UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos de sobrevivência induzidos por fragilidade

Michele Maciel Sacramento

Tese de Doutorado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Michele Maciel Sacramento

Modelos de sobrevivência induzidos por fragilidade

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutora em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Vicente Garibay Cancho

USP – São Carlos
Janeiro de 2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Sacramento, Michele Maciel
S123m Modelos de sobrevivência induzidos por fragilidade / Michele Maciel Sacramento; orientador Vicente Garibay Cancho. -- São Carlos, 2025.
148 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. Modelo de fragilidade. 2. Distribuições PVF. 3. Modelo de longa duração. 4. Modelo defeituoso. 5. Inferência bayesiana. I. Cancho, Vicente Garibay, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Michele Maciel Sacramento

Model survival induced frailty

Thesis submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Doctor in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Vicente Garibay Cancho

USP – São Carlos
January 2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Tese de Doutorado da candidata Michele Maciel Sacramento, realizada em 22/11/2024.

Comissão Julgadora:

Prof. Dr. Vicente Garibay Cancho (USP)

Profa. Dra. Elizabeth Mie Hashimoto (UTFPR)

Prof. Dr. Edwin Moises Marcos Ortega (ESALQ/USP)

Profa. Dra. Daiane de Souza Santos (USP)

Prof. Dr. Roberto Vila Gabriel (UnB)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Este trabalho é dedicado aos meus pais, irmãos e minha avó Lia,
por seu amor incondicional, e ao meu anjo Caike, que se eternizou em nossos corações.

AGRADECIMENTOS

Agradeço a Deus e a Virgem Maria por serem meus guias em todas as jornadas, minha força, por iluminar meu caminho e me proporcionar momentos e vitórias inesquecíveis.

Agradeço aos meus familiares pelo apoio incondicional, em especial, a meus pais, Paulo Cesar e Gorete e a minha avó Lia, que são minha base e minha maior inspiração. Obrigada pelo amor, compreensão e força! Por lutar minhas lutas e sonhar meus sonhos.

Aos meus irmãos, Julio e Milena, e meus primos, Heitor e João Victor, agradeço pela confiança depositada em mim, o amor dado, a presença constante e os sorrisos roubados. Vocês são o motivo da minha existência!

Aos amigos que tive o prazer de conquistar em São Carlos, principalmente a Luana, Leonardo, Camila, Marcílio, Fernando e Eliane. Obrigada por cada dia, cada história, por cada sorriso, cada apoio e cada diversão. Vocês são os maiores presentes que ganhei nesses anos.

Aos amigos de longa estrada, em especial a Elizandra, agradeço por todo amparo, conselhos, compreensão e amizade. Sem vocês minha vida não teria sentido!

Aos colegas e amigos que conheci durante o doutorado agradeço pelo tempo de convívio, por cada descoberta e conquista. Dividimos incertezas e inseguranças, mas sempre com força e alegria.

Aos professores do programa PIPGEs agradeço pela ajuda e apoio em minha formação. Em especial, ao meu orientador, o professor Dr. Vicente Garibay Cancho pelo respeito, paciência, empenho, dedicação e competência com que conduziu a orientação deste trabalho.

Aos professores, membros da banca examinadora, agradeço por aceitarem o convite e pelas sugestões e contribuições ao nosso trabalho.

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

“A persistência é o caminho do êxito.”
(Charles Chaplin)

RESUMO

SACRAMENTO, M. M. **Modelos de sobrevivência induzidos por fragilidade**. 2025. 148 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Aplicações em análise de sobrevivência podem incorporar covariáveis, como idade, sexo, gravidade da doença ou dados laboratoriais, que são conhecidas. Mas existem muitos outros fatores desconhecidos que podem influenciar na sobrevivência do indivíduo, incluindo o estado de saúde, estilo de vida, tabagismo, ocupação e fatores de risco genéticos. Estes fatores são denominados como fragilidade e controlam a heterogeneidade não observada dos indivíduos do estudo. Deste modo, o objetivo deste trabalho é desenvolver modelos de fragilidade para modelar essa heterogeneidade não observada em dados de sobrevivência. Neste contexto, a Família de Funções de Variância de Potência (PVF) será considerada para a modelagem da heterogeneidade não observada destes dados e o modelo Exponencial por Partes (MEP) será admitido como função de risco base. A proposta é flexível, pois a distribuição PVF inclui os principais modelos de fragilidade como casos particulares e, por sua vez, o modelo MEP constitui uma alternativa semiparamétrica às distribuições paramétricas, além de ser um recurso amplamente utilizado devido à sua capacidade de acomodar funções de risco em diferentes formas, sem a necessidade de impor restrições para obter um ajuste adequado do modelo aos dados. Como consequência, os modelos apresentados são estendidos para permitir a construção de modelos defeituosos e de longa duração univariados, resultando em uma fragilidade zero, no qual, é possível determinar a proporção de indivíduos imunes ao evento de interesse nos estudos de sobrevivência. Além disso, é introduzido um modelo de longa duração bivariado com fragilidade, utilizando as distribuições de mistura de Poisson e da família PVF. Este modelo é aprimorado, gerando um modelo de regressão que permite avaliar o impacto das covariáveis em dados de sobrevivência. A abordagem inferencial é baseada em métodos bayesianos através da utilização do método Hamiltoniano de Monte Carlo implementado em R-Stan, onde alguns resultados de simulação são fornecidos para avaliar o desempenho de algumas propriedades dos estimadores de Bayes. A importância dos modelos é ilustrada através de aplicações em conjuntos de dados reais.

Palavras-chave: Modelo de fragilidade, Distribuições PVF, Modelo de longa duração, Modelo defeituoso, Inferência bayesiana.

ABSTRACT

SACRAMENTO, M. M. **Model survival induced frailty**. 2025. 148 p. Tese (Doutorado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2025.

Applications in survival analysis can incorporate covariates, such as age, sex, disease severity, or laboratory data, which are known. However, there are many other unknown factors that can influence an individual's survival, including health status, lifestyle, smoking, occupation, and genetic risk factors. These factors are referred to as frailty and control the unobserved heterogeneity of the study's individuals. Thus, the objective of this work is to develop frailty models for modeling unobserved heterogeneity in survival data. In this context, the Family of Power Variance Functions (PVF) will be considered for modeling the unobserved heterogeneity of these data and the Piecewise Exponential Model (MEP) will be adopted as the baseline hazard function. The proposed approach is flexible, as the PVF distribution includes major frailty models as special cases and, in turn, the MEP model provides a semiparametric alternative to parametric distributions, being widely used due to its ability to accommodate hazard functions in different forms, without the need to impose restrictions to achieve a good fit to the data. Consequently, the presented models are extended to allow the construction of defective and long-duration univariate models, resulting in zero frailty, where it is possible to determine the proportion of individuals immune to the event of interest in survival studies. Furthermore, a bivariate long-duration frailty model is introduced, using Poisson mixture and the PVF family distributions. This model is enhanced, generating a regression model that allows the assessment of the impact of covariates on survival data. The inferential approach is based on bayesian methods using the Hamiltonian Monte Carlo method implemented in R-Stan, where some simulation results are provided to evaluate the performance of certain properties of the Bayes estimators. The importance of models is illustrated through applications to real data sets.

Keywords: Fragility Model, PVF Distributions, Long-term model, Defective Model, Bayesian Inference.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARVs	antirretrovirais
BCH	*Bounded cumulative hazard*
CPO	*Conditional Predictive Ordinate*
DP	desvio padrão
f.d	função densidade
f.d.p	função densidade de probabilidade
f.g.p	função geradora de probabilidade
HMC	Hamiltoniano de Monte Carlo
HPD	highest probability density
IG	Inversa Gaussiana
LPML	logaritmo da função de verossimilhança pseudo marginal
MCMC	Monte Carlo via Cadeias de Markov
MEP	Modelo Exponencial por Partes
MH	Metropolis-Hastings
PE	Positiva Estável
POI	*prior opportunistic infection*
PVF	Power Variance Function
REQM	raiz do erro quadrático médio
TRI	Teoria de Resposta ao Item

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

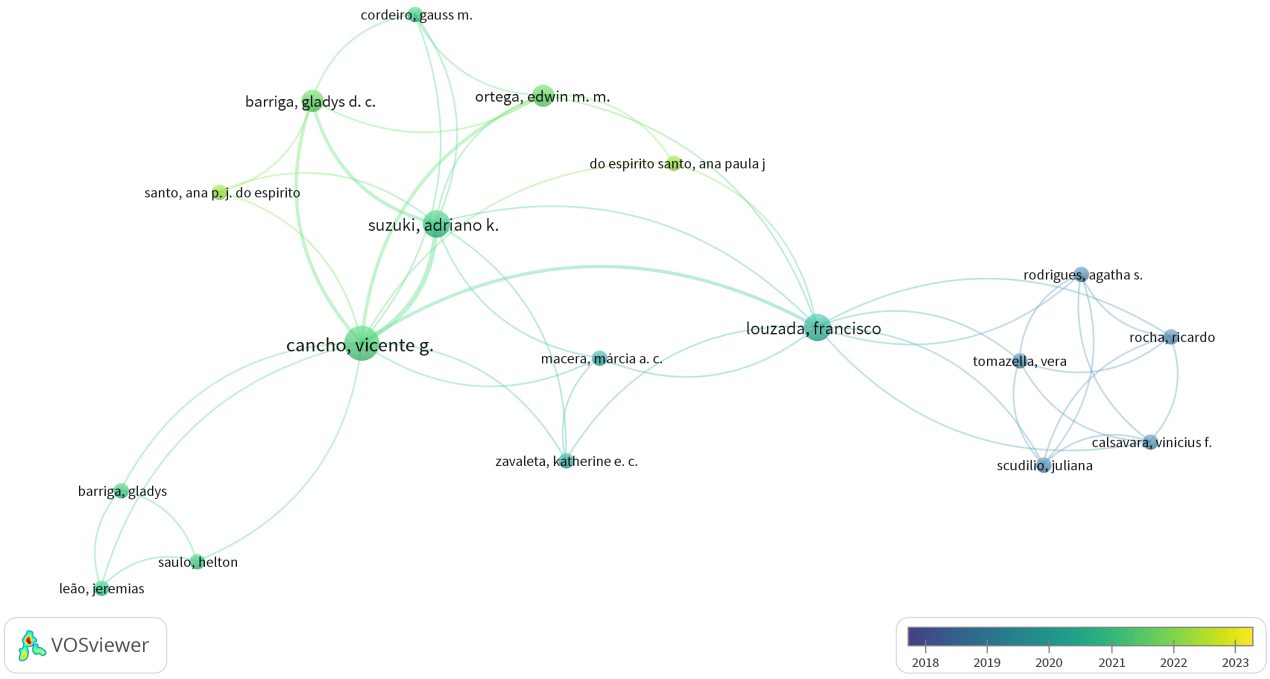
Os conjuntos de dados de sobrevivência são caracterizados por variáveis que indicam os tempos até a ocorrência dos eventos de interesse e, geralmente, por variáveis indicadoras de censuras. Além disso, os dados analisados ainda podem incorporar outras variáveis observadas, denominadas covariáveis ou variáveis explicativas, que estão relacionadas com seus respectivos tempos. Contudo, apenas algumas destas covariáveis são conhecidas e medidas, como por exemplo, idade, sexo, dentre outras. Os fatores desconhecidos ou não incluídos na análise, mas que podem influenciar na sobrevivência dos indivíduos, são denominados na literatura como fragilidade. O principal objetivo deste conceito é controlar a heterogeneidade não observada dos indivíduos do estudo. Desse modo, modelos que buscam captar a existência e quantificar essa fragilidade são conhecidos como modelos de fragilidade.

Vaupel, Manton e Stallard (1979) consideram o modelo de fragilidade como um modelo caracterizado pela presença de um efeito aleatório durante o período de estudo. Em outras palavras, há a existência de uma variável aleatória não observada que representa informações não disponíveis, não observadas ou não consideradas durante o período de análise e planejamento do experimento. Este efeito aleatório é representado pela variável de fragilidade, sendo introduzida na modelagem através da função de risco. Assim, a heterogeneidade incorporada ao modelo decorre de covariáveis individuais não observadas ou ainda desconhecidas como fatores de risco.

Na literatura muitos trabalhos descrevem o impacto dos modelos de fragilidade em estudos estatísticos e análises de dados. Para enfatizar este fato foram exploradas bases de dados disponíveis no repositório *online Scopus*, reconhecido como um banco de dados de resumos e citações revisados de forma independente e imparcial. Segundo Ribeiro e Tavares (2017) é possível, por meio deste repositório, mensurar a produção acadêmica de temas e áreas do conhecimento, com foco na contagem de autoria e coautoria, além da análise de publicações, citações, co-citações, dentre outros fatores. Mediante a isso, a Figura 1 apresenta o mapa de rede que ressalta os principais autores de forma qualitativa, com base no número de citações, e de

forma quantitativa, com base no número de publicações. Estes autores desenvolveram estudos sobre modelos de sobrevivência induzidos por fragilidade, mostrando o avanço literário do tema entre os anos de 2018 a 2023.

Figura 1 – Mapa de rede de coautorias com o agrupamentos por co-citação dos principais autores sobre o tema modelos de sobrevivência induzidos por fragilidade.



Fonte: Elaborada pelo autor.

Note que a intensidade do *link* entre os autores indica o número de publicações feitas em conjunto sobre o tema modelos de fragilidade aplicados em dados de sobrevivência. O tamanho dos "nós" refere-se ao volume de publicações nos últimos anos sobre este tema, enquanto os *clusters* indicam os grupos co-autorais. De acordo com a Figura 1, autores como Cancho *et al.* (2018) e Louzada *et al.* (2020) se destacam entre os autores que abordam o uso de modelos de fragilidade, possuindo um maior número de citações em outras pesquisas e uma maior quantidade de publicações referentes ao tema. Desta forma, alguns trabalhos envolvendo modelos de fragilidade são apresentados a seguir, com o intuito de enfatizar a importância do tema e de sua aplicação, além de propor e mostrar os recursos substanciais para o desenvolvimento deste no âmbito acadêmico.

1.1 Revisão de literatura

O estudo de modelos de fragilidade vem ganhando espaço no âmbito acadêmico tendo destaque em aplicações que enfatizam tanto variáveis discretas de fragilidade quanto contínuas. Em particular, modelos que consideram uma fragilidade discreta possuem uma maior flexibilidade, além de evidenciar a existência de uma proporção de indivíduos dos quais não se espera a ocorrência do evento de interesse, bem como a possível observação de indivíduos com fragilidade zero. Estas características podem ser observadas nos trabalhos de Souza *et al.* (2017), que propõem uma extensão da distribuição de Conway-Maxwell e da distribuição Hiper-Poisson para a modelagem da sobrevivência induzida por uma variável discreta de fragilidade e de Molina (2020), que destaca diferentes modelos de fragilidade usando distribuições pertencentes à família de Série de Potência Zero-Modificada, observando indivíduos com fragilidade zero. Contudo, diferentes modelos empregam distribuições de fragilidades como sendo contínuas e não-negativas, como é o caso das distribuições Gama (VAUPEL; MANTON; STALLARD, 1979), Inversa Gaussiana (IG) (HOUGAARD, 1984), Positiva Estável (PE) (HOUGAARD, 1986) e Gama Generalizada (CHEN; ZHANG; ZHANG, 2013). Na literatura muitos são os trabalhos que exploram desta continuidade em seus estudos. Por exemplo, Giussani e Bonetti (2019) usam técnicas multivariadas de sobrevivência para a análise de dados de tempo de falha com censura à direita, em que é investigado uma nova família de modelos paramétricos de fragilidade bivariada e Bunyatisai, Prasitwattanaseree e Ingsrisawang (2017) investigam o desempenho do modelo de Cox com e sem o uso da fragilidade Gama e estabelecem um modelo paramétrico Weibull para esta mesma variável de fragilidade.

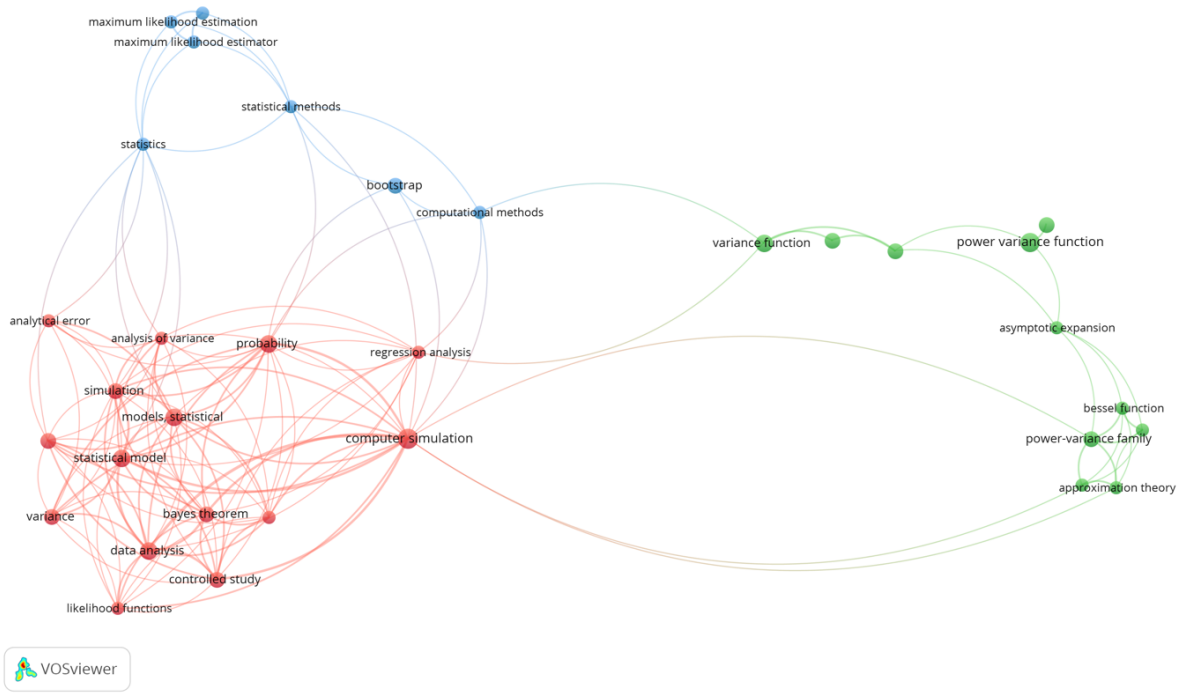
Além desses trabalhos, é comum encontrar estudos que utilizam da fragilidade para analisar conjuntos de dados, especialmente voltados para a área da saúde. Por exemplo, Isidro *et al.* (2020) aplicam modelos de fragilidade Gama e Log-Normal para avaliar o risco e os fatores que contribuem para a morte de pacientes com leucemia, enquanto Gurmu (2018) usam o método de Kaplan-Meier, o modelo de Cox e o modelo paramétrico de fragilidade compartilhada para estudar a sobrevivência de pacientes com câncer cervical. Já Banbeta *et al.* (2015) utilizam o modelo de fragilidade para analisar o tempo de cura de vítimas de desnutrição aguda grave, considerando as distribuições Exponencial, Weibull e Log-Logística para as funções de risco e as distribuições Gama e IG para as distribuições de fragilidade.

O efeito da fragilidade em variáveis contínuas é observado em abordagens semiparamétricas e não-paramétricas, nas quais esses modelos capturam a variabilidade de forma mais eficaz para análise de dados, em comparação com modelos tradicionais de sobrevivência. Estes relatos são presenciados em Brito *et al.* (2020) e Zhou *et al.* (2015). Em Brito *et al.* (2020), é realizada uma comparação entre os modelos de Cox e semiparamétricos, além de uma simulação do modelo de fragilidade Gama com diferentes níveis de censura e heterogeneidade. Já Zhou *et al.* (2015) propõem um modelo bayesiano não-paramétrico dependente, ajustando a distribuição da fragilidade a covariáveis contínuas e categóricas de maneira flexível.

A abordagem bayesiana tem sido amplamente usada em estudos de modelos de fragilidade, destacada no trabalho de Zhou *et al.* (2015) pela sua eficiência. A inferência bayesiana também é aplicada tanto a variáveis de fragilidade discretas quanto contínuas. Exemplos de pesquisas com variável discreta de fragilidade incluem Fortes (2020), Zhou, Hanson e Zhang (2017), Cancho *et al.* (2021), Cancho *et al.* (2018), Macera (2015), entre outros. Por exemplo, Fortes (2020) propõe um modelo de sobrevivência com fragilidade Weibull, baseado na distribuição Birnbaum-Saunders, enquanto Zhou, Hanson e Zhang (2017) utilizam desta inferência para desenvolver uma abordagem generalizada para o modelo de Tempo de Falha Acelerado. Já Wheeler *et al.* (2021), Gasparini *et al.* (2019) e Amorim (2014) são exemplos que autores que usam a inferência bayesiana considerando a variável de fragilidade contínua. Wheeler *et al.* (2021) usam fragilidade para incluir efeitos aleatórios no risco de exposição a alérgenos alimentares, enquanto Amorim (2014) compara diferentes abordagens de verossimilhança em modelos de fragilidade, baseados no algoritmo EM, cadeias de Markov de Monte Carlo e em processos de estimação usando verossimilhança parcial, verossimilhança penalizada, dentre outros, e Gasparini *et al.* (2019) avalia o impacto de uma especificação incorreta da função de risco base em cenários clínicos via simulação de Monte Carlo.

O impacto enfatizado por Gasparini *et al.* (2019) é muito importante nos estudos sobre fragilidade, uma vez que, este recurso é empregado com a finalidade de controlar a heterogeneidade não observada das unidades do estudo, melhorando significativamente os modelos. Um recurso para que haja este controle é o uso da família de distribuições Power Variance Function (PVF), sugerida por Tweedie *et al.* (1984) , nos modelos. Devido a isso, o mapa de rede ilustrado na Figura 2 ressalta a relação desta família de funções com diversos temas estatísticos e probabilísticos em publicações atuais, mostrando-se assim um recurso viável para modelar dados e propor uma variedade crescente de estudos. Ainda, por meio desta figura, é presenciada a existência de algumas palavras-chaves que caracterizam alguns tópicos e/ou modelos que foram analisados em conjunto com a família PVF nos trabalhos publicados até 2023, sendo agrupados de acordo com as suas ocorrências em diferentes áreas de pesquisa. Note que é possível observar três *clusters* principais (vermelho, verde e azul), no qual, o *cluster* verde apresenta trabalhos que relacionam a família PVF com funções de base, a teoria da aproximação e a expansão assintótica; o *cluster* azul exibe a ligação desta família com a maximização, estimação do estimador da máxima verossimilhança e métodos estatísticos e computacionais, com ênfase para o método de *Bootstrap*, e o *cluster* vermelho faz a associação desta família com temas teóricos, como por exemplo verossimilhança, probabilidade, modelos estatísticos, teorema de Bayes e análise de regressão.

Figura 2 – Mapa de rede da associação entre conteúdos estatísticos e probabilísticos com a família PVF.



Fonte: Elaborada pelo autor.

A família PVF é uma classe generalizada de modelos de fragilidade que inclui distribuições conhecidas, como Gama, IG, Log-Normal e PE. Segundo Wienke (2010), essa classe se destaca pelo uso e benefícios das distribuições que a compõem, além da importância durante o processo de estimação. Um exemplo disso é o trabalho de Monaco, Gorfine e Hsu (2018), que apresentam o pacote R denominado *fragiltySurv* para simular e ajustar modelos de fragilidade semiparamétricos, utilizando estimadores consistentes para distribuições PVF. Além disso, outra característica destacada por Wienke (2010) é a possibilidade da representação de algumas distribuições mediante a restrições paramétricas nas distribuições PVF. Isso é exemplificado nos estudos de Rodrigues, Calsavara e Tomazella (2018), Calsavara *et al.* (2017) e Rodrigues *et al.* (2021), que propõem um modelo de taxa de cura com fragilidade usando distribuições PVF. O trabalho de Cancho *et al.* (2021) também destaca a aplicabilidade da distribuição PVF como a variável de fragilidade, permitindo uma estimativa mais precisa da taxa de cura.

Os trabalhos mencionados anteriormente compartilham uma característica comum, chamada "taxa de cura", que retratam estudos que possuem uma fragilidade zero, indicando um subgrupo de indivíduos não suscetíveis, onde o evento de interesse não ocorre. Esses dados possuem a estrutura de longa duração sendo determinada por ter algumas unidades amostrais com fator de risco zero, conforme estudos de Berkson e Gage (1952), Tsodikov, Yakovlev e Asselain (1996), Tsodikov, Ibrahim e Yakovlev (2003) e Rodrigues *et al.* (2009a). Um exemplo de trabalho que destaca a existência de uma fragilidade zero é o estudo de Bedia (2022), que

propõe um modelo de sobrevivência bivariada de longa duração cujo o número de causas de risco para diferentes tipos de eventos é modelado mediante a um modelo de fragilidade da família PVF. Contudo, essa taxa de cura também é analisada por meio de modelos denominados modelos defeituosos. Alguns autores ressaltam a utilização destes modelos com o efeito de fragilidade. Por exemplo, Tesema *et al.* (2022) e Tomazella, Milani e Dias (2018) estudam a mortalidade infantil na África Oriental e investigam a heterogeneidade não observada, respectivamente, usando o modelo de fragilidade compartilhada Gama-Gompertz. Já Lima *et al.* (2021) aplicam o modelo Gama-Gompertz para corrigir a heterogeneidade não observada em cabras anglo-nubianas e Scudilio *et al.* (2019) propõem um modelo defeituoso com termo de fragilidade para modelar a taxa de cura, usando as distribuições Gama-Gompertz Defeituosa e Gama-IG Defeituosa.

Outro recurso ressaltado na literatura é o Modelo Exponencial por Partes (MEP) devido a sua grande flexibilidade em diversos estudos estatísticos e probabilísticos, possuindo a capacidade de acomodar funções taxa de falha com diversas formas (IBRAHIM; CHEN; SINHA, 2001). Outra vantagem é a possibilidade de se trabalhar com este modelo tanto na versão paramétrica quanto na versão não-paramétrica, uma vez que, é caracterizado pela aproximação da função taxa de falha por segmentos de retas cujos comprimentos são determinados através de partições do eixo do tempo, resultando em intervalos em que a função taxa de falha é considerada constante. Desta forma, a versão não-paramétrica do MEP é obtida ao tomar uma partição do eixo do tempo com tantos intervalos quanto for o número de falhas, enquanto a versão paramétrica do MEP é determinada ao considerar um número de intervalos inferior ao número de falhas, permitindo que haja mais de uma falha por intervalo (DEMARQUI, 2006). Os trabalhos de Sibim (2011), Mello *et al.* (2016), Demarqui (2006) e Demarqui (2010) destacam a importância do modelo MEP na modelagem de sobrevivência segundo a abordagem bayesiana. Sibim (2011) desenvolve procedimentos para modelos com e sem taxa de cura, baseados no modelo MEP. Mello *et al.* (2016) propõem uma abordagem obtendo uma distribuição suavizada para os parâmetros, usando o modelo MEP para representar a função de risco. Demarqui (2006) apresenta um estudo do MEP focando na partição intervalar do eixo dos tempos. Por sua vez, Demarqui (2010) reúne artigos que exploram o modelo MEP em diversos contextos, incluindo um com taxa de cura. Esses estudos são aplicáveis a modelagem de dados de sobrevivência em diferentes áreas do conhecimento.

Apesar das inovações acadêmicas, ainda existem desafios práticos de inferência, especialmente devido a complexidade dos modelos como ocorre na abordagem bayesiana, onde estimar parâmetros a partir da distribuição a posteriori pode ser difícil. Como resultado, métodos computacionais eficientes, como os Monte Carlo via Cadeias de Markov (MCMC), são essenciais para melhorar o desempenho dos modelos. O principal objetivo do MCMC é gerar uma amostra da distribuição a posteriori e determinar as estimativas amostrais dos parâmetros desta distribuição de forma similar ao que ocorre no método de Monte Carlo original. Ao contrário do método de Monte Carlo, as técnicas de simulação utilizam cadeias de Markov, em que o processamento de dados destas cadeias são caracterizados por diversos métodos de simulação,

como os métodos Hamiltoniano de Monte Carlo (HMC), Variância-Zero e o Metropolis-Hastings (MH) (PAIXÃO, 2021). Em particular, segundo Spiegelhalter *et al.* (1996), o HMC combina os métodos Gibbs e MH e evita o comportamento de passeio aleatório, reduzindo a autodependência das cadeias. Torres *et al.* (2018) também destaca que diferente dos métodos MCMC de caminhada aleatória, o HMC considera o gradiente da distribuição de probabilidade a posteriori, o que é uma vantagem. Estudos como os de Xavier (2019), Chatzilena *et al.* (2019), Paixão (2021) e Hartmann (2015) enfatizam a importância do HMC, aplicando-o em diversos estudos como na estimação de parâmetros em modelos GARCH univariado e multivariado, na análise de dados de doenças infecciosas dinâmicas, na modelagem de dados GJR-GARCH e na abordagem bayesiana não-paramétrica para dados de comportamento extremo, no qual, o método HMC é aplicado como uma solução alternativa para lidar com a intratabilidade analítica de distribuições a posteriori com altas dimensões.

Outro recurso promissor em estudos de simulação é a utilização de *softwares* com a linguagem em $C++$, como a linguagem de programação Stan. Esta linguagem pode ser executada via linha de comando, sendo compatível com programas como R, Python, Matlab, Stata e Julia, e funciona em plataformas Linux, Mac e Windows, conforme apontado por Jiang e Carter (2019), Luo e Jiao (2018), Xavier (2019) e Gelman, Lee e Guo (2015). Autores como Jiang e Carter (2019) e Luo e Jiao (2018) destacam as vantagens do Stan, enfatizando que esta é uma linguagem fácil, de sintaxe direta, flexibilidade na modelagem e rapidez nas estimativas, devido ao uso de funções otimizadas em $C++$, além de executar o método HMC oferecendo múltiplas opções de estimativa (JIANG; CARTER, 2019). Ainda, Luo e Jiao (2018) também mencionam a eficiência do Stan em comparação com programas bayesianos tradicionais, como o BUGS, pois o Stan exige menos iterações para boa mixagem (1000 contra 100.000 no BUGS), além da possibilidade de permitir o uso de diversos tipos de distribuições. Como consequência, os estudos de Luo e Jiao (2018), Chatzilena *et al.* (2019), Ng'ombe e Lambert (2021) e Jiang e Carter (2019) evidenciam a eficácia do Stan em suas pesquisas, aplicando-o para modelos da Teoria de Resposta ao Item (TRI), epidemiológicos, estocásticos e log-lineares. Portanto, estes trabalhos validam o uso do método HMC via Stan e R-Stan mostrando como estes recursos têm sido eficazes, benéficos e proeminentes nos estudos atuais.

1.2 Objetivos

O presente trabalho tem como objetivo desenvolver modelos de fragilidade para modelar a heterogeneidade não observada em dados de sobrevivência. Os objetivos específicos são:

- Realizar uma revisão teórica sobre modelos univariados e multivariados de fragilidade, associando-os ao modelo de Cox, com foco em modelos da família PVF. Também será abordado um estudo sobre modelos MEP, defeituosos e de longa duração, além dos conceitos de inferência bayesiana e métodos computacionais de estimação.

- Apresentar modelos univariados para a modelagem da heterogeneidade não observada, considerando uma composição entre uma distribuição de fragilidade da família PVF e o modelo MEP, gerando modelos defeituosos e de longa duração induzidos por fragilidade.
- Apresentar modelos bivariados para a modelagem da heterogeneidade não observada, considerando uma composição entre uma distribuição de fragilidade da família PVF e um modelo com suporte em $\mathbb{R}$, gerando modelos bivariados de longa duração induzidos por fragilidade.
- Estudar as principais propriedades dos modelos apresentados, destacando funções importantes na Análise de Sobrevivência.
- Abordar procedimentos inferenciais bayesianos através do método HMC implementado no R-Stan, avaliando o desempenho de propriedades dos estimadores de bayes mediante a estudos de simulações.
- Verificar a aplicabilidade dos modelos apresentados em conjuntos de dados com estatísticas reais.

1.3 Organização do trabalho

Os capítulos deste trabalho são organizados da seguinte forma:

- No Capítulo 2 é realizada uma revisão teórica sobre os temas principais que subsidiaram este trabalho, incluindo modelos de fragilidade, PVF e MEP, além de conceitos de inferência bayesiana destacando os principais métodos computacionais de estimação. Esta revisão fundamenta o desenvolvimento da tese.
- No Capítulo 3 é apresentado um modelo estatístico de fragilidade para modelar a heterogeneidade não observada em dados de sobrevivência, resultando em modelos defeituosos. A abordagem inferencial é baseada em métodos bayesianos mediante ao uso do método HMC implementado no R-Stan.
- O Capítulo 4 decorre de forma similar ao capítulo 3. Contudo, foca na determinação de modelos de longa duração univariados.
- No Capítulo 5 é introduzido um modelo de sobrevivência bivariado que incorpora fragilidade. A abordagem inferencial usada é baseada em métodos bayesianos através do método HMC implementado no R-Stan.
- No Capítulo 6 é exibido um resumo dos principais resultados da tese e propostas para futuras pesquisas relacionadas aos modelos apresentados.

CAPÍTULO

2

CONTEÚDOS PRELIMINARES

Neste capítulo, é apresentada uma revisão teórica sobre os principais temas que fundamentam este trabalho, destacando modelos de fragilidade e a distribuição de fragilidade da família PVF . Os resultados são discutidos no contexto univariado e multivariado, sendo aplicáveis nos modelos descritos nos estudos posteriores, fundamentados por Aalen (1992), Hougaard (2000), Wienke (2010), Balan e Putter (2020), Duchateau e Janssen (2008) e Bedia (2022). Além disso, o modelo MEP é abordado segundo Ibrahim, Chen e Sinha (2001), Sibim (2011) e Demarqui (2010), sendo usado como função de risco base nos próximos capítulos. Por fim, é introduzido os conceitos de inferência bayesiana, com foco nas técnicas computacionais de estimação e nos critérios de comparação e seleção de modelos, conforme os pressupostos de Xavier (2019), Paixão (2021), Hartmann (2015), Pardo (2018), Sibim (2011), Ibrahim, Chen e Sinha (2001), Meng e Wong (1996), Meng e Schilling (2002), Gronau *et al.* (2019), Gronau *et al.* (2017) e Gronau, Singmann e Wagenmakers (2017).

2.1 Heterogeneidade no modelo de Cox

Os estudos propostos por Balan e Putter (2020) destacam a função de risco como o conceito central de Análise de Sobrevivência, uma vez que, esta função descreve a ocorrência do risco instantâneo para o evento de interesse em um indivíduo, considerando que este indivíduo ainda não tenha vivenciado este evento. Contudo, como indivíduos distintos geram probabilidades de sobrevivência e de riscos diferentes é necessário que estas diferenças sejam contabilizadas nos modelos. Na literatura, o modelo mais famoso que realiza este tipo de análise é o modelo de riscos proporcionais de Cox, na qual, a suposição de riscos proporcionais indica que a razão dos riscos entre quaisquer dois indivíduos é constante ao longo do tempo, sendo representado por meio à um “risco de base” não-paramétrico.

O modelo de Cox especifica que o risco em relação ao tempo até a ocorrência do evento

de interesse é dado por

$$h(t \mid \mathbf{X}) = h_0(t)\ \exp\{\beta^T \mathbf{X}\}; \quad t \geq 0, \tag{2.1}$$

em que $h_0(\cdot)$ é a função de risco base comum para todos os indivíduos, $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ é o vetor de covariàveis e $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ é o vetor dos parâmetros de regressão associados a $\mathbf{X}$.

Observe que ao assumir o vetor $\mathbf{X}$ com tempo constante, o risco definido é composto por indivíduos que ainda não vivenciaram o evento de interesse ou que ainda não foram removidos do estudo por outros motivos, ou seja, por censura. Ainda, a definição (2.1) ressalta que quando dois indivíduos apresentam vetores de covariàveis diferentes como $\mathbf{X}^*$ e $\tilde{\mathbf{X}}$, por exemplo, suas funções de risco serão iguais apenas se $\beta^T \mathbf{X}^* = \beta^T \tilde{\mathbf{X}}$. Assim, a suposição de riscos proporcionais do modelo de Cox (2.1) indica que a razão entre os riscos tal que $h(t|\mathbf{X}^*)$ é dividido por $h(t|\tilde{\mathbf{X}})$, é igual a $\exp(\beta^T(\mathbf{X}^* - \tilde{\mathbf{X}}))$, independente do tempo. Entretanto, quando esta suposição é violada, o efeito das covariàveis passam a ser dependentes do tempo e o modelo é reescrito, para $\beta(t)$ não constante, como

$$h(t \mid \mathbf{X}) = h_0(t)\ \exp\{\beta(t)\mathbf{X}\}.$$

De maneira geral, suponha que o modelo (2.1) seja válido para um vetor de covariàveis $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ cujo vetor de coeficientes de regressão associado é dado por $\beta = (\beta_1, \beta_2)$. Consequentemente, esse modelo é reescrito como

$$h(t \mid \mathbf{X}) = h_0(t)\ \exp\{\beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2\}. \tag{2.2}$$

Na prática, raramente é possível contabilizar todas as covariàveis de forma relevante. Neste caso, é considerado que as covariàveis omitidas induzem a heterogeneidade não observada e que as diferenças entre os indivíduos são explicadas somente pelas covariàveis que foram examinadas no modelo. Como consequência, é definida uma variàvel aleatória que atua multiplicativamente sobre o risco e que caracteriza as variàveis não observadas. Por exemplo, ao supor que o modelo de Cox (2.2) inclua somente a variàvel $\mathbf{X}_1$, pode-se definir a variàvel aleatória $Z = \exp\{\beta_2 \mathbf{X}_2\}$ que estará associada a variàvel não observada $\mathbf{X}_2$. Assim, o modelo (2.2) é reescrito como $h(t \mid \mathbf{X}) = Z h_0(t)\ \exp(\beta_1 \mathbf{X}_1)$, em que Z é o termo de "fragilidade" do modelo.

2.2 Modelos de fragilidades

2.2.1 Modelo de fragilidade univariado

A análise dos dados de sobrevivência é geralmente baseada na suposição de que a população em estudo é homogênea, tal que ao condicionar as covariàveis, todo indivíduo possui o mesmo risco de experimentar o evento de interesse, além de assumir tempos independentes para o estudo. Contudo, estas hipóteses não podem ser generalizadas, uma vez que, muitas aplicações exigem o uso de uma amostra heterogênea, nos quais, os indivíduos apresentam

diferentes riscos e funções de risco. Deste modo, embora os indivíduos possam parecer idênticos em alguns aspectos, eles ainda podem diferir em algumas características. Consequentemente, Wienke (2010), Vaupel, Manton e Stallard (1979) e Balan e Putter (2020) sugerem modelos de efeitos aleatórios para explicar a heterogeneidade não observada, sendo denominados modelos de fragilidade.

O modelo de fragilidade clássico e aplicado com mais frequência na literatura assume a estrutura de riscos proporcionais e está condicionado a um efeito aleatório Z, denotado como variável de fragilidade, sendo não observado e independente do tempo. Neste contexto e para este trabalho é adotada a estrutura de riscos multiplicativos, em que Z atua multiplicativamente na função de risco de base, ou seja

$$h\,(t \mid Z) = Z\,h_0(t); \quad t > 0, \tag{2.3}$$

em que Z é a variável aleatória de fragilidade não-negativa e $h_0(\cdot)$ é a função de risco base comum para todos os indivíduos.

A identificabilidade do modelo (2.3) é enfatizada por Wienke (2010), na qual, as distribuições de fragilidade são padronizadas para $\mathbb{E}[Z] = 1$, se o valor esperado da distribuição de fragilidade existir, e a variância $Var[Z] = \theta > 0$ (se existir) é interpretada como uma medida de heterogeneidade entre a população. Assim, quando θ é pequeno, os valores de Z estão localizados perto de um. Caso contrário, estes valores estão mais dispersos, induzindo uma maior heterogeneidade nos riscos individuais (WIENKE, 2010). Além disso, note que a adesão de covariàveis pode ser introduzida ao modelo (2.3) por

$$h\,(t \mid Z, \mathbf{X}) = Z\,h_0(t) \exp\{\beta^T \mathbf{X}\}, \tag{2.4}$$

com $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ e $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ correspondendo aos vetores de covariàveis $\mathbf{X}$ e seus respectivos parâmetros de regressão. Observe que, ao considerar a presença de covariàveis, o modelo (2.4) será composto pelo produto de duas componentes, uma não-paramétrica e outra paramétrica. Assim, ao interpretá-lo é notado que indivíduos que possuem uma maior fragilidade podem ser considerados mais frágeis, consequentemente, podendo experimentar o evento de interesse mais cedo do que os demais indivíduos do estudo (BALAN; PUTTER, 2020).

Desta forma, com intuito de simplicidade, neste capítulo será usado o modelo (2.3) na formalização das definições posteriores. Logo, ao tomar a função de risco condicional (2.3) e dada a variàvel de fragilidade Z, a probabilidade que representará a proporção dos indivíduos sobreviventes após o tempo $t > 0$, condicionada a variàvel Z, é dada por

$$S(t \mid Z) = \exp\left\{-\int_0^t h\,(u \mid Z)\,du\right\} = \exp\left\{-Z\int_0^t h\,(u)\,du\right\} = \exp\{-Z\,H_0\,(t)\}, \tag{2.5}$$

em que $H_0(\cdot)$ é a função de risco acumulado base definida por: $H_0(t) = \int_0^t h_0(u)\,du$. Contudo, como nem sempre os dados para o modelo condicional são observáveis, é necessário considerar

o modelo marginal (ou populacional) determinado ao calcular a esperança do modelo (2.5) com respeito a variàvel de fragilidade Z, ou seja

$$S(t) = \int_0^\infty S(t \mid Z)\, g(z)dz = \int_0^\infty \exp\{-Z\, H_0(t)\}\, g(z)dz = \mathbb{E}\left[\exp\{-Z\, H_0(t)\}\right], \tag{2.6}$$

em que $g(z)$ é a função densidade de probabilidade (f.d.p) da variàvel Z. Observe que o modelo marginal (2.6) corresponde a transformada de Laplace de Z, ao avaliar $s = H_0(\cdot)$, tal que $\mathscr{L}_Z(s) = \mathbb{E}[\exp\{-Zs\}]$. Como consequência, o modelo (2.6) pode ser reescrito como

$$S(t) = \mathscr{L}_Z(H_0(t)), \quad t > 0. \tag{2.7}$$

Na literatura, autores como Aalen (1992), Wienke (2010) e Balan e Putter (2020) enfatizam a relação entre a função de sobrevivência marginal e a transformada de Laplace, sendo uma ferramenta útil para descrever as relações entre as função de risco e de sobrevivência, além da função densidade (f.d). Ainda, Aalen (1992) retrata esta utilidade mediante ao emprego das derivadas da transformada de Laplace, visto que proporcionam resultados gerais sobre a distribuição de sobrevivência. Consequentemente, a r-ésima derivada da transformada de Laplace relacionada ao modelo de sobrevivência marginal (2.7) é dada por

$$\mathscr{L}_Z^{(r)}(s) = (-1)^r\, \mathbb{E}[Z^r \exp\{-Zs\}]. \tag{2.8}$$

Já o r-ésimo momento da fragilidade é dado por

$$\mathbb{E}[Z^r] = (-1)^r\, \mathscr{L}_Z^{(r)}(0). \tag{2.9}$$

Wienke (2010) também define relações que destacam a associação entre a transformada de Laplace e a função de sobrevivência, definindo a f.d do tempo de falha como

$$f\,(t) = -\frac{dS(t)}{dt} = -h_0(t)\, \mathscr{L}_Z'\,(H_0\,(t)), \tag{2.10}$$

em que $\mathscr{L}_Z'$ denota a primeira derivada da transformada de Laplace e $t > 0$. Já a função de risco é determinada pela razão entre a f.d e a função de sobrevivência, sendo dada por

$$h\,(t) = -h_0(t)\, \frac{\mathscr{L}_Z'\,(H_0\,(t))}{\mathscr{L}_Z\,(H_0\,(t))}. \tag{2.11}$$

Somado a isso, mediante ao r-ésimo momento (2.9), a esperança e a variância podem ser caracterizadas, respectivamente, por

$$\mathbb{E}\,[Z] = -\mathscr{L}_Z'\,(0) \tag{2.12}$$

e

$$Var\,[Z] = \mathscr{L}_Z''\,(0) - [-\mathscr{L}_Z'\,(0)]^2, \tag{2.13}$$

com $\mathscr{L}_Z''$ denotando a segunda derivada da transformada de Laplace. Balan e Putter (2020) também ressaltam como características complementares a estes estudos que:

- $\mathscr{L}_Z(0) = 1$.
- O coeficiente de variação da variàvel de fragilidade Z é: $CV^2(Z) = \dfrac{\mathscr{L}_Z''(0)}{(\mathscr{L}_Z'(0))^2} - 1$.

Portanto, é presenciado que são muitas as pesquisas que descrevem, analisam e propõem estudos sobre modelos de fragilidade na literatura. Contudo, destaca-se que uma distribuição de fragilidade que descreve a fragilidade da população no início de um estudo ou análise, geralmente, supõe que essa fragilidade é fixa para cada indivíduo ao longo do tempo. No entanto, é evidenciado que a composição da população muda com o passar do tempo. Consequentemente, em média, os indivíduos mais frágeis morrem mais cedo. Devido a este fato, é notado que a distribuição da fragilidade na população de risco também muda ao longo do tempo. Este fato é destacado por Vaupel, Manton e Stallard (1979) ao estabelecer uma ligação entre o modelo condicional e o marginal, conforme o teorema a seguir.

Teorema 1. (VAUPEL; MANTON; STALLARD, 1979) Assuma o modelo de fragilidade dado por (2.3). A função de risco populacional, $h(t) = \dfrac{f(t)}{S(t)}$ é geralmente dada por: $h(t) = \mathbb{E}[h(t|Z) \mid T > t]$. Ou mais especificamente:

$$h(t) = \int_0^\infty h(t \mid z)\, f(z \mid T > t) dz \;=\; h_0(t) \int_0^\infty z\, f(z \mid T > t) dz,$$

com $f(z \mid T > t)$ representando a f.d de fragilidade dos sobreviventes por um período de tempo t.

A prova deste teorema não será apresentada neste trabalho, mas pode ser encontrada em Wienke (2010). Além disso, observe que o risco marginal pode ser visto como a média ponderada dos riscos individuais de sujeitos vivos em um determinado tempo, onde essa ponderação dependerá da distribuição da variàvel de fragilidade Z. Contudo, é evidenciado que indivíduos frágeis e com altos valores de Z tendem a morrer primeiro. Ainda, a f.d de fragilidade para os sobreviventes até o tempo t, usada nesta demonstração e considerando as definições (2.5) e (2.7), é dada por

$$f(z \mid T > t) = \frac{S(t \mid z)\, f_Z(z)}{S(t)} = \frac{\exp\{-zH_0(t)\}\, f_Z(z)}{\mathscr{L}_Z(H_0(t))}.$$

2.2.2 Modelo de fragilidade discreto

Bedia (2022) ressalta que geralmente distribuições de fragilidade contínuas não permitem a suposição de risco zero em suas análises, fato este que indica a existência de um subgrupo de indivíduos não suscetíveis ao evento de interesse mesmo após o período de observação. Dentre as abordagens utilizadas na literatura empregadas para o estudo de amostras que possuem esta característica, Wienke (2010) aborda o uso do modelo de fragilidade Poisson Composto como um recurso útil para a modelagem de dados de sobrevivência que contém uma proporção de indivíduos para os quais o evento de interesse não ocorre. Outro recurso também apontado por

este autor são os modelos de fragilidade discreta cuja massa de probabilidade é definida nos inteiros não-negativos, ou seja, em $\mathbb{Z}_+ = \{0,1,2,\ldots\}$.

Deste modo, para os modelos de fragilidade discreta, suponha que a distribuição de probabilidade de Z é dada por $\mathbb{P}[Z=z] = p_z$, com $\sum p_z = 1$ e $z = 0,1,2,\ldots$. Observe que a função de sobrevivência marginal deste modelo é obtida ao considerar um somatório aplicado na Equação (2.5) sobre o suporte da distribuição de Z (BEDIA, 2022). Ou ainda

$$S(t) = \sum_{z=0}^{\infty} S(t \mid Z) p_z = \sum_{z=0}^{\infty} (S_0(t))^Z p_z = \mathbb{E}[(S_0(t))^Z] = \psi_Z(S_0(t)), \tag{2.14}$$

em que $\psi_Z(\cdot)$ é a função geradora de probabilidade (f.g.p) da variàvel de fragilidade Z e $S_0(t) = \exp\{-H_0(t)\}$ é a função de sobrevivência de base. Esta autora ainda realça que desde que $\mathbb{P}[Z=0] > 0$, a função de sobrevivência (2.14) é imprópria, ou seja

$$\lim_{t \longrightarrow \infty} S(t) = \mathbb{P}[Z=0] = \psi_Z(0) = p_0 > 0, \tag{2.15}$$

implicando que existe uma probabilidade $p_0 > 0$ de que algum indivíduo seja não suscetível ao evento de interesse mesmo após um período de tempo.

2.2.3 Modelo de fragilidade multivariado

Algumas vezes os dados de sobrevivência são observados em grupos de indivíduos, como por exemplo, indivíduos de uma mesma família, de pacientes tratados em um mesmo hospital, dentre outros. O fato é que nestas situações é esperado que os tempos observados em cada grupo apresentem semelhanças entre si e que não são observadas em outros grupos, além da possibilidade de um único indivíduo repetir o mesmo evento de interesse várias vezes, sendo que este número de repetições pode ser aleatória durante o período de acompanhamento (eventos recorrentes) ou previamente prefixada (medidas repetidas), ou de um mesmo indivíduo ter mais de um evento de interesse durante o estudo. Consequentemente, ao abordar estas situações, é coerente supor a existência de uma associação entre os tempos de sobrevivência. Esta associação é uma característica marcante em dados de sobrevivência multivariados (BEDIA, 2022).

