UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelagem de dados de sobrevivência induzida por fragilidade discreta via versão hurdle da distribuição série de potência zero-modificada

Katy Rocio Cruz Molina

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

ICMC USP
SÃO CARLOS

Katy Rocio Cruz Molina

Modelagem de dados de sobrevivência induzida por fragilidade discreta via versão hurdle da distribuição série de potência zero-modificada

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

Orientadora: Profa. Dra. Vera Lucia Damasceno Tomazella

USP – São Carlos
Setembro de 2025

Katy Rocio Cruz Molina

Modeling discrete frailty-induced survival data via the hurdle version of the zero-modified power series distribution

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Vera Lucia Damasceno Tomazella

USP – São Carlos
September 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 Katy Rocio Cruz Molina, realizada em 11/08/2025.

Comissão Julgadora:

Profa. Dra. Vera Lucia Damasceno Tomazella (UFSCar)

Prof. Dr. Oilson Alberto Gonzatto Junior (USP)

Profa. Dra. Daiane de Souza Santos (USP)

Prof. Dr. Danilo Alvares da Silva (CAMBRIDGE)

Prof. Dr. Pedro Rafael Diniz Marinho (UFPB)

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

Com amor infinito, dedico esta tese à minha família: meus pais, Emilio e Victoria, e meu irmão, George.

Sou imensamente grata a Deus e à vida por ter vocês ao meu lado. Obrigada pelo vosso amor sem limites, pela fé que depositaram em mim quando mais precisei, e por serem meu eterno porto seguro, onde sempre posso me refugiar. Escolheria vocês, sempre, uma e mil vezes.

Amo vocês infinitamente.

AGRADECIMENTOS

Neste momento culminante, ao encerrar uma etapa tão significativa da minha vida, é com imensa gratidão que expresso meu mais profundo reconhecimento a cada pessoa e instituição que tornou possível a concretização desta tese de doutorado.

Agradeço, em primeiro lugar, a Deus, meu guia constante e silencioso. Sou grata por zelar por cada detalhe em minha vida e por, com infinita sabedoria, posicionar as pessoas certas em minha jornada, transformando desafios em oportunidades.

A trajetória do doutorado foi, sem dúvida, uma montanha-russa de emoções. Enfrentei momentos em que o cansaço e a incerteza foram tão intensos que ameaçaram paralisar-me, fazendo-me crer que eu não conseguiria prosseguir. No entanto, foi precisamente nesses períodos que precisei descobrir e cultivar uma força interna até então desconhecida. Por isso, sou imensamente grata a mim mesma: pela persistência incansável, pela resiliência que emergiu da adversidade, pela coragem de acreditar em meu potencial quando tudo parecia difícil, e por ter abraçado cada desafio como uma oportunidade de crescimento. Esta experiência foi, em si, uma profunda lição de vida que transcende o acadêmico, moldando quem sou hoje.

Aos meus amados pais, Emilio e Victoria, e ao meu irmão George, dedico todo o meu amor e uma gratidão infinita. O amor incondicional de vocês, a confiança depositada em mim e o apoio inabalável foram meu alicerce e a principal fonte de minha fortaleza. Obrigada por acreditarem em mim, especialmente quando minha própria fé fraquejava; por cada palavra de ânimo sempre oportuna; e pelas orações e energias positivas enviadas que me impulsionaram. Obrigada também por serem meu eterno porto seguro, meu refúgio para renovar as forças. O amor de vocês foi o impulso silencioso e poderoso que me moveu a seguir em frente.

À minha querida orientadora, Vera, expresso minha mais profunda e sincera gratidão. Sua orientação e seu acompanhamento foram pilares inestimáveis desde as etapas iniciais até a concretização deste complexo projeto. Agradeço imensamente por seu apoio constante, sua valiosa parceria, seu carinho e pelo tempo generosamente investido em guiar-me ao longo de todo o processo. A jornada do doutorado apresenta desafios inerentes, e saber que podia contar com a sua presença nos momentos mais difíceis foi absolutamente fundamental para a minha perseverança.

Ao meu coorientador do intercâmbio sanduíche, Joaquín, estendo minha mais profunda gratidão. Sou grata pelos conhecimentos, ensinamentos e orientações que ele compartilhou, os quais foram cruciais para o desenvolvimento e aprofundamento desta pesquisa. Sua escuta atenta,

seu apoio constante e sua motivação, aliados às nossas conversas sempre repletas de energia positiva, transformaram minha estadia em Valência em uma das melhores e mais enriquecedoras experiências da minha vida.

Minha gratidão estende-se aos membros da banca examinadora: Danilo Alvares, Oilson Gonzato, Daiane de Souza e Pedro Rafael. Suas perspicazes sugestões e valiosas contribuições foram o alicerce para o aprimoramento desta tese.

Da mesma forma, minha gratidão estende-se a todas as pessoas que generosamente me ofereceram sua ajuda em diferentes etapas desta pesquisa: Claudio Tablada, Adriano Suzuki, Vinicius Calsavara e Eder Milani. Agradeço sinceramente pela valiosa colaboração de vocês, pelas ideias perspicazes que compartilharam e pelo precioso tempo dedicado a auxiliar-me na superação dos diversos obstáculos que surgiram em meu caminho.

Dedico, por fim, um agradecimento imenso e repleto de carinho aos meus queridos amigos, esses presentes valiosos que Deus e a vida me deram, conectando Peru, Brasil, Espanha, México, Argentina e Equador. Minha gratidão se estende a todos: os que mantenho desde a escola e a universidade, os que me acompanharam no doutorado, os que fiz durante o intercâmbio sanduíche e todos aqueles que conheci em minhas viagens. Sou infinitamente grata por cada palavra de incentivo que me sustentou, pela torcida constante, pelas conversas carregadas de boas energias, pelas risadas compartilhadas que dissiparam as sombras e alegraram os dias cinzentos. Agradeço pela motivação incansável para que eu seguisse em frente, pela disposição em me ajudar de tantas formas, por estarem ao meu lado nos dias difíceis e por fazerem esta longa jornada parecer muito mais leve. Obrigada por vibrarem comigo em cada etapa e pelos valiosos ensinamentos que cada um agregou à minha trajetória. Mesmo com oceanos ou fronteiras me separando, a amizade genuína de vocês foi um verdadeiro conforto para minha alma e enriqueceu minha vida de maneiras incontáveis, mostrando-me o quão abençoada sou por estar cercada por pessoas que me apreciam e valorizam minha presença. Da mesma forma, saibam que estarei aqui para vocês, sempre.

Aos professores do PIPGEs, minha gratidão não apenas pelos conhecimentos compartilhados em aula, que ampliaram meus horizontes, mas também pela atenção em sanar cada dúvida, o que foi fundamental para meu aprendizado e desenvolvimento. À equipe de funcionários do PIPGEs, sou igualmente grata pela cordialidade, gentileza e pelo suporte prestado com tanta prontidão sempre que precisei, tornando o ambiente acadêmico mais acolhedor.

Ao Departamento de Estadística e Investigación Operativa Aplicadas y Calidad da Universitat Politècnica de València, agradeço por me receberem de braços abertos e por todo o apoio que tornou minha estadia de intercâmbio em Valência uma experiência inesquecível e profissionalmente enriquecedora.

A todos vocês, que fizeram parte desta jornada de alguma forma, muito obrigada de coração. Esta conquista não é apenas minha; ela pertence a cada um de vocês também, que com

seus gestos e apoio, tornaram este sonho uma realidade compartilhada.

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

"Si vas a mirar atrás, que sea para ver lo que has trabajado para llegar donde estás."

(Mireia Belmonte)

RESUMO

MOLINA, K. R. C. **Modelagem de dados de sobrevivência induzida por fragilidade discreta via versão hurdle da distribuição série de potência zero-modificada**. 2025. 133 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.

A modelagem da heterogeneidade não observada em análises de sobrevivência é classicamente realizada por meio de modelos de fragilidade. No entanto, sua arquitetura tradicional, baseada em distribuições contínuas, mostra-se restritiva em contextos que exibem uma fração de cura estrutural. Neste cenário, a presente tese propõe um inovador modelo de fragilidade discreta, fundamentado na distribuição Hurdle Poisson Generalizada Zero-Modificada (HZMGP). Este modelo distingue-se por uma estrutura multifacetada: sua natureza discreta permite uma representação natural de indivíduos com fragilidade nula interpretados como individuos curados, sua capacidade de discernir diferentes padrões de zeros (inflação, deflação e truncamento); e a inclusão de um parâmetro de dispersão captura a variabilidade decorrente de fatores de risco latentes. A inferência para o modelo proposto foi estabelecida e validada sob os paradigmas frequentista e Bayesiano. Em ambos os enfoques, realizaram-se extensos estudos de simulação que confirmaram as propriedades assintóticas dos estimadores e, crucialmente, demonstraram a flexibilidade do modelo em acomodar com precisão diferentes. A potência e a versatilidade do modelo foram, por fim, ilustradas através de sua aplicação a dois conjuntos de dados oncológicos, o que valida sua relevância para a prática clínica.

Palavras-chave: Modelo de fragilidade discreta, fração de cura, versão hurdle, enfoque Bayesiano, zero-modificação, câncer.

ABSTRACT

MOLINA, K. R. C. **Modeling discrete frailty-induced survival data via the hurdle version of the zero-modified power series distribution**. 2025. 133 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.

Modeling unobserved heterogeneity in survival analysis traditionally relies on frailty models. However, their conventional architecture, based on continuous distributions, is often restrictive in contexts exhibiting a structural cure fraction. To address this limitation, this thesis proposes an innovative discrete frailty model based on the Hurdle Zero-Modified Generalized Poisson (HZMGP) distribution. The proposed model is distinguished by a multifaceted architecture that provides three key advantages: its discrete nature offers a natural representation of cured individuals (zero frailty); its zero-modification structure flexibly accommodates diverse patterns of zeros (inflated, deflated, or truncated); and its inclusion of a dispersion parameter captures variability arising from latent risk factors. Inference for the model was developed and validated under both frequentist and Bayesian paradigms. Extensive simulation studies in both frameworks confirmed the estimators' asymptotic properties and, crucially, demonstrated the model's flexibility to accurately capture scenarios both with and without a cure fraction. The model's utility and versatility are ultimately demonstrated through its application to two contrasting oncology datasets, validating its relevance for clinical practice. Furthermore, the Bayesian framework yielded two of the thesis's most significant contributions: a probabilistic classification rule for individual prognostic stratification, and a novel conceptual interpretation of the dispersion parameter as a statistical proxy for unobserved biological heterogeneity. These advances represent a significant contribution to the analysis of complex survival data.

Keywords: Discrete frailty model, cure rate, hurdle version, Bayesian approach, zero-modified, cancer.

LISTA DE ILUSTRAÇÕES

LISTA DE ALGORITMOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AJCC	Comitê Conjunto Americano sobre Câncer
AS	Análise de Sobrevivência
DP	Desvio Padrão
EMV	Estimativa de Máxima Verossimilhança
EP	Erro Padrão
f.g.p.	função geradora de probabilidade
FOSP	Fundação Oncocentro de São Paulo
GP	Poisson Generalizada
HMC	Monte Carlo Hamiltoniano
HZMB	Hurdle Binomial Zero-Modificada
HZMG	Hurdle Geométrica Zero-Modificada
HZMGP	Hurdle Poisson Generalizada Zero-Modificada
HZMNB	Hurdle Binomial Negativa Zero-Modificada
HZMP	Hurdle Poisson Zero-Modificada
HZMPS	Hurdle Série de Potência Zero-Modificada
IC	Intervalo de Confiança
ICr	Intervalos de Credibilidade
INLA	Aproximações de Laplace Aninhadas Integradas
K-M	Kaplan-Meier
LUAD	Adenocarcinoma de pulmão
MC	Monte Carlo
MCMC	Monte Carlo via Cadeias de Markov
MEMV	Média das Estimativas de Máxima Verossimilhança
NB	Binomial Negativa
NSCLC	câncer de pulmão de células não pequenas
NUTS	No-U-Turn Sampler
PC	Probabilidade de Cobertura
PS	Série de Potência
REQM	Raiz do Erro Quadrático Médio
WHO	Organização Mundial da Saúde
ZDB	Binomial Zero-Deflacionada

ZDG	Geométrica Zero-Deflacionada
ZDGP	Poisson Generalizada Zero-Deflacionada
ZDNB	Binomial Negativa Zero-Deflacionada
ZDP	Poisson Zero-Deflacionada
ZDPS	Série de Potência Zero-Deflacionada
ZIB	Binomial Zero-Inflacionada
ZIG	Geométrica Zero-Inflacionada
ZIGP	Poisson Generalizada Zero-Inflacionada
ZINB	Binomial Negativa Zero-Inflacionada
ZIP	Poisson Zero-Inflacionada
ZIPS	Série de Potência Zero-Inflacionada
ZMPS	Série de Potência Zero-Modificada
ZTB	Binomial Zero-Truncada
ZTG	Geométrica Zero-Truncada
ZTGP	Poisson Generalizada Zero-Truncada
ZTNB	Binomial Negativa Zero-Truncada
ZTP	Poisson Zero-Truncada
ZTPS	Série de Potência Zero-Truncada

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Análise de Sobrevivência (AS) é uma área fundamental da Estatística, dedicada à modelagem do tempo até a ocorrência de um evento de interesse. Este evento pode representar uma vasta gama de transições de estado, como o diagnóstico ou complicação de uma doença, o óbito de um paciente, a recidiva de uma condição, a concepção, o matrimônio, ou a falha de um componente eletrônico, entre outros. A variável central nestes estudos é o tempo decorrido até o evento, tipicamente uma variável aleatória contínua e positiva, cuja mensuração exige a definição precisa de um marco temporal inicial para o acompanhamento.

Uma característica distintiva dos dados de sobrevivência é a frequente presença de censura, que representa uma observação incompleta do tempo até o evento para alguns indivíduos. Para lidar com dados censurados, uma das técnicas mais consolidadas é o estimador produto-limite de Kaplan-Meier, proposto por Kaplan e Meier (1958). Trata-se de um método não paramétrico que estima a função de sobrevivência a partir das ocorrências de eventos e censuras registradas ao longo do tempo.

Para avaliar o impacto de fatores explicativos nos tempos de sobrevivência, recorre-se a modelos de regressão. Dentre estes, destaca-se o modelo de riscos proporcionais de Cox (COX, 1972), amplamente utilizado na literatura. Contudo, o modelo de Cox padrão assume que a função de risco de base é idêntica para todos os indivíduos, uma vez condicionada às covariáveis observadas. Esta suposição pode ser restritiva, pois frequentemente existe uma heterogeneidade não observada entre os sujeitos, mesmo após o controle pelas covariáveis disponíveis. Fatores como histórico médico detalhado, variações genotípicas ou estilos de vida específicos podem não ser mensuráveis ou estar indisponíveis, o que introduz uma variabilidade não explicada nos tempos de sobrevivência.

Para superar essa limitação, Vaupel, Manton e Stallard (1979) introduziram o conceito de fragilidade: uma variável aleatória latente e não observada que atua multiplicativamente sobre o risco individual. Esta variável representa o efeito conjunto de fatores de risco não mensurados

ou desconhecidos. A ideia central dos modelos de fragilidade é que os indivíduos possuem níveis intrínsecos e distintos de suscetibilidade, de modo que aqueles considerados mais frágeis tendem a experimentar o evento de interesse mais precocemente do que os menos frágeis (WIENKE, 2010).

Diante da necessidade de modelar a fragilidade, uma questão fundamental reside na escolha da distribuição de probabilidade para esta variável latente. Historicamente, diversas distribuições de probabilidade contínuas e não negativas têm sido propostas para modelar a fragilidade. Entre as mais comuns estão a gamma (VAUPEL; MANTON; STALLARD, 1979; SCUDILIO *et al.*, 2019), inversa gaussiana (HOUGAARD, 1984), log-normal (MCGILCHRIST; AISBETT, 1991; SANTOS; DAVIES; FRANCIS, 1995), estável positiva (PHILIP, 1986), gamma generalizada (BALAKRISHNAN; PENG, 2006; CHEN; ZHANG; ZHANG, 2013), Lomax (TOMAZELLA; MARTINS; BERNARDO, 2008), Birnbaum-Saunders (LEÃO *et al.*, 2017; LEÃO *et al.*, 2018), a família função de variação de potência (CALSAVARA *et al.*, 2017; CALSAVARA *et al.*, 2020) e a Lindley ponderada (GAZON *et al.*, 2022). A preferência por distribuições contínuas é, em grande parte, motivada por suas vantagens computacionais, pois muitas delas, por meio de suas transformadas de Laplace, permitem obter expressões de forma fechada para as funções de sobrevivência e de risco. No entanto, uma limitação importante desses modelos emerge em cenários onde existe uma fração de cura, ou seja, uma proporção de indivíduos que, potencialmente, jamais experimentarão o evento de interesse. A maioria das distribuições contínuas para fragilidade implica um risco que, embora possa diminuir, nunca se torna permanentemente nulo, sendo, portanto, inerentemente incompatíveis com a presença de indivíduos imunes ou curados (ONCHERE, 2013).

Nos últimos anos, avanços nos tratamentos médicos e intervenções têm gerado cenários em que uma proporção significativa de indivíduos pode reverter a progressão de uma doença ou mantê-la sob controle indefinidamente. Tal fato evidencia a existência de subpopulações que podem ser consideradas imunes ou curadas, ou seja, que não apresentarão o evento de interesse mesmo após um longo período de acompanhamento (MALLER; ZHOU, 1996; IBRAHIM; CHEN; SINHA, 2001). Precisamente para abordar tais cenários, os modelos com fragilidade discreta emergem como uma alternativa metodológica promissora. A natureza discreta dessa abordagem permite acomodar uma probabilidade positiva no valor zero da fragilidade, o que pode ser interpretado naturalmente como imunidade ou risco nulo para uma fração da população.

Diversos autores têm explorado distribuições discretas para modelar a fragilidade na literatura recente. Exemplos notáveis incluem as distribuições Poisson, geométrica e binomial negativa (CARONI; CROWDER; KIMBER, 2010); o processo de Poisson composto discreto (ATA; ÖZEL, 2013); a distribuição hiper-Poisson (SOUZA *et al.*, 2017); a família Série de Potência Zero-Inflacionada (ZIPS)[1] (CANCHO *et al.*, 2020); e, mais recentemente, a família

[1] Do inglês: *Zero-Inflated Power Series*

Série de Potência Zero-Modificada (ZMPS)[2] (MOLINA *et al.*, 2021). Adicionalmente, Cancho *et al.* (2021) empregou a distribuição Poisson mista, e a distribuição de Katz foi explorada por Santo *et al.* (2022). A pesquisa e a aplicação de modelos de fragilidade baseados em distribuições discretas constituem, portanto, um campo de pesquisa ativo e relevante, motivado pela necessidade de modelar adequadamente populações com frações de cura.

Neste contexto, propõe-se um novo modelo de análise de sobrevivência com fragilidade discreta baseado na versão Hurdle da distribuição Poisson Generalizada Zero-Modificada (HZMGP). O modelo é motivado por contextos em que há evidência de uma fração de indivíduos imunes ao evento de interesse, como em estudos clínicos com presença de cura. A estrutura discreta da fragilidade permite a representação explícita de indivíduos com risco nulo (fragilidade zero), acomodando diferentes configurações de dados com zeros estruturais (zero-inflação, zero-deflação e zero-truncamento). O modelo também incorpora um parâmetro de dispersão, que capta a variabilidade não explicada pelas covariáveis observadas.

A estimação dos parâmetros em modelos de sobrevivência, especialmente naqueles que incorporam termos de fragilidade, pode ser complexa. Embora abordagens baseadas em máxima verossimilhança sejam tradicionalmente utilizadas, o enfoque Bayesiano tem ganhado proeminência, oferecendo uma alternativa flexível e poderosa para a inferência (IBRAHIM; CHEN; SINHA, 2001). A abordagem Bayesiana permite incorporar conhecimento prévio por medio de distribuições a priori, embora prioris não informativas sejam comuns, e lida de forma natural com a complexidade inerente a este tipo de modelagem. Frequentemente, recorre-se a métodos computacionais intensivos, como as técnicas de Monte Carlo via Cadeias de Markov (MCMC)[3], que viabilizam a amostragem da distribuição a posteriori conjunta de todos os parâmetros do modelo, incluindo os termos de fragilidade individuais. Isso não só facilita a estimação em modelos complexos, mas também fornece uma interpretação direta da incerteza por meio de intervalos de credibilidade e permite obter previsões e avaliar a adequação do modelo de forma integrada. A flexibilidade da inferência Bayesiana é particularmente vantajosa ao lidar com estruturas hierárquicas, dados censurados e modelos com características não padrão, como as frações de cura, representando uma área de pesquisa e aplicação contínua e ativa, como evidenciado por trabalhos recentes (SOUZA *et al.*, 2017; SANTOS; CANCHO; RODRIGUES, 2019; RODRIGUES *et al.*, 2021; ALVARES *et al.*, 2021; CANCHO *et al.*, 2024; RODRIGUES *et al.*, 2024; DABADE, 2024).

Finalmente, para demonstrar a aplicabilidade prática e a robustez da metodologia desenvolvida, tanto sob o enfoque frequentista quanto Bayesiano, o modelo de fragilidade proposto será aplicado a dois conjuntos de dados oncológicos distintos: um estudo sobre câncer de melanoma e outro referente a câncer de pulmão. Estas aplicações visam evidenciar a versatilidade do modelo em diferentes cenários clínicos.

---

[2] Do inglês: *Zero-Modified Power Series*

[3] Do inglês: *Monte Carlo via Markov Chains*

1.1 Objetivos

Objetivo Geral

O objetivo central deste trabalho é propor e desenvolver um novo modelo de análise de sobrevivência que utiliza a distribuição Hurdle Poisson Generalizada Zero-Modificada (HZMGP)[4] para modelar a fragilidade individual. O propósito é explorar a flexibilidade desta distribuição para capturar e interpretar diferentes cenários de zero (zero-inflação, zero-deflação ou zero-truncado) associados ao prognóstico clínico dos pacientes, bem como quantificar a heterogeneidade não observada através do seu parâmetro de dispersão.

Objetivos Específicos

Os objetivos específicos deste trabalho são:

1. Desenvolver os procedimentos de inferência para a estimação dos parâmetros do modelo, abrangendo tanto a abordagem clássica quanto a Bayesiana e detalhando, para a metodologia Bayesiana, a especificação da função de verossimilhança, as distribuições a priori e a implementação computacional de algoritmos MCMC, utilizando a plataforma Stan.

2. Avaliar o desempenho das metodologias de inferência (frequentista e Bayesiana) por meio de estudos de simulação de Monte Carlo, a fim de verificar a capacidade de recuperação dos parâmetros do modelo sob diferentes cenários de zeros e tamanhos amostrais.

3. Analisar a interpretação clínica associada aos diferentes cenários de zeros (zero-inflação, zero-deflação, zero-truncamento) na fragilidade, investigando como essa informação pode fornecer *insights* sobre o prognóstico dos pacientes e auxiliar na identificação de subgrupos clinicamente relevantes.

4. Investigar o papel e a interpretação do parâmetro de dispersão da distribuição HZMGP no contexto da heterogeneidade individual.

5. Aplicar a metodologia desenvolvida na análise de dois conjuntos de dados reais de oncologia, câncer de melanoma e câncer de pulmão, interpretando os resultados obtidos à luz do contexto biomédico.

1.2 Conjunto de dados

O câncer engloba um vasto grupo de doenças caracterizadas pelo crescimento descontrolado de células anormais que podem invadir tecidos adjacentes e metastizar para outras partes do corpo (National Cancer Institute, 2021). Constitui uma das principais causas de morbidade

[4] Do inglês: *Hurdle Zero-Modified Generalized Poisson*

e mortalidade em todo o mundo, representando um desafio significativo para a saúde pública global. Segundo a Organização Mundial da Saúde (WHO)[5], o câncer foi responsável por quase 10 milhões de mortes em 2020, e sua incidência global continua a aumentar, impulsionada por fatores como o envelhecimento populacional e a prevalência de fatores de risco conhecidos (World Health Organization, 2024).

Observaram-se avanços notáveis no diagnóstico e tratamento do câncer, que resultaram em melhorias gerais nas taxas de sobrevida para muitos tipos de tumores, especialmente quando detectados em estágios iniciais (SIEGEL; MILLER; JEMAL, 2020). No entanto, o prognóstico permanece altamente variável, dependendo criticamente do tipo específico de câncer, do estágio no momento do diagnóstico, das características moleculares do tumor, da disponibilidade de tratamentos eficazes e de fatores individuais do paciente. Apesar desses progressos, muitos tipos de câncer continuam a apresentar taxas de mortalidade elevadas, particularmente quando diagnosticados em fases avançadas ou metastáticas (SUNG *et al.*, 2021).

A compreensão dos fatores que influenciam a sobrevida, considerando a heterogeneidade entre pacientes e a possibilidade de cura para alguns, exige modelos estatísticos sofisticados. A análise de sobrevivência, incluindo modelos que incorporam fragilidade e outras características complexas, torna-se, portanto, uma ferramenta essencial na pesquisa oncológica para avaliar prognósticos e o impacto de intervenções terapêuticas.

Os dados analisados neste trabalho, compreendendo coortes de pacientes com diagnóstico de câncer de melanoma e câncer de pulmão, foram obtidos do Registro Hospitalar de Câncer (RHC) mantido pela Fundação Oncocentro de São Paulo (FOSP). Ressalta-se que informações agregadas provenientes do RHC da FOSP encontram-se disponíveis para consulta pública em seu portal eletrônico: <https://fosp.saude.sp.gov.br/>.

1.2.1 Câncer de melanoma

Dentre as diversas neoplasias malignas de pele, o melanoma cutâneo ocupa uma posição de destaque devido à sua gravidade. Sua importância clínica é desproporcional devido à sua alta agressividade e elevado potencial metastático (SUNG *et al.*, 2021). Globalmente, a incidência do melanoma tem apresentado um crescimento expressivo nos últimos anos, tornando-o um foco crescente de preocupação em saúde pública, sendo a exposição excessiva à radiação ultravioleta solar ou artificial o principal fator de risco ambiental identificável (ORGANIZATION, 2024).

A suspeita clínica de melanoma frequentemente se baseia na avaliação de lesões melanocíticas que apresentam características sugestivas de malignidade, como assimetria, bordas irregulares, variação de cores, diâmetro usualmente superior a 6 mm, e principalmente, mudanças na aparência ao longo do tempo. Qualquer lesão com estas características, ou uma lesão nova com aspecto incomum, justifica avaliação dermatológica especializada, frequentemente culminando

[5] Do inglês: *World Health Organization*

em uma biópsia excisional para análise histopatológica (INSTITUTE, 2024). A detecção precoce é um pilar fundamental no manejo do melanoma, pois quando diagnosticado em seus estágios iniciais, o tratamento cirúrgico oferece altas taxas de cura.

O estadiamento preciso, utilizando o sistema proposto pelo Comitê Conjunto Americano sobre Câncer (AJCC)[6], é essencial para determinar o prognóstico e orientar a estratégia terapêutica (GERSHENWALD *et al.*, 2017). Existe uma forte correlação entre o estadiamento clínico no momento do diagnóstico e a sobrevida do paciente. Os estágios iniciais (I e II), onde a doença está restrita à pele, associam-se a um prognóstico alentador, com taxas de sobrevida superiores a 90% para o estágio I após ressecção cirúrgica. Em contraste, pacientes diagnosticados em estágios mais avançados, como o estágio III (com envolvimento de linfonodos regionais) ou estágio IV (presença de metástases), enfrentam um prognóstico significativamente mais reservado, com taxas de sobrevida em 10 anos que podem variar amplamente, caindo para menos de 30% em alguns subgrupos do estágio IV (GERSHENWALD *et al.*, 2017). Historicamente, o tratamento do estadiamento clínico IV apresentava resultados limitados.

Contudo, a última década testemunhou uma revolução terapêutica com o advento e a consolidação da imunoterapia e das terapias-alvo. Estas abordagens melhoraram significativamente as taxas de resposta e a sobrevida em pacientes com doença avançada, embora a gestão de toxicidades, a resistência ao tratamento e a durabilidade da resposta permaneçam como desafios e áreas de intensa pesquisa (DAVIS; SHALIN; TACKETT, 2019).

No contexto brasileiro, é o tipo de câncer mais frequente em ambos os gêneros, correspondendo a cerca de 30% de todos os tumores malignos registrados no país. Dentro deste cenário, o melanoma cutâneo representa uma proporção menor, estimada em torno de 3% a 4% dos casos de câncer de pele (VIGILÂNCIA, 2018). Apesar dessa frequência relativa mais baixa, a alta letalidade associada ao melanoma, especialmente quando diagnosticado tardiamente, o consolida como um relevante problema de saúde pública no Brasil, reforçando a importância de estratégias de prevenção primária, rastreamento em populações de risco e diagnóstico precoce para melhorar os desfechos clínicos.

Para aprofundar a compreensão dos fatores prognósticos e da sobrevida de pacientes com câncer de melanoma no cenário paulista, foi analisada uma coorte de um total de 6741 pacientes diagnosticados com melanoma cutâneo entre os anos de 2000 e 2014. O período de seguimento dos pacientes se estendeu até o ano de 2018, resultando em um tempo máximo de observação de aproximadamente 18 anos. O evento de interesse definido para a análise de sobrevida foi o óbito especificamente atribuído ao melanoma. Durante o período de acompanhamento, observou-se que $71,67\%$ dos pacientes foram censurados.

É relevante notar que este conjunto de dados foi objeto de análise prévia por Calsavara *et al.* (2020). No entanto, o estudo mencionado focou especificamente na avaliação do efeito

[6] Do inglês: *American Joint Committee on Cancer*

prognóstico da realização de cirurgia ao longo do tempo, empregando um modelo de riscos não proporcionais. O presente trabalho, por sua vez, expande a análise ao incorporar um conjunto mais amplo de covariáveis disponíveis no registro da FOSP. Consideraram-se outras covariáveis demográficas e clínicas, como: idade ao diagnóstico (em anos), gênero do paciente, estadiamento clínico (EC), e o recebimento ou não de tratamentos adjuvantes como radioterapia (Rad) e quimioterapia (Qui).

A caracterização inicial da coorte, detalhada na Tabela 1, revela que a idade média ao diagnóstico foi de 58,11 anos e uma proporção de gênero equilibrada (49,40% dos pacientes do sexo masculino). Em relação ao EC no momento do diagnóstico, se observa que 44,61% dos pacientes foram classificados como EC I, enquanto 14,35% diagnosticados no EC IV. Quanto aos tratamentos realizados, a grande maioria dos pacientes (88,58%) foi submetida à cirurgia. Ademais, constata-se que apenas uma minoria recebeu tratamentos sistêmicos durante o período de seguimento, sendo 8,71% tratados com radioterapia e 16,36% com quimioterapia.

Tabela 1 – Descrição das covariáveis para os dados de câncer de melanoma.

$\boldsymbol{X}$	Covariável	Descrição	Categoria	%
X_1	Idade	média $= 58,11$ DP $= 16,26$	-	
X_2	Gênero do paciente	Masculino Feminino	0 1	49,40% 50,60%
X_3	Estadiamento Clínico (EC)	EC I EC II EC III EC IV	1 2 3 4	44,61% 22,83% 18,21% 14,35%
X_4	Cirurgia (Cir)	Não Sim	0 1	11,42% 88,58%
X_5	Radioterapia (Rad)	Não Sim	0 1	91,29% 8,71%
X_6	Quimioterapia (Qui)	Não Sim	0 1	83,64% 16,36%

Fonte: Elaborada pelo autor.

A análise exploratória por meio do estimador de K-M, apresentada na Figura 1, revela um platô pronunciado na curva de sobrevivência, que se estabiliza aproximadamente a partir dos 16 anos de seguimento. Este padrão sugere fortemente a existência de uma fração de cura no conjunto de dados, indicando que o modelo de fragilidade HZMGP proposto pode ser particularmente adequado para esta análise. Com base na estimativa de K-M, estima-se que a probabilidade de um paciente com melanoma sobreviver por mais de 16 anos seja de

aproximadamente $0,60$.

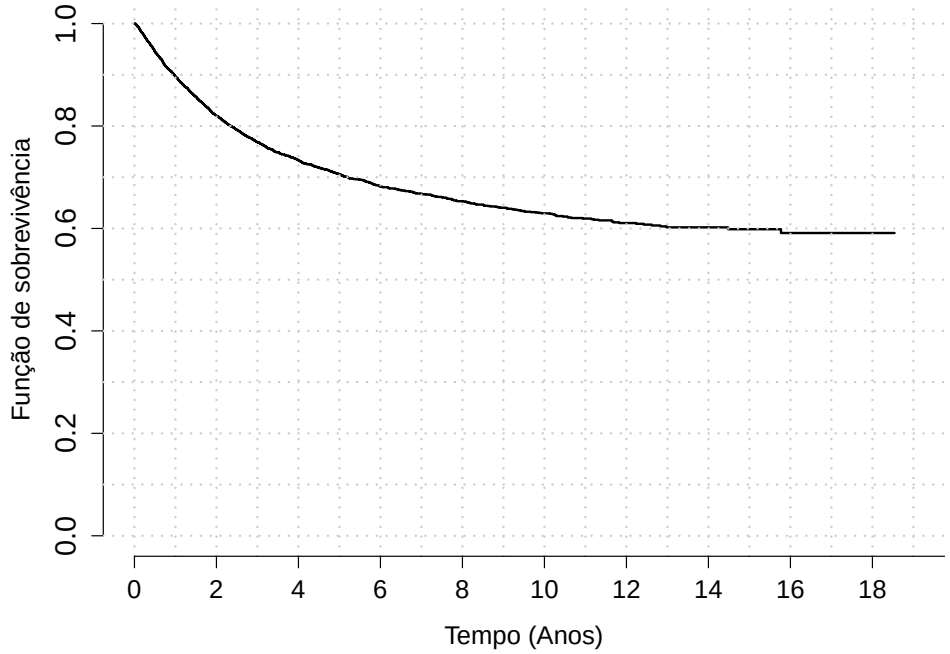

Figura 1 – Curva de Kaplan-Meier para os dados de câncer de melanoma.

Fonte: Elaborada pelo autor.

1.2.2 Câncer de pulmão

O câncer de pulmão representa um desafio proeminente na oncologia mundial, destacando-se como uma das neoplasias mais incidentes e a principal causa de morte por câncer em escala global, afetando significativamente tanto homens quanto mulheres (TAO, 2019; SUNG *et al.*, 2021). Sua alta letalidade é frequentemente atribuída ao diagnóstico tardio, ocorrendo muitas vezes quando a doença já se disseminou, e à agressividade biológica inerente a diversos subtipos histológicos (THAI *et al.*, 2021). Estimativas indicam que o câncer de pulmão foi responsável por aproximadamente 1,8 milhão de mortes em todo o mundo no ano de 2020 (SUNG *et al.*, 2021).

