1. \end{cases} \tag{3.5}$$

Se $x = 0$, a função assume o risco de indivíduos do sexo feminino, por outro lado, se $x = 1$ a equação 3.4 é multiplicado ao risco de base. O risco de um indivíduo do sexo masculino vir a sofrer o evento corresponde á 3.4 ou $\exp\{\gamma\}$ vezes o risco do sexo feminino.

Estudos que têm por finalidade identificar índices prognósticos, geralmente não apresentam interesse no risco basal ($\lambda_0(t)$). Se existir alguma relevância por parte da pesquisa em determiná-lo, isso pode ser feito de forma não paramétrica ou paramétrica. Ao contrário do risco basal, o fator paramétrico da função, ($\exp\{x\gamma\}$), deve ser bem especificado pois este descreve o efeito das covariáveis no risco (LEE; GO, 1997).

Ao generalizar a expressão 3.5 se tem:

$$\begin{aligned} \lambda(t|x_1, x_2, \cdots, x_p) &= \lambda_0(t)\exp(x_1\gamma_1 + x_2\gamma_2 + \cdots + x_p\gamma_p) \\ \lambda(t|\mathbf{x}) &= \lambda_0(t)\exp(\mathbf{x}\boldsymbol{\gamma}), \end{aligned} \tag{3.6}$$

observe que se $\mathbf{x}\boldsymbol{\gamma} = 0$, então $\lambda(t|\mathbf{x}) = \lambda_0(t)$. O vetor de covariáveis do modelo é expresso por $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ e $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_p)$ o vetor de parâmetros da regressão que indica a magnitude do efeito das covariáveis no risco de ocorrência do evento.

O modelo permite determinar a razão entre os riscos do i-ésimo e j-ésimo indivíduos, sendo possível constatar a proporcionalidade devido ao risco ser constante no decorrer do tempo.

$$\begin{aligned} \frac{\lambda(t|\mathbf{x_i})}{\lambda(t|\mathbf{x_j})} &= \frac{\lambda_0(t)\exp(\mathbf{x_i}\gamma)}{\lambda_0(t)\exp(\mathbf{x_j}\gamma)} \\ &= \frac{\exp(\mathbf{x_i}\gamma)}{\exp(\mathbf{x_j}\gamma)} \\ &= \exp(\mathbf{x_i}\gamma - \mathbf{x_j}\gamma). \end{aligned} \tag{3.7}$$

A razão pode ser interpretada como a determinação do risco referente ao i-ésimo indivíduo, dado suas covariáveis $\mathbf{x_i}$ quando comparado ao j-ésimo indivíduo condicionado à suas covariáveis.

Os casos que em a razão assumir valor igual um indicam igualdade de risco para ambos os indivíduos. Se a razão for maior que um, o i-ésimo paciente tem $\lambda(t|\mathbf{x})$ vezes mais risco de sofrer o evento, o mesmo que $1 - \lambda(t|\mathbf{x})$ mais risco. Em contrapartida, valores inferiores a unidade indicam, em muitos casos, fator de proteção ou simplesmente redução do risco. Tais valores dependem do efeito das covariáveis no risco

$$\begin{cases} \gamma < 0; \text{Redução de risco} \\ \gamma = 0; \text{Mesmo risco} \\ \gamma > 0; \text{Aumento de risco} \end{cases} \tag{3.8}$$

As estimativas dos parâmetros do modelo de Cox, são obtidas por meio da Verossimilhança Parcial.

3.6.1 Estimação de parâmetros

O modelo de riscos proporcionais, contrariamente aos modelos paramétricos, não considera o método de máxima verossimilhança adequado. Para melhor compreensão, faz-se necessário observar a probabilidade da amostra dos modelos paramétricos, que tem forma dada por

$$\begin{aligned} L(\gamma) &= \prod_{i=1}^{n} \left[f(t_i|\mathbf{x}_i)\right]^{\delta_i} \left[S(t_i|\mathbf{x}_i)\right]^{1-\delta_i} \\ L(\gamma) &= \prod_{i=1}^{n} \left[\lambda_0(t_i)\exp(\mathbf{x}\gamma)\right]^{\delta_i} \left(S_0(t_i)\right)^{\exp(\mathbf{x_i}\gamma)}, \end{aligned} \tag{3.9}$$

sendo que o primeiro fator corresponde à contribuição dos indivíduos que sofreram o evento de interesse e o segundo aos que foram censurados.

Em 3.9 pode ser observada a presença de $\lambda_0(t_i)$, função de perturbação, evidenciando que os processos usuais de estimação não servem. O método de estimação utilizado para o modelo de riscos proporcionais consiste em determinar a probabilidade do evento ocorrer no tempo t_i dado o espaço amostral dos indivíduos sob risco, denotado por $R(t_i)$.

A estimação dos parâmetros é obtida a partir de verossimilhanças individuais, L_i, descrita pela razão entre o risco do i-ésimo indivíduo sofrer o evento exatamente no instante anterior a t_i e pela soma dos demais indivíduos sob risco. É natural verificar que a censura contribui somente com o denominador da expressão

$$\begin{aligned} L_i &= \frac{\lambda_0(t_i|\mathbf{x}_i)}{\sum_{j\in R(t_i)} \lambda_j(t_j|\mathbf{x}_j)} \\ &= \frac{\exp(\mathbf{x}\gamma)}{\sum_{j\in R(t_i)} \exp(\mathbf{x_j}\gamma)}. \end{aligned} \tag{3.10}$$

Para uma ilustração mais detalhada, supõe-se uma coorte em que o objetivo seja observar o risco apenas no intervalo de tempo [0, 50] dias. Os indivíduos foram submetidos a um processo cirúrgico e a verossimilhança individual, L_i, pode ser formulada sob observação da Figura 6.

No $20.^o$ dia, apenas o paciente A sofreu o evento de interesse, sendo o vetor de tempo especificado por $t = (20, 30, 40)$ e de pacientes por $i = (1, 2, 3, 4, 5, 6)$. É verificado três tempos para seis pacientes, constatando assim a presença de empates que serão ilustrados a seguir.

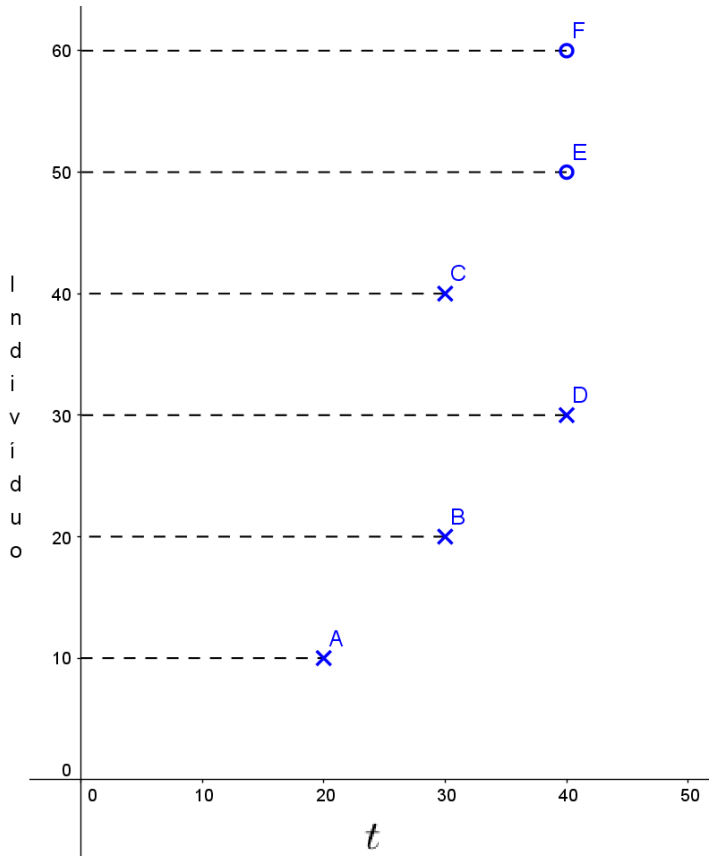

Figura 6 – Tempo de sobrevivência de seis indivíduos, os segmentos horizontais representam a trajetória dos indivíduos, $\times$ representa a ocorrência do evento e $\circ$ censura.

A probabilidade do evento de interesse ocorrer em t_1 corresponde a probabilidade de que o primeiro paciente venha a sofrer o evento dado que os demais indivíduos no momento t_1 estejam sob risco $R(t_1)$.

$$L_1 = \frac{e^{\mathbf{x_1}\gamma}}{e^{\mathbf{x_1}\gamma} + e^{\mathbf{x_2}\gamma} + e^{\mathbf{x_3}\gamma} + e^{\mathbf{x_4}\gamma} + e^{\mathbf{x_5}\gamma} + e^{\mathbf{x_6}\gamma}} = \frac{e^{\mathbf{x_1}\gamma}}{\sum_{j \in R(t_i)} e^{\mathbf{x_j}\gamma}}$$

No caso do tempo $t_1 = 20$, houve apenas um evento, assim A não faz mais parte de conjunto $R(t_i)$.

Contudo, no tempo seguinte ocorreram dois empates, ou seja, em $t = (2, 3)$ foi verificado o evento. Logo, a probabilidade de que nesse tempo tenham morrido os indivíduos $i = 2$ ou $i = 3$, ambos com probabilidades $P(K)$ dado $R(t_2)$ e $P(J)$ dado $R(t_3)$, respectivamente, é dado por $L_2 = P(K) \cup P(J)$, sendo que

$$P(K) = \frac{e^{\mathbf{x_2}\gamma}}{e^{\mathbf{x_2}\gamma} + e^{\mathbf{x_3}\gamma} + e^{\mathbf{x_4}\gamma} + e^{\mathbf{x_5}\gamma} + e^{\mathbf{x_6}\gamma}} \times \frac{e^{\mathbf{x_3}\gamma}}{e^{\mathbf{x_3}\gamma} + e^{\mathbf{x_4}\gamma} + e^{\mathbf{x_5}\gamma} + e^{\mathbf{x_6}\gamma}}.$$

e

$$P(J) = \frac{e^{\mathbf{x_3}\gamma}}{e^{\mathbf{x_2}\gamma} + e^{\mathbf{x_3}\gamma} + e^{\mathbf{x_4}\gamma} + e^{\mathbf{x_5}\gamma} + e^{\mathbf{x_6}\gamma}} \times \frac{e^{\mathbf{x_2}\gamma}}{e^{\mathbf{x_2}\gamma} + e^{\mathbf{x_4}\gamma} + e^{\mathbf{x_5}\gamma} + e^{\mathbf{x_6}\gamma}}.$$

Logo

$$L_2 = \frac{e^{\mathbf{x_2}\gamma}}{\sum_{j \in R(t_i)} e^{\mathbf{x_j}\gamma}} \times \frac{e^{\mathbf{x_3}\gamma}}{\sum_{j \in R(t_i)} e^{\mathbf{x_j}\gamma}} + \frac{e^{\mathbf{x_3}\gamma}}{\sum_{j \in R(t_i)} e^{\mathbf{x_j}\gamma}} \times \frac{e^{\mathbf{x_2}\gamma}}{\sum_{j \in R(t_i)} e^{\mathbf{x_j}\gamma}}. \tag{3.11}$$

Para determinar L_3, foi verificado a ocorrência de um evento e duas censuras, o evento deve ser considerado primeiro como

$$L_3 = \frac{e^{\mathbf{x_4}\gamma}}{e^{\mathbf{x_4}\gamma} + e^{\mathbf{x_5}\gamma} + e^{\mathbf{x_6}\gamma}}.$$

Então a verossimilhança parcial para os 6 primeiros pacientes da coorte é escrita por

$$\begin{aligned} L_p(\boldsymbol{\gamma}) &= L_1 L_2 L_3 \\ &= \frac{e^{\boldsymbol{x_i}\gamma}}{\sum_{j \in R(t_i)} e^{\boldsymbol{x_j}\gamma}}. \end{aligned} \tag{3.12}$$

Que pode ser expressa na forma

$$L_p(\boldsymbol{\gamma}) = \left\{ \frac{\exp\left(\sum_{i=1}^{n} \sum_{k=1}^{p} \mathbf{x}_i \boldsymbol{\gamma}_k\right)}{\sum_{j \in R(t_i)} \exp\left(\sum_{i=1}^{n} \sum_{k=1}^{p} \mathbf{x}_j \boldsymbol{\gamma}_k\right)} \right\}, \tag{3.13}$$

onde $i = 1, ..., n$, denota o número de indivíduos em estudo e $k = 1, ..., p$, representa os parâmetros que especificam $\mathbf{x}$.