Uma abordagem popular utilizada para estes dados é a aplicação de modelos de fragilidade em que é apontado a inclusão de um ou mais efeitos aleatórios na função de risco base, destacando a dependência entre as observações. Para isso, alguns conceitos são fundamentais. O primeiro conceito refere-se a fragilidade compartilhada, na qual, a função risco de base para cada indivíduo no mesmo grupo é a mesma que no modelo padrão de fragilidade univariada, ou seja

$$h(t \mid Z_k) = Z_k h_0(t); \quad t \geq 0,$$

com Z_k representando a fragilidade comum para todos os indivíduos no grupo k e $h_0(\cdot)$ é a função de risco de base comum para todos os indivíduos deste grupo. Este modelo assume que todos os tempos de eventos no mesmo grupo k são independentes dadas as variàveis de

fragilidade Z_k, introduzindo correlação entre os tempos de eventos dos indivíduos do mesmo grupo. Ainda, Wienke (2010) destaca que se $Var(Z_k = 0)$, então há independência entre os tempos de eventos no grupo k. Caso contrário, há dependência positiva deles. Já Duchateau e Janssen (2008) assumem a existência de independência entre as observações de diferentes grupos nestes modelos.

Deste modo, suponha que $q = 1, 2, \ldots, n_k$ tal que n_k representa o número de indivíduos no grupo k. A função de sobrevivência condicional conjunta, dado Z_k, para o grupo k é dada por

$$S(t_1, \ldots, t_{n_k} \mid Z_k) = \exp\left\{-Z_k \sum_{q=1}^{n_k} H_0(t_q)\right\}; \quad t_q \geq 0, \tag{2.16}$$

cujo valor da fragilidade Z_k será constante ao longo do tempo. Já a função de sobrevivência conjunta marginal é obtida a partir da integração da função (2.16) em relação a fragilidade, ou seja

$$S(t_1, \ldots, t_{n_k}) = \int_0^\infty \exp\left\{-Z_k \sum_{q=1}^{n_k} H_0(t_q)\right\} g(z_k) dz_k = \mathscr{L}_{Z_k}\left(\sum_{q=1}^{n_k} H_0(t_q)\right), \tag{2.17}$$

em que $g(z_k)$ é a f.d.p e $\mathscr{L}_{Z_k}(\cdot)$ a transformada de Laplace, ambas associadas a fragilidade Z_k do grupo k. Desta forma, ao aplicar a propriedade (2.8) na função de sobrevivência conjunta (2.17), é determinada a f.d conjunta para um grupo k de tamanho n_k dada por

$$f(t_1, \ldots, t_{n_K}) = (-1)^{n_k} \frac{\partial^{n_k}}{\partial_{t_1} \ldots \partial_{t_{n_k}}} S(t_1, \ldots, t_{n_K}) = (-1)^{n_k} \prod_{q=1}^{n_K} h_0(t_q) \mathscr{L}_{Z_k}^{(n_k)}\left(\sum_{q=1}^{n_k} H_0(t_q)\right).$$

Agora, ao supor que os tempos de sobrevivência estão sujeitos a censura não informativa à direita, a função de verossimilhança completa para o k-ésimo grupo de tamanho n_k, é dada por

$$\left(\prod_{q=1}^{n_K} [h_0(t_{kq})]^{\delta_{kq}}\right) Z_k^{dk} \exp\left\{Z_k \sum_{q=1}^{n_k} H_0(t_q)\right\}, \tag{2.18}$$

em que δ_{kq} é o indicador de censura do q-ésimo indivíduo no k-ésimo grupo e $dk = \sum_{q=1}^{n_k} \delta_{kq}$. Já a função de verossimilhança marginal para o k-ésimo grupo é determinada ao integrar a verossimilhança (2.18) em relação a fragilidade, ou seja

$$\left(\prod_{q=1}^{n_K} [h_0(t_{kq})]^{\delta_{kq}}\right) (-1)^{dk} \mathscr{L}_{Z_k}^{(dk)}\left(\sum_{q=1}^{n_k} H_0(t_q)\right).$$

Ainda, Wienke (2010) enfatiza que, similar aos modelos univariados, a função de sobrevivência conjunta marginal (2.17) pode ser definida para o q-ésimo indivíduo do grupo k em termos da transformada de Laplace como

$$S(t_{q_k}) = \mathscr{L}_{Z_k}(H_0(t_{qk})).$$

Observe que a função de sobrevivência conjunta (2.17) também pode ser estabelecida em termos das funções de sobrevivência marginais dos indivíduos no grupo k de tamanho n_k. Assim, ao aplicar o modelo (2.7) segue que

$$H_0(t_{qk}) = \mathscr{L}_{Z_k}^{-1}\left(S_q(t_{qk})\right); \quad q = 1,2,\ldots,n_k.$$

Consequentemente, a função de sobrevivência conjunta marginal (2.17) é reescrita como

$$S(t_1,\ldots,t_{n_k}) = \mathscr{L}_{Z_k}(\mathscr{L}_{Z_k}^{-1}\left(S_1(t_1)\right) + \mathscr{L}_{Z_k}^{-1}\left(S_2(t_2)\right) + \ldots + \mathscr{L}_{Z_k}^{-1}\left(S_{n_k}(t_{n_k})\right)). \tag{2.19}$$

A expressão (2.19) vincula a função de sobrevivência conjunta as funções de sobrevivência marginais, sendo conhecida como a representação de cópula da função de sobrevivência conjunta com base no modelo de fragilidade. Ainda, observe que as funções de sobrevivência marginais nesta representação são derivadas do modelo de fragilidade e, portanto, também será uma função de distribuição de fragilidade (BEDIA, 2022).

2.2.3.1 *Medida de dependência local*

Clayton (1978) introduz uma medida de associação local entre os tempos T_1 e T_2 que permite a investigação da mudança da dependência com o decorrer do tempo. Esta medida é definida por

$$v^*(t_1,t_2) = \frac{h(t_2 \mid T_1 = t_1)}{h(t_2 \mid T_1 > t_1)} = \frac{S(t_1,t_2)\frac{\partial^2}{\partial_{t_1}\partial_{t_2}}S(t_1,t_2)}{\left(\frac{\partial}{\partial_{t_1}}S(t_1,t_2)\right)\left(\frac{\partial}{\partial_{t_2}}S(t_1,t_2)\right)}, \tag{2.20}$$

em que se $v^*(t_1,t_2) = 1$, então há independência entre os tempos T_1 e T_2. Caso contrário, se $v^*(t_1,t_2) > 1$, então haverá a associação positiva entre os tempos T_1 e T_2.

Um estudo mais detalhado sobre os tópicos desta seção é encontrado nos trabalhos de Bedia (2022), Clayton (1978), Aalen (1992), Wienke (2010) e Duchateau e Janssen (2008).

2.3 Principais modelos de fragilidade

O estudo de modelos de fragilidade vêm ganhando espaço no âmbito acadêmico. Em particular, autores como Cancho *et al.* (2021), Cancho *et al.* (2018), Macera (2015), Zavaleta (2016) e Fortes (2020) ressaltam o uso destes modelos em casos contínuos, não-negativos e multiplicativos, expondo os mais diferentes estudos probabilísticos e estatísticos, onde usualmente são empregadas distribuições já conhecidas na literatura como a Gama, IG, Log-Normal e PE. Por sua vez, propondo uma maior difusão, Wienke (2010) destaca uma classe generalizada de modelos de fragilidade que inclui todas estas distribuições, sendo denominada família PVF. Deste modo, nesta seção haverá a caracterização desta família e das principais distribuições que geralmente são utilizadas em modelos de fragilidade.

2.3.1 Distribuição de fragilidade Gama

O modelo de fragilidade Gama é destacado na literatura por ser usado como uma distribuição de mistura para dados de falha, além da facilidade em derivar sua função de sobrevivência marginal, devido a simplicidade da sua transformada de Laplace. Outra vantagem é que para uma grande classe de modelos de fragilidade univariados, as distribuições de fragilidade entre os sobreviventes convergem para uma distribuição Gama, tornando-a uma ótima sugestão para realizar modelagens (WIENKE, 2010).

Desta maneira, a f.d.p de uma variável aleatória Z com distribuição Gama é dada por

$$g(z) = \frac{1}{\Gamma(\mu)}\sigma^{\mu}\ z^{\mu-1}\exp\{-\sigma z\},$$

com $z > 0$, $\sigma > 0$ e $\mu > 0$. Já transformada de Laplace é definida por

$$\mathscr{L}_Z(s) = \left(1+\frac{s}{\sigma}\right)^{-\mu}. \tag{2.21}$$

Observe que ao utilizar os resultados (2.12) e (2.13), considerando a transformada de Laplace (2.21), é possível estabelecer que $\mathbb{E}[Z] = \frac{\mu}{\sigma}$ e $Var[Z] = \frac{\mu}{\sigma^2}$. Além disso, a identificabilidade deste modelo é confirmada ao assumir a restrição de que $\mu = \sigma$, resultando que $\mathbb{E}[Z] = 1$ e $\theta := Var[Z] = \frac{1}{\sigma}$. Deste modo, ao tomar $Z \sim \Gamma\left(\frac{1}{\theta}, \frac{1}{\theta}\right)$, a transformada de Laplace descrita em (2.21) é reescrita como

$$\mathscr{L}_Z(s) = (1+\theta s)^{-\frac{1}{\theta}}. \tag{2.22}$$

2.3.2 Distribuição de fragilidade Inversa Gaussiana (IG)

A distribuição IG foi introduzida por Hougaard (1984) como uma alternativa a distribuição Gama. Esse destaque também é enfatizado por Wienke (2010) que ressalta que assim como no modelo de fragilidade Gama, existem expressões simples e de forma fechada para as funções marginais de sobrevivência e de risco desta distribuição, o que torna o modelo IG tão atraente. Assim, descreve a f.d.p de uma variável aleatória de distribuição IG com parâmetros $\mu > 0$ e $\sigma > 0$ dada por

$$g(z) = \sqrt{\frac{\sigma}{2\pi z^3}}\exp\left\{-\frac{\sigma}{2\mu^2 z}(z-\mu)^2\right\}; \quad z > 0.$$

Já a transformada de Laplace é definida como

$$\mathscr{L}_Z(s) = \exp\left\{\frac{\sigma}{\mu}\left(1-\sqrt{1+\frac{2\,\mu^2\,s}{\sigma}}\right)\right\}. \tag{2.23}$$

Note que ao utilizar os resultados (2.12) e (2.13), considerando a transformada de Laplace (2.23), é possível estabelecer que $\mathbb{E}[Z] = \mu$ e $Var[Z] = \frac{\mu^3}{\sigma}$. Ainda, a identificabilidade

deste modelo é confirmada ao assumir a restrição de que $\mu = 1$, resultando que $\mathbb{E}[Z] = 1$ e $\theta := Var[Z] = \frac{1}{\sigma}$. Desta maneira, a transformada de Laplace é reescrita como

$$\mathscr{L}_Z(s) = \exp\left\{\frac{1}{\theta}\left(1 - \sqrt{1+2\theta s}\right)\right\}. \tag{2.24}$$

2.3.3 Distribuição de fragilidade Positiva Estável (PE)

Bedia (2022) destaca que uma distribuição é PE se a soma normalizada de n variàveis aleatórias independentes e identicamente distribuídas tiver a mesma distribuição que um fator de escala multiplicado por uma única variàvel aleatória. Ou seja, uma distribuição PE tem a propriedade que afirma que para as variàveis aleatórias independentes e identicamente distribuídas $Z_1, \ldots, Z_n$ e, para cada n, existe uma constante normalizadora $c(n)$ de modo que

$$c(n)Z_1 = Z_1 + \ldots + Z_n,$$

tal que o termo $c(n)Z_1$ possuirá a mesma distribuição que o termo $(Z_1 + \ldots + Z_n)$. Além disso, esta autora ainda realça que a constante $c(n)$ possui a forma $n^{\frac{1}{\gamma}}$ com $\gamma \in (0,2]$. Em particular, se $\gamma = 2$ é obtida a distribuição Normal. Entretanto, para garantir uma distribuição com números positivos é tomado que $\gamma \in (0,1)$ (BEDIA, 2022).

Logo, sob essa parametrização, a f.d.p de uma variàvel aleatória uniparamétrica de uma distribuição PE com $z > 0$ e $0 < \gamma \leq 1$ é dada por

$$g(z) = \frac{1}{\pi}\sum_{k=1}^{\infty}(-1)^{k+1}\frac{\Gamma(k\gamma+1)}{k!}\, z^{-k\gamma-1}\sin(k\gamma\pi). \tag{2.25}$$

Note que a expressão (2.25) é uma série de potências que converge rapidamente quando z possui grandes valores e lentamente para z com valores pequenos. Ainda, se $\gamma = 1$, essa distribuição da fragilidade se degenera na massa pontual com $Z = 1$ (WIENKE, 2010). Este autor também ressalta que todos os momentos dessa distribuição são infinitos, consequentemente, a esperança da fragilidade é infinita e a variância não existe.

Por sua vez, a transformada de Laplace da distribuição de fragilidade PE é definida, segundo Wienke (2010), como

$$\mathscr{L}_Z(s) = \exp\{-s^{\gamma}\}. \tag{2.26}$$

2.3.4 Distribuição de fragilidade para a família PVF

A família PVF é caracterizada por Wienke (2010) como uma classe composta de distribuições que possuem a transformada de Laplace explícita, consequentemente, sendo mais fácil de se obter uma forma fechada para as funções de sobrevivência, de risco e densidade de probabilidade, o que simplifica a estimativa do parâmetro. Outra vantagem é o seu uso para estruturar algumas distribuições famosas mediante a algumas restrições paramétricas, sendo

evidenciado nos trabalhos de Calsavara *et al.* (2017) e Rodrigues *et al.* (2021). Ainda, fazem parte desta família distribuições como: Gama, IG, Hougaard, PE, Normal, dentre outras.

A importância da família PVF é destacada na literatura por muitos autores. Deste modo, mesmo Wienke (2010) apresentando um estudo detalhado sobre este tópico, neste trabalho serão utilizadas as definições estabelecidas por Hougaard (2000). Para isso, este autor constrõe uma nova parametrização para esta classe, sendo esta analisada em Bedia (2022) que denota a família de distribuições PVF como PVF(γ,μ,σ), admitindo que $\gamma \leq 1$, $\mu > 0$ e σ caracterizado para dois casos: se $0 < \gamma \leq 1$ é tomado que $\sigma \geq 0$, caso contrário, se $\gamma \leq 0$ é assumido que $\sigma > 0$. Como consequência, mesmo com esta nova parametrização, ainda há a inclusão das distribuições Gama, PE e IG nesta família. Outra característica marcante deste modelo é que a variância da variável aleatória será uma função de potência da sua média (BEDIA, 2022).

Em suas pesquisas, Hougaard (2000) representa a f.d.p da família PVF sob duas parametrizações. O primeiro caso, retrata quando a distribuição PVF está concentrada em números positivos, ao considerar $0 < \gamma \leq 1$. Assim, define a f.d.p da família PVF para este caso como

$$g(z) = \exp\left\{-\sigma z + \frac{\mu\sigma^{\gamma}}{\gamma}\right\}\left(-\frac{1}{\pi z}\right)\sum_{k=1}^{\infty}\frac{\Gamma(k\gamma+1)}{k!}\left(-z^{-\gamma}\frac{\mu}{\gamma}\right)^{k}\sin(\gamma k\pi).$$

Já o segundo caso, quando $\gamma \leq 0$, representa quando a distribuição PVF está concentrada nos números positivos ou zero, implicando na existência de alguns grupos que possuem risco nulo. Deste modo, a f.d.p da família PVF deste caso é dada por

$$g(z) = \exp\left\{-\sigma z + \frac{\mu\sigma^{\gamma}}{\gamma}\right\}\left(\frac{1}{z}\right)\sum_{k=1}^{\infty}\frac{(\frac{-\mu z^{-\gamma}}{\gamma})^{k}}{k!\,\Gamma(-k\gamma)}. \tag{2.27}$$

Note que, a função Gama não é necessariamente definida, garantindo assim a existência e a validade da f.d.p (2.27). Ainda, observe que essa expressão é válida para todos os valores de γ, exceto 0 e 1, ao considerar a convenção de que quando a função Gama no denominador é indefinida, isto é, quando $k\gamma$ é um inteiro positivo, o termo inteiro na soma será igual a zero (HOUGAARD, 2000).

Somado a isso, a transformada de Laplace do modelo PVF é dada por

$$\mathscr{L}_Z(s) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+s)^{\gamma}-\sigma^{\gamma}\right]\right\};\ \ \gamma \leq 1,\ \ \mu > 0\ \text{ e }\ \sigma \geq 0. \tag{2.28}$$

Hougaard (2000) destaca que a média e a variância da variável aleatória de fragilidade $Z \sim \text{PVF}(\gamma,\mu,\sigma)$ é determinada ao aplicar as derivadas da transformada de Laplace (2.28) nas definições (2.12) e (2.13), obtendo que

$$\mathbb{E}[Z] = \mu\sigma^{\gamma-1} \quad \text{e} \quad Var[z] = \mu(1-\gamma)\sigma^{\gamma-2}.$$

2.3.5 Distribuição de fragilidade Poisson Composta

A distribuição Poisson Composta foi introduzida por Aalen (1992) como uma distribuição de fragilidade sendo construída como a soma de distribuições de Poisson de variàveis aleatórias independentes e identicamente distribuídas por uma distribuição Gama. Consequentemente, como enfatiza Bedia (2022), mesmo que a densidade da parte contínua seja dada apenas como uma série infinita que deve ser calculada numericamente, esta distribuição é matematicamente conveniente. Esta autora constroí a distribuição Poisson Composta da seguinte forma: toma-se a fragilidade Z dos indivíduos da população dada por

$$Z = \begin{cases} X_1 + X_2 + \ldots + X_N; & N > 0 \\ 0; & N = 0, \end{cases}$$

tal que $N \sim Poisson(\rho)$ com $\rho > 0$, as variàveis $X_1, X_2, \ldots, X_N$ são independentes e identicamente distribuídas de forma que $X_i \sim \Gamma(k, \lambda)$ com $k > 0$, $\lambda > 0$, $i \geq 1$ e N é independente de X_i. Assim, ao considerar a transformada de Laplace da distribuição de fragilidade Gama descrita em (2.21) para as variàveis X_i e a função geradora de probabilidades da distribuição Poisson como $\phi_N(s) = \exp\{-\rho(1-e^{-s})\}$ é mostrado que

$$\mathscr{L}_Z(s) = \exp\left\{-\rho\left(1-\left(1+\frac{s}{\lambda}\right)^{-k}\right)\right\}. \tag{2.29}$$

Autores como Wienke (2010) e Hougaard (2000) ressaltam a caracterização desta distribuição em seus trabalhos. Em particular, Hougaard (2000) sugere uma parametrização para o modelo de fragilidade Poisson Composta associada a transformada de Laplace (2.28) ao tomar $\gamma < 0$. Desta forma, ao considerar $k = -\gamma$, $\lambda = \sigma$ e $\rho = -\dfrac{\mu\sigma^{\gamma}}{\gamma}$ em (2.29) será obtido do modelo de fragilidade PVF (2.28) com $\gamma < 0$.

Um estudo mais detalhado sobre os tópicos desta seção pode ser encontrado em Wienke (2010), Hougaard (2000) e Bedia (2022).

2.4 Modelo exponencial por partes (MEP)

O modelo MEP é um recurso viável na literatura durante o processo de modelagem e análise de dados, uma vez que, constitui uma alternativa semi-paramétrica às distribuições paramétricas, além de ser um recurso muito empregado devido a sua capacidade de acomodar funções de risco com diversas formas, não havendo a necessidade de impor restrições para obter um ajuste adequado do modelo aos dados. Deste modo, este modelo é caracterizado mediante a aproximação da função de risco por segmentos de retas, sendo definidos através de J intervalos determinados na partição $\tau = \{s_0, \ldots, s_J\}$, segundo Ibrahim, Chen e Sinha (2001) e Sibim (2011). Em outras palavras, inicialmente, é considerada uma variàvel aleatória não-negativa T que representa o tempo de sobrevivência de interesse em que, posteriormente, será especificada uma partição $\tau = \{s_0, \ldots, s_J\}$ de forma que $0 = s_0 < s_1 < \ldots < s_J < \infty$. Isso implica que é

tomado para cada intervalo disjunto $I_j = (s_{j-1}, s_j]$ $(j = 1, \ldots, J)$ uma função de risco constante. Consequentemente, a função de risco acumulado que é associada ao j-ésimo intervalo é expressa por

$$H_{MEP}(t) = \sum_{m=1}^{j-1} \lambda_m(s_m - s_{m-1}) + \lambda_j(t - s_{j-1}) \; ; \quad \forall t \in I_j \quad \text{e} \quad \forall \lambda_j > 0. \tag{2.30}$$

Observe que esta função é composta pela soma das áreas dos retângulos cujas bases são determinadas pelos intervalos definidos em $\tau = \{s_0, \ldots, s_J\}$ e as alturas são representadas pela função de risco. Já a função de sobrevivência do modelo MEP é definida, mediante a identidade $S(t) = \exp\{-H(t)\}$, por

$$S_{MEP}(t) = \begin{cases} \exp\{-\lambda_1 t\}; & t \in I_1 \\ exp\{-\left[\sum_{m=1}^{j-1} \lambda_m(s_m - s_{m-1}) + \lambda_j(t - s_{j-1})\right]\}; & t \in I_j. \end{cases} \tag{2.31}$$

Por sua vez, a f.d.p correspondente a este modelo é dada por

$$f_{MEP}(t) = \begin{cases} \lambda_1 \exp\{-\lambda_1 t\}; & t \in I_1 \\ \lambda_j exp\{-\left[\sum_{m=1}^{j-1} \lambda_m(s_m - s_{m-1}) + \lambda_j(t - s_{j-1})\right]\}; & t \in I_j. \end{cases} \tag{2.32}$$

Já as funções de risco e de distribuição acumulada do modelo MEP são apresentadas, respectivamente, por

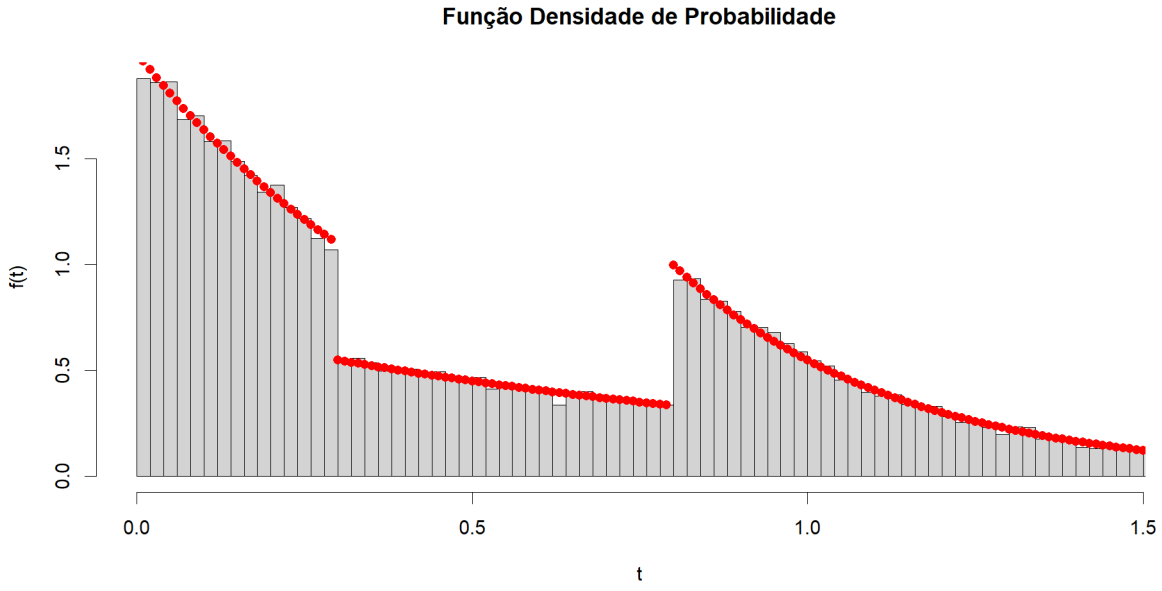
$$h_{MEP}(t) = \begin{cases} \lambda_1; & t \in I_1 \\ \lambda_j; & t \in I_j \end{cases} \tag{2.33}$$

e

$$F_{MEP}(t) = \begin{cases} 1 - \exp\{-\lambda_1 t\}; & t \in I_1 \\ 1 - exp\{-\left[\sum_{m=1}^{j-1} \lambda_m(s_m - s_{m-1}) + \lambda_j(t - s_{j-1})\right]\}; & t \in I_j. \end{cases} \tag{2.34}$$

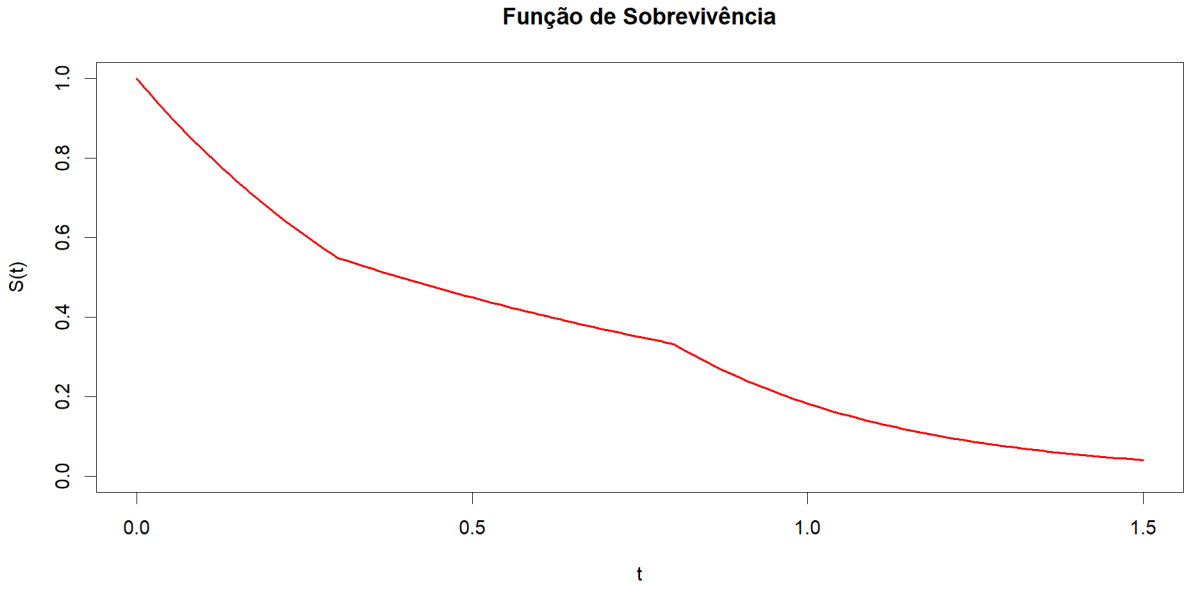
Observe que o modelo MEP se reduz ao modelo Exponencial, quando $J = 1$. Ainda, ao aumentar a quantidade de intervalos J's, este modelo é capaz de capturar qualquer forma do risco subjacente nos intervalos, tornando-se uma abordagem flexível e muito usada (SIBIM, 2011). As Figuras 3, 4, 5 e 6 ilustram os gráficos da f.d.p e das funções de sobrevivência, da distribuição acumulada e do risco acumulado, respectivamente, para o modelo MEP com $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$. Os mesmos foram gerados via o *software* R atráves do pacote **PWEXP** (TEAM, 2020) considerando três partições no eixo dos tempos (J = 3) com $I_1 = (0; 0,3]$, $I_2 = (0,3; 0,8]$ e $I_3 = (0,8; \infty)$ para $n = 151$.

Figura 3 – Gráfico da f.d.p do modelo MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



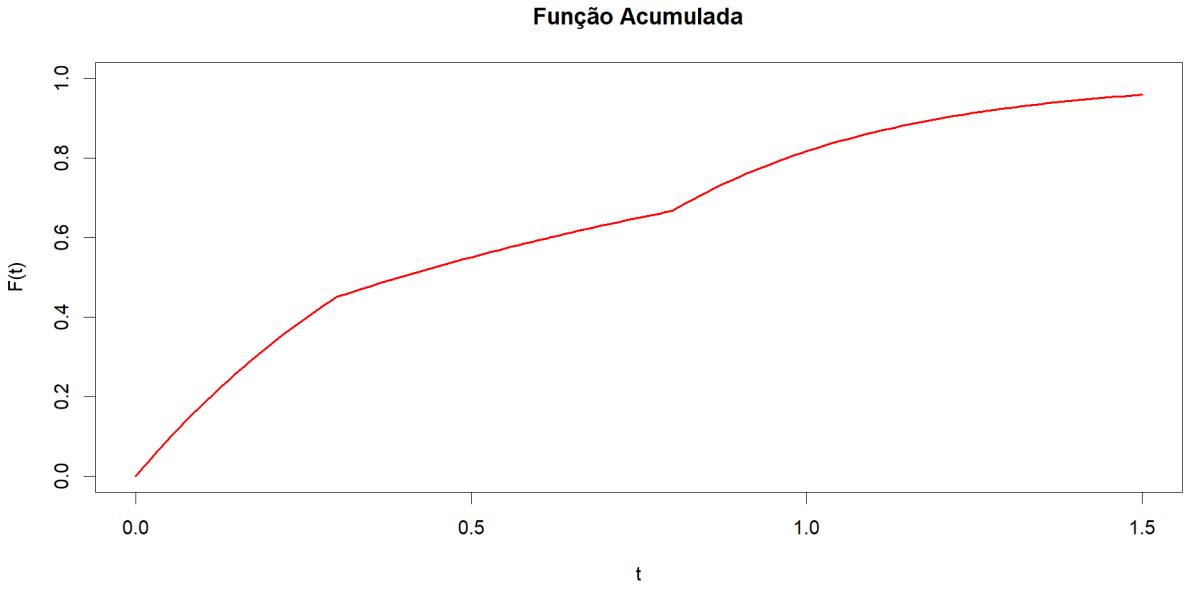
Fonte: Elaborada pelo autor.

Figura 4 – Gráfico da função de sobrevivência do modelo MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



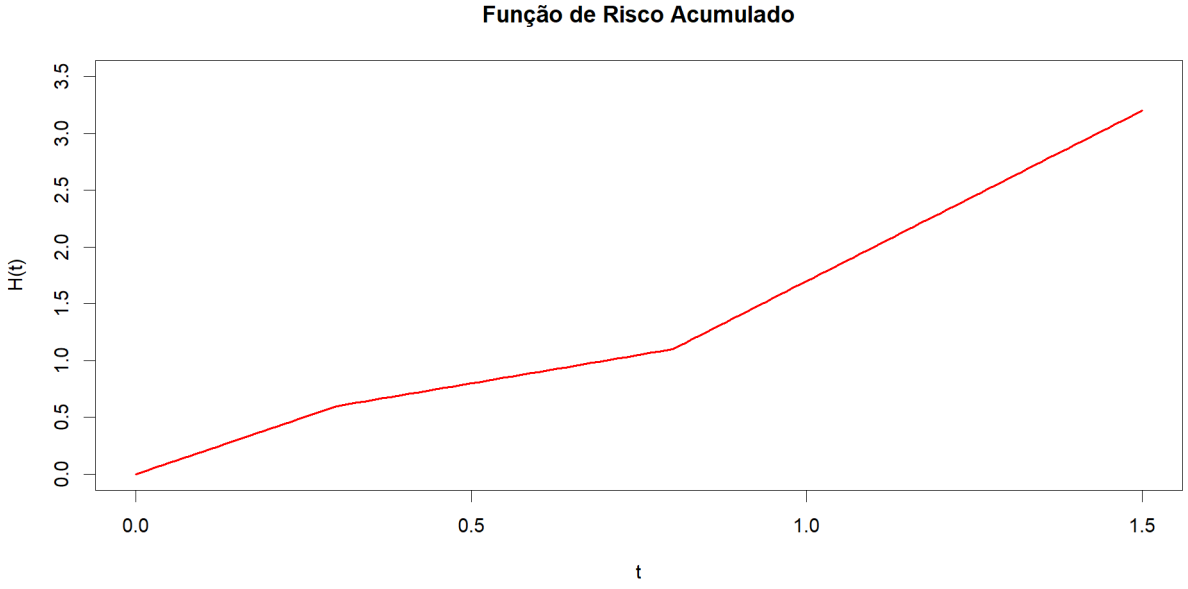
Fonte: Elaborada pelo autor.

Figura 5 – Gráfico da função de distribuição acumulada do modelo MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



Fonte: Elaborada pelo autor.

Figura 6 – Gráfico da função de risco acumulado do modelo MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



Fonte: Elaborada pelo autor.

Contudo, a partição $\tau = \{s_0, \ldots, s_J\}$ é fundamental para a qualidade do ajuste do modelo MEP. Uma partição com um grande número de intervalos pode fornecer estimativas instáveis para as taxas de risco, enquanto uma partição com poucos intervalos pode produzir uma aproximação ineficaz para a verdadeira função de sobrevivência. Deste modo, é sempre desejado a determinação de uma partição que proporcione um equilíbrio entre boas aproximações para as funções de risco e de sobrevivência. Esta questão tem sido um dos maiores desafios ao se trabalhar com o modelo MEP (DEMARQUI, 2010). Um estudo detalhado sobre este modelo é encontrado em Ibrahim, Chen e Sinha (2001), Sibim (2011) e Demarqui (2010).

2.5 Inferência bayesiana

A inferência estatística tem o objetivo de realizar o estudo das características de uma população através da informação sobre uma quantidade de interesse, podendo ser determinada mediante as abordagens bayesiana e frequentista. Em particular, a inferência bayesiana descreve a incerteza sobre o parâmetro por meio dos modelos probabilísticos, assumindo que a probabilidade é subjetiva, ou seja, cada pesquisador definirá o modelo probabilístico que melhor descreve o parâmetro conforme sua visão de mundo (PAIXÃO, 2021). Deste modo, esta abordagem utiliza do teorema de Bayes para tratar das características de interesse desconhecidas como variàveis aleatórias, além de permitir que a informação externa seja incorporada na análise dos dados. Para isso, considere uma amostra $\mathbf{x} = (x_1, \ldots, x_n)$ coletada de um conjunto de dados de tamanho n e $\theta = (\theta_1, \ldots, \theta_d)$ representando os d parâmetros associados ao modelo assumido para $\mathbf{x}$. Logo, ao usar o teorema de Bayes, a distribuição conjunta a posteriori de θ dado $\mathbf{x}$ é dada por

$$\pi(\theta \mid \mathbf{x}) = \frac{\pi(\theta, \mathbf{x})}{\pi(\mathbf{x})} = \frac{\pi(\mathbf{x} \mid \theta)\pi(\theta)}{\pi(\mathbf{x})} \propto \pi(\mathbf{x} \mid \theta)\pi(\theta), \tag{2.35}$$

em que $\pi(\theta, \mathbf{x})$ é a distribuição conjunta de θ e $\mathbf{x}$, $\pi(\mathbf{x} \mid \theta)$ é a função de verossimilhança, $\pi(\theta)$ é a distribuição a priori e $\pi(\mathbf{x})$ é a constante normalizadora de $\pi(\theta \mid \mathbf{x})$ para que a posteriori possa integrar em 1 (quando θ é contínua) ou somar em 1 (caso θ seja discreta), sendo também denominada de verossimilhança marginal.

Note que a expressão (2.35) representa toda a informação atualizada sobre o parâmetro θ por meio de um modelo probabilístico. Assim, a distribuição a priori representará o conhecimento atual sobre θ antes de se considerar qualquer tipo de informação relacionada às características de interesse (HARTMANN, 2015). Contudo, inferências sobre os parâmetros do modelo podem ser obtidas através do cálculo da esperanças a posteriori, ou seja

$$\mathbb{E}_{\theta|\mathbf{x}}[g(\theta)] = \int_{\Theta} g(\theta)\pi(\theta \mid \mathbf{x})d\theta, \tag{2.36}$$

em que Θ é o espaço paramétrico e $g : \Theta \longrightarrow \mathbb{R}$. Entretanto, em muitas aplicações práticas nem sempre essa integração possui solução analítica sendo necessário o emprego de métodos de aproximação para sua determinação. Dentre estes métodos, o método de Monte Carlo se destaca na literatura como um recurso viável para realizar esta tarefa.

O método de Monte Carlo foi proposto por Metropolis e Ulam (1949) sendo baseado em aproximações estocásticas para integrais multivariadas. Este método é fundamentado na Lei Forte dos Grandes Números e garante que, ao possuir uma amostra aleatória de $\pi(\theta|\mathbf{x})$ é possível aproximar o valor de (2.36) da seguinte forma

$$\widehat{\mathbb{E}_{\theta|\mathbf{x}}[g(\theta)]} = \frac{1}{N}\sum_{i=1}^{N} g(\theta^{(i)}) \longrightarrow \mathbb{E}_{\theta|\mathbf{x}}[g(\theta)].$$

Contudo, em dimensões muito elevadas a obtenção de valores para a distribuição a posteriori ainda continua não sendo uma tarefa fácil, rápida e prática. Esta dificuldade impulsionou na criação de novos métodos e recursos que facilitam este processo. Dentre os mais famosos e utilizados estão: o método MCMC (METROPOLIS *et al.*, 1953), o MH (HASTINGS, 1970) e o método HMC (DUANE *et al.*, 1987).

2.5.1 Método de Monte Carlo via Cadeias de Markov (MCMC)

Os estudos propostos por Xavier (2019) caracterizam uma cadeia de Markov como um processo estocástico em que dado o estado presente, os estados passado e futuro são independentes. Este autor define e escreve formalmente esta propriedade em seus trabalhos mediante a teoria de conjuntos e de probabilidades. Similar a ele, Hartmann (2015) formaliza e descreve importantes definições que serão destacadas a seguir.

Definição 1. (Processo Estocástico) Um processo estocástico é uma função de dois argumentos $X(t,\omega) : T\ X\ \Omega \longrightarrow \mathbb{R}$, em que T é um subconjunto dos reais. Assim, para $t^* \in T$ fixo, $X(t^*,\omega)$ é uma variável aleatória e para $\omega^* \in \Omega$ fixo, $X(t,\omega^*)$ é uma realização do processo, ou seja, uma função determinística.

Agora, assuma uma dada coleção de tempos $t = (t_1,\ldots,t_n)$. A função $\mathbf{X} = (X_{t_1},\ldots,X_{t_n})$ segue uma f.d.p conjunta tal que

$$\mathbf{X} \sim g(\mathbf{x}); \quad \forall t \in T \ \text{ e } \ \forall n \in \mathbb{N},$$

em que $\mathbf{x} = (x_{t_1},\ldots,x_{t_n})$ e $X(t_n) = X_{t_n}$.

Definição 2. (Cadeia de Markov) Um processo estocástico é dito ser uma Cadeia de Markov se

$$g(x_{t_s} \mid x_{t_{s-1}},\ldots,x_{t_1}) = g(x_{t_s} \mid x_{t_{s-1}}); \quad \forall t_s \in T \ \text{ e } \ \forall s \in \mathbb{N}.$$

Consequentemente, a f.d.p conjunta pode ser escrita como:

$$g(\mathbf{x}) = g(x_{t_1})\prod_{i=2}^{n} g(x_{t_i} \mid x_{t_{i-1}}).$$

Ademais, são necessárias algumas condições em relação à cadeia Markoviana para que o método de Monte Carlo seja adequado. Por meio destas condições, é garantido que

independente do valor inicial da cadeia, com o limite de $t \longrightarrow \infty$, a função X_t será amostrada de uma distribuição invariante. Em consequência disso, é possível destacar outras três definições importantes: probabilidade de transição, distribuição invariante e equação de balanço detalhada. Estas definições são formalmente descritas a seguir, segundo Hartmann (2015).

Definição 3. (Probabilidade de transição) Para todo $x \in S$ e $A \subseteq S$ é definido $T^n(x,A)$ como a probabilidade condicional da cadeia se encontrar numa região A depois de n passos dado o início em x, ou seja, $T^n(x,A) = \mathbb{P}(X_n \in A \mid X_0 = x)$. Ainda, $T^n(x,y)$ é dita a f.d.p condicional, isto é, $T^n(x,y) = g_{X_n}(y \mid X_0 = x)$.

Definição 4. (Distribuição invariante) A Cadeia de Markov com probabilidade de transição $T(x,A)$ possui uma distribuição invariante $v(x)$ se, para todo conjunto $A \subset S$, segue que

$$\int_A v(x)dx = \int T(x,A)v(x)dx.$$

Definição 5. (Equação de balanço detalhada) A probabilidade de transição $T(x,y)$ satisfaz a equação de balanço detalhada com respeito a densidade $v(x)$ se, para todo $x,y \in S$, é válido que

$$T(x,y)v(x) = T(y,x)v(y).$$

Esta última definição ressalta que após um determinado tempo, a taxa com que a cadeia passa de x para y é mesma que passa de y para x (HARTMANN, 2015).

2.5.2 O algoritmo de Metropolis-Hastings (MH)

Hartmann (2015) destaca que há situações em que a amostragem direta de valores para a distribuição a posteriori não é uma tarefa fácil de se realizar. Contudo, Hastings (1970) desenvolve um algoritmo que promove a simulação de valores aleatórios de uma distribuição alvo mediante a uma distribuição auxiliar, consequentemente, fazendo com que a tarefa de gerar valores para uma distribuição se torne simplificada e fácil. Como exemplo desta situação, Hartmann (2015) supôs que se deseja gerar valores de uma distribuição qualquer $p(\mathbf{x})$. Para isso, é assumida uma distribuição auxiliar $q(\mathbf{x})$ que pertença ao mesmo domínio de $p(\mathbf{x})$ e que seja gerada de forma simplificada. Logo, é possível aplicar o algoritmo de MH segundo os seguintes passos:

1. Tome $\mathbf{X}_0 = \mathbf{x}^{(0)}$ gerado da distribuição auxiliar $q(\mathbf{x})$.
2. Para $n = 1,2,\ldots,n$ segue que:
   - Faça $\mathbf{X}_{n-1} = \mathbf{x}^{(n-1)}$.
   - Gere $\mathbf{y}$ a partir de $q(\mathbf{y} \mid \mathbf{x}^{(n-1)})$ e $u \sim U(0,1)$.
   - Calcule $\mathbb{P}(\mathbf{x}^{(n-1)},\mathbf{y}) = \min\left\{1, \dfrac{p(\mathbf{y})q(\mathbf{x}^{(n-1)} \mid \mathbf{y})}{p(\mathbf{x})q(\mathbf{y} \mid \mathbf{x}^{(n-1)})}\right\}$.

- Faça

$$\mathbf{X}_n = \mathbf{x}^{(n)} = \begin{cases} \mathbf{y}, & \text{com probabilidade } \mathbb{P}(\mathbf{x}^{(n-1)}, \mathbf{y}) > u; \\ \mathbf{x}^{(n-1)}, & \text{caso contrário.} \end{cases}$$

Observe que o algoritmo MH faz uma transição de $\mathbf{x}$ para $\mathbf{y}$ de acordo com a distribuição auxiliar $q(\cdot)$, além de aceitar este valor com probabilidade $\mathbb{P}(\mathbf{x}, \mathbf{y})$. Consequentemente, a probabilidade de transição é dada por

$$T(\mathbf{x}, \mathbf{y}) = q(\mathbf{y} \mid \mathbf{x}) \mathbb{P}(\mathbf{x}, \mathbf{y}).$$

Note que a probabilidade de transição satisfaz as equações de balanço detalhada, pois

$$\begin{aligned} T(\mathbf{x}, \mathbf{y})\ p(\mathbf{x}) &= q(\mathbf{y} \mid \mathbf{x})\ \mathbb{P}(\mathbf{x}, \mathbf{y})\ p(\mathbf{x}) \\ &= q(\mathbf{y} \mid \mathbf{x})\ \min\left\{1, \frac{p(\mathbf{y}) q(\mathbf{x} \mid \mathbf{y})}{p(\mathbf{x}) q(\mathbf{y} \mid \mathbf{x})}\right\}\ p(\mathbf{x}) \\ &= \min\left\{p(\mathbf{y}) q(\mathbf{x} \mid \mathbf{y}), p(\mathbf{x}) q(\mathbf{y} \mid \mathbf{x})\right\} \\ &= q(\mathbf{x} \mid \mathbf{y})\ \min\left\{1, \frac{p(\mathbf{x}) q(\mathbf{y} \mid \mathbf{x})}{p(\mathbf{y}) q(\mathbf{x} \mid \mathbf{y})}\right\}\ p(\mathbf{y}) \\ &= q(\mathbf{x} \mid \mathbf{y})\ \mathbb{P}(\mathbf{y}, \mathbf{x})\ p(\mathbf{y}) \\ &= T(\mathbf{y}, \mathbf{x})\ p(\mathbf{y}). \end{aligned}$$

Hartmann (2015) também ressalta que quando as condições de balanço são satisfeitas em relação a uma medida de probabilidade p então a Definição 4 é válida em relação a esse p. Deste modo, o algoritmo MH gera uma Cadeia de Markov com distribuição invariante p. Ainda, este autor enfatiza que a Definição 5 é uma condição suficiente que fornece a estacionariedade, a irredutibilidade e a periodicidade da cadeia Markoviana.

2.5.3 Monte Carlo Hamiltoniano (HMC)

Hartmann (2015) aponta que a essência do HMC é simular ao movimento de uma partícula que se desloca sob uma energia potencial. Assim, a cada iteração, a velocidade desta partícula é aleatorizada e é simulado o seu movimento por algum tempo. Ao final, é obtida a nova posição desta partícula que representará o novo valor proposto da distribuição alvo no algoritmo do método HMC. Posteriormente, será aplicada a regra de MH para determinar se este valor deve ou não ser aceito.

Na literatura, autores como Xavier (2019), Paixão (2021) e Hartmann (2015) descrevem o algoritmo do método HMC para a geração de dados. Deste modo, similar a descrição apresentada em Hartmann (2015) e Paixão (2021) é considerado o vetor paramétrico aleatório $\boldsymbol{\theta} \in \mathbb{R}^d$ e o vetor aleatório auxiliar e independente $\mathbf{p} \in \mathbb{R}^d$ tal que $\mathbf{p} \sim N_d\ (0, M)$, com M representando a matriz de covariância referente a distribuição Normal Multivariada. Desta maneira, define-se a

função Hamiltoniana de forma que

$$H\ (\theta, \mathbf{p}) = U(\theta) + C(\mathbf{p}),$$

com

$$U\ (\theta) = -l(\theta) = -\log\left[\pi(\theta \mid x)\right] - \log\left[\pi(\theta)\right] \qquad \text{e} \qquad C\ (\mathbf{p}) = \mathbf{p}^T M^{-1}\mathbf{p},$$

de forma que $\pi(\theta \mid x)$ representa a distribuição a posteriori e $\pi(\theta)$ descreve a distribuição a priori empregadas no estudo. Assim, sob estas hipóteses, define-se a f.d conjunta de $(\theta, \mathbf{p})$ por

$$\pi(\theta, \mathbf{p}) \propto \exp\{-H\ (\theta, \mathbf{p})\} \propto \pi(\theta \mid x)\ \pi(\theta)\ \exp\{-\mathbf{p}^T M^{-1}\mathbf{p}\}.$$

O método HMC propõe valores iniciais para que o vetor $(\theta, \mathbf{p})$ obtenha valores gerados por meio a duas etapas antes que estes valores sejam submetidos ao passo de aceitação do método de MH. Assim, na primeira etapa, o vetor $\mathbf{p}$ é simulado por meio de uma distribuição Normal com média $\mathbf{0}$ e matriz de covariância M, sendo esta última independente de θ. Já na segunda etapa será simulado um sistema conjunto de $(\theta, \mathbf{p})$ que obedeça a dinâmica Hamiltoniana. Esta dinâmica é definida através do sistema de equações diferenciais definidos por

$$\frac{\partial \theta}{\partial t} = \frac{\partial H(\theta, \mathbf{p})}{\partial \mathbf{p}} = \nabla_{\mathbf{p}} C(\mathbf{p})$$

e

$$\frac{\partial \mathbf{p}}{\partial t} = -\frac{\partial H(\theta, \mathbf{p})}{\partial \theta} = \nabla_{\theta} U(\theta),$$

em que $\nabla_{\mathbf{p}}$ e ∇_{θ} representam os vetores gradientes em relação aos vetores $\mathbf{p}$ e θ, respectivamente.