Embora o tabagismo seja o fator de risco predominante, associado a cerca de 85% dos casos diagnosticados, outros fatores ambientais e genéticos também contribuem para sua etiologia (THAI *et al.*, 2021; VIGILÂNCIA, 2018). Nos últimos anos, avanços notáveis na compreensão molecular da doença impulsionaram o desenvolvimento de terapias-alvo e imunoterapias. Essas abordagens têm demonstrado capacidade de melhorar o prognóstico para subgrupos específicos de pacientes, especialmente aqueles com doença avançada que possuem biomarcadores preditivos. No entanto, apesar desses progressos significativos, a sobrevida global em cinco anos para o conjunto dos pacientes com câncer de pulmão ainda representa um desafio considerável (THAI *et al.*, 2021).

No contexto brasileiro, o câncer de pulmão ocupa uma posição de destaque epidemiológico, figurando como o terceiro tipo mais incidente entre homens e o quarto entre mulheres, de acordo com estimativas do Instituto Nacional de Câncer (INCA) (VIGILÂNCIA, 2018). Observa-se uma tendência de diminuição nas taxas de incidência, particularmente acentuada no sexo masculino desde meados da década de 1980, reflexo provável das políticas de controle do tabagismo. Apesar dessa tendência favorável na incidência, a doença continua a representar uma carga substancial de mortalidade no país. A forte associação com o consumo de derivados de tabaco, responsável pela grande maioria dos casos diagnosticados no Brasil, reforça o caráter evitável desta neoplasia e a importância contínua das estratégias de prevenção primária e secundária (VIGILÂNCIA, 2018).

Neste contexto, o presente estudo analisou dados de 30900 pacientes diagnosticados com câncer de pulmão no estado de São Paulo, Brasil. A coorte inclui diagnósticos realizados entre os anos 2000 e 2014, com um período de acompanhamento que se estendeu até o ano de 2018. O evento de interesse foi definido como o óbito decorrente de câncer de pulmão. Uma característica marcante desta população é a alta taxa de mortalidade específica pela doença: observa-se que $88,93\%$ dos pacientes incluídos na análise faleceram devido ao câncer de pulmão durante o período do estudo.

É pertinente mencionar que este conjunto de dados foi previamente analisado por Gazon *et al.* (2022). Naquele estudo, os autores examinaram os efeitos de diversas covariáveis, tratando a idade de forma categorizada (< 60 vs. ≥ 60 anos), dividindo o estadiamento clínico em quatro grupos e classificando as variáveis de tratamento (cirurgia, radioterapia, quimioterapia) como recebido/não recebido, sob a ótica de um modelo de riscos não proporcionais com termo de fragilidade.

No presente estudo, contudo, optou-se por uma abordagem distinta na operacionalização de algumas destas covariáveis. A idade foi analisada como uma variável contínua. Adicionalmente, o estadiamento clínico foi reclassificado em duas categorias: I-II e III-IV. Especificamente, os estágios I e II foram agrupados em uma categoria, enquanto os estágios III e IV foram combinados em outra. Esta decisão de agrupar os estágios mais avançados (III e IV) foi motivada pela observação da ausência de diferenças significativas entre eles. Maiores detalhes sobre a definição e categorização das covariáveis utilizadas nesta análise são apresentados na Tabela 2.

A caracterização da coorte em estudo demonstrou que a idade média dos pacientes é aproximadamente 63 anos, sendo 36,39% do sexo feminino. Quanto ao perfil clínico, constatou-se que 84,05% dos pacientes foram diagnosticados em estágio avançado da doença (EC III-IV). Em relação às intervenções terapêuticas, a quimioterapia foi administrada a 65,16% dos indivíduos, enquanto a cirurgia e a radioterapia foram menos frequentes, não sendo realizadas em 80,91% e 61,84% dos casos, respectivamente.

Tabela 2 – Descrição das covariáveis para os dados de câncer de pulmão.

$\boldsymbol{X}$	Covariável	Descrição	Categoria	%
X_1	Idade	média $= 63,21$ DP $= 10,98$	-	
X_2	Gênero do paciente	Masculino Feminino	0 1	$63,61\%$ $36,39\%$
X_3	Estadiamento Clínico (EC)	EC I-II EC III-IV	0 1	15.95% 84.05%
X_4	Cirurgia (Cir)	Não Sim	0 1	$80,91\%$ $19,09\%$
X_5	Radioterapia (Rad)	Não Sim	0 1	$61,84\%$ $38,16\%$
X_6	Quimioterapia (Qui)	Não Sim	0 1	$34,84\%$ $65,16\%$

Fonte: Elaborada pelo autor.

A análise exploratória inicial por meio da curva de K-M para estes dados, apresentada na Figura 2, exibe um comportamento distinto: um rápido declínio inicial na probabilidade de sobrevivência, seguido por uma diminuição mais gradual, porém contínua, ao longo de todo o período de acompanhamento. A ausência de um platô e a clara tendência da curva em direção a zero sugerem a inexistência de uma fração de cura clinicamente significativa para estes pacientes. O declínio contínuo indica que o risco de óbito atribuível à doença permanece presente mesmo para os indivíduos que sobrevivem por períodos mais longos, apresentando um desafio de modelagem diferente do cenário anterior.

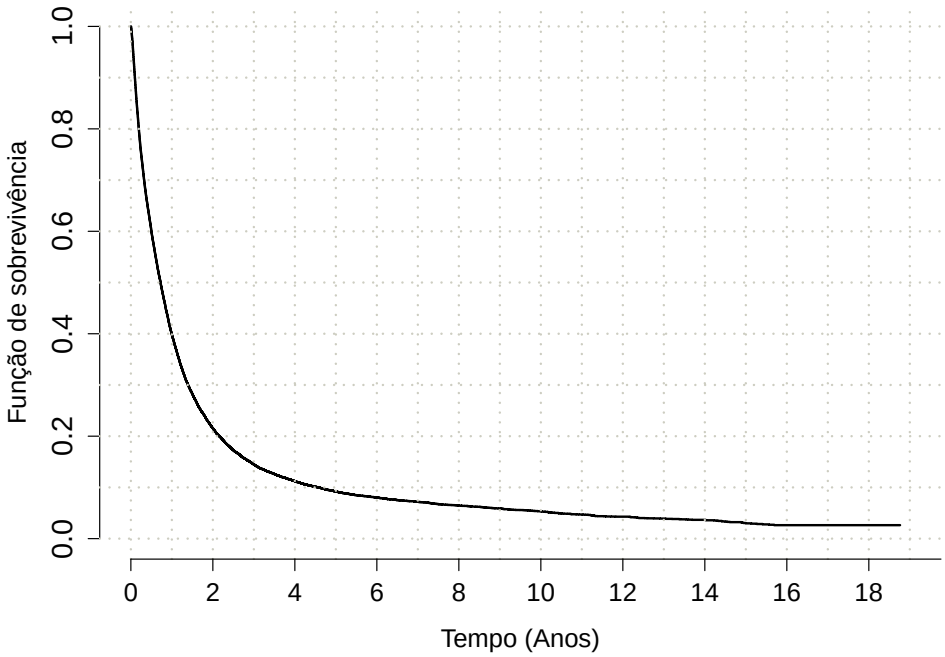

Figura 2 – Curva de Kaplan-Meier para os dados de câncer do pulmão.

Fonte: Elaborada pelo autor.

1.3 Organização dos capítulos

A presente tese está organizada da seguinte forma

O Capítulo 2 é dedicado a uma revisão dos conceitos fundamentais e da literatura pertinente que sustentam este trabalho, abordando tópicos essenciais para a compreensão dos capítulos subsequentes. O Capítulo 3, por sua vez, detalha a formulação matemática e as propriedades da família de distribuições Hurdle Série de Potência Zero-Modificada (HZMPS)[7]. Subsequentemente, o Capítulo 4 introduz os modelos de análise de sobrevivência que incorporam a fragilidade discreta baseada na distribuição HZMGP, desenvolvendo a metodologia de inferência sob a perspectiva frequentista. O Capítulo 5 aborda a mesma estrutura de modelagem sob o enfoque Bayesiano, detalhando os procedimentos de estimação e as técnicas de inferência correspondentes. Ressalta-se que ambos os capítulos, 4 e 5, incluem estudos de simulação de Monte Carlo (MC), projetados para avaliar o desempenho das respectivas metodologias, bem como a aplicação prática dos modelos a dois conjuntos de dados reais: câncer de melanoma e câncer de pulmão. Por fim, o Capítulo 6 sumariza as principais contribuições e conclusões desta tese, discutindo as implicações dos resultados e apontando direções para pesquisas futuras.

[7] Do inglês: *Hurdle Zero-Modified Power Series*

1.4 Produtos da tese

A disseminação dos resultados e da metodologia desenvolvidos nesta tese constitui um de seus produtos fundamentais. Nesse contexto, foi redigido e submetido para publicação o artigo intitulado «A Bayesian survival model induced by hurdle zero-modified power series discrete frailty with dispersion: an application in lung cancer». O manuscrito foi encaminhado para avaliação pela revista *Statistical Methods in Medical Research*, um periódico de referência na área de bioestatística e metodologia médica. Este trabalho articula a estrutura inferencial Bayesiana do modelo HZMGP e ilustra sua potência na análise de dados oncológicos, materializando, assim, a contribuição desta tese para a comunidade científica.

CAPÍTULO

2

NOÇÕES BÁSICAS INTRODUTÓRIAS

Para uma compreensão abrangente do desenvolvimento apresentado nesta tese, é fundamental estabelecer o contexto teórico que o sustenta. Este Capítulo 2 tem como objetivo introduzir conceitos e terminologias fundamentais que facilitarão a leitura e o entendimento das análises subsequentes. Serão abordados, de forma introdutória, os princípios da análise de sobrevivência, as particularidades dos modelos de sobrevivência com fração de cura, os modelos de fragilidade e os fundamentos da inferência Bayesiana. A familiarização com estes tópicos fornecerá a base conceitual indispensável para acompanhar a metodologia e os resultados discutidos nos capítulos posteriores.

2.1 Fundamentos da Análise de Sobrevivência

Análise de Sobrevivência constitui um campo fundamental da Estatística, cujo objeto central de estudo é o tempo de falha. Este termo designa o lapso temporal decorrido desde um ponto inicial bem estabelecido até a ocorrência de um evento de interesse específico. Formalmente, este tempo é modelado como uma variável aleatória não negativa, denotada por T. Assim, este ramo da Estatística engloba um conjunto de ferramentas metodológicas desenvolvidas para observar, modelar e analisar dados referentes a esses tempos até o evento de interesse. A relevância e a aplicabilidade destas técnicas são vastas, estendendo-se a diversas áreas como engenharias (tempo até a falha de componentes), ciências sociais (tempo até o casamento ou desemprego), economia (tempo até a falência de empresas) e, notadamente, a área biomédica (tempo até a cura ou recidiva de uma doença, por exemplo) (KLEINBAUM; KLEIN, 2012; COLOSIMO; GIOLO, 2021).

Uma característica distintiva e desafiadora dos dados de sobrevivência é a ocorrência de censura, que se manifesta quando o tempo exato do evento não pode ser observado para todos os indivíduos do estudo. Isso ocorre, por exemplo, quando a pesquisa termina antes da ocorrência do evento ou quando o indivíduo é perdido durante o seguimento. Para distinguir entre eventos

observados e tempos censurados, utiliza-se um indicador de status, δ_i. As formas mais comuns de censura são: $\delta_i = 1$ se o evento foi observado para o indivíduo i no tempo t_i, e $\delta_i = 0$ se o tempo de acompanhamento t_i foi interrompido antes do evento. Os principais tipos de censura são detalhados a seguir.

- Censura à direita: É o tipo mais frequente. Ocorre quando o estudo termina antes que o indivíduo experimente o evento de interesse, ou quando o indivíduo é perdido durante o acompanhamento (por exemplo, abandono do estudo, mudança). Nesses casos, sabe-se apenas que o tempo de sobrevivência T_i do indivíduo é maior que o tempo de observação C_i. Observa-se, portanto $\min(T_i, C_i)$ e um indicador δ_i tal que

$$\delta_i = \begin{cases} 1, \text{ se } T_i \leq C_i, \\ 0, \text{ se } T_i > C_i. \end{cases} \qquad i = 1, \ldots, n,$$

- Censura à esquerda: Ocorre quando se sabe que o evento ocorreu antes de um determinado tempo de observação C, mas o momento exato é desconhecido. Por exemplo, em um estudo sobre a idade de início de uma doença crônica, um participante pode já apresentar a condição no momento de sua inclusão no estudo.

- Censura intervalar: Ocorre quando se sabe que o evento ocorreu dentro de um intervalo de tempo $(L, R]$, mas o momento exato não foi registrado. Essa situação é comum em estudos com avaliações periódicas, nos quais o evento é detectado entre duas visitas consecutivas.

A presença de censura impede, em geral, a aplicação direta de técnicas estatísticas padrão, como os métodos de regressão usuais, que pressupõem observações completas para a variável resposta. Portanto, AS emprega uma metodologia própria, com técnicas específicas capazes de lidar adequadamente com a natureza particular dos tempos de falha e a informação censurada.

2.1.1 Caracterização do tempo de sobrevivência

A variável aleatória não-negativa T, pode ser caracterizada matematicamente de diferentes maneiras. Na AS, duas funções são fundamentais para essa caracterização: a função de sobrevivência e a função de risco. Dada a sua importância, detalham-se a seguir estas e outras funções associadas.

2.1.1.1 Função de sobrevivência

A função de sobrevivência, denotada por $S(t)$, é um conceito central na AS. Matematicamente, ela representa a probabilidade de um indivíduo não experimentar o evento de interesse até um tempo específico t, sendo definida como

$$S(t) = \mathbb{P}(T > t) = 1 - \mathbb{F}(t),$$

em que $\mathbb{F}(t) = \mathbb{P}(T \leq t)$ é a função de distribuição acumulada de T.

A função $S(t)$ possui propriedades importantes: é uma função não crescente em t, com $S(0) = 1$. Adicionalmente, o limite $\lim_{t\to\infty} S(t) = 0$ é válido, pressupondo-se que o evento certamente ocorrerá para todos os indivíduos em algum momento.

2.1.1.2 Função de risco

A função de risco, também conhecida como taxa de falha, denotada por $h(t)$, representa o potencial instantâneo de ocorrência do evento no tempo t, dado que o indivíduo sobreviveu até esse tempo t. A função de risco de T é então definida como

$$h(t) = \lim_{\Delta t\to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}.$$

É importante notar que $h(t) \geq 0$. Uma característica fundamental desta função é sua flexibilidade para assumir diversas formas (constante, crescente, decrescente, não-monotônica), refletindo assim diferentes padrões de risco ao longo do tempo.

2.1.1.3 Função de risco acumulada

A função de risco acumulada, denotada por $H(t)$, quantifica o risco total acumulado até o tempo t. É definida por

$$H(t) = \int_0^t h(u)\,du.$$

A função $H(t)$ é não decrescente, com $H(0) = 0$.

2.1.1.4 Relações entre as funções

Definem-se, a seguir, as principais relações matemáticas entre as funções anteriormente apresentadas

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t)) = -\frac{S'(t)}{S(t)}, \\ H(t) &= -\log S(t), \\ S(t) &= e^{-H(t)}, \\ f(t) &= -\frac{dS(t)}{dt} = h(t)S(t), \end{aligned} \tag{2.1}$$

em que $f(t)$ é a função densidade associada a T.

2.1.2 Métodos de estimação

A análise de dados de sobrevivência envolve a aplicação de métodos estatísticos específicos para estimar funções de interesse e avaliar o impacto de covariáveis, levando em

conta características como a censura. Diversas abordagens foram desenvolvidas para esses fins, podendo ser classificadas, em linhas gerais, em três categorias principais que variam em suas suposições e objetivos. Esta subseção apresentará uma visão geral desses métodos, começando pelas técnicas não paramétricas, passando pelos modelos semiparamétricos e culminando nos modelos paramétricos.

2.1.2.1 Métodos não paramétricos

Esses métodos são caracterizados pela ausência de suposições sobre a forma da distribuição de probabilidade do tempo de sobrevivência T. O estimador mais proeminente nesta categoria é o de Kaplan-Meier (K-M), proposto por Kaplan e Meier (1958). Sua finalidade é obter uma estimativa não paramétrica da função de sobrevivência $S(t)$ diretamente a partir dos dados observados, mesmo na presença de censura à direita.

O estimador K-M, denotado por $\hat{S}(t)$, é uma abordagem empírica que constrói a curva de sobrevivência como um produto de probabilidades condicionais de sobreviver a cada intervalo definido pelos tempos de evento observados

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

onde t_i, são os tempos distintos em que ocorrem os eventos ($t_1 < t_2 < \ldots$), d_i é o número de eventos em t_i, e n_i é o número de indivíduos que estavam em risco imediatamente antes do tempo t_i. A estimativa $\hat{S}(t)$ resultante é uma função escada, constante entre os tempos de evento e que decresce apenas nos instantes em que eventos ocorrem.

Devido à sua simplicidade, o estimador K-M é amplamente utilizado para análise exploratória e visualização de padrões de sobrevivência. Para a comparação formal de curvas de K-M entre diferentes grupos, o teste log-rank é a técnica não paramétrica mais comum. A principal limitação do método reside na incapacidade de incorporar diretamente o efeito de covariáveis (especialmente as contínuas), o que motiva o uso de modelos de regressão.

Embora o K-M seja o método mais comum para estimar $S(t)$, vale mencionar também o estimador de Nelson-Aalen (AALEN, 1978), que fornece uma estimativa não paramétrica da função de risco acumulado, $H(t)$.

2.1.2.2 Modelos semiparamétricos

O modelo mais popular nesta categoria é o modelo de riscos proporcionais, introduzido por Cox (1972). Sua principal característica é a capacidade de avaliar o efeito de um vetor de covariáveis, denotado por $\mathbf{x}$, sobre a função de risco, sem a necessidade de especificar a forma da função de risco base, $h_0(t)$. O modelo postula que a função de risco para um indivíduo com covariáveis $\mathbf{x}$ é expressa por

$$h(t \mid \mathbf{x}) = h_0(t) e^{(\mathbf{x}^\top \boldsymbol{\beta})}, \tag{2.2}$$

em que $\mathbf{x}^\top = (1, x_1, \ldots, x_p)$ é o vetor de p covariáveis observadas, e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ é o vetor de coeficientes de regressão a serem estimados.

A suposição fundamental deste modelo é a de riscos proporcionais, que estabelece que a razão entre as funções de risco de dois indivíduos com vetores de covariáveis distintos permanece constante ao longo do tempo. O termo $e^{(\mathbf{x}^\top \boldsymbol{\beta})}$ representa a razão de risco que compara o risco de um indivíduo com covariáveis $\mathbf{x}$ ao de um indivíduo de referência (cujo vetor de covariáveis é $\mathbf{x} = \mathbf{0}$). A estimação dos coeficientes $\boldsymbol{\beta}$ é tipicamente realizada por meio da maximização da função de verossimilhança parcial, conforme proposto (COX, 1975).

2.1.2.3 Modelos paramétricos

Em contraste com as abordagens anteriores, os modelos paramétricos para AS partem da suposição de que o tempo até o evento, T, segue uma distribuição de probabilidade específica, pertencente a uma família paramétrica conhecida. Distribuições comumente utilizadas neste contexto incluem a exponencial, Weibull, log-normal, log-logística, gama, entre outras. Essa suposição permite a construção de uma função de verossimilhança completa para os dados observados, incluindo as censuras, o que subsequentemente permite obter expressões analíticas diretas para as funções de sobrevivência e de risco, bem como para outras quantidades de interesse. A principal vantagem desta abordagem reside na eficiência da estimação e na possibilidade de uma descrição completa da distribuição do tempo de falha. Contudo, sua validade depende crucialmente da adequação da distribuição paramétrica escolhida aos dados. Para uma exploração aprofundada de modelos paramétricos, suas propriedades e métodos de diagnóstico, recomenda-se a consulta ao trabalho abrangente de Colosimo e Giolo (2021).

2.2 Modelos de sobrevivência de longa duração

Em diversas aplicações da AS, particularmente em estudos oncológicos, o pressuposto convencional de que todos os indivíduos eventualmente experimentarão o evento de interesse pode não ser realista. Frequentemente, observa-se uma proporção de indivíduos que são efetivamente curados ou se tornam imunes ao evento. Modelos de sobrevivência padrão, que assumem $\lim_{t\to\infty} S(t) = 0$, não conseguem acomodar o platô observado na cauda da curva de sobrevivência em tais cenários. Para abordar esta questão, duas classes principais de modelos foram historicamente propostas: os modelos de mistura e os modelos baseados em riscos competitivos latentes, nos quais o evento ocorre apenas se um ou mais riscos não observáveis forem ativados.

A abordagem mais tradicional é o modelo de mistura, proposto por Boag (1949) e formalizado por Berkson e Gage (1952). A premissa central é a de que a população é uma heterogeneidade composta por dois subgrupos: os indivíduos suscetíveis, que permanecem sob risco, e os imunes (ou curados). A função de sobrevivência populacional é expressa como

$$S_{pop}(t) = p_0 + (1 - p_0)S(t),$$

em que p_0 representa a fração de cura e $S(t)$ é a função de sobrevivência própria para a subpopulação suscetível.

A função de sobrevivência populacional possui as seguintes propriedades:

- Se $p_0 = 1$, então $S_{pop}(t) = S(t)$;
- $S_{pop}(0) = 1$;
- $S_{pop}(t)$ é não crescente;
- $\lim_{t\to\infty} S_{pop}(t) = p_0$.

Ambas as componentes, p_0 e $S(t)$, podem ser modeladas em função de covariáveis, permitindo investigar os fatores que influenciam tanto a probabilidade de cura quanto o tempo de sobrevida dos indivíduos suscetíveis.

Modelos mais complexos de longa duração, surgiram com o objetivo de explicar melhor os efeitos biológicos envolvidos. Rodrigues *et al.* (2009) propuseram uma teoria unificada que oferece uma interpretação biológica mais robusta, baseada em um processo de dois estágios:

Estágio de iniciação: seja N uma variável aleatória representando o número de causas ou riscos competitivos da ocorrência do evento de interesse. A causa de ocorrência do evento é desconhecida, e a variável N não é observada, com distribuição de probabilidade p_n e sua cauda dadas, respectivamente, por

$$p_n = \mathbb{P}[N = n] \quad \text{e} \quad q_n = \mathbb{P}[N > n], \quad n = 0, 1, 2, \ldots.$$

Estágio de maturação: dado $N = n$, sejam Z_k, $k = 1, ..., n$, variáveis aleatórias contínuas (não-negativas) independentes com função distribuição acumulada $F(z) = 1 - S(z)$ e independentes de N, que representam o tempo de ocorrência do evento de interesse devido à k-ésima causa. A fim de incluir os indivíduos que não são suscetíveis ao evento de interesse, seu tempo de ocorrência é definido como

$$T = \min(Z_0, Z_1, Z_2, ..., Z_N),$$

em que $\mathbb{P}[Z_0 = \infty] = 1$, admitindo a possibilidade que uma proporção p_0 da população não apresenta a ocorrência do evento de interesse.

A formulação reside no uso de funções geradoras de probabilidade. Se $A(s) = \sum_{n=0}^{\infty} p_n s^n$ é a função geradora de N, a função de sobrevivência populacional, $S_{pop}(t)$, pode ser expressa da

seguinte forma

$$\begin{aligned} S_{pop}(t) &= \mathbb{P}[N=0] + \mathbb{P}[Z_1 > t, Z_2 > t, ..., Z_N > t, N \geq 1] \\ &= \mathbb{P}(N=0) + \sum_{n=1}^{\infty} \mathbb{P}(N=n)\mathbb{P}(Z_1 > t, \ldots, Z_N > t) \\ &= p_0 + \sum_{n=1}^{\infty} p_n [S(t)]^n \\ &= A(S(t)). \end{aligned}$$

Desta forma, a função de sobrevivência populacional, $S_{pop}(t)$, emerge da composição da função geradora de N e da função de sobrevivência. A fração de cura, p_0, é uma propriedade intrínseca desta estrutura, definida como a probabilidade de um indivíduo ser imune ao evento, $\mathbb{P}(N=0)$. Este valor é diretamente obtido pelo comportamento assintótico da função de sobrevivência e tem-se que $\lim_{t\to\infty} S_{pop}(t) = A\left(\lim_{t\to\infty} S(t)\right) = A(0) = p_0$. Assim, p_0 quantifica precisamente a proporção de indivíduos que nunca experimentarão o evento de interesse.

Em síntese, os modelos de longa duração constituem uma classe de ferramentas estatísticas fundamentais para a análise de dados de sobrevivência em contextos nos quais uma fração dos indivíduos jamais experimenta o evento de interesse. Ao relaxarem a suposição tradicional de que todos os indivíduos eventualmente falharão, esses modelos incorporam de forma explícita a possibilidade de cura ou imunidade estrutural, oferecendo estimativas mais realistas e interpretações mais adequadas para diversas aplicações. O modelo unificado, em particular, destaca-se por sua flexibilidade teórica e por proporcionar uma base conceitual sólida para a modelagem de fenômenos complexos de sobrevivência.

2.3 Modelo de fragilidade

Os modelos tradicionais de sobrevivência partem da premissa de que a população de estudo é homogênea, tratando os dados como independentes e identicamente distribuídos. Contudo, essa suposição nem sempre é sustentada, pois os indivíduos são inerentemente distintos. Embora participantes de um estudo possam compartilhar características observáveis como diagnóstico, idade, gênero, tratamento, entre outros, cada um apresenta características únicas. As reações individuais ao tratamento variam, os fatores psicológicos influenciam de maneira distinta, e muitos outros aspectos individuais podem afetar os resultados. Assim, é importante reconhecer que, mesmo em grupos aparentemente homogêneos, existem diferenças significativas a serem consideradas. Essa variabilidade, também conhecida como heterogeneidade, constitui uma fonte significativa de incerteza em estudos epidemiológicos, médicos e em diversas aplicações biológicas. Observa-se que tal variabilidade frequentemente não é diretamente observável na AS, o que pode dificultar sua identificação e avaliação adequada (WIENKE, 2010).

A incorporação do termo de fragilidade na análise é relevante, pois ignorar a heterogeneidade não observada pode levar a conclusões enviesadas. Flinn e Heckman (1982) demonstraram que a omissão de variáveis relevantes (potencialmente não mensuráveis) pode distorcer a estimação dos efeitos das covariáveis incluídas, afetando sua magnitude e até mesmo sua significância estatística. De fato, covariáveis que pareciam significativas podem perder seu efeito aparente após a inclusão da fragilidade, e vice-versa. Ignorar tal heterogeneidade pode, portanto, comprometer substancialmente o ajuste do modelo e a validade de suas aplicações práticas. Lancaster (1990), por exemplo, observou uma subestimação dos efeitos das covariáveis na análise de taxas de desemprego quando a fragilidade era omitida, enquanto Pickles *et al.* (1994) sugeriram que essa omissão tende a atenuar as estimativas dos parâmetros das covariáveis, aproximando-as de zero. Portanto, a modelagem adequada da heterogeneidade não observada é crucial para a validade das inferências.

Neste contexto, o modelo de fragilidade, introduzido por Vaupel, Manton e Stallard (1979), surge como uma abordagem coerente, incorporando um termo aleatório multiplicativo na função de risco de cada indivíduo para capturar essa heterogeneidade latente. Esse termo reflete informações que não podem ser ou não foram consideradas no estudo, proporcionando uma forma de capturar a heterogeneidade não observada entre os participantes. Adotando a abordagem de riscos proporcionais (COX; OAKES, 1984), o risco condicionado à fragilidade na ausência de covariáveis observáveis, é definido por

$$h(t \mid Z) = Zh_0(t), \tag{2.3}$$

em que Z denota a variável aleatória de fragilidade. Na expressão (2.3), valores de $Z > 1$ indicam um risco aumentado em relação à linha de base, enquanto $Z < 1$ indica um risco diminuído. O caso $Z = 1$ recupera o modelo de riscos proporcionais padrão sem covariáveis. O fato de a variável de fragilidade influenciar multiplicativamente a função de risco significa que, quanto maior o valor dessa variável, maior será a probabilidade de o indivíduo experimentar o evento de interesse. Em outras palavras, quanto maior for o valor de z_i, mais suscetível o indivíduo i estará a falhar. Dessa forma, espera-se que o evento de interesse ocorra com maior frequência nos indivíduos mais frágeis (WIENKE, 2010).

A função de sobrevivência condicionada à fragilidade é, portanto, expressa como

$$S(t \mid Z) = e^{-ZH(t)} = S_0(t)^Z, \tag{2.4}$$

em que $S_0(t)$ é a função de sobrevivência base, válida para todos os indivíduos do estudo.

A função de sobrevivência marginal, obtida a partir da função de sobrevivência condicionada à fragilidade, corresponde à probabilidade de sobrevivência incondicional para um indivíduo da população em estudo e é expressa como segue

$$S(t) = \int_0^\infty S(t \mid Z)\, f_Z(z)\, dz = \mathscr{L}_Z\left[-\log S_0(t)\right],$$

em que $f_Z(z)$ e $\mathscr{L}_Z[\cdot]$ representam, respectivamente, a função de densidade e a transformada de Laplace da distribuição de fragilidade. Para modelar o termo latente Z, é comum assumir uma distribuição contínua e não negativa. A escolha por distribuições contínuas oferece vantagens computacionais significativas, pois, através de suas transformadas de Laplace, frequentemente se obtêm expressões de forma fechada para as funções de sobrevivência e de risco marginais.

A identificabilidade é um atributo essencial em modelagem estatística, pois governa a capacidade de recuperar de forma única os parâmetros do modelo a partir de dados observados (MCCALL, 1994). Essa relevância é acentuada em modelos semiparamétricos com variáveis latentes, que frequentemente permitem a especificação de estruturas complexas cuja estimação, contudo, pode ser inviável sem uma rigorosa análise de identificabilidade (IACHINE, 2004). No contexto do modelo proposto em (2.3), a garantia de identificabilidade para distribuições de fragilidade contínuas é obtida por meio de uma padronização. Especificamente, restringe-se a esperança da variável de fragilidade a ser unitária, $\mathbb{E}(Z) = 1$, pressupondo sua existência. A variância da fragilidade, denotada por $\mathrm{Var}(Z) = \nu > 0$, quantifica o grau de heterogeneidade não observada no risco basal da população. Uma variância $\nu \longrightarrow 0$ implica que os indivíduos são relativamente homogêneos em seu risco, ao passo que uma variância maior reflete uma heterogeneidade substancial entre os riscos individuais (WIENKE, 2010).

Um problema importante em modelos de fragilidade é a escolha da distribuição para o efeito aleatório. Devido à forma como o termo de fragilidade atua na função de risco, as candidatas à distribuição de fragilidade são supostamente não negativas, usualmente contínuas e não dependentes do tempo. No entanto, em diversas aplicações, especialmente em contextos biomédicos ou de confiabilidade, pode ser mais apropriado representar a fragilidade por uma variável aleatória discreta, permitindo incorporar de forma explícita a existência de subgrupos com diferentes níveis de suscetibilidade ao evento de interesse. Essa abordagem é conhecida como modelo de fragilidade discreta. Um dos principais atrativos desse tipo de modelagem é a capacidade de representar indivíduos com fragilidade nula, isto é, que não estão sujeitos ao evento, o que permite, por exemplo, interpretar tais indivíduos como curados ou estruturalmente imunes.

2.4 Inferência Bayesiana

A inferência Bayesiana oferece um paradigma fundamental e distinto para a análise estatística. Em contraste com a abordagem frequentista, que trata os parâmetros do modelo, $\boldsymbol{\vartheta}$, como quantidades fixas e desconhecidas, o enfoque Bayesiano os concebe como variáveis aleatórias. Isso permite representar a incerteza sobre seus verdadeiros valores por meio de distribuições de probabilidade.

O pilar desta abordagem é o Teorema de Bayes, que fornece um mecanismo formal para atualizar o conhecimento sobre os parâmetros à medida que novas evidências (dados, $\mathscr{D}$) se

tornam disponíveis. O teorema combina a distribuição a priori, $\pi(\boldsymbol{\vartheta})$, que codifica o conhecimento sobre $\boldsymbol{\vartheta}$ antes da observação dos dados, com a informação contida nos dados através da função de verossimilhança, $\pi(\mathscr{D} \mid \boldsymbol{\vartheta})$, que mede quão prováveis são os dados observados para diferentes valores dos parâmetros. O resultado dessa combinação é a distribuição a posteriori, $\pi(\boldsymbol{\vartheta} \mid \mathscr{D})$, expressa por

$$\pi(\boldsymbol{\vartheta} \mid \mathscr{D}) = \frac{\pi(\mathscr{D} \mid \boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta})}{\pi(\mathscr{D})}, \tag{2.5}$$

em que o denominador, $\pi(\mathscr{D}) = \int \pi(\mathscr{D} \mid \boldsymbol{\vartheta})\,\pi(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta}$, é a evidência ou verossimilhança marginal dos dados. Essa quantidade atua como uma constante de normalização, garantindo que a posteriori integre (ou some, no caso discreto) a 1. A distribuição $\pi(\boldsymbol{\vartheta} \mid \mathscr{D})$ sintetiza toda a informação disponível sobre $\boldsymbol{\vartheta}$ após a observação de $\mathscr{D}$, e serve como base para todas as inferências, como estimativas pontuais, Intervalos de Credibilidade (ICr) e testes de hipóteses.

Apesar da elegância conceitual do Teorema de Bayes, sua aplicação enfrenta desafios computacionais significativos. Em particular, a avaliação direta da distribuição a posteriori frequentemente envolve o cálculo de integrais multidimensionais complexas, que raramente possue solução analítica fechada, especialmente em modelos estatísticos com muitos parâmetros.

Para superar essa dificuldade computacional, recorre-se a métodos de simulação estocástica, notavelmente os algoritmos MCMC. A ideia central do MCMC não é calcular analiticamente a distribuição $\pi(\boldsymbol{\vartheta} \mid \mathscr{D})$, mas sim gerar amostras de uma cadeia de Markov cuja distribuição estacionária coincide com a distribuição a posteriori desejada. Ao simular essa cadeia por um número suficiente de iterações, obtém-se uma amostra representativa da posteriori, a partir da qual as inferências podem ser facilmente calculadas.