Aplicando o logaritmo na função de verossimilhança tem-se

$$l(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{p} \mathbf{x}_i \boldsymbol{\gamma}_k - log \left[\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \boldsymbol{\gamma}_k\right) \right] \right\}. \tag{3.14}$$

Por $l(\boldsymbol{\gamma})$, é possível determinar os estimadores de máxima verossimilhança do vetor de $\boldsymbol{\gamma}'$s.

3.6.2 Função Escore

Os estimadores de máxima verossimilhança são os valores que compõem o vetor $\hat{\boldsymbol{\gamma}}$, os mesmos satisfazem o sistema de equações abaixo

$$U\left[l(\boldsymbol{\gamma})\right] = \begin{cases} \frac{\partial l(\boldsymbol{\gamma})}{\partial \gamma_1} = \sum_{i=1}^{n} \left[x_{i,1} - \frac{\sum_{j=R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_{j,1}}{\sum_{j=R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\boldsymbol{\gamma}}\right)} \right] = 0 \\ \vdots \\ \frac{\partial l(\boldsymbol{\gamma})}{\partial \gamma_p} = \sum_{i=1}^{n} \left[x_{i,p} - \frac{\sum_{j=R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_{j,p}}{\sum_{j=R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\boldsymbol{\gamma}}\right)} \right] = 0. \end{cases} \tag{3.15}$$

A matriz de covariância dos parâmetros pode ser determinada a partir do negativo da derivada parcial de segunda ordem do logaritmo da função de verossimilhança, a qual é descrita pela matriz $I\left[l(\boldsymbol{\gamma})\right]$, encontrada a partir de

$$I\left[l(\boldsymbol{\gamma})\right] = -\frac{\partial^2 l(\boldsymbol{\gamma})}{\partial \gamma_k \partial \gamma_p}. \tag{3.16}$$

A solução do sistema de equações é obtido pelo método numérico de Newton-Raphson, baseado na expansão de $U\left[l\left(\hat{\boldsymbol{\gamma}}^m\right)\right]$ em séries de Taylor na vizinhança de $\hat{\boldsymbol{\gamma}}^m$.

A ideia do método de otimização em duas dimensões para obter uma aproximação de $\hat{\gamma}$, é ilustrada na sequência. Atribuindo ao processo um tentativa inicial, $\hat{\gamma}^0$, é possível obtendo o ponto $\left(\hat{\gamma^0}, l\left(\gamma^0\right)\right)$, por onde passa a reta tangente (tg) à curva $l\left(\boldsymbol{\gamma}\right)$, a reta intercepta o eixo dos $\boldsymbol{\gamma}'$s em γ^0. O valor de $\hat{\gamma}^1$ encontrado passa a ser uma segunda aproximação de $\hat{\gamma}$ como segue

$$\begin{aligned} U\left[l\left(\gamma\right)\right] - U\left[l\left(\hat{\gamma}^0\right)\right] &= U'\left[l\left(\hat{\gamma}^0\right)\right]\left(\gamma - \hat{\gamma^0}\right) \\ 0 - U\left[l\left(\hat{\gamma}^0\right)\right] &= U'\left[l\left(\hat{\gamma}^0\right)\right]\left(\hat{\gamma} - \hat{\gamma^0}\right) \\ \hat{\gamma}^1 &= \hat{\gamma}^0 - I\left[l\left(\hat{\gamma}^0\right)\right]^{-1} U\left[l\left(\hat{\gamma}^0\right)\right] \end{aligned}$$

O intercepto γ da tg_1, $\hat{\gamma}^1$, é determinado tomando $\gamma = \hat{\gamma}^1$ e $U\left[l\left(\gamma\right)\right] = 0$, sendo assim

$$\hat{\gamma}^{m+1} = \hat{\gamma}^m - I\left[l\left(\hat{\gamma}^m\right)\right]^{-1} U\left[l\left(\hat{\gamma}^m\right)\right],$$

onde $\hat{\boldsymbol{\gamma}}$ tem dimensão p, com matriz hessiana dada por $I\left[l\left(\hat{\boldsymbol{\gamma}}^m\right)\right]^{-1}$ e gradiente $U\left[l\left(\hat{\boldsymbol{\gamma}}^m\right)\right]$. A ilustração em duas dimenssões pode ser vista na Figura 7.

O processo faz uso da matriz de informação $I\left[l\left(\hat{\boldsymbol{\gamma}}^m\right)\right]$. O Método Escore de Fisher proposto substitui a matriz de informação pelo seu valor esperado.

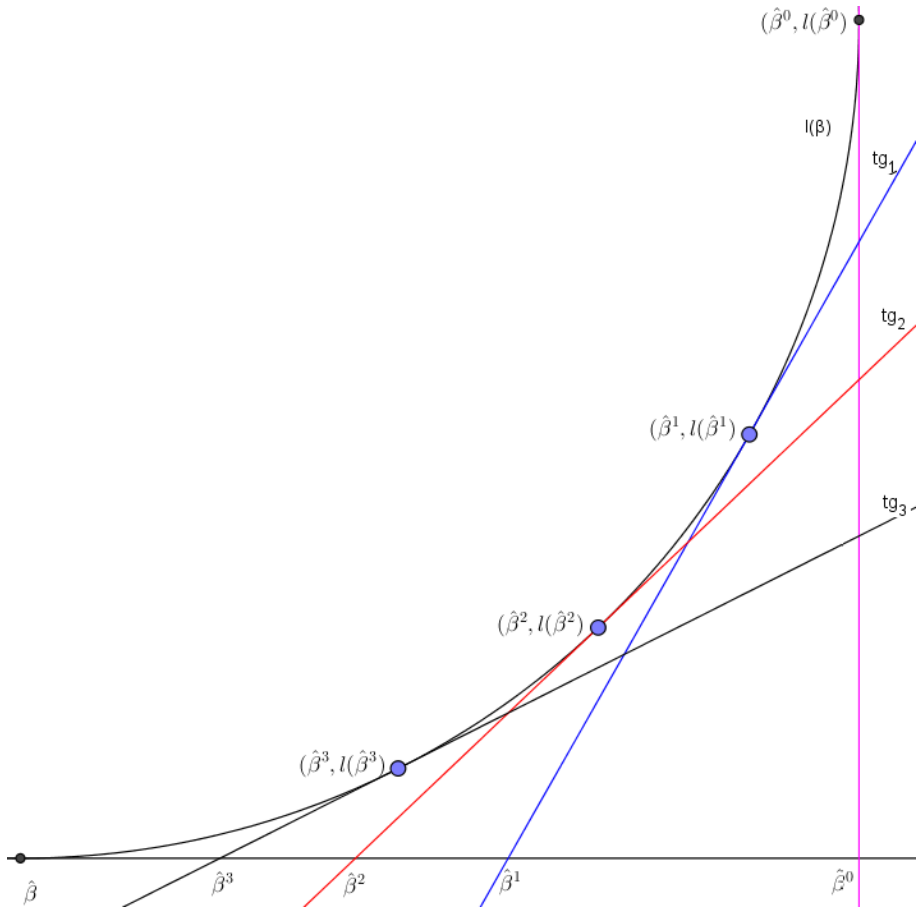

Figura 7 – Representação do método numérico por meio das tangentes em duas dimensões.

3.6.3 Intervalo de Confiança IC

O intervalo de confiança para o vetor γ pode ser determinado por meio do inverso da informação de Fisher, com limites inferior e superior dados por

$$IC(\boldsymbol{\gamma}, 100(1-\alpha)\%) \;=\; \hat{\boldsymbol{\gamma}} \pm Z_{\alpha/2} se(\hat{\boldsymbol{\gamma}}). \tag{3.17}$$

Inferência por meio do IC também pode ser verificada, isto é, intervalos que contemplam o valor zero indicam não haver diferença significativa no risco de diferentes grupos de tratamento, para os caso em que o teste de hipótese for de Wald.

3.6.4 O Risco Relativo-HR

O Risco Relativo é escrito por

$$\frac{\lambda(t|\mathbf{x})}{\lambda_0(t)} \;=\; \exp(\mathbf{x}\boldsymbol{\gamma}), \tag{3.18}$$

o qual pode ter sua variabilidade estimada, por intermédio do método Delta. Supondo que $HR = \psi = \exp(\mathbf{x}\boldsymbol{\gamma})$, uma função dependente com $p = 1$ parâmetro, a variância da função

é dada por

$$\begin{aligned} Var(\hat{\psi}) &= \left(\frac{\partial \psi}{\partial \hat{\gamma}}\right)^2 \cdot Var(\hat{\gamma}) \\ Var(\hat{\psi}) &= [\exp(\hat{\gamma})]^2 \cdot Var(\hat{\gamma}) \\ se(\hat{\psi}) &= \exp(\hat{\gamma}) \cdot se(\hat{\gamma}). \end{aligned} \tag{3.19}$$

O método Delta aproxima uma função de variáveis aleatórias através da expansão em séries de Taylor, o método se baseia na eficiência assintótica dos estimadores de Máxima Verossimilhança. Vale salientar que o método é aplicável para funções multiparamétricas, caso em que mais covariáveis estão envolvidas.

3.6.5 Estimação da Função de Sobrevivência Basal ($\lambda_0(t)$)

Conhecidos os coeficientes do modelo é possível estimar a função de sobrevivência basal subjacente. Segundo Colosimo e Giolo (2006), como a função de risco basal não é conhecida, pode-se usar o estimador proposto por Breslow, uma estimador não-paramétrico de risco basal acumulado, que leva em consideração as observações em distintos tempos e considera o risco constante entre elas, o estimador é dado por

$$\hat{H}_0(t) = \sum_{i:t_i<T} \frac{d_i}{\sum_{j=R(t_i)} \exp(\boldsymbol{x\gamma})}, \tag{3.20}$$

onde d_i corresponde ao número de eventos no tempo t_i, de modo que a função de sobrevivência possa ser determinada por

$$\hat{S}(t) = [S_0(t)]^{\exp(\boldsymbol{x\gamma})}. \tag{3.21}$$

3.7 Análise de Resíduos e Diagnóstico

Os resíduos são importantes ferramentas para a análise do modelo, tornando-se fundamental, tanto no processo de identificação de pontos discrepantes quanto em descriminar se o modelo não está adequado aos dados.

3.7.1 Resíduos de Schoenfeld

Os resíduos Schoenfeld são obtidos a partir da primeira derivada do logaritmo da função de verossimilhança parcial em relação aos parâmetros (COLLETT, 2015). Sendo

utilizados com intuito de verificar quais covariáveis não deveriam pertencer ao modelo, ou seja, quais são dependentes do tempo violando os pressupostos de proporcionalidade, o resíduo pode ser expresso por

$$\mathbf{r}_{Si}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} x_i - \frac{\sum\limits_{j \in R(t_i)} \exp\left(\sum\limits_{k=1}^{p} \mathbf{x}_j \hat{\boldsymbol{\gamma}}\right) \mathbf{x}_j}{\sum\limits_{j \in R(t_i)} \exp\left(\sum\limits_{k=1}^{p} \mathbf{x}_j \hat{\boldsymbol{\gamma}}\right)}, \tag{3.22}$$

que corresponde à diferença entre o valor esperado da covariável para o i-ésimo indivíduo e o valor da soma ponderada das probabilidades de selecionar uma determinada pessoa a partir do risco definido no instante t_i (MOORE, 2016). Em sua análise deseja-se que a forma gráfica do tempo versus resíduos apresente comportamento quase constante ou uma curva muito suave em torno do zero.

3.7.2 Resíduos Martingale

Os resíduos Martingale são definidos pela diferença entre δ_i e o negativo do log da estimativa da função de sobrevivência. Traçando os resíduos contra os preditores contínuos e passando uma curva $loess$ através dos pontos plotados o resultado obtido não pode ser diferente substancialmente de uma linha horizontal. Nos casos em que isso não ocorre, indicam que o modelo não foi especificado corretamente (MAINDONALD; BRAUN, 2006).

$$\begin{aligned} \hat{n}_i &= \delta_i - \hat{H}(t_i|\mathbf{x}_i) \\ &= \delta_i - \hat{H}_0(t_i) \cdot \exp(\mathbf{x}_{\boldsymbol{i}}\boldsymbol{\gamma}) \\ &= \delta_i - log\left[S(\hat{t}_i)\right] \\ &= \delta_i + r_{C_i}, \end{aligned} \tag{3.23}$$

onde, $\delta_i = 0$ ou 1, para o qual, 0 representa censura, 1 ocorrência do evento em t_i e r_{C_i} determina os resíduos de Cox-Snell. Espera-se a partir da análise gráfica que $\hat{n}_i$ do modelo nulo versus covariável contínua identifique uma não linearidade na dispersão dos resíduos. Transformações na variável podem ser realizadas no intuito de suavizar discrepâncias de modelos mal especificados.