Contudo, Hartmann (2015) destaca a necessidade de utilizar métodos numéricos para obter uma aproximação destas equações Hamiltonianas. Logo, o método de *Stormer-Verlet* ou método *leapfrog*, é destacado na literatura como um recurso valioso ao garantir a convergência das Cadeias de Markov para a distribuição alvo. Em suma, inicialmente, é determinado valores iniciais $(\theta^{(I)}, \mathbf{p}^{(I)})$ com um tempo fictício $I = 0$ para que posteriormente, seja aplicado o método de *Stormer-Verlet* nas equações diferenciais da dinâmica Hamiltoniana de forma que

$$\mathbf{p}^{(I+\frac{\varepsilon}{2})} = \mathbf{p}^{(I)} + \frac{\varepsilon}{2}\ \nabla_{\theta} l(\theta^{(I)}),$$

$$\theta^{(I+\varepsilon)} = \theta^{(I)} + \varepsilon\ \nabla_{\mathbf{p}} C(\mathbf{p}^{(I+\frac{\varepsilon}{2})})$$

e

$$\mathbf{p}^{(I+\varepsilon)} = \mathbf{p}^{(I+\frac{\varepsilon}{2})} + \frac{\varepsilon}{2}\ \nabla_{\theta} l(\theta^{(I+\varepsilon)}).$$

Ao final é simulado o valor do sistema em relação ao tempo fictício, sendo denotado por $(\theta^{(t_1)}, \mathbf{p}^{(t_1)})$ tal que a diferença entre os Hamiltonianos será próxima de zero devido a discretização do sistema de equações diferenciais (HARTMANN, 2015). Como resultado, é aplicada a etapa de aceitação do método MH para corrigir o erro introduzido no sistema de

equações, no qual, $(\theta^{(t_1)}, \mathbf{p}^{(t_1)})$ será aceito como a próxima etapa da Cadeia de Markov com probabilidade dada por

$$\mathbb{P}[(\theta^{(t_1)}, \mathbf{p}^{(t_1)}) \ , \ (\theta^{(I)}, \mathbf{p}^{(I)})] = \min\{1, \exp\{H(\theta^{(I)}, \mathbf{p}^{(I)}) - H(\theta^{(t_1)}, \mathbf{p}^{(t_1)})\}\}.$$

Observe que as variáveis ε e M são parâmetros livres que determinam a rapidez com que a cadeia alcança a distribuição estacionária, com ε representando a discretização do sistema de equações diferenciais. Consequentemente, se ε possui valores muito grande, a solução do sistema se torna sem sentido. Caso contrário, não há uma mistura adequada de valores da posteriori. Além disso, M geralmente é tomada como a matriz identidade, pois é algo muito difícil de especificar, além de ser mais significativa em problemas complexos (HARTMANN, 2015).

Para exemplificar o algoritmo do método HMC, vamos supor que se deseja simular valores de $\pi(\theta|x)$ com $\theta \in \mathbb{R}^d$ e que M é a matriz identidade. Desta forma, as etapas desde algoritmo são descritas a seguir.

1. Forneça uma posição inicial: $\theta^{(0)}$.

2. Inicie um contador que represente o tamanho da cadeia: $i = 1, 2, \ldots, n$.

   - Gere $\mathbf{p}^* \sim N_d(\mathbf{0}, I)$ e $u \sim U(0,1)$.
   - Para a etapa inicial, faça $(\theta^{(I)}, \mathbf{p}^{(I)}) = (\theta^{(i-1)}, \mathbf{p}^*)$ e $H_0 = H(\theta^{(I)}, \mathbf{p}^{(I)})$.
   - Repita em número adequado de *loops* a solução numérica de *Stormer-Verlet* ao tomar

   $$\mathbf{p}^* = \mathbf{p}^* + \frac{\varepsilon}{2} \nabla_\theta l(\theta^{(i-1)}),$$

   $$\theta^{(i-1)} = \theta^{(i-1)} + \varepsilon \nabla_{\mathbf{p}} C(\mathbf{p}^*)$$

   e

   $$\mathbf{p}^* = \mathbf{p}^* + \frac{\varepsilon}{2} \nabla_\theta l(\theta^{(i-1)}).$$

   - Ao final da etapa, faça $(\theta^{(F)}, \mathbf{p}^{(F)}) = (\theta^{(i-1)}, \mathbf{p}^*)$ e $H_1 = H(\theta^{(F)}, \mathbf{p}^{(F)})$.
   - Determine a probabilidade de aceitação com

   $$\mathbb{P}[(\theta^{(F)}, \mathbf{p}^{(F)}) \ , \ (\theta^{(I)}, \mathbf{p}^{(I)})] = \min\{1, \exp\{H_0 - H_1\}\}.$$

   - Faça

   $$\theta^{(i)} = \begin{cases} \theta^{(F)}, & \text{com probabilidade } \mathbb{P}[(\theta^{(F)}, \mathbf{p}^{(F)}), (\theta^{(I)}, \mathbf{p}^{(I)})] > u; \\ \theta^{(I)}, & \text{caso contrário.} \end{cases}$$

Um estudo mais detalhado sobre os tópicos destas subseções pode ser encontrado em Paixão (2021), Xavier (2019), Hartmann (2015), Metropolis *et al.* (1953), Hastings (1970) e Ibrahim, Chen e Sinha (2001).

2.5.4 Critério de comparação de modelos bayesianos

Os critérios de comparação de modelos é um recurso usado em inferência bayesiana para casos em que as amostras das distribuições a posteriori dos parâmetros são obtidas ao usar os métodos MCMC. Dentre os diferentes critérios de comparação destacados na literatura, este trabalho utiliza o *Bridge Sampling* e a densidade preditiva ordenada condicional (*Conditional Predictive Ordinate* (CPO)). Autores como Gronau, Singmann e Wagenmakers (2017) destacam o *bridge sampling* como um método que funciona com precisão mesmo em espaços paramétricos de alta dimensão, além da abordagem de validação cruzada bayesiana muito ressaltada na literatura, sendo obtida mediante ao CPO (CANCHO *et al.*, 2018). Ambos os critérios são descritos a seguir.

2.5.4.1 Bridge sampling

Gronau *et al.* (2017) ressalta a importância da probabilidade marginal na teoria bayesiana, sendo fundamental para a determinação da estimativa de parâmetros, comparação de modelos e na estimação da média dos modelos. Em particular, para comparação de modelos, são considerados m $(m \in \mathbb{N})$ modelos concorrentes, no qual, o principal interesse de estudo é a plausibilidade relativa de um modelo particular M_l $(l = 1,\ldots,m)$ dada a probabilidade do modelo a priori em relação aos dados $\mathbf{x}$. Autores como Ibrahim, Chen e Sinha (2001) e Gronau *et al.* (2017) enfatizam que esta plausibilidade relativa é determinada mediante a distribuição a posteriori (2.35) ao considerar o modelo M_l dado os dados $\mathbf{x}$, sendo definida por

$$\pi(M_l \mid \mathbf{x}) \propto \frac{\pi(\mathbf{x} \mid M_l)\pi(M_l)}{\sum_{a=1}^{m} \pi(\mathbf{x} \mid M_a)\pi(M_a)}, \tag{2.37}$$

em que o denominador é a soma da verossimilhança marginal (ou constante normalizadora) multiplicada pela probabilidade a priori de todos os m modelos. Como consequência, se essa comparação ocorrer apenas para dois modelos M_1 e M_2, a equação (2.37) é usada para quantificar a plausibilidade relativa dos modelos a posterior M_1 em comparação ao modelo M_2, sendo dada por

$$\frac{\pi(M_1 \mid \mathbf{x})}{\pi(M_2 \mid \mathbf{x})} = \frac{\pi(\mathbf{x} \mid M_1)}{\pi(\mathbf{x} \mid M_2)} \times \frac{\pi(M_1)}{\pi(M_2)}, \tag{2.38}$$

em que o primeiro fator é a razão das probabilidades marginais de ambos os modelos, denominado como fator Bayes, e o segundo fator é a razão entre as probabilidades a priori dos modelos.

Gronau *et al.* (2017) destaca que a verossimilhança marginal é a probabilidade dos dados observados $\mathbf{x}$, dado um modelo específico de interesse M_l, sendo definida por

$$\pi(\mathbf{x} \mid M_l) = \int \pi(\mathbf{x} \mid \theta, M_l)\pi(\theta \mid M_l)d\theta, \tag{2.39}$$

tal que θ, $\pi(\mathbf{x}|\theta, M_l)$ e $\pi(\theta|M_l)$ são, respectivamente, o vetor paramétrico, a verossimilhança e a densidade a priori associadas ao modelo M_l. Ainda, note que a verossimilhança marginal (2.39)

pode ser reescrita como

$$\pi(\mathbf{x} \mid M_l) = \mathbb{E}_{priori}[\pi(\mathbf{x} \mid \theta, M_l)],$$

tendo a esperança tomada com relação a distribuição a priori. Esse conceito é muito utilizado em métodos de amostragem computacionais, uma vez que, como nem sempre a verossimilhança marginal é analiticamente tratável, promover sua aproximação por meio ao uso de métodos numéricos é um recurso viável enfatizado na literatura (GRONAU *et al.*, 2017). Autores como Meng e Wong (1996), Meng e Schilling (2002), Gronau *et al.* (2017) e Gronau, Singmann e Wagenmakers (2017) ressaltam o *bridge sampling* como um dos métodos mais promissores para a estimação destas verossimilhanças ou constantes. Em particular, Gronau, Singmann e Wagenmakers (2017) considera o *bridge sampling* como a generalização de métodos mais simples usados para estimar constantes de normalização. Para este autor, a diferença destes recursos é que os métodos mais simples utilizam de amostras de uma única distribuição, enquanto *bridge sampling* combina amostras de duas distribuições.

Um exemplo famoso desta abordagem é que na formulação original de Meng e Wong (1996), o *bridge sampling* é usado para estimar a proporção de duas constantes de normalização, como o fator de Bayes. Como consequência, as duas distribuições aplicadas foram as posteriores para cada um dos modelos envolvidos. Neste cenário, a precisão do estimador dependerá da sobreposição entre as duas distribuições envolvidas (GRONAU *et al.*, 2017).

Para este exemplo, é usada a ideia central de que a verossimilhança marginal pode ser escrita como um valor esperado com respeito a distribuição a priori, ou seja

$$\pi(\mathbf{x}) = \mathbb{E}_{priori}[\pi(\mathbf{x} \mid \theta)].$$

Assim, segundo Gronau, Singmann e Wagenmakers (2017), a função de verossimilhança marginal da formulação de Meng e Wong (1996) é definida por

$$\pi(\mathbf{x}) = \frac{\mathbb{E}_{g(\theta)}[h(\theta)\pi(\mathbf{x} \mid \theta)\pi(\theta)]}{\mathbb{E}_{\pi(\theta|\mathbf{x})}[h(\theta)g(\theta)]}, \tag{2.40}$$

tal que $h(\theta)$ é a função *bridge* e $g(\theta)$ é uma distribuição proposta que deve ser escolhida. Já o estimador do *bridge sampling* associado a função (2.40) é dado por

$$\hat{\pi}(\mathbf{x}) = \frac{\frac{1}{n_2}\sum_{b=1}^{n_2} h(\theta_{\mathbf{b}}^{**})\pi(\mathbf{x} \mid \theta_{\mathbf{b}}^{**})\pi(\theta_{\mathbf{b}}^{**})}{\frac{1}{n_1}\sum_{c=1}^{n_1} h(\theta_{\mathbf{c}}^{*})g(\theta_{\mathbf{c}}^{*})}, \tag{2.41}$$

com $\{\theta_1^*, \theta_2^*, \ldots, \theta_{n_1}^*\}$ representando o espaço n_1 para a distribuição a posteriori $\pi(\theta|\mathbf{x})$ e $\{\theta_1^{**}, \theta_2^{**}, \ldots, \theta_{n_2}^{**}\}$, o espaço n_2 para a distribuição proposta $g(\theta)$.

Observe que é preciso especificar as funções $h(\theta)$ e $g(\theta)$. Desta forma, Gronau, Singmann e Wagenmakers (2017) indica como exemplos de distribuições para serem usadas como

distribuições propostas: a distribuição normal multivariada com vetor médio e matriz de covariância que corresponderá às respectivas quantidades das amostras posteriores e a distribuição normal multivariada padrão combinada com uma distribuição a posteriori *Warped*. Segundo estes autores, ambas as escolhas aumentam a eficiência do estimador tornando a distribuição proposta e a posteriori o mais semelhantes possível, além promover um estimador de *bridge sampling* robusto.

Já a função *bridge* deverá ser uma função que minimiza o erro quadrático médio relativo do estimador, como por exemplo, a função ótima apresentada por Meng e Wong (1996) definida como

$$h(\theta) = \frac{C}{s_1\pi(\mathbf{x} \mid \theta)\pi(\theta) + s_2\pi(\mathbf{x})g(\theta)}, \tag{2.42}$$

com $s_1 = \dfrac{n_1}{n_1+n_2}$, $s_2 = \dfrac{n_2}{n_1+n_2}$ e C constante. Contudo, note que esta função depende justamente da verossimilhança marginal $\pi(\mathbf{x})$ que queremos aproximar. Como consequência, é realizada a junção das funções (2.41) e (2.42), além de um esquema de iterações que atualiza uma estimativa inicial da verossimilhança marginal com $t = 0$ até a estimativa da verossimilhança marginal convergir para um nível de tolerância predefinido (MENG; WONG, 1996). Logo, essa estimativa na iteração $t+1$ é obtida por

$$\hat{\pi}(\mathbf{x})^{(t+1)} = \frac{\dfrac{1}{n_2}\sum_{b=1}^{n_2} \dfrac{\pi(\mathbf{x} \mid \theta_{\mathbf{b}}^{**})\pi(\theta_{\mathbf{b}}^{**})}{s_1\pi(\mathbf{x} \mid \theta_{\mathbf{b}}^{**})\pi(\theta_{\mathbf{b}}^{**}) + s_2\hat{\pi}(\mathbf{x})^{(t)}g(\theta_{\mathbf{b}}^{**})}}{\dfrac{1}{n_1}\sum_{c=1}^{n_1} \dfrac{g(\theta_{\mathbf{c}}^{*})}{s_1\pi(\mathbf{x} \mid \theta_{\mathbf{c}}^{*})\pi(\theta_{\mathbf{c}}^{*}) + s_2\hat{\pi}(\mathbf{x})^{(t)}g(\theta_{\mathbf{c}}^{*})}},$$

com $\hat{\pi}(\mathbf{x})^{(t)}$ representando a estimativa da verossimilhança marginal na iteração t.

A utilidade do *bridge sampling* vem se ampliando nos mais diversos estudos estatísticos e probabilísticos, principalmente após a construção um pacote em R, denominado **bridgesampling** (TEAM, 2020), que realiza a estimativa direta da função de verossimilhança marginal e, consequentemente, de constantes de normalização por meio de técnicas de amostragem de *bridge*. Este pacote é apresentado por Gronau *et al.* (2017) e Gronau, Singmann e Wagenmakers (2017) que enfatizam a não exigência de que os usuários programem suas próprias rotinas MCMC para obter amostras a posteriores, além da produzição de uma computação não supervisionada da função de verossimilhança marginal como vantagens do uso do *bridge sampling*.

2.5.4.2 Conditional predictive ordinate (CPO)

Seja $\mathscr{D}$ os dados completos e $\mathscr{D}_{(-i)}$ os dados com a i-ésima observação excluída para $i = 1, \ldots, n$. Desta maneira, é denotada a densidade a posteriori de θ dado $\mathscr{D}_{(-i)}$ como $\pi(\theta|\mathscr{D}_{(-i)})$, em que θ representa o vetor paramétrico. Logo, a CPO_i para a i-ésima observação é dada por

$$CPO_i = \left(\int_\Theta \frac{\pi(\theta \mid \mathscr{D})}{g(t_i \mid \theta)} d\theta \right)^{-1},$$

com $g(t_i \mid \theta)$ representando a f.d.p para a i-ésima observação. Já a estimativa de Monte Carlo do CPO_i pode ser obtida através de uma amostra MCMC da distribuição a posteriori $\pi(\theta|\mathscr{D})$. Deste modo, segundo Ibrahim, Chen e Sinha (2001), ao considerar uma amostra de tamanho N referente a esta distribuição a posteriori, a estimativa de CPO_i é dada por

$$C\hat{P}O_i = \left(\frac{1}{N} \sum_{n=1}^{N} \frac{1}{g(t_i \mid \theta^{(n)})} \right)^{-1}.$$

O logaritmo da função de verossimilhança pseudo marginal (LPML) é definido como $LPML = \sum_{i=1}^{N} log(C\hat{P}O_i)$, no qual, quanto maior for o valor do LPML mais adequado será o modelo analisado (SIBIM, 2011).

2.5.5 Influência bayesiana global

A metodologia sobre a influência bayesiana global é empregada para identificar a presença de *outliers* ou observações influentes por meio do uso da divergência. Para isso, seja $D_\phi\ (\pi, \pi_{(-i)})$ a ϕ-divergência existente entre π e $\pi_{(-i)}$, sendo definida por

$$D_\phi\ (\pi, \pi_{(-i)}) = E_{\theta\ |\mathscr{D}} = \left[\phi \left(\frac{CPO_i}{g(t_i|\theta^{(n)})} \right) \right]; \quad i = 1, \ldots, n, \tag{2.43}$$

em que π é a distribuição a posteriori de θ para o conjunto de dados completos, $\pi_{(-i)}$ é a distribuição a posteriori de θ sem a i-ésima observação e ϕ é uma função convexa tal que $\phi(1) = 0$. Em consequência, diferentes tipos de funções podem ser definidas para ϕ, como aborda o estudo de Pardo (2018). Em particular, para este trabalho é considerada a divergência K-L como $\phi(z) = -log(z)$, a distância como $\phi(z) = (z-1)log(z)$, a distância variacional da norma L_1 como $\phi(z) = 0.5|z-1|$ e a divergência Qui-quadrada como $\phi(z) = z(1/z-1)^2$. Logo, a estimativa da divergência K-L para a função (2.43) é dada por

$$\hat{D}_\phi\ (\pi, \pi_{(-i)}) = -log(C\hat{P}O_i) + \frac{1}{N} \sum_{n=1}^{N} log(g(t_i \mid \theta^{(n)})).$$

Um estudo mais detalhado sobre os tópicos destas subseções pode ser encontrado em Meng e Wong (1996), Meng e Schilling (2002), Gronau *et al.* (2019), Gronau *et al.* (2017), Gronau, Singmann e Wagenmakers (2017), Sibim (2011), Ibrahim, Chen e Sinha (2001), Cancho, Rodrigues e Castro (2011), Cancho *et al.* (2018) e Pardo (2018).

2.6 Conclusão

Neste capítulo, foram apresentados os principais conteúdos para o desenvolvimento da tese, incluindo definições, referências e métodos utilizados. As Seções 2.1, 2.2 e 2.3 abordam a ideia de fragilidade ou heterogeneidade não observada, no qual, o conceito da transformada de Laplace é explorado, sendo fundamental para determinar alguns resultados dos modelos

de fragilidade. Já a família PVF é discutida como uma rica família de distribuições usadas na modelagem da variàvel de fragilidade na Subseção 2.3.4.

A Seção 2.4 trata do modelo MEP, essencial para este trabalho, que será utilizado como função de risco base nos modelos descritos nos capítulos seguintes. Finalmente, a Seção 2.5 apresenta a inferência bayesiana, detalhando as técnicas computacionais para estimação paramétrica, enfatizando o método HMC, e os critérios de comparação e seleção de modelos bayesianos, destacando os métodos CPO e *bridge sampling*.

CAPÍTULO

3

MODELO DEFEITUOSO INDUZIDO POR FRAGILIDADE

3.1 Introdução

Dentre as pesquisas realizadas em análise de sobrevivência é comum encontrar estudos, no qual, uma parcela de indivíduos não são suscetíveis ao evento de interesse. Estes estudos são geralmente analisados mediante a modelos denominados como modelos de longa duração, de fração de cura ou de mistura cuja característica marcante é a existência de uma proporção de indivíduos imunes ou sobreviventes ao evento de interesse do estudo. Na literatura os modelos de mistura padrão, propostos por Berkson e Gage (1952), são frequentemente utilizados para modelar conjuntos de dados que incorporam a existência destes indivíduos. Contudo, Balka, Desmond e McNicholas (2009) introduzem uma forma alternativa para realizar essa modelagem por meio das denominadas distribuições defeituosas. Assim, em vez de estimar diretamente a proporção de indivíduos imunes ao evento de interesse, como ocorre nos modelos de mistura padrão, são utilizadas distribuições que através da modificação do domínio dos seus parâmetros também conseguem determinar esta proporção. Estas distribuições recebem o nome de distribuições defeituosas, sendo caracterizadas por meio da integral de sua f.d.p que possuirá um valor $p_0 \in (0,1)$, quando o domínio dos seus parâmetros é modificado (ROCHA, 2016).

Neste cenário, autores como Rocha *et al.* (2016), Rocha *et al.* (2017a), Rocha *et al.* (2017b) e Scudilio *et al.* (2019) definem os modelos defeituosos como modelos que possuem uma distribuição defeituosa, em que é possível estimar a proporção de indivíduos imunes ao evento de interesse através de uma distribuição imprópria. Como consequência, em vez de estimar a proporção p_0 diretamente é usada uma distribuição, onde se é alterado o domínio de seus parâmetros, resultando em modelos que podem ser empregados como modelos de mistura. Assim, esta proporção é obtida ao calcular o limite da função de sobrevivência (da distribuição imprópria) usando os parâmetros estimados. Ainda, outra característica apresentada por estas

distribuições é que suas funções acumuladas não se aproximarão de 1, mas sim do valor p_0. Este fato implica que a função de sobrevivência relacionada a estes modelos se aproximará do valor correspondente à $(1-p_0)$ (ROCHA, 2016).

Na literatura, um exemplo de distribuição defeituosa é a distribuição Gompertz (GOMPERTZ, 1825). Observe que autores como Scudilio *et al.* (2019) e Rocha *et al.* (2016) enfatizam o uso desta distribuição para modelar dados de sobrevivência em diversas áreas do conhecimento, principalmente quando há a suspeita de risco exponencial. Assim, definem a f.d.p para a distribuição de Gompertz como

$$f_G(t) = be^{at}e^{-\frac{b}{a}(e^{at}-1)}; \quad t > 0, \tag{3.1}$$

com $a \in \mathbb{R}$ e $b > 0$. Já a função de sobrevivência da distribuição Gompertz é dada por

$$S_G(t) = e^{-\frac{b}{a}(e^{at}-1)}. \tag{3.2}$$

Consequentemente, a distribuição Gompertz defeituosa é definida ao permitir valores negativos para o parâmetro a, tornando-a uma distribuição imprópria cuja proporção de indivíduos imunes ao evento de interesse na população é calculada mediante ao limite da função de sobrevivência quando $a < 0$, ou seja

$$p_{0G} = \lim_{t\longrightarrow\infty} S(t) = \lim_{t\longrightarrow\infty} e^{-\frac{b}{a}(e^{at}-1)} = e^{\frac{b}{a}} \in (0,1).$$

Outro exemplo de distribuição defeituosa é a distribuição IG, proposta por Whttmore (1979). Esta distribuição também apresentará parâmetros positivos, contudo ao considerar valores negativos em relação ao seu parâmetro de forma, se torna uma distribuição defeituosa. Uma descrição mais detalhada sobre a distribuição IG e IG defeituosa pode ser encontrada nos trabalhos de Scudilio *et al.* (2019) e Rocha *et al.* (2016).

Na literatura, diversos autores destacam a importância das distribuições Gompertz e IG defeituosas em suas pesquisas. Santos, Achcar e Martinez (2017) propõem uma abordagem bayesiana para o modelo Gompertz defeituoso, comparando-o com o modelo de máxima verossimilhança, enquanto Borges (2017) usa o algoritmo EM para generalizar o modelo de regressão Gompertz defeituoso. Já Rocha *et al.* (2016) estendem os modelos Gompertz e IG por meio da família de distribuições Marshall-Olkin, possibilitando uma nova forma de gerar distribuições defeituosas construída por Rocha *et al.* (2017a). Por sua vez, Rocha *et al.* (2017b) aplicam as distribuições Gompertz e IG defeituosas em estudos sobre câncer, utilizando a família Kumaraswamy para sua extensão. Por fim, Scudilio *et al.* (2019) propõem um modelo defeituoso induzido por um termo de fragilidade para modelar a proporção de indivíduos imunes ao evento de interesse usando as distribuições Gompertz e IG defeituosas como funções de base.

Em um contexto geral, Bedia (2022) enfatiza a importância de escolher uma distribuição adequada para o termo de fragilidade, a fim de obter uma boa descrição da estrutura de dependência dos dados e evitar resultados tendenciosos. A escolha da distribuição de fragilidade

é um desafio na literatura, pois ela influencia tanto a simplicidade analítica e computacional quanto as propriedades específicas que o modelo pode exibir. Embora certas distribuições possam ser atrativas, elas nem sempre representam adequadamente os dados. Uma solução para esse problema é o uso de famílias de distribuições de fragilidade, como a família PVF. Essa família é considerada uma família exponencial natural cuja variância é uma função potência da média que inclui distribuições bem conhecidas e amplamente utilizadas, além de ser flexível para modelar a dependência e permitir a descrição de grupos com risco zero, representando indivíduos imunes ao evento de interesse (HOUGAARD, 2000).

Neste contexto, o objetivo deste capítulo é apresentar um modelo de fragilidade para a modelagem da heterogeneidade não observada em dados de sobrevivência. Este modelo é proveniente da composição entre uma distribuição de fragilidade caracterizada por meio da família PVF, segundo Hougaard (2000), e o modelo MEP, sendo estendido para permitir a construção de modelos defeituosos, resultando em uma fragilidade zero e enfatizando o impacto da presença de covariáveis nos estudos. A abordagem inferencial é baseada em métodos bayesianos mediante ao uso do método HMC implementado no R-Stan, no qual, alguns resultados de simulação são fornecidos para avaliar o desempenho das propriedades dos estimadores de Bayes. A importância deste modelo é ilustrada por meio de aplicações em conjuntos de dados reais. As demonstrações dos principais resultados deste capítulo estão descritas no Apêndice A.

3.2 Modelo PVF defeituoso

Nesta seção será apresentado o modelo proveniente da composição entre uma distribuição de fragilidade da família PVF e uma função de risco base com suporte em $\mathbb{R}$ que, posteriormente, será estendido para permitir a construção de modelos defeituosos. Para isso, considere T uma variável aleatória e não-negativa que representa o tempo até a ocorrência de um evento de interesse e Z uma variável de fragilidade aleatória, contínua e não-negativa. Deste modo, devido a estrutura de riscos proporcionais, é definida a função de risco condicional de T, dada a fragilidade Z, como

$$h\,(t \mid Z) = Z\, h_0(t); \quad t > 0,$$

com $h_0(\cdot)$ representando a função de risco base comum para todos os indivíduos. Consequentemente, a probabilidade que representará a proporção dos indivíduos sobreviventes condicionada a Z é expressa por (2.5), tendo seu respectivo modelo marginal caracterizado pela expressão (2.6).

Assuma que a variável Z no modelo marginal (2.6) é definida por meio de uma distribuição PVF, denotada como PVF(γ, μ, σ), com $\gamma \leq 1$, $\mu > 0$ e σ caracterizado para dois casos: se $0 < \gamma \leq 1$ é tomado que $\sigma \geq 0$, caso contrário, se $\gamma \leq 0$ é admitido que $\sigma > 0$ (HOUGAARD, 2000). Como consequência, sob esta parametrização, é definida a transformada de Laplace para

o modelo PVF(γ,μ,σ) dada por

$$\mathscr{L}_Z(s) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+s)^{\gamma}-\sigma^{\gamma}\right]\right\}, \tag{3.3}$$

tendo a média e a variância da variável Z dadas, respectivamente, por: $\mathbb{E}[Z] = \mu\sigma^{\gamma-1}$ e $Var[Z] = \mu(1-\gamma)\sigma^{\gamma-2}$. Observe que por meio a restrições paramétricas, a função (3.3) corresponderá a algumas funções de distribuições famosas da literatura, sendo esta uma característica marcante do modelo proposto por Hougaard (2000). Dentre estas distribuições estão:

- A distribuição Gama: ao tomar $\gamma \longrightarrow 0$, resultando em $\mathscr{L}_Z(s) = \left(1+\frac{s}{\sigma}\right)^{-\mu}$ com $\mathbb{E}[Z] = \frac{\mu}{\sigma}$ e $Var[Z] = \frac{\mu}{\sigma^2}$.
- A distribuição IG: ao considerar $\gamma = 0,5$, obtendo $\mathscr{L}_Z(s) = e^{\left\{2\mu\sigma^{\frac{1}{2}}\left[1-\sqrt{1+\frac{s}{\sigma}}\right]\right\}}$ e tendo $\mathbb{E}[Z] = \frac{\mu}{\sigma^{\frac{1}{2}}}$ e $Var[Z] = \frac{\mu}{2\sigma^{\frac{3}{2}}}$.
- A distribuição PE: ao assumir $\gamma = \mu$ e $\sigma = 0$ tal que $\mathscr{L}_Z(s) = e^{\{-s^{\mu}\}}$.

A função de sobrevivência marginal é definida ao aplicar a transformada de Laplace (3.3) na relação (2.7) por

$$S(t) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+H_0(t))^{\gamma}-\sigma^{\gamma}\right]\right\}; \quad t>0, \quad \gamma\leq 1, \quad \mu>0 \quad \text{e} \quad \sigma>0, \tag{3.4}$$

com $H_0(\cdot)$ representando a função de risco acumulado base comum para todos os indivíduos.

Observação 1. Ao considerar a função de sobrevivência marginal (3.4) com $\gamma \leq 1$, $\mu > 0$ e $\sigma > 0$ são válidas as seguintes afirmações:

1. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{\frac{b\gamma(e^{at}-1)+\sigma^{\gamma}}{a\mu}} - \sigma$, com $t>0$, $b>0$ e $a\in\mathbb{R}$ de forma que: $b\neq\mu$ e $a\neq\gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.
2. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{e^{\gamma t}-1+\sigma^{\gamma}} - \sigma$, com $t>0$, $b>0$ e $a\in\mathbb{R}$ de forma que: $b=\mu$ e $a=\gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

Para evitar problemas de identificabilidade, é considerada uma restrição no modelo (3.4) de forma que $\mu = \sigma = 1$, obtendo que $\mathbb{E}[Z] = 1$ e $Var(Z) = 1-\gamma$. Como consequência, a função de sobrevivência marginal em relação ao tempo T é dada por

$$S(t) = \exp\left\{-\frac{1}{\gamma}\left[(1+H_0(t))^{\gamma}-1\right]\right\}; \quad t>0 \quad \text{e} \quad \gamma\leq 1, \tag{3.5}$$

com $H_0(\cdot)$ representando a função de risco acumulado base comum para todos os indivíduos. Note que a função de sobrevivência (3.5) é um caso particular do modelo (3.4), sendo uma função não crescente, além de ter $\lim_{t\longrightarrow 0} S(t) = 1$ e

$$\lim_{t\longrightarrow\infty} S(t) = \begin{cases} 0, & 0 < \gamma \leq 1; \\ \exp\left\{\dfrac{1}{\gamma}\right\}, & \gamma < 0. \end{cases}$$

Consequentemente, a função (3.5) é uma função de sobrevivência própria ao assumir $0 < \gamma \leq 1$, mas será imprópria, nos casos em que $\gamma < 0$. Devido a essa caracterização é possível estabelecer a proporção de indivíduos imunes ao evento de interesse dada por

$$p_0 = \lim_{t\longrightarrow\infty} S(t) = \exp\left\{\frac{1}{\gamma}\right\} > 0; \quad \gamma < 0. \tag{3.6}$$

Além disso, devido a essa restrição paramétrica, note que a função de sobrevivência (3.5) pode ser caracterizada como um modelo defeituoso quando $\gamma < 0$. Como consequência, o modelo (3.5) será denominado como "modelo PVF defeituoso", quando $\gamma < 0$ e "modelo PVF", caso $0 < \gamma \leq 1$.

A f.d associada ao modelo (3.5) é dada por

$$f(t) = h_0(t)\ (1 + H_0(t))^{\gamma-1}\ \exp\left\{-\frac{1}{\gamma}[(1 + H_0(t))^{\gamma} - 1]\right\},$$

em que $h_0(\cdot)$ é a função de risco base comum para todos os indivíduos. Já a função de risco associada a (3.5) é definida como

$$h(t) = h_0(t)\ (1 + H_0(t))^{\gamma-1}.$$

Por sua vez, a função de risco acumulado correspondente ao modelo (3.5) é dada por

$$H(t) = \frac{(1 + H_0(t))^{\gamma} - 1}{\gamma}. \tag{3.7}$$

Note que ao assumir $\gamma < 0$ segue que $H(t) \longrightarrow -\log(p_0)$, quando $t \longrightarrow \infty$ e com p_0 definido por (3.6). Este fato implica que, sob esta hipótese, a função (3.7) é limitada por $-\log(p_0)$, ou seja, $H(t) \leq -\log(p_0)$, sendo esta mais uma característica proveniente em modelos defeituosos. Ainda, repare que são preservados os casos especiais das distribuições Gama e IG, uma vez que:

- Para a distribuição Gama, a identificabilidade do modelo provêm ao assumir que $\sigma = \mu$. Isso implica que $\mathbb{E}[Z] = 1$ e $\theta := Var[Z] = \dfrac{1}{\sigma}$, além de resultar na função $\mathscr{L}_Z(s) = (1 + \theta s)^{-\frac{1}{\theta}}$, correspondendo a transformada de Laplace da distribuição Gama (2.22).
- Já a distribuição IG, a identificabilidade do modelo prevalece ao tomar $\sigma = \mu^{\frac{1}{2}}$ implicando que $\mathbb{E}[Z] = 1$ e $\theta := Var[Z] = \dfrac{1}{2\sigma}$, além de obter $\mathscr{L}_Z(s) = \exp\left\{\dfrac{1}{\theta}\left(1 - \sqrt{1 + 2\theta s}\right)\right\}$ que corresponde a transformada de Laplace da distribuição IG (2.24).

Observação 2. Ao considerar a função de sobrevivência marginal (3.5) com $\gamma \leq 1$ são válidas as seguintes afirmações:

1. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{\frac{b\gamma(e^{at}-1)+1}{a}} - 1$, com $t > 0$, $b > 0$ e $a \in \mathbb{R}$ de forma que: $b \neq 1$ e $a \neq \gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

2. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = e^{\gamma t} - 1$, com $t > 0$, $b > 0$ e $a \in \mathbb{R}$ de forma que: $b = 1$ e $a = \gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

Devido às vantagens expressadas na literatura do uso do modelo MEP, principalmente devido ao fato deste modelo ser uma alternativa semiparamétrica às distribuições paramétricas, além de ser um recurso flexível para acomodar funções de taxa de falha com diversas formas, o modelo MEP passa a ser considerado como função de risco de base para o modelo (3.5) neste capítulo. Deste modo, assuma as hipóteses estabelecidas na Seção 2.4, suponha uma partição no eixo dos tempos $\tau = \{s_0, \ldots, s_J\}$ de forma que $0 = s_0 < s_1 < \ldots < s_J < \infty$, gerando J intervalos disjuntos, e tome $\lambda_j > 0$ $(j = 1, \ldots, J)$. Consequentemente, a função de sobrevivência marginal em relação ao tempo T com base segundo o modelo MEP é dada por

$$S_D(t) = \exp\left\{-\frac{1}{\gamma}[(1 + H_{MEP}(t))^{\gamma} - 1]\right\}; \quad t > 0 \quad \text{e} \quad \gamma \leq 1, \tag{3.8}$$

em que $H_{MEP}(\cdot)$ é a função de risco acumulado do modelo MEP definida em (2.30) para todo $t \in I_j$ intervalos disjuntos com $\lambda_j > 0$. Note que a função de sobrevivência (3.8) está bem definida, uma vez que: $S_D(0) = 1$ e, ao tomar $t \longrightarrow \infty$, segue que $S_D(t) = 0$, para $0 < \gamma \leq 1$ e $S_D(t) = p_0$, caso $\gamma < 0$ com p_0 definido em (3.6). Por consequência, observe que ainda são preservadas as características dos modelos defeituosos destacadas anteriormente. Devido a isso, a função de sobrevivência (3.8) é denominada "modelo PVF-MEP defeituoso", quando $\gamma < 0$, e "modelo PVF-MEP", quando $0 < \gamma \leq 1$.

A f.d associada ao modelo (3.8) é dada por

$$f_D(t) = h_{MEP}(t)\ (1 + H_{MEP}(t))^{\gamma - 1}\ \exp\left\{-\frac{1}{\gamma}[(1 + H_{MEP}(t))^{\gamma} - 1]\right\},$$

tal que $h_{MEP}(\cdot)$ é a função de risco do modelo MEP definida em (2.33) para todo $t \in I_j$ intervalos disjuntos com $\lambda_j > 0$. Já a função de risco associada a (3.8) é definida como

$$h_D(t) = h_{MEP}(t)\ (1 + H_{MEP}(t))^{\gamma - 1}. \tag{3.9}$$

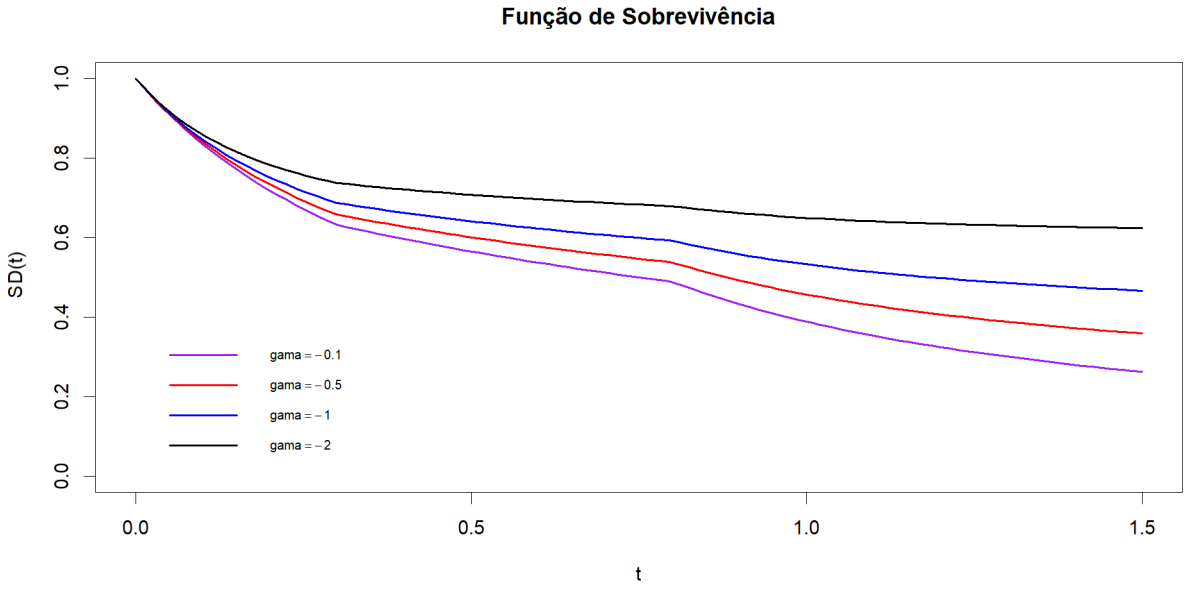
Por sua vez, a função de risco acumulado correspondente ao modelo (3.8) é dada por

$$H_D(t) = \frac{(1 + H_{MEP}(t))^{\gamma} - 1}{\gamma}, \tag{3.10}$$

no qual, ao assumir $\gamma < 0$ temos que $H_D(t) \longrightarrow -\log(p_0)$, quando $t \longrightarrow \infty$ e com p_0 definido por (3.6), implicando que a função (3.10) também é limitada por $-\log(p_0)$.

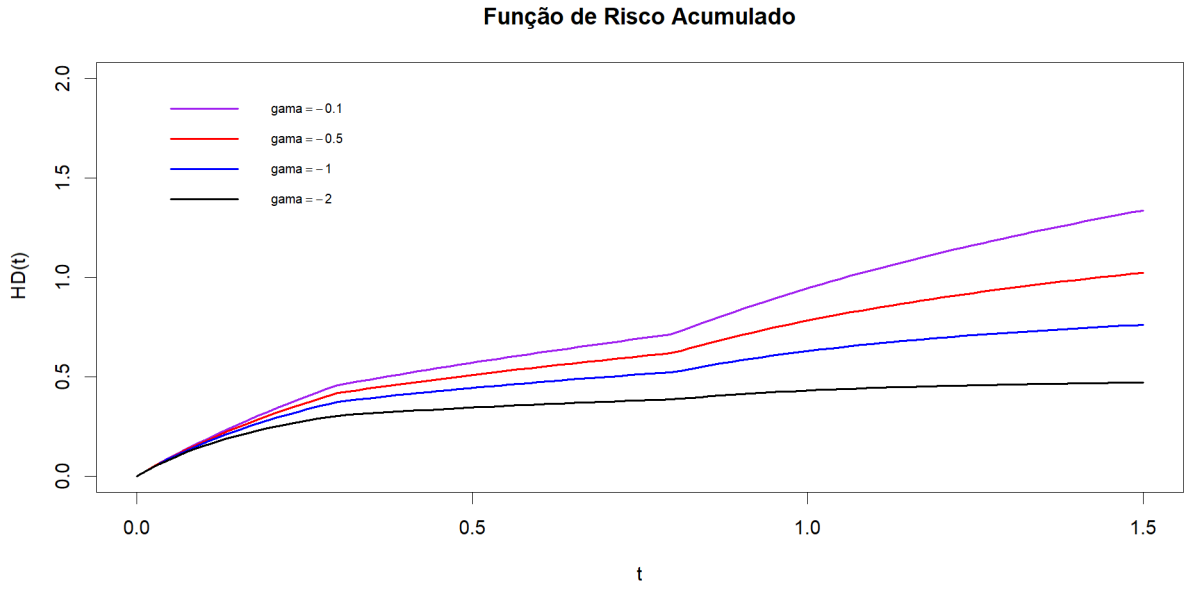
As Figuras 7, 8, 9 e 10 ilustram os gráficos das funções de sobrevivência e de risco acumulado para os modelos PVF-MEP defeituoso e PVF-MEP, respectivamente. Os mesmos foram gerados via o *software* R atráves do pacote **PWEXP** (TEAM, 2020) considerando três partições no eixo dos tempos (J = 3) com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$ com $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\gamma \leq 1$. Deste modo, as Figuras 7 e 8 ilustram os casos correspondentes ao modelo PVF-MEP defeituoso com $\gamma < 0$. Observe que, quando $t \longrightarrow \infty$, as curvas da função de sobrevivência se aproximarão do valor p_0 e as curvas da função de risco acumulado serão limitadas por $-log(p_0)$, sendo estas características marcantes de modelos defeituosos. Já as Figuras 9 e 10 exibem o comportamento das curvas do modelo PVF-MEP, quando $0 < \gamma \leq 1$, e enfatizam características de curvas que retratam modelos próprios. Note que com o aumento do tempo, as curvas da função de sobrevivência tendem a zero e as curvas da função de risco acumulado não serão limitadas.

Figura 7 – Gráfico da função sobrevivência do modelo PVF-MEP defeituoso considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



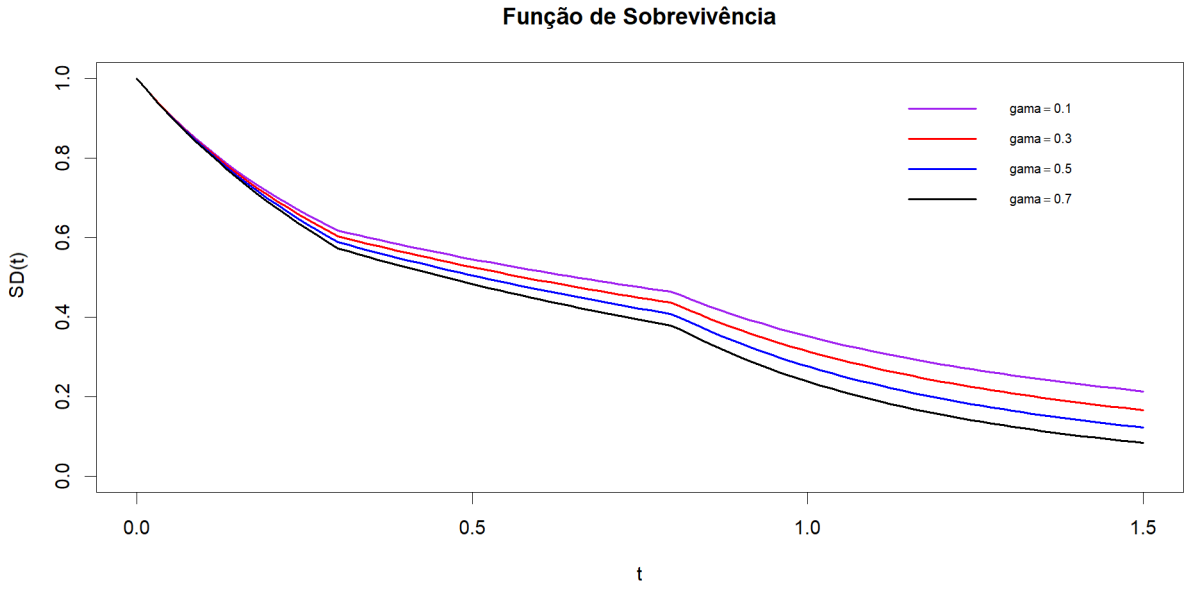
Fonte: Elaborada pelo autor.

Figura 8 – Gráfico da função de risco acumulado do modelo PVF-MEP defeituoso considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



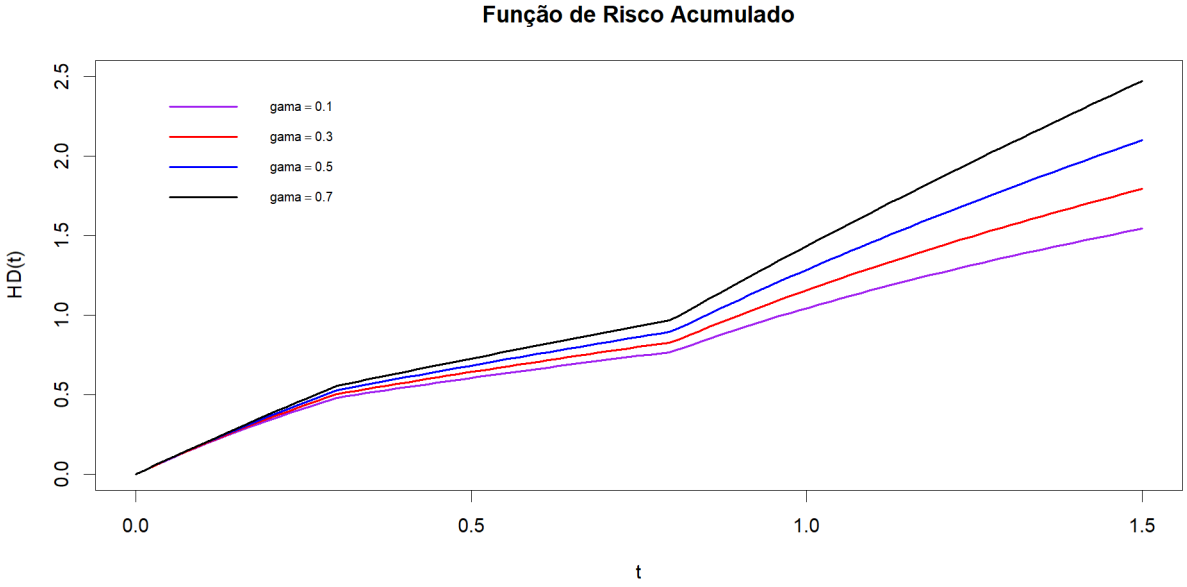
Fonte: Elaborada pelo autor.

Figura 9 – Gráfico da função sobrevivência do modelo PVF-MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



Fonte: Elaborada pelo autor.

Figura 10 – Gráfico da função de risco acumulado do modelo PVF-MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$ e $\lambda_3 = 3$.



Fonte: Elaborada pelo autor.

3.2.1 Propriedades matemáticas do modelo

Na literatura, autores como Cancho *et al.* (2021) destacam o uso de séries de potência aplicadas em modelos de sobrevivência, mostrando a eficácia desta abordagem em suas pesquisas. Para isso, considere a função de sobrevivência definida (3.5) com $0 < \gamma < 1$, caracterizando uma distribuição própria. Observe que ao analisar esta função é possível descrevê-la por meio a séries de potência atráves da seguinte relação matemática

$$[(1+H_0(t))^{\gamma} - 1] = \sum_{i=1}^{\infty} s_i H_0(t)^i, \tag{3.11}$$

com

$$s_1 = \gamma, \quad s_2 = \frac{-\gamma\,(1-\gamma)}{2}, \quad s_3 = \frac{\gamma\,(1-\gamma)\,(2-\gamma)}{6}, \quad \ldots, \quad s_i = \frac{(-1)^{2i-2}\,\gamma!}{i!}, \quad \ldots.$$

Assim, ao utilizar a relação matemática (3.11), a função de sobrevivência marginal para os indivíduos em risco definida em (3.5) é reescrita como

$$S_{P1}(t) = exp\left\{-\frac{1}{\gamma}\sum_{i=1}^{\infty} s_i\, H_0(t)^i\right\}; \quad t > 0 \quad \text{e} \quad 0 < \gamma < 1. \tag{3.12}$$

Por meio a manipulações algébricas, o somatório exibido na função de sobrevivência (3.12) pode ser caracterizado por intermédio dos Polinômios Exponenciais de Bell, definidos a seguir.

Definição 6. (CANCHO *et al.*, 2021) Os Polinômios Exponenciais de Bell são definidos por

$$\exp\left(u\sum_{i\geq 1}x_i\frac{z^i}{i!}\right)=\sum_{n,k\geq 0}B_{n,k}\,u^k\frac{z^n}{n!},$$

com

$$B_{n,k}=B_{n,k}(x_1,x_2,\ldots,x_{n-k+1})=\sum\frac{n!}{c_1!c_2!\ldots(1!)^{c_1}(2!)^{c_2}\ldots}x_1^{c_1}x_2^{c_2},\ldots,$$

sendo que o somatório é válido para $c_1,c_2,\ldots\geq 0$, tal que $c_1+c_2+c_3+\ldots=k$ e $c_1+2c_2+3c_3+\ldots=n$. Usando recursos computacionais temos que BellY$[n,k,\{1,\ldots,n-k+1\}]$ em MATHEMATICA e IncompleteBellB$(n,k,z[1],z[2],\ldots,z[n=k+1])$ no MAPLE. Assim, têm-se que $B_{0,0}=1$, $B_{n,0}=0$ (for $n\geq 1$) e $B_{0,k}=0$ (for $k\geq 1$).