Algoritmos MCMC amplamente utilizados incluem o algoritmo de Metropolis-Hastings e suas variantes, bem como o Gibbs Sampling, que é particularmente eficiente quando as distribuições condicionais completas dos parâmetros são tratáveis. Após um período de aquecimento (burn-in), a média amostral pode ser usada como estimativa pontual para um parâmetro, e os quantis amostrais podem ser empregados para construir ICr. Contudo, os algoritmos MCMC tradicionais, como Metropolis-Hastings ou Gibbs Sampling, podem apresentar ineficiências, especialmente quando os parâmetros estão altamente correlacionados ou quando a geometria da posteriori é complexa, resultando em cadeias que exploram lentamente o espaço de parâmetros, exigindo um grande número de iterações para convergir adequadamente (BROOKS *et al.*, 2011).

2.4.1 Stan

Para contornar as limitações dos métodos tradicionais, plataformas modernas, como o Stan, implementam algoritmos MCMC avançados, com destaque para o Monte Carlo Hamiltoniano (HMC)[1] e sua extensão adaptativa, o No-U-Turn Sampler (NUTS) (HOFFMAN; GELMAN, 2014). Stan implementa o NUTS, uma versão adaptativa de última geração do HMC, particular-

[1] Do inglês: *Hamiltonian Monte Carlo*

mente eficaz na exploração de distribuições a posteriori correlacionadas e de alta dimensão. O HMC utiliza gradientes da densidade logarítmica da posteriori para guiar a exploração do espaço de parâmetros, simulando uma dinâmica Hamiltoniana que permite transições eficientes e reduz a autocorrelação entre as amostras. O NUTS aprimora esse processo ao ajustar adaptativamente os parâmetros da simulação, eliminando a necessidade de calibração manual e melhorando a eficiência em geometrias desafiadoras.

O Stan destaca-se por sua capacidade de realizar diferenciação automática, calculando gradientes exatos de forma interna, o que é crucial para modelos com funções de verossimilhança não padrão. Adicionalmente, seus robustos diagnósticos de convergência — como o fator de redução de escala potencial ($\hat{R}$), o tamanho efetivo da amostra (ESS)[2] e as verificações de divergências — fornecem ferramentas essenciais para avaliar a qualidade do ajuste do modelo (CARPENTER *et al.*, 2017). O uso de Stan e NUTS viabiliza, assim, a realização de inferências Bayesianas precisas e escaláveis para uma vasta gama de modelos complexos.

2.5 Síntese

Este capítulo apresentou os principais fundamentos teóricos que sustentam o desenvolvimento metodológico desta tese. Foram discutidos os conceitos essenciais da análise de sobrevivência, incluindo os diferentes tipos de censura e os métodos de estimação. Também foi destacada a importância dos modelos de fragilidade para capturar a heterogeneidade não observada entre os indivíduos, com ênfase na fragilidade discreta como ferramenta para representar subpopulações imunes ao evento de interesse. Por fim, abordou-se a inferência Bayesiana como uma alternativa poderosa e flexível para a estimação em modelos complexos, destacando suas vantagens em cenários com estruturas hierárquicas e dados censurados. Esses conceitos fornecem a base necessária para a formulação e desenvolvimento do modelo proposto nos capítulos seguintes.

[2] Do inglês: *Effective Sample Size*

CAPÍTULO

3

FAMÍLIA DE DISTRIBUIÇÕES HZMPS

Este Capítulo 3, aprofunda o estudo da família de distribuições HZMPS. Esta classe de distribuições foi introduzida e detalhadamente estudada por Conceição (2013), Conceição *et al.* (2017a), Conceição *et al.* (2017b). Essa família caracteriza-se por possuir suporte nos inteiros não-negativos e a incorporação de um parâmetro que controla a dispersão dos dados. A sua generalidade evidencia-se por abranger, como casos particulares, outras distribuições como Hurdle Poisson Zero-Modificada (HZMP)[1], HZMGP, Hurdle Geométrica Zero-Modificada (HZMG)[2], Hurdle Binomial Zero-Modificada (HZMB)[3] e Hurdle Binomial Negativa Zero-Modificada (HZMNB)[4]. Para uma adequada formulação e compreensão da HZMPS, é fundamental o conhecimento prévio das distribuições Série de Potência (PS)[5] e ZMPS, uma vez que a HZMPS é construída a partir de ambas.

3.1 Família de distribuições Série de Potência

As distribuições de probabilidade discretas frequentemente se baseiam em expansões de série de potência, que dão origem a distribuições como Poisson, Poisson generalizada, geométrica, binomial, binomial negativa, logarítmica e outras. Essas distribuições integram a família de distribuições PS, cujo desenvolvimento é creditado a Noack (1950), que a definiu e estudou suas propriedades em detalhe. Posteriormente, diversos autores propuseram reparametrizações nesta família de distribuições em torno da média, como apresentado em Consul e Jain (1973), Conceição (2013). Neste trabalho, utiliza-se a parametrização em torno da média.

Considera-se, assim, uma variável aleatória Y que segue uma distribuição PS, $Y \sim \text{PS}(\mu, \phi)$, com parâmetro de média $\mu > 0$ e parâmetro de dispersão $\phi \geq 0$, cuja função de

[1] Do inglês: *Hurdle Zero-Modified Poisson*
[2] Do inglês: *Hurdle Zero-Modified Geometric*
[3] Do inglês: *Hurdle Zero-Modified Binomial*
[4] Do inglês: *Hurdle Zero-Modified Negative Binomial*
[5] Do inglês: *Power Series*

probabilidade é definida por

$$\pi_{PS}(Y=y;\mu,\phi)=\frac{\alpha(y,\phi)\,(g(\mu,\phi))^{y}}{f(\mu,\phi)},\quad y\in\mathscr{A}_s, \tag{3.1}$$

em que $\alpha(y,\phi)$ é uma função positiva, $g(\mu,\phi)$ e $f(\mu,\phi)$ são funções positivas, finitas e duas vezes diferenciáveis, sendo $f(\mu,\phi)=\sum_{y\in\mathscr{A}_s}\alpha(y,\phi)\,g(\mu,\phi)^{y}$ e $\mathscr{A}_s$ o suporte no subconjunto dos inteiros não negativos $\{0,1,2,\dots\}$.

A suposição $\phi\geq 0$ é necessária para garantir que a variável aleatória Y tenha uma distribuição de probabilidade completamente determinada por sua função de variância (CORDEIRO; ANDRADE; CASTRO, 2009).

A família de distribuições PS definida em (3.1) abrange uma ampla variedade de distribuições conhecidas, incluindo as já mencionadas Poisson, Poisson generalizada, geométrica, binomial e binomial negativa. Na Tabela 3, apresentam-se as funções $\alpha(y,\phi)$, $g(\mu,\phi)$ e $f(\mu,\phi)$ para algumas distribuições pertencentes a esta família.

Tabela 3 – Distribuições pertencentes à família PS.

Distribuição	$\alpha(y,\phi)$	$g(\mu,\phi)$	$f(\mu,\phi)$	$\mathscr{A}_s$
Poisson	$\frac{1}{y!}$	μ	e^{μ}	$\{0,1,\dots\}$
Poisson generalizada	$\frac{(1+\phi y)^{y-1}}{y!}$	$\frac{\mu e^{-\mu\phi(1+\mu\phi)^{-1}}}{1+\mu\phi}$	$e^{\mu(1+\mu\phi)^{-1}}$	$\{0,1,\dots\}$
Geométrica	1	$\frac{\mu}{1+\mu}$	$1+\mu$	$\{0,1,\dots\}$
Binomial	$\binom{m}{y}$	$\frac{\mu}{m-\mu}$	$\left(\frac{m}{m-\mu}\right)^{m}$	$\{0,1,\dots,m\}$
Binomial negativa	$\frac{\Gamma(\phi+y)}{y!\Gamma(\phi)}$	$\frac{\mu}{\mu+\phi}$	$\left(\frac{\phi}{\mu+\phi}\right)^{-\phi}$	$\{0,1,\dots\}$

Fonte: Conceição (2013).

Com base na definição da distribuição PS, a média e a variância são definidas por

$$\mathbb{E}(Y)=\mu_{PS}=\frac{f'(\mu,\phi)\,g(\mu,\phi)}{f(\mu,\phi)\,g'(\mu,\phi)},$$
$$\mathrm{Var}(Y)=\sigma^2_{PS}=\frac{g(\mu,\phi)}{g'(\mu,\phi)},$$

em que $f'(\mu,\phi)$ e $g'(\mu,\phi)$ representam as primeiras derivadas das funções em relação a μ (GUPTA, 1974). A Tabela 4 apresenta a média e a variância para algumas distribuições desta família.

Tabela 4 – Média e variância de algumas distribuições da família PS.

Distribuição	Média	Variância
Poisson	μ	μ
Poisson generalizada	μ	$\mu(1+\mu\phi)^2$
Geométrica	μ	$\mu(1+\mu)$
Binomial	μ	$\mu\left(1-\frac{\mu}{m}\right)$
Binomial negativa	μ	$\mu\left(1+\frac{\mu}{\phi}\right)$

Fonte: Elaborada pelo autor.

A análise de dados de contagem ocupa um lugar de destaque na estatística aplicada. Embora a distribuição de Poisson seja tradicionalmente empregada, dados de contagem frequentemente exibem um excesso de zeros, uma característica que modelos mais flexíveis precisam acomodar. A necessidade de lidar com essa variação na probabilidade no ponto zero motiva o desenvolvimento de distribuições modificadas, capazes de capturar adequadamente tais padrões (JOHNSON; KOTZ; BALAKRISHNAN, 1995). A família PS, por sua vez, pode ser generalizada para incluir parâmetros adicionais, aumentando sua flexibilidade (JOHNSON; KEMP; KOTZ, 2005). A próxima seção detalha uma dessas generalizações, a família de distribuições ZMPS.

3.2 Família de distribuições Série de Potência Zero-Modificada

Com base na equação (3.1), Conceição (2013) propôs uma modificação na probabilidade de ocorrência do zero na distribuição PS, dando origem à família de distribuições ZMPS. Esta família de distribuições é especialmente recomendada para modelar conjuntos de dados com diferentes padrões de ocorrência de zeros, seja em excesso (zero-inflação), escassez (zero-deflação) ou ausência (zero-truncado).

Para formalizar a definição, considera-se uma variável aleatória Y que segue uma distribuição ZMPS, $Y \sim \text{ZMPS}\ (\mu,\phi,\rho)$, com $\mu > 0$ como parâmetro de média, $\phi \geq 0$ como parâmetro de dispersão e ρ como o parâmetro encarregado da modificação da probabilidade de zero. Sua função de probabilidade é definida como

$$\pi_{ZMPS}(Y=y;\mu,\phi,\rho) = (1-\rho)\,\mathbf{I}(y) + \rho\,\pi_{PS}(Y=y;\mu,\phi), \quad y \in \mathscr{A}_0, \tag{3.2}$$

em que $\mathscr{A}_0 = \{0,1,\ldots\}$ é o suporte formado pelo subconjunto dos inteiros não negativos,

$\pi_{PS}(Y=y;\mu,\phi)$ é a distribuição PS definida em (3.1) e $\mathbf{I}(y)$ é uma função indicadora tal que

$$\mathbf{I}(y)=\begin{cases}1, \text{ se } y=0,\\ 0, \text{ se } y>0.\end{cases}$$

O parâmetro ρ obedece à seguinte restrição

$$0\leq\rho\leq\frac{1}{1-\pi_{PS}(Y=0;\mu,\phi)}, \tag{3.3}$$

em que $\pi_{PS}(Y=0;\mu,\phi)$ é a probabilidade da distribuição PS em zero.

Para obter as diferentes distribuições desta ampla família, basta substituir a distribuição PS correspondente na expressão (3.2), o que gera a função de probabilidade específica para cada caso.

Outra característica importante da família de distribuições ZMPS é que a diferença na probabilidade no ponto zero entre a distribuição ZMPS e a distribuição PS é dada por

$$\begin{aligned}\pi_{ZMPS}(Y=0;\mu,\phi,\rho)-\pi_{PS}(Y=0;\mu,\phi) &= 1-\rho+\rho\,\pi_{PS}(Y=0;\mu,\phi)-\pi_{PS}(Y=0;\mu,\phi)\\ &= 1-\rho-\pi_{PS}(Y=0;\mu,\phi)(1-\rho)\\ &= (1-\rho)\left[1-\pi_{PS}(Y=0;\mu,\phi)\right]. \end{aligned}\tag{3.4}$$

Com base na expressão (3.4), diferentes valores de ρ definem casos particulares da distribuição ZMPS, permitindo identificar as seguintes situações

1) Para $\rho=0$:
   Resulta em $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)=1$. Portanto, $\pi_{ZMPS}(Y=y;\mu,\phi,\rho)$ é uma distribuição degenerada com toda a massa de probabilidade em zero.

2) Para $0<\rho<1$:
   Obtém-se $(1-\rho)(1-\pi_{PS}(Y=0;\mu,\phi))>0$, o que implica $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)>\pi_{PS}(Y=0;\mu,\phi)$. Neste caso, $\pi_{ZMPS}(Y=y;\mu,\phi,\rho)$ corresponde à distribuição ZIPS.

3) Para $\rho=1$:
   Tem-se que $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)-\pi_{PS}(Y=0;\mu,\phi)=0$. Assim, $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)=\pi_{PS}(Y=0;\mu,\phi)$, e $\pi_{ZMPS}(Y=y;\mu,\phi,\rho)$ reduz-se à distribuição PS.

4) Para $1<\rho<\dfrac{1}{1-\pi_{PS}(0;\mu,\phi)}$:
   A diferença $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)-\pi_{PS}(Y=0;\mu,\phi)$ é negativa. Portanto, $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)<\pi_{PS}(Y=0;\mu,\phi)$, e $\pi_{ZMPS}(Y=y;\mu,\phi,\rho)$ corresponde à distribuição Série de Potência Zero-Deflacionada (ZDPS)[6].

[6] Do inglês: *Zero-Deflated Power Series*

5) Para $\rho = \dfrac{1}{1-\pi_{PS}(Y=0;\mu,\phi)}$:
Resulta em $\pi_{ZMPS}(Y=0;\mu,\phi,\rho)=0$. Então, $\pi_{ZMPS}(Y=y;\mu,\phi,\rho)$ é a distribuição Série de Potência Zero-Truncada (ZTPS)[7].

A análise dos casos particulares evidencia que o parâmetro ρ pode assumir valores superiores a 1, o que constitui a principal diferença entre a família ZMPS e as distribuições de mistura tradicionais, pois enquanto na distribuição ZMPS o parâmetro ρ pode ultrapassar 1, na distribuição mistura padrão isso não é possível. Essa flexibilidade é um diferencial importante, pois, dependendo do valor de ρ, as distribuições ZIPS, ZDPS, ZTPS e PS configuram-se como casos particulares da ZMPS, influenciando diretamente a probabilidade de zero.

A família ZMPS oferece medidas de variabilidade que desempenham um papel importante na sua caracterização, entre as quais se destaca o coeficiente de dispersão. Segundo Kurnia, Sadik *et al.* (2021), Kamalja e Wagh (2018), uma das causas da superdispersão é o excesso de zeros nos dados, um problema para o qual se recomenda o uso de distribuições mais flexíveis.

Nesta reparametrização, a família de distribuições ZMPS é composta por diferentes distribuições, detalhadas na Tabela 5. Adicionalmente, esta família possui um conjunto de propriedades importantes, sumarizadas na Tabela 6, cujas demonstrações e informações adicionais podem ser encontradas em Consul e Famoye (2006), Conceição (2013), Conceição *et al.* (2017). Complementarmente, a Tabela 7 apresenta a média, a variância e a função geradora de probabilidade (f.g.p.) para estas distribuições.

[7] Do inglês: *Zero-Truncanted Power Series*

Tabela 5 – Função de probabilidade e restrição de ρ para distribuições da família ZMPS.

Distribuição	Função de probabilidade	Restrição de ρ
ZMP	$(1-\rho)\mathbf{I}(y)+\rho\dfrac{e^{-\mu}\mu^{y}}{y!}$	$0\leq\rho\leq\dfrac{1}{1-e^{-\mu}}$
ZMGP	$(1-\rho)\mathbf{I}(y)+\rho\dfrac{\dfrac{(1+\phi y)^{y-1}}{y!}\left[\dfrac{\mu e^{-\mu\phi(1+\mu\phi)^{-1}}}{1+\mu\phi}\right]^{y}}{e^{\mu(1+\mu\phi)^{-1}}}$	$0\leq\rho\leq\dfrac{1}{1-e^{-\frac{\mu}{1+\mu\phi}}}$
ZMG	$(1-\rho)\mathbf{I}(y)+\rho\dfrac{\left(\dfrac{\mu}{1+\mu}\right)^{y}}{1+\mu}$	$0\leq\rho\leq\dfrac{1+\mu}{\mu}$
ZMB	$(1-\rho)\mathbf{I}(y)+\rho\dfrac{\binom{m}{y}\left(\dfrac{\mu}{m-\mu}\right)^{y}}{\left(\dfrac{m}{m-\mu}\right)^{m}},\quad m\geq 1$	$0\leq\rho\leq\dfrac{m^{m}}{m^{m}-(m-\mu)^{m}}$
ZMNB	$(1-\rho)\mathbf{I}(y)+\rho\dfrac{\dfrac{\Gamma(\phi+y)}{y!\Gamma(\phi)}\left(\dfrac{\mu}{\mu+\phi}\right)^{y}}{\left(\dfrac{\phi}{\mu+\phi}\right)^{-\phi}}$	$0\leq\rho\leq\dfrac{(\mu+\phi)^{\phi}}{(\mu+\phi)^{\phi}-\phi^{\phi}}$

Fonte: Elaborada pelo autor.

Tabela 6 – Principais propriedades da família de distribuições ZMPS.

Notação	Característica	Descrição
μ_{ZMPS}	Média	$\rho\mu_{PS}$
$\sigma^{2}{}_{ZMPS}$	Variância	$\rho\sigma^{2}_{PS}+\rho(1-\rho)\mu^{2}_{PS}$
$\mathbb{F}_{ZMPS}(k)$	Função de distribuição	$1-\rho+\rho\mathbb{F}_{PS}(k)$
$\mathbb{G}_{ZMPS}(t)$	Função geradora de probabilidade	$1-\rho+\rho\mathbb{G}_{PS}(t)$, $0\leq t\leq 1$
$\mathbb{M}_{ZMPS}(t)$	Função geradora de momentos	$1-\rho+\rho\mathbb{M}_{PS}(t)$
$\Psi_{ZMPS}(t)$	Função característica	$1-\rho+\rho\Psi_{PS}(t)$

Fonte: Elaborada pelo autor.

Tabela 7 – Média, variância e f.g.p. para distribuições da família ZMPS.

	Média	Variância	f.g.p.
ZMP	$\rho\mu$	$\rho\mu+\rho(1-\rho)\mu^2$	$1-\rho\left[1-e^{-\mu(1-t)}\right]$
ZMGP	$\rho\mu$	$\rho\mu(1+\mu\phi)^2+\rho(1-\rho)\mu^2$	$1-\rho\left[1-e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}te^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}\right]$
ZMG	$\rho\mu$	$\rho\mu\left(1+\mu\right)+\rho\left(1-\rho\right)\mu^2$	$1-\rho\left[1-\frac{1}{1+\mu-\mu t}\right]$
ZMB	$\rho\mu$	$\rho\mu\left(1-\frac{\mu}{m}\right)+\rho\left(1-\rho\right)\mu^2$	$1-\rho\left[1-\left(\frac{m-\mu+\mu t}{m}\right)^m\right]$
ZMNB	$\rho\mu$	$\rho\mu\left(1+\frac{\mu}{\phi}\right)+\rho(1-\rho)\mu^2$	$1-\rho\left[1-\left(\frac{\phi}{\mu-\mu t+\phi}\right)^\phi\right]$

Fonte: Elaborada pelo autor.

3.3 Versão Hurdle das distribuições Série de Potência Zero-Modificada

Conceição *et al.* (2017a) propuseram uma reparametrização em ρ que se mostra vantajosa em certas aplicações, ao oferecer maior flexibilidade no ajuste da distribuição.

A expressão original da família ZMPS em (3.2) pode ser reescrita como

$$\begin{aligned}\pi_{ZMPS}(Y=y;\mu,\phi,\rho) &= (1-\rho)\,\mathbf{I}(y)+\rho\,\pi_{PS}(Y=y;\mu,\phi)\\ &= \underbrace{1-\rho+\rho\,\pi_{PS}(Y=0;\mu,\phi)}_{y=0}+\underbrace{\rho\,\pi_{PS}(Y=y;\mu,\phi)}_{y>0}\\ &= \pi_{ZMPS}(Y=0;\mu,\phi,\rho)\,\mathbf{I}(y)+\pi_{ZMPS}(Y=y;\mu,\phi,\rho)\,(1-\mathbf{I}(y))\\ &= \{1-\rho+\rho\,\pi_{PS}(Y=0;\mu,\phi)\}\,\mathbf{I}(y)\\ &+\{\rho\,\pi_{PS}(Y=y;\mu,\phi)\}\,(1-\mathbf{I}(y)),\quad y\in\mathscr{A}_0,\end{aligned}\tag{3.5}$$

nota-se que $\mathbf{I}_{\mathscr{A}_1}(y)=1-\mathbf{I}(y)$. A restrição sobre ρ dada em (3.3) pode ser reescrita em termos do novo parâmetro, como se segue

$$\begin{aligned}&0\leq\rho\,(1-\pi_{PS}(Y=0;\mu,\phi))\leq 1\\ &0\leq\omega\leq 1.\end{aligned}\tag{3.6}$$

A estrutura em (3.5), combinada com a restrição em (3.6), permite definir a distribuição HZMPS. A função de probabilidade de uma variável aleatória Y que segue a HZMPS, $Y\sim$

HZMPS (μ, ϕ, ω), é dada por

$$\pi_{HZMPS}(Y = y; \mu, \phi, \omega) = (1-\omega)\,\mathbf{I}(y) + \omega\pi_{ZTPS}(Y = y; \mu, \phi), \quad y \in \mathscr{A}_0, \tag{3.7}$$

em que $\pi_{ZTPS}(Y = y; \mu, \phi) = \left\{ \dfrac{\pi_{PS}(Y = y; \mu, \phi)}{1 - \pi_{PS}(Y = 0; \mu, \phi)} \right\} \mathbf{I}_{\mathscr{A}_1}(y)$ é a distribuição ZTPS com suporte $\mathscr{A}_1 = \{1, 2, 3, \ldots\}$. A função $\mathbf{I}_{\mathscr{A}_1}(y)$ é uma função indicadora para o conjunto $\mathscr{A}_1$ tal que $\mathbf{I}_{\mathscr{A}_1}(y) = 1$ se $y \in \mathscr{A}_1$ e 0 se $y \notin \mathscr{A}_1$. Note-se que $\mathscr{A}_1 \subset \mathscr{A}_0$.

A média, a variância e a f.g.p. da distribuição HZMPS são dadas, respectivamente, por

$$\begin{aligned}
\mu_{HZMPS} &= \frac{\omega\mu_{PS}}{1 - \pi_{PS}(Y = 0; \mu, \phi)}, \\
\sigma^2_{HZMPS} &= \frac{\omega\left(\sigma^2_{PS} + \mu^2_{PS}\right)}{1 - \pi_{PS}(Y = 0; \mu, \phi)} - (\mu_{HZMPS})^2, \\
\mathbb{G}_{HZMPS}(t) &= 1 - \frac{\omega}{1 - \pi_{PS}(Y = 0; \mu, \phi)}\left[1 - \mathbb{G}_{PS}(t)\right], \quad 0 \leq t \leq 1.
\end{aligned}$$

Conforme discutido por Conceição *et al.* (2017a), a HZMPS contém a distribuição ZTPS como componente, diferindo da estrutura de mistura usualmente empregada em modelos inflacionados de zeros. A HZMPS pode ser interpretada como resultante de um processo de dois estágios: um primeiro estágio gera apenas zeros $(Y = 0)$ com probabilidade $(1 - \omega)$, e um segundo gera valores estritamente positivos $(Y > 0)$ com probabilidade $\omega\pi_{ZTPS}(Y = y; \mu, \phi)$.

Uma vantagem desta reparametrização é a ortogonalidade entre os parâmetros ω e μ, o que permite a estimação de ω independentemente de μ. Por outro lado, a parametrização original expressa em (3.2) com o parâmetro ρ facilita a identificação direta do tipo de modificação nos zeros (inflação, deflação ou truncamento), como apontado por Conceição *et al.* (2017b). A reparametrização com ω, no entanto, preserva uma interpretação clara, pois ω também indica a natureza da modificação no zero.

Uma característica importante da família HZMPS reside na sua capacidade de modificar a frequência de zeros por meio do parâmetro ω. Essa flexibilidade pode ser analisada quantitativamente ao se examinar a diferença entre as probabilidades no ponto zero da distribuição HZMPS e da distribuição PS correspondente, como se detalha a seguir

$$\pi_{HZMPS}(Y = 0; \mu, \phi, \omega) - \pi_{PS}(Y = 0; \mu, \phi) = 1 - \omega - \pi_{PS}(Y = 0; \mu, \phi). \tag{3.8}$$

Analisando-se a diferença expressa em (3.8), identificam-se as seguintes situações:

1) Para $\omega = 0$:
   Resulta em $\pi_{HZMPS}(Y = 0; \mu, \phi, \omega) = 1$. Portanto, $\pi_{HZMPS}(Y = y; \mu, \phi, \omega)$ é uma distribuição degenerada com toda a massa de probabilidade em zero.

2) Para $0 < \omega < 1 - \pi_{PS}(Y = 0; \mu, \phi)$:
   A diferença $1 - \omega - \pi_{PS}(Y = 0; \mu, \phi)$ é positiva, o que implica $\pi_{HZMPS}(Y = 0; \mu, \phi, \omega) > \pi_{PS}(Y = 0; \mu, \phi)$. Neste caso, $\pi_{HZMPS}(Y = y; \mu, \phi, \omega)$ corresponde à distribuição ZIPS.

3) Para $\omega = 1 - \pi_{PS}(Y = 0; \mu, \phi)$:
A diferença é nula, resultando em $\pi_{HZMPS}(Y = 0; \mu, \phi, \omega) = \pi_{PS}(Y = 0; \mu, \phi)$. Consequentemente, $\pi_{HZMPS}(Y = y; \mu, \phi, \omega)$ reduz-se à distribuição PS.

4) Para $1 - \pi_{PS}(0; \mu, \phi) < \omega < 1$:
A diferença é negativa, o que implica $\pi_{HZMPS}(Y = 0; \mu, \phi, \omega) < \pi_{PS}(Y = 0; \mu, \phi)$. Neste caso, $\pi_{HZMPS}(Y = y; \mu, \phi, \omega)$ corresponde à distribuição ZDPS.

5) Para $\omega = 1$:
Resulta em $\pi_{HZMPS}(Y = 0; \mu, \phi, \omega) = 0$. Então, $\pi_{HZMPS}(Y = y; \mu, \phi, \omega)$ é a distribuição ZTPS.

3.4 Distribuições pertencentes à família HZMPS

Nesta seção, apresentam-se as distribuições que constituem a família HZMPS. A construção destas distribuições parte das funções de probabilidade da família ZMPS, detalhadas na Tabela 5. A seguir, exploram-se as principais características e os casos particulares de cada distribuição HZMPS.

3.4.1 Distribuição Hurdle Poisson Zero-Modificada

Seja Y uma variável aleatória que segue a distribuição HZMP, denotada por $Y \sim \text{HZMP}(\mu, \omega)$. Sua função de probabilidade é definida como

$$\pi_{HZMP}(Y = y; \mu, \omega) = (1 - \omega)\,\mathbf{I}(y) + \omega \frac{e^{-\mu}\mu^{y}}{y!\,(1 - e^{-\mu})}, \quad y = 0, 1, \ldots. \tag{3.9}$$

A diferença entre as probabilidades no ponto zero $(Y = 0)$ da distribuição HZMP e da distribuição Poisson correspondente é dada por

$$\pi_{HZMP}(Y = 0; \mu, \omega) - \pi_{P}(Y = 0; \mu) = 1 - \omega - \left(e^{-\mu}\right). \tag{3.10}$$

A expressão resultante, apresentada em (3.10), indica como o parâmetro ω modifica a probabilidade de ocorrência do zero. Com base no valor desta diferença, distinguem-se as seguintes situações:

1) Para $0 < \omega < 1 - e^{-\mu}$:
A diferença em (3.10) é positiva, implicando que $\pi_{HZMP}(Y = 0; \mu, \omega) > \pi_{P}(Y = 0; \mu)$. Portanto, $\pi_{HZMP}(Y = y; \mu, \omega)$ corresponde à distribuição Poisson Zero-Inflacionada (ZIP)[8].

2) Para $\omega = 1 - e^{-\mu}$:
A diferença em (3.10) é nula, resultando em $\pi_{HZMP}(Y = 0; \mu, \omega) = \pi_{P}(Y = 0; \mu)$. Assim, $\pi_{HZMP}(Y = y; \mu, \omega)$ reduz-se à distribuição de Poisson.

[8] Do inglês: *Zero-Inflated Poisson*

3) Para $1-e^{-\mu}<\omega<1$:
A diferença em (3.10) é negativa, o que implica $\pi_{HZMP}(Y=0;\mu,\omega)<\pi_P(Y=0;\mu)$. Desta forma, $\pi_{HZMP}(Y=y;\mu,\omega)$ corresponde à distribuição Poisson Zero-Deflacionada (ZDP)[9].

4) Para $\omega=1$:
Em (3.10), obtém-se $\pi_{HZMP}(Y=0;\mu,\omega)=0$. Logo, $\pi_{HZMP}(Y=y;\mu,\omega)$ corresponde à distribuição Poisson Zero-Truncada (ZTP)[10].

Com base na definição da distribuição HZMP, a média, a variância e a f.g.p. são dadas, respectivamente, por

$$\begin{aligned}\mu_{HZMP}&=\frac{\omega\mu}{1-e^{-\mu}},\\ \sigma^2_{HZMP}&=\frac{\omega\left(\mu+\mu^2\right)}{1-e^{-\mu}}-\left(\frac{\omega\mu}{1-e^{-\mu}}\right)^2, \quad \text{e}\\ \mathbb{G}_{HZMP}(t)&=1-\frac{\omega}{1-e^{-\mu}}\left[1-e^{-\mu(1-t)}\right], \quad 0\leq t\leq 1.\end{aligned}$$

3.4.2 Distribuição Hurdle Poisson Generalizada Zero-Modificada

Seja Y uma variável aleatória que segue a distribuição HZMGP, denotada por $Y\sim$ HZMGP(μ,ϕ,ω). Sua função de probabilidade é definida como

$$\pi_{HZMGP}(Y=y;\mu,\phi,\omega)=(1-\omega)\,\mathbf{I}(y)+\omega\frac{\dfrac{(1+\phi y)^{y-1}}{y!}\left[\dfrac{\mu e^{-\mu\phi(1+\mu\phi)^{-1}}}{1+\mu\phi}\right]^y}{\dfrac{e^{\mu(1+\mu\phi)^{-1}}}{1-e^{-\frac{\mu}{1+\mu\phi}}}}, \quad y=0,1,\ldots \tag{3.11}$$

A diferença entre as probabilidades no ponto zero $(Y=0)$ da distribuição HZMGP e da distribuição Poisson Generalizada (GP)[11] é dada por

$$\pi_{HZMGP}(Y=0;\mu,\phi,\omega)-\pi_{GP}(Y=0;\mu,\phi)=1-\omega-e^{-\frac{\mu}{1+\mu\phi}}. \tag{3.12}$$

A expressão resultante, apresentada em (3.12), indica como o parâmetro ω modifica a probabilidade de ocorrência do zero. Com base no valor desta diferença, distinguem-se as seguintes situações:

1) Para $0<\omega<1-e^{-\frac{\mu}{1+\mu\phi}}$:
A diferença em (3.12) é positiva, implicando que $\pi_{HZMGP}(Y=0;\mu,\phi,\omega)>\pi_{GP}(Y=0;\mu,\phi)$.

[9] Do inglês: *Zero-Deflated Poisson*
[10] Do inglês: *Zero-Truncated Poisson*
[11] Do inglês: *Generalized Poisson*

Portanto, $\pi_{HZMGP}(Y=y;\mu,\phi,\omega)$ corresponde à distribuição Poisson Generalizada Zero-Inflacionada (ZIGP)[12].

2) Para $\omega = 1-e^{-\frac{\mu}{1+\mu\phi}}$:
A diferença em (3.12) é nula, resultando em $\pi_{HZMGP}(Y=0;\mu,\phi,\omega) = \pi_{GP}(Y=0;\mu,\phi)$. Assim, $\pi_{HZMGP}(Y=y;\mu,\phi,\omega)$ reduz-se à distribuição GP.

3) Para $1-e^{-\frac{\mu}{1+\mu\phi}} < \omega < 1$:
A diferença em (3.12) é negativa, o que implica $\pi_{HZMGP}(Y=0;\mu,\phi,\omega) < \pi_{GP}(Y=0;\mu,\phi)$. Desta forma, $\pi_{HZMGP}(Y=y;\mu,\phi,\omega)$ corresponde à distribuição Poisson Generalizada Zero-Deflacionada (ZDGP)[13].

4) Para $\omega = 1$:
Em (3.12), obtém-se $\pi_{HZMGP}(Y=0;\mu,\phi,\omega) = 0$. Logo, $\pi_{HZMGP}(Y=y;\mu,\phi,\omega)$ corresponde à distribuição Poisson Generalizada Zero-Truncada (ZTGP)[14].

Com base na definição da distribuição HZMGP, a média, a variância e a f.g.p. são dadas, respectivamente, por

$$\mu_{HZMGP} = \frac{\omega\mu}{1-e^{-\frac{\mu}{1+\mu\phi}}},$$

$$\sigma^2_{HZMGP} = \frac{\omega\left[\mu(1+\mu\phi)^2+\mu^2\right]}{1-e^{-\frac{\mu}{1+\mu\phi}}} - \left(\frac{\omega\mu}{1-e^{-\frac{\mu}{1+\mu\phi}}}\right)^2, \quad \text{e}$$

$$\mathbb{G}_{HZMGP}(t) = 1 - \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}}\left[1-e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}te^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}\right], \quad 0\leq t\leq 1.$$

A f.g.p. da distribuição GP adotada neste trabalho é a versão desenvolvida por Ambagaspitiya e Balakrishnan (1994), a qual é expressa em termos da função $\mathbb{W}$ de Lambert.