3.7.3 Resíduos $Deviance$

Os resíduos $Deviance$ permitem determinar possíveis $outliers$, sendo expresso por

$$D_i = sinal(\hat{n}_i)\left\{-2\left[\hat{n}_i + \delta_i log(\delta_i - \hat{n}_i)\right]\right\}^{\frac{1}{2}}, \tag{3.24}$$

onde $sinal(\hat{n}_i)$ representa o sinal dos resíduos martingale, para os quais, valores negativos são verificados para indivíduos que demoram a apresentar o evento e valores positivos caso contrário. Esse fato pode ser facilmente observado pelos gráficos de resíduos Martingale.

3.7.4 Resíduos escore ($Dfbetas$)

O resíduo escore permite avaliar quais dentre as observações influenciam no ajuste do modelo. Para Cleves (2008), é de suma importância avaliar se algum sujeito ou grupo exerce influência nas estimativas dos parâmetros da regressão. Esses sujeitos ou grupos geralmente têm efeito de alavanca sobre as estimativas.

No intuito de verificar a influência de cada observação sobre a estimativa $\hat{\gamma}$ de γ, considera-se $\hat{\gamma_{(i)}}$ denotando o vetor estimado de coeficientes calculados na amostra com a i-ésima observação excluída. Em seguida, são verificados quais componentes do vetor $\hat{\gamma} - \hat{\gamma_{(i)}}$ têm valores absolutos indevidamente grandes. Isso é feito para cada uma das n observações.

$$dfbeta = I(\hat{\gamma})^{-1}(r_{S1i}, ..., r_{Ski}) \tag{3.25}$$

onde $I(\hat{\gamma})^{-1}$, é a inversa da matriz de informação de Fisher, para $k = 1, ..., p$.

$$\mathbf{r}_{Sji} = \{x_{i,k} - a_{i,k}\} - \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\gamma}\right) \sum_{t_i \leq y_i} \frac{\{x_{i,k} - a_{i,k}\}}{\sum_{l \in R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_l \hat{\gamma}\right)}, \tag{3.26}$$

em que

$$a_{i,k} = \frac{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\gamma}\right) \mathbf{x}_{i,j}}{\sum_{j \in R(t_i)} \exp\left(\sum_{k=1}^{p} \mathbf{x}_j \hat{\gamma}\right)}. \tag{3.27}$$

O primeiro componente de 3.27 corresponde ao i-ésimo resíduo Schoenfeld e o segundo mede o efeito combinado sobre todos os conjuntos de risco que incluem o i-ésimo sujeito. Assim, para os primeiros tempos de morte o primeiro componente domina e para

tempos de morte posteriores, o segundo geralmente é de maior magnitude. Isso significa que para pacientes que morrem tarde (pacientes que vivem por muito tempo) estão presentes em muitos conjuntos de risco, tem mais efeito sobre $\hat{\gamma}$ do que pacientes que morreram em vez de serem censurados (TABLEMAN; KIM, 2003).

3.8 Extensão do modelo de Cox

O modelo estendido é basicamente uma alteração do modelo padrão de Cox. As alterações são evidenciadas por meio de inclusão de covariáveis que durante um período de tempo estão sujeitas a mudanças. Como, por exemplo, no acompanhamento clínico de pacientes, o pesquisador tem interesse em observar variáveis cujos valores se alteram no decorrer do tempo, sejam essas alterações decorrentes da ação de reagentes químicos no organismo ou de reações naturais da própria condição do indivíduo. Covariáveis que descrevem os registros da carga viral, contagem de células T CD4 em indivíduos com HIV são exemplos de covariáveis dependentes do tempo que podem ser consideradas no modelo de Cox estendido, expresso por

$$\lambda(t|x(t)) = \lambda_0(t)\exp(\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{K}(t)), \tag{3.28}$$

o mesmo deixa de ser um modelo de riscos proporcionais no tempo, a covariável dependente pode assumir valores tanto contínuos como discretos.

O método de estimação, seleção de modelos e análises de resíduos são semelhantes ao modelo de Cox original.

Ao incorporar no modelo estendido de Cox um efeito aleatório (A_i), é possível obter

$$\begin{aligned} \lambda(t|x(t)) &= \lambda_0(t)\exp(\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{K}(t) + A_i), \\ \lambda(t|x(t)) &= \lambda_0(t)\exp(\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{K}(t)).\exp(A_i), \end{aligned} \tag{3.29}$$

onde $\boldsymbol{K}$ é uma matriz de covariáveis associada ao vetor de parâmetros $\boldsymbol{\gamma}$, $\exp(A_i)$ específica a fragilidade do i-ésimo indivíduo agindo de forma multiplicativa sob $\lambda_0(t)$. O termo de fragilidade corresponde a uma amostra aleatória independente e identicamente distribuída com distribuição de probabilidade com média igual a um e variância não especificada, se $\exp(A_i)$ for um valor maior que 1, a sobrevivência se torna mais curta. De acordo com Colosimo e Giolo (2006), a não inclusão de covariáveis importantes ao modelo produzem um aumento na heterogeneidade, afetando a inferência, de modo que ao ser incluído um termo de fragilidade ao modelo esse problema pode ser amenizado.

4 Modelo Conjunto

O Modelo Conjunto para estimar associações entre respostas longitudinais e de sobrevivência são, em síntese, compostos por dois submodelos, os quais especificam os processos longitudinal e de risco (RIZOPOULOS, 2012). A contagem de CD4, biomarcador longitudinal, é utilizada para descrever o processo de evolução ou história natural do HIV e, considerando seu valor como verdadeiro, serve como informação para estimar o tempo de sobrevivência.

A estimação dos parâmetros foi obtida ao maximizar o logaritmo da função de verossimilhança da distribuição conjunta dos resultados observados do tempo de evento e da resposta longitudinal. Uma função única que incorpora na estimação o método de otimização EM resolvendo problemas de dados faltantes aleatórios (MAR) presentes nos dados longitudinais. No caso da falta não aleatória (MNAR), o último tempo observado é utilizado pela função como informação para a sobrevivência (RIZOPOULOS, 2010; RIZOPOULOS, 2012; TERRIN et al., 2015). Essa metodologia fornece estimativas mais eficientes para o processo longitudinal e reduz o viés da sobrevivência (GUO; CARLIN, 2004; IBRAHIM et al., 2010; SUÁREZ et al., 2017).

Para modelar conjuntamente a medida de associação entre o nível do biomarcador longitudinal e o risco de evento, o modelo conjunto especificado neste estudo seguirá a abordagem proposta por Rizopoulos (2012), a expressão é dada por

$$\begin{aligned} \lambda(t|m_i(t), k_i) &= \lambda_0(t)\exp(\boldsymbol{\gamma}^T k_i)\exp(\alpha m_i(t)) \\ \lambda(t|m_i(t), k_i) &= \lambda_0(t)\exp(\boldsymbol{\gamma}^T k_i + \alpha m_i(t)), t > 0, \end{aligned} \tag{4.1}$$

onde $\lambda(t|M_i(t), K_i)$, corresponde ao risco do evento, dado o verdadeiro valor da covariável medida no tempo ($m_i(t)$). O parâmetro α especifica a intensidade de ligação da função trajetória de $m_i(t)$ ao risco do evento. As informações basais são especificadas por K_i, matriz de covariáveis explicativas, com parâmetros de regressão dados pelo vetor $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_n)$, representando o efeito das covariáveis no risco. A função de risco basal $\lambda_0(t)$, geralmente não é especificada, pois de acordo com Hsieh et al. (2006, apud RIZOPOULOS, 2012) leva a análise a subestimar os erros padrão das estimativas dos parâmetros. Em muitos casos pode ser adequado utilizar distribuições tipicamente conhecidas, como Weibull, Log-normal e a Gamma (RIZOPOULOS, 2012).

Para determinar o grau de associação denotado por α, entre o nível do biomarcador longitudinal e o risco de ocorrência do evento faz-se necessário determinar $m_i(t)$, que corresponde ao verdadeiro valor não observado no processo longitudinal de tempo t.

O verdadeiro valor do biomarcador longitudinal $m_i(t)$ isento de erros residuais é composto por componentes de efeitos fixos, $x_{ij}\boldsymbol{\beta}$, e aleatórios, $z_{ij}\boldsymbol{b_i}$. Por sua vez, $m_i(t) = \{m_i(t), 0 \leq s \leq t\}$, denota a história da verdadeira observação longitudinal para cada indivíduo até t. Então para um modelo linear de efeitos aleatórios temos

$$\begin{cases} y_{ij}(t) = m_i(t) + \epsilon_i(t), \\ m_i(t) = x_i^T(t)\boldsymbol{\beta} + z_i^T(t)\boldsymbol{b_i}, \\ \boldsymbol{b_i} \sim N(0, G), \boldsymbol{\epsilon}_i(t) \sim N(0, R), \end{cases} \tag{4.2}$$

assim, $y_{ij}(t)$, denota a j-ésima observação do marcador longitudinal para o i-ésimo indivíduo, $m_i(t)$, o verdadeiro nível do marcador (sem erros residuais) no tempo t, $x_i^T(t)$, representa o vetor de covariáveis do i-ésimo indivíduo medido no j-ésimo tempo, sendo explicado completamente por β, e $\boldsymbol{b_i}$ descreve a taxa do nível do marcador longitudinal do i-ésimo indivíduo.

4.1 Estimação

Seguindo a abordagem de Rizopoulos (2010), a estimação dos parâmetros do modelo conjunto se baseia na maximização da função de verossimilhança da distribuição conjunta dos resultados observados (t_i, δ_i, y_i), ou seja, do tempo até o evento e da resposta longitudinal. A função densidade de probabilidade conjunta da amostra é dada por

$$p(t_i, \delta_i, y_i|b_i; \theta) = p(t_i, \delta_i|b_i; \theta)p(y_i|b_i; \theta) \tag{4.3}$$

onde

$$p(y_i|b_i; \theta) = \prod_j p\left\{y_{ij}(t_{ij})|b_i; \theta\right\},$$

em que, $\boldsymbol{\theta}$ denota um vetor de parâmetros constituído por: $\boldsymbol{\theta_t}$, subvetor dos parâmetros do resultado do tempo até o evento $(\gamma^T, \alpha, \theta_{\lambda_0}^T)$, $\boldsymbol{\theta_y}$, subvetor dos parâmetros da resposta longitudinal $(\boldsymbol{\beta}^T, \sigma^2)$ e $\boldsymbol{\theta}_b$ subvetor dos parâmetros da matriz de covariância dos efeitos aleatórios (RIZOPOULOS, 2010). Assume-se que o mesmo vetor de efeitos aleatórios é compartilhado entre os submodelos de sobrevivência e longitudinal. O logaritmo da função de máxima verossimilhança conjunta para o i-ésimo indivíduo é expresso por

$$\log p(t_i, \delta_i, y_i; \theta) = \log \int p(t_i, \delta_i|b_i; \theta_t, \beta)p(y_i|b_i; \theta)db_i. \tag{4.4}$$

Dessa forma, a primeira parcela da integral corresponde a função densidade de probabilidade condicional para a sobrevivência

$$\begin{aligned} p(t_i, \delta_i | b_i; \theta_t, \beta) &= \lambda_i(t_i | m_i; \theta_t, \beta)^{\delta}.S_i(t_i | m_i(t_i); \theta_t, \beta) \\ &= \lambda_0(t_i) \exp\left\{\gamma^T k_i + \alpha m_i(t_i)\right\}^{\delta_i} \\ &\times \exp\left(-\int_0^{t_i} \lambda_0(s) \exp\left\{\gamma^T k_i + \alpha m_i(s)\right\} ds\right), \end{aligned} \tag{4.5}$$

de modo que a função basal de risco possa admitir qualquer função positiva no tempo (RIZOPOULOS, 2010).

Para a função densidade de probabilidade conjunta da resposta longitudinal com o efeito aleatório, é escrita como segue

$$\begin{aligned} p(y_i | b_i; \theta) &= \prod_j p\left\{y_i(t_{ij} | b_i; \theta_y)\right\} p(b_i; \theta_b) \\ &= (2\pi\sigma^2)^{-n_i/2} \exp\left\{- \parallel y_i - X_i\beta - Z_i b_i \parallel^2 / 2\sigma^2\right\} \\ &\times (2\pi)^{-q/2} det(G)^{-1/2} \exp\left(-b_i^T G^{-1} b_i / 2\right), \end{aligned} \tag{4.6}$$

onde q denota a dimensionalidade dos efeitos aleatórios e $\|x\| = \{\sum_i x_i^2\}$ é a norma vetorial euclidiana (RIZOPOULOS, 2012).