Portanto, ao considerar a Definição 6, é possível reescrever a função de sobrevivência marginal definida em (3.12) por

$$S_{P1}(t)=\sum_{n=0}^{\infty}q_n\,\frac{H_0(t)^n}{n!},\tag{3.13}$$

com $q_n=\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n,k}$ tal que $B^{\star}_{n,k}=B_{n,k}(1!s_1,2!s_2,\ldots,(n-k+1)!s_{n-k+1})$ e $n,k\geq 0$. Assuma que $w_n=\dfrac{-q_{n+1}}{(n+1)!}$ e tome a função $h^*_{n+1}(t)=(n+1)\,H_0(t)^n\,h_0(t)$, definida segundo uma distribuição Exponenciada com parâmetro de potência $(n+1)$ para $n\geq 0$. Desta forma, ao derivar a função (3.13), é possível determinar a f.d da variável T, sendo dada por

$$f_{P1}(t)=\sum_{n=0}^{\infty}w_n\,h^*_{n+1}(t).\tag{3.14}$$

Na literatura, a distribuição Exponenciada-H_0 (exp $-H_0$) é empregada para determinar algumas propriedades matemáticas da distribuição de T, como por exemplo, o r-ésimo momento (CANCHO *et al.*, 2021). Para isso, considere a variável aleatória Y tal que $Y_0=T\sim H_0$ e $Y_{n+1}\sim \exp -H_0(n+1)$ com $n\geq 0$. Desta forma, o r-ésimo momento da variável T cuja f.d.p é dada por (3.14) é definido como

$$\begin{aligned}
\mu'_r&=\mathbb{E}\,(T^r)\\
&=\int_{-\infty}^{\infty}t^r\,f_{P1}(t)\,dt\\
&=\int_{-\infty}^{\infty}t^r\sum_{n=0}^{\infty}w_n\,h^*_{n+1}(t)\,dt\\
&=\int_{0}^{\infty}t^r\sum_{n=0}^{\infty}w_n\,(n+1)\,H_0(t)^n\,h_0(t)\,dt\\
&=\sum_{n=0}^{\infty}w_n\,(n+1)\int_{0}^{\infty}t^r\,H_0(t)^n\,h_0(t)\,dt.
\end{aligned}$$

Para ilustrar esta propriedade serão simuladas a média e a variância de dados aleatórios, assumindo diferentes valores para os parâmetros γ e λ. Deste modo, considere o modelo MEP com $J = 1$ (distribuição Exponencial) como função de base, com $H_0(t) = \lambda t$ e $h_0(t) = \lambda$. Assim

$$\mu'_r = \sum_{n=0}^{\infty} w_n\,(n+1)\int_0^{\infty} t^r\,H_0(t)^n\,h_0(t)\,dt = \sum_{n=0}^{\infty} w_n\,(n+1)\int_0^{\infty} t^r\,(\lambda t)^n\,\lambda\,dt.$$

Note que, por meio a algumas manipulações algébricas, segue que

$$\mu_r^{'} = \sum_{n=0}^{\infty} w_n(n+1)\frac{\lambda^{n+1}t^{r+n+1}}{r+n+1}; \quad \lambda > 0, \quad t,n \geq 0 \quad \text{e} \quad r+n \neq -1. \tag{3.15}$$

Os valores simulados ao assumir o vetor $t = (2,49; 1,80; 3,28; 2,08; 2,55; 2,52; 2,52)$ com $n = 7$ são exibidos na Tabela 1. Observe que para todas as combinações testadas foi possível estimar os valores da média e da variância, ilustrando a aplicabilidade do r-ésimo momento (3.15). Essa propriedade também se aplica ao considerar um vetor t contendo valores entre 0 e 1.

Tabela 1 – Estimativas da média e variância para diferentes valores de $0 < \gamma < 1$ e $\lambda > 0$.

	$\gamma = 0.2$		$\gamma = 0.5$		$\gamma = 0.8$	
	Média	Variância	Média	Variância	Média	Variância
$\lambda = 0,4$	1,781	4,920	0,002	0,007	0,000	0,000
$\lambda = 1$	2.730,698	7.518,652	4,313	11,897	0,165	0,461
$\lambda = 1,2$	11.741,491	32.328,844	18,567	51,125	0,726	2,001

Fonte: Elaborada pelo autor.

3.3 Modelo de regressão PVF-MEP

Nesta subseção será apresentado o modelo de regressão usado nas análises dos modelos PVF-MEP e PVF-MEP defeituoso, considerando o modelo de riscos proporcionais. Este modelo modela a função de risco, tratando-a como uma função promissora com duas componentes: uma destacando o risco dos indivíduos do estudo, enquanto a outra se relaciona com os efeitos mensurados das covariáveis. Para isso, sejam n indivíduos de forma que T_i $(i = 1,\ldots,n)$ são condicionalmente independentes e $t_i = min(T_i, \delta_i)$ representando as observações consideradas tal que δ_i é o tempo de censura com $\delta_i = 1$, se o i-ésimo indivíduo for falha e 0, caso contrário. Assuma o vetor paramétrico de covariáveis $\mathbf{x_i} = (x_{i1},\ldots,x_{ip})^T$ e seu respectivo vetor dos coeficientes de regressão $\beta = (\beta_1,\ldots,\beta_p)^T$. Sob esta parametrização, o modelo de riscos proporcionais que descreve o modelo PVF-MEP é dado por

$$h_R(t \mid \mathbf{x_i}) = h_D(t)\ \exp\{\beta^T\mathbf{x_i}\}, \tag{3.16}$$

em que $h_D(\cdot)$ é a função de risco do modelo PVF-MEP definida em (3.9). Ou ainda,

$$h_R(t \mid \mathbf{x_i}) = h_{MEP}(t)\ (1+H_{MEP}(t))^{\gamma-1}\ \exp\{\beta^T\mathbf{x_i}\}, \tag{3.17}$$

para todo $t \in I_j$ intervalos disjuntos com $\lambda_j > 0$. Já a função de sobrevivência associada ao modelo (3.16) é dada por

$$S_R(t \mid \mathbf{x_i}) = [S_D(t)]^{\exp\{\beta^T \mathbf{x_i}\}}, \tag{3.18}$$

em que $S_D(\cdot)$ é a função de sobrevivência do modelo PVF-MEP definida em (3.8). Como consequência, é possível reescrever a função (3.18) como

$$S_R(t \mid \mathbf{x_i}) = \left[\exp\left\{-\frac{1}{\gamma}[(1+H_{MEP}(t))^{\gamma}-1]\right\}\right]^{\exp\{\beta^T \mathbf{x_i}\}}.$$

Note que a função de sobrevivência (3.18) está bem definida, pois a função $S_D(\cdot)$ é bem definida. Ainda, observe que quando $\gamma < 0$ é possível determinar a proporção de indivíduos imunes ao evento de interesse, sendo dada por

$$p_{OR} = \lim_{t \longrightarrow \infty} S_R(t) = \left[\exp\left\{\frac{1}{\gamma}\right\}\right]^{\exp\{\beta^T \mathbf{x_i}\}} > 0,$$

novamente, ressaltando uma das características mais importantes dos modelos defeituosos. Em particular, para $\gamma < 0$, este modelo de regressão será denominado como modelo de regressão PVF-MEP defeituoso. Caso contrário, para $0 < \gamma \leq 1$, será chamado de modelo de regressão PVF-MEP.

Ademais, os coeficientes β's no modelo de regressão medem os efeitos das covariáveis sobre a taxa de falha, sendo que estas covariáveis podem acelerar ou desacelerar a função de risco. Assim, a estrutura de riscos proporcionais considera que a razão de risco entre dois indivíduos distintos K e W independente do tempo t (SANTO, 2022). Em particular, ao tomar a razão das funções de risco dadas em (3.17), para dois indivíduos distintos K e W, segue que

$$\frac{h_{RK}(t \mid \mathbf{x_i})}{h_{RW}(t \mid \mathbf{x_i})} = \frac{h_{MEP}(t)\ (1+H_{MEP}(t))^{\gamma-1}\ \exp\{\beta^T \mathbf{x_K}\}}{h_{MEP}(t)\ (1+H_{MEP}(t))^{\gamma-1}\ \exp\{\beta^T \mathbf{x_W}\}} = \frac{\exp\{\beta^T \mathbf{x_K}\}}{\exp\{\beta^T \mathbf{x_W}\}} = \exp\{\beta^T(\mathbf{x_K}-\mathbf{x_W})\},$$

em que $\exp\{\beta^T\}$ representa o efeito multiplicativo da diferença $(\mathbf{x_K}-\mathbf{x_W})$ no risco de morte.

3.4 Inferência bayesiana

Considere o modelo de regressão PVF-MEP e assuma n indivíduos de forma que T_i $(i = 1,\ldots,n)$ são condicionalmente independentes dado Z_i, onde Z_i representa as componentes de fragilidade admitindo uma distribuição PVF, segundo Hougaard (2000), com $\mathbb{E}(Z_i) = 1$ e $Var(Z_i) = 1-\gamma$. Suponha que a distribuição de T_i é caracterizada por um modelo básico de fragilidade com

$$h(T_i \mid Z_i) = Z_i h_0(T_i \mid \eta),$$

em que $h_0(T_i \mid \eta)$ é a função de risco de base para T_i com vetor de parâmetros η. Sejam as observações consideradas $t_i = min(T_i, \delta_i)$ tal que δ_i é o tempo de censura com $\delta_i = 1$, se o

i-ésimo indivíduo for falha e 0, caso contrário. Tome $\mathbf{t} = (t_1, \ldots, t_n)^T$, $\delta = (\delta_1, \ldots, \delta_n)^T$ e $\mathbf{X}$ representando a matriz de covariáveis de ordem (n ×p) cujo vetor de covariáveis de ordem (p ×1) para o i-ésimo indivíduo é dado por $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})^T$, com $\beta = (\beta_1, \ldots, \beta_p)^T$ sendo o seu respectivo vetor paramétrico dos coeficientes de regressão. Sob estas hipóteses, a função de verossimilhança relacionada ao modelo de regressão PVF-MEP para $\vartheta = (\gamma, \lambda_j, \beta)$ dado o vetor D é dada por

$$\begin{aligned} L(\vartheta \mid D) = \prod_{i=1}^{n}\prod_{j=1}^{J} \left[h_{MEP}(t_i \mid \lambda_j)\ (1 + H_{MEP}(t_i \mid \lambda_j))^{\gamma-1} \exp\{\beta^T \mathbf{x_i}\}\right]^{\delta_i v_{ij}} \\ \times \left[\left(\exp\left\{-\frac{1}{\gamma}\left[(1 + H_{MEP}(t_i \mid \lambda_j))^{\gamma} - 1\right]\right\}\right)^{\exp\{\beta^T \mathbf{x_i}\}}\right]^{v_{ij}}, \end{aligned} \tag{3.19}$$

com $D = (n, \mathbf{t}, \delta, \mathbf{X}, \mathbf{v})$ de forma que $\mathbf{v} = (v_{11}, \ldots, v_{nJ})^T$, no qual, para $j = 1, \ldots, J$ intervalos disjuntos temos que $v_{ij} = 1$, se $s_{j-1} < t_i \leq s_j$ e $v_{ij} = 0$, caso contrário e $\lambda_j > 0$. Este indicador v_{ij} é usado para definir a função de verossimilhança sobre cada um dos J intervalos.

Para a realização da investigação bayesiana, é tomado que os parâmetros γ, β e $\lambda_j's$ são independentes de forma que a densidade a priori conjunta é definida por

$$\pi\ (\gamma, \lambda_j, \beta) = \pi(\gamma)\ \pi(\lambda_j)\ \pi(\beta), \tag{3.20}$$

em que $\gamma \sim N(0,1)$ truncada em 1 e $\beta \sim N(\mu_\beta, \sigma_\beta^2)$. No entanto, os parâmetros $\lambda_j's$ são modelados segundo suas correlações em intervalos adjacentes, definidos por Gamerman (1991), com: $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j | \varepsilon_{j-1} \sim N(\varepsilon_{k-1}, \tau)$, para $j = 1, \ldots, J$, $k = 1, \ldots, p$ e $\varepsilon_0 = 0$.

Deste modo, ao combinar a densidade a priori (3.20) com a função de verossimilhança (3.19) segue que a densidade a posteriori conjunta de $\gamma, \lambda_j's$ e β é dada por

$$\pi\ (\vartheta \mid D) \propto L(\vartheta \mid D)\ \pi(\gamma)\ \pi(\lambda_j)\ \pi(\beta). \tag{3.21}$$

Note que como a densidade a posteriori (3.21) não é uma densidade padrão, sendo analiticamente intratável, a inferência realizada é baseada no método HMC. De acordo com Hartmann (2015), uma das principais vantagens do HMC é sua eficiência em simular valores aleatórios de uma distribuição alvo, percorrendo rapidamente seu suporte sem a necessidade de distribuições auxiliares, o que o torna mais direto e sistemático em comparação com outros métodos de simulação. Já os cálculos deste estudo são implementados com o R-Stan, no qual, autores como Jiang e Carter (2019) e Ng'ombe e Lambert (2021) ressaltam que o Stan oferece uma linguagem probabilística expressiva para especificar modelos estatísticos, além de possuir um código aberto para a realização da inferência bayesiana, de fazer o ajuste de modelos grandes e complexos e de calcular diretamente a densidade log-posteriori, permitindo o uso de distribuições próprias e impróprias. Por sua vez, autores como McElreath (2020), Tian *et al.* (2018) e Ng'ombe e Lambert (2021) destacam que o uso do HMC por meio ao Stan melhora a velocidade, estabilidade e escalabilidade dos modelos em comparação com os métodos MCMC e

o amostrador Gibbs, além de ser um método de computação avançado, sendo usado para calcular as distribuições desejadas de forma mais eficiente, com estimativas de simulações precisas e resultados consistentes. Para finalizar, os critérios de comparação usados nas próximas análises deste capítulo estão descritos nas Subseções 2.5.4.1, 2.5.4.2 e 2.5.5.

3.4.1 Estudo de simulação

O estudo de simulação é feito para avaliar algumas propriedades frequentistas dos estimadores de Bayes com base na média, desvio padrão (DP), viés e a raiz do erro quadrático médio (REQM). Para isso, foram realizadas $M = 1.000$ simulações para cada configuração paramétrica tendo tamanhos de amostras diferentes com $n = (200; 500; 800; 1.000)$. Os tempos censurados δ_i são amostrados segundo uma distribuição Uniforme, na qual, a porcentagem média de observações censuradas varia de 7% a 20%. A componente de fragilidade é gerada a partir de uma distribuição PVF, conforme descrito na Seção 3.2, para $\gamma = -0,5$ e $\gamma = 0,5$. Assim, é adotado o modelo PVF-MEP com $J = 1$, denominado como "modelo PVF-Exponencial", com $\log(\lambda) = 2$.

Para a adesão de covariáveis são consideradas duas covariáveis, sendo x_{i1} gerada por uma distribuição Binomial com probabilidade de sucesso $0,5$ e x_{i2} gerada mediante a uma distribuição Qui-Quadrada com média $\frac{5}{2}$ e variância $\frac{1}{2}$, ambas covariáveis tendendo a 1. Além disso, são assumidas prioris independentes para os parâmetros do modelo tal que para os λ_j's é tomado que $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j|\varepsilon_{j-1} \sim N(\varepsilon_{k-1}, 1)$ com $j = 1, \ldots, J$, $k = 1, \ldots, p$ e $\varepsilon_0 = 0$. Ainda, para

- Modelo PVF-Exponencial com $\gamma = 0,5$: Considera-se $\gamma \sim N\ (0,1)$ truncada em 1 e $\beta_i \sim N\ (0,100)$ em que $i = 1,2$.

- Modelo PVF-Exponencial defeituoso com $\gamma = -0,5$: Suponha que $\gamma \sim N\ (0,1)$ truncada em 0 e $\beta_i \sim N\ (0,10)$ em que $i = 1,2$.

Note que devido a parametrização definida para o modelo de Hougaard (2000) é possível estabelecer as restrições ao parâmetro γ. Este recurso foi usado durante o truncamento das prioris de γ para garantir a convergência dos modelos PVF-Exponencial e PVF-Exponencial defeituoso. Os resultados deste estudo de simulação são mostrados na Tabela 2. Estes revelam que à medida que o tamanho da amostra aumenta, as estimativas tendem aos valores verdadeiros dos parâmetros em média. Ainda, devido a este aumento amostral, o módulo do viés, DP e a REQM tendem a zero. Estes são aspectos esperados quando o esquema de estimativa está funcionando corretamente.

Tabela 2 – Estimativas da média, DP, viés e REQM dos parâmetros dos modelos PVF-Exponencial e PVF-Exponencial defeituoso com a presença de covariáveis para $\gamma = -0,5$, $\gamma = 0,5$, $\log(\lambda) = 2$ e $\beta_1 = \beta_2 = 1$.

		$\gamma = -0,5$				$\gamma = 0,5$			
n	Parâmetro	Média	DP	Viés	REQM	Média	DP	Viés	REQM
200	γ	-0,487	0,007	0,013	0,015	0,513	0,013	0,013	0,018
	$\log(\lambda)$	2,002	0,034	0,002	0,034	2,015	0,063	0,015	0,064
	β_1	0,996	0,176	-0,004	0,176	1,045	0,397	0,045	0,398
	β_2	0,971	0,169	-0,029	0,171	1,001	0,407	0,001	0,406
500	γ	-0,488	0,006	0,012	0,014	0,510	0,012	0,012	0,017
	$\log(\lambda)$	2,001	0,033	0,001	0,033	2,009	0,063	0,009	0,063
	β_1	0,996	0,166	-0,004	0,166	0,967	0,386	-0,033	0,401
	β_2	0,949	0,182	-0,051	0,188	0,980	0,401	-0,020	0,387
800	γ	-0,488	0,006	0,011	0,014	0,508	0,012	0,012	0,016
	$\log(\lambda)$	1,997	0,031	-0,003	0,031	2,010	0,064	0,010	0,064
	β_1	1,007	0,179	0,007	0,178	0,958	0,350	0,017	0,389
	β_2	0,969	0,176	-0,031	0,179	1,017	0,393	-0,042	0,394
1.000	γ	-0,489	0,005	0,010	0,011	0,505	0,011	0,011	0,015
	$\log(\lambda)$	1,999	0,031	-0,001	0,031	2,009	0,062	0,009	0,062
	β_1	1,001	0,175	0,001	0,174	0,989	0,346	-0,011	0,346
	β_2	0,975	0,165	-0,029	0,172	0,993	0,306	-0,037	0,310

Fonte: Elaborada pelo autor.

3.5 Aplicações

3.5.1 Aplicação: dados AIDS/HIV

3.5.1.1 Descrição dos dados

A AIDS é uma doença que interfere na capacidade do organismo do paciente de combater infecções, sendo causada pelo HIV. O vírus do HIV pode ser transmitido pelo contato com sangue infectado, sêmen ou fluidos vaginais, no qual, após várias semanas o indivíduo infectado pode apresentar sintomas semelhantes aos de uma gripe. Consequentemente, esse tipo de infecção tende a ser considerada assintomática até evoluir para AIDS. Dentre os sintomas da AIDS estão a perda de peso, febre, sudorese noturna, fadiga e infecções recorrentes. Infelizmente, ainda não há cura para a AIDS, mas a adesão estrita aos antirretrovirais (ARVs) pode retardar significativamente o progresso da doença e prevenir infecções secundárias e complicações.

No Brasil, o primeiro caso de AIDS foi diagnosticado na década de 1980 e desde então foram desenvolvidos sistemas de informação que permitem conhecer o perfil epidemiológico da doença. Além disso, a propagação da infecção pelo HIV no país passou por múltiplas dimensões com diversas características epidemiológicas abrangendo todas as cinco regiões do país. Entre os estudos feitos referente as dimensões provocadas por esta doença é concluído que a região Sul possui a maior taxa de detecção do HIV, tendo o estado do Paraná como o estado que apresenta a menor taxa de detecção desta região. Em suma, as notificações compulsórias e universais dos

casos de HIV, bem como as análises provenientes na literatura, tem permitido conhecer melhor a magnitude da infecção e caracterizar o perfil epidemiológico, os riscos, as vulnerabilidades e as tendências da população infectada, permitindo a formulação de políticas públicas para enfrentar esta epidemia.

Nesse contexto, o banco de dados utilizado representa um estudo de coorte de pacientes com diagnóstico de AIDS/HIV no período de 2002 a 2006 no Paraná, totalizando 272 observações. Assim, este estudo irá analisar os tempos de sobrevida destes pacientes relacionando-os com diversos fatores, como por exemplo, idade, raça, infecções, dentre outros. Uma análise mais detalhada destes dados é encontrada em Cancho *et al.* (2021), contudo a Tabela 3 apresenta as medidas descritivas do tempo em anos até a morte do paciente por AIDS/HIV.

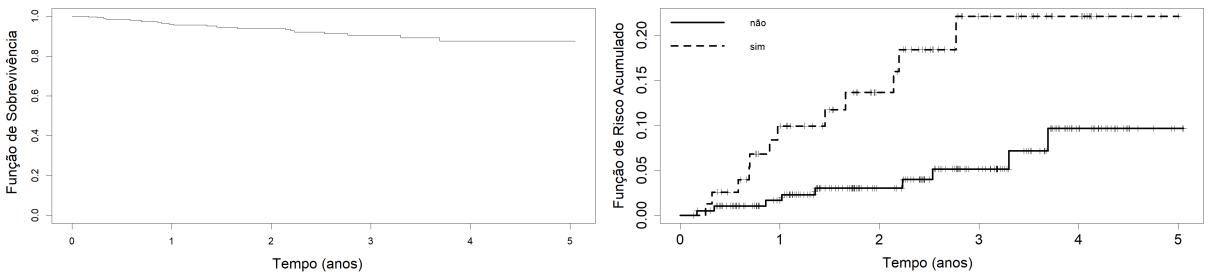
Tabela 3 – Medidas descritivas do tempo em anos até a morte do paciente por AIDS/HIV.

Mínimo	1 Quantil	Mediana	Média	3 Quantil	Máximo
0,13	1,18	2,29	2,32	3,40	5,05

Fonte: Elaborada pelo autor.

Observe que no painel esquerdo da Figura 11 é exibida a estimativa de Kaplan-Meier da função de sobrevivência para estes dados. Note que há a existência de uma apreciável proporção de indivíduos com risco zero, pelo menos no que concerne ao intervalo de tempo abrangido pelo estudo. Já no painel direito desta figura é mostrada a estimativa de risco acumulado para pacientes que desenvolveram infecção oportunista prévia (*prior opportunistic infection* (POI)). Assim, é evidenciado que o risco de morte é maior em pacientes que possuem esta infecção, além de retratar que os pacientes destas duas categorias (com e sem POI) provavelmente terão riscos de morte diferentes. Somado a isso, as funções de risco acumulativo são limitadas, indicando que estes dados podem ser modelados mediante modelos defeituosos.

Figura 11 – Estimativa de Kaplan-Meier da função de sobrevivência para os dados de AIDS/HIV (painel esquerdo) e função de risco acumulado estratificada para pacientes com ou sem POI (painel direito).



Fonte: Elaborada pelo autor.

3.5.1.2 Análise dos dados

Neste estudo foram observadas as seguintes variáveis: y_i: tempo até a morte por AIDS/HIV ou tempo censurado (em anos), x_{i_1}: faixa etária (16 a 65 anos) e x_{i_2}: POI na inclusão no estudo (não=0: sem infecção; sim=1: com infecção) para $i = 1,\ldots,272$. Além disso, este conjunto de dados possui taxa de censura de, aproximadamente, 8%. Desta forma, para estes dados, é ajustado o modelo PVF-MEP defeituoso descrito na Seção 3.2 com todas as covariáveis de forma que

$$\log(\exp\{\mathbf{x}_\mathbf{i}'\beta\}) = \beta_{idade}x_{i1} + \beta_{POI}x_{i2_1}; \quad i = 1,\ldots,272,$$

em que os efeitos das covariáveis são definidos por meio de variáveis fictícias, sendo que x_1 corresponde a idade dos pacientes em anos e:

$$x_{2_1} = \begin{cases} 1, & \text{se há a presença de POI;} \\ 0, & \text{caso contrário.} \end{cases}$$

Para o modelo proposto são adotadas a densidade a posteriori (3.21) e as seguintes priores independentes para os cálculos bayesianos: $\gamma \sim N\,(0,1)$ truncada em 0, $\beta_i \sim N\,(0,10)$ com $i = 1,\ldots,272$ e para parâmetros λ_j's é tomado que $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j|\varepsilon_{j-1} \sim N(\varepsilon_{k-1},1)$ tal que $\varepsilon_0 = 0$, $k = 1,\ldots,p$ e $j = 1,\ldots,J$ intervalos disjuntos. A escolha destas prioris é enfatizada na literatura nos trabalhos de Cancho *et al.* (2018) e Sibim (2011) que usam do modelo MEP para promover análises mediante a aplicação destas ou de prioris semelhantes em dados de sobrevivência. Ainda, os cálculos bayesianos foram baseados em amostras HMC obtidas de quatro cadeias independentes com 4.000 observações para cada parâmetro. Para eliminar o efeito dos valores iniciais, as primeiras 2.000 iterações foram desconsideradas. O modelo MEP foi investigado usando $J = 1,\ldots,5$ intervalos disjuntos na partição do eixo dos tempos, consequentemente, as estatísticas do LPML para os dados de HIV/AIDS são relatados na Tabela 4. Estas estatísticas são utilizadas para determinar uma partição J apropriada do eixo dos tempos. Observe que, com base nas estatísticas do LPML, o modelo com $J = 1$ (modelo PVF-Exponencial defeituoso) é considerado o modelo de melhor ajuste ao dados de AIDS/HIV.

Tabela 4 – Estatísticas do LPML para os dados de HIV/AIDS.

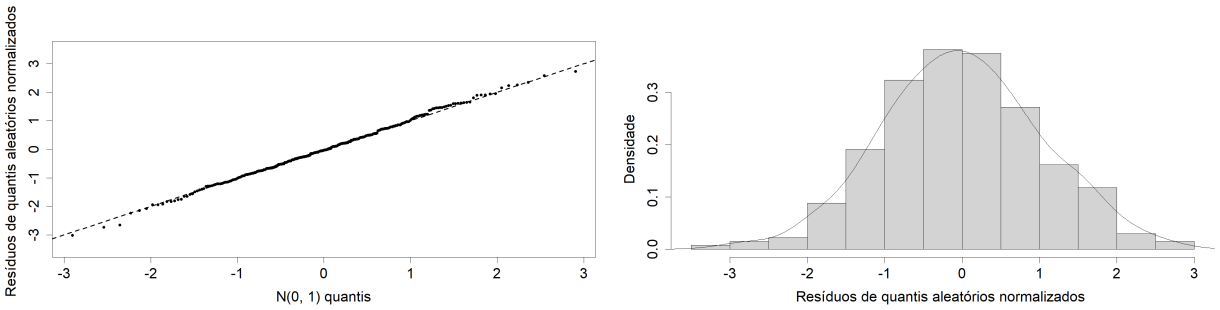
J	1	2	3	4	5
LPML	**-89,92**	-92,90	-94,85	-94,80	-96,67

Fonte: Elaborada pelo autor.

Para avaliar a adequação do ajuste do modelo PVF-Exponencial defeituoso são obtidos os resíduos de quantis aleatórios normalizados a posteriores, sendo este um recurso empregado no trabalho de Rigby e Stasinopoulos (2005). O gráfico QQ-plot dos resíduos destes quantis aleatórios são exibidos na Figura 12 e sugerem que o modelo determinado possui um ajuste aceitável ao dados de AIDS/HIV. Já a convergência das cadeias bayesianas são monitoradas pelos

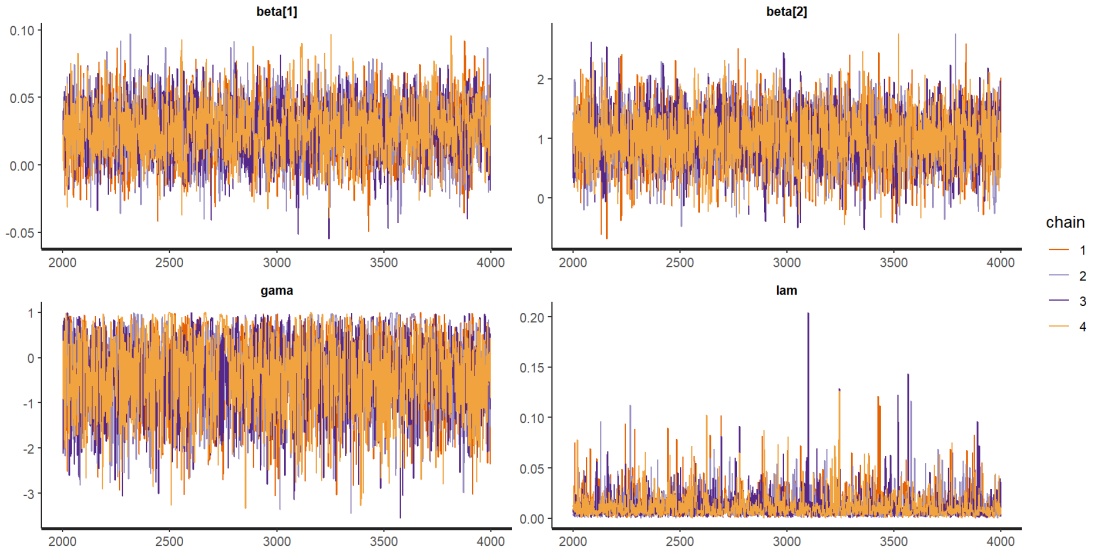
métodos de Cowles e Carlin (1996) por meio aos gráficos de rastreamento para os parâmetros do modelo, sendo ilustrado na Figura 13. Note que o comportamento dos gráficos indicam a convergência das cadeias.

Figura 12 – Gráfico QQ-plot dos resíduos de quantis normalizados a posteriores com linha de identidade (a esquerda) e histograma do modelo PVF-Exponencial defeituoso sob as hipóteses definidas (a direita).



Fonte: Elaborada pelo autor.

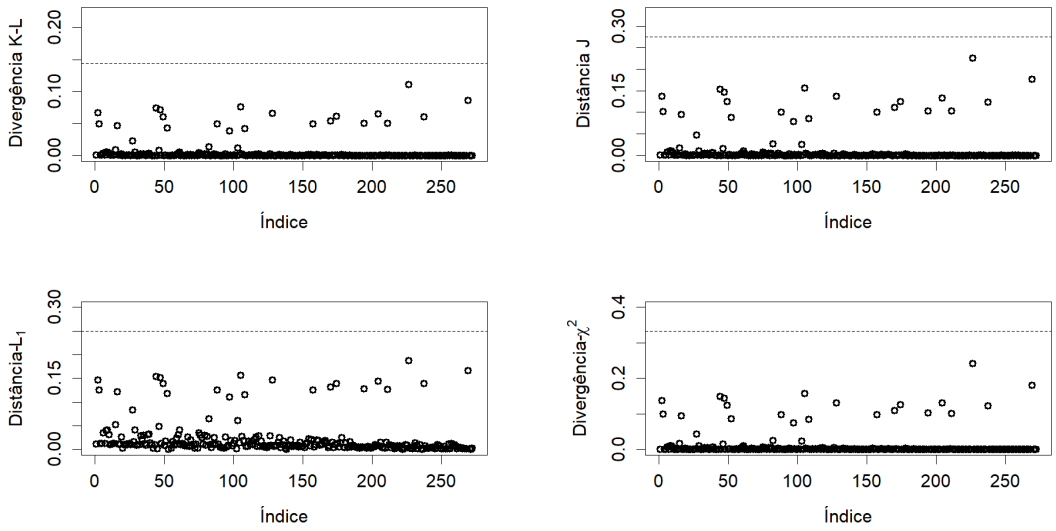
Figura 13 – Gráficos de rastreamento para parâmetros do modelo PVF-Exponencial defeituoso referente aos dados de HIV/AIDS.



Fonte: Elaborada pelo autor.

Considerando as amostras das distribuições a posteriores dos parâmetros foram calculadas as medidas de ϕ-divergência descritas na Subseção 2.5.5. A Figura 14 mostra o gráfico do índice das quatro medidas de divergência ϕ, em que é notado que não há possíveis observações influentes na distribuição a posteriori dos parâmetros do modelo de regressão, consequentemente, não haverá mudanças inferenciais a serem removidas nas observações.

Figura 14 – Gráfico de índice das medidas de divergência ϕ relacionadas aos dados da AIDS/HIV.



Fonte: Elaborada pelo autor.

As estimativas bayesianas paramétricas, bem como, as médias a posteriores, medianas, desvios padrões e 95% de confiança dos intervalos HPD (highest probability density (HPD)) para o modelo PVF-Exponencial defeituoso são exibidas na Tabela 5. Observe que é possível determinar a proporção de pacientes curados/imunes ao evento de interesse, sendo de $p_{OR} = 0,423$, além de investigar o impacto das covariáveis sobre o risco de morte dos pacientes. Note que a estimativa de Bayes de $\beta_{idade} > 0$ implicando que a cada ano adicional do paciente haverá um aumento em seu risco de morte, correspondendo a $\exp\{0,026\} = 1,02$ vezes. Além disso, como $\beta_{POI} > 0$ temos que o risco de morte de um paciente que possui POI é $\exp\{0,984\} = 2,67$ vezes maior ao se comparar com os pacientes que não possuem esta infecção.

Tabela 5 – Estatísticas LPML, médias a posteriores, medianas, desvios padrões e 95% de confiança dos intervalos HPD para os parâmetros do modelo PVF-Exponencial defeituoso relacionado aos dados de AIDS/HIV.

LPML	Parâmetros	Média	Mediana	DP	L	U
-89,92	γ	-0,321	-0,230	0,809	-1,801	0,999
	$\log\lambda$	0,012	0,008	0,012	0,000	0,035
	β_{idade}	0,026	0,026	0,021	-0,015	0,066
	β_{POI}	0,984	0,979	0,455	0,087	1,875

Fonte: Elaborada pelo autor.

Para analisar a proporção de indivíduos imunes ao evento de interesse foram considerados quatro pacientes hipotéticos A, B, C e D. Estes pacientes são caracterizados com valores diferenciados para a covariável idade, contudo é assumido que todos possuem a existência de POI. Consequentemente, as estimativas bayesianas paramétricas, bem como, as médias a

posteriores, medianas, DP's e 95% de confiança dos intervalos HPD para estes indivíduos são retratados na Tabela 6. Note que, por exemplo, ao comparar o paciente A com 16 anos e o paciente D com 52 anos, é visto que estes possuem probabilidades de não morrer diferentes e distantes, sendo de 0,282 para o paciente A e 0,056 para o paciente D. Contudo, ao considerar os pacientes A e B, que descrevem pacientes com idades próximas como 16 e 24 anos, respectivamente, observa-se que mesmo tendo probabilidades de não morrer diferentes, correspondendo a 0,282 e 0,212, respectivamente, ambas estão próximas.

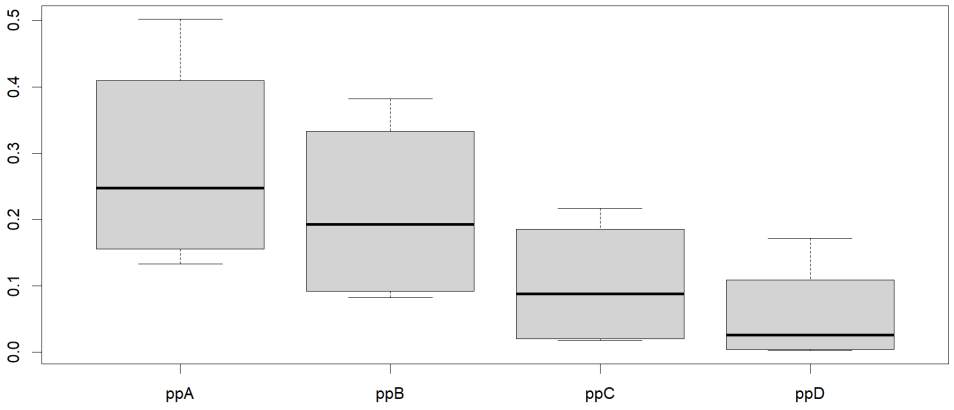
Tabela 6 – Estimativas de Bayes da probabilidade de não morrer e 95% de confiança dos intervalos HPD para os quatro hipopacientes terapêuticos com AIDS/HIV.

Pacientes	Idade	POI	Média	Mediana	DP	L	U
A	16	sim	0,282	0,247	0,166	0,132	0,502
B	24	sim	0,212	0,192	0,144	0,082	0,382
C	40	sim	0,102	0,087	0,098	0,017	0,216
D	52	sim	0,056	0,025	0,079	0,002	0,171

Fonte: Elaborada pelo autor.

A Figura 15 fornece os boxplots das médias a posteriores das proporções de indivíduos imunes ao evento de interesse para o modelo PVF-Exponencial defeituoso, relacionados aos pacientes hipotéticos descritos anteriormente. Desta forma, observa-se que a taxa mediana de cura para os pacientes A e B são próximas, sendo de aproximadamente 0,247 e 0,192, respectivamente. Porém, os pacientes A e D possuem um valor mais diferenciado para estas taxas correspondendo a aproximadamente 0,247 e 0,025, respectivamente. As faixas dos bigodes que representam os pacientes A, B, C e D são 0,282, 0,212, 0,102 e 0,056, respectivamente, indicando uma heterogênidade para a recuperação de cada paciente.

Figura 15 – Boxplot das médias posteriores para pacientes hipotéticos A, B, C e D segundo o modelo PVF-Exponencial defeituoso.



Fonte: Elaborada pelo autor.

Além disso, foram determinadas as estimativas bayesiana paramétricas, bem como, as médias a posteriores e 95% de confiança dos intervalos HPD para o modelo Gompertz defeituoso, sendo exibidas na Tabela 7. Para isso, foram consideradas a função de sobrevivência (3.2) e a f.d.p (3.1) com $a < 0$ durante a composição da função de verossimilhança do modelo Gompertz defeituoso, além de assumir as seguintes prioris independentes para os cálculos bayesianos: $a \sim N\ (0,1)$ truncada em 0, $b \sim N\ (0,1)$ e $\beta_i \sim N\ (0,100)$ com $i = 1,\ldots,272$. Os cálculos bayesianos foram baseados em amostras HMC obtidas de quatro cadeias independentes com 4.000 observações para cada parâmetro. Para eliminar o efeito dos valores iniciais, as primeiras 2.000 iterações foram desconsideradas. Note que as estimativas do modelo Gompertz defeituoso são similares ao modelo PVF-Exponencial defeituoso, enfatizando que ainda haverá um aumento do risco de morte para cada paciente a cada ano adicional, além do aumento deste risco para pacientes que possuem POI. Contudo, ao analisar as estatísticas do LPML é observado que, aparentemente, o modelo PVF-Exponencial defeituoso se destaca como o modelo de melhor ajuste para os dados de HIV/AIDS.

Tabela 7 – Estatísticas LPML, médias a posteriores e 95% de confiança dos intervalos HPD para os parâmetros nos dados de AIDS/HIV referente aos modelos PVF-Exponencial e Gompertz defeituosos.

PVF-Exponencial defeituoso			Gompertz defeituoso		
Parâmetros	Média	Intervalos HPD (95%)	Parâmetros	Média	Intervalos HPD (95%)
γ	-0,290	(-1,762; 0,995)	a	-0,215	(-0,496 ; 0,000)
$\log\lambda$	0,012	(0,001; 0,034)	b	0,016	(0,001 ; 0,045)
β_{idade}	0,026	(-0,016; 0,067)	β_{idade}	0,026	(-0,015 ; 0,065)
β_{POI}	0,977	(0,070; 1,850)	β_{POI}	0,986	(0,126 1,903)
LPML	**-89,92**		-90,65		

Fonte: Elaborada pelo autor.

Como as estimativas a posteriores dos modelos PVF-Exponencial e Gompertz defeituosos são bem próximas, é realizada a estimação da verossimilhança marginal por meio de *bridge sampling*, enfatizado na Subseção 2.5.4.1. Este método determina as probabilidades marginais dos modelos a posteriores comparando-os por meio das funções (2.38) e (2.39). Estas estimativas são apresentadas na Tabela 8. Portanto, é ressaltado que o modelo que melhor se ajusta aos dados de HIV/AIDS é realmente o modelo PVF-Exponencial defeituoso com uma probabilidade de 84%.

Tabela 8 – Probabilidades dos modelos PVF-Exponencial e Gompertz defeituosos a posterioris mediante as verossimilhanças marginais.

	PVF-Exponencial defeituoso	Gompertz defeituoso
Probabilidade	**0,84**	0,16

Fonte: Elaborada pelo autor.

3.5.2 Aplicação: dados sobre diarreia

3.5.2.1 Descrição dos dados

Este banco de dados retrata um ensaio comunitário randomizado realizado entre dezembro de 1990 a dezembro de 1991, sendo cedido pelo Instituto de Saúde Coletiva da Universidade Federal da Bahia e possuindo uma análise detalhada em Carvalho *et al.* (2011). Esta análise tem como objetivo proporcionar um estudo do efeito da suplementação de vitamina A na diarreia no nordeste do Brasil em uma coorte de crianças de 6 a 48 meses. O banco de dados consiste em 860 crianças, em que 426 pertencem ao grupo placebo e 434 ao grupo vitamina A, totalizando 5.592 observações, uma vez que, cada criança pode ter um ou mais episódios de diarreia ao longo de sua permanência no estudo. A Tabela 9 mostra as medidas descritivas do tempo em anos até o final do estudo, ou seja, período em que as crianças não apresentam casos de diarreia.

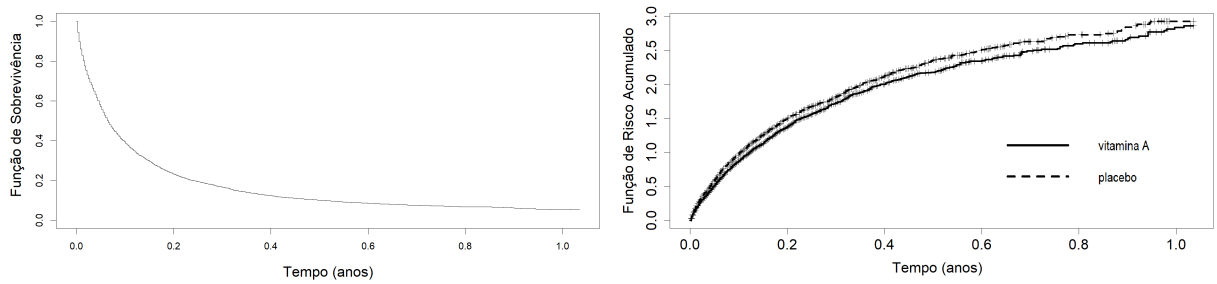
Tabela 9 – Medidas descritivas do tempo em anos até o final do estudo sobre diarreia.

Mínimo	1 Quantil	Mediana	Média	3 Quantil	Máximo
0,002	0,019	0,060	0,133	0,158	1,034

Fonte: Elaborada pelo autor.

Observe que no painel esquerdo da Figura 16 é exibida a estimativa de Kaplan-Meier da função de sobrevivência para estes dados. Note que não há a existência de uma proporção de indivíduos com risco zero, pois o estudo em questão é finalizado após um período sem diarreia. Já no painel direito desta figura é mostrada a estimativa do risco acumulado para pacientes que receberam vitamina A ou placebo como forma de tratamento. Como consequência, é evidenciado que o risco do paciente ter mais episódios de diarreia é maior em pacientes que recebem placebo, além de retratar que os pacientes destas duas categorias (vitamina A e placebo) provavelmente terão riscos diferentes.

Figura 16 – Estimativa de Kaplan-Meier da função de sobrevivência para os dados de diarreia (painel esquerdo) e função de risco acumulado estratificada para os pacientes que receberam vitamina A ou placebo como forma de tratamento (painel direito).



Fonte: Elaborada pelo autor.

3.5.2.2 Análise dos dados

Neste estudo foram observadas as seguintes variáveis: y_i: tempo sem diarreia ou tempo censurado (em anos), x_{i_1}: grupo (vit=0: vitamina A ; pla=1: placebo) e x_{i_2}: duração do episódio de diarreia anterior, para $i = 1,\ldots,5.592$. Além disso, este conjunto de dados possui uma taxa de censura de, aproximadamente, 15%. Deste modo, é ajustado o modelo PVF-MEP descrito na Seção 3.2 com todas as covariáveis de forma que

$$\log(\exp\{\mathbf{x_i'}\beta\}) = \beta_{pla}x_{i1_1} + \beta_{diar}x_{i2}; \quad i = 1,\ldots,5.592,$$

em que os efeitos das covariáveis são definidos por meio de variáveis fictícias, sendo que x_2 corresponde a duração do episódio de diarreia anterior e:

$$x_{1_1} = \begin{cases} 1, & \text{se o tratamento é o placebo;} \\ 0, & \text{caso contrário.} \end{cases}$$

Para o modelo proposto são adotadas a densidade a posteriori (3.21) e as seguintes priores independentes para os cálculos bayesianos: $\gamma \sim N\,(0,1)$ truncada em 0, $\beta_i \sim N\,(0,100)$ com $i = 0,\ldots,5.592$ e para parâmetros λ_j's é tomado que $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j|\varepsilon_{j-1} \sim N(\varepsilon_{k-1},1)$ com $\varepsilon_0 = 0$, $k = 1,\ldots,p$ e $j = 1,\ldots,J$ intervalos disjuntos. Os cálculos bayesianos foram baseados em amostras HMC obtidas de quatro cadeias independentes com 2.000 observações para cada parâmetro. Para eliminar o efeito dos valores iniciais, as primeiras 1.000 iterações foram desconsideradas. O modelo MEP foi investigado usando $J = 1,\ldots,5$ intervalos disjuntos na partição do eixo dos tempos, tendo as estatísticas do LPML para os dados de diarreia relatados na Tabela 10. Estas estatísticas são utilizadas para determinar uma partição J apropriada do eixo dos tempos. Observe que, com base nas estatísticas do LPML, o modelo com $J = 1$ (modelo PVF-Exponencial) é considerado o modelo de melhor ajuste para estes dados.

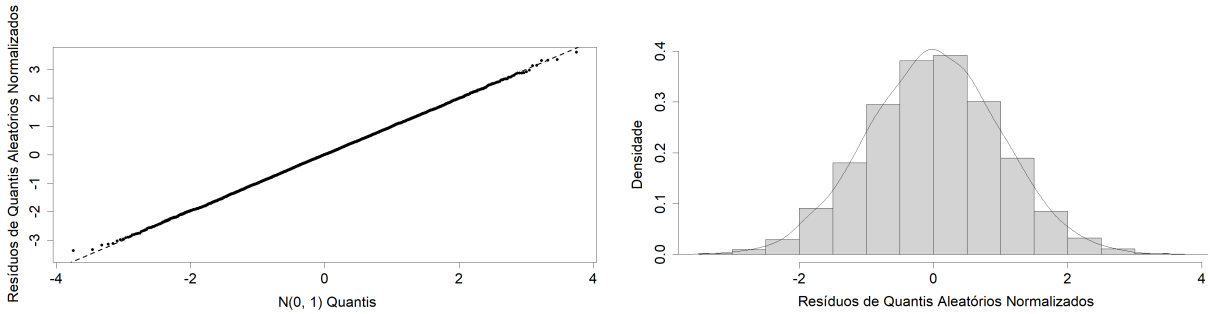
Tabela 10 – Estatísticas do LPML para os dados de diarreia.

J	1	2	3	4	5
LPML	**-5.023,08**	-5.029,78	-5.039,22	-5.048,08	-5.057,73

Fonte: Elaborada pelo autor.

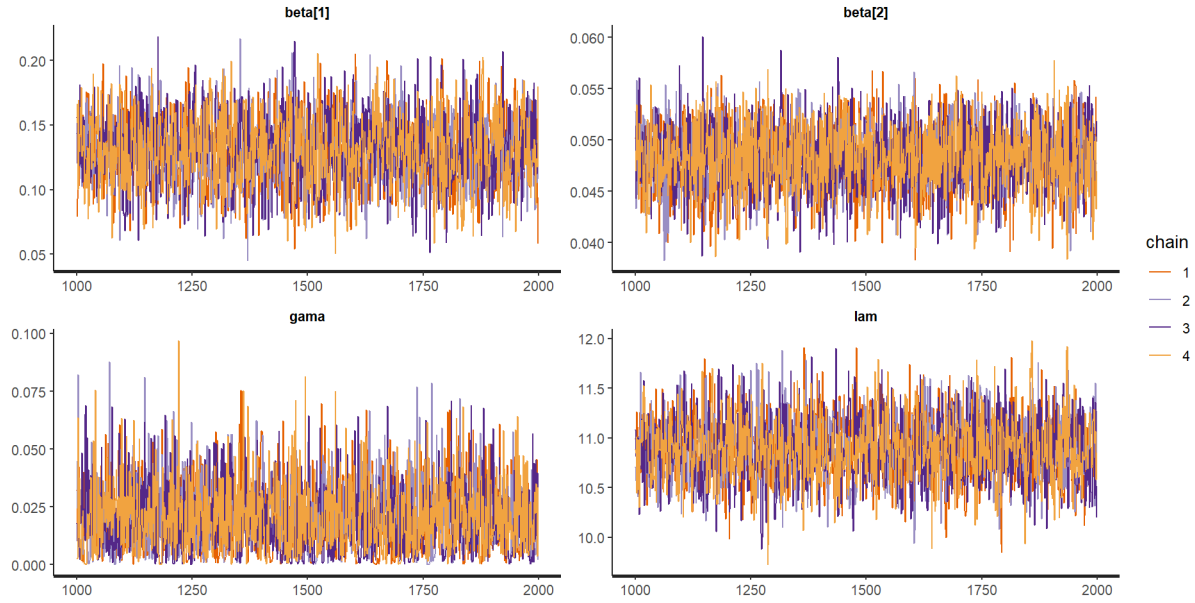
Para avaliar a adequação do ajuste do modelo PVF-Exponencial são obtidos os resíduos de quantis aleatórios normalizados a posteriores. O gráfico QQ-plot dos resíduos destes quantis aleatórios são exibidos na Figura 17 e sugerem que o modelo PVF-Exponencial possui um ajuste aceitável ao dados de diarreia. Já a convergência das cadeias bayesianas são monitoradas pelos métodos de Cowles e Carlin (1996) por meio aos gráficos de rastreamento para os parâmetros do modelo, sendo ilustrado na Figura 18. Logo, o comportamento gráfico indica a convergência das cadeias.