3.4.3 Distribuição Hurdle Geométrica Zero-Modificada

Seja Y uma variável aleatória que segue a distribuição HZMG, denotada por $Y \sim \text{HZMG}(\mu,\omega)$. Sua função de probabilidade é definida como

$$\pi_{HZMG}(Y=y;\mu,\omega) = (1-\omega)\mathbf{I}(y) + \omega\frac{\mu^{(y-1)}}{(1+\mu)^y}, \quad y=0,1,\ldots. \tag{3.13}$$

A diferença entre as probabilidades no ponto zero $(Y=0)$ da distribuição HZMG e da distribuição geométrica correspondente é dada por

$$\pi_{HZMG}(Y=0;\mu,\omega) - \pi_G(Y=0;\mu) = 1-\omega-\left(\frac{1}{1+\mu}\right). \tag{3.14}$$

[12] Do inglês: *Zero-Inflated Generalized Poisson*
[13] Do inglês: *Zero-Deflated Generalized Poisson*
[14] Do inglês: *Zero-Truncated Generalized Poisson*

A expressão resultante, apresentada em (3.14), indica como o parâmetro ω modifica a probabilidade de ocorrência do zero. Com base no valor desta diferença, distinguem-se as seguintes situações:

1) Para $0 < \omega < 1 - \dfrac{1}{1+\mu}$:
A diferença em (3.14) é positiva, implicando que $\pi_{HZMG}(Y=0;\mu,\omega) > \pi_G(Y=0;\mu)$. Portanto, $\pi_{HZMG}(Y=y;\mu,\omega)$ corresponde à distribuição Geométrica Zero-Inflacionada (ZIG)[15].

2) Para $\omega = 1 - \dfrac{1}{1+\mu}$:
A diferença em (3.14) é nula, resultando em $\pi_{HZMG}(Y=0;\mu,\omega) = \pi_G(Y=0;\mu)$. Assim, $\pi_{HZMG}(Y=y;\mu,\omega)$ reduz-se à distribuição geométrica.

3) Para $1 - \dfrac{1}{1+\mu} < \omega < 1$:
A diferença em (3.14) é negativa, o que implica $\pi_{HZMG}(Y=0;\mu,\omega) < \pi_G(Y=0;\mu)$. Desta forma, $\pi_{HZMG}(Y=y;\mu,\omega)$ corresponde à distribuição Geométrica Zero-Deflacionada (ZDG)[16].

4) Para $\omega = 1$:
Em (3.14), obtém-se $\pi_{HZMG}(Y=0;\mu,\omega) = 0$. Logo, $\pi_{HZMG}(Y=y;\mu,\omega)$ corresponde à distribuição Geométrica Zero-Truncada (ZTG)[17].

Com base na definição da distribuição HZMG, a média, a variância e a f.g.p. são dadas, respectivamente, por

$$\begin{aligned}
\mu_{HZMG} &= \omega(1+\mu),\\
\sigma^2_{HZMG} &= \frac{\omega\left[\mu(1+\mu)+\mu^2\right]}{\dfrac{\mu}{1+\mu}} - \left[\omega(1+\mu)\right]^2, \quad \text{e}\\
\mathbb{G}_{HZMG}(t) &= 1 - \frac{\omega(1+\mu)}{\mu}\left[1 - \frac{1}{1+\mu-\mu t}\right], \quad 0 \le t \le 1.
\end{aligned}$$

3.4.4 Distribuição Hurdle Binomial Zero-Modificada

Seja Y uma variável aleatória que segue a distribuição HZMB, denotada por $Y \sim \text{HZMB}(\mu,\omega)$. Sua função de probabilidade é definida como

$$\pi_{HZMB}(Y=y;\mu,\omega) = (1-\omega)\,\mathbf{I}(y) + \omega\frac{\binom{m}{y}\mu^y(m-\mu)^{m-y}}{m^m-(m-\mu)^m}, \quad y = 0,1,\ldots.m. \tag{3.15}$$

[15] Do inglês: *Zero-Inflated Geometric*
[16] Do inglês: *Zero-Deflated Geometric*
[17] Do inglês: *Zero-Truncated Geometric*

A diferença entre as probabilidades no ponto zero $(Y = 0)$ da distribuição HZMB e da distribuição binomial correspondente é dada por

$$\pi_{HZMB}(Y = 0; \mu, \omega) - \pi_B(Y = 0; \mu) = 1 - \omega - \left(\frac{m-\mu}{m}\right)^m. \tag{3.16}$$

A expressão resultante, apresentada em (3.16), indica como o parâmetro ω modifica a probabilidade de ocorrência do zero. Com base no valor desta diferença, distinguem-se as seguintes situações:

1) Para $0 < \omega < 1 - \left(\frac{m-\mu}{m}\right)^m$:
A diferença em (3.16) é positiva, implicando que $\pi_{HZMB}(Y = 0; \mu, \omega) > \pi_B(Y = 0; \mu)$. Portanto, $\pi_{HZMB}(Y = y; \mu, \omega)$ corresponde à distribuição Binomial Zero-Inflacionada (ZIB)[18].

2) Para $\omega = 1 - \left(\frac{m-\mu}{m}\right)^m$:
A diferença em (3.16) é nula, resultando em $\pi_{HZMB}(Y = 0; \mu, \omega) = \pi_B(Y = 0; \mu)$. Assim, $\pi_{HZMB}(Y = y; \mu, \omega)$ reduz-se à distribuição binomial.

3) Para $1 - \left(\frac{m-\mu}{m}\right)^m < \omega < 1$:
A diferença em (3.16) é negativa, o que implica $\pi_{HZMB}(Y = 0; \mu, \omega) < \pi_B(Y = 0; \mu)$. Desta forma, $\pi_{HZMB}(Y = y; \mu, \omega)$ corresponde à distribuição Binomial Zero-Deflacionada (ZDB)[19].

4) Para $\omega = 1$:
Em (3.16), obtém-se $\pi_{HZMB}(Y = 0; \mu, \omega) = 0$. Logo, $\pi_{HZMB}(Y = y; \mu, \omega)$ corresponde à distribuição Binomial Zero-Truncada (ZTB)[20].

Com base na definição da distribuição HZMB, a média, a variância e a f.g.p. são dadas, respectivamente, por

$$\mu_{HZMB} = \frac{\omega\mu}{1 - \left(\frac{m-\mu}{m}\right)^m},$$

$$\sigma^2_{HZMB} = \frac{\omega\left[\mu\left(1 + \frac{\mu}{m}\right) + \mu^2\right]}{1 - \left(\frac{m-\mu}{m}\right)^m} - \left[\frac{\omega\mu}{1 - \left(\frac{m-\mu}{m}\right)^m}\right]^2, \quad \text{e}$$

$$\mathbb{G}_{HZMB}(t) = 1 - \frac{\omega}{1 - \left(\frac{m-\mu}{m}\right)^m}\left[1 - \left(\frac{m - \mu + \mu t}{m}\right)^m\right], \quad 0 \leq t \leq 1.$$

[18] Do inglês: *Zero-Inflated Binomial*
[19] Do inglês: *Zero-Deflated Binomial*
[20] Do inglês: *Zero-Truncated Binomial*

3.4.5 Distribuição Hurdle Binomial Negativa Zero-Modificada

Seja Y uma variável aleatória que segue a distribuição HZMNB, denotada por $Y \sim \text{HZMNB}(\mu, \phi, \omega)$. Sua função de probabilidade é definida como

$$\pi_{HZMNB}(Y = y; \mu, \phi, \omega) = (1-\omega)\mathbf{I}(y) + \omega \frac{\dfrac{\Gamma(\phi+y)}{y!\Gamma(\phi)}\left(\dfrac{\mu}{\mu+\phi}\right)^{y}}{1-\left(\dfrac{\phi}{\mu+\phi}\right)^{\phi}}, \quad y = 0, 1, \ldots. \tag{3.17}$$

A diferença entre as probabilidades no ponto zero $(Y = 0)$ da distribuição HZMNB e da Binomial Negativa (NB)[21] correspondente é dada por

$$\pi_{HZMNB}(Y = 0; \mu, \phi, \omega) - \pi_{NB}(Y = 0; \mu, \phi) = 1 - \omega - \left(\frac{\phi}{\mu+\phi}\right)^{\phi}. \tag{3.18}$$

A expressão resultante, apresentada em (3.18), indica como o parâmetro ω modifica a probabilidade de ocorrência do zero. Com base no valor desta diferença, distinguem-se as seguintes situações:

1) Para $0 < \omega < 1 - \left(\frac{\phi}{\mu+\phi}\right)^{\phi}$:
A diferença em (3.18) é positiva, implicando que $\pi_{HZMNB}(Y = 0; \mu, \phi, \omega) > \pi_{NB}(Y = 0; \mu, \phi)$. Portanto, $\pi_{HZMNB}(Y = y; \mu, \phi, \omega)$ corresponde à distribuição Binomial Negativa Zero-Inflacionada (ZINB)[22].

2) Para $\omega = 1 - \left(\frac{\phi}{\mu+\phi}\right)^{\phi}$:
A diferença em (3.18) é nula, resultando em $\pi_{HZMNB}(Y = 0; \mu, \phi, \omega) = \pi_{NB}(Y = 0; \mu, \phi)$. Assim, $\pi_{HZMNB}(Y = y; \mu, \phi, \omega)$ reduz-se à distribuição NB.

3) Para $1 - \left(\frac{\phi}{\mu+\phi}\right)^{\phi} < \omega < 1$:
A diferença em (3.18) é negativa, o que implica $\pi_{HZMNB}(Y = 0; \mu, \phi, \omega) < \pi_{NB}(Y = 0; \mu, \phi)$. Desta forma, $\pi_{HZMNB}(Y = y; \mu, \phi, \omega)$ corresponde à distribuição Binomial Negativa Zero-Deflacionada (ZDNB)[23].

4) Para $\omega = 1$:
Em (3.18), obtém-se $\pi_{HZMNB}(Y = 0; \mu, \phi, \omega) = 0$. Logo, $\pi_{HZMNB}(Y = y; \mu, \phi, \omega)$ corresponde à distribuição Binomial Negativa Zero-Truncada (ZTNB)[24].

[21] Do inglês: *Negative Binomial*
[22] Do inglês: *Zero-Inflated Negative Binomial*
[23] Do inglês: *Zero-Deflated Negative Binomial*
[24] Do inglês: *Zero-Truncated Negative Binomial*

Com base na definição da distribuição HZMNB, a média, a variância e a f.g.p. são dadas, respectivamente, por

$$\mu_{HZMNB} = \frac{\omega\mu}{1-\left(\frac{\phi}{\mu+\phi}\right)^{\phi}},$$

$$\sigma^2_{HZMNB} = \frac{\omega\left[\mu\left(1+\frac{\mu}{\phi}\right)+\mu^2\right]}{1-\left(\frac{\phi}{\mu+\phi}\right)^{\phi}} - \left[\frac{\omega\mu}{1-\left(\frac{\phi}{\mu+\phi}\right)^{\phi}}\right]^2, \quad \text{e}$$

$$\mathbb{G}_{HZMNB}(t) = 1 - \frac{\omega}{1-\left(\frac{\phi}{\mu+\phi}\right)^{\phi}}\left[1-\left(\frac{\phi}{\mu-\mu t+\phi}\right)^{\phi}\right], \quad 0 \leq t \leq 1.$$

3.5 Síntese

Neste capítulo, apresentamos o desenvolvimento progressivo de famílias de distribuições para dados de contagem. Partiu-se da família PS, que constitui a base para a introdução da família ZMPS. Esta última representa uma extensão flexível, capaz de modelar explicitamente diferentes padrões de frequência de zeros (inflação, deflação ou truncamento) sem impor pressupostos restritivos sobre os dados. Demonstrou-se como diferentes valores do parâmetro ρ, que controla a modificação na probabilidade de zero, definem casos particulares, incluindo as distribuições ZIPS, ZDPS, ZTPS e a própria PS.

Posteriormente, explorou-se a reparametrização que origina a família HZMPS. Esta família admite uma representação como modelo de mistura, mas com uma diferença fundamental em relação às abordagens tradicionais: o parâmetro ω não se restringe ao intervalo $[0,1)$, o que permite acomodar o caso truncado ($\omega = 1$). Apesar desta estrutura, a interpretabilidade é preservada, pois ω continua a governar a modificação na probabilidade de zero, definindo os casos HZIPS, HZDPS e HZTPS. Adicionalmente, destacou-se que o parâmetro ϕ amplia a flexibilidade do modelo para acomodar a dispersão dos dados. Finalmente, detalharam-se as distribuições notáveis pertencentes a esta família, como HZMP, HZMGP, HZMG, HZMB e HZMNB.

CAPÍTULO

4

O MODELO DE FRAGILIDADE HZMGP: ABORDAGEM CLÁSSICA E APLICAÇÕES

Neste Capítulo 4, é proposto um novo modelo de fragilidade para análise de sobrevivência, fundamentado na distribuição HZMGP. A escolha dessa distribuição se justifica pela sua notável flexibilidade em modelar fragilidade discreta, superando alternativas como a distribuição ZMP ao incorporar um parâmetro de dispersão. Este capítulo apresenta uma análise detalhada das propriedades do modelo de fragilidade HZMGP e suas implicações. Além disso, desenvolve-se o processo de estimação dos parâmetros do modelo sob a perspectiva da inferência clássica, por meio do método de máxima verossimilhança.

4.1 Modelo de fragilidade discreta HZMGP

Em diversas áreas, notadamente na medicina, observa-se a existência de um subgrupo da população que não experimenta o evento de interesse, mesmo após longos períodos de seguimento. Este fenômeno, que pode ser atribuído a fatores como a eficácia de tratamentos ou a resiliência imunológica individual, motiva a utilização de modelos de sobrevivência que contemplem uma fração de indivíduos não suscetíveis, ou curados.

Neste contexto, a modelagem da fragilidade por meio de uma distribuição discreta é particularmente adequada. Conforme discutido por Caroni, Crowder e Kimber (2010), a heterogeneidade nos tempos de sobrevivência pode, em certas situações, ser naturalmente representada por uma variável aleatória discreta, como quando esta decorre da exposição a um número aleatório de danos. Assume-se, portanto, que a variável de fragilidade Z pertence ao conjunto dos inteiros não negativos $\{0,1,\dots\}$, com função de massa de probabilidade dada por $\mathbb{P}(Z=z)=q_z$, para $z\in\{0,1,\dots\}$ e $\sum q_z=1$. A função de sobrevivência não condicional (ou marginal) é então obtida pela média da sobrevivência condicional sobre todas as possíveis

realizações de Z

$$S(t) = \sum_{z=0}^{\infty} S(t \mid Z = z)\, q_z = \sum_{z=0}^{\infty} q_z [S_0(t)]^z = \mathbb{G}_Z[S_0(t)], \tag{4.1}$$

em que $\mathbb{G}_Z(\cdot)$ é a f.g.p. da variável aleatória Z. A estrutura do modelo em (4.1) é formalmente equivalente à do modelo de sobrevivência com fração de cura proposto por Rodrigues *et al.* (2009).

A modelagem discreta da fragilidade permite representar de forma direta a existência de indivíduos imunes ao evento de interesse, correspondentes à realização $Z = 0$. A probabilidade deste evento, $q_0 = \mathbb{P}(Z = 0)$, é denominada fração de cura, e pode ser obtida como o limite da função de sobrevivência quando o tempo tende ao infinito, ou seja,

$$\lim_{t\to\infty} S(t) = \lim_{t\to\infty} \sum_{z=0}^{\infty} q_z [S_0(t)]^z = q_0.$$

Assumindo-se que a variável de fragilidade Z segue a distribuição HZMGP, $Z \sim \text{HZMGP}(\mu, \phi, \omega)$, as funções de sobrevivência, risco, densidade e a fração de cura do modelo de fragilidade resultante são dadas, respectivamente, por

$$S(t) = 1 - \frac{\omega}{1 - e^{-\frac{\mu}{1+\mu\phi}}} + \frac{\omega}{1 - e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right]}, \tag{4.2}$$

$$h(t) = -\frac{\omega h_0(t)\, e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right]}}{\left(1 - e^{-\frac{\mu}{1+\mu\phi}}\right) \phi S(t)} \frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1 + \mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}, \tag{4.3}$$

$$f(t) = -\frac{\omega h_0(t)\, e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right]}}{\left(1 - e^{-\frac{\mu}{1+\mu\phi}}\right) \phi} \frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1 + \mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}, \tag{4.4}$$

$$q_0 = 1 - \omega. \tag{4.5}$$

Os cálculos detalhados para a derivação das expressões (4.2), (4.3) e (4.4) são apresentados no Apêndice A. Neste trabalho, as funções de sobrevivência e risco basais, $S_0(t)$ e $h_0(t)$, são especificadas pela distribuição Weibull, cujas formas são

$$\begin{aligned} S_0(t) &= e^{-(\lambda t)^\gamma}, \\ h_0(t) &= \gamma \lambda^\gamma t^{\gamma-1}, \end{aligned} \quad t > 0, \lambda > 0, \gamma > 0. \tag{4.6}$$

Segundo Li, Taylor e Sy (2001), Hanin e Huang (2014), Balan e Putter (2020), uma curva de sobrevivência imprópria pode levar a problemas de identificabilidade do modelo. Uma estratégia eficaz para mitigar esta questão é a incorporação de covariáveis, cuja informação

adicional é frequentemente essencial para assegurar a identificabilidade em modelos de cura complexos.

Em aplicações de análise de sobrevivência, é comum dispor de covariáveis observáveis para cada indivíduo. O modelo de fragilidade pode ser estendido para incorporar os efeitos desses preditores, existindo diferentes abordagens para tal. Uma alternativa consiste em associar as covariáveis diretamente aos parâmetros da função de risco basal. Contudo, conforme apontado por Cancho *et al.* (2020), esta abordagem pode restringir a flexibilidade do modelo em capturar a proporção de curados que varia entre os diferentes níveis das covariáveis. Diante disso, e para maximizar a flexibilidade interpretativa, opta-se por introduzir as covariáveis nos parâmetros da distribuição de fragilidade HZMGP, especificamente em μ e ω. Essa incorporação é realizada por meio de funções de ligação apropriadas, que respeitam o domínio de cada parâmetro. Para $\mu_i > 0$ e $0 < \omega_i < 1$, utilizam-se, respectivamente, as funções de ligação exponencial e logística, detalhadas a seguir

$$\begin{aligned} \mu_i &= e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{\boldsymbol{\mu}}}, \\ \omega_i &= \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{\boldsymbol{\omega}}}}{1+e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{\boldsymbol{\omega}}}}, \end{aligned} \qquad i = 1,\ldots,n, \tag{4.7}$$

em que $\mathbf{x}_i^\top = (1, x_{i1}, \ldots, x_{ip})$ é o vetor $(p+1)\times 1$ contendo os valores das p covariáveis observadas e intercepto para o $i-$ésimo indivíduo, $\boldsymbol{\beta}_{\boldsymbol{\mu}}^\top = (\beta_{01}, \beta_{11}, \ldots, \beta_{p1})$ e $\boldsymbol{\beta}_{\boldsymbol{\omega}}^\top = (\beta_{00}, \beta_{10}, \ldots, \beta_{p0})$ são os vetores dos coeficientes de regressão correspondentes. É importante observar que ambas as funções de ligação incluem a mesma covariável ou o mesmo conjunto de covariáveis.

O uso da distribuição HZMGP em modelos de fragilidade, oferece flexibilidade significativa. Especificamente, dependendo do valor assumido pelo parâmetro ω, originando casos particulares relacionados à frequência de zero, os quais são sumarizados na Figura 3.

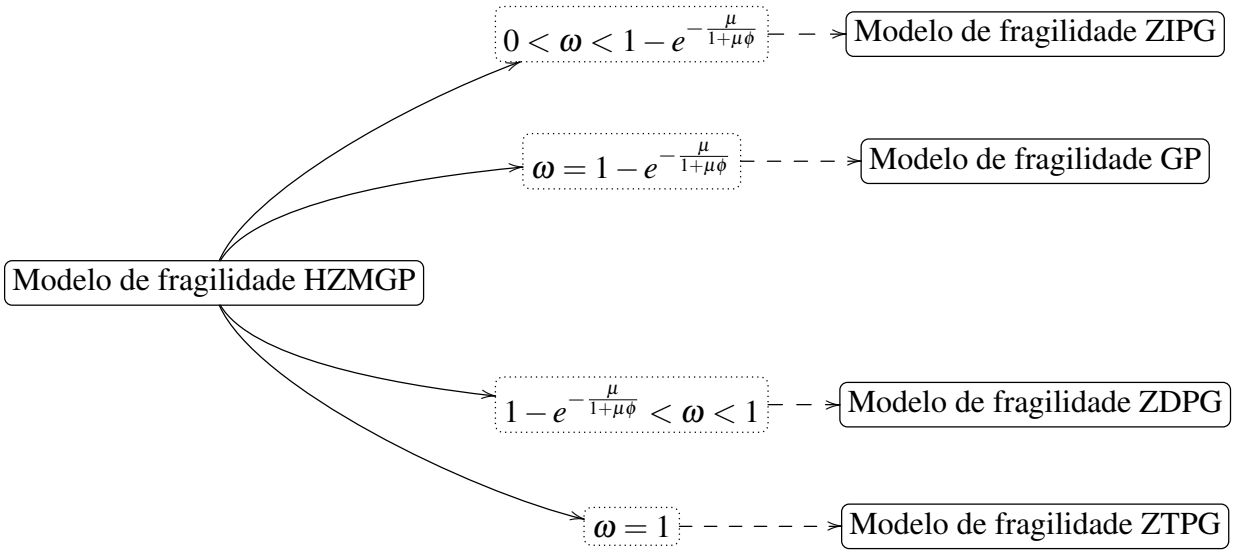

Figura 3 – Modelo de fragilidade HZMGP: Casos particulares segundo a frequência de zero.

Para ilustrar a flexibilidade do modelo de fragilidade HZMGP, realiza-se uma análise gráfica de suas funções características, $S(t)$, $h(t)$ e $f(t)$, conforme definidas em (4.2), (4.3) e (4.4). A análise explora o impacto da inclusão de uma covariável dicotômica (presente/ausente) sob duas configurações paramétricas distintas, permitindo visualizar como as formas das curvas respondem a diferentes valores dos parâmetros. As configurações utilizadas são:

- Configuração 1: $\beta_{01} = 3,70$, $\beta_{11} = -2,30$, $\phi = 0,24$, $\beta_{00} = 0,87$, $\beta_{10} = -1,40$, $\lambda = 0,09$ e $\gamma = 1,30$,
- Configuração 2: $\beta_{01} = 3,70$, $\beta_{11} = -2,30$. $\phi = 0,24$, $\beta_{00} = 2,70$, $\beta_{10} = -0,60$, $\lambda = 0,09$ e $\gamma = 1,30$.

As Figuras 4, 5 e 6 apresentam as curvas resultantes, nas quais a cor verde indica a presença da covariável e a roxa, sua ausência.

A análise da função de sobrevivência na Figura 4 revela a principal virtude do modelo HZMGP. Sob a configuração 1 (painel esquerdo), as curvas $S(t)$ estabilizam-se em um platô após certo tempo, indicando a presença de uma fração de cura. Em contrapartida, na configuração 2 (painel direito), as curvas tendem a zero, sugerindo uma proporção menor ou inexistente de indivíduos curados. Essa dualidade de comportamento demonstra a capacidade do modelo de se adaptar a dados com ou sem evidência de cura, uma flexibilidade essencial em análise de sobrevivência que valida a escolha da distribuição HZMGP.

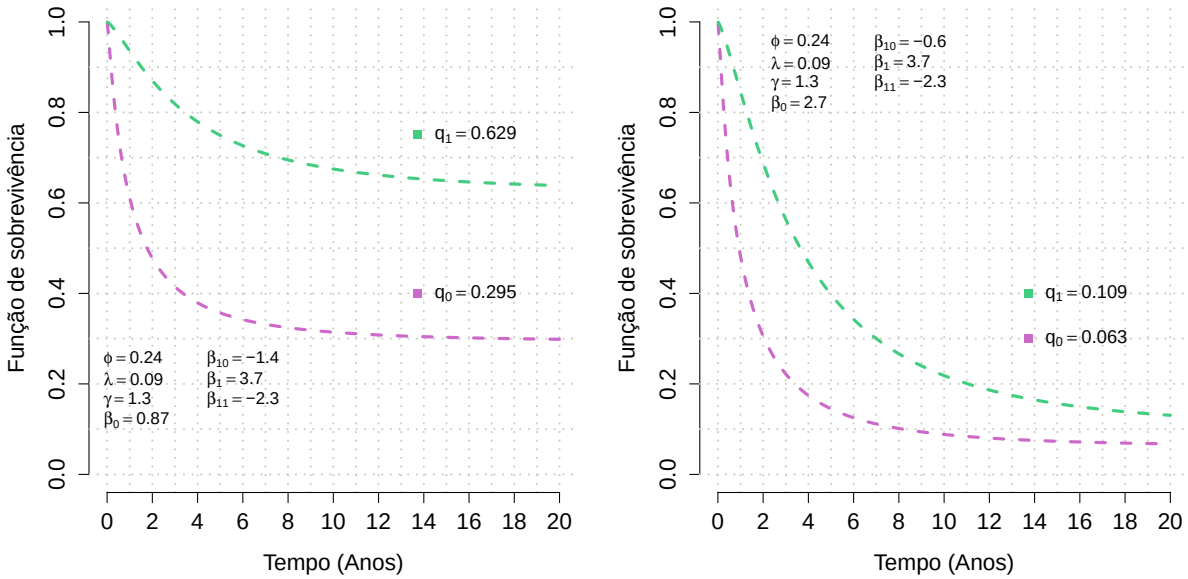

Figura 4 – Comportamento da curva de sobrevivência $S(t)$ do modelo de fragilidade HZMGP.

Fonte: Elaborada pelo autor.

A Figura 5 ilustra os diversos formatos que a função de risco $h(t)$ pode assumir para as configurações. A curva associada à ausência da covariável (cor roxa) revela que o maior

risco se concentra, portanto, no início do período de observação, diminuindo progressivamente. Este padrão sugere um perfil de risco com um perigo inicial agudo que se atenua com o tempo. Por outro lado, a curva associada à presença da covariável (cor verde) inicia com uma taxa de risco baixa. Verificamos uma ligeira tendência de aumento do risco durante aproximadamente o primeiro ano, culminando num pico que pode indicar o período de maior vulnerabilidade para os indivíduos deste grupo. Após este pico, a taxa de risco diminui de forma mais gradual, aproximando-se de zero nos anos subsequentes. Este comportamento dinâmico permite modelar perfis de risco complexos, onde a vulnerabilidade ao evento varia ao longo do tempo.

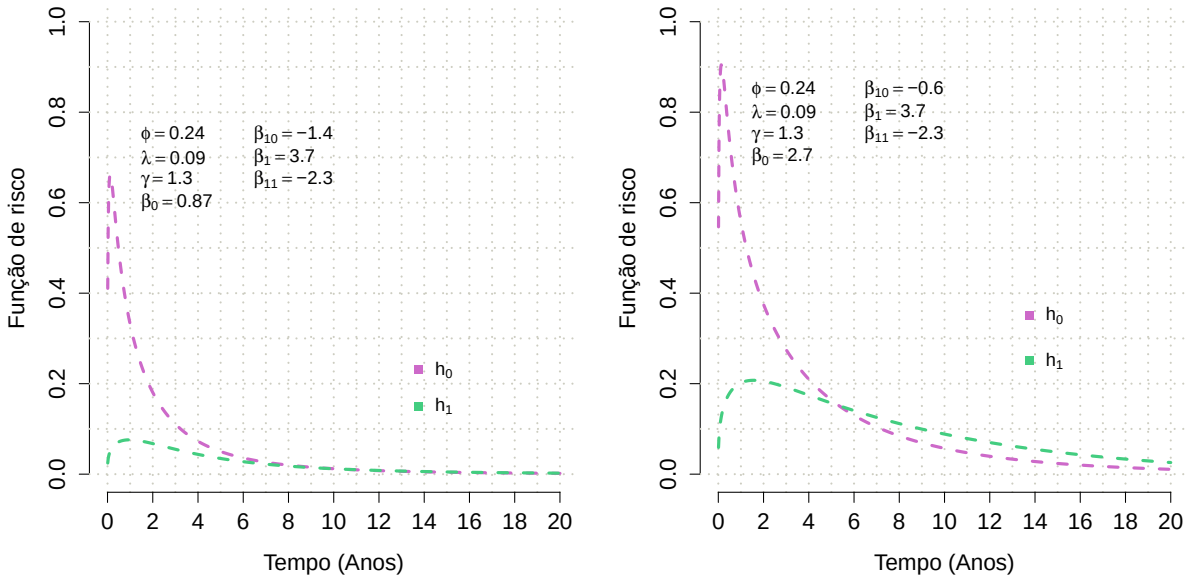

Figura 5 – Comportamento da curva de risco $h(t)$ do modelo de fragilidade HZMGP.

Fonte: Elaborada pelo autor.

Finalmente, a Figura 6 apresenta as funções de densidade $f(t)$, que refletem a distribuição dos tempos de evento. A curva correspondente à ausência da covariável (roxa) caracteriza-se por uma elevada densidade de probabilidade no início, que decai com o tempo, indicando que a ocorrência do evento é mais provável nos momentos iniciais. Em contrapartida, a curva associada à presença da covariável (verde) inicia com uma densidade baixa, que aumenta progressivamente até atingir um pico por volta de 1,5 a 2 anos, para depois diminuir. Este perfil unimodal sugere que, para este grupo, os eventos são menos prováveis no início, concentrando-se em um período posterior.

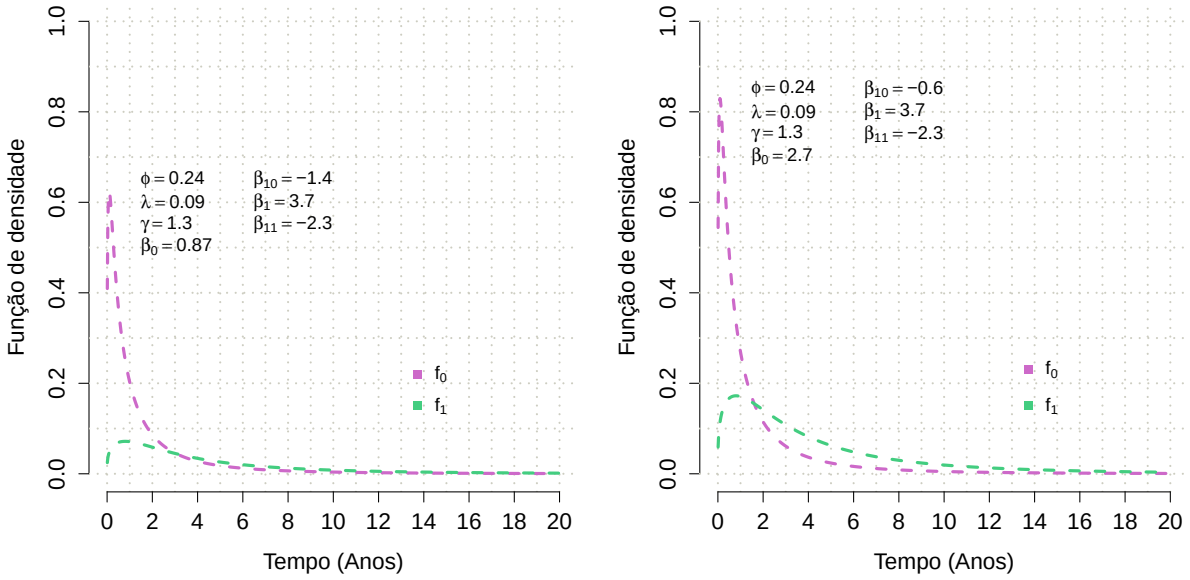

Figura 6 – Comportamento da curva de densidade $f(t)$ do modelo de fragilidade HZMGP.

Fonte: Elaborada pelo autor.

4.2 Inferência

A estimação dos parâmetros do modelo de fragilidade HZMGP é conduzida sob o arcabouço da máxima verossimilhança. Para uma amostra de n indivíduos, o conjunto de dados observado é denotado por $\mathscr{D} = (\mathbf{t}, \delta, \mathbf{x})$, em que $\mathbf{t} = (t_1, \ldots t_n)^\top$ representa o vetor dos tempos observados (tempos de evento ou de censura), $\delta = (\delta_1, \ldots, \delta_n)^\top$ é o vetor dos indicadores de censura correspondentes e $\mathbf{x}$ é a matriz $n \times (p+1)$ contendo os valores das covariáveis associadas a cada indivíduo. O vetor de parâmetros de interesse do modelo é $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_{\boldsymbol{\mu}}^\top, \phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}^\top, \lambda, \gamma)^\top$.

Assumindo-se que os tempos de evento são independentes dos tempos de censura, condição conhecida como censura não informativa, a contribuição do i-ésimo indivíduo para a função de verossimilhança é o produto da função de sobrevivência $S(t_i;\vartheta)$ e da função de risco $h(t_i;\vartheta)$ se o evento não foi censurado. A função de verossimilhança para a amostra completa é, portanto, dada por

$$\mathscr{L}(\mathscr{D} \mid \boldsymbol{\vartheta}) = \prod_{i=1}^{n} S(t_i;\boldsymbol{\vartheta}) \left[h(t_i;\boldsymbol{\vartheta})\right]^{\delta_i}. \tag{4.8}$$

A correspondente log-verossimilhança do modelo de fragilidade HZMGP é $\ell(\boldsymbol{\vartheta}) =$

$\log\left(\mathscr{L}\left(\mathscr{D} \mid \boldsymbol{\vartheta}\right)\right)$, é expressa como

$$\ell\left(\boldsymbol{\vartheta}\right) = \sum_{i=1}^{n} \log\left\{1 - \frac{\omega_i}{1-e^{-\frac{\mu_i}{1+\mu_i\phi}}} + \frac{\omega_i}{1-e^{-\frac{\mu_i}{1+\mu_i\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0(t_i) e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right) + \frac{\mu_i\phi}{1+\mu_i\phi}\right]}\right\} +$$
$$\sum_{i=1}^{n} \delta_i \left\{\log\left[\omega_i h_0\left(t_i\right) e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0(t_i) e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right) + \frac{\mu_i\phi}{1+\mu_i\phi}\right]}\right] - \log\left[\left(e^{-\frac{\mu_i}{1+\mu_i\phi}} - 1\right)\phi S\left(t_i\right)\right]\right\}$$
$$+ \sum_{i=1}^{n} \delta_i \left\{\log\left[\mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0\left(t_i\right) e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right)\right] - \log\left[1 + \mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0\left(t_i\right) e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right)\right]\right\}. \tag{4.9}$$

A obtenção da Estimativa de Máxima Verossimilhança (EMV) para o vetor de parâmetros $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_{\boldsymbol{\mu}}^{\top}, \phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}^{\top}, \lambda, \gamma)^{\top}$ é fundamental para a inferência estatística do modelo proposto. Dada a complexidade da expressão descrita em (4.9), que impede uma solução analítica direta, a maximização é realizada mediante algoritmos de otimização numérica. Neste trabalho, empregou-se o software R Core Team (2025) para este fim, especificamente o algoritmo BFGS.