A estimação é realizada por meio da maximização do logaritmo da função de verossimilhança, em que o vetor escore é dado por

$$\begin{aligned} U(\theta) &= \sum_i \frac{\partial}{\partial \theta^T} \log \int p(t_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) p(b_i; \theta) db_i \\ &= \sum_i \frac{1}{p(t_i, \delta_i, y_i; \theta)} \frac{\partial}{\partial \theta^T} \log \int p(t_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) p(b_i; \theta) db_i \\ &= \sum_i \frac{1}{p(t_i, \delta_i, y_i; \theta)} \int \frac{\partial}{\partial \theta^T} p(t_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) p(b_i; \theta) db_i \\ &= \sum_i \int \left[\frac{\partial}{\partial \theta^T} \log\left\{p(t_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) p(b_i; \theta)\right\}\right] \\ &\times \frac{p(t_i, \delta_i | b_i; \theta) p(y_i | b_i; \theta) p(b_i; \theta)}{p(t_i, \delta_i, y_i; \theta)} db_i \\ &= \sum_i \int A(\theta, b_i) p(b_i | t_i, \delta_i, y_i; \theta) db_i, \end{aligned} \tag{4.7}$$

sendo que, $A(\theta, b_i) = \partial\left\{\log p(t_i, \delta_i | b_i; \theta) + \log p(y_i | b_i; \theta) + \log p(b_i; \theta) / \partial \theta^T\right\}$, denota o vetor de escore dos dados completos. Derivando 4.7, é possível determinar a matriz de

informação de Fisher $I(\hat{\theta})$, dado por

$$\begin{aligned} I(\hat{\theta}) = \frac{\partial U_i(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \int A(\theta, b_i) p(b_i|t_i, \delta_i, y_i; \theta) db_i \\ &= \int \frac{\partial A(\theta, b_i)}{\partial \theta} p(b_i|t_i, \delta_i, y_i; \theta) db_i \\ &+ \underbrace{\int A(\theta, b_i) \frac{\partial p(b_i|t_i, \delta_i, y_i; \theta)}{\partial \theta} db_i}, \end{aligned} \tag{4.8}$$

onde a segunda parcela de 4.8 corresponde a

$$\begin{aligned} &= \int A(\theta, b_i) \left\{ \frac{\partial \log p(b_i|t_i, \delta_i, y_i; \theta)}{\partial \theta} \right\}^T p(b_i|t_i, \delta_i, y_i; \theta) db_i \\ &= \int A(\theta, b_i) \left\{ \frac{\partial \left\{ \log p(t_i, \delta_i|b_i; \theta) + \log p(y_i|b_i; \theta) + \log p(b_i; \theta) \right\}}{\partial \theta} - \partial \log p(t_i, \delta_i, y_i; \theta) \right\}^T \\ &\times \; p(b_i|t_i, \delta_i, y_i; \theta) db_i. \end{aligned} \tag{4.9}$$

Estimado $I(\hat{\theta})$, que permite determinar os erros padrão das estimativas,

$$\begin{aligned} Var(\hat{\theta}) &= \left\{ I(\hat{\theta}) \right\}^{-1} \\ &= \left\{ -\sum_{i=1}^{n} \frac{\partial U(\theta)}{\partial \theta} \right\}^{-1}, \end{aligned} \tag{4.10}$$

a utilização da matriz de informação observada é preferida em relação à esperada, devido ao alto número de dados faltantes (RIZOPOULOS, 2012).

A solução das integrais com respeito ao tempo e efeito aleatório não apresentam solução analítica exigindo aproximações por métodos numéricos. Um método numérico para a aproximação das integrais em relação a $\boldsymbol{b_i}$ é a quadratura de Gauss-Hermite ou quadratura gaussiana, como pode ser visto em Pinheiro e Bates (1995).

Para a maximização do logaritmo da verossimilhança são necessários algoritmos como EM, Newton e híbrido que inicia com EM e ao passo que a convergência demora continua com quasi-Newton.

4.1.1 Problemas de convergencia

Em muitos casos podem ser verificados problemas de convergência, isto se dá devido a falta de linearidade dos perfis individuais. Para burlar esse problema são utilizados polinômios para aproximar as integrais com respeito ao tempo. O spline cúbico tem função

dada por

$$\log \lambda_0(t) = k_0 + \sum_{d=1}^{m} k_d \mathrm{B}_d(t, q), \tag{4.11}$$

onde k_d são os coeficientes *spline*, q denota o grau da função de base B-*spline* $\mathrm{B}(.)$, $m = m' + q - 1$, com número de nós internos dado por m'. Quanto maior for o número de nós internos maior é a flexibilidade em estimar o risco basal.

4.2 Inferência

4.2.1 Teste da razão de verossimilhança

A razão de verossimilhanças é basicamente utilizada para comparar modelos aninhados e o ajuste global como visto na Seção 2.4.1. O teste é um pouco mais trabalhoso computacionalmente, devido a necessidade de se ajustar os modelos sob-hipótese nula $l(\hat{\boldsymbol{\theta}}_{\mathbf{0}})$ e alternativa $l(\hat{\boldsymbol{\theta}})$. As hipóteses testadas são

$$\begin{cases} H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_{\mathbf{0}} \\ H_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\mathbf{0}}. \end{cases} \tag{4.12}$$

A estatística de teste é dada pela *Likelihood Ratio* (LRT) e escrita por

$$LRT = -2\left\{l(\hat{\boldsymbol{\theta}}_{\mathbf{0}}) - l(\hat{\boldsymbol{\theta}})\right\} \tag{4.13}$$

em que $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ e $\hat{\boldsymbol{\theta}}$ são os estimadores de máxima verossimilhança das respectivas hipóteses.

A estatística segue assintoticamente uma distribuição qui-quadrado χ^2_{gl}, com graus de liberdade igual ao número de covariáveis do modelo sob hipótese alternativa, isto se analisado o ajuste do modelo aos dados e a diferença entre o número de covariáveis dos modelos em questão, para a seleção do modelo. Segundo (RIZOPOULOS, 2012), o teste da razão de verossimilhança é preferido, devido o fato de levar em consideração a matriz de covariância dos efeitos aleatórios que não é considerada pelo teste de Wald. Na *library* JM, implementado por Rizopoulos (2010), a função $anova()$ é utilizada para o teste, onde deve ser especificado o modelo sob hipóteses nula e alternativa. A função retorna o AIC, BIC, o negativo do logaritmo da verossimilhança, de ambos modelos, além da LRT com seus graus de liberdade e valor p.

4.2.2 Teste de Wald

Mais simples que a LRT, a estatística de Wald testa o verdadeiro valor dos parâmetros com base no modelo de interesse, em que são utilizadas as estimativas obtidas pela amostra

e erros padrão das mesmas. Sendo assim, seja um vetor de parâmetros, $\boldsymbol{\theta} = (\theta_0, \theta_1, \cdots, \theta_p)$, as seguintes hipóteses devem ser testadas

$$\begin{cases} H_0 : \boldsymbol{\theta} = \theta_0 \\ H_a : \boldsymbol{\theta} \neq \theta_0. \end{cases} \tag{4.14}$$

O teste para um único parâmetro é dado por

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\hat{S}e(\hat{\theta})}, \tag{4.15}$$

W segue aproximadamente uma distribuição qui-quadrado.

Para p parâmetros, a estatística testa todos os parâmetros simultaneamente em relação à θ_0.

$$W = (\hat{\boldsymbol{\theta}} - \theta_0)^T I(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \theta_0), \tag{4.16}$$

sendo $I(\hat{\boldsymbol{\theta}})$, a matriz de informação de Fisher, a estatística segue assintoticamente uma distribuição qui-quadrado χ^2_{gl}, com graus de liberdade igual ao número de parâmetros testados. De acordo com Rizopoulos (2012), devido aos efeitos aleatórios não serem levados em consideração, a variabilidade dos estimadores acaba sendo subestimada fazendo com que a estatística seja menos confiável, além disso existem poucos estudos relacionados as propriedades assintóticas do teste para dados finitos quando se trata da metodologia dos modelos conjuntos. A estatística também pode ser obtida pela *library* JM, com função $anova()$, especificando o modelo e na sequência o processo, testando assim os parâmetros do processo longitudinal e de sobrevivência.

4.2.3 Intervalos de confiança

Os intervalos de confiança assintóticos para os modelos conjuntos podem ser baseados na estatística de Wald. A função $confint()$, com argumentos que especificam o processo como $"Longitudinal"$ ou de $"Event"$ e o nível de confiança adotado, resulta no intervalo descrito a baixo.

$$IC(\boldsymbol{\theta}, 100(1-\alpha)\%) \;=\; \hat{\boldsymbol{\theta}} \pm Z_{\alpha/2} Se(\hat{\boldsymbol{\theta}}). \tag{4.17}$$

Para o intervalo de confiança da razão de risco, basta exponenciar os valores obtidos pela fórmula 4.16.

4.3 Análise de Resíduos e Diagnóstico

Os resíduos utilizados para os modelos conjuntos estão relacionados basicamente com os modelos que fazem parte da articulação conjunta, isto é, de acordo com a natureza das variáveis resposta que se deseja relacionar, modelos específicos devem ser utilizados, de modo que, seus resíduos também sejam aproveitados. Será abordada a análise de resíduos por meios gráficos, entre os quais são apresentados os resíduos condicionais, marginais e de martingale.

4.3.1 Resíduos marginais

De acordo com Rizopoulos (2010), entre os resíduos mais utilizados estão os resíduos marginais padronizados, escritos por

$$r_i^{ym}(t_{ij}) = \frac{y_i(t_{ij}) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}}{\boldsymbol{V_i}^{-1/2}}, \tag{4.18}$$

que avaliam os erros do modelo marginal e os pressupostos da estrutura de correlação dentro do sujeito, com $\boldsymbol{V_i} = \boldsymbol{Z_i} Var(\boldsymbol{b_i})\boldsymbol{Z_i'} + Var(\boldsymbol{\epsilon}_i)$.

4.3.2 Resíduos condicionais

Os resíduos condicionais são expressos por

$$r_i^{yc} = y_i(t_{ij}) - \boldsymbol{X_i(t_{ij})}\hat{\boldsymbol{\beta}} - \boldsymbol{Z_i(t_{ij})}\hat{\boldsymbol{b}}_{\boldsymbol{i}}, \tag{4.19}$$

cuja função é analisar os pressupostos de homocedasticidade e normalidade dos dados.

4.3.3 Resíduos martingale

Para a parte de sobrevivência são verificados os resíduos de martingale, os quais buscam identificar sujeitos mal ajustados pelo modelo e verificar se a forma funcional da covariável é apropriada. O resíduo é escrito por

$$r_i^{tm} \;=\; N_i(t) - \left(\int_0^t R_i(s)\hat{\lambda}_0(s)\exp\left\{\gamma^T k_i + \alpha m_i(s)\right\} ds\right), \tag{4.20}$$

onde $N_i(t)$ é o número de eventos observados, o restante corresponde ao total de eventos esperados. Os possíveis valores aberrantes podem ser identificados de duas maneiras: i) Nos casos em que uma concentração de resíduos se posiciona acima da linha do zero, ou seja, quando o número de eventos observados é superior ao estimado pelo modelo e, ii) Valores abaixo de zero corresponde a valores observados inferiores aos estimados pelo

modelo. Para a forma funcional é desejada a existência de uma relação de linearidade entre os resíduos e a covariável, se isto não ocorrer podem ser necessárias transformações na covariável.

5 Aplicação à coorte de HIV coinfectada pelos vírus HBV e HCV

A análise dos dados, bem como o ajuste do modelo conjunto foram obtidos por meio do aplicativo R versão 3.3.3. Os comandos utilizados na análise descritiva foram: $survfit()$, $xyplot()$ dos *librarys* $survival$ e $latticeExtra$. Para o ajuste dos modelos foram utilizados: i) $lme()$ do *library* $nlme$, ii) $coxph()$ do *library* $survival$ e iii) $jointModel()$ do *library* JM, todos disponíveis em rede.

5.1 Análise descritiva dos dados

A Tabela 1 apresenta as cinco primeiras linhas da base de dados, que contém 147 indivíduos, dos quais, 109 pertencem ao grupo controle (somente HIV), 21 ao grupo de coinfecção por HBV e 18 aos coinfectados por HCV. É importante salientar que cada indivíduo da amostra tem múltiplas linhas. As colunas da base são compostas pelo indivíduo (pac), o tempo em dias (obstime) que ocorreu a contagem do biomarcador longitudinal, a contagem de células CD4, a contagem de CD8, a coluna indicadora de censura e evento (event), os grupos (controle, HBC e HCV) e tempo de ocorrência do evento (time).

Tabela 1 – Resumo dos dados contendo as 5 primeiras linhas.

pac	obstime	CD4	CD8	event	group	time
105	0	57	1530	0	HCV	4463
105	158	191	2000	0	HCV	4463
105	268	225	1946	0	HCV	4463
105	399	286	1624	0	HCV	4463
105	468	287	1430	0	HCV	4463
105	577	371	1890	0	HCV	4463

Pela variável resposta CD4 se tratar de dados de contagem, assumiu-se que sua variância era proporcional a média de modo que ao transformar a variável resposta utilizando a raiz quadrada, sendo possível estabilizar a variância (MILLAR, 2011) e obter uma distribuição aproximadamente normal.