Figura 17 – Gráfico QQ-plot dos resíduos de quantis normalizados a posteriores com linha de identidade (a esquerda) e histograma do modelo PVF-Exponencial sob as hipóteses definidas (a direita).



Fonte: Elaborada pelo autor.

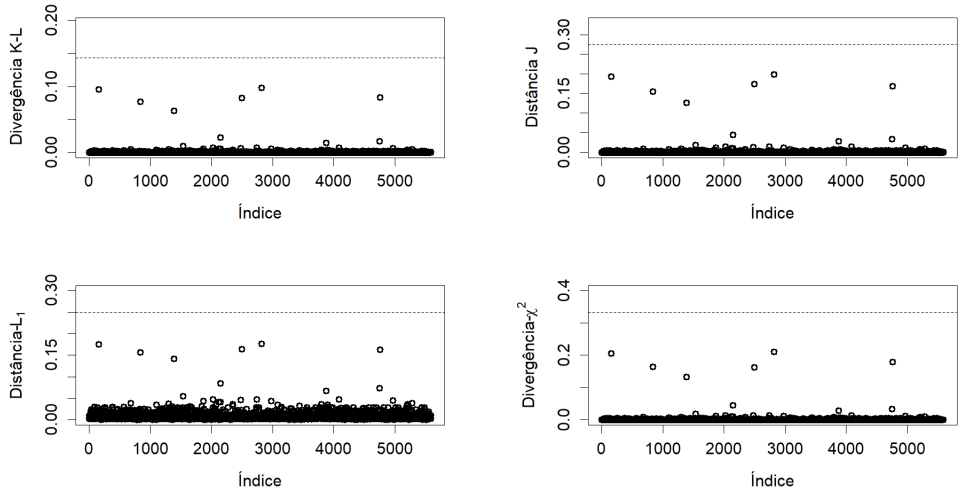
Figura 18 – Gráficos de rastreamento para parâmetros do modelo PVF-Exponencial referente aos dados de diarreia.



Fonte: Elaborada pelo autor.

Considerando as amostras das distribuições a posteriores dos parâmetros foram calculadas as medidas de ϕ-divergência descritas na Subseção 2.5.5. A Figura 19 mostra o gráfico do índice das quatro medidas de divergência ϕ, em que é notado que não há possíveis observações influentes na distribuição a posteriori dos parâmetros deste modelo de regressão. Consequentemente, não haverá mudanças inferenciais a serem removidas nas observações.

Figura 19 – Gráfico de índice das medidas de divergência ϕ relacionadas aos dados de diarreia.



Fonte: Elaborada pelo autor.

As estimativas bayesianas paramétricas, bem como, as médias a posteriores, medianas, desvios padrões e 95% de confiança dos intervalos HPD do modelo PVF-Exponencial são exibidas na Tabela 11. Note que a estimativa de Bayes de $\beta_{diar} > 0$ implicando que quanto maior for a duração do episódio de diarreia anterior, maior será o risco do paciente ter outros episódios de diarreia ($\exp\{0,048\} = 1,05$ vezes) durante o estudo. Além disso, como $\beta_{pla} > 0$ é relatado que pacientes que usam placebo também terão um maior risco de ter mais episódios ($\exp\{0.13\} = 1,14$ vezes) ao se comparar com pacientes que tomam vitamina A.

Tabela 11 – Estatísticas LPML, médias a posteriores, desvios padrão e 95% de confiança dos intervalos HPD para os parâmetros nos dados de diarreia.

LPML	Parâmetros	Média	Mediana	DP	L	U
-5.023,08	γ	0,020	0,017	0,015	0,000	0,048
	$\log\lambda$	10,888	10,881	0,318	10,294	11,536
	β_{pla}	0,130	0,130	0,026	0,076	0,178
	β_{diar}	0,048	0,048	0,003	0,042	0,054

Fonte: Elaborada pelo autor.

A Tabela 12 mostra as estimativas bayesiana paramétricas, as médias a posteriores e 95% de confiança dos intervalos HPD para os modelos Gompertz, IG e Gama. Em particular, para os modelos IG e Gama foram preservadas as hipóteses estipuladas anteriormente, contudo tomando $\gamma = 0,5$ (IG) e $\gamma = 0,00001$ (Gama). Por sua vez, para o modelo Gompertz foram consideradas a função de sobrevivência (3.2) e a f.d.p (3.1) com $a > 0$ durante a composição da função de verossimilhança, além de assumir as seguintes prioris independentes: $a \sim N(0,1)$ truncada em 1, $b \sim N(0,1)$ e $\beta_i \sim N(0,100)$ com $i = 1,\ldots,5.592$. Os cálculos bayesianos foram baseados em amostras HMC obtidas de quatro cadeias independentes com 2.000 observações

para cada parâmetro. Para eliminar o efeito dos valores iniciais, as primeiras 1.000 iterações foram desconsideradas. Note que as estimativas referentes aos parâmetros β's destes modelos são similares ao modelo PVF-Exponencial, destacando o aumento do risco de ter outros episódios de diarreia em pacientes que tomam placebo e que tiveram uma maior duração do episódio anterior desta doença. Portanto, ao analizar as estatísticas do LPML destes modelos, é observado que, aparentemente, o modelo de melhor ajuste para estes dados é o PVF-Exponencial.

Tabela 12 – Estatísticas LPML, médias a posteriores e 95% de confiança dos intervalos HPD para os parâmetros nos dados de diarreia, segundo os modelos Gompertz, IG e Gama.

Parâmetros	Gompertz		IG		Gama	
	Média	Intervalos HPD (95%)	Média	Intervalos HPD (95%)	Média	Intervalos HPD (95%)
γ	0,002	(0,000 ; 0,006)	—	—	—	—
$\log\lambda$	5,118	(4,893 ; 5,345)	8,293	(7,840 ; 8,697)	10,915	(10,261 ; 11,504)
β_{idade}	0,114	(0,058 ; 0,171)	0,026	(-0,025 ; 0,085)	0,139	(0,087 ; 0,191)
β_{POI}	0,065	(0,059 ; 0,070)	0,050	(0,044 ; 0,056)	0,049	(0,043 ; 0,055)
LPML	-5.217,96		-5.765,70		-5.023,93	

Fonte: Elaborada pelo autor.

A estimação da verossimilhança marginal é realizada para comparar os modelos PVF-Exponencial, Gompertz, IG e Gama mediante as funções (2.37) e (2.39). Estas estimativas são apresentadas na Tabela 13. Como consequência, é ressaltado que o modelo que melhor se ajusta aos dados de diarreia é o modelo PVF-Exponencial com uma probabilidade de $86,05\%$.

Tabela 13 – Probabilidades dos modelos PVF-Exponencial, Gompertz, IG e Gama a posteriori mediante as verossimilhanças marginais.

	PVF-Exponencial	Gompertz	IG	Gama
Probabilidade	**86,05**	0,01	0,08	13,86

Fonte: Elaborada pelo autor.

3.6 Conclusão

Neste capítulo é apresentado um modelo de fragilidade para modelar a heterogeneidade não observada nos dados de sobrevivência. Esse modelo é formado pela composição de uma distribuição de fragilidade baseada na família PVF e do modelo MEP, que atua como função de risco base, sendo estendido para permitir a construção de modelos defeituosos, resultando em uma fragilidade zero, conforme detalhado nas Seções 3.2 e 3.3.

A abordagem inferencial, discutida na Seção 3.4, utiliza do método HMC implementado no R-Stan, no qual, o estudo de simulação revela que o modelo possui boas propriedades dos estimadores de Bayes. Por fim, na Seção 3.5, a utilidade do modelo é validada por meio de aplicações em dados reais, o estudo de casos de AIDS/HIV e de diarreia, mostrando que o modelo é uma boa alternativa para a análise desses dados.

CAPÍTULO

4

MODELO DE LONGA DURAÇÃO INDUZIDO POR FRAGILIDADE

4.1 Introdução

O emprego de modelos de sobrevivência de longa duração vêm se destacando devido ao aumento da aplicação deste tema em estudos que envolvem as àreas de confiabilidade e sobrevivência, uma vez que, analisam as situações em que há a existência de unidades de amostragem que não são susceptíveis a ocorrência de um evento de interesse. Na literatura, a proporção destas unidades é denominada como taxa de cura (fração/proporção de curados, proporção de indivíduos imunes ao evento de interesse, proporção de sobreviventes de longo prazo, dentre outros). Em particular, autores como Cancho *et al.* (2023), Oliveira *et al.* (2023) e Brandão *et al.* (2023) estudam estes modelos ressaltando a sua abordagem em aplicações de dados de sobrevivência onde, por exemplo, os eventos de interesse consistem, respectivamente, em avaliar os efeitos das covariàveis na proporção de indivíduos imunes em um conjunto de dados de câncer cervical, em analisar a um conjunto de dados médicos reais obtidos de um estudo de coorte tendo como foco a avaliação das condições clínicas que afetam a vida de pacientes com diabetes e em aplicar dados médicos de um estudo sobre melanoma cutâneo.

Para estas situações, os modelos de sobrevivência induzidos pela fragilidade contínua não são apropriados durante o processo de modelagem. Logo, deve-se considerar modelos que possuem a fragilidade nula, além de descrever e analisar dados que contém unidades experimentais nas quais o evento de interesse não ocorre mesmo após um longo período de observação. Consequentemente, modelos que apresentam a fragilidade discreta são exemplos de modelos que se destacam por ter tais características, tornando-os adequados para a modelagem dos dados de sobrevivência que retratam estas situações.

O modelo de fragilidade discreto foi enfatizado neste trabalho na Subseção 2.2.2. Em

suma, é considerada a distribuição de probabilidade de Z sendo especificada por $\mathbb{P}[Z=z]=p_z$, com $\sum p_z=1$ e $z=0,1,2,\ldots$. Como consequência, é possível obter a função de sobrevivência não condicional (2.14) proposta por Tsodikov, Ibrahim e Yakovlev (2003) e que pode ser utilizada em estudos que abrangem tanto modelos próprios quanto impróprios (CANCHO *et al.*, 2021). Em particular, quando esta função de sobrevivência é imprópria, como mostrado em (2.15), haverá a existência de uma probabilidade $p_0>0$ que retrata que algum indivíduo é não suscetível ao evento de interesse mesmo após um período de tempo, ou seja, há uma proporção de indivíduos imunes ao evento de interesse do estudo. Um exemplo deste tipo de modelagem é ressaltado por Cancho *et al.* (2021) ao admitir uma variàvel de fragilidade Z segundo a distribuição de Poisson com média $\rho>0$, obtendo a f.g.p correspondente $\psi_Z(s)=\exp\{-\rho(1-s)\}$. Como resultado, a função de sobrevivência marginal para este modelo é dada por

$$S(t)=\exp\{-\rho F_0(t)\}, \tag{4.1}$$

em que $F_0(\cdot)$ é a função de distribuição acumulada base para o tempo T, quando Z é igual à constante um. Note que a proporção de indivíduos imunes ao evento de interesse correspondente ao modelo (4.1) é dada por

$$\lim_{t\longrightarrow\infty} S(t)=e^{-\rho}>0.$$

Cancho *et al.* (2021) destacam que a função (4.1) caracteriza o modelo *Bounded cumulative hazard* (BCH), definido por Tsodikov, Yakovlev e Asselain (1996), além de salientar que este modelo considera que os pacientes não imunes ao evento de interesse estarão expostos a Z fatores desconhecidos de riscos competitivos ou causas do evento de interesse com $\mathbb{E}(Z)=Var(Z)$. Contudo, afirmam que caso haja a superdispersão desses fatores haverá o aumento do risco de morte dos pacientes após um longo período de tratamento, havendo assim a necessidade de distribuições flexíveis para evitar esta situação (CANCHO *et al.*, 2021). Portanto, como recurso a esta situação, desenvolvem o modelo BCH-PVF que representa uma classe de modelos de sobrevivência induzidos por uma fragilidade discreta com distribuição de Poisson mista que pode explicar a dispersão não observada.

Neste contexto, o objetivo deste capítulo é exibir modelos de sobrevivência similares ao de Cancho *et al.* (2021), estendendo o modelo de fragilidade desenvolvido para permitir o uso de uma distribuição discreta, resultando em uma fragilidade zero que será analisada mediante a construção de modelos de longa duração. Assim, são apresentados modelos de sobrevivência induzidos por fragilidade discreta cuja distribuição de Poisson mista é usada para explicar a dispersão não observada e a componente de fragilidade segue uma distribuição PVF segundo Hougaard (2000). A abordagem inferencial é baseada em métodos bayesianos mediante o uso do método HMC implementado no R-Stan, no qual, resultados de simulação são fornecidos para avaliar o desempenho de algumas propriedades dos estimadores de Bayes. A importância deste modelo é ilustrada por meio de uma aplicação em um conjunto de dados reais. As demonstrações dos principais resultados deste capítulo estão descritas no Apêndice B.

4.2 Modelo de longa duração PVF

Nesta seção serão exibidos modelos de sobrevivência similares ao de Cancho *et al.* (2021), na qual, a distribuição de Poisson mista é usada para explicar a dispersão não observada do modelo construído e a sua componente de fragilidade segue uma distribuição PVF segundo Hougaard (2000). Para isso, seja $T > 0$ uma variável aleatória e não-negativa que representa o tempo até o acontecimento de um evento de interesse. Assuma que a distribuição de T é caracterizada por um modelo de fragilidade definido segundo a estrutura de riscos multiplicativos dada em (2.3) e que Z é uma variável aleatória discreta que denota o número não observado de causas de risco para o acontecimento do evento de interesse. Tome Z segundo uma distribuição de Poisson com média $\omega\rho$, em que $\rho > 0$ constante e ω é a componente de fragilidade variando ao longo da reta real positiva, sendo constante e com distribuição acumulada $G(\omega)$. Sob estas hipóteses, a distribuição marginal de Z é a classe de distribuições mistas de Poisson com probabilidade dada por

$$\mathbb{P}(Z = z) = \frac{1}{z!}\int_0^\infty \exp\{-\rho\omega\}(\rho\omega)^Z \, dG(\omega). \tag{4.2}$$

A f.g.p de Z é descrita mediante a equação (4.2) como

$$\psi_Z(s) = \int_0^\infty \exp\{-\rho\omega(1-s)\}dG(\omega) = \mathscr{L}_\omega(\rho(1-s)), \tag{4.3}$$

em que $\mathscr{L}_\omega(\cdot)$ é a transformada de Laplace da distribuição de ω. Assuma que a variável aleatória ω é definida por uma distribuição PVF, denotada como PVF(γ, μ, σ), com $\gamma \leq 1$, $\mu > 0$ e σ caracterizado para dois casos: se $0 < \gamma \leq 1$ é tomado que $\sigma \geq 0$, caso contrário, se $\gamma \leq 0$ é admitido que $\sigma > 0$ (HOUGAARD, 2000). Como consequência, sob esta parametrização, é definida a transformada de Laplace para o modelo PVF(γ, μ, σ) como

$$\mathscr{L}_\omega(s) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+s)^\gamma - \sigma^\gamma\right]\right\}. \tag{4.4}$$

Observe que, similar ao apresentado na Seção 3.2, por meio a algumas restrições paramétricas a função (4.4) corresponde as distribuições Gama ($\gamma \longrightarrow 0$), PE ($\gamma = \mu$ e $\sigma = 0$) e IG ($\gamma = 0,5$). Além disso, sob estas hipóteses, é possível reescrever a f.g.p de Z estabelecida pela função (4.3) como

$$\psi_Z(s) = \exp\left\{\frac{\mu}{\gamma}\left[\sigma^\gamma - (\sigma+\rho(1-s))^\gamma\right]\right\} = \mathscr{L}_\omega(\rho(1-s)). \tag{4.5}$$

Mediante a este cenário, considere que a variável de fragilidade Z da função (2.14) possua uma distribuição mista de Poisson com f.g.p definida por (4.5). Logo, a função de sobrevivência marginal em relação ao tempo T para este modelo é

$$S(t) = \exp\left\{\frac{\mu}{\gamma}\left[\sigma^\gamma - (\sigma+\rho F_0(t))^\gamma\right]\right\}; \quad \gamma \leq 1, \quad \mu > 0, \quad \sigma > 0 \quad \text{e} \quad \rho > 0, \tag{4.6}$$

em que $F_0(t) = 1 - \exp\left\{-\int_0^t h_0(u)du\right\}$ é a função de distribuição acumulada base para o tempo T, quando Z é a constante igual a um. Note que as funções de base $f_0(\cdot)$ e $F_0(\cdot)$ estão associadas a função de risco base $h_0(\cdot)$ quando Z é constante, implicando que todos os indivíduos em risco têm a mesma fragilidade constante.

Ainda, observe que quando $\sigma = 0$ e $\mu = \gamma = 1$, o modelo (4.6) se reduz ao modelo BCH introduzido por Tsodikov, Yakovlev e Asselain (1996). Também retrata os modelos de Cancho *et al.* (2021), pois descreve o modelo BCH-Gama, quando $\gamma \longrightarrow 0$ e $\mu = \sigma = \dfrac{1}{\varepsilon}$, e o modelo BCH-IG, ao assumir $\gamma = 0,5$, $\sigma = \dfrac{1}{2\varepsilon}$ e $\mu = \dfrac{\sqrt{2\varepsilon}}{2\varepsilon}$, ambos com $\varepsilon > 0$. A Tabela 14 fornece as funções de sobrevivência e de risco correspondente a estes três casos.

Tabela 14 – Funções de sobrevivência e de risco correspondente aos casos especiais do modelo (4.6).

Modelo	$S(t)$	$h(t)$
BCH $(\sigma = 0; \mu = \gamma = 1)$	$e^{-\rho F_0(t)}$	$\rho f_0(t)$
BCH-Gama $\left(\gamma \longrightarrow 0; \mu = \sigma = \frac{1}{\varepsilon}\right)$	$(1+\varepsilon\rho F_0(t))^{-\frac{1}{\varepsilon}}$	$\frac{\rho f_0(t)}{1+\varepsilon\rho F_0(t)}$
BCH-IG $\left(\gamma = 0,5; \sigma = \frac{1}{2\varepsilon}; \mu = \frac{\sqrt{2\varepsilon}}{2\varepsilon}\right)$	$e^{\frac{1}{\varepsilon}\left(1-\sqrt{1+2\varepsilon\rho F_0(t)}\right)}$	$\frac{\rho f_0(t)}{\sqrt{1+2\varepsilon\rho F_0(t)}}$

Fonte: Elaborada pelo autor.

Para evitar problemas de identificabilidade, é considerada uma restrição no modelo (4.6) com $\mu = \sigma = 1$ de forma que $\mathbb{E}[\omega] = 1$ e $Var(\omega) = 1 - \gamma$. Logo, $\mathbb{E}[Z] = \rho$ e $Var[Z] = \rho^2(1-\gamma)$. Como consequência, é possível reescrever a função de sobrevivência marginal em relação ao tempo T como

$$S(t) = \exp\left\{\frac{1}{\gamma}\left[1 - (1+\rho F_0(t))^{\gamma}\right]\right\}; \quad \gamma \leq 1 \quad \text{e} \quad \rho > 0, \tag{4.7}$$

em que $F_0(\cdot)$ é a função de distribuição acumulada base para o tempo T. Note que a função de sobrevivência (4.7) é um caso particular do modelo (4.6). Ainda, através desta função é possível determinar a proporção de indivíduos imunes ao evento de interesse dada por

$$p_0 = \lim_{t \longrightarrow \infty} S(t) = \exp\left\{\frac{1}{\gamma}\left[1 - (1+\rho)^{\gamma}\right]\right\} > 0, \tag{4.8}$$

pois $F_0(t) \longrightarrow 1$, quando $t \longrightarrow \infty$. Como consequência, o modelo (4.7) será denominado "modelo de longa duração PVF".

Já probabilidade de indivíduos imunes ao evento de interesse após $t > 0$ anos do período de acompanhamento é dada por

$$\mathbb{P}(t) = \mathbb{P}\left(Z = 0 \mid T > t\right) = \left[\frac{\exp\{(1+\rho F_0(t))^{\gamma}\}}{\exp\{(1+\rho)^{\gamma}\}}\right]^{\frac{1}{\gamma}},$$

em que $\mathbb{P}(0) = p_0$, caracterizando a proporção de indivíduos imunes ao evento de interesse após o término do período de observação.

Observação 3. A validade da função de sobrevivência (4.7) é comprovada ao considerar o seguinte teorema:

Teorema 2. (CANCHO *et al.*, 2021) Seja $h(t \mid Z) = Zh_0(t)$ o modelo de fragilidade básica tal que $h_0(t)$ é a função de risco base comum para todos os indivíduos e Z é a variável aleatória de fragilidade discreta possuindo uma distribuição Poisson com média $\eta\theta$, onde $\theta > 0$ é constante real e η é uma componente aleatória de fragilidade. Se a distribuição de η tem a transformada de Laplace $\mathscr{L}_\eta(s)$, então a função de sobrevivência não condicional é dada por

$$S(t) = \mathscr{L}_\eta(\theta F_0(t)),$$

com $F_0(t) = 1 - \exp\left\{-\int_0^t h_0(u)du\right\}$ representando a função de distribuição acumulada base para o tempo T.

A f.d associada ao modelo (4.7) é dada por

$$f(t) = \rho f_0(t)\,(1+\rho F_0(t))^{\gamma-1}\,\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\}, \tag{4.9}$$

em que $f_0(\cdot)$ é a função de distribuição base para o tempo T que está associada a função de risco base comum $h_0(\cdot)$. Por sua vez, a função de risco correspondente a (4.7) é definida como

$$h(t) = \rho f_0(t)\,(1+\rho F_0(t))^{\gamma-1}. \tag{4.10}$$

Já a função de risco acumulado associado a (4.7) é dada por

$$H(t) = \frac{(1+\rho F_0(t))^{\gamma}-1}{\gamma}. \tag{4.11}$$

Note que, ao tomar p_0 definido em (4.8), é constatado que $H(t) \longrightarrow -\log(p_0)$, quando $t \longrightarrow \infty$. Este fato implica que a função (4.11) é limitada por $-\log(p_0)$. Ainda, observe que os casos especiais destacados na Tabela 14 são preservados, pois ao admitir que $\gamma = 1$, $\gamma \longrightarrow 0$ e $\gamma = 0,5$, o modelo (4.7) se reduz aos modelos BCH, BCH-Gama com $\varepsilon = 1$ e BCH-IG com $\varepsilon = 0,5$, respectivamente. A Tabela 15 mostra as funções de sobrevivência e de risco, além da proporção de indivíduos imunes ao evento de interesse correspondente a estes casos.

Tabela 15 – Funções de sobrevivência e de risco e a proporção de indivíduos imunes ao evento de interesse correspondente aos casos especiais do modelo (4.7).

Modelo	$S(t)$	$h(t)$	p_0
BCH $(\gamma = 1)$	$e^{-\rho F_0(t)}$	$\rho f_0(t)$	$e^{-\rho}$
BCH-Gama com $\varepsilon = 1$ $(\gamma \longrightarrow 0)$	$(1+\rho F_0(t))^{-1}$	$\frac{\rho f_0(t)}{1+\rho F_0(t)}$	$(1+\rho)^{-1}$
BCH-IG com $\varepsilon = 0,5$ $(\gamma = 0,5)$	$e^{2(1-\sqrt{1+\rho F_0(t)})}$	$\frac{\rho f_0(t)}{\sqrt{1+\rho F_0(t)}}$	$e^{2(1-\sqrt{1+\rho})}$

Fonte: Elaborada pelo autor.

No trabalho de Cancho *et al.* (2021), o modelo proposto ressalta o uso da função de sobrevivência marginal tanto para casos próprios quanto impróprios. Deste modo, este trabalho propõe uma análise da função de sobrevivência (4.7), destacando o modelo próprio, similar ao que é apresentado por estes autores. Para isso, observe que é possível escrever a função de sobrevivência própria para os indivíduos sob risco na população como

$$S_{\mathrm{P}}(t) = \mathbb{P}(T > t \mid Z \geq 1) = \frac{\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\} - p_0}{1-p_0}; \quad \gamma \leq 1, \tag{4.12}$$

com p_0 definido por (4.8). Observe que esta função está bem definida, uma vez que, $S_P(0) = 1$ e $S_P(t) = 0$, quando $t \longrightarrow \infty$. Além disso, a f.d associada ao modelo (4.12) é dada por

$$f_P(t) = \frac{f(t)}{1-p_0},$$

tal que $f(\cdot)$ é a f.d definida em (4.9). Já a função de risco correspondente a (4.12) é descrita por

$$h_P(t) = \frac{S(t)}{S(t)-p_0}\, h(t),$$

em que $S(\cdot)$ e $h(\cdot)$ são as funções de sobrevivência e de risco estabelecidas em (4.7) e (4.10), respectivamente.

Observação 4. O modelo (4.7) pode ser reescrito conforme o modelo de longa duração de Berkson e Gage (1952) por

$$S(t) = p_0 + (1-p_0)S_P(t),$$

com $S_P(t)$ indicando a função de sobrevivência própria definida em (4.12). Este fato implica que todo modelo de longa duração poderá ser escrito mediante a um modelo (4.7) para algum γ, ρ e $F_0(\cdot)$ que devem ser determinados. A demonstração deste resultado é imediata ao substituir a função $S_P(t)$ na igualdade acima.

Em particular, o modelo MEP é considerado como a função de risco de base para o modelo (4.7) neste capítulo, assumindo as hipóteses estabelecidas na Seção 2.4. Como consequência, tome uma partição no eixo dos tempos $\tau = \{s_0, \ldots, s_J\}$ de forma que $0 = s_0 < s_1 < \ldots < s_J < \infty$, gerando J intervalos disjuntos, e seja $\lambda_j > 0$ $(j = 1, \ldots, J)$. Desta forma, é possível definir a função de sobrevivência marginal em relação ao tempo T com base segundo o modelo MEP como

$$S_L(t) = \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_{MEP}(t))^{\gamma}]\right\}; \quad \gamma \leq 1 \quad \text{e} \quad \rho > 0, \tag{4.13}$$

em que $F_{MEP}(\cdot)$ é a função de distribuição acumulada (2.34) referente ao modelo MEP para todo $t \in I_j$ intervalos disjuntos e $\lambda_j > 0$. Note que a função (4.13) está bem definida, pois $S_L(0) = 1$ e, ao tomar p_0 definido por (4.8), $S_L(t) = p_0$, quando $t \longrightarrow \infty$. Também observe que são preservadas

as características dos modelos de longa duração destacadas anteriormente. Devido a isso, a função de sobrevivência (4.13) é denominada "modelo de longa duração PVF-MEP".

A f.d associada ao modelo (4.13) é dada por

$$f_L(t) = \rho f_{MEP}(t)\,(1+\rho F_{MEP}(t))^{\gamma-1}\,\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_{MEP}(t))^{\gamma}]\right\}, \tag{4.14}$$

em que $f_{MEP}(\cdot)$ é a f.d.p definida em (2.32) relacionada ao modelo MEP. Já a função de risco associada ao modelo (4.13) é definida como

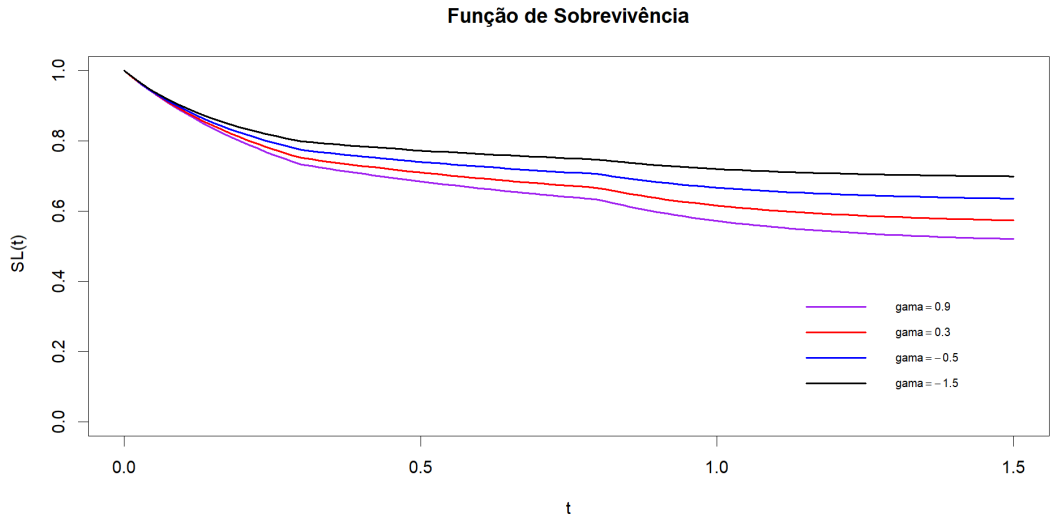
$$h_L(t) = \rho f_{MEP}(t)\,(1+\rho F_{MEP}(t))^{\gamma-1}. \tag{4.15}$$

Por sua vez, a função de risco acumulado correspondente a (4.13) é dada por

$$H_L(t) = \frac{(1+\rho F_{MEP}(t))^{\gamma}-1}{\gamma}. \tag{4.16}$$

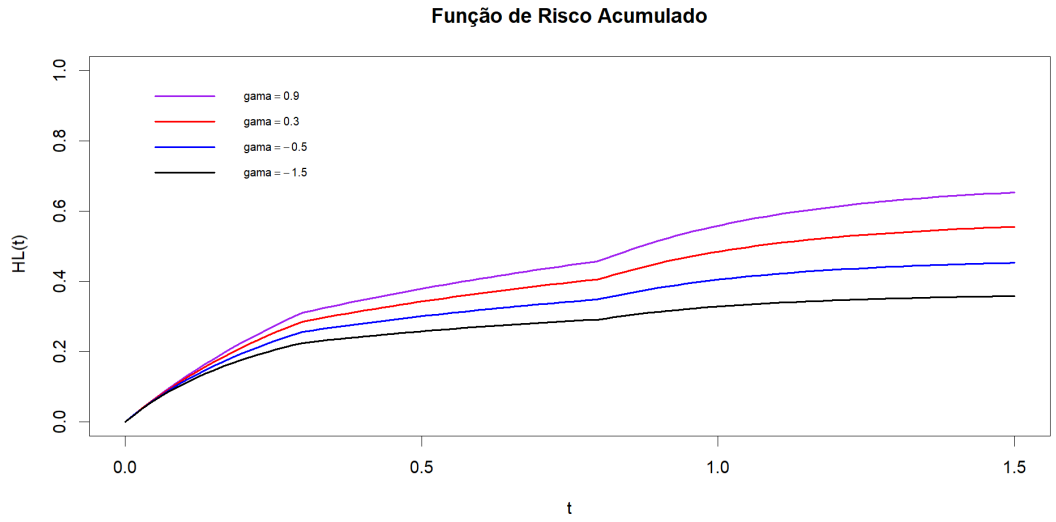
Note que a limitação da função (4.16) por $-\log(p_0)$, com p_0 definido em (4.8), também é preservada, pois $H_L(t) \longrightarrow -\log(p_0)$, quando $t \longrightarrow \infty$, resultando em $H_L(t) \leq -\log(p_0)$. As Figuras 20 e 21 ilustram os gráficos das funções de sobrevivência e de risco acumulado, respectivamente, para o modelo de longa duração PVF-MEP. Os mesmos foram gerados via o *software* R atráves do pacote **PWEXP** (TEAM, 2020) considerando três partições no eixo dos tempos (J = 3) com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$ com $\gamma \leq 1$, $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\rho = 0,7$ fixos. Note que a função de sobrevivência (4.13) tende ao valor de p_0, quando $t \longrightarrow \infty$. Já a função de risco acumulado (4.16) será limitada por $-log(p_0)$.

Figura 20 – Gráfico da função de sobrevivência do modelo de longa duração PVF-MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\rho = 0,7$ fixo.



Fonte: Elaborada pelo autor.

Figura 21 – Gráfico da função de risco acumulado do modelo de longa duração PVF-MEP considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\rho = 0,7$ fixo.



Fonte: Elaborada pelo autor.

A construção da função de sobrevivência própria ocorre de maneira similar ao que foi apresentado para a função (4.12). Assim, a função de sobrevivência própria para os indivíduos sob risco na população com base segundo o modelo MEP é dada por

$$S_{\mathrm{M}}(t) = \mathbb{P}(T > t \mid Z \geq 1) = \frac{\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_{MEP}(t))^{\gamma}]\right\} - p_0}{1-p_0}; \quad \gamma \leq 1 \quad \text{e} \quad \rho > 0, \quad (4.17)$$

com p_0 definido por (4.8) e para todo $t \in I_j$ intervalos disjuntos com $\lambda_j > 0$. Já a f.d associada ao modelo (4.17) é dada por

$$f_M(t) = \frac{f_L(t)}{1-p_0},$$

tal que $f_L(\cdot)$ é a f.d definida em (4.14). Por sua vez, a função de risco correspondente a (4.17) é descrita como

$$h_M(t) = \frac{S_L(t)}{S_L(t)-p_0} h_L(t).$$

em que $S_L(\cdot)$ e $h_L(\cdot)$ são as funções de sobrevivência e risco estabelecidas em (4.13) e (4.15), respectivamente.

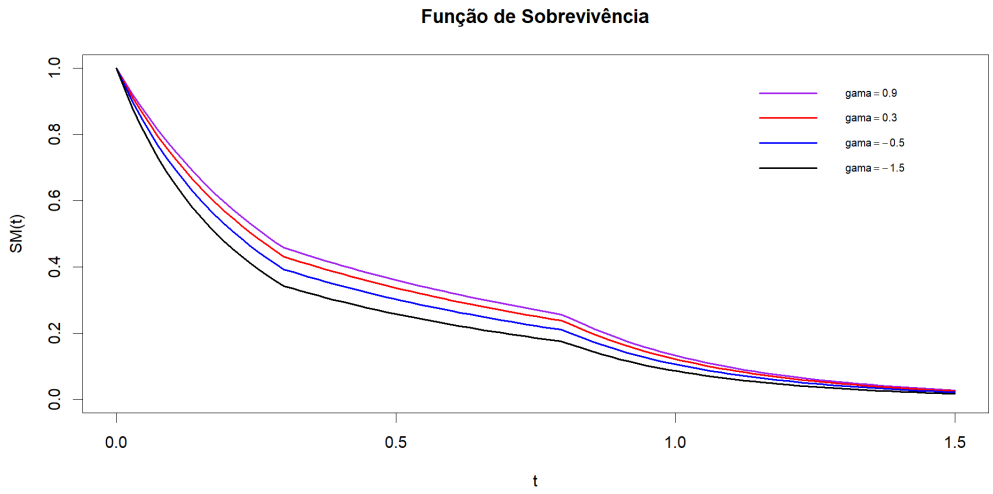
Observação 5. O modelo (4.13) pode ser reescrito conforme o modelo de longa duração de Berkson e Gage (1952) por

$$S_L(t) = p_0 + (1-p_0)S_M(t),$$

em que $S_M(t)$ é a função de sobrevivência própria definida em (4.17). A demonstração deste resultado é imediata ao substituir a função $S_M(t)$ na igualdade acima.

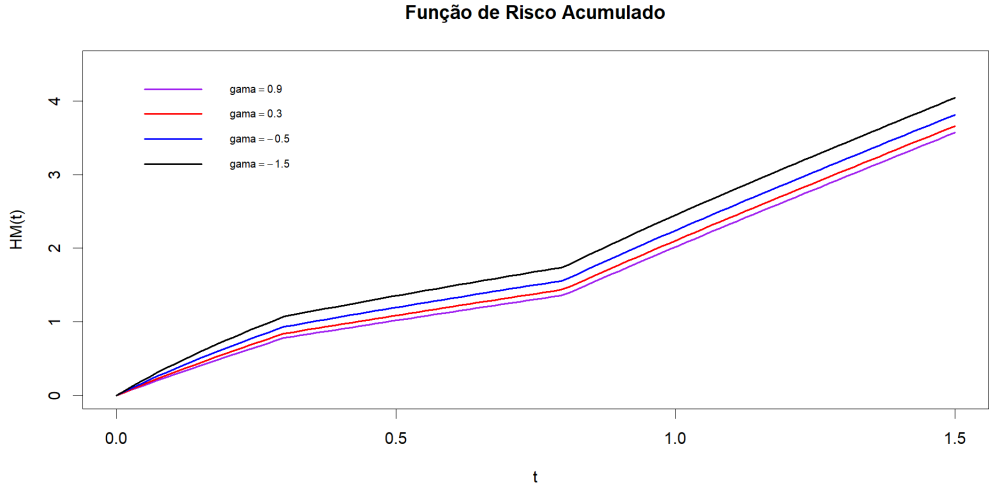
As Figuras 22 e 23 ilustram os gráficos das funções de sobrevivência e de risco acumulado, respectivamente, para o modelo próprio. Os mesmos foram gerados via o *software* R atráves do pacote **PWEXP** (TEAM, 2020) considerando três partições no eixo dos tempos (J = 3) com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$ com $\gamma \leq 1$, $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\rho = 0,7$ fixos. Note que como as funções são próprias, a função de sobrevivência tenderá a zero, enquanto a função de risco acumulado não será limitada.

Figura 22 – Gráfico da função de sobrevivência do modelo próprio considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\rho = 0,7$ fixos.



Fonte: Elaborada pelo autor.

Figura 23 – Gráfico da função de risco acumulado do modelo próprio considerando $J = 3$ partições no eixo dos tempos com $I_1 = (0;0,3]$, $I_2 = (0,3;0,8]$ e $I_3 = (0,8;\infty)$ para $n = 151$, $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 3$ e $\rho = 0,7$ fixos.



Fonte: Elaborada pelo autor.

4.2.1 Propriedades matemáticas do modelo

Similar ao realizado no Capítulo 3, nesta seção será mostrado o uso de séries de potência para descrever algumas propriedades matemáticas importantes dos modelos de sobrevivência. Deste modo, observe que é possível descrever a função de sobrevivência própria definida em (4.12) com $\rho > 0$ e $\gamma \leq 1$ por meio de séries de potência através da seguinte relação matemática

$$[1-(1+\rho F_0(t))^{\gamma}] = \sum_{i=1}^{\infty} s_i F_0(t)^i; \tag{4.18}$$

com

$$s_1 = -\gamma\,\rho, \quad s_2 = \frac{\gamma\,\rho^2\,(1-\gamma)}{2}, \quad s_3 = -\frac{\gamma\,\rho^3\,(1-\gamma)\,(2-\gamma)}{6}, \quad \ldots \quad, s_i = -\left(\frac{\gamma!\,\rho^i}{i!}\right), \quad \ldots.$$

Logo, ao utilizar a relação matemática (4.18), a função de sobrevivência própria para os indivíduos em risco (4.12) pode ser reescrita como

$$S_P(t) = \frac{\exp\left\{\frac{1}{\gamma}\sum_{i=1}^{\infty} s_i\,F_0(t)^i\right\} - p_0}{1-p_0}; \quad \rho > 0, \quad \gamma \leq 1. \tag{4.19}$$

com p_0 dado em (4.8). Note que o somatório da função (4.19) pode ser caracterizado segundo os Polinômios Exponenciais de Bell, conforme a Definição 6. Portanto, ao utilizar esta definição, é possível reescrever a função de sobrevivência marginal (4.19) como

$$S_P(t) = \frac{1}{1-p_0}\sum_{n=0}^{\infty} z_n\,\frac{F_0(t)^n}{n!} - \frac{p_0}{1-p_0}, \tag{4.20}$$

com $z_n = \sum_{k=0}^{\infty}\left(\frac{1}{\gamma}\right)^k B^{\star}_{n,k}$ e $B^{\star}_{n,k} = B_{n,k}(1!s_1, 2!s_2, \ldots, (n-k+1)!s_{n-k+1})$ e $n,k \geq 0$.

Assuma que $y_n = \dfrac{-z_{n+1}}{(1-p_0)(n+1)!}$ e tome a função $h_{n+1}(t) = (n+1)\,F_0(t)^n\,f_0(t)$ definida mediante a f.d.p base $F_0(\cdot)$ da distribuição Exponenciada com parâmetro de potência $(n+1)$ para $n \geq 0$. Desta forma, ao derivar a função (4.20), é possível determinar a f.d da variável T, sendo dada por

$$f_P(t) = \sum_{n=0}^{\infty} y_n\,h_{n+1}(t). \tag{4.21}$$

Agora, considere a variável aleatória Y tal que $Y_0 = T \sim F_0$ e $Y_{n+1} \sim \exp -F_0(n+1)$ com $n \geq 0$. Deste modo, o r-ésimo momento de T correspondente a função (4.21) é dado por

$$\mu'_r = \mathbb{E}\,(T^r) = \sum_{t=0}^{\infty} t^r f_P(t) = \sum_{t=0}^{\infty}\sum_{n=0}^{\infty} y_n\,(n+1)\,t^r\,[F_0(t)]^n\,f_0(t), \quad t > 0.$$

Para ilustrar esta propriedade serão simuladas a média e a variância de dados aleatórios, assumindo diferentes valores para os parâmetros γ e λ. Assim, considere o modelo MEP com

$J = 1$ (distribuição Exponencial) como função de base tal que $F_0(t) = 1 - e^{-\lambda t}$ e $f_0(t) = \lambda e^{-\lambda t}$. Deste modo

$$\mu_r^{'} = \sum_{t=0}^{\infty}\sum_{n=0}^{\infty} y_n\,(n+1)\,t^r\,[F_0(t)]^n\,f_0(t) = \sum_{t=0}^{\infty}\sum_{n=0}^{\infty} y_n(n+1)\;t^r\;(1-e^{-\lambda t})^n\;\lambda\;e^{-\lambda t}, \qquad (4.22)$$

com $\lambda > 0$, $t \geq 0$ e $n \geq 0$. A Tabela 16 mostra os valores simulados ao assumir $n = 10$, o vetor $t = (2,73; 2,08; 2,10; 2,34; 2,04; 1,98; 1,95; 2,63; 3,44; 2,82)$, $\rho = 2$ e $p_0 = 0,4$. Observe que para todas as combinações testadas foi possível estimar os valores da média e da variância, ilustrando a aplicabilidade do r-ésimo momento definido em (4.22). Essa propriedade também se aplica ao considerar um vetor t contendo valores entre 0 e 1.

Tabela 16 – Estimativas da média e variância para diferentes valores de $\gamma \leq 1$ e $\lambda > 0$.

	$\gamma = -0,5$		$\gamma = 0,4$		$\gamma = 0,8$	
	Média	Variância	Média	Variância	Média	Variância
$\lambda = 0,4$	0,0002	0,0008	0,0011	0,0041	0,0000	0,0000
$\lambda = 1$	0,0041	0,0124	0,0202	0,0505	0,0002	0,0005
$\lambda = 1,2$	0,0052	0,0127	0,0203	0,0506	0,0002	0,0006

Fonte: Elaborada pelo autor.

4.3 Inferência bayesiana

Considere n indivíduos e Z_i o número de riscos latentes para o i-ésimo indivíduo com $i = 1, \ldots, n$. Assuma que as variàveis Z_i são aleatórias e independentes segundo a distribuição Poisson com média $\rho\omega_i$, em que $\rho > 0$ é constante e ω_i representa as componentes de fragilidade admitindo uma distribuição PVF, segundo Hougaard (2000), com $\mathbb{E}\,(\omega_i) = 1$ e $Var\,(\omega_i) = 1 - \gamma$. Suponha que a distribuição de T_i é representada por um modelo básico de fragilidade com

$$h(T_i \mid Z_i) = Z_i\,h_0(T_i \mid \eta); \quad i = 1, \ldots, n,$$

em que $h_0(T_i \mid \eta)$ é a função de risco de base para T_i com vetor de parâmetros η. Tome T_i condicionalmente independentes dado Z_i e $t_i = min\{T_i, C_i\}$ representando as observações consideradas tal que C_i é os tempos de censura de forma que: $\delta_i = I(T_i \leq C_i) = 1$, se T_i é tempo de falha e 0, se o tempo de falha for censurado à direita. Seja $D = (n, \mathbf{t}, \delta, \mathbf{Z}, \omega)$ o vetor de dados completos, em que $\mathbf{t} = (t_1, \ldots, t_n)^T$, $\delta = (\delta_1, \ldots, \delta_n)^T$, $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ e $\omega^T = (\omega_1, \ldots, \omega_n)$ de forma que $\mathbf{Z}$ e ω são vetores aleatórios não observàveis. Considere $\vartheta = (\gamma, \rho, \eta)$ o vetor de parâmetros a serem estimados. Logo, a função de verossimilhança para ϑ dado os dados completos é dada por

$$L(\vartheta \mid D) = \prod_{i=1}^{n}[S_0(t_i \mid \eta)]^{Z_i - \delta_i}[Z_i f_0(t_i \mid \eta)]^{\delta_i} \exp\left\{\sum_{i=1}^{n}[Z_i \log(\omega_i\rho) - \log(Z_i) - \omega_i\rho]\right\}, \qquad (4.23)$$

tal que $f_0(t_i \mid \eta)$ é a f.d de base e $S_0(t_i \mid \eta)$ é a função de sobrevivência de base.

Agora, suponha $D_{OBS} = (n, \mathbf{t}, \delta)$ como o vetor dos dados observados. Note que a função de verossimilhança dos dados observados é obtida ao integrar a função (4.23) em relação a $(\mathbf{Z}, \omega)$. Logo, ao assumir a distribuição PVF para a variável ω_i com $\mathbb{E}(\omega_i) = 1$ e $Var(\omega_i) = 1-\gamma$, é possível determinar a função de verossimilhança para ϑ dado os dados observados como

$$L(\vartheta \mid D_{OBS}) = \prod_{i=1}^{n} \left[\rho f_0(t_i \mid \eta)\,(1+\rho F_0(t_i \mid \eta))^{\gamma-1}\right]^{\delta_i} \exp\left\{\sum_{i=1}^{n} \frac{1}{\gamma}\left[1-(1+\rho F_0(t_i \mid \eta))^{\gamma}\right]\right\}, \tag{4.24}$$

com $f_0(t_i \mid \eta)$ e $F_0(t_i \mid \eta)$ representando as funções de distribuição e distribuição acumulada de base.

Para incorporar covariàveis no modelo, considere $\mathbf{X}$ a matriz de covariàveis de ordem (n×p) tendo o vetor de covariàveis de ordem (p×1) para o i-ésimo indivíduo dado por $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})^T$, em que $\beta = (\beta_1, \ldots, \beta_p)^T$ é o vetor paramétrico dos coeficientes de regressão associados a $\mathbf{x_i}$, e que ρ é relacionado a estas covariàveis por meio da função de ligação logarítmica: $\rho_i = \exp\{\mathbf{x'}_i\beta\}$. Consequentemente, a função de sobrevivência com a adesão das covariàveis é dada por

$$S(t_i \mid \eta) = \exp\left\{\frac{1}{\gamma}\left[1-(1+\exp\{\beta^T\mathbf{x_i}\}F_0(t_i \mid \eta))^{\gamma}\right]\right\}. \tag{4.25}$$

Observe que, por meio da função de sobrevivência (4.25), é possível determinar a proporção de indivíduos imunes ao evento de interesse, sendo dada por

$$p_{0L} = \lim_{t\longrightarrow\infty} S(t_i \mid \eta) = \exp\left\{\frac{1}{\gamma}\left[1-(1+\exp\{\beta^T\mathbf{x_i}\})^{\gamma}\right]\right\} > 0, \tag{4.26}$$

pois $F_0(t_i \mid \eta) \longrightarrow 1$, quando $t \longrightarrow \infty$.

Deste modo, ao assumir o modelo MEP como a função de risco de base considerando as hipóteses estabelecidas na Seção 2.4, a função de verossimilhança para $\vartheta^* = (\gamma, \lambda_j, \beta)$ dado o vetor de dados observados é dada por

$$\begin{aligned} L(\vartheta^* \mid D_{OBS}) = &\prod_{i=1}^{n}\prod_{j=1}^{J} \left[\exp\{\beta^T\mathbf{x_i}\}f_{MEP}(t_i \mid \lambda_j)\,(1+\exp\{\beta^T\mathbf{x_i}\}F_{MEP}(t_i \mid \lambda_j))^{\gamma-1}\right]^{\delta_i v_{ij}} \\ &\times \left[\exp\left\{\sum_{i=1}^{n}\frac{1}{\gamma}\left[1-(1+\exp\{\beta^T\mathbf{x_i}\}F_{MEP}(t_i \mid \lambda_j))^{\gamma}\right]\right\}\right]^{v_{ij}}, \end{aligned} \tag{4.27}$$

com $D_{OBS} = (n, \mathbf{t}, \delta, \mathbf{X}, \mathbf{v})$ de forma que $\mathbf{v} = (v_{11}, \ldots, v_{nJ})^T$ com $v_{ij} = 1$, se $s_{j-1} < t_i \leq s_j$ e $v_{ij} = 0$, caso contrário, para $j = 1, \ldots, J$ intervalos disjuntos.