Sob condições de regularidade padrão, os estimadores de máxima verossimilhança possuem propriedades assintóticas desejáveis. Notavelmente, a distribuição amostral do estimador $\hat{\boldsymbol{\vartheta}}$ converge assintoticamente para uma distribuição Normal multivariada, centrada no verdadeiro vetor de parâmetros $\boldsymbol{\vartheta}_0$ e com uma matriz de variâncias e covariâncias cujo estimador é $\widehat{\Sigma}(\boldsymbol{\vartheta}) = \left\{-\left.\frac{\partial^2 l(\boldsymbol{\vartheta} \mid \mathbf{x})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}^T}\right|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}}\right\}^{-1}$. Assim, à medida que $n \longrightarrow \infty$, tem-se que

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \xrightarrow{D} \mathscr{N}_k\left(\mathbf{0}, \Sigma_{\hat{\boldsymbol{\vartheta}}}\right),$$

em que k é a dimensão do vetor de parâmetros $\boldsymbol{\vartheta}$.

A propriedade de normalidade assintótica permite a construção de um Intervalo de Confiança (IC) aproximado de $100(1-\alpha)\%$ para cada componente ϑ_i do vetor de parâmetros

$$\mathrm{IC}_{100(1-\alpha)\%}(\vartheta_i) = \widehat{\vartheta}_i \pm z_{1-\alpha/2}\sqrt{\widehat{\Sigma}_{ii}},$$

em que $\hat{\vartheta}_i$ é a EMV, $\widehat{\Sigma}_{ii}$ é o i-ésimo elemento da diagonal da matriz de variâncias-covariâncias estimada, e $z_{1-\alpha/2}$ é o quantil correspondente da distribuição Normal padrão.

4.3 Estudo de simulação

Para avaliar o desempenho do método de estimação e as propriedades dos estimadores de máxima verossimilhança para o modelo de fragilidade HZMGP, foi conduzido um extenso estudo de simulação de MC. O principal objetivo é investigar o comportamento dos estimadores em amostras finitas sob diferentes condições, com foco na consistência, precisão e na cobertura dos IC.

O estudo de simulação foi estruturado em torno de dois cenários principais, projetados para reproduzir condições de zero-inflação (ZIGP) e zero-deflação (ZDGP), que representam os casos particulares mais relevantes do modelo HZMGP. Em ambos os cenários, a geração de dados incorpora uma covariável binária, seguindo a estrutura de regressão definida em (4.7), e utiliza a distribuição Weibull como modelo de risco basal, conforme especificado em (4.6).

O desempenho dos estimadores é avaliado para um conjunto variado de tamanhos amostrais (n). Para cada combinação de cenário e tamanho amostral, os dados $\mathscr{D} = \{(\mathbf{t}_i, \delta_i, \mathbf{x}_i)\}_{i=1}^{n}$ são gerados por meio do método da inversa da função de sobrevivência, detalhado no Algoritmo 1.

Algoritmo 1 – Geração de uma amostra para o modelo de fragilidade HZMGP.

Entradas:

n: tamanho da amostra.

$\boldsymbol{\vartheta} = (\boldsymbol{\beta}_{\boldsymbol{\mu}}^{\top}, \phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}^{\top}, \lambda, \gamma)^{\top}$: vetor de parâmetros verdadeiros.

U_{max}: limite superior para a busca da raiz (tempo máximo de seguimento).

Procedimento:

para $i = 1, \ldots, n$ **faça**

  Gerar a covariável: $x_i \sim \text{Bernoulli}(0,5)$.

  Gerar um número aleatório: $u_i \sim \mathscr{U}(0,1)$.

  Resolver a equação $S(t; x_i, \boldsymbol{\vartheta}) = u_i$ para t no intervalo $[0, U_{max}]$, onde $S(\cdot)$ é a função de sobrevivência em (4.2).

  se uma raiz t_i^* for encontrada em $[0, U_{max}]$ **então**

    $t_{\text{potencial},i} \leftarrow t_i^*$. ▷ Tempo de evento potencial

  senão

    $t_{\text{potencial},i} \leftarrow \infty$. ▷ Indivíduo curado ou censurado além de U_{max}

  fim se

fim para

Definir o tempo máximo de evento observado: $t_c \leftarrow \max\{t_{\text{potencial},i} \mid t_{\text{potencial},i} < \infty\}$. ▷ Mecanismo de censura do tipo I adaptativo

para $i = 1, \ldots, n$ **faça**

  se $t_{\text{potencial},i} \leq t_c$ **então**

    $t_i \leftarrow t_{\text{potencial},i}$. ▷ Tempo de evento

    $\delta_i \leftarrow 1$.

  senão

    $t_i \leftarrow t_c$. ▷ Tempo de censura

    $\delta_i \leftarrow 0$.

  fim se

fim para

Saídas:

O conjunto de dados simulado: $\mathscr{D} = \{(t_i, \delta_i, x_i)\}_{i=1}^{n}$.

A fim de conferir maior realismo ao estudo, definiram-se os parâmetros para os dois cenários de simulação com base nos resultados do ajuste de um modelo aos conjuntos de dados reais, detalhados na Seção 1.2. Em cada ajuste, incorporou-se a covariável cirurgia e, subsequentemente, utilizaram-se as EMV obtidas como os verdadeiros parâmetros para os respectivos cenários.

Em cada cenário, realizaram-se estudos de simulação via Monte Carlo com 100 réplicas para os tamanhos amostrais $n \in \{100, 500, 1000, 5000, 10000, 15000, 20000\}$. Avaliou-se o desempenho dos estimadores de máxima verossimilhança por meio das seguintes métricas: Média das Estimativas de Máxima Verossimilhança (MEMV), Desvio Padrão (DP), Raiz do Erro Quadrático Médio (REQM) e Probabilidade de Cobertura (PC). A consistência da PC com o nível nominal de 95% foi aferida pela pertença do valor nominal ao IC (0,907; 0,993), obtido a partir da distribuição Binomial$(100, 0,95)$.

Para facilitar a análise visual dos resultados, padronizou-se a apresentação dos gráficos com elementos de referência. O conjunto completo de gráficos pode ser consultado no Apêndice B. A padronização consiste em uma linha tracejada fúcsia que representa o valor verdadeiro do parâmetro (na análise da MEMV), o valor-alvo zero (para DP e REQM) e o nível nominal de 95% (para a PC). Complementarmente, nesta última, o referido IC (0,907; 0,993) é delimitado por duas linhas verdes tracejadas.

4.3.1 Cenário 1: modelo de fragilidade ZIGP

O primeiro cenário investiga o desempenho do modelo em uma condição de zero-inflação por meio do modelo de fragilidade ZIGP, utilizando a seguinte configuração dos parâmetros: $\beta_{01} = 3,70$, $\beta_{11} = -2,30$, $\phi = 0,24$, $\beta_{00} = 0,87$, $\beta_{10} = -1,40$, $\lambda = 0,09$ e $\gamma = 1,30$. Os resultados detalhados são apresentados na Tabela 8.

A análise da Tabela 8 confirma as boas propriedades assintóticas dos estimadores. A MEMV converge para os valores verdadeiros à medida que o tamanho amostral n aumenta, evidenciando a consistência dos estimadores, como se visualiza na Figura 9. A precisão, avaliada pelo DP e pela REQM, melhora sistematicamente com o aumento de n, com ambas as métricas tendendo a zero, conforme ilustrado nas Figuras 10 e 11. Nota-se, contudo, que o parâmetro β_{01} apresenta uma REQM mais elevada em amostras menores, indicando uma convergência mais lenta para este coeficiente. A PC dos IC de 95%, apresentada na Figura 12, mantém-se próxima do nível nominal na maioria das configurações. Uma exceção pontual é observada para β_{01} com $n = 1000$, cuja PC se situa ligeiramente abaixo do esperado. Para os demais parâmetros e tamanhos amostrais, a cobertura nominal foi adequadamente alcançada.

Tabela 8 – Resultados da simulação para o modelo de fragilidade ZIGP.

	Parâmetros	MEMV	DP	REQM	PC		MEMV	DP	REQM	PC
n=100	$\beta_{01} = 3,70$	3,905	1,412	1,420	0,970	n=10000	3,682	0,430	0,428	0,940
	$\beta_{11} = -2,30$	-2,713	1,358	1,413	0,920		-2,322	0,120	0,121	0,960
	$\phi = 0,24$	0,254	0,519	0,517	0,990		0,242	0,034	0,034	0,950
	$\beta_{00} = 0,87$	0,899	0,356	0,356	0,950		0,868	0,031	0,031	0,930
	$\beta_{10} = -1,40$	-1,259	0,536	0,552	0,960		-1,398	0,044	0,044	0,910
	$\lambda = 0,09$	0,087	0,029	0,029	1,000		0,092	0,021	0,021	0,930
	$\gamma = 1,30$	1,314	0,250	0,249	0,930		1,294	0,035	0,036	0,960
n=500	$\beta_{01} = 3,70$	3,571	0,950	0,954	0,980	n=15000	3,658	0,385	0,385	0,930
	$\beta_{11} = -2,30$	-2,367	0,708	0,708	0,930		-2,306	0,109	0,109	0,950
	$\phi = 0,24$	0,213	0,086	0,090	0,980		0,242	0,032	0,032	0,930
	$\beta_{00} = 0,87$	0,892	0,143	0,144	0,950		0,868	0,026	0,026	0,905
	$\beta_{10} = -1,40$	-1,400	0,197	0,196	0,960		-1,399	0,036	0,036	0,940
	$\lambda = 0,09$	0,094	0,045	0,044	0,970		0,092	0,020	0,020	0,940
	$\gamma = 1,30$	1,259	0,129	0,135	0,920		1,290	0,028	0,029	0,950
n=1000	$\beta_{01} = 3,70$	3,619	0,996	0,994	0,890	n=20000	3,744	0,379	0,380	0,940
	$\beta_{11} = -2,30$	-2,513	0,784	0,809	0,960		-2,303	0,103	0,103	0,930
	$\phi = 0,24$	0,240	0,075	0,075	0,970		0,238	0,034	0,033	0,940
	$\beta_{00} = 0,87$	0,850	0,104	0,105	0,950		0,868	0,022	0,022	0,940
	$\beta_{10} = -1,40$	-1,391	0,135	0,135	0,960		-1,397	0,031	0,031	0,950
	$\lambda = 0,09$	0,099	0,046	0,046	0,920		0,089	0,020	0,020	0,930
	$\gamma = 1,30$	1,268	0,111	0,114	0,910		1,297	0,026	0,026	0,950
n=5000	$\beta_{01} = 3,70$	3,778	0,500	0,503	0,960					
	$\beta_{11} = -2,30$	-2,284	0,179	0,179	0,950					
	$\phi = 0,24$	0,229	0,040	0,041	0,960					
	$\beta_{00} = 0,87$	0,875	0,044	0,044	0,950					
	$\beta_{10} = -1,40$	-1,402	0,058	0,058	0,960					
	$\lambda = 0,09$	0,085	0,024	0,025	0,940					
	$\gamma = 1,30$	1,288	0,046	0,047	0,970					

Fonte: Elaborada pelo autor.

4.3.2 Cenário 2: modelo de fragilidade ZDGP

O segundo cenário avalia o modelo sob uma condição de zero-deflação por meio do modelo de fragilidade modelo ZDGP, com os seguintes valores fixados para os parâmetros: $\beta_{01} = 2,50$, $\beta_{11} = -1$,$\phi = 0,13$, $\beta_{00} = 4,80$, $\beta_{10} = -3,30$, $\lambda = 0,14$ e $\gamma = 1,07$. Os resultados numéricos são apresentados na Tabela 9.

A análise da PC, ilustrada na Figura 16, mostra que os IC performam bem para amostras de tamanho moderado a grande. No entanto, para $n = 100$, as PCs para os parâmetros γ e β_{00} situam-se abaixo do limite inferior do intervalo esperado, sugerindo que a aproximação assintótica para a distribuição destes estimadores pode ser menos confiável quando a informação

Tabela 9 – Resultados da simulação para o modelo de fragilidade ZDGP.

	Parâmetros	EMV	DP	REQM	PC		EMV	DP	REQM	PC
n=100	$\beta_{01} = 2,50$	2,770	1,845	1,855	0,990	n=10000	2,589	0,321	0,312	0,940
	$\beta_{11} = -1$	-1,174	1,382	1,386	0,980		-1,010	0,042	0,045	0,960
	$\phi = 0,13$	0,071	0,320	0,324	0,990		0,128	0,013	0,015	0,960
	$\beta_{00} = 4,80$	4,072	0,825	1,097	0,870		4,847	0,186	0,191	0,980
	$\beta_{10} = -3,30$	-2,360	0,863	1,273	0,910		-3,352	0,189	0,195	0,950
	$\lambda = 0,14$	0,150	0,131	0,131	0,950		0,142	0,030	0,024	0,930
	$\gamma = 1,07$	1,113	0,181	0,186	0,900		1,069	0,029	0,028	0,970
n=500	$\beta_{01} = 2,50$	2,835	0,840	0,901	0,950	n=15000	2,544	0,258	0,261	0,930
	$\beta_{11} = -1$	-1,159	0,574	0,593	0,970		-1,009	0,036	0,037	0,960
	$\phi = 0,13$	0,129	0,064	0,063	0,990		0,129	0,009	0,009	0,950
	$\beta_{00} = 4,80$	4,967	0,880	0,891	0,980		4,847	0,158	0,164	0,940
	$\beta_{10} = -3,30$	-3,429	0,857	0,863	0,990		-3,345	0,161	0,166	0,920
	$\lambda = 0,14$	0,131	0,065	0,065	0,960		0,138	0,026	0,026	0,920
	$\gamma = 1,07$	1,088	0,089	0,090	0,960		1,070	0,014	0,014	0,970
n=1000	$\beta_{01} = 2,50$	2,785	0,568	0,633	0,950	n=20000	2,533	0,230	0,231	0,940
	$\beta_{11} = -1$	-1,056	0,149	0,159	0,990		-1,002	0,031	0,031	0,960
	$\phi = 0,13$	0,129	0,037	0,037	0,980		0,128	0,010	0,010	0,950
	$\beta_{00} = 4,80$	5,058	0,768	0,806	0,960		4,799	0,114	0,113	0,960
	$\beta_{10} = -3,30$	-3,537	0,780	0,812	0,960		-3,302	0,112	0,112	0,960
	$\lambda = 0,14$	0,124	0,053	0,055	0,940		0,138	0,024	0,024	0,950
	$\gamma = 1,07$	1,079	0,054	0,054	0,970		1,069	0,013	0,013	0,930
n=5000	$\beta_{01} = 2,50$	2,585	0,321	0,330	0,950					
	$\beta_{11} = -1$	-1,012	0,075	0,076	0,950					
	$\phi = 0,13$	0,128	0,016	0,016	0,970					
	$\beta_{00} = 4,80$	4,849	0,273	0,276	0,960					
	$\beta_{10} = -3,30$	-3,345	0,278	0,280	0,960					
	$\lambda = 0,14$	0,134	0,037	0,037	0,920					
	$\gamma = 1,07$	1,066	0,021	0,021	0,980					

Fonte: Elaborada pelo autor.

amostral é muito limitada. Para $n \geq 500$, a cobertura se alinha com o nível nominal de 95%.

De forma análoga ao cenário anterior, os resultados na Tabela 9 confirmam o comportamento assintótico esperado dos estimadores. A Figura 13 ilustra a convergência da MEMV para os valores teóricos à medida que n aumenta. As Figuras 14 e 15 demonstram a melhoria na precisão com o aumento de n. No entanto, os coeficientes de regressão (β_{00}, β_{10}, β_{01}, β_{11}) exibem uma REQM maior em amostras pequenas ($n \leq 1000$), uma questão que se atenua a partir de $n \geq 5000$, indicando a necessidade de amostras maiores para uma estimação precisa destes efeitos. A análise da PC, ilustrada na Figura 16, mostra que os IC apresentam bom desempenho para amostras de tamanho moderado a grande. Contudo, para $n = 100$, as PCs para γ e β_{00}

situam-se abaixo do limite inferior esperado, sugerindo que a aproximação assintótica pode ser menos confiável quando a informação amostral é muito limitada. Para $n \geq 500$, a cobertura alinha-se com o nível nominal de 95%.

4.4 Aplicação em dados reais

Após a fundamentação teórica e a avaliação por simulação, a aplicabilidade do modelo de fragilidade HZMGP é investigada por meio da análise de dois conjuntos de dados oncológicos. O objetivo é demonstrar a utilidade e a versatilidade do modelo em cenários empíricos contrastantes: um sobre câncer de melanoma, que exibe forte evidência de uma fração de cura, e outro sobre câncer de pulmão, em que tal fração não é aparente na análise exploratória. Estes estudos de caso permitem validar a aplicabilidade do modelo e destacar sua capacidade de fornecer *insights* em diferentes contextos clínicos.

A abordagem metodológica adotada em ambas as aplicações será consistente com as definições apresentadas anteriormente. A modelagem das funções de risco e de sobrevivência base será realizada por meio da distribuição Weibull, conforme especificado em (4.6), e as covariáveis são incorporadas aos parâmetros ω e μ da fragilidade por meio da estrutura de regressão definida em (4.7).

A estimação dos parâmetros é realizada pelo método de máxima verossimilhança, com a otimização numérica conduzida pelo algoritmo quasi-Newton BFGS, implementado na função `optim` do software R. Para cada parâmetro, reportam-se a EMV, o Erro Padrão (EP) e o IC de 95%. Adicionalmente, o EP para as frações de cura estimadas é calculado utilizando o método delta, baseado na aproximação de Taylor de primeira ordem.

4.4.1 Aplicação 1: Câncer de melanoma

Com o objetivo de avaliar o impacto simultâneo das covariáveis na sobrevida dos pacientes, procedeu-se ao ajuste do modelo de fragilidade HZMGP. Contudo, o algoritmo de otimização numérica não alcançou a convergência, uma falha que persistiu mesmo após múltiplas tentativas de ajuste, utilizando-se diferentes conjuntos de valores iniciais. A ausência de convergência foi confirmada pelo código de status retornado pelo método de otimização empregado, o BFGS. Nesse método, um código de retorno igual a 1 indica que o número máximo de iterações foi atingido sem que os critérios de convergência fossem satisfeitos.

Diante disso, é imperativo ressaltar que as estimativas e as medidas de precisão apresentadas na Tabela 10 não constituem EMV válidas. Consequentemente, os valores reportados devem ser interpretados com extrema cautela, servindo apenas como um registro da tentativa de ajuste do modelo. Qualquer inferência estatística formal baseada nestes resultados seria metodologicamente inválida. Acredita-se que a falha na convergência possa ser atribuída a uma

combinação de fatores, como a elevada complexidade do modelo ou um tamanho amostral insuficiente para estimar de forma estável todos os seus parâmetros.

Longe de ser um resultado puramente negativo, a falha de convergência corrobora as conclusões do estudo de simulação. O estudo prévio já indicava a necessidade de amostras robustas para garantir a estabilidade das estimativas em modelos complexos. Portanto, a dificuldade encontrada no ajuste com os dados reais serve como uma demonstração prática dessa exigência metodológica, validando os resultados da simulação.

Tabela 10 – EMV, EP, IC de 95% para os parâmetros do modelo de fragilidade HZMGP, ajustado aos dados de câncer de melanoma utilizando todas as covariáveis.

			IC 95%	
Parâmetros	EMV	EP	Inferior	Superior
β_{01}	0,092	0,899	-1,670	1,854
β_{11} (Idade)	0,000	0,069	-0,137	0,136
β_{21} (Feminino)	-0,258	0,118	-0,490	-0,027
β_{31} (EC II)	0,604	0,731	-0,828	2,036
β_{41} (EC III)	2,308	0,768	0,802	3,813
β_{51} (EC IV)	4,196	0,806	2,616	5,776
β_{61} (Cir: S)	-1,199	0,137	-1,467	-0,931
β_{71} (Rad: S)	-0,135	0,144	-0,417	0,147
β_{81} (Qui: S)	-0,962	0,151	-1,259	-0,666
ϕ	0,185	0,021	0,144	0,226
β_{00}	-1,296	0,174	-1,636	-0,956
β_{10} (Idade)	0,331	0,046	0,240	0,421
β_{20} (Feminino)	-0,503	0,088	-0,676	-0,331
β_{30} (EC II)	1,386	0,124	1,143	1,629
β_{40} (EC III)	2,046	0,125	1,801	2,291
β_{50} (EC IV)	3,105	0,170	2,771	3,438
β_{60} (Cir: S)	-0,481	0,152	-0,780	-0,183
β_{70} (Rad: S)	1,188	0,201	0,793	1,582
β_{80} (Qui: S)	1,254	0,144	0,971	1,536
λ	0,160	0,014	0,132	0,188
γ	1,451	0,041	1,370	1,531
$\max \ell(\cdot)$		$-5392,181$		

Fonte: Elaborada pelo autor.

4.4.2 Aplicação 2: Câncer de pulmão

Com o intuito de validar a aplicabilidade do modelo proposto em um novo cenário, ajustou-se o modelo de fragilidade HZMGP a um conjunto de dados de sobrevida de pacientes

com câncer de pulmão. A análise procurou mensurar o efeito conjunto das covariáveis observadas. Os resultados deste ajuste são detalhados na Tabela 11.

Tabela 11 – EMV, EP, IC de 95% para os parâmetros do modelo de fragilidade HZMGP, ajustado aos dados de câncer de pulmão utilizando todas as covariáveis.

			IC 95%	
Parâmetro	EMV	EP	Inferior	Superior
β_{01}	2,407	0,069	2,271	2,542
β_{11} (Idade)	-0,061	0,011	-0,082	-0,039
β_{21} (Feminino)	-0,267	0,022	-0,311	-0,223
β_{31} (EC III-IV)	1,662	0,054	1,555	1,768
β_{41} (Cir: S)	-1,291	0,044	-1,378	-1,204
β_{51} (Rad: S)	-0,657	0,023	-0,703	-0,612
β_{61} (Qui: S)	-1,936	0,032	-1,998	-1,873
ϕ	0,140	0,003	0,133	0,147
β_{00}	2,357	0,106	2,150	2,565
β_{10} (Idade)	0,452	0,039	0,376	0,529
β_{20} (Feminino)	-0,601	0,077	-0,753	-0,449
β_{30} (EC III-IV)	2,319	0,096	2,131	2,507
β_{40} (Cir: S)	-1,953	0,097	-2,143	-1,763
β_{50} (Rad: S)	0,355	0,108	0,144	0,566
β_{60} (Qui: S)	0,739	0,088	0,566	0,911
λ	0,266	0,007	0,252	0,280
γ	1,266	0,008	1,251	1,281
max $\ell(\cdot)$		$-28727,500$		

Fonte: Elaborada pelo autor.

As estimativas dos parâmetros, apresentadas na Tabela 11, permitem as seguintes inferências:

- Todas as covariáveis associadas ao parâmetro μ, ($\hat{\beta}_{01}$, $\hat{\beta}_{11}$, $\hat{\beta}_{21}$, $\hat{\beta}_{31}$, $\hat{\beta}_{41}$, $\hat{\beta}_{51}$ e $\hat{\beta}_{61}$), mostraram-se estatisticamente significativas. Como exemplo, o diagnóstico de um paciente em estágio avançado (EC III-IV) é associado a um aumento no risco de óbito.

- A estimativa $\hat{\phi} = 0,140$ (IC 95%: [0,133; 0,147]) é estatisticamente significativa, confirmando a presença de heterogeneidade não capturada pelas covariáveis e justificando o uso de um modelo de fragilidade.

- De maneira similar, verificou-se que todas as covariáveis associadas ao parâmetro ω, ($\hat{\beta}_{00}$, $\hat{\beta}_{10}$, $\hat{\beta}_{20}$, $\hat{\beta}_{30}$, $\hat{\beta}_{40}$, $\hat{\beta}_{50}$ e $\hat{\beta}_{60}$), exerceram influência estatisticamente significativa. A análise

dos coeficientes indicou que, por exemplo, ser do sexo feminino e ter realizado cirurgia estão associados a uma maior probabilidade de sobreviver ao câncer de pulmão.

- Com base em $\hat{\gamma} = 1,266$ (IC 95%: [1,251; 1,281]), rejeita-se a hipótese de um risco basal constante (modelo Exponencial). O valor sugere que o risco de óbito, na ausência de fragilidade, aumenta com o tempo.

Uma das principais características do modelo HZMGP é a sua capacidade de transcender a análise de coeficientes, permitindo uma estratificação prognóstica individualizada. Essa estratificação é realizada a partir das estimativas dos parâmetros e do perfil de covariáveis de cada paciente. Para cada indivíduo, calcularam-se os termos ω_i e μ_i e, subsequentemente, procedeu-se à alocação de cada paciente ao regime de zero-inflação (ZIGP) ou zero-deflação (ZDGP), comparando-se esses termos com o limite crítico $1 - e^{(-\mu_i/(1+\mu_i\phi))}$.

Os resultados dessa classificação, sintetizados na Tabela 12, revelam um panorama particularmente interessante. Ainda que a análise exploratória inicial não sugerisse a existência de uma fração de cura, a aplicação do modelo HZMGP foi capaz de discernir dois subgrupos prognósticos latentes na população. Verificou-se que 33,50% dos pacientes foram alocados ao regime ZIGP, o qual está associado a uma maior propensão à cura (com $q_{0,i} \geq 0,20$), enquanto os 66,50% restantes foram classificados no regime ZDGP, caracterizado por uma menor probabilidade de cura ($q_{0,i} < 0,20$). Essa capacidade de revelar uma heterogeneidade prognóstica não aparente evidencia o valor do modelo como uma ferramenta de análise mais sensível, capaz de identificar estruturas complexas nos dados que seriam omitidas por abordagens convencionais.

Tabela 12 – Proporções da classificação do modelo de fragilidade HZMGP para o conjunto de dados de câncer de pulmão.

	ZIGP	GP	ZDGP	ZTGP
Proporção	0,335	0	0,665	0

Fonte: Elaborada pelo autor.

Para além da estratificação prognóstica, avaliou-se também a capacidade preditiva do modelo por meio das frações de cura estimadas para distintos perfis de pacientes, cujos valores são detalhados na Tabela 13. A análise dessas estimativas permitiu destacar os seguintes pontos de relevância clínica:

- Observa-se uma vantagem prognóstica consistente para o gênero feminino: mantendo-se a idade em 63 anos, as mulheres demonstram probabilidades de cura sistematicamente superiores às dos homens em todos os regimes terapêuticos.

- Ressalta-se o papel da intervenção cirúrgica como o fator mais associado a um resultado favorável. Como exemplo, uma paciente de 63 anos, no estadiamento I-II, tratada exclusi-

vamente com cirurgia, alcança uma fração de cura estimada de 0,549. Este valor contrasta drasticamente com a probabilidade de apenas 0,055 para uma paciente com o mesmo perfil que foi submetida a quimioterapia e radioterapia, mas não à cirurgia.

Este segundo resultado, embora aparentemente contraintuitivo, pode ser explicado pela própria estrutura do modelo de fragilidade HZMGP. O valor estimado do parâmetro de dispersão, ϕ, reflete precisamente a heterogeneidade não observada entre os pacientes, a qual pode englobar, por exemplo, diferenças biológicas subjacentes, como subtipos moleculares tumorais ou a condição geral do paciente, que não foram explicitamente avaliadas.

Sem invalidar a eficácia das terapias oncológicas, propõe-se que o parâmetro de dispersão ϕ do modelo HZMGP represente a heterogeneidade biológica não observada, como a variabilidade genética e molecular entre os indivíduos. Esta interpretação, que contextualiza biologicamente o componente de fragilidade, é robustamente suportada por estudos multiômicos seminais (ZHANG; WANG; QIAN, 2024; WANG *et al.*, 2025). Tais pesquisas demonstram a existência de subtipos moleculares no câncer de pulmão, como o Adenocarcinoma de pulmão (LUAD)[1], com prognósticos e respostas a terapias-padrão intrinsecamente distintos (ZHANG; WANG; QIAN, 2024). Adicionalmente, foram identificadas assinaturas moleculares que predizem a recorrência pós-cirúrgica em câncer de pulmão de células não pequenas (NSCLC)[2] em estágio I, validando que a diversidade biológica do tumor é um determinante crítico dos desfechos clínicos (WANG *et al.*, 2025).

Nesta perspectiva, as frações de cura mais elevadas em perfis de pacientes submetidos apenas a cirurgia não constituem um paradoxo, mas sim um reflexo da interação entre a terapia e a heterogeneidade biológica latente. É plausível que esses pacientes possuam subtipos tumorais com prognóstico inerentemente favorável e/ou menor sensibilidade às terapias adjuvantes. Em contrapartida, a coorte que recebeu tratamento sistêmico é, por natureza, heterogênea, englobando tanto respondentes quanto não respondentes, além de indivíduos com perfis biológicos mais agressivos, cujo prognóstico permanece reservado apesar da intervenção terapêutica.

[1] Do Inglês: *Lung adenocarcinoma*
[2] Do Inglês: *Non-small cell lung cancer*

Tabela 13 – Estimativas da fração de cura (q_0), EP e IC de 95% para perfis de pacientes definidos por idade, gênero, EC, cirurgia, radioterapia e quimioterapia, com base no conjunto de dados de câncer de pulmão.

Idade	Gênero	EC	Cir.	Rad.	Qui.	q_0	EP	IC 95%	
								LI	LS
63	M	I ou II	N	N	N	0,086	0,009	0,068	0,104
					S	0,043	0,018	0,008	0,078
				S	N	0,062	0,025	0,013	0,111
					S	0,031	0,020	0,000	0,070
			S	N	N	0,400	0,016	0,370	0,430
					S	0,242	0,031	0,181	0,303
				S	N	0,319	0,028	0,265	0,373
					S	0,183	0,041	0,103	0,263
		III ou IV	N	N	N	0,009	0,006	0,000	0,021
					S	0,004	0,004	0,000	0,012
				S	N	0,006	0,006	0,000	0,018
					S	0,003	0,003	0,000	0,010
			S	N	N	0,062	0,035	0,000	0,130
					S	0,030	0,028	0,000	0,085
				S	N	0,044	0,035	0,000	0,112
					S	0,022	0,024	0,000	0,068
63	F	I ou II	N	N	N	0,147	0,023	0,103	0,191
					S	0,076	0,029	0,019	0,133
				S	N	0,108	0,035	0,039	0,177
					S	0,055	0,032	0,000	0,118
			S	N	N	0,549	0,013	0,523	0,575
					S	0,368	0,027	0,316	0,420
				S	N	0,460	0,022	0,417	0,503
					S	0,290	0,037	0,217	0,363
		III ou IV	N	N	N	0,017	0,012	0,000	0,041
					S	0,008	0,008	0,000	0,024
				S	N	0,012	0,012	0,000	0,035
					S	0,006	0,007	0,000	0,019
			S	N	N	0,107	0,045	0,018	0,196
					S	0,054	0,042	0,000	0,136
				S	N	0,077	0,048	0,000	0,171
					S	0,039	0,037	0,000	0,112

Fonte: Elaborada pelo autor.

4.5 Síntese

Neste capítulo, introduziu-se e desenvolveu-se um novo modelo de fragilidade baseado na distribuição HZMGP. Sua formulação representa uma generalização de trabalhos anteriores, como os de Cancho *et al.* (2020), Molina *et al.* (2021), por meio da inclusão de um parâmetro de dispersão, ϕ. Com efeito, os modelos mencionados são casos particulares da HZMGP quando este parâmetro é nulo. Além dessa generalização, o modelo preserva a flexibilidade de acomodar múltiplos padrões de zeros, o que o torna uma ferramenta robusta para a análise de heterogeneidade em dados de sobrevida.

A validação das propriedades do modelo foi realizada por meio de estudos de simulação via MC, nos quais se avaliou o desempenho dos estimadores sob uma gama de cenários, incluindo-se as estruturas de zero-inflação e zero-deflação. Os resultados não apenas confirmaram as esperadas propriedades assintóticas dos estimadores de máxima verossimilhança, mas, de forma crucial, revelaram um requisito prático para sua aplicação bem-sucedida: a elevada complexidade do modelo HZMGP demanda um tamanho amostral substancial para assegurar a estabilidade e a precisão da inferência estatística.

A aplicabilidade e os limites práticos do modelo foram então examinados mediante o ajuste a dois conjuntos de dados oncológicos. O ajuste ao conjunto de dados de câncer de pulmão foi bem-sucedido, demonstrando a notável capacidade do modelo de estratificar os pacientes em grupos prognósticos distintos e de revelar uma heterogeneidade latente. Em contrapartida, na aplicação aos dados de melanoma, o algoritmo de otimização não alcançou a convergência. Longe de invalidar o modelo, este resultado serviu como uma poderosa validação empírica das conclusões da simulação. Ele demonstrou na prática que a estabilidade da inferência no modelo HZMGP é sensível a uma combinação de fatores. Primeiramente, a complexidade do modelo demanda um volume de dados substancial, que não foi suficiente no conjunto de dados de melanoma. Em segundo lugar, acredita-se que esta dificuldade de estimação tenha sido exacerbada pela alta dimensionalidade do espaço paramétrico, uma consequência da inclusão de um grande número de regressores nos componentes ω e μ. A combinação de uma amostra relativamente pequena com um modelo ricamente parametrizado gerou, portanto, a instabilidade que impediu a convergência.

Portanto, as contribuições deste capítulo são duplas. Do ponto de vista metodológico, apresenta-se um modelo inovador, flexível e com utilidade preditiva comprovada, cuja aplicação é ideal para grandes coortes. Conceitualmente, avança-se ao propor uma nova interpretação para o parâmetro de dispersão ϕ como um proxy estatístico da heterogeneidade biológica não observada, oferecendo assim uma perspectiva inovadora para a análise de sobrevivência.