Pelas Figuras 8 (a) e 8 (b), é possível verificar que a distribuição da variável CD4 é assimétrica em (a) e a variável transformada em (b) apresenta-se normalmente distribuída.

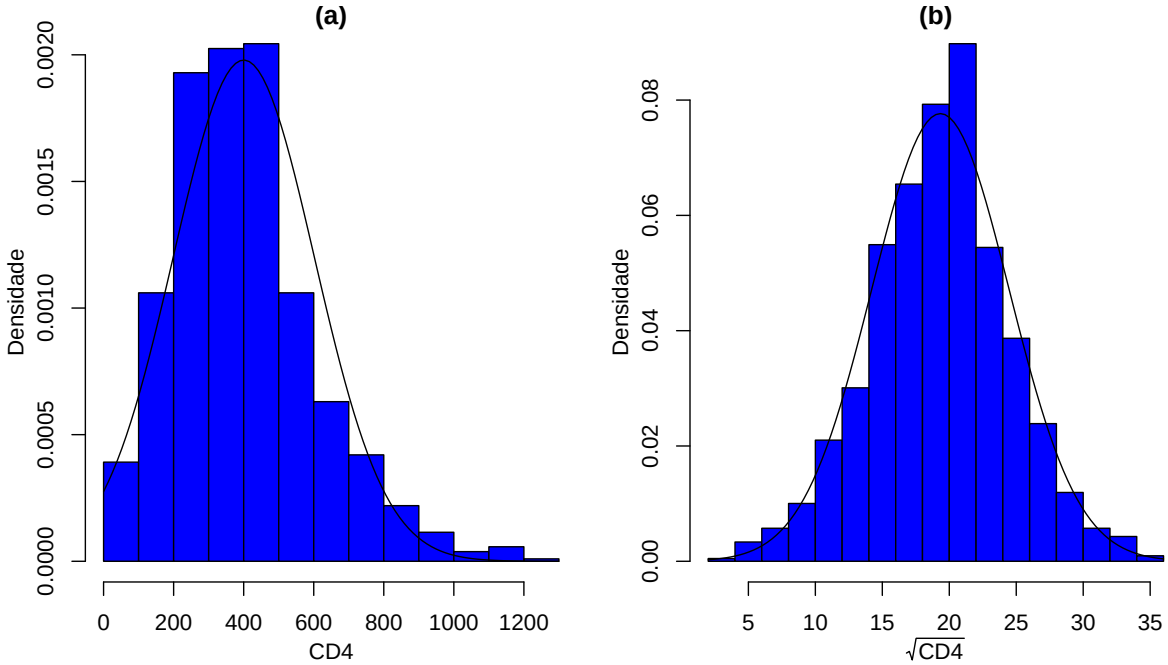

Figura 8 – Distribuição da variável CD4; (a) dados observados, (b) dados transformados por raiz quadrada.

Os perfis na Figura 9, são referentes aos 147 indivíduos que estão separados em: controle (HIV), coinfecções por HBV e HCV.

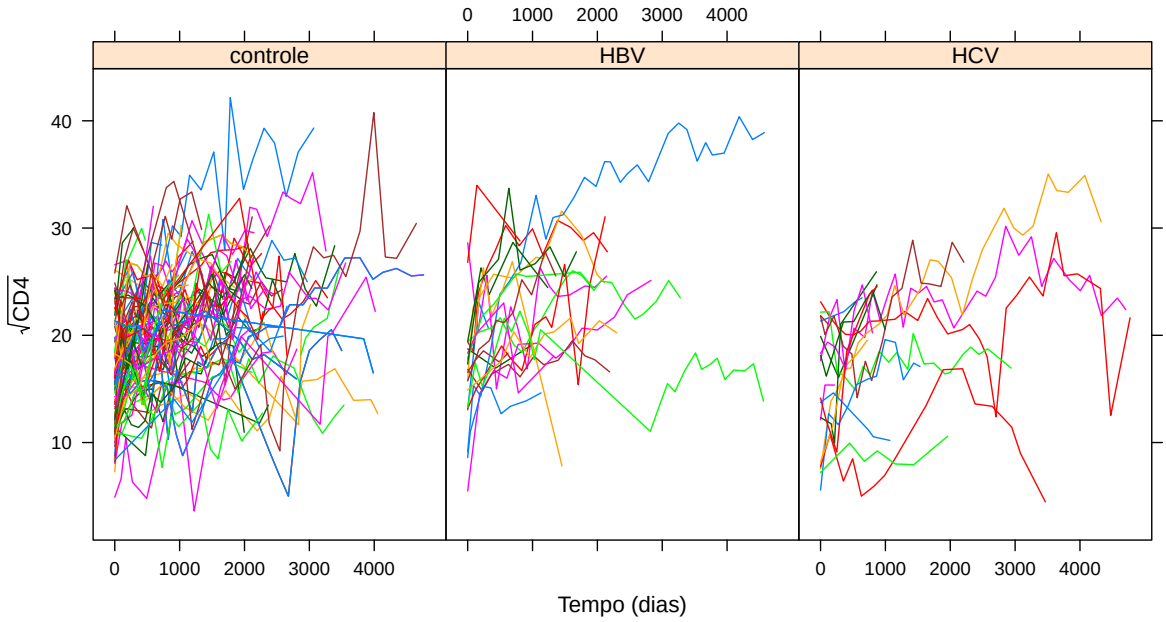

Figura 9 – Gráfico de perfis da raiz quadrada da contagem de CD4 para os diferentes grupos de coinfecção.

Na análise foi possível observar que alguns perfis individuais ao final do estudo se mantêm praticamente isolados. A partir de 1460 dias (4 anos), para o grupo controle

o número restante de indivíduos foi de 48, para a coinfecção por HBV restaram 11 e na coinfecção por HCV 7. Sendo assim é notório que o número de indivíduos para cada grupo teve uma redução maior que 50% no número total de indivíduos por grupo. Dessa forma, para que o comportamento solitário desses indivíduos ao final do estudo não pudessem influenciar a resposta média da raiz quadrada de CD4, a análise longitudinal considerou as observações coletadas até 1460 dias.

A partir dos perfis médios obtidos pelos dados observados de cada grupo, na Figura 10, pode-se perceber que em média indivíduos coinfectados por HBV apresentam níveis superiores ao controle e o grupo de coinfecção por HCV, valores inferiores. Além disso, parece haver uma recuperação da CD4 com o passar do tempo em ambos os grupos, quanto à inclinação não mostra ter grande variabilidade. Embora o comportamento dos perfis médios demonstrem uma não linearidade, buscou-se uma aproximação linear para os mesmos, mostrando ser possivelmente necessário o uso de funções B-*splines* para tornar o modelo mais flexível.

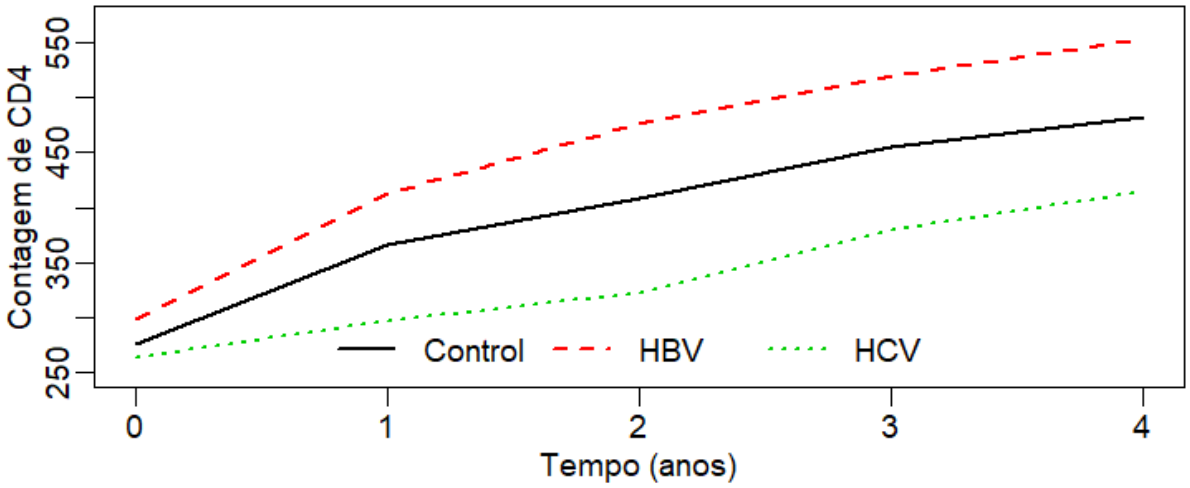

Figura 10 – Representação dos perfis médios dos dados observados.

A partir do diagrama de caixas (Figura 11), pode-se perceber heterogeneidade entre os grupos, assumindo que, em média, a coinfecção por HBV apresenta nível médios superior ao controle e o grupo de coinfecção por HCV, valores inferiores. Além disso, ambos os grupos (HBV e HCV) apresentam menor variabilidade em relação ao controle.

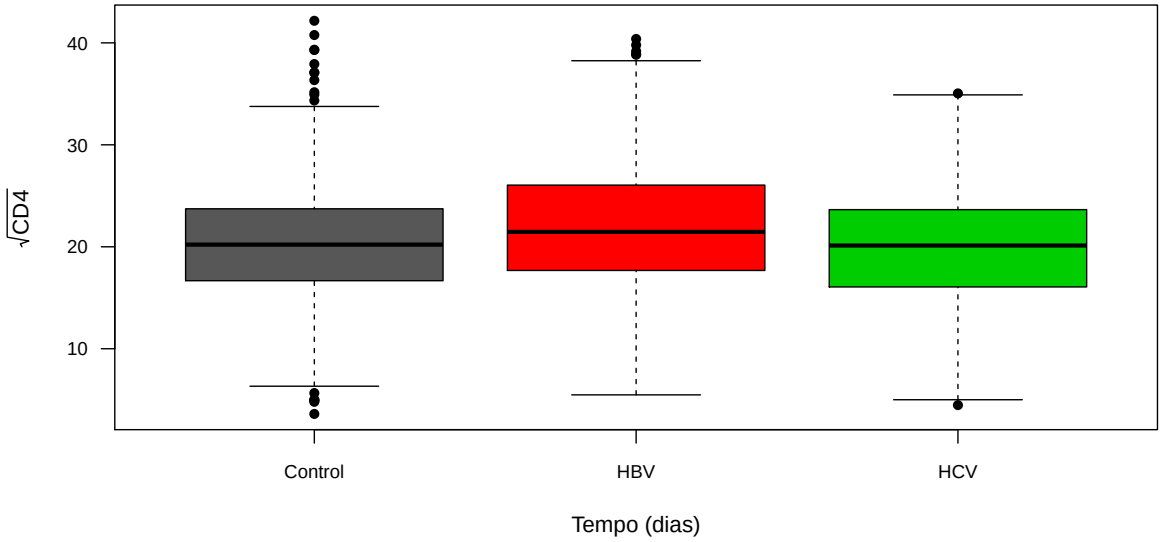

Figura 11 – Boxplot da raiz quadrada da contagem de CD4 para os diferentes grupos de coinfecção.

As curvas de sobrevivência, obtidas pelo estimador de Kaplan-Meier (Figura 12 (a)) dão uma ideia inicial das curvas de sobrevivência entre os grupos. É possível verificar que o controle e coinfecção por HCV têm menores probabilidades de sobrevivência ao longo do tempo com relação ao HBV. Neste estudo, curvas de sobrevivência inferiores implicam maior risco de atingir razões CD4:CD8 = 0,9. Além disso, é possível observar que a partir do dia 3000 parece haver uma leve violação do pressuposto de proporcionalidade.