Para a realização da investigação bayesiana, é tomado que os parâmetros γ, β e $\lambda'_j s$ são independentes de forma que a densidade a priori conjunta é definida por

$$\pi(\gamma, \lambda_j, \beta) = \pi(\gamma)\,\pi(\lambda_j)\,\pi(\beta), \tag{4.28}$$

para $\gamma \sim N(0,1)$ truncada em 1 e $\beta \sim N(\mu_\beta, \sigma_\beta^2)$. No entanto, os parâmetros λ_j's são modelados segundo suas correlações em intervalos adjacentes, definidos por Gamerman (1991), com $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j|\varepsilon_{j-1} \sim N(\varepsilon_{k-1}, \tau)$ para $j = 1,\ldots,J$, $k = 1,\ldots,p$ e $\varepsilon_0 = 0$.

Deste modo, ao combinar a densidade a priori (4.28) com a função de verossimilhança (4.27), segue que a densidade a posteriori conjunta é dada por

$$\pi\left(\vartheta^* \mid D_{OBS}\right) \propto L(\vartheta^* \mid D_{OBS})\ \pi(\gamma)\ \pi(\lambda_j)\ \pi(\beta). \tag{4.29}$$

Novamente, a densidade a posteriori (4.29) não é uma densidade padrão sendo analiticamente intratável, consequentemente, a inferência realizada é baseada nos métodos MCMC. Assim, será usado o método HMC implementado no R-Stan, além dos critérios de comparação e seleção de modelos, descritos nas Subseções 2.5.4.1, 2.5.4.2 e 2.5.5, para as próximas análises deste capítulo.

4.3.1 Estudo de simulação

O estudo de simulação é feito para avaliar algumas propriedades frequentistas dos estimadores de Bayes com base na média, DP, viés e na REQM. Para isso, foram realizadas $M = 1.000$ simulações para cada configuração paramétrica tendo tamanhos de amostras diferentes com $n = (200; 500; 800; 1.000)$. A componente de fragilidade é gerada a partir de uma distribuição PVF, conforme descrito na Seção 4.2, para $\gamma = -0,5$ e $\gamma = 0,5$. Além disso, é adotado o modelo PVF-MEP com $J = 1$, denominado "modelo PVF-Exponencial", com $\log(\lambda) = 1$. Os tempos observados são gerados da seguinte forma:

- Gere $u_i \sim U(0,1)$.
- Se $u_i < p_{0L_i}$, então $t = \infty$. Caso contrário, é tomado $y_i = F^{-1}\left(\dfrac{(1-\gamma\log(u_1))^{\frac{1}{\gamma}} - 1}{\rho_i} \mid \lambda\right)$, em que $F^{-1}(\cdot \mid \lambda)$ corresponde a função quantil da distribuição Exponencial e p_{0L_i} é definido em (4.26), $\forall i$.
- Gere o tempo de censura C_i segundo uma distribuição Uniforme. Consequentemente, se $T_i \leq c_i$, então $t_i = T_i$ e $\delta_i = 1$. Caso contrário, se $t_i = C_i$ então $\delta_i = 0$. Em particular, a porcentagem média de observações censuradas neste estudo de simulação varia, aproximadamente, entre 7% e 20%.

Para a adesão das covariàveis são consideradas duas covariàveis, sendo x_{i1} gerada por uma distribuição Binomial com probabilidade de sucesso $0,5$ e x_{i2} mediante a distribuição Normal. Tome $\rho_i = \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\}$, para $i = 1,\ldots,n$, e $\beta_0 = \beta_1 = \beta_2 = 1$. Ainda, assuma prioris independentes para os parâmetros do modelo tal que para os λ_j's é tomado que $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j|\varepsilon_{j-1} \sim N(\varepsilon_{k-1}, 1,2)$ com $j = 1,\ldots,J$, $k = 1,\ldots,p$ e $\varepsilon_0 = 0$, além de $\beta_i \sim N\ (0,100)$, com $i = 0,1,2$, e para

- Modelo de longa duração PVF-Exponencial com $\gamma = 0,5$: Considera-se $\gamma \sim N\ (0,1)$ truncada em 1.

- Modelo de longa duração PVF-Exponencial com $\gamma = -0,5$: Suponha que $\gamma \sim N\ (0,1)$ truncada em 0.

Os resultados deste estudo de simulação são mostrados na Tabela 17. Estes resultados revelam que a medida que o tamanho da amostra aumenta, as estimativas tendem aos valores verdadeiros dos parâmetros em média. Ainda, devido a este aumento amostral, o módulo do viés, DP e a REQM tendem a zero. Estes são aspectos esperados quando o esquema de estimativa está funcionando corretamente.

Tabela 17 – Estimativas da média, DP, viés e REQM de Bayes dos parâmetros do modelo de longa duração PVF-Exponencial com a presença de covariàveis para $\gamma = -0,5$, $\gamma = 0,5$, $\log(\lambda) = 1$ e $\beta_0 = \beta_1 = \beta_2 = 1$.

		$\gamma = -0,5$				$\gamma = 0,5$			
n	Parâmetro	Média	DP	Viés	REQM	Média	DP	Viés	REQM
200	γ	-0,497	0,014	0,003	0,014	0,512	0,013	0,012	0,017
	$\log(\lambda)$	0,958	0,046	-0,042	0,045	0,964	0,045	-0,036	0,050
	β_0	0,998	0,035	-0,002	0,035	0,989	0,075	-0,011	0,075
	β_1	1,008	0,036	0,008	0,036	1,005	0,077	0,005	0,076
	β_2	1,003	0,047	0,003	0,046	1,002	0,083	-0,000	0,082
500	γ	-0,496	0,013	0,004	0,013	0,510	0,045	0,010	0,018
	$\log(\lambda)$	0,958	0,045	-0,042	0,044	0,961	0,040	-0,039	0,049
	β_0	1,004	0,037	0,004	0,037	0,991	0,081	-0,009	0,081
	β_1	0,997	0,037	-0,003	0,037	1,009	0,078	0,009	0,078
	β_2	0,996	0,032	-0,004	0,032	0,986	0,076	-0,014	0,076
800	γ	-0,497	0,014	0,003	0,014	0,511	0,015	0,011	0,019
	$\log(\lambda)$	0,960	0,039	-0,040	0,044	0,966	0,032	-0,044	0,035
	β_0	0,999	0,046	-0,001	0,046	1,001	0,076	0,003	0,075
	β_1	0,996	0,039	-0,004	0,039	1,001	0,073	-0,000	0,072
	β_2	1,002	0,037	0,002	0,037	0,989	0,073	-0,011	0,073
1.000	γ	-0,498	0,011	0,004	0,012	0,509	0,012	0,009	0,015
	$\log(\lambda)$	0,961	0,038	-0,039	0,042	0,969	0,031	-0,021	0,034
	β_0	0,999	0,045	0,003	0,045	0,998	0,062	-0,009	0,062
	β_1	0,997	0,037	-0,003	0,037	0,999	0,068	-0,019	0,070
	β_2	1,001	0,031	0,001	0,031	0,991	0,072	-0,029	0,072

Fonte: Elaborada pelo autor.

4.4 Aplicação: dados AIDS/HIV

Para esta análise foram retomados o conjunto de dados sobre AIDS/HIV descrito na Subseção 3.5.1.1.

4.4.1 Análise dos dados

Neste estudo foram observadas as seguintes variàveis: y_i: tempo até a morte por AIDS/HIV ou tempo censurado (em anos), x_{i_1}: faixa etária (16 a 65 anos) e x_{i_2}: POI na inclusão no estudo (não=0: sem infecção; sim=1: com infecção) para $i = 1,\ldots,272$. Além disso, este conjunto de dados possui a taxa de censura de, aproximadamente, 8%. Desta forma, para estes dados, é ajustado o modelo de longa duração PVF-MEP descrito na Seção 4.2 com todas as covariàveis no parâmetro ρ, ou seja

$$\log(\rho_i) = \beta_0 + \beta_{idade} x_{i1} + \beta_{POI} x_{i2_1}; \quad i = 1,\ldots,272,$$

em que os efeitos das covariàveis são definidos por meio de variàveis fictícias, sendo que x_1 corresponde a idade dos pacientes em anos e

$$x_{2_1} = \begin{cases} 1, & \text{se há a presença de POI;} \\ 0, & \text{caso contrário.} \end{cases}$$

Para o modelo proposto são adotadas a densidade a posteriori (4.29) e as seguintes priores independentes para os cálculos bayesianos: $\gamma \sim N\,(0,1)$ truncada em 1, $\beta_i \sim N\,(0,100)$ com $i = 1,\ldots,272$ e para os parâmetros λ_j's é tomado que $\log(\lambda_k) = \varepsilon_k$ e $\varepsilon_j|\varepsilon_{j-1} \sim N(\varepsilon_{k-1},1,2)$ com $\varepsilon_0 = 0$, $k = 1,\ldots,p$ e $j = 1,\ldots,J$ intervalos disjuntos. Os cálculos bayesianos foram baseados em amostras HMC obtidas de quatro cadeias independentes com 4.000 observações para cada parâmetro. Para eliminar o efeito dos valores iniciais, as primeiras 2.000 iterações foram desconsideradas. O modelo MEP foi investigado usando $J = 1,\ldots,6$ intervalos disjuntos na partição do eixo dos tempos, consequentemente, as estatísticas do LPML para os dados de HIV/AIDS são relatados na Tabela 18. Estas estatísticas são utilizadas para determinar uma partição J apropriada do eixo dos tempos. Observe que, com base nas estatísticas do LPML, o modelo com $J = 1$ (modelo de longa duração PVF-Exponencial) é considerado o modelo de melhor ajuste para estes dados.

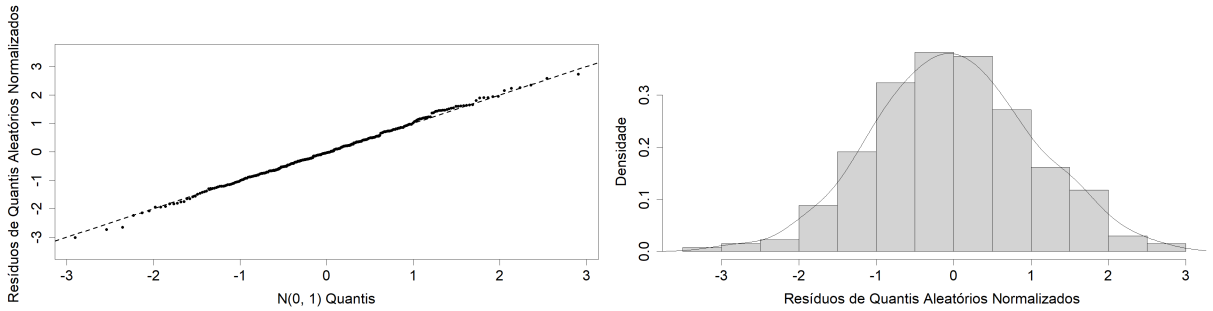
Tabela 18 – Estatísticas do LPML para os dados de HIV/AIDS.

J	1	2	3	4	5	6
LPML	**-89,89**	-92,60	-94,12	-94,99	-96,41	-96,46

Fonte: Elaborada pelo autor.

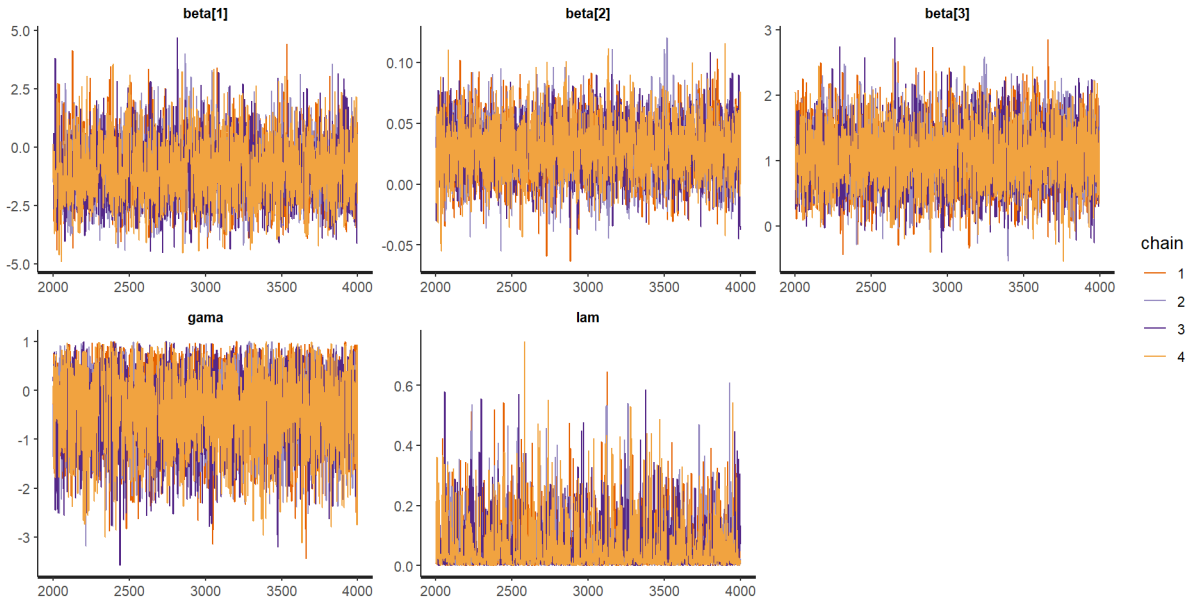
Para avaliar a adequação do ajuste do modelo de longa duração PVF-Exponencial são obtidos os resíduos de quantis aleatórios normalizados a posteriores. O gráfico QQ-plot dos resíduos destes quantis aleatórios são exibidos na Figura 24 e sugerem que o modelo determinado possui um ajuste aceitável ao dados de AIDS/HIV. Já a convergência das cadeias bayesianas são monitoradas pelos métodos de Cowles e Carlin (1996), sendo ilustrado na Figura 25. Note que o comportamento dos gráficos de rastreamento indicam a convergência das cadeias.

Figura 24 – Gráfico QQ-plot dos resíduos de quantis normalizados a posteriores com linha de identidade (a esquerda) e histograma do modelo de longa duração PVF-Exponencial sob as hipóteses definidas (a direita).



Fonte: Elaborada pelo autor.

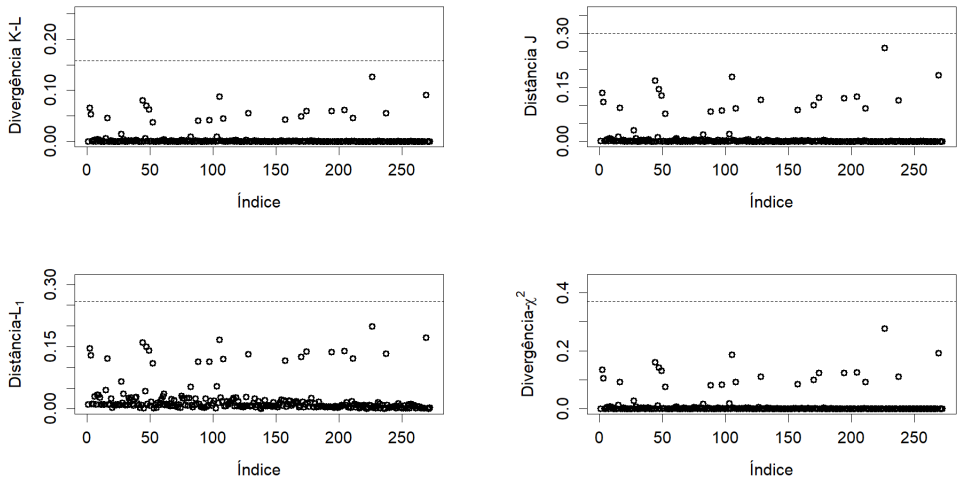
Figura 25 – Gráficos de rastreamento para os parâmetros do modelo de longa duração PVF-Exponencial referente aos dados de AIDS/HIV.



Fonte: Elaborada pelo autor.

Considerando as amostras das distribuições a posteriores dos parâmetros foram calculadas as medidas de ϕ-divergência descritas na Subeção 2.5.5. A Figura 26 mostra o gráfico do índice das quatro medidas de divergência ϕ, em que é notado que não há possíveis observações influentes na distribuição a posteriori dos parâmetros do modelo de regressão deste estudo, consequentemente, não haverá mudanças inferenciais a serem removidas nas observações.

Figura 26 – Gráfico de índice das medidas de divergência ϕ relacionadas aos dados da AIDS/HIV.



Fonte: Elaborada pelo autor.

As estimativas bayesianas paramétricas, bem como, as médias a posteriores, medianas, DP's e 95% de confiança dos intervalos HPD para o modelo de longa duração PVF-Exponencial são exibidas na Tabela 19. Observe que é possível determinar a proporção de pacientes curados/imunes ao evento de interesse, sendo de $p_{OL} = 0,52$, além de investigar o impacto que as covariàveis possuem sobre esta proporção. Note que a estimativa de Bayes de $\beta_{idade} > 0$ implicando que a cada ano adicional do paciente haverá uma diminuição da sua proporção de cura, correspondendo a $\exp\{0,028\} = 1,02$ vezes. Além disso, como $\beta_{POI} > 0$ temos que a proporção de pacientes curados e que possuem POI é $\exp\{1,074\} = 2,92$ vezes menor ao se comparar com os pacientes que não possuem esta infecção.

Tabela 19 – Estatísticas LPML, médias a posteriores, medianas, DP's e 95% de confiança dos intervalos HPD para os parâmetros do modelo de longa duração PVF-Exponencial nos dados de AIDS/HIV.

LPML	Parâmetros	Média	Mediana	DP	L	U
-89,89	γ	-0,319	-0,251	0,777	-1,718	0,998
	$\log\lambda$	0,055	0,024	0,077	0,000	0,221
	β_0	-0,964	-1,026	1,319	-3,481	1,634
	β_{idade}	0,028	0,028	0,023	-0,018	0,072
	β_{POI}	1,074	1,062	0,469	0,188	2,023

Fonte: Elaborada pelo autor.

Para analisar a proporção de indivíduos imunes ao evento de interesse do estudo foram considerados quatro pacientes hipotéticos A, B, C e D. Estes pacientes são caracterizados com valores diferenciados para a covariàvel idade, contudo é assumido que todos possuem a existência de POI, devido a relevância deste fator na análise do estudo. Consequentemente, as estimativas bayesianas paramétricas, bem como, as médias a posteriores, medianas, DP's e 95% de confiança

dos intervalos HPD para estes indivíduos são retratados na Tabela 20. Note que, por exemplo, ao comparar o paciente A com 16 anos e o paciente D com 52 anos, é visto que estes possuem probabilidades de não morrer diferentes e distantes, sendo de 0,462 para o paciente A e 0,343 para o paciente D. Contudo, ao considerar os pacientes A e B, que descrevem pacientes com idades próximas como 16 e 24 anos, respectivamente, nota-se que mesmo estes tendo probabilidades de não morrer diferentes, correspondendo a 0,462 e 0,430, respectivamente, ambas estão próximas.

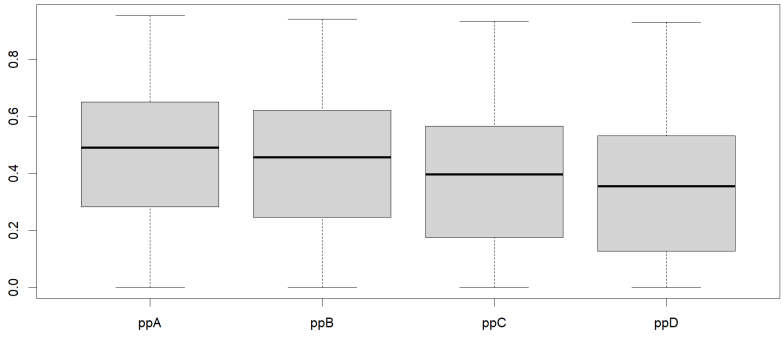
Tabela 20 – Estimativas de Bayes da probabilidade de não morrer e intervalos de HPD de 95% de confiança, para quatro hipopacientes terapêuticos com AIDS.

Pacientes	Idade	POI	Média	Mediana	DP	L	U
A	16	sim	0,462	0,490	0,239	0,000	0,891
B	24	sim	0,430	0,457	0,236	0,000	0,786
C	40	sim	0,375	0,396	0,232	0,000	0,735
D	52	sim	0,343	0,355	0,232	0,000	0,707

Fonte: Elaborada pelo autor.

A Figura 27 fornece os boxplots das médias posteriores das proporções de indivíduos imunes ao evento de interesse para o modelo de longa duração PVF-Exponencial, relacionados aos pacientes hipotéticos descritos anteriormente. Desta forma, observa-se que a taxa mediana de cura para os pacientes A, B, C e D possuem um valor pequeno mas diferenciado entre si, sendo de aproximadamente 0,490, 0,457, 0,396 e 0,355, respectivamente, decaindo a medida que a idade aumenta. As faixas dos bigodes que representam os pacientes A, B, C e D são 0,462, 0,430, 0,375 e 0,343, respectivamente, indicando uma pequena heterogênidade na recuperação dos pacientes.

Figura 27 – Boxplot das médias posteriores para pacientes hipotéticos A, B, C e D segundo o modelo de longa duração PVF-Exponencial.



Fonte: Elaborada pelo autor.

A Tabela 21 exibe as estimativas bayesiana paramétricas, bem como, as médias a posteriores e 95% de confiança dos intervalos HPD para os sub-modelos do modelo de longa

duração PVF-MEP retratado na Tabela 15. Para isso, foram preservadas as hipóteses estipuladas anteriormente, contudo tomando $\gamma = 1$ (BCH-MEP), $\gamma \longrightarrow 0$ (BCH-Gama-MEP) e $\gamma = 0,5$ (BCH-IG-MEP). Note que como $\beta_{idade} > 0$ e $\beta_{POI} > 0$ ainda é constatado que a cada ano adicional haverá uma diminuição na proporção de cura dos pacientes, além da diminuição desta proporção para os pacientes que possuem infecção prévia. Contudo, ao analisar a estatística do LPML, é observado que, aparentemente, o modelo de longa duração PVF-Exponencial se destaca como o modelo de melhor ajuste para os dados de HIV/AIDS.

Tabela 21 – Estatísticas LPML, médias a posteriores e 95% de confiança dos intervalos HPD para os parâmetros nos dados de AIDS/HIV.

Parâmetros	BCH-MEP		BCH-Gama-MEP		BCH-IG-MEP	
	Média	Intervalos HPD (95%)	Média	Intervalos HPD (95%)	Média	Intervalos HPD (95%)
γ	—	—	—	—	—	—
$\log\lambda$	0,063	(0,000; 0,251)	0,057	(0,000 ; 0,221)	0,057	(0,000 ; 0,226)
β_0	-1,114	(-3,508 ; 1,432)	-1,041	(-3,441 ; 1,586)	-1,042	(-3,372 ; 1,517)
β_{idade}	0,027	(-0,013 ; 0,067)	0,029	(-0,014 ; 0,072)	0,029	(-0,013 ; 0,070)
β_{POI}	0,979	(0,143 ; 1,827)	1,073	(0,125 ; 1,922)	1,020	(0,186 ; 1,907)
LPML	-89,97		-90.09		-90.07	

Fonte: Elaborada pelo autor.

Como as estimativas a posteriores do modelo de longa duração PVF-Exponecial e de seus sub-modelos são próximas, é realizada a estimação da verossimilhança marginal por meio de *bridge sampling*. Este método determina as probabilidades marginais dos modelos a posteriores comparando-os por meio das funções (2.37) e (2.39). Estas estimativas são exibidas na Tabela 22. Portanto, é ressaltado que o modelo que melhor se ajusta aos dados de HIV/AIDS é o modelo de longa duração PVF-Exponencial com uma probabilidade de 44%.

Tabela 22 – Probabilidades do modelo de regressão a posteriori mediante as verossimilhanças marginais.

	PVF-Exponencial	BCH-MEP	BCH-Gama-MEP	BCH-IG-MEP
Probabilidade	**0,44**	0,18	0,20	0,18

Fonte: Elaborada pelo autor.

4.5 Conclusão

Neste capítulo é apresentado um modelo de fragilidade para modelar a heterogeneidade não observada nos dados de sobrevivência, onde o modelo desenvolvido é estendido para permitir o uso de uma distribuição discreta, resultando em uma fragilidade zero analisada por meio de modelos de longa duração. Desta forma, na Seção 4.2, são estruturados modelos de sobrevivência induzidos por fragilidade discreta, utilizando de uma distribuição de Poisson mista para explicar a dispersão não observada, e tendo como componente de fragilidade uma distribuição PVF cujo modelo MEP é usado como função base.

A Seção 4.3 descreve a abordagem inferencial, baseada em métodos bayesianos utilizando o método HMC implementados no R-Stan, no qual, o estudo de simulação revela que este modelo possui boas propriedades dos estimadores de Bayes. Por fim, na Seção 4.4, a utilidade do modelo é validada por meio de uma aplicação em dados reais, mostrando que este modelo é uma boa alternativa para estudos e análises sobre dados de AIDS/HIV.

CAPÍTULO

5

MODELO BIVARIADO BAYESIANO PARA DADOS DE SOBREVIVÊNCIA DE LONGA DURAÇÃO

5.1 Introdução

O capítulo anterior ressalta o emprego de modelos de sobrevivência de longa duração em estudos que envolvem, principalmente, as àreas de confiabilidade e sobrevivência cuja famosa proporção de indivíduos imunes retratam a existência de unidades de amostragem que não são susceptíveis a ocorrência do evento de interesse. Contudo, estes modelos são enfatizados mediante a abordagem univariada, além de obter um grande destaque na literatura devido à suas contribuições nos mais diversos estudos, destacando os modelos de mistura de Berkson e Gage (1952), o modelo BCH de Tsodikov, Yakovlev e Asselain (1996) e os estudos de Tsodikov, Ibrahim e Yakovlev (2003), Cancho *et al.* (2021), Rodrigues *et al.* (2009b) e Souza *et al.* (2017).

Deste modo, é evidenciado que na literatura há pouca atenção voltada para pesquisas sobre modelos de longa duração multivariado (BEDIA, 2022). Dentre os autores que estudam este tema, ressaltando a abordagem frequentista, estão Chatterjee e Shih (2001), Price e Manatunga (2001), Gallardo, Bolfarine e Pedroso-De-Lima (2016) e Kim (2017). Ainda, Bedia (2022) também propõe um novo modelo de sobrevivência multivariada com a proporção de indivíduos imunes em termos de modelos de fragilidade e Giussani e Bonetti (2019) usam técnicas multivariadas de sobrevivência para a análise de dados de tempo de falha com censura à direita, em que é investigado uma nova família de modelos paramétricos de fragilidade bivariada.

Este cenário não é muito diferente ao considerar a abordagem bayesiana, sendo geralmente caracterizada por meio da extensão do modelo BHC para a modelagem de dados de sobrevivência multivariada. Exemplos de autores que tratam esse tema incluem Cancho *et al.* (2018), Martins, Silva e Andreozzi (2017) e Cancho *et al.* (2022). Cancho *et al.* (2018) propõem

um modelo de sobrevivência multivariado com fração de cura, assumindo múltiplos tipos de causas latentes, onde o número de causas de cada tipo segue uma distribuição Poisson multivariada. Já Martins, Silva e Andreozzi (2017) desenvolvem um modelo multivariado para dados longitudinais e de sobrevivência, incorporando a proporção de indivíduos imunes ao evento de interesse em seu estudo, e Cancho *et al.* (2022) apresentam um modelo de longa duração utilizando a fragilidade como uma mistura de distribuições Poisson e Gama.

Neste contexto, o objetivo deste capítulo é apresentar um modelo de sobrevivência bivariado com fragilidade, no qual uma distribuição de Poisson é utilizada para explicar a dispersão não observada e a componente de fragilidade segue uma distribuição PVF, segundo Hougaard (2000). O modelo é semelhante ao proposto por Bedia (2022), mas incorpora os modelos de Chen, Ibrahim e Sinha (2002) e Cancho *et al.* (2022), além de ser estendido para incluir os modelos de Cancho, Rodrigues e Castro (2011) e de regressão. A inferência é realizada por métodos bayesianos, usando o método HMC implementado no R-Stan, no qual, alguns resultados de simulação são apresentados para avaliar o desempenho dos estimadores de Bayes. A relevância do modelo bivariado é demonstrada por meio de uma aplicação a dados reais, oferecendo suporte estratégico para reduzir o *churn* e aumentar a lealdade dos clientes.

5.2 Modelo bivariado de longa duração PVF

Nesta seção será exibido o modelo bivariado para dados de sobrevivência de longa duração, sendo embasado no modelo apresentado por Bedia (2022). Para isso, considere T_1 e T_2 os tempos de sobrevivência para dois indivíduos relacionados ou os tempos para um indivíduo que possui dois eventos de interesse. Consequentemente, para um indivíduo arbitrário na população do estudo é assumida uma distribuição T_k $(k = 1,2)$ que é representada por um modelo de fragilidade descrito por

$$h_k(t_k \mid Z_k) = Z_k\, h_{0k}(t_k); \quad t_k > 0 \quad \text{e} \quad k = 1,2,$$

em que $h_{0k}(t_k)$ é a função de risco de base para T_k. Assuma Z_k uma variàvel aleatória e discreta que denota o número não observado de causas de risco para o k-ésimo evento de interese, sendo caracterizada segundo uma distribuição de Poisson com média $\omega\rho_k$, tal que $\rho_k > 0$ constante e ω é a componente de fragilidade variando ao longo da reta real positiva. Suponha que a variàvel ω é definida por uma distribuição PVF, denotada como PVF(γ,μ,σ), com $\gamma \leq 1$, $\mu > 0$ e σ caracterizado para dois casos: se $0 < \gamma \leq 1$ é tomado que $\sigma \geq 0$, caso contrário, se $\gamma \leq 0$ é admitido que $\sigma > 0$ (HOUGAARD, 2000). Como consequência, sob esta parametrização, é definida a transformada de Laplace para o modelo PVF(γ,μ,σ) como

$$\mathscr{L}_\omega(s) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+s)^\gamma - \sigma^\gamma\right]\right\}. \tag{5.1}$$

Seja Z_1 e Z_2 condicionalmente independentes dado ω. Note que a variàvel ω é responsável por induzir uma correlação entre as variáveis latentes Z_1 e Z_2, além da heterogeneidade

não observada de cada variável Z_k com $k = 1,2$ (BEDIA, 2022). Como consequência, sob estas suposições, é possível definir a função de sobrevivência bivariada condicional, dado ω, como

$$S(t_1,t_2 \mid \omega) = \exp\{-\omega[\rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2)]\}; \quad t_1 > 0 \quad \text{e} \quad t_2 > 0, \tag{5.2}$$

com $\rho_k > 0$ e $F_{0k}(t_k) = 1 - \exp\left\{-\int_0^{t_k} h_{0k}(u)du\right\}$ representando a função de distribuição acumulada base para o tempo T_k tal que $k = 1,2$. A demonstração que determina a função de sobrevivência (5.2) é encontrada no trabalho de Bedia (2022), sendo utilizado para fundamentar este capítulo.

Observe que, ao utilizar a transformada de Laplace (5.1), é possível determinar a função de sobrevivência marginal conjunta para T_1 e T_2 como

$$S(t_1,t_2) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma + \rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))^{\gamma} - \sigma^{\gamma}\right]\right\}; \quad \gamma \leq 1, \quad \mu > 0 \quad \text{e} \quad \sigma > 0, \tag{5.3}$$

com $t_k > 0$ e $\rho_k > 0$ para $k = 1,2$. Logo, as funções de sobrevivência marginal ralacionadas ao modelo (5.3) são dadas por

$$S(t_k) = \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma + \rho_k F_{0k}(t_k))^{\gamma} - \sigma\right]\right\}; \ \gamma \leq 1, \ \mu > 0, \ \sigma > 0 \ \text{e} \ t_k > 0 \ \text{com} \ k = 1,2. \tag{5.4}$$

Note que o modelo (5.4) se estenderá a outros modelos destacados na literatura, como por exemplo:

- O modelo de Cancho *et al.* (2022): ao assumir $\gamma \longrightarrow 0$ e $\mu = \sigma = \frac{1}{\theta}$ com $\theta > 0$, obtendo $S(t) = (1 + \theta\rho F_0(t))^{\frac{1}{\theta}}$ que corresponde a distribuição Gama.
- O modelo de Chen, Ibrahim e Sinha (2002): ao tomar $\sigma = 0$ e $\mu = \gamma$ no modelo (5.3).
- A um novo modelo, ao considerar $\gamma = \dfrac{1}{2}$, $\mu = \sigma^{\frac{1}{2}}$ e $\sigma^{-1} = 2\theta$, com $\theta > 0$, resultando em

$$S(t_1,t_2) = \exp\left\{\frac{1}{\theta}\left[1 - \sqrt{1 + 2\theta(\rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))}\right]\right\}.$$

Para evitar problemas de identificabilidade, é considerada uma restrição no modelo (5.2) de forma que $\mu = \sigma = 1$ obtendo que $\mathbb{E}[\omega] = 1$ e $Var(\omega) = 1 - \gamma := \theta$. Como consequência, $\mathbb{E}[Z_k] = \rho_k$, $Var[Z_k] = \rho_k(1 + \theta\rho_k)$ e $Cov(Z_1,Z_2) = \theta\rho_1\rho_2$ para $k = 1,2$. Além disso, sob estas suposições, a função de sobrevivência marginal conjunta de T_1 e T_2 é reescrita por

$$S(t_1,t_2) = \exp\left\{-\frac{1}{\gamma}\left[(1 + \rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))^{\gamma} - 1\right]\right\}; \quad \gamma \leq 1, \quad t_1 > 0 \quad \text{e} \quad t_2 > 0. \tag{5.5}$$

Observe que, por meio da função de sobrevivência (5.5), é possível determinar a proporção de indivíduos imunes ao evento de interesse, sendo dada por

$$p_{00} = \lim_{t \longrightarrow \infty} S(t_1,t_2) = \exp\left\{-\frac{1}{\gamma}\left[(1 + \rho_1 + \rho_2)^{\gamma} - 1\right]\right\} > 0, \tag{5.6}$$

pois, $F_{0k}(t_k) \longrightarrow 1$, quando $t_k \longrightarrow \infty$ e $k = 1,2$. Este fato implica que a função (5.5) é uma função de sobrevivência imprópria e que p_{00} é a proporção conjunta de indivíduos imunes ao evento de interesse. Portanto, o modelo (5.5) será denominado como "modelo bivariado de longa duração PVF".

Em um contexto geral, as funções de sobrevivência marginal relacionadas ao modelo (5.5) são definas por

$$S(t_k) = \exp\left\{-\frac{1}{\gamma}[(1+\rho_k F_{0k}(t_k))^{\gamma} - 1]\right\}; \quad \gamma \leq 1, \tag{5.7}$$

com $t_k > 0$ e $\rho_k > 0$ para $k = 1,2$. Isso implica que esta função possuirá a proporção de indivíduos imunes ao evento de interesse, sendo dada por

$$p_{0k} = \exp\left\{-\frac{1}{\gamma}[(1+\rho_k)^{\gamma} - 1]\right\} > 0; \quad \gamma \leq 1 \quad \text{e} \quad \rho_k > 0 \quad \text{com} \quad k = 1,2. \tag{5.8}$$

Já a probabilidade de indivíduos imunes a ambos os eventos de interesse após $t > 0$ anos do período de acompanhamento é dada por

$$\mathbb{P}(t) = \mathbb{P}(Z_1 = 0, Z_2 = 0 \mid T_1 > t, T_2 > t) = \left[\frac{\exp\{(1+\rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))^{\gamma}\}}{\exp\{(1+\rho_1+\rho_2)^{\gamma}\}}\right]^{\frac{1}{\gamma}},$$

implicando que $\mathbb{P}(0) = p_{00}$, caracterizando a proporção de indivíduos imunes a ambos os eventos de interesse após o término do período de observação.

A f.d marginal conjunta da distribuição bivariada de T_1 e T_2 é determinada mediante a função (5.5) por

$$\begin{aligned} f(t_1,t_2) = {} & \rho_1\rho_2 f_{01}(t_1) f_{02}(t_2) S(t_1,t_2)(1+\rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))^{\gamma-2} \\ & \times\ [1-\gamma+(1+\rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))^{\gamma}], \end{aligned} \tag{5.9}$$

com $t_k > 0$, $1-\gamma > 0$, $\rho_k > 0$ e $f_{0k}(t_k) = \dfrac{\partial}{\partial t_k} F_{0k}(t_k)$ para $k = 1,2$.

Por sua vez, Clayton (1978) calcula a medida de associação local entre T_1 e T_2, sendo definida em (2.20) por

$$v^*(t_1,t_2) = 1 + (1-\gamma)(1+\rho_1 F_{01}(t_1) + \rho_2 F_{02}(t_2))^{-\gamma}.$$

5.3 Inferência bayesiana

Nesta seção será determinada a função de verossimilhança utilizada para os parâmetros do modelo bivariado (5.5), sendo baseada no trabalho de Bedia (2022), além das distribuições a priori e posteriori necessárias para a realização da investigação bayesiana. Para isso, suponha n indivíduos na amostra e seja Z_{kj} o número de riscos latentes do k-ésimo tipo de evento de interesse para o j-ésimo indivíduo com $j = 1,\ldots,n$ e $k = 1,2$. Assuma que Z_{kj} são variàveis

aleatórias de Poisson independentes com média $\omega_j\rho_k$ de forma que $\rho_k > 0$ é constante e ω_j são variàveis aleatórias e identicamente distribuidas segundo uma distribuição PVF com $\mathbb{E}(\omega_j) = 1$ e $Var(\omega_j) = 1-\gamma > 0$. Seja T_{kj} o tempo de falha para o j-ésimo indivíduo no k-ésimo tipo de evento de interesse, onde T_{kj} pode ser censurado pela direita, sendo representado pelo modelo básico de fragilidade dado por

$$h_k(t_{kj} \mid Z_{kj}) = Z_{kj} h_{0k}(t_{kj} \mid \eta_k),$$

em que $h_{0k}(t_{kj} \mid \eta_k)$ é a função de risco de base comum para todos os indivíduos com vetor de parâmetros η_k. Suponha que T_{kj} e Z_{kj} são condicionalmente independentes e que $t_{kj} = min\{T_{kj}, C_{kj}\}$ são as observações consideradas tal que C_{kj} são os tempos de censura de forma que: $\delta_{kj} = I(T_{kj} \leq C_{kj}) = 1$, se T_{kj} é tempo de falha e 0, se o tempo de falha for censurado à direita. Seja $D = (n, \mathbf{t_1}, \mathbf{t_2}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2}, \mathbf{Z_1}, \mathbf{Z_2}, \omega)$ o vetor dos dados completos, com $\mathbf{t_k} = (t_{k1}, \ldots, t_{kn})$, $\boldsymbol{\delta_k} = (\delta_{k1}, \ldots, \delta_{kn})$ e $\mathbf{Z_k} = (Z_{k1}, \ldots, Z_{kn})$ para $k = 1, 2$ e $\omega = (\omega_1, \ldots, \omega_n)$ de forma que $\mathbf{Z_1}$, $\mathbf{Z_2}$ e ω são vetores aleatórios não observàveis. Considere $\vartheta = (\gamma, \rho_1, \rho_2, \eta_1, \eta_2)$ o vetor de parâmetros a serem estimados. Logo, a função de verossimilhança para ϑ, dado os dados completos, é dada por

$$\begin{aligned} L(\vartheta \mid D) = & \prod_{k=1}^{2}\prod_{j=1}^{n}[S_{0k}(t_{kj} \mid \eta_k)]^{Z_{kn}-\delta_{kn}}[Z_{kj} f_{0k}(t_{kj} \mid \eta_k)]^{\delta_{kn}} \\ & \times \prod_{k=1}^{2} \exp\left\{\sum_{j=1}^{n}[Z_{kj}\log(\omega_j\rho_k) - \log(Z_{kj}!) - \omega_j\rho_k]\right\} \prod_{j=1}^{n} g(\omega_j), \end{aligned} \tag{5.10}$$

em que $f_{0k}(t_{kj} \mid \eta_k)$ é a f.d de base, $S_{0k}(t_{kj} \mid \eta_k)$ é a função de sobrevivência de base e $g(\cdot)$ é a f.d da distribuição PVF.

Agora, suponha $D_{OBS} = (n, \mathbf{t_1}, \mathbf{t_2}, \boldsymbol{\delta_1}, \boldsymbol{\delta_2})$ como o vetor dos dados observados. Note que a função de verossimilhança dos dados observados é obtida ao integrar a função (5.10) em relação a $(\mathbf{Z_1}, \mathbf{Z_2}, \omega)$. Como consequência, é possível determinar a função de verossimilhança para ϑ, dado os dados observados, como

$$\begin{aligned} L(\vartheta \mid D_{OBS}) = & \prod_{j=1}^{n}(\rho_1 f_{01}(t_{1j} \mid \eta_1))^{\delta_{1j}}(\rho_2 f_{02}(t_{2j} \mid \eta_2))^{\delta_{2j}} \\ & \times \prod_{j=1}^{n} \exp\left\{-\frac{1}{\gamma}\left[(1 + \rho_1 F_{01}(t_{1j} \mid \eta_1) + \rho_2 F_{02}(t_{2j} \mid \eta_2))^{\gamma} - 1\right]\right\} \\ & \times \prod_{j=1}^{n}(1 + \rho_1 F_{01}(t_{1j} \mid \eta_1) + \rho_2 F_{02}(t_{2j} \mid \eta_2))^{(1-\gamma)(\delta_{1j}+\delta_{2j}) - \gamma\delta_{1j}\delta_{2j}} \\ & \times \prod_{j=1}^{n}[1 - \gamma + (1 + \rho_1 F_{01}(t_{1j} \mid \eta_1) + \rho_2 F_{02}(t_{2j} \mid \eta_2))^{\gamma}]^{\delta_{1j}\delta_{2j}}, \end{aligned} \tag{5.11}$$

em que $f_{0K}(t_{Kj} \mid \eta_K)$ e $F_{0K}(t_{Kj} \mid \eta_K)$ são as funções de distribuição e distribuição acumulada de base, respectivamente, para $j = 1, \ldots, n$ e $k = 1, 2$. As demonstrações das funções (5.10) e (5.11) podem ser encontradas em Bedia (2022).

Para incorporar as covariàveis no modelo (5.5) é necessário considerar a inclusão destas covariàveis para cada indivíduo indexado por $j = 1,\ldots,n$, além de ter uma média distinta para cada causa latente denotada como ρ_{kj}. Assim, assuma o vetor paramétrico p-dimensional de covariàveis para o j-ésimo indivíduo dado por $\mathbf{x_j^T} = (x_{j1},\ldots,x_{jp})$ e o seu respectivo vetor paramétrico dos coeficientes de regressão $\beta_{\mathbf{k}} = (\beta_{k1},\ldots,\beta_{kp})^T$ com $k = 1,2$. Suponha que ρ_{kj} é relacionado as covariàveis por meio da função de ligação logarítmica: $\rho_{kj} = \exp\{\mathbf{x_j^T}\beta_k\}$. Consequentemente, a função de sobrevivência associada ao modelo (5.5) com a adesão das covariàveis é dada por

$$S_j(t_1,t_2) = \exp\left\{-\frac{1}{\gamma}\left[(1+\exp\{\mathbf{x_j^T}\beta_1\}F_{01}(t_1)+\exp\{\mathbf{x_j^T}\beta_2\}F_{02}(t_2))^{\gamma}-1\right]\right\};\quad \gamma\leq 1. \quad (5.12)$$

Ou ainda

$$S_{kj}(t_k) = \exp\left\{-\frac{1}{\gamma}\left[(1+\exp\{\mathbf{x_j^T}\beta_k\}F_{0k}(t_k))^{\gamma}-1\right]\right\};\quad \gamma\leq 1,\quad j=1,\ldots,n\ \text{ e }\ k=1,2. \quad (5.13)$$

Observe que, através da função de sobrevivência (5.13), é possível determinar a proporção de indivíduos imunes ao evento de interesse, sendo dada por

$$p_{0kj} = \exp\left\{-\frac{1}{\gamma}\left[(1+\exp\{\mathbf{x_j^T}\beta_k\})^{\gamma}-1\right]\right\};\quad \gamma\leq 1,\quad j=1,\ldots,n\ \text{ e }\ k=1,2. \quad (5.14)$$

Sob as suposições definidas nesta seção e ao tomar $\beta = (\beta_1^T,\beta_2^T)^T$, a função de verossimilhança para $\vartheta^* = (\gamma,\beta,\eta_1,\eta_2)$, dado os dados observados, é dada por

$$\begin{aligned} L(\vartheta^* \mid D_{OBS}) = &\prod_{j=1}^{n}\prod_{k=1}^{2}(\exp\{\mathbf{x_j^T}\beta_k\}f_{0k}(t_{kj}\mid\eta_k))^{\delta_{kj}} \\ &\times\prod_{j=1}^{n}\exp\left\{-\frac{1}{\gamma}\left[(1+\exp\{\mathbf{x_j^T}\beta_1\}F_{01}(t_{1j}\mid\eta_1)+\exp\{\mathbf{x_j^T}\beta_2\}F_{02}(t_{2j}\mid\eta_2))^{\gamma}-1\right]\right\} \\ &\times\prod_{j=1}^{n}(1+\exp\{\mathbf{x_j^T}\beta_1\}F_{01}(t_{1j}\mid\eta_1)+\exp\{\mathbf{x_j^T}\beta_2\}F_{02}(t_{2j}\mid\eta_2))^{(1-\gamma)(\delta_{1j}+\delta_{2j})-\gamma\delta_{1j}\delta_{2j}} \\ &\times\prod_{j=1}^{n}[1-\gamma+(1+\exp\{\mathbf{x_j^T}\beta_1\}F_{01}(t_{1j}\mid\eta_1)+\exp\{\mathbf{x_j^T}\beta_2\}F_{02}(t_{2j}\mid\eta_2))^{\gamma}]^{\delta_{1j}\delta_{2j}}. \end{aligned} \quad (5.15)$$

Por meio a este cenário, considere o modelo bivariado de longa duração PVF com distribuição Weibull como função de base para tempos de falha T_K, tendo a f.d representada por $f_{0K}(t_{Kj}\mid\eta_K)$ com vetor de parâmetros $\eta_k = (\eta_1,\eta_2) = (\alpha_k,\phi_k)$, onde $\exp\{\alpha_k\}$ é o parâmetro de escala e ϕ_k é o parâmetro de forma com $k = 1,2$. Sob essa parametrização, a função de distribuição acumulada é dada por: $F(t_k\mid\eta_k) = 1-\exp\{-\exp\{\alpha_K\}t_K^{\phi_k}\}$, com $\alpha_k\in\mathbb{R}$ e $\phi_k > 0$. Como consequência, para a realização da investigação bayesiana, assuma que os parâmetros

γ, $\eta_1 = (\alpha_1, \phi_1)$, $\eta_2 = (\alpha_2, \phi_2)$, β_1 e β_2 são independentes de forma que a densidade a priori conjunta imprópria é definida por

$$\pi\left(\gamma, \eta_1, \eta_2, \beta_1, \beta_2\right) = \pi(\gamma) \prod_{k=1}^{2} \pi\left(\eta_k, \beta_k\right), \tag{5.16}$$

implicando que as prioris de η_k e β_k também são independentes, ou seja

$$\pi\left(\eta_k, \beta_k\right) \propto \pi\left(\eta_k\right) \pi(\beta_k); \quad k = 1, 2.$$

Deste modo, tome uma priori própria para $\pi(\gamma)$ e uma priori uniforme imprópria para $\pi(\beta_k)$, com $k = 1, 2$, além de supor que os parâmetros η_1 e η_2 são independentes e identicamente distribuidos de forma que

$$\pi(\eta_k) = \pi\left(\alpha_k, \phi_k\right) = \pi\left(\alpha_k \mid a_0\right) \pi\left(\phi_k \mid b_0, c_0\right); \quad k = 1, 2,$$

em que

$$\pi\left(\alpha_k\right) \propto \exp\{-a_0 \alpha_k^2\}$$

e

$$\pi\left(\phi_k\right) \propto \phi_k^{b_0 - 1} \exp\{-c_0 \phi_k\}$$

em que a_0, b_0 e c_0 são hiperparâmetros especificados. Assim, ao combinar a densidade a priori (5.16) com a função de verossimilhança (5.15), segue que a densidade a posteriori conjunta é dada por

$$\pi\left(\vartheta^* \mid D_{OBS}\right) \propto L(\vartheta^* \mid D_{OBS})\ \pi(\gamma) \prod_{k=1}^{2} \pi\left(\alpha_k \mid a_0\right) \pi\left(\phi_k \mid b_0, c_0\right) \pi\left(\beta_k\right). \tag{5.17}$$

Note que a densidade a posteriori (5.17) será uma função própria. Este fato é comprovado por meio do teorema a seguir.

Teorema 3. (CANCHO *et al.*, 2022) Seja X_k^* uma matriz $n \times p$ com linhas $\delta_{kj} x_j^T$ para $k = 1, 2$ e $\vartheta = (\gamma, \eta_1, \eta_2, \beta_1, \beta_2)$. Logo, se X_k^* tem posto completo e $\pi(\gamma)$ é própria, então a densidade a posteriori (5.17) é própria.

A demonstração desse teorema não será realizada neste trabalho, contudo é feita de forma similar ao apresentado por Cancho *et al.* (2022). Ainda, novamente, observe que como a densidade a posteriori (5.17) não é uma densidade padrão sendo analiticamente intratável, a inferência realizada é baseada nos métodos MCMC. Assim, será empregado o método HMC cujos cálculos do estudo serão implementados usando o R-Stan, além dos critérios de comparação e seleção de modelos, descritos na Subseção 2.5.4.1, para as próximas análises deste capítulo.