CAPÍTULO

5

O MODELO DE FRAGILIDADE HZMGP: ABORDAGEM BAYESIANA E APLICAÇÕES

Neste Capítulo 5, desenvolve-se a abordagem Bayesiana para a inferência dos parâmetros do modelo de fragilidade HZMGP. A inferência Bayesiana fundamenta-se na combinação de duas fontes de informação: os dados, representados pela função de verossimilhança, e o conhecimento prévio sobre os parâmetros, formalizado por meio de distribuições a priori. A síntese dessas duas fontes de informação, governada pelo Teorema de Bayes, culmina na distribuição a posteriori, que representa a incerteza atualizada sobre os parâmetros do modelo (GELMAN *et al.*, 2013). Com o objetivo de detalhar essa construção, as seções subsequentes são dedicadas a apresentar cada um desses componentes: a função de verossimilhança, a especificação das distribuições a priori e a estratégia computacional adotada para a amostragem da distribuição a posteriori, cuja implementação completa para os estudos de simulação e as aplicações pode ser encontrada em <www.github.com/Katy-RCM/HZMGP>.

5.1 Estrutura do modelo Bayesiano HZMGP

A construção do modelo Bayesiano inicia-se com a função de verossimilhança, o componente que formaliza a contribuição da informação contida nos dados observados ($\mathscr{D}$). Adotando-se a mesma estrutura de modelo definida na Seção 4.1, tem-se que

$$T_i \mid \boldsymbol{x_i} \sim \text{HZMGP}(\mu_i, \phi, \omega_i, \lambda, \gamma),$$

em que os parâmetros individuais μ_i e ω_i estão vinculados às covariáveis por meio das funções de ligação logarítmica e logito, respectivamente, conforme descrito em (4.7). Consequentemente, a função de verossimilhança para o vetor de parâmetros $\boldsymbol{\vartheta} = \left(\boldsymbol{\beta}_{\boldsymbol{\mu}}^{\top}, \phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}^{\top}, \lambda, \gamma\right)^{\top}$ do modelo de

fragilidade HZMGP, fundamentada na expressão (4.8), é detalhada a seguir

$$\mathscr{L}(\mathscr{D} \mid \boldsymbol{\vartheta}) = \prod_{i=1}^{n} \left[1 - \frac{\omega_i}{1 - e^{-\frac{\mu_i}{1+\mu_i\phi}}} + \frac{\omega_i}{1 - e^{-\frac{\mu_i}{1+\mu_i\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0(t_i) e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right) + \frac{\mu_i\phi}{1+\mu_i\phi}\right]} \right]$$
$$\times \left[- \frac{\omega_i h_0(t_i)\, e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0(t_i) e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right) + \frac{\mu_i\phi}{1+\mu_i\phi}\right]}}{\left(1 - e^{-\frac{\mu_i}{1+\mu_i\phi}}\right) \phi S(t_i)} \frac{\mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0(t_i)\, e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right)}{1 + \mathbb{W}\left(-\frac{\mu_i\phi}{1+\mu_i\phi} S_0(t_i)\, e^{-\frac{\mu_i\phi}{1+\mu_i\phi}}\right)} \right]^{\delta_i}. \tag{5.1}$$

5.1.1 Especificação das distribuições a priori

A especificação das distribuições a priori é a etapa que completa a formulação do modelo Bayesiano. Para os parâmetros (ϕ, λ, γ), que são restritos ao domínio dos números reais positivos, optou-se por uma reparametrização logarítmica:

$$\phi = e^{\theta_\phi}, \quad \lambda = e^{\theta_\lambda}, \quad \gamma = e^{\theta_\gamma}.$$

Essa transformação oferece uma dupla vantagem: garante que as restrições de positividade sejam satisfeitas e melhora a performance dos algoritmos de amostragem, ao mapear o espaço paramétrico para o domínio irrestrito dos reais, onde a geometria da posterior tende a ser mais simétrica e numericamente estável para a computação (ALVARES *et al.*, 2024).

A inferência é, portanto, conduzida sobre o espaço dos parâmetros transformados, $\boldsymbol{\vartheta} = (\boldsymbol{\beta}_{\boldsymbol{\mu}}^\top, \theta_\phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}^\top, \theta_\lambda, \theta_\gamma)^\top$. Assumindo-se independência a priori entre os componentes de $\boldsymbol{\vartheta}$, a distribuição a priori conjunta pode ser fatorada como

$$\pi(\boldsymbol{\vartheta}) = \pi\left(\boldsymbol{\beta}_{\boldsymbol{\mu}}, \theta_\phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}, \theta_\lambda, \theta_\gamma\right) = \pi\left(\boldsymbol{\beta}_{\boldsymbol{\mu}}\right) \pi\left(\theta_\phi\right) \pi\left(\boldsymbol{\beta}_{\boldsymbol{\omega}}\right) \pi\left(\theta_\lambda\right) \pi\left(\theta_\gamma\right). \tag{5.2}$$

Seguindo as recomendações para a especificação de prioris fracamente informativas em modelos Bayesianos com parâmetros de escala positivos (GELMAN *et al.*, 2013), atribui-se uma distribuição a priori Normal com média zero e variância 10^2 a cada um dos componentes do vetor de parâmetros $\boldsymbol{\vartheta}$. A escolha de uma priori pouco informativa reflete um conhecimento prévio limitado, permitindo que a inferência seja predominantemente guiada pelos dados, ao mesmo tempo que penaliza valores extremos para os parâmetros, conferindo regularização ao modelo. Com isso, o modelo Bayesiano HZMGP completo pode ser expresso hierarquicamente da seguinte forma

$$\begin{aligned}
(t_i, \delta_i, \mathbf{x}_i \mid \mu_i, \phi, \omega_i, \lambda, \gamma) &\sim \text{HZMGP}(\mu_i, \phi, \omega_i, \lambda, \gamma), \\
\mu_i &= e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{\boldsymbol{\mu}}}, \\
\omega_i &= \frac{e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{\boldsymbol{\omega}}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\beta}_{\boldsymbol{\omega}}}}, \qquad i = 1, \ldots, n, \\
\boldsymbol{\beta}_{\boldsymbol{\mu}}, \theta_\phi, \boldsymbol{\beta}_{\boldsymbol{\omega}}\ \theta_\lambda,\ \theta_\gamma &\sim \mathcal{N}(0, 10^2).
\end{aligned} \tag{5.3}$$

A formulação hierárquica apresentada em (5.3) encapsula a estrutura completa do modelo de fragilidade Bayesiano HZMGP. Com a verossimilhança e as prioris completamente especificadas, o próximo passo consiste em obter a distribuição a posteriori, $\pi(\boldsymbol{\vartheta} \mid \mathscr{D})$, cuja obtenção será realizada por meio de métodos computacionais baseados em HMC (NEAL *et al.*, 2011).

5.1.2 Distribuição a posteriori e classificação

Com a função de verossimilhança definida em (5.1) e as distribuições a priori especificadas em (5.2), pode-se agora formular a distribuição a posteriori dos parâmetros. Seguindo o Teorema de Bayes, o kernel da distribuição a posteriori é dado pelo produto da verossimilhança pelas prioris, resultando na seguinte expressão

$$\pi\left(\boldsymbol{\vartheta} \mid \mathscr{D}\right) \propto \mathscr{L}\left(\mathscr{D} \mid \boldsymbol{\vartheta}\right) \pi\left(\boldsymbol{\beta}_{\boldsymbol{\mu}}\right) \pi\left(\theta_{\phi}\right) \pi\left(\boldsymbol{\beta}_{\boldsymbol{\omega}}\right) \pi\left(\theta_{\lambda}\right) \pi\left(\theta_{\gamma}\right). \tag{5.4}$$

A complexidade analítica da distribuição a posteriori, decorrente da forma funcional da verossimilhança, que envolve a função W de Lambert e múltiplas estruturas de regressão, impede a sua derivação em forma fechada. Essa intratabilidade torna imprescindível o uso de métodos computacionais para realizar a inferência. Para este fim, empregou-se a linguagem de programação probabilística Stan (CARPENTER *et al.*, 2017), uma escolha justificada por sua reconhecida eficiência na estimação de modelos hierárquicos complexos e por sua implementação de algoritmos de amostragem de última geração baseados em gradiente, como o HMC e o seu derivado adaptativo, o NUTS.

Uma vez estabelecida a estratégia computacional, avança-se para uma das contribuições metodológicas centrais desta tese: a proposição de uma regra de classificação prognóstica individualizada. Esta regra explora a rica estrutura de zeros do modelo de fragilidade HZMGP. A partir das amostras obtidas da distribuição a posteriori de $\boldsymbol{\vartheta}$, é possível derivar a distribuição a posteriori completa para os componentes individuais ω_i e μ_i para cada paciente.

A alocação de um paciente aos regimes de zero-inflação (ZIGP), zero-deflação (ZDGP) ou tradicional (GP) é governada pela relação entre ω_i e o limiar crítico $1 - e^{-\mu_i/(1+\mu_i\phi)}$ (ver Figura 3). A principal vantagem do paradigma Bayesiano, neste contexto, é permitir a quantificação da incerteza inerente a esta classificação. Assim, para atribuir cada paciente a um regime de zero-modificação com um grau de confiança pré-especificado, propõe-se a seguinte regra de decisão Bayesiana, baseada nos limites de probabilidade a posteriori $(1-\alpha)$

$$\begin{cases} \mathbb{P}\left(\omega_i < 1 - e^{-\frac{\mu_i}{1+\mu_i\phi}}\right) \geq 1-\alpha \text{ e } \mathbb{P}\left(\omega_i < 1\right) \geq 1-\alpha & \text{ZIGP,} \\ \alpha < \mathbb{P}\left(\omega_i < 1 - e^{-\frac{\mu_i}{1+\mu_i\phi}}\right) < 1-\alpha \text{ e } \mathbb{P}\left(\omega_i < 1\right) \geq 1-\alpha & \text{GP,} \\ \mathbb{P}\left(\omega_i > 1 - e^{-\frac{\mu_i}{1+\mu_i\phi}}\right) \geq 1-\alpha \text{ e } \mathbb{P}\left(\omega_i < 1\right) \geq 1-\alpha & \text{ZDGP,} \\ \mathbb{P}\left(\omega_i \geq 1\right) \geq 1-\alpha & \text{ZTGP.} \end{cases} \tag{5.5}$$

Dessa forma, a regra de decisão proposta eleva o modelo de fragilidade HZMGP de uma ferramenta descritiva a um instrumento de estratificação prognóstica. A força desta abordagem reside em alavancar o perfil a posteriori completo de cada paciente; ao fazê-lo, unifica-se a teoria do modelo com o poder da inferência Bayesiana, permitindo não apenas classificar cada indivíduo, mas também quantificar a certeza associada a essa classificação. Isso viabiliza a extração de conclusões com credibilidade formal, tanto em nível individual quanto populacional. Constitui-se, portanto, uma estrutura metodológica que não apenas modela com flexibilidade dados de sobrevivência complexos, mas também gera uma base quantitativa para a tomada de decisão.

A validação empírica e a utilidade desta abordagem de classificação serão rigorosamente examinadas nas seções subsequentes, por meio de um estudo de simulação abrangente e da aplicação a dados oncológicos.

5.2 Estudos de simulação

Para avaliar o desempenho da abordagem Bayesiana sob condições controladas, conduziu-se um extenso estudo de simulação. O objetivo principal consistiu em examinar a capacidade do modelo de recuperar com precisão os valores verdadeiros dos parâmetros sob diferentes cenários e tamanhos amostrais. Adicionalmente, avaliou-se a performance da regra de classificação prognóstica individualizada, considerando-se distintas especificações de zero-modificação.

A estrutura geral do modelo, com a incorporação de covariáveis nos parâmetros ω e μ, manteve-se. No entanto, para a abordagem Bayesiana, adotou-se uma parametrização alternativa para a distribuição Weibull basal. As funções de sobrevivência e risco basais foram, portanto, definidas como

A estrutura de regressão para os parâmetros μ e ω, por sua vez, permaneceu inalterada em relação à especificação frequentista detalhada em (4.7). No entanto, para a função de sobrevivência e risco basal, embora se tenha mantido a escolha da distribuição Weibull, adotou-se uma parametrização ligeiramente diferente, comumente empregada na literatura Bayesiana para melhorar a eficiência computacional. As funções de sobrevivência e risco basais foram, portanto, definidas como

$$\begin{aligned} S_0(t) &= e^{-\lambda(t^\gamma)}, \\ h_0(t) &= \gamma\lambda t^{(\gamma-1)}, \end{aligned} \qquad t>0, \lambda>0, \gamma>0. \tag{5.6}$$

Para a condução do estudo de simulação, foram elaborados dois cenários distintos, cada um projetado para avaliar o desempenho do modelo HZMGP sob um regime específico de zero-modificação (inflação e deflação). Em ambos os cenários, a estrutura de regressão incluiu duas covariáveis: uma contínua, gerada a partir da distribuição Normal, e uma binária, oriunda da distribuição de Bernoulli. Para cada combinação de cenário e tamanho amostral, $n \in$

$\{1000, 10000, 20000\}$, replicou-se o processo 100 vezes, gerando-se os tempos de sobrevivência a partir do modelo HZMGP, em conformidade com o procedimento detalhado no Algoritmo 1.

Para garantir que os cenários de simulação emulassem condições realistas, a especificação dos valores verdadeiros dos parâmetros foi diretamente informada pelos ajustes do modelo aos dados reais (detalhados na Seção 1.2), os quais incorporaram as covariáveis idade e cirurgia. Desta forma, as médias a posteriori obtidas em cada análise empírica foram adotadas como os valores de referência para os respectivos cenários de simulação, conforme especificado a seguir:

- **Cenário 1:** $\beta_{01} = 3,90$, $\beta_{11} = 0,13$, $\beta_{21} = -2,40$, $\phi = 0,25$, $\beta_{00} = 0,90$, $\beta_{10} = 0,23$, $\beta_{20} = -1,40$, $\lambda = 0,04$ e $\gamma = 1,30$

- **Cenário 2:** $\beta_{01} = 1,40$, $\beta_{11} = 0,10$, $\beta_{21} = 1,20$, $\phi = 0,13$, $\beta_{00} = 1,50$, $\beta_{10} = 0,60$, $\beta_{20} = 3,50$, $\lambda = 0,11$ e $\gamma = 1,08$.

Para cada um dos conjuntos de dados simulados, procedeu-se ao ajuste do modelo hierárquico definido em (5.3). A amostragem da distribuição a posteriori foi realizada por meio de um algoritmo HMC, executando-se três cadeias de Markov independentes, cada uma com 1000 iterações. As 300 iterações iniciais de cada cadeia foram descartadas como período de *burn-in*, totalizando 2100 amostras para a inferência de cada parâmetro. A Tabela 14 sintetiza o desempenho do modelo por meio das seguintes métricas de avaliação: a média a posteriori, o DP a posteriori, a PC dos intervalos de credibilidade de 95% e a probabilidade de acerto da classificação de zero-modificação, adotando-se um limiar de decisão de $\alpha = 0,1$. A análise consolidada desses resultados permite destacar os seguintes pontos:

- Em ambos os cenários, observa-se que as médias a posteriori convergem para os valores verdadeiros à medida que o tamanho amostral n aumenta, enquanto os DP a posteriori diminuem correspondentemente. Tal comportamento corrobora, respectivamente, a consistência e o ganho de precisão das estimativas.

- A PC de 95% situou-se, em geral, próxima do nível nominal. Contudo, uma exceção notável foi observada para o parâmetro γ no Cenário 2, cuja cobertura se mostrou consistentemente inferior à esperada. Este achado sugere que, para certas configurações de parâmetros, a distribuição a posteriori de γ pode exibir uma forma mais complexa, a qual não é plenamente capturada pelo intervalo de credibilidade baseado em quantis.

- A proporção da classificação de zero-modificação também melhora com o aumento do tamanho amostral. No Cenário I com $n = 20000$, a totalidade dos termos de fragilidade foi classificada como ZIGP, indicando um prognóstico favorável. No Cenário II, para o mesmo tamanho amostral, a proporção de termos ZIGP estabilizou-se em 29%, enquanto 2% foram classificados como ZDGP.

- A análise conjunta dos resultados evidencia que o desempenho geral do modelo, tanto em termos de precisão das estimativas quanto de acurácia da classificação, melhora substancialmente com amostras maiores ($n \geq 10000$). Essa necessidade de um grande volume de dados é acompanhada, contudo, por um custo computacional significativo. O tempo de ajuste para cada réplica, mesmo com a paralelização em 3 núcleos, foi de aproximadamente 20-30 minutos para $n = 1000$, aumentando para 2 a 3 horas para $n = 10000$ e atingindo de 4 a 5 horas para $n = 20000$. Este custo individual, ao ser multiplicado pelas 100 réplicas de cada cenário, sublinha a intensidade computacional da abordagem proposta. Fica, portanto, claro que a complexidade e a flexibilidade do modelo HZMGP impõem um compromisso prático entre a robustez da inferência e a viabilidade computacional, tornando sua aplicação mais indicada para contextos com grandes bases de dados e recursos computacionais adequados.

Tabela 14 – Resultados dos estudos de simulação Bayesiano para os dois cenários.

	Parâmetros	n = 1000		n = 10000		n = 20000	
		Média (DP)	PC	Média (DP)	PC	Média (DP)	PC
Primeiro cenário	$\beta_{01} = 3,90$	4,168 (2,244)	0,740	4,285 (0,349)	0,990	4,220 (0,329)	0,980
	$\beta_{11} = 0,13$	0,144 (0,198)	0,910	0,127 (0,037)	0,950	0,125 (0,024)	0,950
	$\beta_{21} = -2,40$	-5,385 (2,593)	0,900	-2,368 (0,173)	0,960	-2,361 (0,132)	0,910
	$\phi = 0,25$	0,371 (0,560)	0,950	0,214 (0,028)	0,990	0,217 (0,027)	0,990
	$\beta_{00} = 0,90$	0,931 (0,187)	0,960	0,901 (0,064)	0,950	0,905 (0,050)	0,930
	$\beta_{10} = 0,23$	0,234 (0,064)	0,950	0,229 (0,022)	0,980	0,229 (0,015)	0,960
	$\beta_{20} = -1,40$	-1,440 (0,209)	0,950	-1,397 (0,070)	0,960	-1,401 (0,053)	0,930
	$\lambda = 0,04$	0,095 (0,068)	0,800	0,033 (0,010)	0,990	0,033 (0,008)	0,990
	$\gamma = 1,30$	1,211 (0,143)	0,720	1,285 (0,030)	0,910	1,290 (0,021)	0,940
	Classificação	ZIGP	0,490	ZIGP	0,990	ZIGP	1
		GP	0,460	GP	0,010	GP	0
		ZDGP	0,050	ZDGP	0	ZDGP	0
		ZTGP	0	ZTGP	0	ZTGP	0
Segundo cenário	$\beta_{01} = 1,40$	1,719 (0,972)	1,000	1,623 (0,266)	0,990	1,612 (0,242)	0,940
	$\beta_{11} = 0,10$	0,118 (0,063)	0,960	0,107 (0,017)	0,950	0,105 (0,012)	0,950
	$\beta_{21} = 1,20$	1,665 (0,471)	0,900	1,246 (0,060)	0,930	1,235 (0,041)	0,910
	$\phi = 0,13$	0,116 (0,039)	0,990	0,130 (0,011)	0,980	0,131 (0,008)	0,970
	$\beta_{00} = 1,50$	1,536 (0,260)	0,930	1,516 (0,075)	0,930	1,511 (0,046)	0,960
	$\beta_{10} = 0,60$	0,595 (0,234)	0,920	0,597 (0,069)	0,930	0,603 (0,050)	0,940
	$\beta_{20} = 3,50$	4,441 (1,552)	0,930	3,525 (0,160)	0,970	3,525 (0,126)	0,930
	$\lambda = 0,11$	0,101 (0,042)	1,000	0,087 (0,019)	0,990	0,085 (0,018)	0,930
	$\gamma = 1,08$	1,118 (0,071)	0,930	1,099 (0,022)	0,790	1,097 (0,016)	0,790
	Classificação	ZIGP	0,030	ZIGP	0,150	ZIGP	0,290
		GP	0,970	GP	0,840	GP	0,690
		ZDGP	0	ZDGP	0,010	ZDGP	0,020
		ZTGP	0	ZTGP	0	ZTGP	0

Fonte: Elaborada pelo autor.

5.3 Aplicação em dados reais

Tendo sido apresentada a fundamentação teórica e a performance do modelo em cenários de simulação, esta seção dedica-se à sua aplicação prática. A abordagem Bayesiana para o modelo de fragilidade HZMGP é, assim, empregada na análise dos mesmos dois conjuntos de dados oncológicos, com o duplo objetivo de demonstrar sua versatilidade em cenários empíricos contrastantes e de destacar as vantagens da inferência probabilística em um contexto clínico real.

A metodologia adotada para ambas as aplicações espelha a do estudo de simulação, garantindo consistência. A inferência foi conduzida com base na formulação hierárquica completa (5.3), que incorpora a estrutura de regressão para os parâmetros ω e μ e a parametrização Weibull basal, conforme definida em (5.6).

A amostragem da distribuição a posteriori foi realizada no software Stan, mantendo-se a mesma configuração computacional adotada no estudo de simulação. Isto incluiu a execução de três cadeias de Markov paralelas (utilizando-se 3 núcleos), cada uma com 1000 iterações e 300 descartadas como período de *burn-in*, o que resultou em 2100 amostras para a inferência de cada parâmetro. A partir destas amostras, os resultados são sumarizados pela média, DP e ICr de 95% a posteriori. Como ponto central da análise, a regra de decisão Bayesiana (5.5) é empregada para realizar a classificação prognóstica individual, quantificando-se a incerteza associada. A validação gráfica do processo de amostragem, que assegura a robustez dos resultados, encontra-se detalhada no Apêndice C. Este apêndice inclui tanto os diagnósticos de convergência das cadeias quanto os histogramas de cada distribuição a posteriori de todos os parâmetros envolvidos.

5.3.1 Aplicação 1: Câncer de Melanoma

Procedeu-se ao ajuste do modelo de fragilidade HZMGP Bayesiano ao conjunto de dados de câncer de melanoma, com o propósito de quantificar o impacto conjunto das covariáveis na sobrevida dos pacientes. Os resultados da inferência a posteriori, detalhados na Tabela 15, permitem destacar os seguintes pontos principais:

- Observou-se que diversas covariáveis associadas a μ, ($\hat{\beta}_{11}$, $\hat{\beta}_{21}$, $\hat{\beta}_{31}$ e $\hat{\beta}_{71}$), não apresentaram um efeito estatisticamente credível, uma vez que seus respectivos ICr de 95% continham o valor zero. Especificamente, este foi o caso para as variáveis idade, sexo, EC II e radioterapia. Adicionalmente, a elevada amplitude do ICr para os estágios EC III e EC IV indica uma considerável incerteza sobre a magnitude de seus efeitos no risco.

- A estimativa da dispersão ($\hat{\phi}$), com média a posteriori de 0,189 (ICr 95%: [0,144; 0,239]), revelou-se estatisticamente relevante, pois seu ICr exclui o zero. Este resultado corrobora a presença de uma heterogeneidade não observada entre os pacientes, validando a pertinência da estrutura de fragilidade do modelo.

- Em contraste, todas as covariáveis incorporadas no parâmetro ω mostraram-se relevantes para modelar a fração de cura, dado que seus ICr de 95% não incluíram o zero. A título de exemplo, o modelo indica que pacientes em estágios mais avançados (EC IV) possuem um prognóstico desfavorável. Por outro lado, a intervenção cirúrgica demonstrou um efeito protetor, aumentando significativamente a probabilidade de um paciente ser considerado curado.

- A estimativa a posteriori do parâmetro de forma da Weibull, $\hat{\gamma} = 1,419$, possui um ICr de 95% ([1,344; 1,499]) que exclui o valor 1. Tal achado fornece forte evidência contra a suposição de um risco basal constante, indicando a inadequação de um modelo Exponencial como distribuição de base.

Tabela 15 – Média a posteriori, DP e ICr de 95% para os parâmetros do modelo de fragilidade HZMGP, ajustado aos dados de câncer de melanoma.

			ICr 95%	
Parâmetros	Média	DP	2.5%	97.5%
β_{01}	-3,879	3,155	-11,196	0,492
β_{11}(Idade)	-0,053	0,069	-0,184	0,089
β_{21}(Gen: F)	-0,230	0,124	-0,469	0,030
β_{31}(EC: II)	-2,209	7,017	-18,989	7,882
β_{41}(EC: III)	5,983	3,079	1,928	13,114
β_{51}(EC: IV)	7,982	3,097	3,822	15,147
β_{61}(Cir: S)	-1,180	0,138	-1,452	-0,918
β_{71}(Rad: S)	-0,199	0,141	-0,478	0,078
β_{81}(Qui: S)	-1,026	0,154	-1,322	-0,724
ϕ	0,189	0,024	0,144	0,239
β_{00}	-1,246	0,169	-1,576	-0,911
β_{10}(Idade)	0,340	0,045	0,246	0,423
β_{20}(Gen: F)	-0,522	0,088	-0,691	-0,349
β_{30}(EC: II)	1,426	0,117	1,195	1,652
β_{40}(EC: III)	2,070	0,124	1,823	2,308
β_{50}(EC: IV)	3,129	0,165	2,821	3,455
β_{60}(Cir: S)	-0,544	0,154	-0,846	-0,246
β_{70}(Rad: S)	1,213	0,201	0,828	1,630
β_{80}(Qui: S)	1,234	0,135	0,983	1,502
λ	0,085	0,011	0,062	0,104
γ	1,419	0,039	1,344	1,499

Fonte: Elaborada pelo autor.

A validade da inferência a posteriori foi assegurada pela verificação da convergência das cadeias de Markov, utilizando-se os diagnósticos da estatística $\hat{R}$ de Gelman-Rubin e do ESS. Em

geral, observou-se uma convergência satisfatória, com todos os parâmetros apresentando valores de $\hat{R} \approx 1$ e ESS ≥ 500, ao final de um processo de ajuste que demandou aproximadamente 3 horas e meia de tempo computacional.

Contudo, uma análise mais detalhada revelou exceções notáveis a esta convergência geral. A inspeção dos gráficos de traço, na Figura 17, mostrou que as cadeias associadas aos coeficientes dos estágios clínicos avançados (EC II, III e IV) no componente de risco μ exibiram um comportamento menos estável. Em contraste, todos os demais parâmetros do modelo, incluindo os associados a ω, ϕ, λ e γ, demonstraram uma convergência robusta. Esta instabilidade no processo de amostragem reflete-se diretamente na forma das distribuições a posteriori resultantes. Conforme se observa nos histogramas da Figura 18, as distribuições para os parâmetros associados principalmente aos estágios clínicos β_{01}, β_{31}, β_{41} e β_{51} apresentam um comportamento assimétrico e não se assemelham a uma distribuição Normal, enquanto os demais parâmetros exibem a esperada normalidade assintótica. Acredita-se que essa dificuldade localizada, tanto na convergência quanto na forma da posterior, possa indicar uma identificabilidade fraca ou multicolinearidade entre os efeitos desses estágios clínicos específicos, dado o número de eventos observados.

A análise culmina com a aplicação da regra de decisão Bayesiana (5.5) para a classificação prognóstica individual dos pacientes. Os resultados desta estratificação para a totalidade da coorte revelam que a grande maioria dos pacientes (83,20%) foi alocada ao regime de zero-inflação (ZIGP). Este achado é consistente com a análise exploratória de Kaplan-Meier e reforça a evidência de uma elevada fração de cura nesta população com câncer de melanoma. O subgrupo restante (16,80%) corresponde à ZDGP, portanto, a indivíduos com um perfil de risco mais acentuado, para os quais um acompanhamento clínico mais intensivo poderia ser indicado.

A eficácia e a granularidade desta classificação são ilustradas na Figura 7, que apresenta as distribuições a posteriori de ω e do limite crítico $1 - e^{(-\mu_i/(1+\mu_i\phi))}$ para quatro pacientes selecionados. A visualização por meio de gráficos de violino evidencia a capacidade do método em discernir perfis de risco distintos: os pacientes 1 e 2 foram classificados no regime ZDGP, enquanto os pacientes 11 e 15 foram classificados no regime ZIGP. Notavelmente, observa-se que as distribuições a posteriori do limite crítico (representado em rosa) exibem uma considerável assimetria, com caudas pronunciadas para os pacientes 1 e 2. Acredita-se que este comportamento possa ser uma consequência direta das dificuldades de convergência previamente identificadas para os coeficientes dos estágios clínicos no componente de risco (μ), cuja incerteza se propaga para a estimativa do limiar.

A análise da classificação individual revelou um achado de grande relevância clínica: certos perfis de pacientes que não receberam tratamentos adjuvantes apresentaram uma probabilidade de cura estimada superior à de pacientes que foram tratados. A explicação para este resultado, aparentemente contraintuitivo, reside na interpretação do parâmetro de dispersão ϕ, que reflete a heterogeneidade biológica não observada. Propõe-se que ϕ captura o impacto de

fatores intrínsecos ao tumor, como sua diversidade genética e molecular, que são determinantes prognósticos fundamentais.

Esta hipótese encontra forte respaldo na literatura oncológica contemporânea. Estudos multiômicos em melanoma têm consistentemente identificado subtipos moleculares com prognósticos intrinsecamente distintos, independentemente do tratamento administrado. Por exemplo, já foram caracterizados subgrupos prognósticos baseados em assinaturas de metilação (SUN *et al.*, 2020; MDPI, 2020) e, de forma crucial, no microambiente imune tumoral. Sabe-se que um alto infiltrado de células T CD8+ e perfis imunes específicos ("Immunity H") se correlacionam com uma melhor sobrevida. Em contrapartida, a desregulação de vias como a de p53 ou a piroptose associa-se a um pior prognóstico e a respostas diferenciais à imunoterapia (HAN *et al.*, 2020; JCANCER, 2021; INSIGHT, 2020; MEDICINE, 2023).

Em suma, estas evidências da biologia molecular validam a premissa de que a heterogeneidade intrínseca do tumor é um fator causal primário na sobrevida diferencial dos pacientes. O parâmetro ϕ do modelo HZMGP oferece, portanto, uma via estatística para capturar e quantificar o efeito agregado dessa complexidade biológica não medida, conferindo ao modelo um poder explicativo que transcende as covariáveis clínicas observadas.

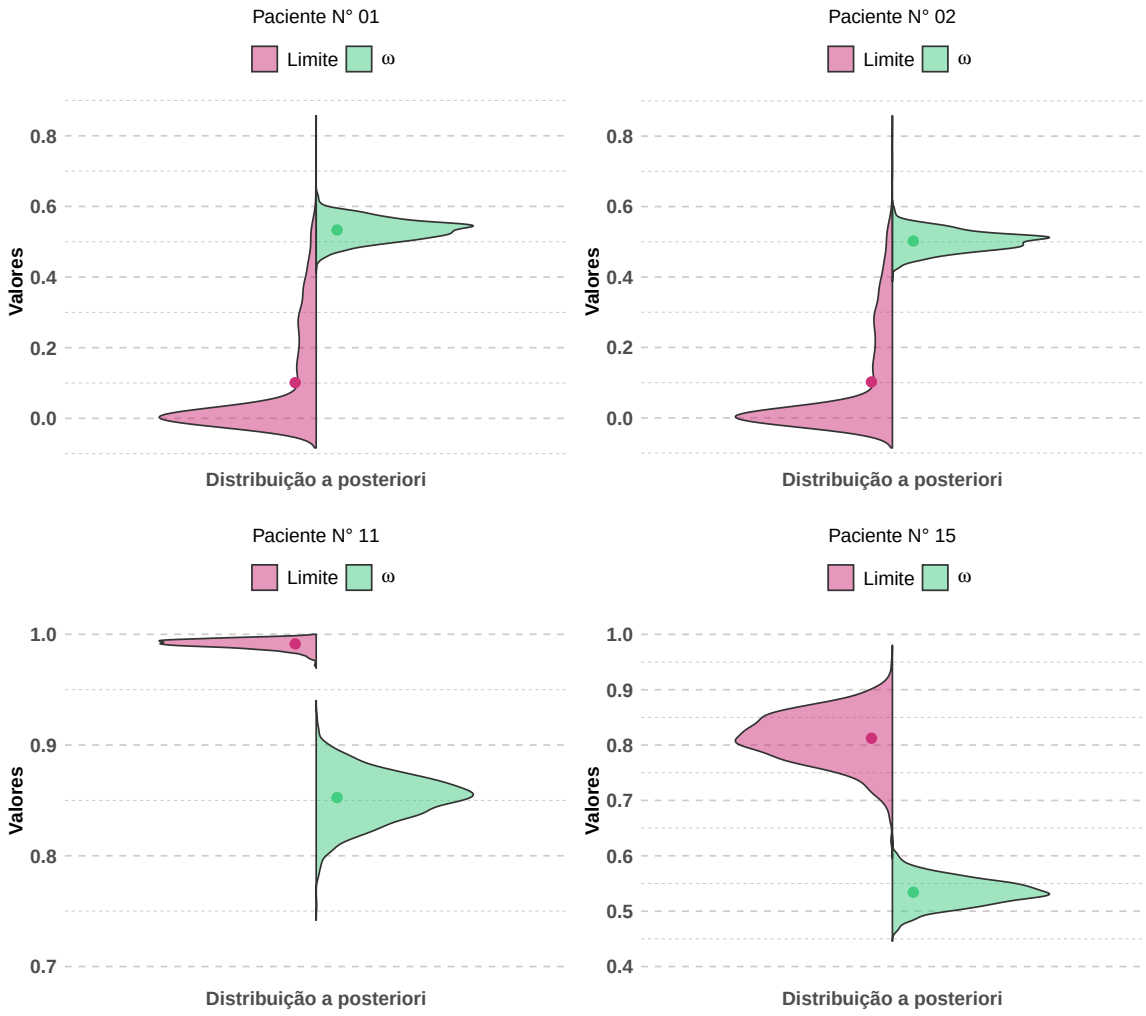

Figura 7 – Classificação da especificação de zero-modificação para os quatro primeiros pacientes nos dados do câncer de melanoma.

5.3.2 Aplicação 2: Câncer de pulmão

Procedeu-se à aplicação do modelo de fragilidade HZMGP Bayesiano ao conjunto de dados de câncer de pulmão. Esta análise teve como objetivo avaliar a influência das covariáveis na sobrevida dos pacientes neste contexto clínico de alta mortalidade. Os resultados da inferência a posteriori, sintetizados na Tabela 16, permitem extrair as seguintes conclusões:

- A análise dos coeficientes de regressão associados a μ e ω revelaram que todas as covariáveis incluídas no modelo apresentaram efeitos credíveis, uma vez que nenhum de seus ICr continham o zero. A estrutura do modelo permite uma interpretação detalhada desses efeitos em dois componentes distintos:
  - Componentes de μ: O modelo identificou que o sexo feminino ($\hat{\beta}_{21} = -0,267$), a cirurgia ($\hat{\beta}_{41} = -1,291$), a radioterapia ($\hat{\beta}_{51} = -0,658$) e a quimioterapia ($\hat{\beta}_{61} = -1,936$) atuam como fatores de proteção, reduzindo significativamente o risco de

óbito. Em contraste, o diagnóstico em estágio clínico avançado (EC III-IV, $\hat{\beta}_{31} = 1,666$) foi o principal fator de risco, associado a um aumento substancial na taxa de mortalidade.