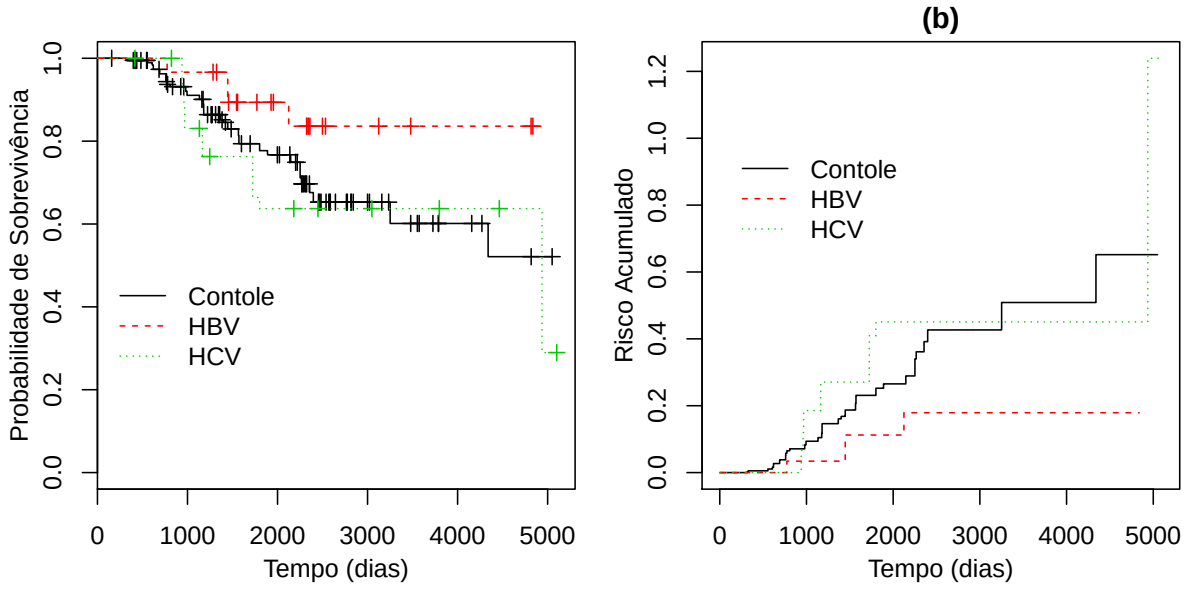

Figura 12 – (a) Curvas de sobrevivência estimadas por Kaplan-Meier segundo tipo de grupo. (b) Curvas de risco segundo tipo de grupo.

As estimativas de risco acumulado apresentadas na Figura 12 (b), obtidas por Nelson-Aalen, comprovam que o controle e a coinfecção por HCV aparentam riscos semelhantes e maiores que a coinfecção por HBV.

5.2 Modelagem conjunta

5.2.1 Seleção do modelo

Os modelos comparados foram:

- Mod1 descrevendo os efeitos do tempo e grupo.
- Mod2 acrescentando a interação entre os efeitos principais (tempo e grupo).

O Teste da razão de verossimilhança pode ser visto na Tabela 2, onde são apresentados os valores do AIC, BIC, LRT e Valor p.

Tabela 2 – Os valores do AIC, BIC e log Verossimilhança para os modelos Mod1 e Mod2.

	AIC	BIC	ln(Verossimil.)	LRT	df	Valor p
Mod1	6349.12	6408.92	-3154.56			
Mod2	6358.50	6418.31	-3159.25	-9.38	0	1

De acordo com os valores apresentados na seleção, é possível identificar o modelo que melhor descreve o comportamento do tempo até que a razão CD4:CD8 = 0,9 seja atingida. O modelo escolhido para continuar a análise foi o modelo Mod1 apresentado menor AIC e BIC.

A análise da *Deviance* também foi utilizada para obter uma estatística global do ajuste do modelo Mod1, o qual mostrou ajustar-se bem aos dados.

5.2.2 Submodelo longitudinal

Ao processo longitudinal, foi adotado um modelo de efeito aleatório normal multivariado, assumindo o uso de de funções B-$splines$ para descrever o comportamento do biomarcador longitudinal no tempo, a notação do modelo pode ser expressa por

$$\begin{aligned} y_i(t) &= m_i(t) + \epsilon_i \\ &= (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})\mathrm{B}_n(t, \lambda_1) + \beta_2 \{\mathrm{grupoHBV}\} + \beta_3 \{\mathrm{grupoHCV}\} + \epsilon_i, \end{aligned} \tag{5.1}$$

onde $y_i(t)$ denota a observação da raiz quadrada da contagem de CD4, para o i-ésimo indivíduo com $i = 1, \cdots, 147$, o real nível da resposta longitudinal é representado por $m_i(t)$, os resíduos $\epsilon_i(t) \sim N(0, \sigma^2 I_{n_i})$, sendo I_{n_i} a matriz identidade n_i-dimensional e n_i

o número de observações do i-ésimo indivíduo. O vetor $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ corresponde ao efeitos fixos, B$_n(t, \lambda_1)$ as funções B-$spline$ sem nós, $\epsilon_i(t)$ e $\boldsymbol{b}_i$ são independentes entre si, $\boldsymbol{b}_i \sim N(0, D_{(2\times 2)})$. A matriz de variância e covariância, $V_i = Z_i D Z_i^T + \sigma^2 I_{n_i}$, fica completamente especificada, sendo Z_i a matriz de variáveis explicativas com efeito aleatório.

No ajuste do modelo foi necessário a utilização de B-$splines$, seu uso é fundamentado devido a alguns indivíduos terem apresentado comportamento não linear em sua trajetória longitudinal. Essa conduta dificulta e gera problemas de convergência, os quais foram resolvidos utilizando *splines* que tornam o modelo mais flexível.

5.2.3 Submodelo de sobrevivência

Para estimar a medida de associação entre o nível do biomarcador e o risco do evento, o submodelo de riscos proporcionais foi adotado para estimar o tempo de sobrevivência, o modelo é expresso por

$$\lambda(t|m_i(t), K_i) \;=\; \lambda_0(t) \exp\left\{\boldsymbol{\gamma}^T \text{group}_i + \alpha m_i(t)\right\}, \tag{5.2}$$

onde $\lambda(t|m_i(t), K_i)$ corresponde ao risco de atingir razão DC4:CD8 = 0,9 (evento), dado o verdadeiro nível da raiz quadrada da contagem de CD4 ($m_i(t)$). O parâmetro α especifica a intensidade de ligação da função trajetória da raiz quadrada da contagem de CD4 ao risco do evento. As informações basais são descritas por K_i, matriz de covariáveis explicativas (grupo), com parâmetros de regressão dados pelo vetor $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)$, representando o efeito do grupo no risco de superar razão inferior a 0,9, e para estimar o risco basal $\lambda_0(t)$, foi utilizado B-$splines$ ao longo do tempo. Os efeitos foram obtidos baseados no tempo de início da CART, até o tempo em que a razão 0,9 foi atingida, os resultados podem ser vistos na Tabela 3.

As estimativas, erros padrão e valor p dos estimadores nos processos longitudinal e de sobrevivência podem ser observados na Tabela 3. O submodelo longitudinal aponta um efeito positivo do tempo sobre a $\sqrt{CD4}$ (5%, valor p $<$ 0,0001). O efeito da coinfecção por HBV mesmo sendo superior ao controle, não apresentou significância estatistica (5%, valor p = 0,0573). A coinfecção por HCV teve redução significativa de 2,52 (5%, valor p $<$ 0,0001) na taxa da $\sqrt{CD4}$. No processo de sobrevivência observou-se que a coinfecção por HBV, obteve um efeito negativo de 1,62 (5%, valor p $<$ 0,0186), apresentando risco relativo de 0,1974, correspondendo a um risco 80% menor de atingir razão CD4:CD8 = 0,9 com relação ao controle. Sendo assim, o grupo controle tem risco 5 vezes maior de atingir a razão igual a 0,9 em relação a coinfecção por HBV. A coinfecção por HCV, mesmo mostrando um efeito positivo no risco de atingir CD4:CD8 = 0,9, não foi estatísticamente significativa (5%, p-value = 0,2012). A associação entre a progressão da $\sqrt{CD4}$ ao longo

Tabela 3 – Estimadores, erros padrão e valor p para o modelo proposto e, os D_{ij} denota o ij-elemento da matriz de covariância dos efeitos aleatórios.

Joint Models				
	Coeficiente	Erro padrão		Valor p
(Intercept)	17,16161	0,3631		$< 0,0001$
ns(obstime, 1)	8,4849	0,6144		$< 0,0001$
groupHBV	1,0623	0,5588		$0,0573$
groupHCV	$-$2,5214	0,5948		$< 0,0001$
$\log(\sigma)$	0,9151	0,0254		
D_{11}	14,8551	2,0598		
D_{12}	$-$6,5043	3,3818		
D_{22}	57,5229	9,8258		
	Coeficiente	Erro padrão	Risco relativo (HR)	Valor p
groupHBV	$-$1,6222	0,6892	0,1974	$< 0,0186$
groupHCV	0,5209	0,4075	1,6835	$0,2012$
Assoct	0,1055	0,0286	1,1113	$< 0,0002$

do tempo e o risco de se atingir a razão igual a 0,9 foi significativa, com risco relativo de 1,11 (5%, p-value $<$ 0,0002).

É importante salientar que ao analisar a contagem da resposta ao longo do tempo sob um aspecto puramente longitudinal, os grupos não obtiveram valores significativos a 5%, dando a falsa impressão de que a recuperação de CD4 é semelhante entre os grupos, informação que se contradiz aos resultados encontrados pelo modelo conjunto. Os erros padrão obtidos para os efeitos estimados da evolução da $\sqrt{CD4}$, ao longo do tempo, foram superiores ao modelo conjunto, com intervalos de confiança maiores, comprovando menor precisão nas estimativas.

Com respeito à sobrevivência, se essa análise fosse realizada pelo modelo de Cox estendido, o qual considera as medidas longitudinais como uma variável medida no tempo, seria possível constatar que, de fato, o modelo de Cox subestima os erros das estimativas quando comparado ao modelo conjunto. De acordo com a extensão de Cox, para o aumento de uma unidade na $\sqrt{CD4}$, acarreta um acréscimo significativo (5%, p = 0.0047) de aproximadamente 8% no risco de atingir o evento, contra 11% estimado pelo modelo conjunto. Além disso, o modelo padrão de Cox com covariável medida no tempo não apresentou diferença significativa a 5% entre os grupos. Essas divergências, poderiam levar profissionais da saúde a alterar dosagens ou até mesmo medicamentos, quando não necessário, além de desconsiderar que uma coinfecção pode ser mais grave que outra.

5.2.4 Análise de Resíduos

O gráfico de resíduos condicionais versus valores estimados de $\sqrt{CD4}$, situado no lado esquerdo superior da Figura 13, mesmo apresentando alguns valores possivelmente

outlier, parece não violar os pressupostos de homocedasticidade e normalidade. De acordo com os resíduos marginais no lado direito superior revelam que a estrutura do modelo se apresenta adequada, representando bem a resposta longitudinal.

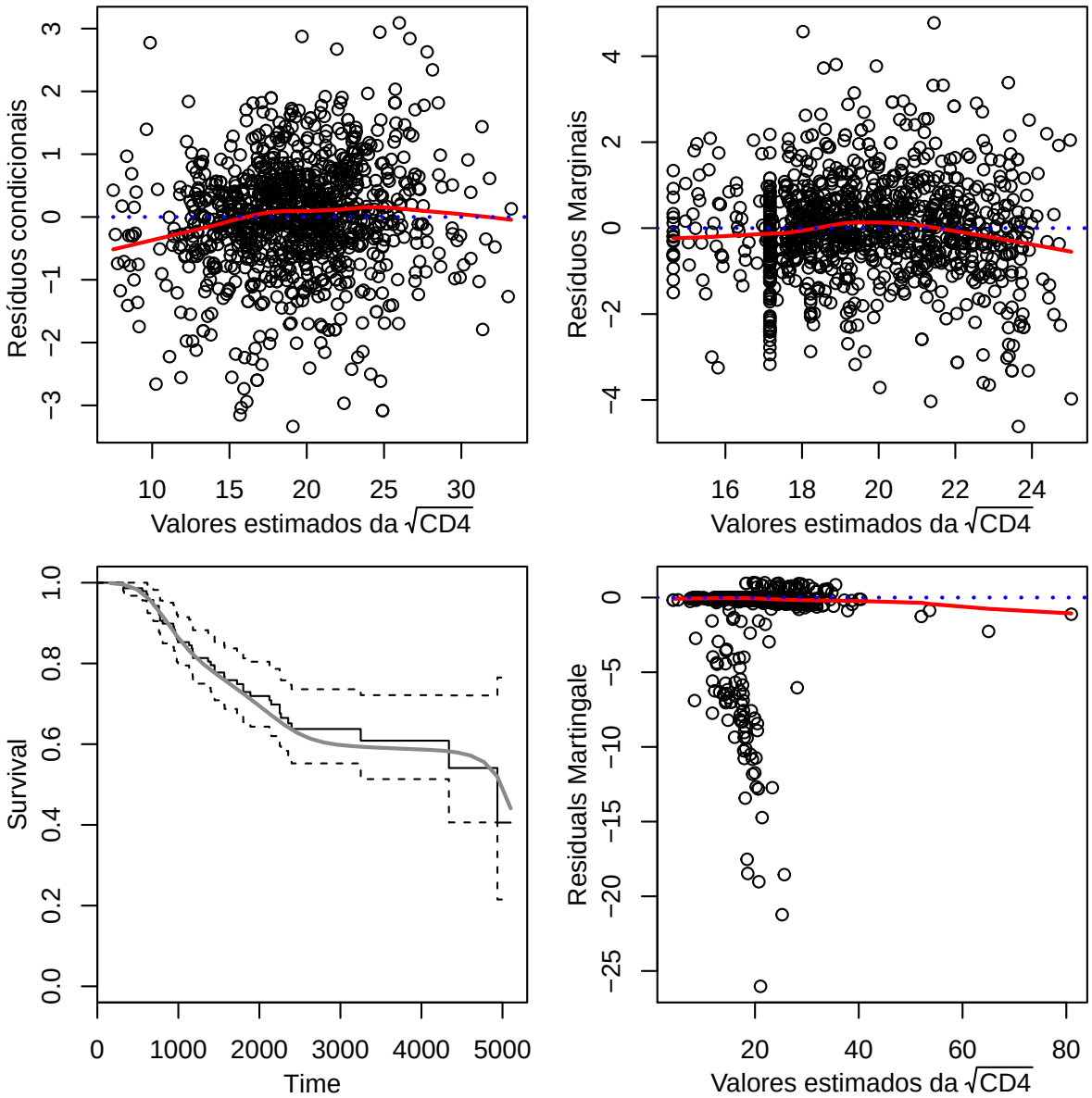

Figura 13 – Gráficos de resíduos: Na parte superior estão respectivamente os resíduos condicionais padronizados e resíduos marginais versus valores estimados para o processo longitudinal. Para a sobreviência a estimativa de Kaplan-Meier sobre a curva de sobrevivência marginal e os resíduos martingale versus as valores estimados.