5.3.1 Estudo de simulação

O estudo de simulação é feito para avaliar algumas propriedades frequentistas dos estimadores de Bayes com base na média, DP, viés e na REQM. Para esse propósito, foram realizadas $M = 1.000$ simulações para cada configuração paramétrica tendo tamanhos amostrais diferentes com $n = (500; 800; 1.000)$. Além disso, é adotado o modelo bivariado de longa duração PVF assumindo a distribuição Weibull como função de base para os tempos de falha T_k, com $k = 1, 2$, e cujos parâmetros são definidos como $\eta_1 = (1, 1)$ e $\eta_2 = (1, 1)$.

Ainda, o número de causas de risco do k-ésimo evento de interesse para o j-ésimo indivíduo, com $j = 1, \ldots, n$, são gerados mediante a uma distribuição de Poisson com média $\omega_j \rho_{kj}$ de forma que: $\rho_{kj} = \exp\{\beta_{k1} + \beta_{k2} x_j\}$, no qual, β_{k1} é o intercepto e $\beta_{k1} = \beta_{k2} = -0,5$. Por sua vez, a componente de fragilidade ω_j é gerada por meio de uma distribuição PVF, segundo Hougaard (2000) e conforme descrito na Subseção 2.3.4, para $\gamma = -0,5$ e $\gamma = 0,5$. Já a covariável x_j é gerada por uma distribuição Binomial com probabilidade de sucesso $0,5$.

Os tempos de censura (C_{1j}, C_{2j}) são gerados segundo uma distribuição Uniforme para o intervalo $(0, T_k)$, com T_k sendo definido para controlar a proporção de observações censuradas. Neste estudo, a proporção de observações censuradas foi em média, aproximadamente, 50% e 55%, respectivamente. Ademais, são consideradas prioris independentes para os parâmetros do modelo bivariado de longa duração PVF-Weibull tal que, para $k = 1, 2$, segue

- $\beta_k \propto 1$.
- $\alpha_K \propto \exp\{-\alpha_k^2\}$.
- $\phi_k \propto \phi_k^{-0,9} \exp\{0,005\phi_k\}$.

Além de

- Modelo bivariado de longa duração PVF-Weibull com $\gamma = 0,5$: Assuma que $\gamma \sim N\,(0, 10)$ $I\,(\gamma < 1)$.
- Modelo bivariado de longa duração PVF-Weibull com $\gamma = -0,5$: Suponha que $\gamma \sim N\,(0, 1)\, I\,(\gamma < 1)$.

Os resultados deste estudo de simulação são mostrados na Tabela 23. Estes resultados revelam que a medida que o tamanho da amostra aumenta, as estimativas tendem aos valores verdadeiros dos parâmetros em média. Ainda, devido ao aumento amostral, o módulo do viés, DP e a REQM tendem a zero. Estes são aspectos esperados quando o esquema de estimativa está funcionando corretamente.

Tabela 23 – Estimativas da média, DP, viés e REQM de Bayes dos parâmetros do modelo bivariado de longa duração PVF com a presença de covariàveis para $\gamma = -0,5$, $\gamma = 0,5$, $\alpha_1 = \alpha_2 = \phi_1 = \phi_2 = 1$ e $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = -0,5$

			$\gamma = -0,5$				$\gamma = 0,5$			
n		Parâmetro	Média	DP	Viés	REQM	Média	DP	Viés	REQM
500		α_1	1,037	0,079	0,037	0,079	1,008	0,070	0,008	0,070
		ϕ_1	1,018	0,072	0,018	0,072	1,015	0,083	0,015	0,084
	T_1	β_{11}	-0,495	0,139	0,005	0,137	-0,507	0,142	-0,007	0,140
		β_{12}	-0,539	0,181	-0,039	0,184	-0,465	0,174	0,035	0,176
		α_2	0,970	0,161	-0,030	0,163	0,997	0,147	-0,003	0,145
	T_2	ϕ_2	0,918	0,201	-0,082	0,216	0,950	0,277	-0,050	0,279
		β_{21}	-0,450	0,167	0,050	0,173	-0,502	0,171	-0,002	0,170
		β_{22}	-0,411	0,239	0,089	0,253	-0,472	0,178	0,028	0,179
		γ	-1,107	0,374	-0,607	0,364	0,434	0,188	-0,066	0,187
800		α_1	1,008	0,052	0,008	0,052	0,991	0,054	-0,009	0,054
	T_1	ϕ_1	1,014	0,059	0,014	0,060	1,022	0,067	0,022	0,070
		β_{11}	-0,468	0,126	0,032	0,129	-0,495	0,098	0,005	0,097
		β_{12}	-0,466	0,134	0,034	0,137	-0,483	0,108	0,017	0,109
		α_2	0,935	0,141	-0,065	0,145	0,996	0,112	-0,004	0,111
	T_2	ϕ_2	0,936	0,177	-0,064	0,177	1,016	0,174	0,016	0,173
		β_{21}	-0,481	0,149	0,009	0,148	-0,512	0,136	-0,012	0,135
		β_{22}	-0,487	0,186	0,013	0,184	-0,511	0,155	-0,011	0,154
		γ	-0,678	0,202	-0,178	0,254	0,448	0,146	-0,052	0,150
1.000		α_1	1,003	0,048	0,005	0,049	1,002	0,045	0,002	0,045
	T_1	ϕ_1	1,011	0,059	0,011	0,059	0,997	0,061	-0,003	0,060
		β_{11}	-0,493	0,093	0,007	0,094	-0,509	0,087	-0,010	0,087
		β_{12}	-0,487	0,120	0,033	0,124	-0,493	0,106	0,016	0,111
		α_2	0,975	0,115	-0,025	0,116	0,992	0,087	-0,018	0,088
	T_2	ϕ_2	0,954	0,173	-0,046	0,177	0,997	0,171	-0,028	0,181
		β_{21}	-0,496	0,128	0,007	0,127	-0,506	0,132	0,008	0,133
		β_{22}	-0,498	0,146	-0,018	0,146	-0,509	0,121	-0,009	0,126
		γ	-0,568	0,149	-0,068	0,162	0,471	0,135	-0,029	0,137

Fonte: Elaborada pelo autor.

5.4 Aplicação: dados de *churn* de clientes brasileiros

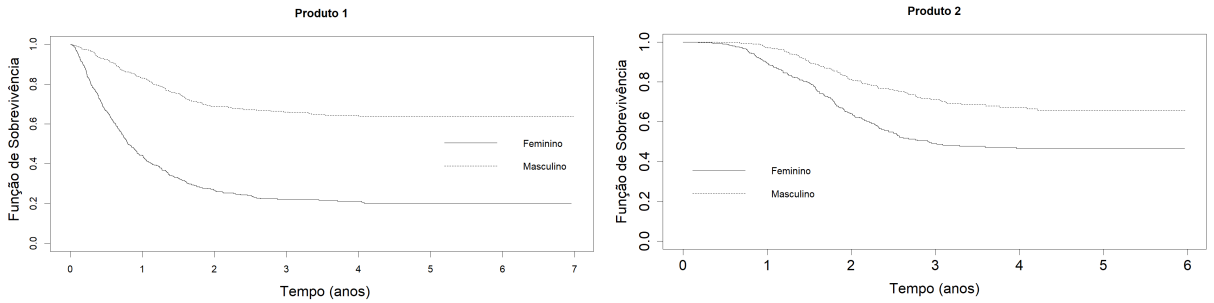
5.4.1 Descrição dos dados

O termo *churn* de clientes é também conhecido como rotatividade de clientes *(customer chur)* ou desistência de clientes, sendo uma grande preocupação para empresas e prestadores de serviços. O foco central deste cenário é denominado como *churn* voluntário que ocorre quando, por decisão pessoal do cliente, este decide migrar para outra empresa ou prestador de serviço. Neste contexto, para os dados de *churn* de clientes brasileiros é observado dois tempos em anos até o *churn*, T_1 e T_2, em dois produtos de cartão de crédito, Produto 1 e Produto 2, durante um período de acompanhamento de 7 anos para ambos produtos. O conjunto de dados possui 900 clientes com taxas de censura sendo de, aproximadamente, 52% e 73% para cada um dos produtos, respectivamente. Além disso, estes dados retratam um ensaio realizado, sendo cedido

pela Faculdade de Engenharia de Bauru da Universidade Estadual Paulista. Contudo, são dados confidenciais e serão mantidos em sigilo, conforme as normas éticas e regulatórias pertinentes.

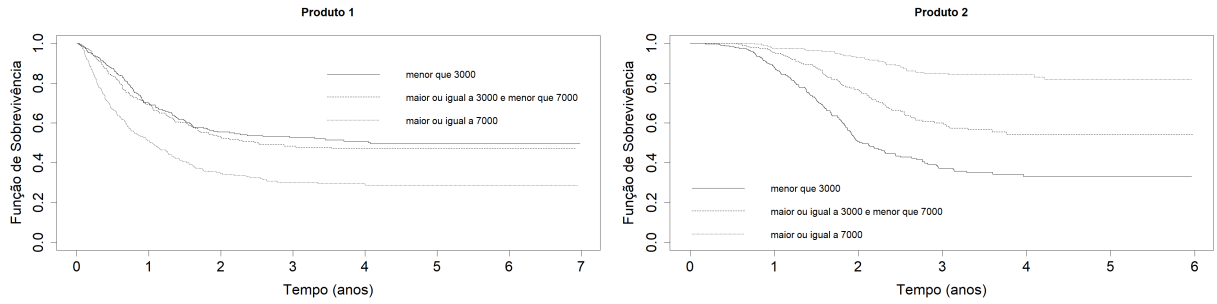
Deste modo, o objetivo desta análise é verificar se existe alguma associação entre os tempos até o *churn* de clientes dos produto 1 e 2, além de investigar se há diferença na proporção de clientes fiéis (taxa de cura) entre estes produtos e avaliar como as covariàveis podem impactar nesta taxa de cura. Observe que as Figuras 28 e 29 exibem as estimativas de Kaplan-Meier da função de sobrevivência para estes dados. Note que há a existência de uma proporção de clientes fiéis tanto ao considerar os níveis de renda, quanto a orientação sexual determinada pelos clientes entrevistados e definidas por eles como homens ou mulheres. Contudo, ao verificar a Figura 28, nota-se que cliente que se definem como mulheres tendem a desistir com mais facilidade dos produtos 1 e 2 ao se comparar com os cliente que se caracterizam como homens. Além disso, pela Figura 29, é evidenciado que pessoas que possuem a renda maior ou igual a R$ 7.000,00 desistem do produto 1 com mais frequência ao se comparar aos demais, mas ao considerar o produto 2 é constatado que esta desistência ocorre, geralmente, com pessoas que possuem a renda inferior a R$ 3.000,00.

Figura 28 – Estimativa de Kaplan-Meier da função de sobrevivência para os dados de o *churn* de clientes do produto 1 e 2, respectivamente, considerando a orientação sexual determinada pelos clientes entrevistados.



Fonte: Elaborada pelo autor.

Figura 29 – Estimativa de Kaplan-Meier da função de sobrevivência para os dados de o *churn* de clientes do produto 1 e 2, respectivamente, considerando a renda dos clientes.



Fonte: Elaborada pelo autor.

5.4.2 Análise dos dados

Para este estudo são consideradas como variàveis de resposta os tempos até o abandono dos produtos 1 e 2. Ainda, foram observadas as seguintes variàveis, para $j = 1, \ldots, 900$:

- t_{kj} : Representa o tempo observado (em anos) para o k-ésimo produto com $k = 1, 2$.
- x_{1j} : Representa o tipo de gênero (0= feminino (51%); 1 = masculino (49%)).
- x_{2j} : Representa os níveis de renda dos clientes (1 = renda < 3.000 ; 2 = $3.000 \leq$ renda < 7.000, 3 = renda ≥ 7.000).

Note que entender a dinâmica de rotatividade dentro do setor financeiro é fundamental para se tomar decisões e manter a satisfação e a fidelidade do cliente, consequentemente, o foco deste estudo não é apenas identificar os padrões de rotatividade, mas também modelá-los usando a distribuição Weibull para prever o futuro comportamento de rotatividade e informar decisões estratégicas de negócios. Desta forma, para este conjunto de dados, é ajustado o modelo descrito na Seção 5.2 tendo todas as covariàveis no parâmetro ρ_{kj}, para $k = 1, 2$ e $j = 1, \ldots, 900$, com

$$\log(\rho_{kj}) = \beta_{k,intercepto} + \beta_{k,genero_1} x_{j,genero_1} + \beta_{k,renda_2} x_{j,renda_2} + \beta_{k,renda_3} x_{j,renda_3}; \qquad (5.18)$$

em que os efeitos das covariàveis são definidos por meio de variàveis fictícias de forma que

$$x_{j,renda_2} = \begin{cases} 1, & \text{se } 3.000 \leq \text{renda} < 7.000; \\ 0, & \text{caso contrário.} \end{cases} \qquad x_{j,renda_3} = \begin{cases} 1, & \text{se renda} \geq 7.000; \\ 0, & \text{caso contrário.} \end{cases}$$

$$x_{j,genero_1} = \begin{cases} 1, & \text{se o gênero é masculino;} \\ 0, & \text{caso contrário.} \end{cases}$$

Para o modelo proposto são adotadas a densidade a posteriori (5.17) e as seguintes priores independentes para os cálculos bayesianos: $\gamma \sim N\,(0, 10)\; I\,(\gamma < 1)$, $\beta_k \propto 1$, $\alpha_K \propto \exp\{-\alpha_k^2\}$ e $\phi_k \propto \phi_k^{-0,9} \exp\{0,005\phi_k\}$ para $k = 1, 2$. Os cálculos bayesianos foram baseados em amostras HMC obtidas de quatro cadeias independentes com 10.000 observações para cada parâmetro. Para eliminar o efeito dos valores iniciais, as primeiras 5.000 iterações foram desconsideradas, além de considerar um espaçamento de tamanho 5 para evitar problemas de correlação, consequentemente, obtendo uma amostra de tamanho 4.000.

Este conjunto de dados é analisado assumindo as distribuições de fragilidade Gama, IG e PVF. Para comparar esses modelos é estimada a verossimilhança marginal por meio do *bridge sampling*, realizado mediante as funções (2.37) e (2.39). As estimativas para estas três distribuições são exibidas na Tabela 24. Como consequência, é ressaltado que o modelo que melhor se ajusta aos dados de *churn* é o modelo bivariado de longa duração PVF. Assim, este modelo é tomado para analisar os dados de rotatividade de clientes.

Tabela 24 – Probabilidades do modelo de regressão bivariado a posteriori mediante as verossimilhanças marginais.

	Gama	IG	PVF
Probabilidade	0,272	0,198	**0,530**

Fonte: Elaborada pelo autor.

As estimativas bayesianas para as distribuições de fragilidade Gama, IG e PVF, bem como, as médias a posteriores e 95% de confiança dos intervalos HPD para os parâmetros γ, (α_k, ϕ_k) e $\beta = (\beta_1^T, \beta_2^T)^T$ com $k = 1, 2$ são mostradas na tabela a seguir.

Tabela 25 – As estimativas bayesianas, bem como, as médias a posteriores e 95% de confiança dos intervalos HPD para os parâmetros no conjunto de dados de bancos brasileiros considerando as distribuições Gama, IG e PVF.

		Gama		IG		PVF	
	Parâmetros	Média	Intervalos HPD (95%)	Média	Intervalos HPD (95%)	Média	Intervalos HPD (95%)
	ϕ_1	1,587	(1,463 ; 1,715)	1,596	(1,478 ; 1,724)	1,581	(1,461 ; 1,704)
	α_1	-0,885	(-1,114 ; -0,679)	-0,843	(-1,041 ; -0,662)	-0,874	(-1,088 ; -0,678)
T_1	$\beta_{1,0}$	0,940	(0,627 ; 1,277)	0,995	(0,663 ; 1,342)	0,922	(0,632 ; 1,215)
	$\beta_{1,genero_1}$	-1,960	(-2,261 ; -1,655)	-1,931	(-2,227 ; -1,644)	-1,949	(-2,243 ; -1,656)
	$\beta_{1,renda_2}$	0,134	(-0,183 ; 0,455)	0,156	(-0,167 ; 0,473)	0,137	(-0,180 ; 0,454)
	$\beta_{1,renda_3}$	1,186	(0,881 ; 1,511)	1,111	(0,819 ; 1,421)	1,178	(0,857 ; 1,501)
	ϕ_2	3,069	(2,779 ; 3,361)	3,117	(2,825 ; 3,411)	3,061	(2,795 ; 3,341)
	α_2	-2,865	(-3,201 ; -2,553)	-2,911	(-3,237 ; -2,595)	-2,854	(-3,155 ; -2,562)
T_2	$\beta_{2,0}$	1,262	(0,930 ; 1,615)	1,347	(0,999 ; 1,701)	1,242	(0,937 ; 1,560)
	$\beta_{2,genero_1}$	-1,023	(-1,346 ; -0,711)	-1,047	(-1,361 ; -0,739)	-1,018	(-1,333 ; -0,715)
	$\beta_{2,renda_2}$	-0,987	(-1,346 ; -0,641)	-0,949	(-1,309 ; -0,602)	-0,973	(-1,325 ; -0,628)
	$\beta_{2,renda_3}$	-2,398	(-2,870 ; -1,950)	-2,441	(-2,934 ; -1,988)	-2,383	(-2,843 ; -1,932)
	θ	0,978	(0,701 ; 1,286)	1,763	(1,086 ; 2,673)	0,959	(0,787 ; 1,121)
	γ	-	-	-	-	0,041	(-0,121 ; 0,213)

Fonte: Elaborada pelo autor.

Pela Tabela 25, é notado que para o modelo PVF, as estimativas de Bayes de todas as covariàveis afetam significativamente a proporção de clientes fiéis para os produtos 1 e 2 a um nível de significância de 5%. Ainda, note que as estimativas $\beta_{1,genero1}, \beta_{2,genero1} < 0$ implicando que a proporção de clientes fiéis masculinos é maior ao se comparar com as clientes femininas. No entanto, como as estimativas associados ao produto 1 são $\beta_{1,renda1}, \beta_{1,renda2} > 0$ observa-se que a proporção clientes fiéis diminuirá para clientes que apresentam a renda maior ou igual a R$ 3.000,00. Isso ocorre de maneira inversa ao considerar o produto 2, ou seja, como $\beta_{2,renda1}, \beta_{2,renda2} < 0$, haverá uma aumento na proporção clientes fiéis para clientes que apresentam a renda maior ou igual a R$ 3.000,00. Ainda, ao comparar os níveis de renda é visto que para o Produto 1, a proporção de clientes fiéis com rendas abaixo de R$ 3.000,00 é semelhante àquelas com renda entre R$ 3.000,00 e R$ 7.000,00, enquanto para o Produto 2, a proporção de clientes fiéis com rendas abaixo de R$ 3.000,00 é menor do que aquelas com rendas entre R$ 3.000,00 e R$ 7.000,00. Para finalizar, $\beta_{k,renda_3} > 0$ mostra que os clientes fiéis

do Produto 2 com renda entre R$ 3.000,00 e R$ 7.000,00 são mais prevalentes do que os clientes com renda superior a R$ 7.000,00. Essa relação é invertida para o Produto 1, pois quanto maiores os valores de renda, maior será a proporção de clientes fiéis no estudo.

Além disso, a probabilidade dos clientes permaneçam leais ao k-ésimo produto $(k = 1,2)$ após $t > 0$ anos do período de acompanhamento é dada por

$$\mathbb{P}_k(t) = \mathbb{P}(Z_k = 0 \mid T_k > t) = \left[\frac{\exp\{(1+\rho_k(1-\exp\{-e^{\alpha_K} t_K^{\phi_k}\}))^{\gamma}\}}{\exp\{(1+\rho_k)^{\gamma}\}}\right]^{\frac{1}{\gamma}}; \quad k = 1,2,$$

onde ρ_k é dado por (5.18). Note que quando $t = 0$ então $\mathbb{P}_k(0) = \rho_{0k}$, correspondendo a taxa de clientes fiéis para o produto 1 ou 2 após o período de observação. Mediante a isso, é estimada a proporção de clientes fiéis para seis clientes hipotéticos A, B, C, D, E e F após $t = 3$ anos do período de observação e considerando todas as covariàveis. Estas estimativas são mostradas na Tabela 26.

Tabela 26 – Estimativa de Bayes da probabilidade de clientes fiéis aos produtos 1 e 2.

Consumidores	Renda	Gênero	Produtos 1 e 2 $Pr_{00}(3)$	Produto 1 $Pr_{01}(3)$	Produto 2 $Pr_{02}(3)$
A	$<$ 3.000	Feminino	0,861	0,929	0,844
			(0,799 ; 0,908)	(0,886 ; 0,960)	(0,770 ; 0,901)
B		Masculino	0,892	0,926	0,888
			(0,848 ; 0,927)	(0,881 ; 0,959)	(0,835 ; 0,930)
C	$\leq$ 3.000 e $<$ 7.000	Feminino	0,904	0,908	0,953
			(0,847 ; 0,945)	(0,850 ; 0,949)	(0,925 ; 0,973)
D		Masculino	0,892	0,975	0,891
			(0,845 ; 0,929)	(0,956 ; 0,987)	(0,839 0,933)
E	$\geq$ 7.000	Feminino	0,929	0,972	0,938
			(0,901 ; 0,952)	(0,952 ; 0,985)	(0,905; 0,963)
F		Masculino	0,940	0,948	0,980
			(0,907 ; 0,962)	(0,914 ; 0,971)	(0,967 ;0,989)

Fonte: Elaborada pelo autor.

A Tabela 27 apresenta as estimativas de Bayes da proporção de clientes fiéis aos produtos 1 e 2 (p_{00}), ao produto 1 (p_{01}) e ao produto 2 (p_{02}). Por exemplo, para clientes com fatores de risco idênticos ao cliente A, a probabilidade de não abandonar ambos os produtos é $p_{00} = 0,133$, já para os produtos 1 e 2 são $p_{01} = 0,278$ e $p_{02} = 0,216$, respectivamente. Por outro lado, as probabilidades desses clientes continuarem com os produtos 1 e 2 após 3 anos são $Pr_{01}(3) = 0,929$ e $Pr_{02}(3) = 0,844$, respectivamente, e para ambos os produtos é de $Pr_{00}(3) = 0,861$. Ainda, em um contexto geral, nota-se que a proporção de clientes que continuam com ambos os produtos é menor para clientes do sexo feminino do que para clientes do sexo masculino. Essa relação de proporções de clientes fiéis (entre mulheres e homens) vale para o produto 2 individualmente. Além disso, observe que a fração de clientes homens que não cancelaram o produto 2 aumenta conforme a renda aumenta.

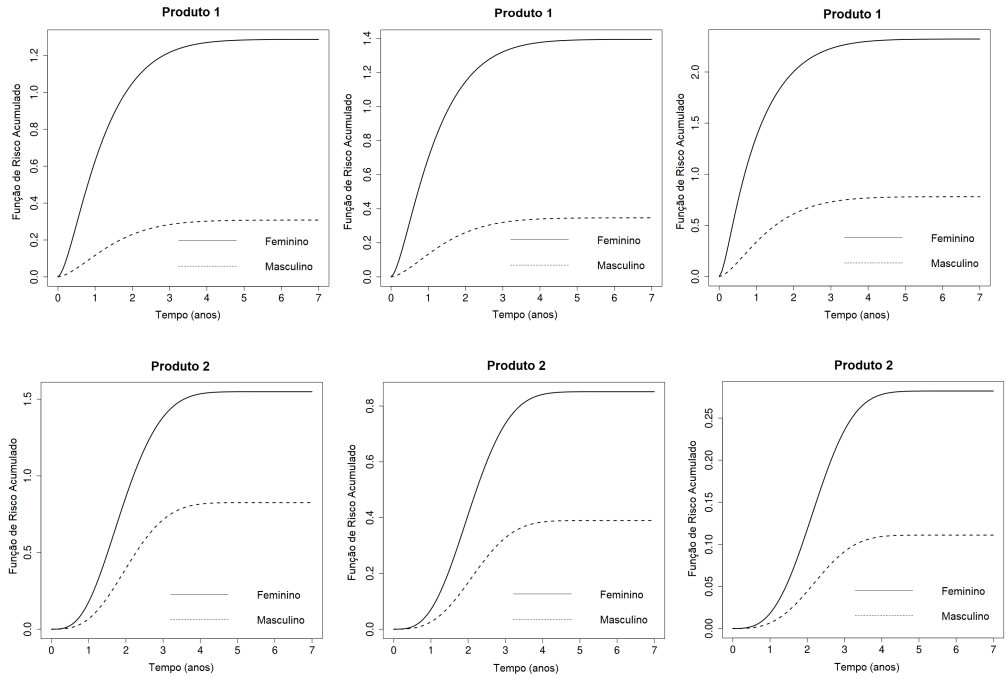
Tabela 27 – Estimativa de Bayes da proporção de clientes fiéis aos produtos 1 e 2.

Consumidores	Renda	Gênero	Produtos 1 e 2 $p_{00}(3)$	Produto 1 $p_{01}(3)$	Produto 2 $p_{02}(3)$
A	< 3.000	Feminino	0,133	0,278	0,216
			(0,094; 0,176)	(0,224; 0,333)	(0,163 ; 0,275)
B		Masculino	0,183	0,250	0,428
			(0,139 ; 0,230)	(0,196 ; 0,310)	(0,350 ; 0,508)
C	≤ 3.000 e < 7.000	Feminino	0,097	0,101	0,753
			(0,065 ; 0,133)	(0,068 ;0,138)	(0,676 0,823)
D		Masculino	0,377	0,734	0,440
			(0,314 ; 0,440)	(0,675 ; 0,789)	(0,363 ; 0,515)
E	≥ 7.000	Feminino	0,525	0,706	0,677
			(0,464 ; 0,587)	(0,643 ; 0,765)	(0,600 ; 0,748)
F		Masculino	0,433	0,458	0,894
			(0,370 ; 0,497)	(0,390 ; 0,525)	(0,849 ; 0,930)

Fonte: Elaborada pelo autor.

A Figura 30 apresenta a função de risco acumulado a posteriori marginal estratificada por gênero para cada nível de renda, sendo que o painel direito retrata os clientes que ganham menos de R$ 3.000,00, o painel do meio os clientes que ganham entre R$ 3.000,00 e R$ 7.000,00 e o painel esquerdo para clientes que ganham mais de R$ 7.000,00 para ambos os produtos 1 e 2. Observe que, independentemente da renda, o risco de abandonar o produto 1 ou 2 é maior entre clientes do sexo feminino do que entre clientes do sexo masculino. Ainda, nota-se que o risco para clientes do sexo feminino abandonarem o produto 1 aumenta com o nível de renda, enquanto essa relação é invertida para o produto 2. Ademais, essas diferenças são menos presenciadas entre clientes do sexo masculino.

Figura 30 – Função de risco acumulado a posteriori marginal estratificada por gênero para cada nível de renda para ambos os produtos 1 e 2.



Fonte: Elaborada pelo autor.

5.5 Conclusão

Neste capítulo é apresentado um modelo de sobrevivência bivariado incorporando fragilidade, onde uma distribuição de Poisson é usada para explicar a dispersão não observada e a componente de fragilidade segue uma distribuição PVF, sendo baseado no trabalho de Bedia (2022). Este modelo bivariado inclui os modelos de Chen, Ibrahim e Sinha (2002) e Cancho *et al.* (2022), além de ser estendido para incluir o modelo de Cancho, Rodrigues e Castro (2011) e modelos de regressão, estando estruturado na Seção 5.2.

A abordagem inferencial construída na Seção 5.3 é baseada em métodos bayesianos mediante ao uso do método HMC implementado no R-Stan, no qual, o estudo de simulação revela que este modelo possui boas propriedades dos estimadores de Bayes. Para finalizar, na Seção 5.4, a importância do modelo bivariado de longa duração PVF-Weibull é ilustrada por meio de uma aplicação a uma conjunto de dados reais, fornecendo bases estratégicas para a redução de *churn* e o aumento da lealdade dos clientes.

CAPÍTULO

6

CONSIDERAÇÕES FINAIS E PESQUISAS FUTURAS

6.1 Considerações finais

Os estudos realizados indicam que os modelos PVF e MEP são eficazes na análise de dados de sobrevivência. O modelo PVF se destaca pela sua flexibilidade em descrever funções amplamente utilizadas na literatura, enquanto o MEP é uma alternativa semiparamétrica às distribuições paramétricas, sendo útil para ajustar funções de taxa de falha com diferentes formas. No entanto, a escolha de uma partição adequada para o eixo dos tempos no modelo MEP é desafiadora, pois pode afetar a convergência dos modelos. Devido a isso, critérios de seleção de modelos se mostraram uma ferramenta importante para determinar a melhor partição durante a análise dos dados.

Os modelos de fragilidade resultantes da composição dos modelos PVF e MEP são atraentes por sua versatilidade paramétrica, permitindo modificações nos parâmetros e resultando em modelos impróprios. Isso possibilitou a criação dos modelos defeituosos e de longa duração univariados, apresentados nos capítulos 3 e 4, cuja eficácia foi comprovada por meio de aplicações a dados reais. Em particular, no estudo sobre AIDS/HIV, o modelo de longa duração teve um ajuste mais adequado em comparação ao modelo defeituoso, com estimativas de LPML de -89,89 e -89,94, respectivamente. Essa vantagem do modelo de longa duração sob o modelo defeituoso também é presenciada durante a comparação dos pacientes hipotéticos, uma vez que, os pacientes modelados com o modelo de longa duração apresentaram uma maior probabilidade de não morrer. Contudo, ambos os modelos mostraram que as covariáveis idade e POI são significativas, e os critérios de comparação indicaram que o modelo PVF-Exponencial foi o melhor modelo que se ajustou para esses dados. Esses dados também foram usados em estudos de Cancho *et al.* (2021), obtendo um LPML de -90,76, um valor ligeiramente superior aos encontrados neste trabalho. Entretanto, as pequenas diferenças entre os valores de LPML indicam que os três modelos não

apresentam variações significativas na modelagem dos dados sobre AIDS/HIV.

No modelo bivariado, a escolha de uma função de base que seja compatível com modelos de longa duração e que se ajuste a um conjunto de dados adequadamente é desafiadora. Contudo, a distribuição Weibull se destaca como uma opção versátil, pois pode descrever curvas com diferentes formas, tornando-se uma ferramenta eficaz na modelagem de dados de sobrevivência.

Por fim, a abordagem inferencial baseada em métodos bayesianos se mostrou apropriada, pois os estudos de simulação demonstraram boas propriedades dos estimadores de Bayes. Além disso, a utilização do método HMC via R-Stan foi eficaz, promissora e rápida para a análise de dados, oferecendo uma linguagem de fácil compreensão, o que facilita a programação dos algoritmos, além de ter muitos recursos teóricos e práticos sobre esta linguagem e uma comunidade muito participativa em fóruns de discussões.

6.2 Sugestões para pesquisas futuras

- Propor um modelo para determinar a função de risco base usando os Polinômios de Bernstein.
- Estender os modelos propostos para considerar dados longitudinais.
- Propor novos modelos de sobrevivência induzidos por fragilidade utilizando outras famílias de distribuições, além das famílias PVF e MEP usadas nesta tese.

REFERÊNCIAS

AALEN, O. O. Modelling heterogeneity in survival analysis by the compound poisson distribution. **The Annals of Applied Probability**, JSTOR, p. 951–972, 1992. Citado nas páginas 35, 38, 42 e 46.

AMORIM, W. N. de. **Verossimilhança hierárquica em modelos de fragilidade**. Tese (Doutorado) — Universidade de São Paulo, 2014. Citado na página 30.

BALAN, T. A.; PUTTER, H. A tutorial on frailty models. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 29, n. 11, p. 3424–3454, 2020. Citado nas páginas 35, 37 e 38.

BALKA, J.; DESMOND, A. F.; MCNICHOLAS, P. D. Review and implementation of cure models based on first hitting times for wiener processes. **Lifetime data analysis**, Springer, v. 15, p. 147–176, 2009. Citado na página 61.

BANBETA, A.; SEYOUM, D.; BELACHEW, T.; BIRLIE, B.; GETACHEW, Y. Modeling time-to-cure from severe acute malnutrition: application of various parametric frailty models. **Archives of Public Health**, BioMed Central, v. 73, n. 1, p. 1–8, 2015. Citado na página 29.

BEDIA, E. C. Modelos de sobrevivência bivariados induzidos por fragilidade. Universidade Federal de São Carlos, 2022. Citado nas páginas 31, 35, 39, 40, 42, 44, 45, 46, 62, 107, 108, 109, 110, 111 e 121.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American statistical association**, JSTOR, p. 501–515, 1952. Citado nas páginas 31, 61, 92, 94 e 107.

BORGES, P. Em algorithm-based likelihood estimation for a generalized gompertz regression model in presence of survival data with long-term survivors: an application to uterine cervical cancer data. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 9, p. 1712–1722, 2017. Citado na página 62.

BRANDÃO, M.; LEÃO, J.; GALLARDO, D. I.; BOURGUIGNON, M. Cure rate models for heterogeneous competing causes. **Statistical Methods in Medical Research**, SAGE Publications Sage UK: London, England, v. 32, n. 9, p. 1823–1841, 2023. Citado na página 87.

BRITO, A. D. L.; JÚNIOR, S. F. X.; MENDONÇA, E. B. D.; XAVIER, É. F. M.; SANTOS, T. T. D. M.; OLIVEIRA, T. A. de. Ajuste de modelos de fragilidade e riscos proporcionais aplicados a dados de retinopatia diabética. **Research, Society and Development**, v. 9, n. 8, p. e478985691–e478985691, 2020. Citado na página 29.

BUNYATISAI, W.; PRASITWATTANASEREE, S.; INGSRISAWANG, L. Assessing frailty survival models in describing variations caused by unobserved covariates. **Chiang Mai J. Sci**, v. 44, p. 1191–1200, 2017. Citado na página 29.

CALSAVARA, V. F.; RODRIGUES, A. S.; TOMAZELLA, V. L. D.; CASTRO, M. de. Frailty models power variance function with cure fraction and latent risk factors negative binomial. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 46, n. 19, p. 9763–9776, 2017. Citado nas páginas 31 e 45.

CANCHO, V. G.; BARRIGA, G. D.; CORDEIRO, G. M.; ORTEGA, E. M.; SUZUKI, A. K. Bayesian survival model induced by frailty for lifetime with long-term survivors. **Statistica Neerlandica**, Wiley Online Library, 2021. Citado nas páginas 30, 31, 42, 69, 70, 76, 88, 89, 90, 91, 92, 107, 123, 141 e 142.

CANCHO, V. G.; BEDIA, E. C.; CORDEIRO, G. M.; PRATAVIERA, F.; ORTEGA, E. M.; SANTO, A. P. A survival regression with cure fraction applied to cervical cancer. **Computational Statistics**, Springer, v. 38, n. 1, p. 403–418, 2023. Citado na página 87.

CANCHO, V. G.; RODRIGUES, J.; CASTRO, M. de. A flexible model for survival data with a cure rate: a bayesian approach. **Journal of Applied Statistics**, Taylor & Francis, v. 38, n. 1, p. 57–70, 2011. Citado nas páginas 59, 108 e 121.

CANCHO, V. G.; SUZUKI, A. K.; BARRIGA, G. D.; SANTO, A. P. d. E. A multivariate survival model induced by discrete frailty. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 51, n. 11, p. 6572–6590, 2022. Citado nas páginas 107, 108, 109, 113 e 121.

CANCHO, V. G.; ZAVALETA, K. E.; MACERA, M. A.; SUZUKI, A. K.; LOUZADA, F. A bayesian cure rate model with dispersion induced by discrete frailty. **Communications for Statistical Applications and Methods**, The Korean Statistical Society, v. 25, n. 5, p. 471–488, 2018. Citado nas páginas 28, 30, 42, 56, 59, 77 e 107.

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. Citado na página 82.

CHATTERJEE, N.; SHIH, J. A bivariate cure-mixture approach for modeling familial association in diseases. **Biometrics**, Oxford University Press, v. 57, n. 3, p. 779–786, 2001. Citado na página 107.

CHATZILENA, A.; LEEUWEN, E. van; RATMANN, O.; BAGUELIN, M.; DEMIRIS, N. Contemporary statistical inference for infectious disease models using stan. **Epidemics**, Elsevier, v. 29, p. 100367, 2019. Citado na página 33.

CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. Bayesian inference for multivariate survival data with a cure fraction. **Journal of Multivariate Analysis**, Elsevier, v. 80, n. 1, p. 101–126, 2002. Citado nas páginas 108, 109 e 121.

CHEN, P.; ZHANG, J.; ZHANG, R. Estimation of the accelerated failure time frailty model under generalized gamma frailty. **Computational Statistics & Data Analysis**, Elsevier, v. 62, p. 171–180, 2013. Citado na página 29.

CLAYTON, D. G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. **Biometrika**, Oxford University Press, v. 65, n. 1, p. 141–151, 1978. Citado nas páginas 42 e 110.

COWLES, M. K.; CARLIN, B. P. Markov chain monte carlo convergence diagnostics: a comparative review. **Journal of the American Statistical Association**, Taylor & Francis, v. 91, n. 434, p. 883–904, 1996. Citado nas páginas 78, 83 e 101.

DEMARQUI, F. N. Modelo exponencial por partes via modelos partição produto. Universidade Federal de Minas Gerais, 2006. Citado na página 32.

____. Uma classe mais flexível de modelos semiparamétricos para dados de sobrevivência. Universidade Federal de Minas Gerais, 2010. Citado nas páginas 32, 35 e 50.

DUANE, S.; KENNEDY, A. D.; PENDLETON, B. J.; ROWETH, D. Hybrid monte carlo. **Physics letters B**, Elsevier, v. 195, n. 2, p. 216–222, 1987. Citado na página 51.

DUCHATEAU, L.; JANSSEN, P. Frailty distributions. **The Frailty Model**, Springer, p. 117–197, 2008. Citado nas páginas 35, 41 e 42.

FORTES, R. d. S. R. Modelo de sobrevivência birnbaum-saunders com fragilidade espacial. 2020. Citado nas páginas 30 e 42.

GALLARDO, D. I.; BOLFARINE, H.; PEDROSO-DE-LIMA, A. C. Promotion time cure rate model with bivariate random effects. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 45, n. 2, p. 603–624, 2016. Citado na página 107.

GAMERMAN, D. Dynamic bayesian models for survival data. **Journal of the Royal Statistical Society Series C: Applied Statistics**, Oxford University Press, v. 40, n. 1, p. 63–79, 1991. Citado nas páginas 73 e 99.

GASPARINI, A.; CLEMENTS, M. S.; ABRAMS, K. R.; CROWTHER, M. J. Impact of model misspecification in shared frailty survival models. **Statistics in medicine**, Wiley Online Library, v. 38, n. 23, p. 4477–4502, 2019. Citado na página 30.

GELMAN, A.; LEE, D.; GUO, J. Stan: A probabilistic programming language for bayesian inference and optimization. **Journal of Educational and Behavioral Statistics**, Sage Publications Sage CA: Los Angeles, CA, v. 40, n. 5, p. 530–543, 2015. Citado na página 33.

GIUSSANI, A.; BONETTI, M. Marshall–olkin frailty survival models for bivariate right-censored failure time data. **Journal of Applied Statistics**, Taylor & Francis, 2019. Citado nas páginas 29 e 107.

GOMPERTZ, B. Xxiv. on the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to francis baily, esq. frs &c. **Philosophical transactions of the Royal Society of London**, The Royal Society London, n. 115, p. 513–583, 1825. Citado na página 62.

GRONAU, Q. F.; SARAFOGLOU, A.; MATZKE, D.; LY, A.; BOEHM, U.; MARSMAN, M.; LESLIE, D. S.; FORSTER, J. J.; WAGENMAKERS, E.-J.; STEINGROEVER, H. A tutorial on bridge sampling. **Journal of mathematical psychology**, Elsevier, v. 81, p. 80–97, 2017. Citado nas páginas 35, 56, 57, 58 e 59.

GRONAU, Q. F.; SINGMANN, H.; WAGENMAKERS, E.-J. bridgesampling: An r package for estimating normalizing constants. **arXiv preprint arXiv:1710.08162**, 2017. Citado nas páginas 35, 56, 57, 58 e 59.

GRONAU, Q. F.; WAGENMAKERS, E.-J.; HECK, D. W.; MATZKE, D. A simple method for comparing complex models: Bayesian model comparison for hierarchical multinomial processing tree models using warp-iii bridge sampling. **Psychometrika**, Springer, v. 84, p. 261–284, 2019. Citado nas páginas 35 e 59.

GURMU, S. E. Assessing survival time of women with cervical cancer using various parametric frailty models: a case study at tikur anbessa specialized hospital, addis ababa, ethiopia. **Annals of Data Science**, Springer, v. 5, n. 4, p. 513–527, 2018. Citado na página 29.

HARTMANN, M. Métodos de monte carlo hamiltoniano na inferência bayesiana não-paramétrica de valores extremos. Universidade Federal de São Carlos, 2015. Citado nas páginas 33, 35, 50, 51, 52, 53, 54, 55 e 73.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. Oxford University Press, 1970. Citado nas páginas 51, 52 e 55.

HOUGAARD, P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. **Biometrika**, Oxford University Press, v. 71, n. 1, p. 75–83, 1984. Citado nas páginas 29 e 43.

____. Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, Oxford University Press, v. 73, n. 2, p. 387–396, 1986. Citado na página 29.

____. **Analysis of multivariate survival data**. [S.l.]: Springer, 2000. v. 564. Citado nas páginas 35, 45, 46, 63, 64, 72, 74, 88, 89, 97, 108, 114 e 139.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. Bayesian semiparametric models for survival data with a cure fraction. **Biometrics**, Wiley Online Library, v. 57, n. 2, p. 383–388, 2001. Citado nas páginas 32, 35, 46, 50, 55, 56 e 59.

ISIDRO, M. J.; ISIDRO, R. de A.; BRITO, A. de L.; OLIVEIRA, T. A. de. Modelos de fragilidade aplicados a análise de fatores contribuintes na morte de pacientes portadores de leucemia. **Brazilian Journal of Development**, v. 6, n. 8, p. 54802–54820, 2020. Citado na página 29.

JIANG, Z.; CARTER, R. Using hamiltonian monte carlo to estimate the log-linear cognitive diagnosis model via stan. **Behavior Research Methods**, Springer, v. 51, n. 2, p. 651–662, 2019. Citado nas páginas 33 e 73.

KIM, Y.-J. Cure rate model with bivariate interval censored data. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 46, n. 9, p. 7116–7124, 2017. Citado na página 107.

LIMA, C. M.; TOMAZELLA, V. L.; CAMPELO, J. E.; FILHO, L. J.; JUNIOR, W. B.; JUNIOR, S. C. S. Gamma-gompertz shared frailty model for analysis of the time of stay in an anglo-nubian goat herd. **Small Ruminant Research**, Elsevier, v. 199, p. 106368, 2021. Citado na página 32.

LOUZADA, F.; CUMINATO, J. A.; RODRIGUEZ, O. M. H.; TOMAZELLA, V. L.; MILANI, E. A.; FERREIRA, P. H.; RAMOS, P. L.; BOCHIO, G.; PERISSINI, I. C.; JUNIOR, O. A. G. *et al.* Incorporation of frailties into a non-proportional hazard regression model and its diagnostics for reliability modeling of downhole safety valves. **IEEE Access**, IEEE, v. 8, p. 219757–219774, 2020. Citado na página 28.

LUO, Y.; JIAO, H. Using the stan program for bayesian item response theory. **Educational and psychological measurement**, SAGE Publications Sage CA: Los Angeles, CA, v. 78, n. 3, p. 384–408, 2018. Citado na página 33.

MACERA, M. A. C. Modelagem de dados de sobrevivência com eventos recorrentes via fragilidade discreta. Universidade Federal de São Carlos, 2015. Citado nas páginas 30 e 42.

MARTINS, R.; SILVA, G. L.; ANDREOZZI, V. Joint analysis of longitudinal and survival aids data with a spatial fraction of long-term survivors: A bayesian approach. **Biometrical Journal**, Wiley Online Library, v. 59, n. 6, p. 1166–1183, 2017. Citado nas páginas 107 e 108.

MCELREATH, R. **Statistical rethinking: A Bayesian course with examples in R and Stan**. [S.l.]: Chapman and Hall/CRC, 2020. Citado na página 73.

MELLO, J. F. de *et al.* Modelo exponencial por partes para dados de sobrevivência com longa duração. Universidade Federal de Minas Gerais, 2016. Citado na página 32.

MENG, X.-L.; SCHILLING, S. Warp bridge sampling. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 11, n. 3, p. 552–586, 2002. Citado nas páginas 35, 57 e 59.

MENG, X.-L.; WONG, W. H. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. **Statistica Sinica**, JSTOR, p. 831–860, 1996. Citado nas páginas 35, 57, 58 e 59.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **The journal of chemical physics**, American Institute of Physics, v. 21, n. 6, p. 1087–1092, 1953. Citado nas páginas 51 e 55.

METROPOLIS, N.; ULAM, S. The monte carlo method. **Journal of the American statistical association**, JSTOR, p. 335–341, 1949. Citado na página 51.

MOLINA, K. R. C. Modelos de sobrevivência induzidos por fragilidade discreta série de potência zero-modificada. Universidade Federal de São Carlos, 2020. Citado na página 29.

MONACO, J. V.; GORFINE, M.; HSU, L. General semiparametric shared frailty model: Estimation and simulation with frailtysurv. **Journal of statistical software**, NIH Public Access, v. 86, 2018. Citado na página 31.

NG'OMBE, J. N.; LAMBERT, D. M. Using hamiltonian monte carlo via stan to estimate crop input response functions with stochastic plateaus. **Journal of Agriculture and Food Research**, Elsevier, v. 6, p. 100226, 2021. Citado nas páginas 33 e 73.

OLIVEIRA, R. P. de; PERES, M. V. de O.; ACHCAR, J. A.; MARTINEZ, E. Z. A new class of bivariate sushila distributions in presence of right-censored and cure fraction. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 37, n. 1, p. 55–72, 2023. Citado na página 87.

PAIXÃO, R. S. Método zero-variance para monte carlo hamiltoniano aplicado a modelos garch univariados e multivariados. Universidade Federal de São Carlos, 2021. Citado nas páginas 33, 35, 50, 53 e 55.

PARDO, L. **Statistical inference based on divergence measures**. [S.l.]: CRC press, 2018. Citado nas páginas 35 e 59.

PRICE, D. L.; MANATUNGA, A. K. Modelling survival data with a cured fraction using frailty models. **Statistics in medicine**, Wiley Online Library, v. 20, n. 9-10, p. 1515–1527, 2001. Citado na página 107.

RIBEIRO, H. C. M.; TAVARES, V. C. M. Comportamento e particularidades da produção acadêmica do tema "contabilidade gerencial" divulgada na base de dados do isi web of science core collection de 1985 a 2014. **Revista de Educação e Pesquisa em Contabilidade (REPeC)**, v. 11, n. 1, 2017. Citado na página 27.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Wiley Online Library, v. 54, n. 3, p. 507–554, 2005. Citado na página 77.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F. Two new defective distributions based on the marshall–olkin extension. **Lifetime data analysis**, Springer, v. 22, p. 216–240, 2016. Citado nas páginas 61 e 62.

____. A new class of defective models based on the marshall–olkin family of distributions for cure rate modeling. **Computational Statistics & Data Analysis**, Elsevier, v. 107, p. 48–63, 2017. Citado nas páginas 61 e 62.

ROCHA, R.; NADARAJAH, S.; TOMAZELLA, V.; LOUZADA, F.; EUDES, A. New defective models based on the kumaraswamy family of distributions with application to cancer data sets. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 26, n. 4, p. 1737–1755, 2017. Citado nas páginas 61 e 62.

ROCHA, R. F. d. Defective models for cure rate modeling. Universidade Federal de São Carlos, 2016. Citado nas páginas 61 e 62.

RODRIGUES, A. S.; CALSAVARA, V. F.; BERTOLLI, E.; PERES, S. V.; TOMAZELLA, V. L. Bayesian long-term survival model including a frailty term: Application to melanoma data. **Chilean Journal of Statistics (ChJS)**, v. 12, n. 1, 2021. Citado nas páginas 31 e 45.

RODRIGUES, A. S.; CALSAVARA, V. F.; TOMAZELLA, V. L. D. Modeling cure fraction with frailty term in latent risk: a bayesian approach. **arXiv preprint arXiv:1803.08128**, 2018. Citado na página 31.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de; LOUZADA-NETO, F. On the unification of long-term survival models. **Statistics & Probability Letters**, Elsevier, v. 79, n. 6, p. 753–759, 2009. Citado na página 31.

RODRIGUES, J.; CASTRO, M. de; CANCHO, V. G.; BALAKRISHNAN, N. Com–poisson cure rate survival models and an application to a cutaneous melanoma data. **Journal of Statistical Planning and Inference**, Elsevier, v. 139, n. 10, p. 3605–3611, 2009. Citado na página 107.

SANTO, A. P. J. d. E. Modelos de sobrevivência induzidos por fragilidade discreta com fração de cura e riscos proporcionais. Universidade Federal de São Carlos, 2022. Citado na página 72.

SANTOS, M. R. d.; ACHCAR, J. A.; MARTINEZ, E. Z. Bayesian and maximum likelihood inference for the defective gompertz cure rate model with covariates: an application to the cervical carcinoma study. **Ciência e Natura**, v. 39, n. 2, p. 244–258, 2017. Citado na página 62.