  - Componentes de ω: Observou-se que o sexo feminino ($\hat{\beta}_{20} = -0,603$) e a realização de cirurgia ($\hat{\beta}_{40} = -1,964$) estão associados a uma maior probabilidade de que o paciente seja curado. Em contrapartida, idade avançada ($\hat{\beta}_{10} = 0,454$) e, notavelmente, o tratamento com quimioterapia ($\hat{\beta}_{60} = 0,744$), diminuiem a probabilidade de cura.

- A média a posteriori do parâmetro de dispersão, $\hat{\phi} = 0,140$ (ICr 95%: [0,134; 0,147]), corrobora a existência de uma significativa variabilidade não observada entre os pacientes.

- O ICr de 95% para o parâmetro de forma γ, com média a posteriori de 1,266, foi de [1,252; 1,280], o que descarta a hipótese de um risco basal constante e, consequentemente, a adequação de uma distribuição Weibull é coerente.

Tabela 16 – Média a posteriori, DP e ICr de 95% para os parâmetros do modelo de fragilidade HZMGP, ajustado aos dados de câncer de pulmão.

			ICr 95%	
Parâmetro	Média	DP	2.5%	97.5%
β_{01}	2,402	0,070	2,267	2,536
β_{11}(Idade)	-0,061	0,011	-0,082	-0,039
β_{21}(Gen: F)	-0,267	0,022	-0,312	-0,224
β_{31}(EC: III-IV)	1,666	0,054	1,563	1,780
β_{41}(Cir: S)	-1,291	0,043	-1,375	-1,207
β_{51}(Rad: S)	-0,658	0,024	-0,704	-0,611
β_{61}(Qui: S)	-1,936	0,032	-2,002	-1,874
ϕ	0,140	0,003	0,134	0,147
β_{00}	2,370	0,107	2,160	2,582
β_{10}(Idade)	0,454	0,039	0,379	0,528
β_{20}(Gen: F)	-0,603	0,078	-0,755	-0,450
β_{30}(EC: III-IV)	2,322	0,098	2,136	2,511
β_{40}(Cir: S)	-1,964	0,097	-2,149	-1,773
β_{50}(Rad: S)	0,361	0,108	0,153	0,570
β_{60}(Qui: S)	0,744	0,091	0,567	0,918
λ	0,187	0,007	0,174	0,201
γ	1,266	0,007	1,252	1,280

Fonte: Elaborada pelo autor.

A robustez da inferência Bayesiana para este conjunto de dados foi confirmada pela análise dos diagnósticos de convergência. Para todos os parâmetros do modelo, os valores de $\hat{R}$

de Gelman-Rubin se aproximaram de 1, com um ESS consistentemente superior a 900, resultado obtido após um tempo de ajuste de aproximadamente 3 horas.

Esta convergência robusta é corroborada visualmente. A inspeção dos gráficos de traço, apresentados na Figura 19, revela uma excelente mistura e estacionariedade para as cadeias de todos os parâmetros. Adicionalmente, e em contraste com a análise anterior, os histogramas da Figura 20 demonstram que todas as distribuições a posteriori exibem um comportamento aproximadamente Normal. Este duplo diagnóstico, convergência estável no processo de amostragem e regularidade na forma da posterior resultante, atesta a confiabilidade e a estabilidade da inferência gerada para esta aplicação específica.

A etapa conclusiva da análise para os dados de câncer de pulmão consiste na estratificação prognóstica individual. A aplicação da regra de decisão Bayesiana (5.5) revelou um perfil prognóstico marcadamente distinto do observado no melanoma. Verificou-se que a maioria dos pacientes (65%) foi alocada ao regime de zero-deflação (ZDGP), indicando um cenário de alta vulnerabilidade. Em contraste, 32% foram classificados no regime ZIGP, correspondendo a um subgrupo com prognóstico mais favorável, enquanto uma pequena fração (3%) se enquadrou no regime GP padrão.

A granularidade desta estratificação individual é visualmente demonstrada na Figura 8, que apresenta as distribuições a posteriori de ω e do limite crítico para quatro pacientes, evidenciando a alocação do paciente 1 ao regime ZDGP e dos pacientes 2, 3 e 4 ao ZIGP.

Este resultado para o câncer de pulmão, dominado pelo regime ZDGP, contrasta fortemente com a análise do melanoma e ressalta a notável flexibilidade do modelo HZMGP em se adaptar a cenários oncológicos com perfis de cura radicalmente diferentes. Enquanto na análise do melanoma o modelo capturou uma forte evidência de uma fração de cura substancial, aqui ele quantifica com precisão um cenário de alta mortalidade, característico deste tipo de câncer, identificando o subgrupo majoritário de alto risco que poderia se beneficiar de um acompanhamento clínico intensivo. Fica, assim, validada a utilidade do modelo não apenas para identificar esperança de cura, mas também para caracterizar e estratificar populações de alto risco com prognóstico reservado.

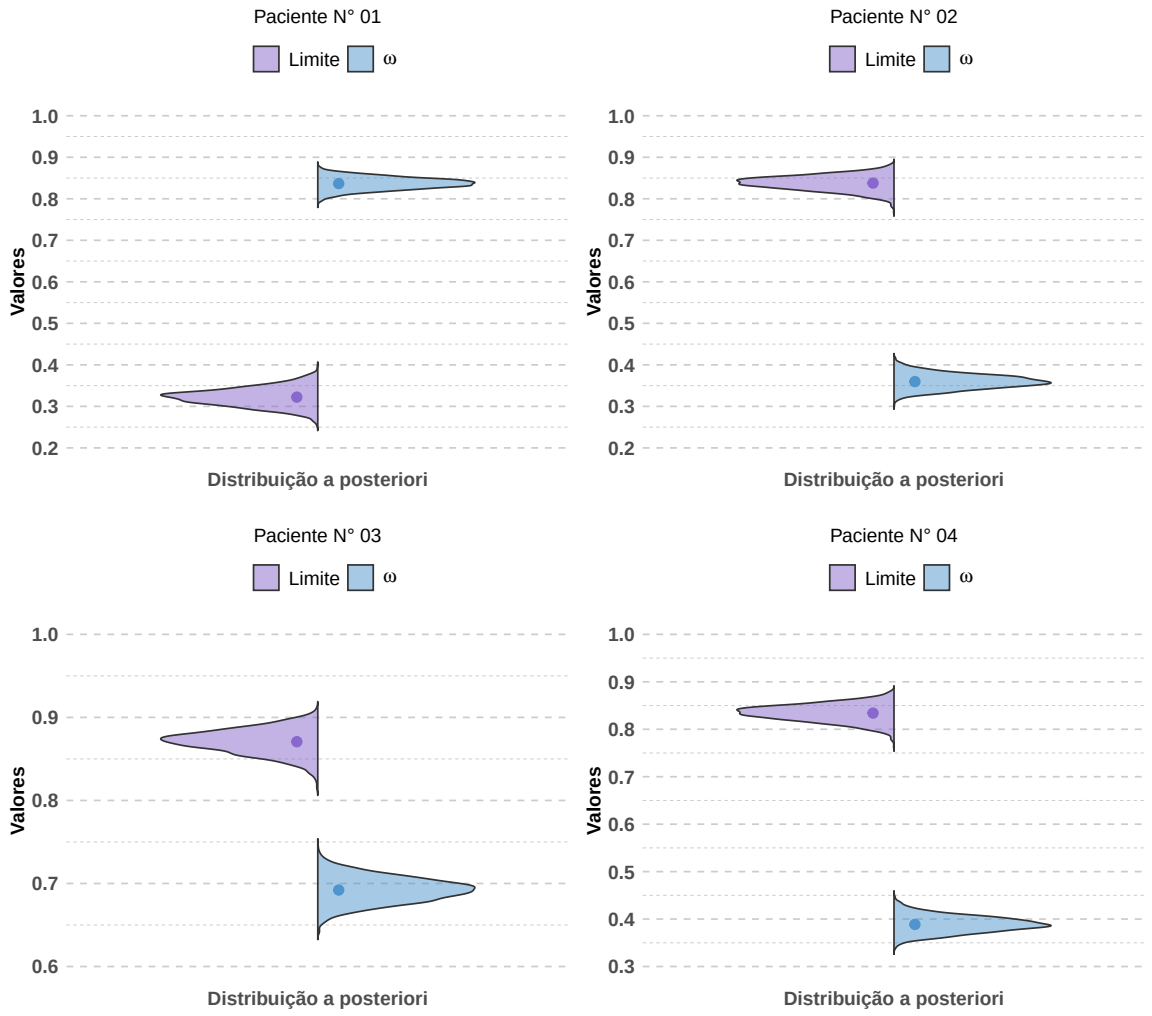

Figura 8 – Gráfico de violino para analisar a situação zero nos quatro primeiros pacientes com câncer de pulmão.

5.4 Síntese

Neste capítulo, desenvolveu-se a abordagem Bayesiana para o modelo de fragilidade HZMGP, com a inferência fundamentada na distribuição a posteriori, que sintetiza a informação da verossimilhança e das prioris. A principal contribuição metodológica foi a proposição de uma regra de classificação probabilística, que utiliza as distribuições a posteriori para estratificar cada paciente nos regimes de zero-inflação, zero-deflação ou tradicional, fornecendo uma poderosa ferramenta para a análise prognóstica individualizada.

A robustez da inferência e da classificação foi rigorosamente avaliada. Os estudos de simulação confirmaram as boas propriedades assintóticas dos estimadores, mas também elucidaram um ponto crítico: a complexidade do modelo HZMGP demanda amostras de tamanho substancial para garantir a estabilidade e a precisão da inferência. Esta conclusão foi fortemente corroborada na prática. A aplicação nos dados de melanoma, com um tamanho amostral menor, apresentou dificuldades de convergência para alguns parâmetros. Em contraste, na aplicação com

os dados de pulmão, uma amostra consideravelmente maior, a convergência foi uniformemente robusta.

Essa diferença no desempenho da convergência também se reflete no custo computacional. Contraintuitivamente, o ajuste para os dados de melanoma demandou um tempo computacional superior ao do pulmão, apesar de sua amostra menor. Acredita-se que este fenômeno ocorra porque a relativa escassez de informação nos dados de melanoma força o algoritmo de amostragem a explorar uma topologia da distribuição a posteriori mais complexa. Esta dificuldade é exacerbada pela alta dimensionalidade do espaço paramétrico, uma consequência da inclusão de múltiplas covariáveis nos componentes ω e μ. A combinação de uma a posteriori desafiadora com um grande número de parâmetros a serem estimados evidencia, portanto, que a estabilidade do ajuste, e não apenas o tamanho da amostra, é o fator determinante do tempo de processamento.

Finalmente, a análise empírica também validou decisões metodológicas cruciais para a flexibilidade do modelo. A inadequação de uma distribuição Exponencial e a subsequente adequação da Weibull como distribuição de base, observada em ambas as aplicações, ressaltam a importância de se escolher uma função de risco basal suficientemente flexível para não conflitar com a estimação dos demais componentes de um modelo tão parametrizado. Em suma, a abordagem Bayesiana não apenas validou a utilidade prática do HZMGP, mas também ofereceu *insights* profundos sobre suas condições de aplicabilidade e os requisitos para uma inferência robusta.

CAPÍTULO

6

CONCLUSÕES E PROPOSTAS FUTURAS

6.1 Conclusões

A análise de sobrevivência em oncologia frequentemente parte da premissa de homogeneidade populacional, uma simplificação que se distancia da realidade clínica, na qual cada paciente possui uma jornada única, influenciada por uma complexa rede de fatores observados e não observados. Com o intuito de abordar esta lacuna, esta tese teve como objetivo central o desenvolvimento, a validação e a aplicação de um novo modelo de fragilidade capaz de capturar com maior fidelidade a heterogeneidade intrínseca aos dados de sobrevivência.

A partir desta motivação, a contribuição teórica fundamental deste trabalho foi a proposição da família de distribuições HZMPS. Esta classe, derivada das famílias PS e ZMPS, oferece uma dupla fonte de flexibilidade: a capacidade de modelar diferentes padrões de zero (inflação, deflação ou truncamento) através do parâmetro ω, e de capturar a dispersão dos dados por meio do parâmetro ϕ. Para a construção do modelo de fragilidade central desta tese, selecionou-se desta família a distribuição HZMGP. A escolha desta distribuição específica se deve às suas características favoráveis para a análise de sobrevivência, sendo uma generalização de trabalhos anteriores e possuindo a flexibilidade necessária para acomodar cenários com fração de cura.

A robustez da metodologia foi rigorosamente investigada sob as perspectivas frequentista e Bayesiana, revelando *insights* consistentes e complementares. Uma das conclusões transversais mais importantes, emergindo tanto dos estudos de simulação quanto das aplicações empíricas, foi a de que a elevada flexibilidade do modelo HZMGP demanda um volume de dados substancial para uma inferência estável. Esta necessidade de amostras grandes foi inequivocamente validada na prática: a análise dos dados de câncer de pulmão resultou em uma convergência robusta, enquanto a aplicação nos dados de melanoma apresentou instabilidades. Este desempenho diferencial, por sua vez, revelou um *insights* profundo sobre o custo computacional, que se mostrou mais dependente da estabilidade do ajuste do que meramente do tamanho da amostra

– uma consequência da interação entre a alta dimensionalidade do espaço paramétrico e a topologia da distribuição a posteriori em cenários com informação limitada. Portanto, a principal implicação prática deste trabalho é o reconhecimento de que o modelo HZMGP, embora potente, possui um domínio de aplicabilidade ótimo em contextos de grandes bases de dados, onde sua complexidade pode ser plenamente suportada.

No âmbito das inovações, esta tese apresentou duas contribuições de destaque. A primeira, de natureza conceitual, foi a proposição de uma nova interpretação para o parâmetro de dispersão ϕ como um proxy estatístico para a heterogeneidade biológica não observada, criando uma ponte entre um parâmetro matemático e a complexidade biológica subjacente. A segunda, e principal contribuição metodológica, foi o desenvolvimento de uma regra de classificação prognóstica individual.

Esta regra de classificação, embora derivada formalmente no contexto Bayesiano pela quantificação da incerteza a posteriori, transcende o paradigma inferencial. A sua lógica de estratificação de pacientes nos regimes de zero-modificação é aplicável independentemente do enfoque, oferecendo uma ferramenta prática para a tomada de decisão clínica e para a sugestão de protocolos de acompanhamento diferenciados. A robustez e o poder deste método de classificação foram adicionalmente validados pela notável concordância observada entre as proporções de classificação obtidas nas abordagens frequentista e Bayesiana, confirmando a capacidade do modelo HZMGP de identificar de forma consistente os subgrupos de risco.

6.2 Propostas futuras

A presente tese, embora tenha atingido seus objetivos centrais, também revelou diversas e promissoras sugestões para a continuidade da pesquisa. A seguir, delineiam-se algumas propostas concretas para trabalhos futuros, agrupadas por áreas temáticas:

- Investigar o uso de métodos de estimação que possam aprimorar a estabilidade do modelo HZMGP em cenários com amostras de tamanho pequeno ou moderado. Isso inclui a exploração da máxima verossimilhança penalizada (regularização) e de abordagens Bayesianas com prioris mais informativas ou robustas, que poderiam conferir maior robustez ao processo de inferência.

- Explorar outras técnicas de inferência como Aproximações de Laplace Aninhadas Integradas (INLA)[1] para o modelo bayesiano HZMGP. A INLA poderia oferecer uma alternativa computacionalmente mais eficiente e potencialmente mais estável que o MCMC, tornando a análise de modelos complexos ou de grandes volumes de dados mais viável. Uma segunda etapa consistiria no desenvolvimento e na implementação destes métodos em pacotes de software de R, a fim de facilitar seu uso e adoção pela comunidade científica.

[1] Do inglês: *Integrated Nested Laplace Approximation*

- Investigar, teórica e empiricamente, o uso de outras distribuições da família HZMPS (como a HZMNB ou HZMGNB) como candidatas para a distribuição de fragilidade. Seria de grande interesse comparar o desempenho, a flexibilidade e a interpretabilidade desses novos modelos com o HZMGP em diferentes cenários de dados de sobrevivência.

- Aprofundar e validar empiricamente a interpretação do parâmetro ϕ. Isso poderia ser alcançado por meio do desenho de estudos que incorporem covariáveis biológicas, marcadores moleculares ou genéticos.

- Estender a aplicação do modelo de fragilidade HZMGP, e potencialmente de outros modelos da família HZMPS, a problemas em outros campos do conhecimento que lidam com dados de sobrevivência com características complexas, como ecologia, economia ou confiabilidade.

REFERÊNCIAS

AALEN, O. Nonparametric inference for a family of counting processes. **The Annals of Statistics**, JSTOR, p. 701–726, 1978. Citado na página 44.

ALVARES, D.; LÁZARO, E.; GÓMEZ-RUBIO, V.; ARMERO, C. Bayesian survival analysis with bugs. **Statistics in Medicine**, Wiley Online Library, v. 40, n. 12, p. 2975–3020, 2021. Citado na página 31.

ALVARES, D.; NIEKERK, J. V.; KRAINSKI, E. T.; RUE, H.; RUSTAND, D. Bayesian survival analysis with inla. **Statistics in medicine**, Wiley Online Library, v. 43, n. 20, p. 3975–4010, 2024. Citado na página 88.

AMBAGASPITIYA, R. S.; BALAKRISHNAN, N. On the compound generalized poisson distributions. **ASTIN Bulletin: the Journal of the IAA**, Cambridge University Press, v. 24, n. 2, p. 255–263, 1994. Citado na página 63.

ATA, N.; ÖZEL, G. Survival functions for the frailty models based on the discrete compound poisson process. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 83, n. 11, p. 2105–2116, 2013. Citado na página 30.

BALAKRISHNAN, N.; PENG, Y. Generalized gamma frailty model. **Statistics in medicine**, Wiley Online Library, v. 25, n. 16, p. 2797–2816, 2006. 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 na página 70.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 47, n. 259, p. 501–515, 1952. Citado na página 45.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 11, n. 1, p. 15–53, 1949. Citado na página 45.

BROOKS, S.; GELMAN, A.; JONES, G. L.; MENG, X.-L. **Handbook of Markov Chain Monte Carlo**. [S.l.]: CRC Press, 2011. Citado na página 50.

CALSAVARA, V. F.; MILANI, E. A.; BERTOLLI, E.; TOMAZELLA, V. Long-term frailty modeling using a non-proportional hazards model: Application with a melanoma dataset. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, v. 29, n. 8, p. 2100–2118, 2020. Citado nas páginas 30 e 34.

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 na página 30.

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, v. 75, n. 3, p. 299–323, 2021. Citado na página 31.

CANCHO, V. G.; MACERA, M. A.; SUZUKI, A. K.; LOUZADA, F.; ZAVALETA, K. E. A new long-term survival model with dispersion induced by discrete frailty. **Lifetime Data Analysis**, Springer, v. 26, n. 2, p. 221–244, 2020. Citado nas páginas 30, 71 e 86.

CANCHO, V. G.; SACRAMENTO, M. M.; ORTEGA, E. M.; MORAES, T. E. de; CORDEIRO, G. M. A bayesian cure rate regression model using hamiltonian monte carlo methods. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, p. 1–18, 2024. Citado na página 31.

CARONI, C.; CROWDER, M.; KIMBER, A. Proportional hazards models with discrete frailty. **Lifetime data analysis**, Springer, v. 16, n. 3, p. 374–384, 2010. Citado nas páginas 30 e 69.

CARPENTER, B.; GELMAN, A.; HOFFMAN, M. D.; LEE, D.; GOODRICH, B.; BETANCOURT, M.; BRUBAKER, M.; GUO, J.; LI, P.; RIDDELL, A. Stan: A probabilistic programming language. **Journal of statistical software**, v. 76, p. 1–32, 2017. Citado nas páginas 51 e 89.

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 30.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2021. Citado nas páginas 41 e 45.

CONCEIÇÃO, K.; LOUZADA, F.; ANDRADE, M. G.; HELOU, E. Zero-modified power series distribution and its hurdle distribution version. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 9, p. 1842–1862, 2017. Citado na página 57.

CONCEIÇÃO, K. S. **Modelos série de potência zero-modificados**. Tese (Doutorado) — Universidade Federal de São Carlos, 2013. Citado nas páginas 53, 54, 55 e 57.

CONCEIÇÃO, K. S.; LOUZADA, F.; ANDRADE, M.; HELOU, E. Zero-modified power series distribution and its Hurdle distribution version. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 9, p. 1842–1862, 2017. Citado nas páginas 53, 59 e 60.

CONCEIÇÃO, K. S.; TOMAZELLA, V.; ANDRADE, M. G.; LOUZADA, F. Biparametric zero-modified power series distributions: Bayesian analysis under a reference prior approach. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 46, n. 21, p. 10518–10536, 2017. Citado nas páginas 53 e 60.

CONSUL, P.; JAIN, G. On some interesting properties of the generalized poisson distribution. **Biometrische Zeitschrift**, Wiley Online Library, v. 15, n. 7, p. 495–500, 1973. Citado na página 53.

CONSUL, P. C.; FAMOYE, F. **Lagrangian probability distributions**. [S.l.]: Springer, 2006. Citado na página 57.

CORDEIRO, G. M.; ANDRADE, M. G.; CASTRO, M. de. Power series generalized nonlinear models. **Computational statistics & data analysis**, Elsevier, v. 53, n. 4, p. 1155–1166, 2009. Citado na página 54.

COX, D. R. Regression models and life-tables. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 2, p. 187–202, 1972. Citado nas páginas 29 e 44.

____. Partial likelihood. **Biometrika**, Oxford University Press, v. 62, n. 2, p. 269–276, 1975. Citado na página 45.

COX, D. R.; OAKES, D. **Analysis of Survival Data**. London: Chapman and Hall, 1984. Citado na página 48.

DABADE, A. Compound negative binomial shared frailty model with random probability of susceptibility. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 94, n. 4, p. 843–861, 2024. Citado na página 31.

DAVIS, L. E.; SHALIN, S. C.; TACKETT, A. J. Current state of melanoma diagnosis and treatment. **Cancer Biology & Therapy**, v. 20, n. 11, p. 1366–1379, 2019. Citado na página 34.

FLINN, C. J.; HECKMAN, J. J. New methods for analyzing individual event histories. **Sociological methodology**, JSTOR, v. 13, p. 99–140, 1982. Citado na página 48.

GAZON, A. B.; MILANI, E. A.; MOTA, A. L.; LOUZADA, F.; TOMAZELLA, V. L.; CALSAVARA, V. F. Nonproportional hazards model with a frailty term for modeling subgroups with evidence of long-term survivors: Application to a lung cancer dataset. **Biometrical Journal**, Wiley Online Library, v. 64, n. 1, p. 105–130, 2022. Citado nas páginas 30 e 37.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian Data Analysis**. 3rd. ed. Boca Raton, FL: CRC Press, 2013. ISBN 9781439840955. Citado nas páginas 87 e 88.

GERSHENWALD, J. E.; SCOLYER, R. A.; HESS, K. R.; SONDAK, V. K.; LONG, G. V.; ROSS, M. I.; LAZAR, A. J.; FARIES, M. B.; KIRKWOOD, J. M.; MCARTHUR, G. A. *et al.* Melanoma staging: Evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual. **CA: A Cancer Journal for Clinicians**, v. 67, p. 472–492, 2017. Citado na página 34.

GUPTA, R. C. Modified power series distribution and some of its applications. **Sankhyā: The Indian Journal of Statistics, Series B**, JSTOR, p. 288–298, 1974. Citado na página 54.

HAN, H.; ZHANG, H.; KIM, S.; LEE, S. Immune subtyping of melanoma using rna-seq: identifying subtypes with distinct prognoses based on immune infiltrates. **Journal of Immunology**, v. 205, p. 1234–1245, 2020. Disponível em: <https://www.jimmunol.org/content/early/2020/10/13/jimmunol.2000394>. Citado na página 96.

HANIN, L.; HUANG, L.-S. Identifiability of cure models revisited. **Journal of Multivariate Analysis**, v. 130, p. 261–274, 2014. ISSN 0047-259X. Citado na página 70.

HOFFMAN, M. D.; GELMAN, A. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. **Journal of Machine Learning Research**, v. 15, n. 1, p. 1593–1623, 2014. Citado na página 50.

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 na página 30.

IACHINE, I. Identifiability of bivariate frailty models. **Preprint**, Citeseer, v. 5, 2004. Citado na página 49.

IBRAHIM, J. G.; CHEN, M.-H.; SINHA, D. **Bayesian Survival Analysis**. [S.l.]: Springer, 2001. Citado nas páginas 30 e 31.

INSIGHT, J. Epigenetic signatures and p53 mutations in melanoma: defining high-risk subtypes. **JCI Insight**, v. 5, p. e141543, 2020. Disponível em: <https://insight.jci.org/articles/view/141543>. Citado na página 96.

INSTITUTE, N. C. **Melanoma Treatment (PDQ)–Health Professional Version**. 2024. <https://www.cancer.gov/types/skin/hp/melanoma-treatment-pdq>. Citado na página 34.

JCANCER. Infiltration of cd8+ t cells and its association with survival in melanoma: A multi-omics approach. **Journal of Cancer**, v. 12, p. 703–711, 2021. Disponível em: <https://www.jcancer.org/v12p0703.htm>. Citado na página 96.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. **Univariate discrete distributions**. [S.l.]: John Wiley & Sons, 2005. v. 444. Citado na página 55.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. **Continuous univariate distributions, volume 2**. [S.l.]: John wiley & sons, 1995. v. 289. Citado na página 55.

KAMALJA, K. K.; WAGH, Y. S. Estimation in zero-inflated generalized poisson distribution. **Journal of Data Science**, v. 16, n. 1, p. 183–206, 2018. Citado na página 57.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958. Citado nas páginas 29 e 44.

KLEINBAUM, D. G.; KLEIN, M. **Survival Analysis: A Self-Learning Text**. 3rd. ed. New York, NY: Springer, 2012. Citado na página 41.

KURNIA, A.; SADIK, K. *et al.* Analysis of overdispersed count data by poisson model. **European Journal of Molecular & Clinical Medicine**, v. 7, n. 10, p. 1400–1409, 2021. Citado na página 57.

LANCASTER, T. **The econometric analysis of transition data**. [S.l.]: Cambridge university press, 1990. Citado na página 48.

LEÃO, J.; LEIVA, V.; SAULO, H.; TOMAZELLA, V. Birnbaum–saunders frailty regression models: Diagnostics and application to medical data. **Biometrical Journal**, Wiley Online Library, v. 59, n. 2, p. 291–314, 2017. Citado na página 30.

LEÃO, J.; LEIVA, V.; SAULO, H.; TOMAZELLA, V. A survival model with birnbaum–saunders frailty for uncensored and censored cancer data. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 32, n. 4, p. 707–729, 2018. Citado na página 30.

LI, C.-S.; TAYLOR, J. M.; SY, J. P. Identifiability of cure models. **Statistics & Probability Letters**, v. 54, n. 4, p. 389–395, 2001. ISSN 0167-7152. Citado na página 70.

MALLER, R. A.; ZHOU, X. **Survival analysis with long-term survivors**. [S.l.]: John Wiley & Sons, 1996. Citado na página 30.

MCCALL, B. P. Testing the proportional hazards assumption in the presence of unmeasured heterogeneity. **Journal of Applied Econometrics**, Wiley Online Library, v. 9, n. 3, p. 321–334, 1994. Citado na página 49.

MCGILCHRIST, C.; AISBETT, C. Regression with frailty in survival analysis. **Biometrics**, JSTOR, p. 461–466, 1991. Citado na página 30.

MDPI. jsvd-based multi-omics analysis of melanoma: Rna-seq and methylation profiling identify distinct subtypes of melanoma cutaneous tumors. **MDPI**, MDPI, 2020. URL: <https://www.mdpi.com/2227-9059/10/12/3240>. Citado na página 96.

MEDICINE, B. Proteomic and transcriptomic profiling in metastatic melanoma: Identifying subgroups linked to pyroptosis and immune response. **BMC Medicine**, v. 21, p. 1–10, 2023. Disponível em: <https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-023-03175-0>. Citado na página 96.

MOLINA, K. C.; CALSAVARA, V. F.; TOMAZELLA, V. D.; MILANI, E. A. Survival models induced by zero-modified power series discrete frailty: Application with a melanoma data set. **Statistical Methods in Medical Research**, SAGE Publications Sage UK: London, England, v. 30, n. 8, p. 1874–1889, 2021. Citado nas páginas 31 e 86.

National Cancer Institute. **What Is Cancer?** 2021. <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>. Citado na página 32.

NEAL, R. M. *et al.* Mcmc using hamiltonian dynamics. **Handbook of markov chain monte carlo**, Chapman and Hall/CRC, v. 2, n. 11, p. 2, 2011. Citado na página 89.

NOACK, A. A class of random variables with discrete distributions. **The Annals of Mathematical Statistics**, JSTOR, v. 21, n. 1, p. 127–132, 1950. Citado na página 53.

ONCHERE, W. O. **Frailty Models Applications In Pension Schemes**. Tese (Doutorado) — University of Nairobi, 2013. Citado na página 30.

ORGANIZATION, W. H. **Skin cancers**. 2024. <https://www.who.int/news-room/fact-sheets/detail/ultraviolet-(uv)-radiation-and-skin-cancer>. Citado na página 33.

PHILIP, H. Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, Oxford University Press, v. 73, n. 2, p. 387–396, 1986. Citado na página 30.

PICKLES, A.; CROUCHLEY, R.; SIMONOFF, E.; EAVES, L.; MEYER, J.; RUTTER, M.; HEWITT, J.; SILBERG, J. Survival models for developmental genetic data: age of onset of puberty and antisocial behavior in twins. **Genetic Epidemiology**, Wiley Online Library, v. 11, n. 2, p. 155–170, 1994. Citado na página 48.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2025. Disponível em: <https://www.R-project.org/>. Citado na página 75.

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 na página 31.

RODRIGUES, J.; CANCHO, V. G.; BALAKRISHNAN, N.; SUZUKI, A. K. A bayesian destructive generalized waring regression cure model with a variance decomposition and application in colorectal cancer data. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 94, n. 14, p. 3111–3130, 2024. 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 nas páginas 46 e 70.

SANTO, A. P. J. do E.; CANCHO, V. G.; LOUZADA, F.; ORTEGA, E. M. A survival model for lifetime with long-term survivors and unobserved heterogeneity. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 36, n. 4, p. 692–703, 2022. Citado na página 31.

SANTOS, D. d. S.; CANCHO, V.; RODRIGUES, J. Hypothesis testing for the dispersion parameter of the hyper-poisson regression model. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 89, n. 5, p. 763–775, 2019. Citado na página 31.

SANTOS, D. M. dos; DAVIES, R. B.; FRANCIS, B. Nonparametric hazard versus nonparametric frailty distribution in modelling recurrence of breast cancer. **Journal of statistical planning and inference**, Elsevier, v. 47, n. 1-2, p. 111–127, 1995. Citado na página 30.

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 na página 30.

SIEGEL, R. L.; MILLER, K. D.; JEMAL, A. Cancer statistics, 2020. **CA: A Cancer Journal for Clinicians**, v. 70, n. 1, p. 7–30, 2020. Citado na página 33.

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 30 e 31.

SUN, X.; LI, X.; WANG, Y.; CHEN, Q. Multi-omics analysis of melanoma: Integrating rna-seq, methylation, and mutation data to reveal subtypes of melanoma cutaneous tumors. **Frontiers in Molecular Biosciences**, v. 7, p. 598725, 2020. Disponível em: <https://www.frontiersin.org/articles/10.3389/fmolb.2020.598725/full>. Citado na página 96.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. **CA: A Cancer Journal for Clinicians**, v. 71, n. 3, p. 209–249, 2021. Citado nas páginas 33 e 36.

TAO, M.-H. Epidemiology of lung cancer. **Lung Cancer and Imaging**, IOP Publishing, 2019. Citado na página 36.

THAI, A.; SOLOMON, B.; SEQUIST, L. V.; GAINOR, J. F.; HEIST, R. S. Lung cancer. **The Lancet**, v. 398, n. 10299, p. 535–554, 2021. Citado na página 36.

TOMAZELLA, V. L. D.; MARTINS, C. B.; BERNARDO, J. M. Inference on the univariate frailty model: A bayesian reference analysis approach. In: AMERICAN INSTITUTE OF PHYSICS. **AIP Conference Proceedings**. [S.l.], 2008. v. 1073, n. 1, p. 340–347. 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 29, 30 e 48.

VIGILÂNCIA Coordenação de Prevenção e. **Instituto Nacional de Câncer José Alencar Gomes da Silva. Estimativa 2018: Incidência de câncer no Brasil**. Rio de Janeiro, 2018. Citado nas páginas 34, 36 e 37.

WANG, C.; LI, J.; CHEN, J.; WANG, Z.; ZHU, G.; SONG, L.; WU, J.; LI, C.; QIU, R.; CHEN, X. *et al.* Multi-omics analyses reveal biological and clinical insights in recurrent stage i non-small cell lung cancer. **Nature Communications**, Nature Publishing Group UK London, v. 16, n. 1, p. 1477, 2025. Citado na página 84.

WIENKE, A. **Frailty models in survival analysis**. [S.l.]: Chapman and Hall/CRC, 2010. Citado nas páginas 30, 47, 48 e 49.

World Health Organization. **Cancer**. 2024. Fact sheet. <https://www.who.int/news-room/fact-sheets/detail/cancer>. Acessado em [Insira a data de acesso]. Citado na página 33.

ZHANG, Y.; WANG, Y.; QIAN, H. Multi-omics characterization and machine learning of lung adenocarcinoma molecular subtypes to guide precise chemotherapy and immunotherapy. **Frontiers in Immunology**, Frontiers Media SA, v. 15, p. 1497300, 2024. Citado na página 84.

APÊNDICE

A

CÁLCULOS DO MODELO HZMGP

Este apêndice detalha o procedimento matemático das funções de risco, $h(t)$, e de densidade de probabilidade, $f(t)$, para o modelo de fragilidade HZMGP. Partindo da função de sobrevivência, $S(t)$, cuja expressão final foi apresentada em (4.2), demonstram-se os passos para obter as formas explícitas de $h(t)$ e $f(t)$.

A função de sobrevivência é dada por:

$$S(t) = 1 - \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} + \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right]}. \quad \text{(A.1)}$$

A derivação da função de risco, $h(t)$, parte da sua relação fundamental com a função de sobrevivência. Conforme definido em (2.1), tem-se que: $h(t) = -\dfrac{S'(t)}{S(t)}$.