O gráfico de sobrevivência marginal sobreposto pelas estimativas de Kaplan-Meier, no lado esquerdo inferior, demostram o quanto a estimativa não paramétrica se aproxima do modelo. Os resíduos martingale versus valores estimados de $\sqrt{CD4}$ mostram que a forma funcional escolhida para o marcador longitudinal se apresenta adequada, identificando uma relação linear, de modo que os parâmetros serão estimados corretamente. A medida de

efeito da $\sqrt{CD4}$ a cada unidade aumentada resulta em um efeito negativo na sobrevivência dos indivíduos. Redução de sobrevivência indica resultado favorável, já que implica maior risco de atingir uma razão considerada saudável. Também pode ser observado que alguns pontos, mesmo não se apresentando bem ajustados, foram mantidos na análise, pois não influenciaram nos efeitos e conclusões finais do estudo.

5.2.5 Discussões

Compreender a evolução natural dos biomarcadores que descrevem o HIV é de fundamental importância para as áreas médicas e farmacêuticas, além de acrescentar qualidade de vida aos pacientes. É por meio dessa evolução que se torna possível prever diferentes desfechos e intervir clinicamente. Neste estudo, foram abordados dois indicadores mais comumente usados para avaliar a saúde do Sistema de Ativação Imune, o CD4 e a razão CD4:CD8. O entendimento dos mesmos e sua relação quando observados no cenário da infecção pelo HIV e coinfecção por HBV e HCV têm se tornado, ao longo dos últimos anos, alvo de constantes estudos. Muitos resultados com respeito a gravidade da coinfecção são encontrados na literatura, entretanto não se tem uma concordância sobre qual apresenta maior risco no desenvolvimento de doenças hepáticas. Estudos que permitam quantificar a intensidade da relação entre a contagem de CD4 com o risco de alcançar uma razão saudável de CD4:CD8 ainda não estão bem definidos.

Bräu et al. (2006) mostraram que indivíduos coinfectados por HCV, quando submetidos a terapia antirretroviral, reduzem a progressão de doenças hepáticas. Pineda et al. (2007) também apontam diminuição de mortes relacionadas ao fígado em grupos sob as mesmas condições. O fato é que a recuperação do Sistema de Ativação Imune em pacientes com HIV positivo depende de níveis normais ou aproximadamente normais de CD4 e razão CD4:CD8.

Esses trabalhos, de certa forma, servem como suporte para o presente estudo. Aqui, mostrou-se que a coinfecção por HCV apresentou uma contagem de CD4 significativamente inferior ao grupo controle. Contudo, para este grupo, o risco de se atingir a razão CD4:CD8 = 0,9 não foi significativamente diferente do grupo controle.

É importante destacar que, na presença do HBV, os antirretrovirais têm eficiência reduzida (BRINKER et al., 2000). Outros estudos alertam para o fato de que o HBV se torna mais agressivo na presença do HIV, com maior reativação e progressão de doenças relacionadas ao fígado e aumento da mortalidade (THIO et al., 2002; KONOPNICKI et al., 2005). A replicação do HBV tem proporcionado maior risco de cirrose e carcinoma hepatocelular (HOFFMANN; THIO, 2007). Neste estudo, apesar de não significativa, a contagem de CD4 para o grupo coinfectado por HBV foi maior que o grupo controle. Com relação ao grupo controle, o HBV não evidenciou diferença significativa na progressão do biomarcador ao longo do tempo, porém o risco de atingir a razão CD4:CD8 = 0.9 foi

significativamente inferior.

Os resultados obtidos neste estudo apontam para uma severidade maior aos pacientes coinfectados por HCV sob o ponto de vista longitudinal (CD4), no entanto, sob o ponto de vista de sobrevivência (CD4:CD8), a maior severidade foi no grupo HBV. Portanto, mostrou-se a importância de se analisar conjuntamente os dois processos sob a perspectiva da análise clínica em indivíduos coinfectados.

5.3 Considerações Finais

O atual estudo analisou, por meio da metodologia dos modelos conjuntos, uma coorte retrospectiva de indivíduos infectados por HIV coinfectados por HBV e HCV. O intuito foi investigar a relação entre a evolução da $\sqrt{CD4}$ com o risco de diferentes grupos em atingir uma relação considerada saudável. Embora nossos dados apresentem algumas restrições, com quantidade de indivíduos e limitações em relação ao intervalo de tempo para as observações longitudinais, o modelo proposto se ajustou bem aos dados. A análise de resíduos se mostrou adequada, mesmo discriminando alguns indivíduos mal ajustados.

Também é notório que em nossos dados, sob uma perspectiva biológica, parecem ter um comportamento não linear, pois a contagem do biomarcador (CD4) tende assintóticamente a um determinado valor. Porém, como nosso objetivo foi capturar uma tendência aproximada para a evolução longitudinal, por meio de um modelo misto normal multivariado, obteve-se resultados satisfatórios na análise. Talvez, estudos futuros de coorte com maior número de indivíduos possam ser realizados agregando ainda mais credibilidade aos achados por essa pesquisa.

6 Apêndice A

```
*********************************************************************
**********************  Pacotes ***************************
*********************************************************************
library(JM)
library(latticeExtra)
```

```
*********************************************************************
********** Histograma e função de densidade de \sqrt(CD4) ***********
*********************************************************************
par(mfrow = c(1, 2), mar = c(3.5,3.5,0.5,0.5)+0.1,
                     mgp = c(2.1,1,0))

hist(cd4data.2$CD4,main = "(a)",
                 breaks = 15,
                    col = "8",
                 border = "black",
                   prob = T,
                   xlab = expression(CD4),
                   ylab = "Densidade")

curve(dnorm(x, mean = mean(cd4data.2$CD4,
              na.rm = TRUE),
                 sd = sd(cd4data.2$CD4, na.rm =TRUE)),
                col = 1,
                add = T)

hist(sqrt(cd4data.2$CD4), breaks = 15,
                             col = "8",
                          border = "black",
                            main = "(b)",
```

```
                            prob = T,
                            xlab = expression(sqrt(CD4)),
                            ylab = "Densidade")

curve(dnorm(x, mean = mean(sqrt(cd4data.2$CD4),na.rm =TRUE),
                 sd = sd(sqrt(cd4data$CD4),
              na.rm = TRUE)),
                col = 1,
                add = T)

*********************************************************************
********************** Gráfico de perfis ***************************
*********************************************************************
library("lattice")
xyplot(sqrt(CD4) ~ obstime | group,
           group = pac,
            data = datta,
            xlab = "Tempo (dias)",
            ylab = expression(sqrt("CD4")),
           strip = function(bg = 'white', ...)
strip.default(bg = 'white', ...),
             col = "black",
            type = 'l',
          scales = list(x=list(at=c(0,1000,2000,3000, 4000))))

*******************************************************************
****************** Curvas de Sobrevivência *************************
*******************************************************************
par(mfrow = c(1, 2),mar=c(3.5,3.5,1.5,0.5)+0.1,mgp=c(2.1,1,0))
plot(survfit(Surv(time, event) ~ group,
                         data = cd4data.2),
                     conf.int = FALSE,
                    mark.time = TRUE,
                         main = "(a)",
                          col = c(gray(0.8), gray(0.1), gray(0.3)),
                          lty = 1:3,
                         ylab = "Probabilidade de Sobrevivência",
                         xlab = "Tempo (dias)")
legend(0,0.5, c("Contole", "HBV","HCV"),
       lty = 1:3,
       col = c(gray(0.8), gray(0.1), gray(0.5)),
       bty = "n")
```

```
plot(survfit(Surv(time, event) ~ group,
            data =cd4data.2),
            lty = 1:3,
            fun = "cumhaz",
            col = c(gray(0.8), gray(0.1), gray(0.4)),
           main = "(b)",
           ylab = "Risco Acumulado",
           xlab = "Tempo (dias)")
legend(0,1, c("Contole", "HBV","HCV"),
       lty = 1:3,
       col = c(gray(0.8), gray(0.1), gray(0.3)),
       bty = "n")

*********************************************************************
********************** Ajuste do modelo ******************************
*********************************************************************
# Submodelo longitudinal
lmeFit  <- lme(sqrt(CD4)~ ns(obstime,1)+group,
                random = ~ns(obstime,1)|pac,
                  data = cd4data.2)

# Submodelo de sobrevivência

cd4data.id  <- cd4data.2[!duplicated(cd4data.2$pac),]
S.Cox       <- coxph(Surv(time, event) ~ group,
                                 data = cd4data.id,
                                    x = TRUE )

# Joint Models
fitJM <- jointModel(lmeFit,S.Cox,
                    timeVar = "obstime",
                     method = "spline-PH-GH")

*********************************************************************
******************** Gráficos de resíduos ****************************
*********************************************************************
par(mfrow = c(2, 2),mar=c(3.5,3.5,0.3,0.5)+0.1,mgp=c(2.1,1,0))

******Resíduos sujeito-específico (condicionais) padronizado**
 resSubY <- residuals(fitJM, process = "Longitudinal",
                                type = "stand-Subject")
#Valores estimados
fitSubY <- fitted(fitJM, process = "Longitudinal",
                            type = "Subject")
plotResid(fitSubY, resSubY,
        xlab = expression(paste("Fitted ", sqrt(CD4), " Values")),
        ylab = "Resíduos condicionais ")
```

```
#*********************** Resíduos Marginais *************************
resMargY <- residuals(fitJM,
            process = "Longitudinal",
               type = "stand-Marginal")
fitMargY <- fitted(fitJM,
            process = "Longitudinal",
               type = "Marginal")
plotResid(fitMargY, resMargY,
               xlab = expression(paste("Fitted ", sqrt(CD4), " Values")),
               ylab = "Marginal Residuals")
#***************** Curva da sobrevivência marginais *****************
plot(fitJM,  which = 3, add.KM = T,
                          col = "gray54",
                      caption = c(),
                          lwd = 2)
#************************ Martingale ********************************
resMartT <- residuals(fitJM,
              process = "Event",
                 type = "Martingale")
fitSubY <- fitted(fitJM,
              process = "Longitudinal",
                 type = "EventTime")
plotResid(fitSubY, resMartT,
             xlab = expression(paste("Fitted ", sqrt(CD4), " Values")),
***************************** FIM ***************************************
```

Referências

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974.

ASAR, Ö.; RITCHIE, J.; KALRA, P. A.; DIGGLE, P. J. Joint modelling of repeated measurement and time-to-event data: an introductory tutorial. **International journal of epidemiology**, Oxford University Press, v. 44, n. 1, p. 334–344, 2015.

BAIN, B. J. Practical flow cytometry in haematology diagnosis. by m. leach, m. drummond and a. doig, wiley-blackwell, oxford, 2013. **British Journal of Haematology**, Wiley Online Library, v. 167, n. 5, p. 727–727, 2014.

BONACINI, M.; PUOTI, M. Hepatitis c in patients with human immunodeficiency virus infection: diagnosis, natural history, meta-analysis of sexual and vertical transmission, and therapeutic issues. **Archives of Internal Medicine**, American Medical Association, v. 160, n. 22, p. 3365–3373, 2000.