SCUDILIO, J.; CALSAVARA, V. F.; ROCHA, R.; LOUZADA, F.; TOMAZELLA, V.; RODRIGUES, A. S. Defective models induced by gamma frailty term for survival data with cured fraction. **Journal of Applied Statistics**, Taylor & Francis, v. 46, n. 3, p. 484–507, 2019. Citado nas páginas 32, 61 e 62.

SIBIM, A. C. **Estimação e diagnóstico na distribuição exponencial por partes em análise de sobrevivência com fração de cura**. Tese (Doutorado) — Universidade de São Paulo, 2011. Citado nas páginas 32, 35, 46, 47, 50, 59 e 77.

SOUZA, D. de; CANCHO, V. G.; RODRIGUES, J.; BALAKRISHNAN, N. Bayesian cure rate models induced by frailty in survival analysis. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 26, n. 5, p. 2011–2028, 2017. Citado nas páginas 29 e 107.

SPIEGELHALTER, D.; THOMAS, A.; BEST, N.; GILKS, W. Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). **MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK**, p. 1–59, 1996. Citado na página 33.

TEAM, R. C. Ra language and environment for statistical computing, r foundation for statistical. **Computing**, 2020. Citado nas páginas 47, 58, 67, 93 e 95.

TESEMA, G. A.; SEIFU, B. L.; TESSEMA, Z. T.; WORKU, M. G.; TESHALE, A. B. Incidence of infant mortality and its predictors in east africa using gompertz gamma shared frailty model. **Archives of Public Health**, BioMed Central, v. 80, n. 1, p. 1–12, 2022. Citado na página 32.

TIAN, W.; HEO, Y.; WILDE, P. D.; LI, Z.; YAN, D.; PARK, C. S.; FENG, X.; AUGENBROE, G. A review of uncertainty analysis in building energy assessment. **Renewable and Sustainable Energy Reviews**, Elsevier, v. 93, p. 285–301, 2018. Citado na página 73.

TOMAZELLA, V. L. D.; MILANI, E. Â.; DIAS, T. C. M. Gompertz regression model with gamma frailty: a study on the application in lung cancer. **Brazilian Journal of Biometrics**, v. 36, n. 4, p. 860–879, 2018. Citado na página 32.

TORRES, M. d. P. *et al.* Determinação de propriedades térmicas em problemas de condução de calor por inferência bayesiana com o método de monte carlo hamiltoniano. Universidade do Estado do Rio de Janeiro, 2018. Citado na página 33.

TSODIKOV, A.; IBRAHIM, J. G.; YAKOVLEV, A. Estimating cure rates from survival data: an alternative to two-component mixture models. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 464, p. 1063–1078, 2003. Citado nas páginas 31, 88 e 107.

TSODIKOV, A. D.; YAKOVLEV, A. Y.; ASSELAIN, B. **Stochastic models of tumor latency and their biostatistical applications**. [S.l.]: World Scientific, 1996. v. 1. Citado nas páginas 31, 88, 90, 107 e 141.

TWEEDIE, M. C. *et al.* An index which distinguishes between some important exponential families. In: **Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference**. [S.l.: s.n.], 1984. v. 579, p. 579–604. Citado na página 30.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, Springer, v. 16, n. 3, p. 439–454, 1979. Citado nas páginas 27, 29, 37 e 39.

WHEELER, M. W.; WESTERHOUT, J.; BAUMERT, J. L.; REMINGTON, B. C. Bayesian stacked parametric survival with frailty components and interval-censored failure times: An application to food allergy risk. **Risk Analysis**, Wiley Online Library, v. 41, n. 1, p. 56–66, 2021. Citado na página 30.

WHTTMORE, G. An inverse gaussian model for labour turnover. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 142, n. 4, p. 468–478, 1979. Citado na página 62.

WIENKE, A. **Frailty models in survival analysis**. [S.l.]: CRC press, 2010. Citado nas páginas 31, 35, 37, 38, 39, 41, 42, 43, 44, 45 e 46.

XAVIER, C. M. Métodos de monte carlo hamiltoniano aplicados em modelos garch. Universidade Federal de São Carlos, 2019. Citado nas páginas 33, 35, 51, 53 e 55.

ZAVALETA, K. E. C. Modelos série de potência com excesso de zeros observáveis e latentes. Universidade Federal de São Carlos, 2016. Citado na página 42.

ZHOU, H.; HANSON, T.; JARA, A.; ZHANG, J. Modelling county level breast cancer survival data using a covariate-adjusted frailty proportional hazards model. **The annals of applied statistics**, NIH Public Access, v. 9, n. 1, p. 43, 2015. Citado nas páginas 29 e 30.

ZHOU, H.; HANSON, T.; ZHANG, J. Generalized accelerated failure time spatial frailty model for arbitrarily censored data. **Lifetime data analysis**, Springer, v. 23, n. 3, p. 495–515, 2017. Citado na página 30.

APÊNDICE

A

DEMONSTRAÇÕES E OPERAÇÕES MATEMÁTICAS DO CAPÍTULO 3

As operações e demonstrações apresentadas a seguir são referentes ao Capítulo 3. Estas foram realizadas para fundamentar o estudo, a interpretação e a caracterização dos modelos PVF e PVF defeituoso descritos neste capítulo.

A.1 Principais resultados e demonstrações

Proposição 1. Considere a função de risco acumulado (3.7) e a proporção de indivíduos imunes ao evento de interesse p_0 definida em (3.6). Se $\gamma < 0$, então a função (3.7) é limitada por $-\log(p_0)$, quando $t \longrightarrow \infty$.

Demonstração. De fato, ao considerar a função (3.7) com $\gamma < 0$ segue que

$$H(t) = \frac{(1+H_0(t))^{\gamma}-1}{\gamma} = \frac{1}{\gamma}[(1+H_0(t))^{\gamma}-1] = \frac{1}{\gamma}\left[\frac{1}{(1+H_0(t))^{-\gamma}}-1\right] \longrightarrow -\frac{1}{\gamma}, \quad \text{(A.1)}$$

pois como $H_0(t) \longrightarrow \infty$, quando $t \longrightarrow \infty$, então $\left[\frac{1}{(1+H_0(t))^{-\gamma}}\right] \longrightarrow 0$. Agora, tome a proporção de indivíduos imunes ao evento de interesse definida em (3.6). Assim, ao aplicar menos o logaritmo nesta proporção têm-se que

$$-\log(p_0) = -\log\left(\exp\left\{\frac{1}{\gamma}\right\}\right) = -\frac{1}{\gamma}; \quad \gamma < 0. \quad \text{(A.2)}$$

Deste modo, (A.1) e (A.2) implicam que $H(t) \longrightarrow -\log(p_0)$, quando $t \longrightarrow \infty$. Consequentemente, $H(t) \leq -\log(p_0)$. Portanto, $H(t)$ é limitada por $-\log(p_0)$, quando $t \longrightarrow \infty$. □

Proposição 2. Considere a função de sobrevivência marginal (3.4) com $\gamma \leq 1$, $\mu > 0$ e $\sigma > 0$. Logo, são válidas as seguintes afirmações:

1. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{\frac{b\gamma(e^{at}-1)+\sigma^{\gamma}}{a\mu}} - \sigma$, com $t>0$, $b>0$ e $a\in\mathbb{R}$ de forma que: $b\neq\mu$ e $a\neq\gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

2. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{e^{\gamma t}-1+\sigma^{\gamma}} - \sigma$, com $t>0$, $b>0$ e $a\in\mathbb{R}$ de forma que: $b=\mu$ e $a=\gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

Demonstração. Considere a função de sobrevivência marginal (3.4) com $\gamma\leq 1$, $\mu>0$ e $\sigma>0$.

1. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{\frac{b\gamma(e^{at}-1)+\sigma^{\gamma}}{a\mu}} - \sigma$, com $t>0$, $b>0$ e $a\in\mathbb{R}$ de forma que: $b\neq\mu$ e $a\neq\gamma$. Assim

$$\begin{aligned}
S(t) &= \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+H_0(t))^{\gamma}-\sigma^{\gamma}\right]\right\}\\
&= \exp\left\{-\frac{\mu}{\gamma}\left[\left(\sigma+\sqrt[\frac{1}{\gamma}]{\frac{b\gamma(e^{at}-1)+\sigma^{\gamma}}{a\mu}}-\sigma\right)^{\gamma}-\sigma^{\gamma}\right]\right\}\\
&= \exp\left\{-\frac{\mu}{\gamma}\left[\left(\frac{b\gamma(e^{at}-1)+\sigma^{\gamma}}{a\mu}\right)-\sigma^{\gamma}\right]\right\}\\
&= \exp\left\{-\frac{b}{a}(e^{at}-1)\right\}\\
&= S_G(t).
\end{aligned}$$

   Portanto, $S(t)=S_G(t)$ com $t>0$, ou seja, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

2. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{e^{\gamma t}-1+\sigma^{\gamma}} - \sigma$, com $t>0$, $b>0$ e $a\in\mathbb{R}$ de forma que: $b=\mu$ e $a=\gamma$. Logo

$$\begin{aligned}
S(t) &= \exp\left\{-\frac{\mu}{\gamma}\left[(\sigma+H_0(t))^{\gamma}-\sigma^{\gamma}\right]\right\}\\
&= \exp\left\{-\frac{b}{a}\left[\left(\sigma+\sqrt[\frac{1}{a}]{e^{at}-1+\sigma^{a}}-\sigma\right)^{a}-\sigma^{a}\right]\right\}\\
&= \exp\left\{-\frac{b}{a}(e^{at}-1+\sigma^{a})-\sigma^{a}\right\}\\
&= \exp\left\{-\frac{b}{a}(e^{at}-1)\right\}\\
&= S_G(t).
\end{aligned}$$

   Portanto, $S(t)=S_G(t)$ com $t>0$, ou seja, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

□

Proposição 3. Considere a função de sobrevivência marginal (3.5) com $\gamma \leq 1$. Logo, são válidas as seguintes afirmações:

1. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = \sqrt[\frac{1}{\gamma}]{\frac{b\gamma(e^{at}-1)+1}{a}} - 1$, com $t > 0$, $b > 0$ e $a \in \mathbb{R}$ de forma que: $b \neq 1$ e $a \neq \gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

2. Assuma a função de sobrevivência do modelo Gompertz (3.2) e suponha que $H_0(t) = e^{\gamma t} - 1$, com $t > 0$, $b > 0$ e $a \in \mathbb{R}$ de forma que: $b = 1$ e $a = \gamma$. Então, as funções de sobrevivência $S(t)$ e $S_G(t)$ são equivalentes.

Demonstração. Essa demonstração é similar a apresentada para a Proposição 2, contudo assumindo que $b = 1$, $a = \gamma$ e $\sigma = 1$. □

Proposição 4. A função de sobrevivência marginal (3.8) está bem definida.

Demonstração. De fato, considere a função de sobrevivência marginal (3.8) dada por

$$S_D(t) = \exp\left\{-\frac{1}{\gamma}[(1+H_{MEP}(t))^{\gamma}-1]\right\}; \quad t > 0 \quad \text{e} \quad \gamma \leq 1,$$

em que $H_{MEP}(\cdot)$ é a função de risco acumulado do modelo MEP definida em (2.30) para todo $t \in I_j$ intervalos disjuntos com $\lambda_j > 0$. Note que

- $S_D(0) = 1$
  Se $t = 0$ então $H_{MEP}(0) = 0$. Assim

  $$S_D(0) = \exp\left\{-\frac{1}{\gamma}[(1+H_{MEP}(0))^{\gamma}-1]\right\} = \exp\left\{-\frac{1}{\gamma}[(1+0)^{\gamma}-1]\right\} = e^0 = 1.$$

  Logo, $S_D(0) = 1$.

- Se $0 < \gamma \leq 1$ então $S_D(t) = 0$, quando $t \longrightarrow \infty$.
  Note que $H_{MEP}(t) \longrightarrow \infty$, quando $t \longrightarrow \infty$. Consequentemente, ao assumir que $0 < \gamma \leq 1$, segue que

  $$-\frac{1}{\gamma}[(1+H_{MEP}(t))^{\gamma}-1] \longrightarrow -\infty.$$

  Logo

  $$S_D(t) = \exp\left\{-\frac{1}{\gamma}[(1+H_{MEP}(t))^{\gamma}-1]\right\} = e^{-\infty} = 0$$

  Portanto, para $0 < \gamma \leq 1$, $S_D(t) = 0$, quando $t \longrightarrow \infty$.

- Se $\gamma < 0$ e p_0 corresponde a proporção de indivíduos imunes ao evento de interesse definida em (3.6), então $S_D(t) = p_0$, quando $t \longrightarrow \infty$.
  Note que $H_{MEP}(t) \longrightarrow \infty$, quando $t \longrightarrow \infty$. Consequentemente, para $0 < \gamma$, segue que

$$S_D(t) = \exp\left\{-\frac{1}{\gamma}\left[\frac{1}{(1+H_{MEP}(t))^{\gamma}} - 1\right]\right\} = \exp\left\{\frac{1}{\gamma}\right\} = p_0.$$

  Logo, para $0 < \gamma$, $S_D(t) = p_0$, quando $t \longrightarrow \infty$.

Portanto, a função de sobrevivência marginal (3.8) está bem definida. □

Proposição 5. Assuma as hipóteses estabelecidas na descrição da propriedade matemática do Capítulo 3. Considere a variável aleatória Y tal que $Y_0 = T \sim H_0$ e $Y_{n+1} \sim \exp -H_0(n+1)$ com $n \geq 0$. Desta forma, o r-ésimo momento de T correspondente a função (3.14) dado por

$$\mu'_r = \sum_{n=0}^{\infty} w_n\,(n+1) \int_0^{\infty} t^r\, H_0(t)^n\, h_0(t)\, dt, \tag{A.3}$$

é convergente.

Demonstração. Nesta demonstração será mostrado que o r-ésimo momento (A.3) é convergente, ou seja, que a série definida para este r-ésimo momento converge. Para isso, assuma as hipóteses estabelecidas na descrição da propriedade matemática do Capítulo 3. Considere a variável aleatória Y tal que $Y_0 = T \sim H_0$ e $Y_{n+1} \sim \exp -H_0(n+1)$ com $n \geq 0$ e tome como função de base a distribuição Exponencial tal que $H_0(t) = \lambda t$ e $h_0(t) = \lambda$ para $\lambda > 0$. Assim, ao aplicar a função de base em (A.3), é possível reescrever o r-ésimo momento como

$$\mu'_r = \sum_{n=0}^{\infty} w_n\,(n+1) \int_0^{\infty} t^r\, (\lambda t)^n\, \lambda\, dt. \tag{A.4}$$

O intuito é realizar operações algébricas em (A.4) organizando-a, para que, posteriormente, seja aplicado o Teste de D'Alembert ou Teste da Razão para séries de potência, garantindo a convergência da série presente na expressão. Desta forma, observe que ao calcular a integral da expressão (A.4), segue que

$$\int_0^t t^r\, (\lambda t)^n\, \lambda\, dt = \int_0^t \lambda^{n+1}\, t^{n+r} dt = \frac{\lambda^{n+1}\, t^{n+r+1}}{n+r+1}; \quad n+r \neq -1 \quad \text{e} \quad 0 \leq t \leq \infty.$$

Logo, substituindo em (A.4) têm-se que

$$\mu'_r = \sum_{n=0}^{\infty} w_n\,(n+1)\, \frac{\lambda^{n+1}\, t^{n+r+1}}{n+r+1}.$$

Agora, defina as sequências

$$(a_n) = w_n\,(n+1)\, \frac{\lambda^{n+1}\, t^{n+r+1}}{n+r+1}$$

e

$$(a_{n+1}) = w_{n+1}\,(n+2)\,\frac{\lambda^{n+2}\,t^{n+r+2}}{n+r+2}.$$

Assim, observe que

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{w_{n+1}\,(n+2)\,(\lambda^{n+2}\,t^{n+r+2})(n+r+1)}{(n+r+2)w_n\,(n+1)\,(\lambda^{n+1}\,t^{n+r+1})}\right| = \left|\frac{w_{n+1}\,(n+2)\,(\lambda t)(n+r+1)}{w_n\,(n+1)(n+r+2)}\right|, \quad \text{(A.5)}$$

em que pelas hipóteses estabelecidas na propriedade, segue que

$$w_n = \frac{-q_{n+1}}{(n+1)!} = -\frac{\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n+1,k}}{(n+1)!}$$

e

$$w_{n+1} = \frac{-q_{n+2}}{(n+2)!} = -\frac{\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n+2,k}}{(n+2)!}$$

de forma que

$$B^{\star}_{n+1,k} = B_{n+1,k}(2!s_2,\ldots,(n-k+2)!s_{n-k+2}) = \sum \frac{(n+1)!}{c_2!\,c_3!\ldots(2!)^{c_2}(3!)^{c_3}\ldots}\,(2!s_2)^{c_2}(3!s_3)^{c_3}\ldots,$$

$$B^{\star}_{n+2,k} = B_{n+2,k}(3!s_3,\ldots,(n-k+3)!s_{n-k+3}) = \sum \frac{(n+2)!}{c_3!\,c_4!\ldots(3!)^{c_3}(4!)^{c_4}\ldots}\,(3!s_3)^{c_3}(4!s_4)^{c_4}\ldots,$$

no qual, o somatório de $B^{\star}_{n+1,k}$ é válido para $c_2, c_3, \ldots \geq 0$, com $c_2 + c_3 + c_4 + \ldots = k$, $2c_2 + 3c_3 + \ldots = n$ e $s_i = \dfrac{(-1)^{2i-2}\,\gamma!}{i!}$, para $i \geq 2$. As hipóteses para $B^{\star}_{n+2,k}$ são similares. Deste modo, ao realizar as devidas manipulações algébricas, têm-se que

$$\frac{w_{n+1}}{w_n} = \left(-\frac{\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n+2,k}}{(n+2)!}\right)\left(-\frac{(n+1)!}{\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n+1,k}}\right) = \frac{B^{\star}_{n+2,k}}{(n+2)B^{\star}_{n+1,k}}.$$

Ou ainda

$$\frac{w_{n+1}}{w_n} = \frac{B^{\star}_{n+2,k}}{(n+2)B^{\star}_{n+1,k}} = \sum \frac{1}{n+2}\,\frac{(n+2)c_2!(2!)^{c_2}}{(2!s_2)^{c_2}} = \frac{c_2!(2!)^{c_2}}{(2!s_2)^{c_2}} = \frac{c_2!}{(s_2)^{c_2}}. \quad \text{(A.6)}$$

Aplicando (A.6) em (A.5) é visto que

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{c_2!}{(s_2)^{c_2}}\,\frac{(n+2)\,(\lambda t)(n+r+1)}{(n+1)(n+r+2)}\right| = \left|\frac{c_2!\lambda t}{(s_2)^{c_2}}\,\frac{nr+n^2+n+2r+2n+2}{nr+n^2+2n+r+n+2}\right|. \quad \text{(A.7)}$$

Agora, ao aplicar o limite em (A.7) segue que

$$\begin{aligned}
\lim_{n\longrightarrow\infty}\left|\frac{a_{n+1}}{a_n}\right| &= \lim_{n\longrightarrow\infty}\left|\frac{c_2!\lambda t}{(s_2)^{c_2}}\,\frac{nr+n^2+n+2r+2n+2}{nr+n^2+2n+r+n+2}\right| \\
&= \left|\frac{c_2!\lambda t}{(s_2)^{c_2}}\right| \lim_{n\longrightarrow\infty}\left|\frac{nr+n^2+n+2r+2n+2}{nr+n^2+2n+r+n+2}\right| \\
&= \left|\frac{c_2!\lambda t}{(s_2)^{c_2}}\right| \lim_{n\longrightarrow\infty}\left|\frac{\frac{r}{n}+1\frac{1}{n}+\frac{2r}{n^2}+\frac{2}{n}+\frac{2}{n^2}}{\frac{r}{n}+1\frac{2}{n}+\frac{r}{n^2}+\frac{1}{n}+\frac{2}{n^2}}\right| \\
&= \left|\frac{c_2!\lambda t}{(s_2)^{c_2}}.\right|
\end{aligned}$$

Portanto, pelo Teste de D'Alembert, é concluído que se

$$\lim_{n\longrightarrow\infty}\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{c_2!\lambda t}{(s_2)^{c_2}}\right| < 1,$$

então a série (A.4) será absolutamente convergente, ou seja

$$\left|\frac{c_2!\lambda t}{(s_2)^{c_2}}\right| = \left|\frac{c_2!\lambda t}{\left(\frac{-\gamma\,(1-\gamma)}{2}\right)^{c_2}}\right| = \left|\frac{2^{c_2}c_2!\lambda t}{(-\gamma(1-\gamma))^{c_2}}\right| < 1.$$

Observe que isso ocorre se, e somente se

1. $t = 0$.
2. $$\left|\frac{2^{c_2}c_2!\lambda t}{(-\gamma(1-\gamma))^{c_2}}\right| < 1 \Longleftrightarrow t < \frac{(-\gamma(1-\gamma))^{c_2}}{2^{c_2}c_2!\lambda},$$

   tal que $\gamma \neq 1$ e $c_2 \geq 0$ é par. Caso contrário, se $\gamma = 1$ ou $c_2 \geq 0$ é impar então $t < 0$, sendo uma contradição as hipóteses da propriedade.

□

APÊNDICE

B

DEMONSTRAÇÕES E OPERAÇÕES MATEMÁTICAS DO CAPÍTULO 4

As operações e demonstrações apresentadas a seguir são referentes ao Capítulo 4. Estas foram realizadas para fundamentar o estudo, a interpretação e a caracterização dos modelos de longa duração PVF e PVF-MEP descritos neste capítulo.

B.1 Principais resultados e demonstrações

Proposição 6. Considere as hipóteses estabelecidas na Seção 4.2 e assuma a variável Z segundo a distribuição de Poisson com média $\omega\rho$, em que $\rho > 0$ constante e ω é a componente de fragilidade definida mediante a distribuição PVF de Hougaard (2000) com $\mu = \sigma = 1$. Então, $\mathbb{E}[Z] = \rho$ e $Var(Z) = (1-\gamma)\rho^2$.

Demonstração. Assuma as hipóteses estabelecidas na Seção 4.2 e tome a variável Z segundo a distribuição de Poisson com média $\omega\rho$, em que $\rho > 0$ constante e ω é a componente de fragilidade definida mediante a distribuição PVF de Hougaard (2000), tendo a transformada de Laplace caracterizada por (4.4) e a f.g.p de Z dada pela função (4.5). Suponha que $\mu = \sigma = 1$ então $\mathbb{E}[\omega] = 1$ e $Var(\omega) = 1-\gamma$. Consequentemente, é possível reescrever a f.g.p de Z como

$$\psi_Z(s) = \exp\left\{\frac{1}{\gamma}[1-(1+\rho(1-s))^{\gamma}]\right\} = \mathscr{L}_{\omega}(\rho(1-s)), \tag{B.1}$$

em que $\mathscr{L}_{\omega}(\cdot)$ é a transformada de Laplace da distribuição de ω. Observe que para determinar as derivadas de primeira e segunda ordem desta função, e com o intuito de simplicidade, é tomada a substituição $(1-s) = x$ na expressão (B.1), além de utilizar algumas regras de derivação como as Regras da Cadeia, do Produto e da derivação da função Exponencial. Como consequência a

derivada de primeira ordem é dada por

$$\begin{aligned}\mathscr{L}'_{\omega}[x] &= \frac{\partial}{\partial x}\left[\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\}\right] \\ &= \exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\}\frac{\partial}{\partial x}\left[\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right] \\ &= -\rho(1+\rho x)^{\gamma-1}\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\},\end{aligned}$$

pois

$$\frac{\partial}{\partial x}\left[\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right] = -\frac{1}{\gamma}\left(\frac{\partial}{\partial x}[(1+\rho x)^{\gamma}]\right) = -\rho(1+\rho x)^{\gamma-1}.$$

Já derivada de segunda ordem é calculada por

$$\begin{aligned}\mathscr{L}''_{\omega}(x) &= \frac{\partial}{\partial x}\left[-\rho(1+\rho x)^{\gamma-1}\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\}\right] \\ &= \frac{\partial}{\partial x}\left[-\rho(1+\rho x)^{\gamma-1}\right]\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\} \\ &\quad -\rho(1+\rho x)^{\gamma-1}\frac{\partial}{\partial x}\left[\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\}\right] \\ &= \left[-\rho^2(\gamma-1)(1+\rho x)^{\gamma-2}\right]\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\} \\ &\quad +\rho^2(1+\rho x)^{2\gamma-2}\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\} \\ &= \left[-\rho^2(\gamma-1)(1+\rho x)^{\gamma-2}+\rho^2(1+\rho x)^{2\gamma-2}\right]\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\},\end{aligned}$$

pois

$$\frac{\partial}{\partial x}\left[-\rho(1+\rho x)^{\gamma-1}\right] = -\rho\frac{\partial}{\partial x}\left[(1+\rho x)^{\gamma-1}\right] = -\rho^2(\gamma-1)(1+\rho x)^{\gamma-2}$$

e

$$\frac{\partial}{\partial x}\left[\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\}\right] = -\rho(1+\rho x)^{\gamma-1}\exp\left\{\frac{1}{\gamma}[1-(1+\rho x)^{\gamma}]\right\}.$$

Portanto, para $(1-s)=x$, segue que

$$\mathscr{L}'_{\omega}[(1-s)] = -\rho(1+\rho(1-s))^{\gamma-1}\exp\left\{\frac{1}{\gamma}[1-(1+\rho(1-s))^{\gamma}]\right\}$$

e

$$\begin{aligned}\mathscr{L}''_{\omega}[(1-s)] &= \left[-\rho^2(\gamma-1)(1+\rho(1-s))^{\gamma-2}+\rho^2(1+\rho(1-s))^{2\gamma-2}\right] \\ &\quad \times\exp\left\{\frac{1}{\gamma}[1-(1+\rho(1-s))^{\gamma}]\right\}.\end{aligned}$$

Note que como a variàvel de fragilidade Z da função (2.14) tem uma distribuição mista de Poisson com f.g.p definida por (4.5) e ao aplicar as derivadas de primeira e segunda ordem nas

expressões (2.12) e (2.13), é possível determinar a esperança e a variância em relação a variável Z de forma que

$$\mathbb{E}[Z] = -\mathscr{L}'_{\omega}(0) = \rho(1)^{\gamma-1} \exp\left\{\frac{1}{\gamma}[1-(1)^{\gamma}]\right\} = \rho e^{0} = \rho$$

e

$$Var[Z] = \mathscr{L}''_{\omega}(0) - [-\mathscr{L}'_{\omega}(0)]^2 = \left[\rho^2(1-\gamma)(1)^{\gamma-2} + \rho^2(1)^{2\gamma-2}\right] e^{0} - \rho^2 = (1-\gamma)\rho^2.$$

Portanto, $\mathbb{E}[Z] = \rho$ e $Var(Z) = (1-\gamma)\rho^2$. □

Proposição 7. Considere as hipóteses definidas para a construção do modelo (4.6) com $\rho > 0$, $\mu > 0$, $\sigma > 0$ e $\gamma \leq 1$. Logo, as seguintes afirmações são válidas:

1. Se $\sigma = 0$ e $\mu = \gamma = 1$, então o modelo (4.6) se reduz ao modelo BCH introduzido por Tsodikov, Yakovlev e Asselain (1996).
2. Se $\gamma \longrightarrow 0$ e $\mu = \sigma = \dfrac{1}{\varepsilon}$ com $\varepsilon > 0$, então o modelo (4.6) se reduz ao modelo BCH-Gama de Cancho *et al.* (2021).
3. Se $\gamma = 0,5$, $\sigma = \dfrac{1}{2\varepsilon}$ e $\mu = \dfrac{\sqrt{2\varepsilon}}{2\varepsilon}$ com $\varepsilon > 0$, então o modelo (4.6) se reduz ao modelo BCH-IG de Cancho *et al.* (2021).

Demonstração. Considere as hipóteses definidas para o modelo (4.6) com $\rho > 0$, $\mu > 0$, $\sigma > 0$ e $\gamma \leq 1$. Logo, a função de sobrevivência marginal em relação ao tempo T que caracteriza este modelo é dada por

$$S(t) = \exp\left\{\frac{\mu}{\gamma}[\sigma^{\gamma} - (\sigma + \rho F_0(t))^{\gamma}]\right\}; \quad t > 0, \tag{B.2}$$

tal que $F_0(t) = 1 - \exp\left\{-\int_0^t h_0(u)du\right\}$ é a função de distribuição acumulada base para o tempo T. Deste modo

1. Assuma a função (B.2) com $\sigma = 0$ e $\mu = \gamma = 1$. Dai

$$\begin{aligned} S(t) &= \exp\left\{\frac{\mu}{\gamma}[\sigma^{\gamma} - (\sigma + \rho F_0(t))^{\gamma}]\right\} \\ &= \exp\left\{\frac{\gamma}{\gamma}[0 - (0 + \rho F_0(t))^{\gamma}]\right\} \\ &= \exp\left\{-(\rho F_0(t))^{\gamma}\right\} \\ &= \exp\left\{-\rho F_0(t)\right\}, \end{aligned}$$

que corresponde a função de sobrevivência do modelo BCH de Tsodikov, Yakovlev e Asselain (1996). Consequentemente, a função de risco e f.d.p também serão similares as definidas para este modelo. Portanto, o modelo (4.6) se reduz ao modelo BCH, quando $\sigma = 0$ e $\mu = \gamma = 1$.

2. Considere a função de risco relacionada ao modelo (4.6) e tome $\gamma \longrightarrow 0$ e $\mu = \sigma = \frac{1}{\varepsilon}$ para $\varepsilon > 0$. Logo

$$\begin{aligned} h(t) &= \mu\rho f_0(t)(\sigma + \rho F_0(t))^{\gamma-1} \\ &= \sigma\rho f_0(t)(\sigma + \rho F_0(t))^{\gamma-1} \\ &= \frac{\rho f_0(t)}{\varepsilon}\left(\frac{1+\varepsilon\rho F_0(t)}{\varepsilon}\right)^{\gamma-1} \\ &= \frac{\frac{\rho f_0(t)}{\varepsilon}}{\left(\frac{1+\varepsilon\rho F_0(t)}{\varepsilon}\right)^{1-\gamma}}. \end{aligned}$$

Como $\gamma \longrightarrow 0$ segue que

$$h(t) = \frac{\frac{\rho f_0(t)}{\varepsilon}}{\left(\frac{1+\varepsilon\rho F_0(t)}{\varepsilon}\right)^{1-\gamma}} \longrightarrow \frac{\frac{\rho f_0(t)}{\varepsilon}}{\left(\frac{1+\varepsilon\rho F_0(t)}{\varepsilon}\right)} = \frac{\rho f_0(t)}{1+\varepsilon\rho F_0(t)},$$

que corresponde a função de risco do modelo BCH-Gama de Cancho *et al.* (2021). Consequentemente, a função de sobrevivência e f.d.p também serão similares as definidas para este modelo. Portanto, o modelo (4.6) se reduz ao modelo BCH-Gama, quando $\gamma \longrightarrow 0$ e $\mu = \sigma = \frac{1}{\varepsilon}$.

3. Suponha a função (B.2) com $\gamma = 0,5$, $\sigma = \frac{1}{2\varepsilon}$ e $\mu = \frac{\sqrt{2\varepsilon}}{2\varepsilon}$ para $\varepsilon > 0$. Assim

$$\begin{aligned} S(t) &= \exp\left\{\frac{\mu}{\gamma}\left[\sigma^\gamma - (\sigma + \rho F_0(t))^\gamma\right]\right\} \\ &= \exp\left\{\frac{\mu}{\gamma}\sigma^\gamma\left[\frac{\sigma^\gamma}{\sigma^\gamma} - \frac{(\sigma+\rho F_0(t))^\gamma}{\sigma^\gamma}\right]\right\} \\ &= \exp\left\{\frac{\mu}{\gamma}\sigma^\gamma\left[1 - \left(1 + \frac{\rho F_0(t)}{\sigma}\right)^\gamma\right]\right\} \\ &= \exp\left\{\frac{\mu}{\gamma}\left(\frac{1}{2\varepsilon}\right)^\gamma\left[1 - (1+2\varepsilon\rho F_0(t))^\gamma\right]\right\} \\ &= \exp\left\{\frac{1}{\varepsilon}\left(1 - \sqrt{1+2\varepsilon\rho F_0(t)}\right)\right\} \end{aligned}$$

que corresponde a função de sobrevivência do modelo BCH-IG de Cancho *et al.* (2021). Consequentemente, a função de risco e f.d.p também serão similares as definidas para este modelo. Portanto, o modelo (4.6) se reduz ao modelo BCH-IG, quando $\gamma = 0,5$, $\sigma = \frac{1}{2\varepsilon}$ e $\mu = \frac{\sqrt{2\varepsilon}}{2\varepsilon}$.

□

Proposição 8. Considere a função de risco acumulado (4.11) e a proporção de indivíduos imunes ao evento de interesse p_0 definida em (4.8). Logo, a função (4.11) é limitada por $-\log(p_0)$, quando $t \longrightarrow \infty$.

Demonstração. De fato, ao considerar a função (4.11) segue que

$$H(t) = \frac{(1+\rho F_0(t))^{\gamma}-1}{\gamma} \longrightarrow \frac{(1+\rho)^{\gamma}-1}{\gamma}, \tag{B.3}$$

pois, $F_0(t) \longrightarrow 1$, quando $t \longrightarrow \infty$. Agora, tome a proporção de indivíduos imunes ao evento de interesse definida em (4.8). Assim, ao aplicar menos o logaritmo nesta proporção é obtido que

$$-\log(p_0) = -\log\left(\exp\left\{\frac{1}{\gamma}[1-(1+\rho)^{\gamma}]\right\}\right) = \frac{(1+\rho)^{\gamma}-1}{\gamma}. \tag{B.4}$$

Deste modo, segue de (B.3) e (B.4) que $H(t) \longrightarrow -\log(p_0)$, quando $t \longrightarrow \infty$. Consequentemente, $H(t) \leq -\log(p_0)$. Portanto, $H(t)$ é limitada por $-\log(p_0)$.

Observe que a demonstração referente a limitação da função de risco acumulado com base segundo o modelo MEP, sendo definida em (4.16), é feita de forma similar ao que foi realizado para esta proposição. □

Proposição 9. Considere as hipóteses definidas na construção do modelo (4.7) com $\rho > 0$ e $\gamma \leq 1$. Logo, as seguintes afirmações são válidas:

1. Se $\gamma = 1$, então o modelo (4.7) se reduz ao modelo BCH.
2. Se $\gamma \longrightarrow 0$, então o modelo (4.7) se reduz ao modelo BCH-Gama com $\varepsilon = 1$.
3. Se $\gamma = 0,5$, então o modelo (4.7) se reduz ao modelo BCH-IG com $\varepsilon = 0,5$.

Demonstração. Considere as hipóteses definidas no modelo (4.7) com $\rho > 0$ e $\gamma \leq 1$. Logo, a função de sobrevivência marginal em relação ao tempo T que caracteriza este modelo é dada por

$$S(t) = \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\}, \tag{B.5}$$

em que $F_0(\cdot)$ é a função de distribuição acumulada base para o tempo T. Deste modo

1. Assuma a função (B.5) com $\gamma = 1$. Logo

$$\begin{aligned} S(t) &= \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\} \\ &= \exp\left\{1\left[1-(1+\rho F_0(t))^{1}\right]\right\} \\ &= \exp\left\{-\rho F_0(t)\right\}, \end{aligned}$$

que corresponde a função de sobrevivência do modelo BCH. Consequentemente, a função de risco e a f.d.p também serão similares as definidas para este modelo. Portanto, o modelo (4.7) se reduz ao modelo BCH, quando $\gamma = 1$.

2. Considere a função de risco (4.10) relacionada ao modelo (4.6) e tome $\gamma \longrightarrow 0$. Dai

$$h(t) = \rho f_0(t)\ (1+\rho F_0(t))^{\gamma-1} = \frac{\rho f_0(t)}{(1+\rho F_0(t))^{1-\gamma}}.$$

Como $\gamma \longrightarrow 0$ e assumindo que $\varepsilon = 1$, segue que

$$h(t) = \frac{\rho f_0(t)}{(1+\rho F_0(t))^{1-\gamma}} \longrightarrow \frac{\rho f_0(t)}{(1+\rho F_0(t))} = \frac{\rho f_0(t)}{1+\varepsilon\rho F_0(t)}.$$

que corresponde a função de risco do modelo BCH-Gama com $\varepsilon = 1$. Consequentemente, a função de sobrevivência e a f.d.p também serão similares as definidas para este modelo. Portanto, o modelo (4.7) se reduz ao modelo BCH-Gama com $\varepsilon = 1$, quando $\gamma \longrightarrow 0$.

3. Seja função (B.5) com $\gamma = 0,5$ e considere $\varepsilon = 0,5$. Assim

$$\begin{aligned} S(t) &= \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\} \\ &= \exp\left\{2\left[1-\sqrt{1+\rho F_0(t)}\right]\right\} \\ &= \exp\left\{\frac{1}{\varepsilon}\left(1-\sqrt{1+2\varepsilon\rho F_0(t)}\right)\right\}, \end{aligned}$$

que corresponde a função de sobrevivência do modelo BCH-IG com $\varepsilon = 0,5$. Consequentemente, a função de risco e f.d.p também serão similares as definidas para este modelo. Portanto, o modelo (4.7) se reduz ao modelo BCH-IG com $\varepsilon = 0,5$, quando $\gamma = 0,5$.

□

Proposição 10. A função de sobrevivência própria (4.12) está bem definida.

Demonstração. De fato, considere a função de sobrevivência própria (4.12) dada por

$$S_{\mathrm{P}}(t) = \frac{\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\} - p_0}{1-p_0}; \quad \gamma \leq 1 \quad \text{e} \quad \rho > 0,$$

com p_0 é definido por (4.8). Note que

$$S_{\mathrm{P}}(0) = \frac{\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(0))^{\gamma}]\right\} - p_0}{1-p_0} = \frac{e^0 - p_0}{1-p_0} = \frac{1-p_0}{1-p_0} = 1,$$

pois, como $t = 0$ então $F_0(0) = 0$. Agora, observe que $F_0(t) = 1$, quando $t \longrightarrow \infty$. Assim

$$S_{\mathrm{P}}(t) = \frac{\exp\left\{\frac{1}{\gamma}[1-(1+\rho F_0(t))^{\gamma}]\right\} - p_0}{1-p_0} = \frac{\exp\left\{\frac{1}{\gamma}[1-(1+\rho)^{\gamma}]\right\} - p_0}{1-p_0} = \frac{p_0 - p_0}{1-p_0} = 0.$$

Logo, $S_P(0) = 1$ e $S_P(t) = 0$, quando $t \longrightarrow \infty$. Portanto, a função de sobrevivência própria (4.12) está bem definida. □

Proposição 11. A função de sobrevivência com base MEP definida em (4.13) está bem definida.

Demonstração. De fato, considere a função de sobrevivência com base MEP definida em (4.13) dada por

$$S_L(t) = \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_{MEP}(t))^{\gamma}]\right\}; \quad \gamma \leq 1 \quad \text{e} \quad \rho > 0,$$

em que $F_{MEP}(\cdot)$ é a função de distribuição acumulada (2.34) referente ao modelo MEP para todo $t \in I_j$ intervalos disjuntos e $\lambda_j > 0$. Note que

$$S_L(0) = \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_{MEP}(0))^{\gamma}]\right\} = \exp\left\{\frac{1}{\gamma}[1-(1+0)^{\gamma}]\right\} = e^0 = 1,$$

pois, como $t = 0$ segue que $F_{MEP}(0) = 0$. Agora, observe que $F_{MEP}(t) \longrightarrow 1$, quando $t \longrightarrow \infty$. Assim

$$S_L(t) = \exp\left\{\frac{1}{\gamma}[1-(1+\rho F_{MEP}(t))^{\gamma}]\right\} = \exp\left\{\frac{1}{\gamma}[1-(1+\rho)^{\gamma}]\right\} = p_0,$$

com p_0 definido por (4.8). Logo, $S_L(0) = 1$ e $S_L(t) = p_0$, quando $t \longrightarrow \infty$. Portanto, a função de sobrevivência com base MEP definida em (4.13) está bem definida. □

Proposição 12. Assuma as hipóteses definidas na descrição da propriedade matemática do Capítulo 4. Considere a variável aleatória Y tal que $Y_0 = T \sim F_0$ e $Y_{n+1} \sim \exp -F_0(n+1)$ com $n \geq 0$. Desta forma, o r-ésimo momento de T correspondente a função (4.21) dado por

$$\mu'_r = \mathbb{E}\,(T^r) = \sum_{t=0}^{\infty} t^r f_C(t) = \sum_{t=0}^{\infty}\sum_{n=0}^{\infty} y_n\,(n+1)\,t^r\,[F_0(t)]^n f_0(t), \tag{B.6}$$

é convergente.

Demonstração. Nesta demonstração será mostrado que o r-ésimo momento (B.6) é convergente, ou seja, que as séries definidas para este r-ésimo momento convergem. Para isso, assuma as hipóteses estabelecidas na descrição da propriedade matemática do Capítulo 4. Considere a variável aleatória Y tal que $Y_0 = T \sim F_0$ e $Y_{n+1} \sim \exp -F_0(n+1)$ com $n \geq 0$. Tome como função de base a distribuição Exponencial tal que $F_0(t) = 1 - e^{-\lambda t}$ e $f_0(t) = \lambda e^{-\lambda t}$ com $\lambda > 0$. Assim, ao aplicar a função de base em (B.6), é possível reescrever o r-ésimo momento como

$$\mu'_r = \sum_{t=0}^{\infty}\sum_{n=0}^{\infty}(n+1)\;y_n\;t^r(1-e^{-\lambda t})^n\lambda e^{-\lambda t}. \tag{B.7}$$

O intuito é realizar operações algébricas em (B.7) para que, posteriormente, seja aplicado o Teste de D'Alembert ou Teste da Razão para séries de potência, garantindo a convergência destas séries. Assim, denota-se as sequências

$$(a_n) = y_n\;(n+1)\;\lambda t^r\;e^{-\lambda t}\;(1-e^{-\lambda t})^n$$

e

$$(a_{n+1}) = y_{n+1}\ (n+2)\ \lambda t^r\ e^{-\lambda t}\ (1-e^{-\lambda t})^{n+1}.$$

Logo,

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{y_{n+1}\ (n+2)\ \lambda t^r\ e^{-\lambda t}\ (1-e^{-\lambda t})^{n+1}}{y_n\ (n+1)\ \lambda t^r\ e^{-\lambda t}\ (1-e^{-\lambda t})^{n}}\right| = \left|\frac{y_{n+1}\ (n+2)\ (1-e^{-\lambda t})}{y_n\ (n+1)}\right|. \tag{B.8}$$

Pelas hipóteses estabelecidas na propriedade, segue que

$$y_n = \frac{-z_{n+1}}{(1-p_0)(n+1)!} = -\frac{\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n+1,k}}{(1-p_0)(n+1)!}$$

e

$$y_{n+1} = \frac{-z_{n+2}}{(1-p_0)(n+2)!} = -\frac{\sum_{k=0}^{\infty}\left(-\frac{1}{\gamma}\right)^k B^{\star}_{n+2,k}}{(1-p_0)(n+2)!},$$

com

$$B^{\star}_{n+1,k} = B_{n+1,k}(2!s_2,\ldots,(n-k+2)!s_{n-k+2}) = \sum \frac{(n+1)!}{c_2!\,c_3!\ldots(2!)^{c_2}(3!)^{c_3}\ldots}\,(2!s_2)^{c_2}(3!s_3)^{c_3}\ldots,$$

$$B^{\star}_{n+2,k} = B_{n+2,k}(3!s_3,\ldots,(n-k+3)!s_{n-k+3}) = \sum \frac{(n+2)!}{c_3!\,c_4!\ldots(3!)^{c_3}(4!)^{c_4}\ldots}\,(3!s_3)^{c_3}(4!s_4)^{c_4}\ldots,$$

de forma que o somatório de $B^{\star}_{n+1,k}$ é válido para $c_2, c_3, \ldots \geq 0$, com $c_2 + c_3 + c_4 + \ldots = k$, $2c_2 + 3c_3 + \ldots = n$ e $s_i = -\left(\frac{\gamma!\ \rho^i}{i!}\right)$, para $i \geq 2$. As hipóteses para $B^{\star}_{n+2,k}$ são similares. Desta forma, ao realizar as devidas manipulações algébricas, segue que

$$\begin{aligned}
\frac{y_{n+1}(n+2)(1-e^{-\lambda t})}{y_n(n+1)} &= \left(\frac{-z_{n+2}(n+2)(1-e^{-\lambda t})}{(1-p_0)(n+2)!}\right)\left(\frac{(1-p_0)(n+1)!}{-z_{n+1}(n+1)}\right) \\
&= \frac{z_{n+2}(1-e^{-\lambda t})}{z_{n+1}(n+1)} \\
&= \frac{(1-e^{-\lambda t})\sum_{k=0}^{\infty}\left(\frac{1}{\gamma}\right)^k B^{\star}_{n+2,k}}{(n+1)\sum_{k=0}^{\infty}\left(\frac{1}{\gamma}\right)^k B^{\star}_{n+1,k}} \\
&= \frac{(1-e^{-\lambda t})B^{\star}_{n+2,k}}{(n+1)B^{\star}_{n+1,k}}
\end{aligned}$$

Ou ainda

$$\frac{(1-e^{-\lambda t})B^{\star}_{n+2,k}}{(n+1)B^{\star}_{n+1,k}} = \sum \frac{(1-e^{-\lambda t})}{n+1}\frac{(n+2)c_2!(2!)^{c_2}}{(2!s_2)^{c_2}} = \frac{(1-e^{-\lambda t})}{n+1}\frac{(n+2)c_2!}{(s_2)^{c_2}}. \tag{B.9}$$

Ao aplicar (B.9) em (B.8) segue que

$$\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{y_{n+1}(n+2)(1-e^{-\lambda t})}{y_n(n+1)}\right| = \left|\frac{(1-e^{-\lambda t})}{n+1}\frac{(n+2)c_2!}{(s_2)^{c_2}}\right|. \tag{B.10}$$

Agora, ao aplicar o limite em (B.10) é notado que

$$\begin{aligned}
\lim_{n\longrightarrow\infty}\left|\frac{a_{n+1}}{a_n}\right| &= \lim_{n\longrightarrow\infty}\left|\frac{(1-e^{-\lambda t})}{n+1}\frac{(n+2)c_2!}{(s_2)^{c_2}}\right| \\
&= \left|\frac{(1-e^{-\lambda t})c_2!}{(s_2)^{c_2}}\right| \lim_{n\longrightarrow\infty}\left|\frac{n+2}{n+1}\right| \\
&= \left|\frac{(1-e^{-\lambda t})c_2!}{(s_2)^{c_2}}\right| \lim_{n\longrightarrow\infty}\left|\frac{\frac{n}{n}+\frac{2}{n}}{\frac{n}{n}+\frac{1}{n}}\right| \\
&= \left|\frac{(1-e^{-\lambda t})c_2!}{(s_2)^{c_2}}\right| \lim_{n\longrightarrow\infty}\left|\frac{1+\frac{2}{n}}{1+\frac{1}{n}}\right| \\
&= \left|\frac{(1-e^{-\lambda t})c_2!}{(s_2)^{c_2}}\right|.
\end{aligned}$$

Portanto, pelo Teste de D'Alembert, é concluído que se

$$\lim_{n\longrightarrow\infty}\left|\frac{a_{n+1}}{a_n}\right| = \left|\frac{(1-e^{-\lambda t})c_2!}{(s_2)^{c_2}}\right| < 1,$$

então a primeira série de (B.7) será absolutamente convergente. Ou seja

$$\left|\frac{(1-e^{-\lambda t})c_2!}{(s_2)^{c_2}}\right| = \left|\frac{(1-e^{-\lambda t})c_2!}{\left(\frac{\gamma\rho^2\,(1-\gamma)}{2}\right)^{c_2}}\right| = \left|\frac{(1-e^{-\lambda t})c_2!(2)^{c_2}}{(\gamma\rho^2\,(1-\gamma))^{c_2}}\right| < 1.$$

Observe que isso ocorre se, e somente se

1. $t = 0$.
2. $$\left|\frac{(1-e^{-\lambda t})c_2!(2)^{c_2}}{(\gamma\rho^2\,(1-\gamma))^{c_2}}\right| < 1 \Longleftrightarrow \left|\frac{-\log\left(1-\frac{(\gamma\rho^2\,(1-\gamma))^{c_2}}{c_2!(2)^{c_2}}\right)}{\lambda}\right| < t.$$

   Note que se $\gamma = 1$ então $t > 0$. Agora, se $\gamma \neq 1$ então pelas definições da função logarítimica segue que $1 > a$, com $a = \frac{(\gamma\rho^2\,(1-\gamma))^{c_2}}{c_2!(2)^{c_2}}$. Como $a \in \mathbb{Q}^-$, então $0 < \frac{-\log(a)}{\lambda} < t$, respeitando as hipóteses definidas. Portanto, a primeira série de (B.7) será absolutamente convergente.

Agora, observe que ao considerar o limite em relação a t é constatado que

$$\lim_{t \longrightarrow \infty} = \left| \frac{(1 - e^{-\lambda t})c_2!}{(s_2)^{c_2}} \right| = \left| \frac{c_2!}{(s_2)^{c_2}} \left(1 - \frac{1}{e^{\lambda t}} \right) \right| = \left| \frac{c_2!}{(s_2)^{c_2}} \right|,$$

que será convergente e menor do que um, se e somente se, $c_2! < (s_2)^{c_2}$. Logo, a segunda série de (B.7) também será absolutamente convergente. Portanto, o r-ésimo momento (B.7) é convergente. □


ICMC
USP
SÃO CARLOS