BRÄU, N.; SALVATORE, M.; RíOS-BEDOYA, C. F.; FERNáNDEZ-CARBIA, A.; PARONETTO, F.; RODRíGUEZ-ORENGO, J. F.; RODRíGUEZ-TORRES, M. Slower fibrosis progression in hiv/hcv-coinfected patients with successful hiv suppression using antiretroviral therapy. **Journal of hepatology**, Elsevier, v. 44, n. 1, p. 47–55, 2006.

BRINKER, M. den; WIT, F. W.; DILLEN, P. M. Wertheim-van; JURRIAANS, S.; WEEL, J.; LEEUWEN, R. van; PAKKER, N. G.; REISS, P.; DANNER, S. A.; WEVERLING, G. J. et al. Hepatitis b and c virus co-infection and the risk for hepatotoxicity of highly active antiretroviral therapy in hiv-1 infection. **Aids**, LWW, v. 14, n. 18, p. 2895–2902, 2000.

BROWN, H.; PRESCOTT, R. **Applied mixed models in medicine**. [S.l.]: John Wiley & Sons, 2014. 53–56 p.

CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA, S. E. **Análise de Sobrevivência: teoria e aplicações em saúde**. [S.l.]: SciELO-Editora FIOCRUZ, 2011.

CHRISTENSEN, R.; PEARSON, L. M.; JOHNSON, W. Case-deletion diagnostics for mixed models. **Technometrics**, Taylor & Francis, v. 34, n. 1, p. 38–45, 1992.

CLEVES, M. **An introduction to survival analysis using Stata**. [S.l.]: Stata Press, 2008. 223 p.

COLLETT, D. **Modelling survival data in medical research**. [S.l.]: CRC press, 2003.

COLLETT, D. **Modelling survival data in medical research**. [S.l.]: CRC press, 2015.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Edgard Blücher, 2006. (ABE - Projeto Fisher).

COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. [S.l.]: CRC Press, 1979.

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.

DAVID, C. R. et al. Regression models and life tables (with discussion). **Journal of the Royal Statistical Society**, v. 34, p. 187–220, 1972.

DEEKS, S. G. Hiv infection, inflammation, immunosenescence, and aging. **Annual review of medicine**, Annual Reviews, v. 62, p. 141–155, 2011.

DEEKS, S. G.; TRACY, R.; DOUEK, D. C. Systemic effects of inflammation on health during chronic hiv infection. **Immunity**, Elsevier, v. 39, n. 4, p. 633–645, 2013.

DIGGLE, P. **Analysis of longitudinal data**. [S.l.]: Oxford University Press, 2002.

DIGGLE, P. J. An approach to the analysis of repeated measurements. **Biometrics**, JSTOR, p. 959–971, 1988.

FAUSTO, M. A.; CARNEIRO, M.; ANTUNES, C. M. d. F.; PINTO, J. A.; COLOSIMO, E. A. O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. 2008.

FITZMAURICE, G. M.; LAIRD, N. M.; WARE, J. H. Applied longitudinal analysis. 2004. **Hoboken Wiley-Interscience**, p. 60,62, 2004.

FOULLEY, J.-L.; JAFFRÉZIC, F.; ROBERT-GRANIÉ, C. Em-reml estimation of covariance parameters in gaussian mixed models for longitudinal data analysis. **Genetics Selection Evolution**, BioMed Central, v. 32, n. 2, p. 129, 2000.

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.

HENRIK, M.; POUL, T. **Introduction to general and generalized linear models**. [S.l.]: CRC Press, 2011.

HOFFMANN, C. J.; THIO, C. L. Clinical implications of hiv and hepatitis b co−infection in asia and africa. **The Lancet infectious diseases**, Elsevier, v. 7, n. 6, p. 402–409, 2007.

HSIEH, F.; TSENG, Y.-K.; WANG, J.-L. Joint modeling of survival and longitudinal data: likelihood approach revisited. **Biometrics**, Wiley Online Library, v. 62, n. 4, p. 1037–1043, 2006, apud RIZOPOULOS, 2012.

IBRAHIM, J. G.; CHU, H.; CHEN, L. M. Basic concepts and methods for joint models of longitudinal and survival data. **Journal of Clinical Oncology**, American Society of Clinical Oncology, v. 28, n. 16, p. 2796–2801, 2010.

KONOPNICKI, D.; MOCROFT, A.; WIT, S. D.; ANTUNES, F.; LEDERGERBER, B.; KATLAMA, C.; ZILMER, K.; VELLA, S.; KIRK, O.; LUNDGREN, J. D. et al. Hepatitis b and hiv: prevalence, aids progression, response to highly active antiretroviral therapy and increased mortality in the eurosida cohort. **Aids**, LWW, v. 19, n. 6, p. 593–601, 2005.

KUNIHOLM, M. H.; O'BRIEN, T. R.; PROKUNINA-OLSSON, L.; AUGENBRAUN, M.; PLANKEY, M.; KARIM, R.; SARKAR, M.; FRENCH, A. L.; PIERCE, C.; STRICKLER, H. D. et al. Association of hepatitis c virus infection with cd4/cd8 ratio in hiv-positive women. **Journal of acquired immune deficiency syndromes (1999)**, Wolters Kluwer Health, v. 72, n. 2, p. 162, 2016.

LANIER, R. **AIDS testing-A comprehensive guide to technical, medical, social, legal and management issues, -Schochetman, G, George, JR**. [S.l.]: GUILFORD PUBLICATIONS INC 72 SPRING STREET, NEW YORK, NY 10012, 1995.

LEE, E. T.; GO, O. T. Survival analysis in public health research. **Annual review of public health**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 18, n. 1, p. 105–134, 1997.

LEE, S. S.; WONG, N. S.; WONG, B. C. K.; WONG, K. H.; CHAN, K. C. W. Combining cd4 recovery and cd4: Cd8 ratio restoration as an indicator for evaluating the outcome of continued antiretroviral therapy: an observational cohort study. **BMJ open**, British Medical Journal Publishing Group, v. 7, n. 9, p. e016886, 2017.

LEE, Y.; NELDER, J. A. Hierarchical generalized linear models. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 619–678, 1996.

LU, W.; MEHRAJ, V.; VYBOH, K.; CAO, W.; LI, T.; ROUTY, J.-P. Cd4: Cd8 ratio as a frontier marker for clinical outcome, immune dysfunction and viral reservoir size in virologically suppressed hiv-positive patients. **Journal of the International AIDS Society**, The International AIDS Society, v. 18, n. 1, 2015.

MAINDONALD, J.; BRAUN, J. **Data analysis and graphics using R: an example-based approach**. [S.l.]: Cambridge University Press, 2006. v. 10.

MCLEAN, R. A.; SANDERS, W. L.; STROUP, W. W. A unified approach to mixed linear models. **The American Statistician**, Taylor & Francis, v. 45, n. 1, p. 54–64, 1991.

MILLAR, R. B. **Maximum likelihood estimation and inference: with examples in R, SAS and ADMB**. [S.l.]: John Wiley & Sons, 2011. v. 111.

MOORE, D. F. **Applied Survival Analysis Using R**. 1. ed. Springer International Publishing, 2016. (Use R!). ISBN 978-3-319-31243-9, 978-3-319-31245-3. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=93D2C227D802FAFFE3154E2F70894777>.

NOBRE, J. S.; SINGER, J. da M. Residual analysis for linear mixed models. **Biometrical Journal**, Wiley Online Library, v. 49, n. 6, p. 863–875, 2007.

ORGANIZATION, W. H. et al. **Global hepatitis report 2017**. [S.l.]: World Health Organization, 2017. 11 p.

PEARSON, J. D.; MORRELL, C. H.; LANDIS, P. K.; CARTER, H. B.; BRANT, L. J. Mixed-effects regression models for studying the natural history of prostate disease. **Statistics in medicine**, Wiley Online Library, v. 13, n. 5-7, p. 587–601, 1994.

PINEDA, J. A.; GARCÍA-GARCÍA, J. A.; AGUILAR-GUISADO, M.; RÍOS-VILLEGAS, M. J.; RUIZ-MORALES, J.; RIVERO, A.; VALLE, J. del; LUQUE, R.; RODRÍGUEZ-BAÑO, J.; GONZÁLEZ-SERRANO, M. et al. Clinical progression of hepatitis c virus–related chronic

liver disease in human immunodeficiency virus–infected patients undergoing highly active antiretroviral therapy. **Hepatology**, Wiley Online Library, v. 46, n. 3, p. 622–630, 2007.

PINHEIRO, J.; BATES, D. Mixed-effects models in s and s-plus springer-verlag new york google scholar. 2000.

PINHEIRO, J. C.; BATES, D. M. Approximations to the log-likelihood function in the nonlinear mixed-effects model. **Journal of computational and Graphical Statistics**, Taylor & Francis, v. 4, n. 1, p. 12–35, 1995.

RIBEIRO, M. H. D. M.; MILANI, H.; PREVIDELLI, I. Mixed models in cerebral ischemia study. **Acta Scientiarum. Technology**, Universidade Estadual de Maringá, v. 38, n. 3, 2016.

RIZOPOULOS, D. Jm: An r package for the joint modelling of longitudinal and time-to-event data. **Journal of Statistical Software (Online)**, v. 35, n. 9, p. 1–33, 2010.

RIZOPOULOS, D. **Joint models for longitudinal and time-to-event data: With applications in R**. [S.l.]: CRC Press, 2012.

ROBINSON, G. K. That blup is a good thing: the estimation of random effects. **Statistical science**, JSTOR, p. 15–32, 1991.

SAÚDE, M. da. **Boletim Epidemiológico de Hepatites Virais**. 2017. 21 p. Disponível em: <http://www.aids.gov.br/pt-br/pub/2017/boletim-epidemiologico-de-hepatites-virais>.

SCHABENBERGER, O.; PIERCE, F. J. **Contemporary statistical models for the plant and soil sciences**. [S.l.]: CRC press, 2001.

SEARLE, S. R. **Linear Models. New York: John Wylie & Sons**. [S.l.]: Inc, 1971.

SEARLE, S. R.; CASELLA, G.; MCCULLOCH, C. E. **Variance components**. [S.l.]: John Wiley & Sons, 2009. v. 391. 138–142 p.

SHERMAN, K. E. Hcv and hiv: a tale of two viruses. **Reviews in gastroenterological disorders**, v. 4, p. S48–54, 2004.

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

STRAM, D. O.; LEE, J. W. Variance components testing in the longitudinal mixed effects model. **Biometrics**, JSTOR, p. 1171–1177, 1994.

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.

SUÁREZ, C. C.; KLEIN, N.; KNEIB, T.; MOLENBERGHS, G.; RIZOPOULOS, D. Editorial "joint modeling of longitudinal and time-to-event data and beyond". **Biometrical Journal**, Wiley Online Library, v. 59, n. 6, p. 1101–1103, 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.

TABLEMAN, M.; KIM, J. S. **Survival analysis using S: analysis of time-to-event data**. [S.l.]: CRC press, 2003. 164–165 p.

TERRIN, N.; RODDAY, A. M.; PARSONS, S. K. Joint models for predicting transplant-related mortality from quality of life data. **Quality of Life Research**, Springer, v. 24, n. 1, p. 31–39, 2015.

THIO, C. L.; SEABERG, E. C.; JR, R. S.; PHAIR, J.; VISSCHER, B.; MUÑOZ, A.; THOMAS, D. L. Hiv-1, hepatitis b virus, and risk of liver-related mortality in the multicenter cohort study (macs). **The Lancet**, Elsevier, v. 360, n. 9349, p. 1921–1926, 2002.

VERBEKE, G. Linear mixed models for longitudinal data. In: **Linear mixed models in practice**. [S.l.]: Springer, 1997. p. 63–153.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models in practice: a SAS-oriented approach. Vol. 126**. [S.l.]: Springer-Verlag New York, LLC, New York, NY, 1997.

VIRAIS, M. da Saúde Secretaria de Políticas de Saúde Coordenação Nacional de DST e Aids Programa Nacional de H. **Recomendações para Tratamento da Co-Infecção entre HIV e Hepatites Virais**. 2002. 7 p. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/hepatite_010.pdf>.

WATERNAUX, C.; LAIRD, N. M.; WARE, J. H. Methods for analysis of longitudinal data: blood-lead concentrations and cognitive development. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 84, n. 405, p. 33–41, 1989.

WEST, B. T.; WELCH, K. B.; GALECKI, A. T. **Linear mixed models: a practical guide using statistical software**. [S.l.]: CRC Press, 2014.

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

ZHANG, D.; CHEN, M.-H.; IBRAHIM, J. G.; BOYE, M. E.; SHEN, W. Jmfit: a sas macro for joint models of longitudinal and survival data. **Journal of statistical software**, NIH Public Access, v. 71, n. 3, 2016.