Derivando-se a função de sobrevivência $S(t)$ em relação a t, obtém-se:

$$S'(t) = \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} \left\{ e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right]} \right\}'$$

$$= \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right]}$$

$$\times \left\{ -\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi} S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right) + \frac{\mu\phi}{1+\mu\phi}\right] \right\}'$$

$$= \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}$$
$$\times\left(-\frac{1}{\phi}\right)\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)\right]'$$

$$= \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[W\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}\left(-\frac{1}{\phi}\right)$$
$$\times\frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)\left[1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)\right]}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)'$$

$$= \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}\left(-\frac{1}{\phi}\right)$$
$$\times\frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}\left(-\frac{1+\mu\phi}{\mu\phi S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}}\right)\left(-\frac{\mu\phi}{1+\mu\phi}\left(S_0'(t)\right)e^{-\frac{\mu\phi}{1+\mu\phi}}\right)$$

$$= \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}\left(\frac{1}{\phi}\right)$$
$$\times\frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}\left(-\frac{S_0'(t)}{S_0(t)}\right)$$

$$S'(t) = \left(\frac{\omega}{\phi\left(1-e^{-\frac{\mu}{1+\mu\phi}}\right)}\right)h_0(t)\,e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}$$
$$\times\frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\,e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}. \tag{A.2}$$

Substituindo-se (A.2) na definição da função de risco, $h(t) = -S'(t)/S(t)$, obtém-se:

$$h(t) = -\frac{\omega h_0(t)\, e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}}{\left(1-e^{-\frac{\mu}{1+\mu\phi}}\right)\phi S(t)} \frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}. \tag{A.3}$$

Com as expressões de (A.1) e (A.3) já definidas, a função de densidade de probabilidade, $f(t)$, é encontrada multiplicando-se ambas, como se segue:

$$\begin{aligned} f(t) &= 1 - \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} + \frac{\omega}{1-e^{-\frac{\mu}{1+\mu\phi}}} e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t) e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]} \\ &\quad \times -\frac{\omega h_0(t)\, e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}}{\left(1-e^{-\frac{\mu}{1+\mu\phi}}\right)\phi S(t)} \frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)} \\ &= -\frac{\omega h_0(t)\, e^{-\frac{1}{\phi}\left[\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)+\frac{\mu\phi}{1+\mu\phi}\right]}}{\left(1-e^{-\frac{\mu}{1+\mu\phi}}\right)\phi} \frac{\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}{1+\mathbb{W}\left(-\frac{\mu\phi}{1+\mu\phi}S_0(t)\, e^{-\frac{\mu\phi}{1+\mu\phi}}\right)}. \end{aligned} \tag{A.4}$$

APÊNDICE

B

GRÁFICOS DO ESTUDO DE SIMULAÇÃO CLÁSSICO

Este apêndice apresenta os resultados gráficos completos do estudo de simulação clássico, detalhados por cenário. Para cada cenário, são exibidos os gráficos de convergência da MEMV, o comportamento do DP e da REQM, e a análise da PC.

B.1 Cenário 1: modelo de fragilidade ZIGP

A seguir, nas Figuras 9, 10, 11 e 12, são apresentados os resultados gráficos referentes ao Cenário 1.

B.2 Cenário 2: modelo de fragilidade ZDGP

Os gráficos a seguir, nas Figuras 13, 14, 15 e 16, correspondem aos resultados obtidos para o Cenário 2.

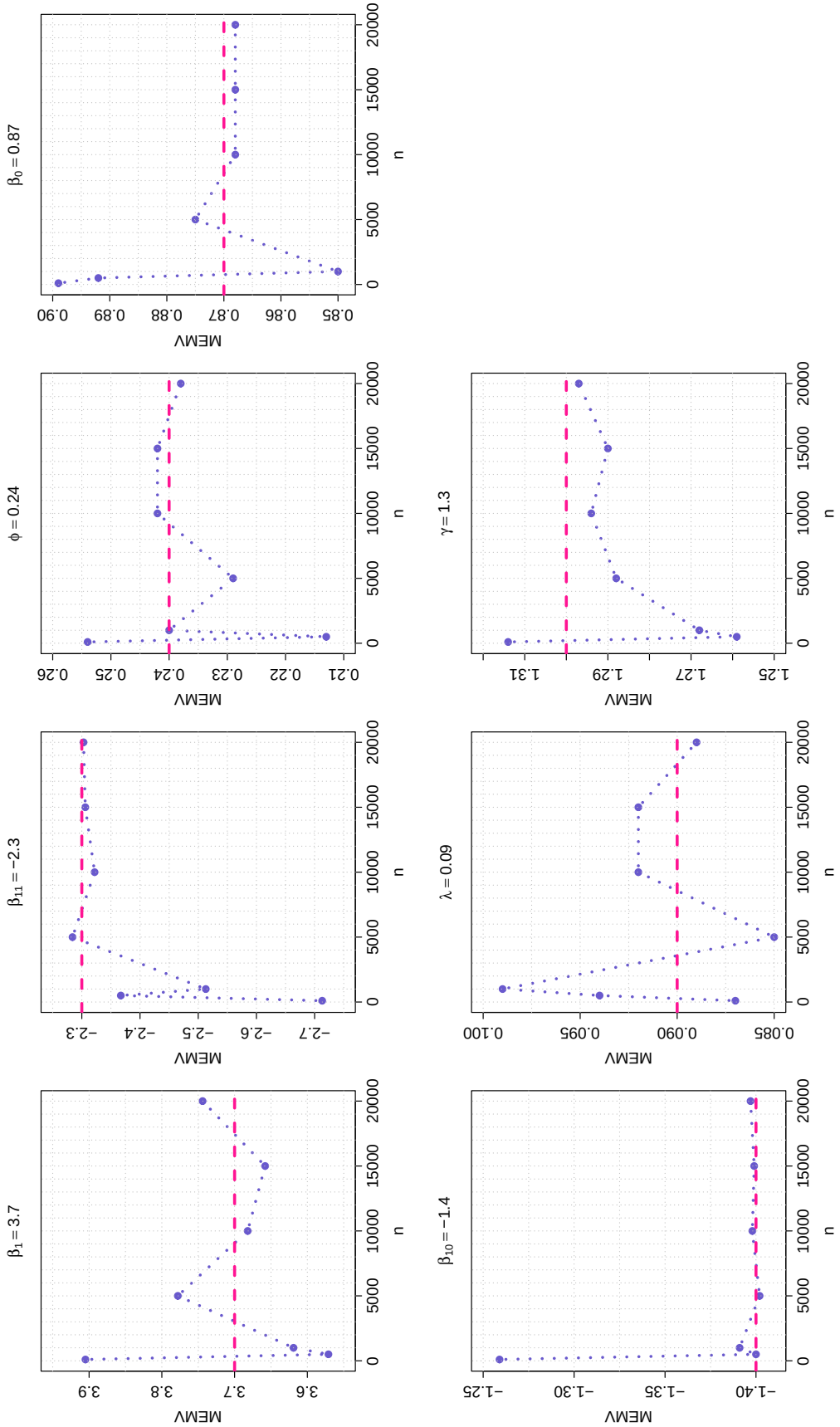

Figura 9 – Convergência da MEMV para os valores verdadeiros (linhas tracejadas cor fúcsia) dos parâmetros do modelo de fragilidade ZIGP, em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

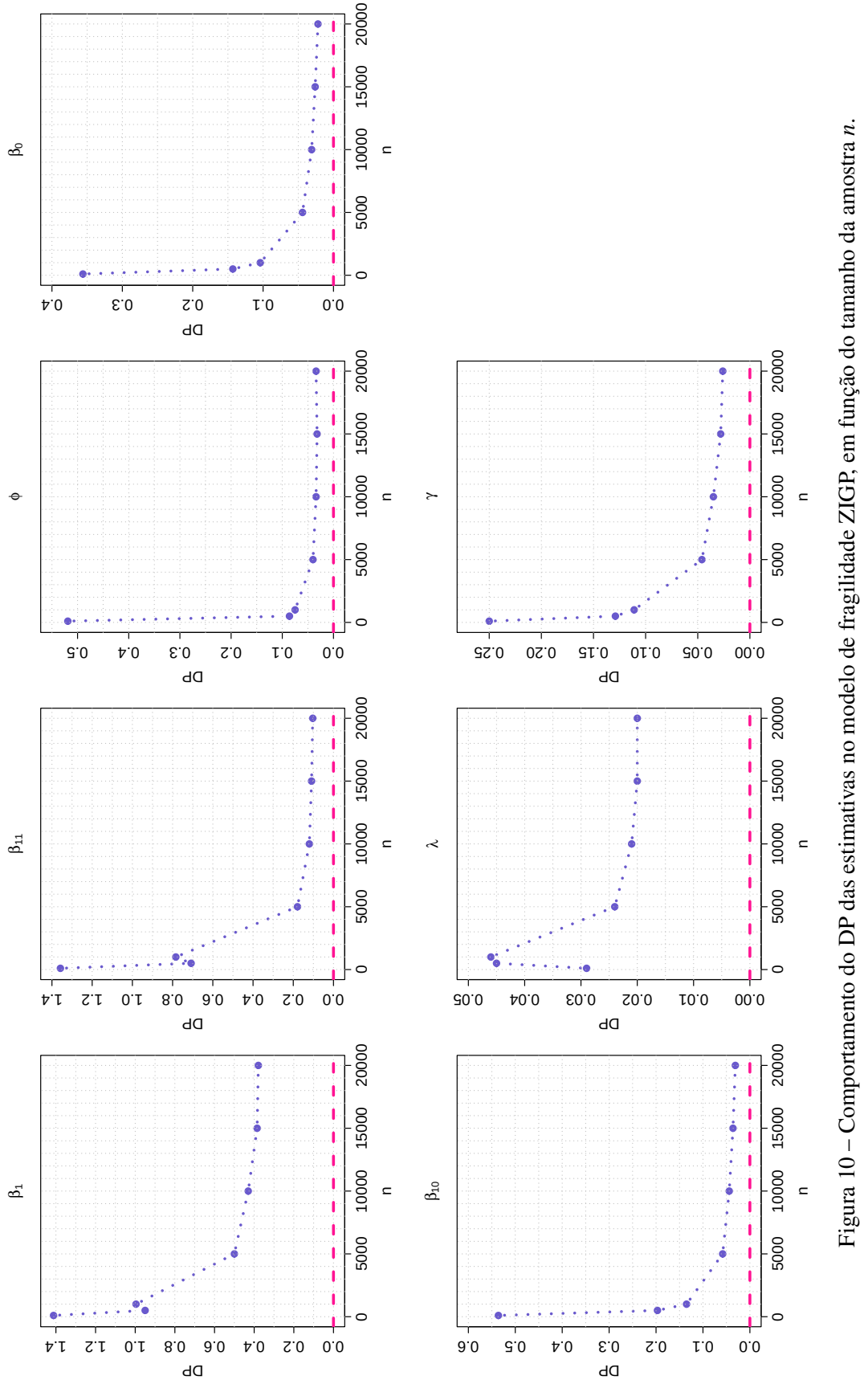

Figura 10 – Comportamento do DP das estimativas no modelo de fragilidade ZIGP, em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

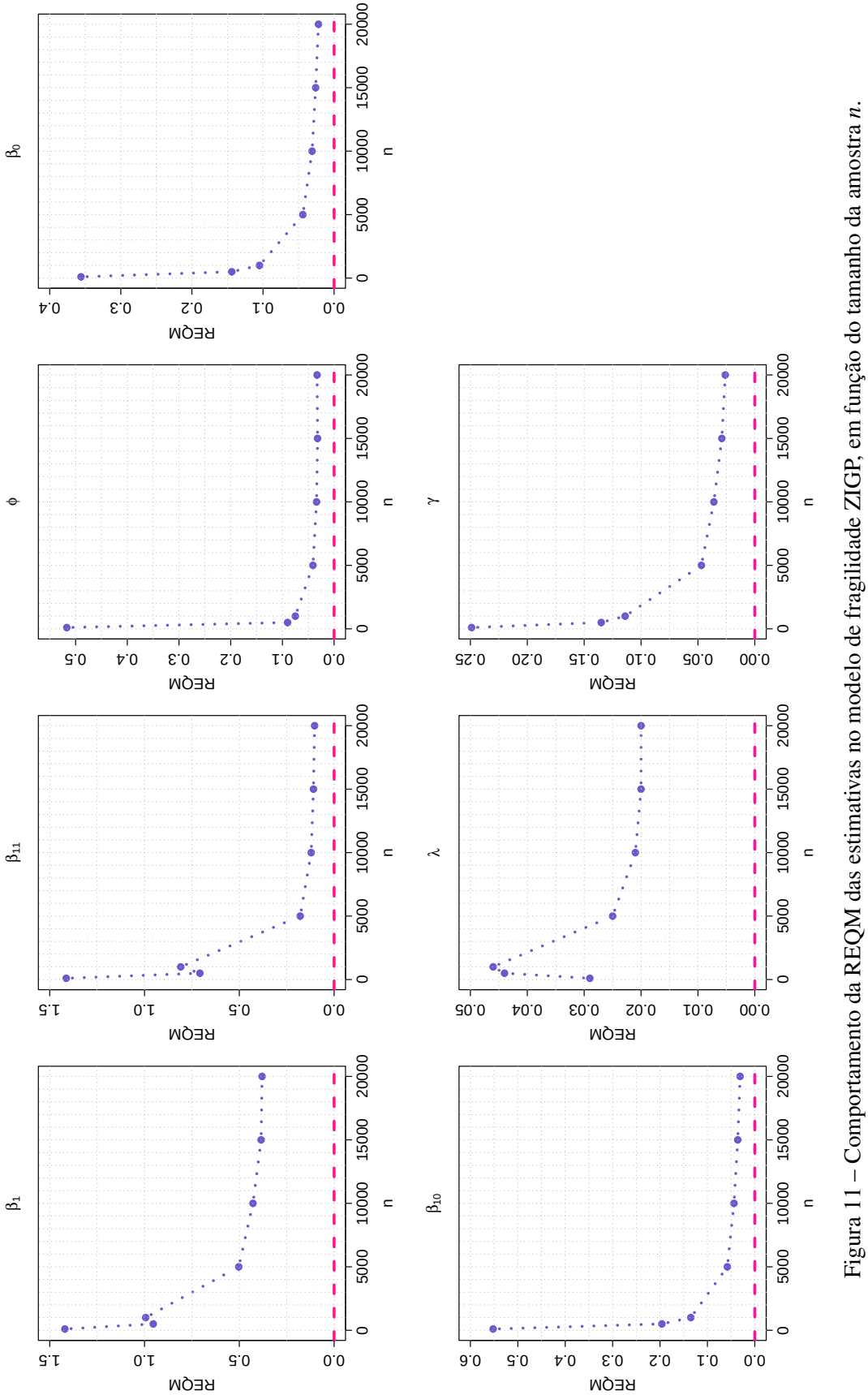

Figura 11 – Comportamento da REQM das estimativas no modelo de fragilidade ZIGP, em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

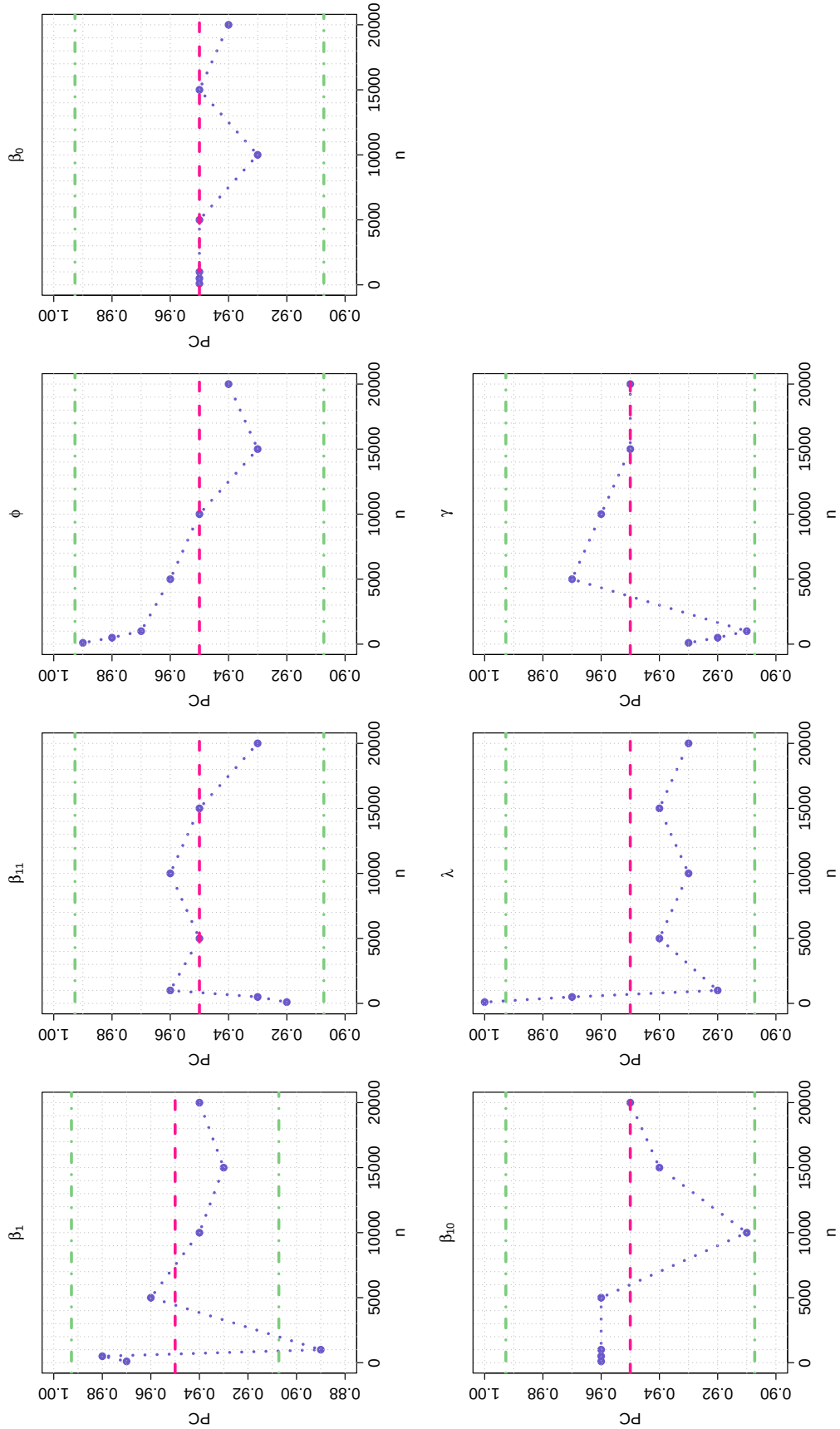

Figura 12 – Comportamento da PC para os parâmetros do modelo de fragilidade ZIGP em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

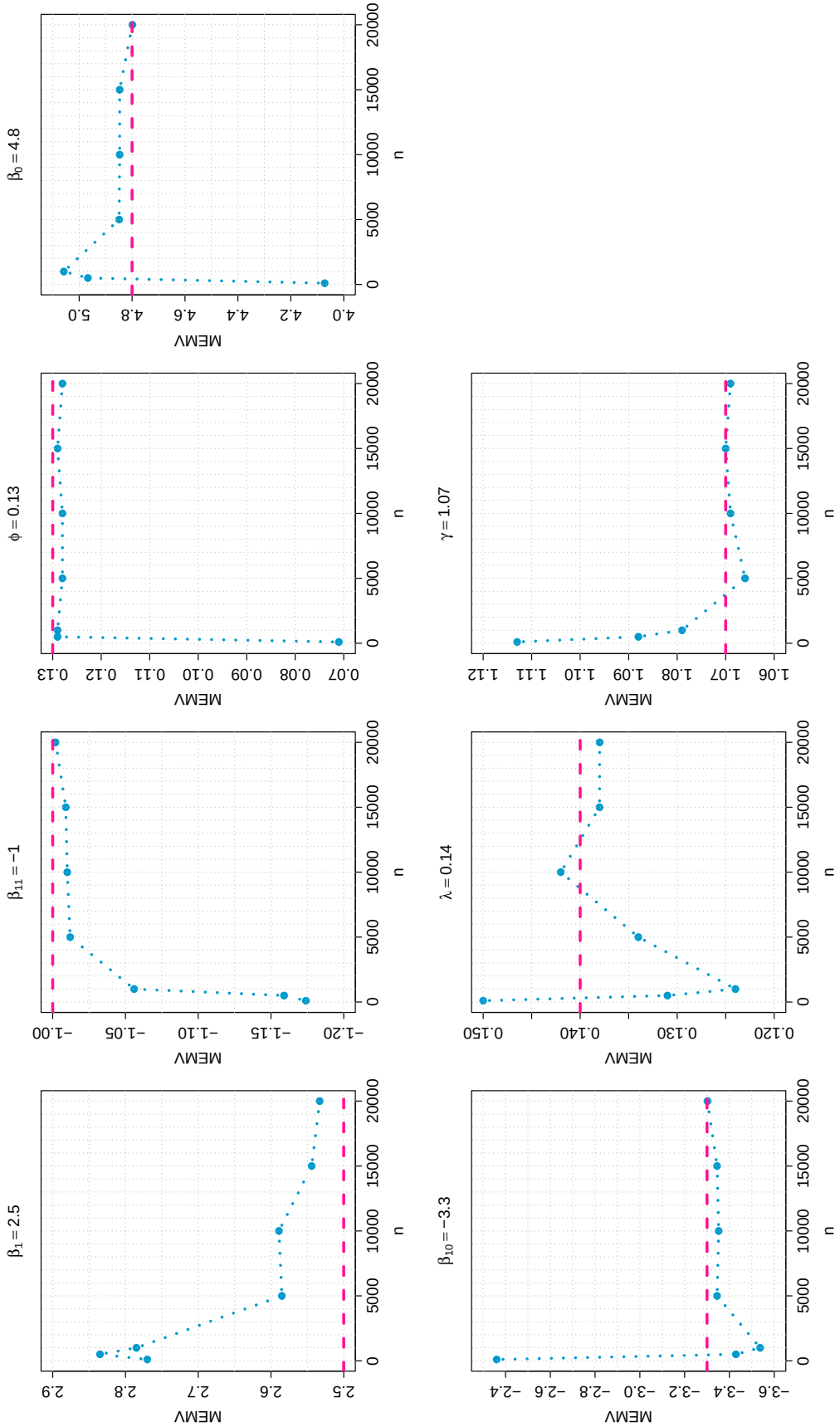

Figura 13 – Convergência da MEMV para os valores verdadeiros (linhas tracejadas cor fúcsia) dos parâmetros do modelo de fragilidade ZDGP, em função do tamanho da amostra *n*.

Fonte: Elaborada pelo autor.

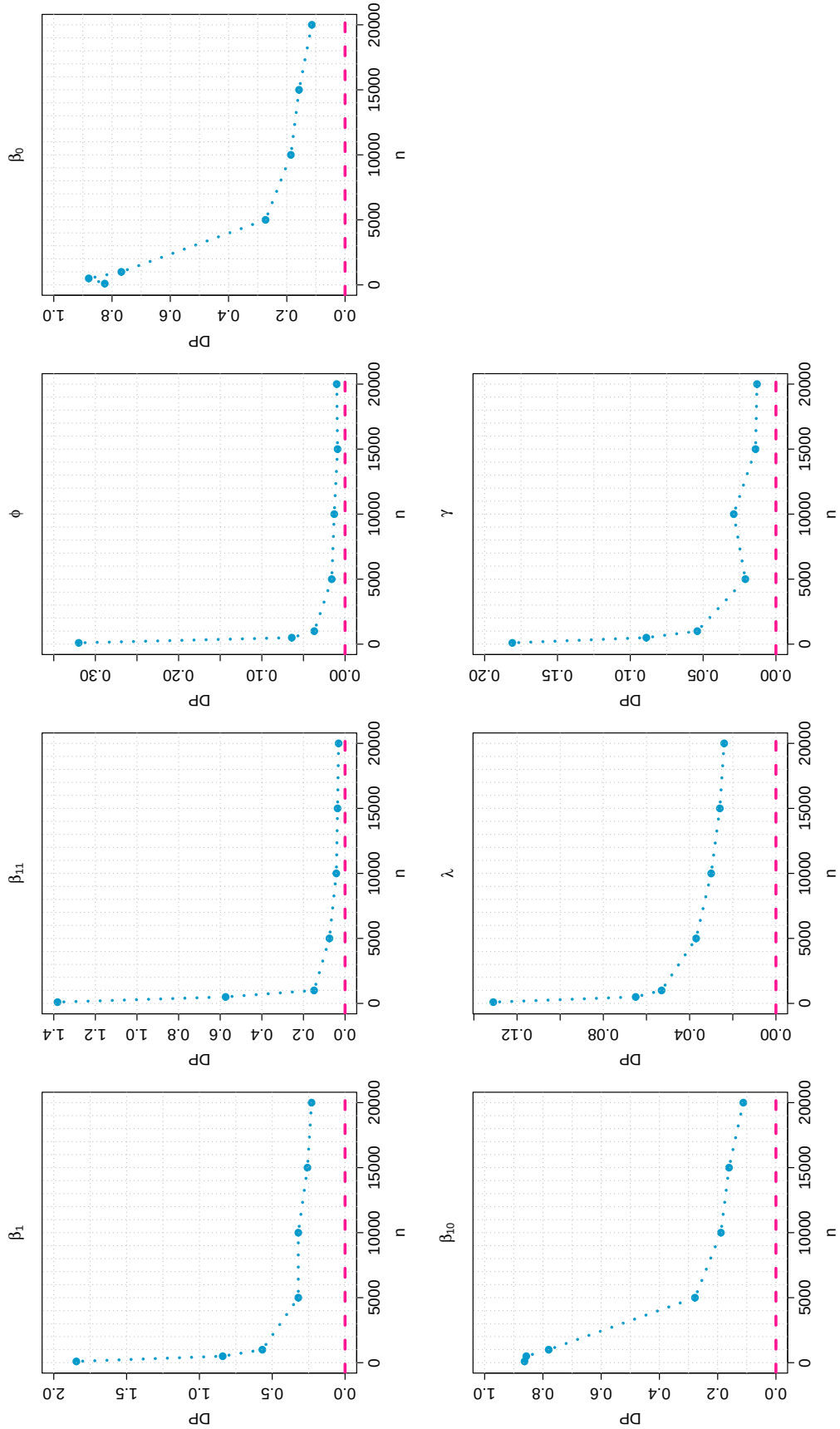

Figura 14 – Comportamento do DP das estimativas no modelo de fragilidade ZDGP, em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

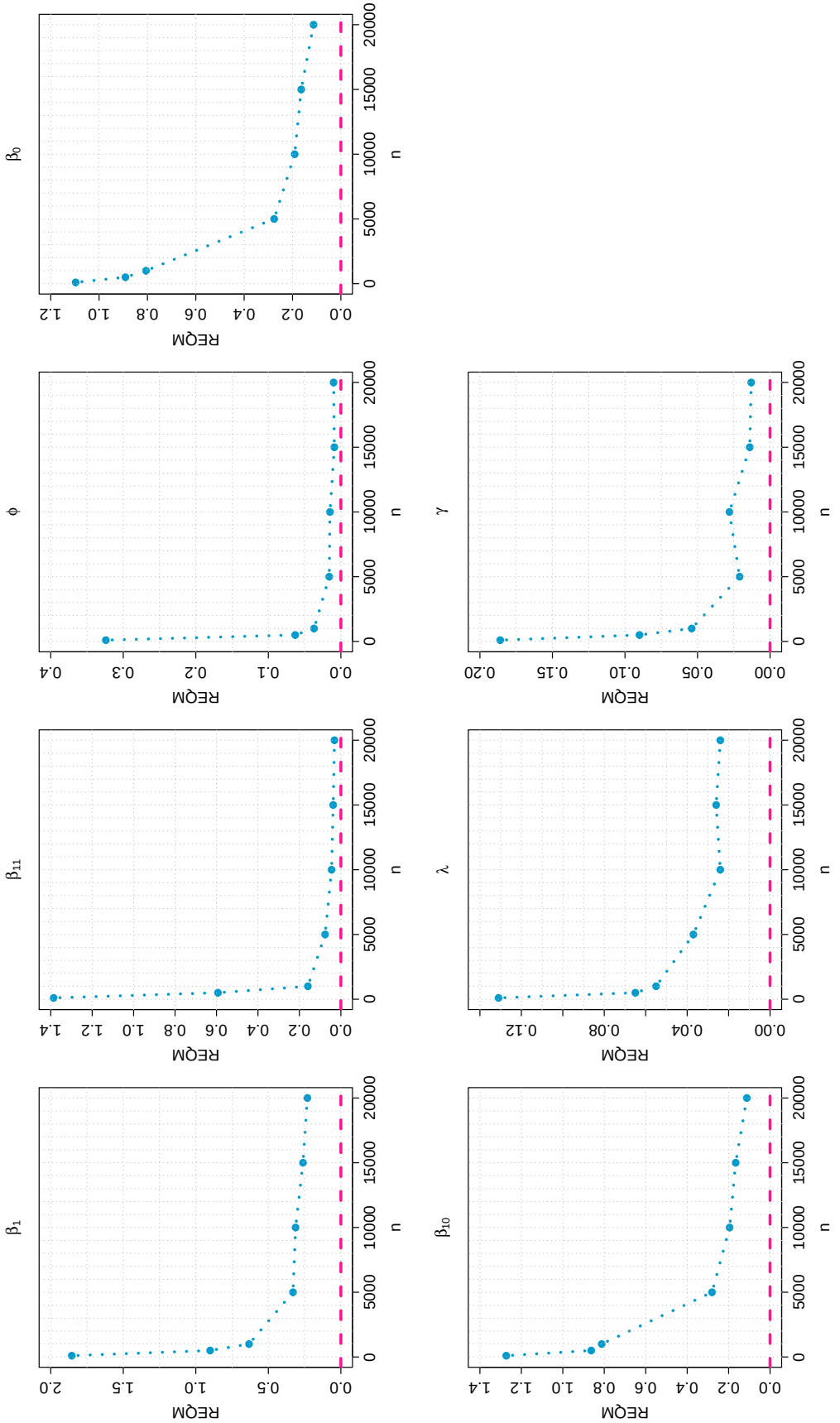

Figura 15 – Comportamento da REQM das estimativas no modelo de fragilidade ZDGP, em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

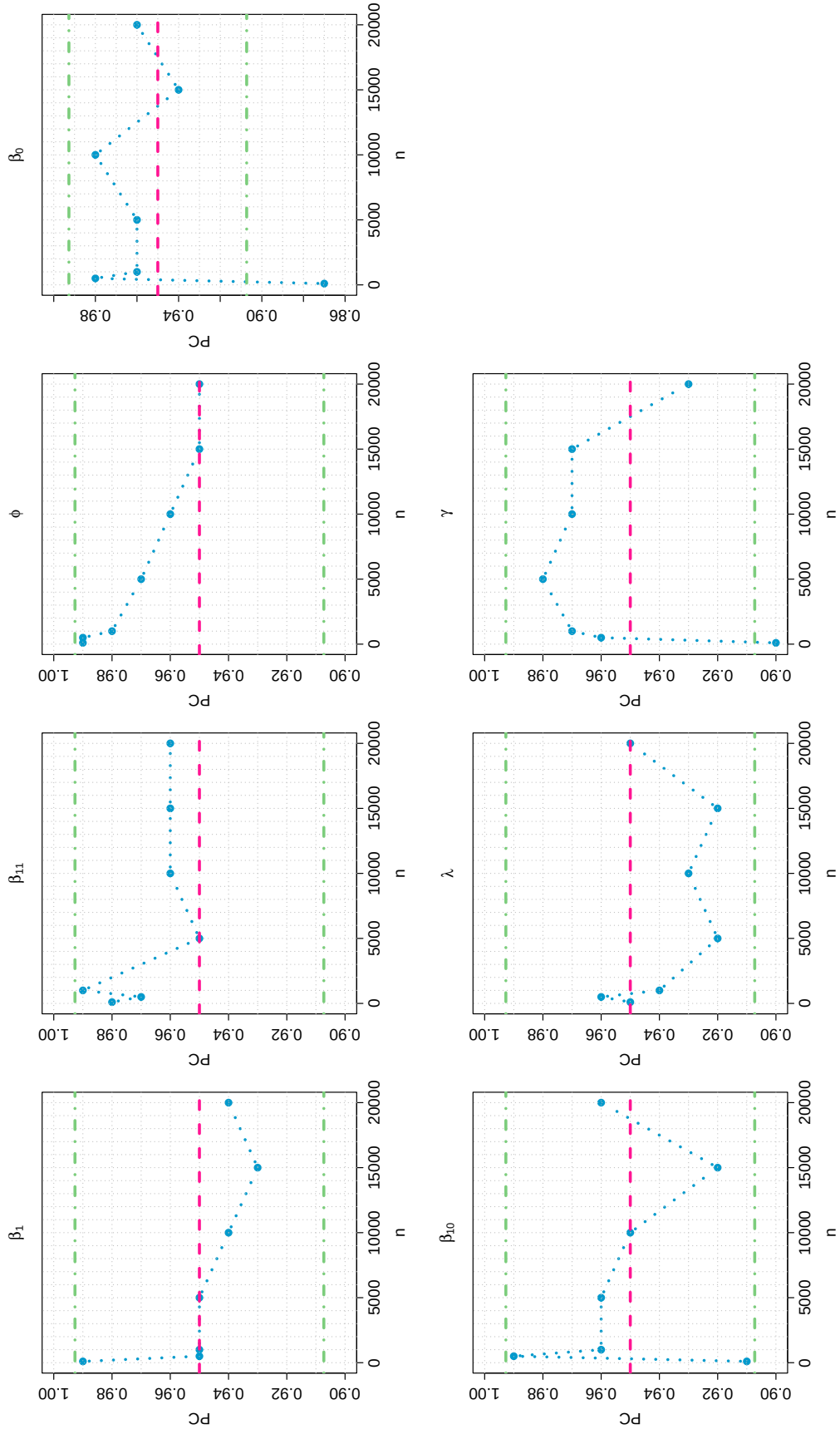

Figura 16 – Comportamento da PC para os parâmetros do modelo de fragilidade ZDGP em função do tamanho da amostra n.

Fonte: Elaborada pelo autor.

APÊNDICE

C

DIAGNÓSTICOS GRÁFICOS DA INFERÊNCIA BAYESIANA

Este apêndice apresenta os diagnósticos gráficos utilizados para validar a robustez da inferência Bayesiana, referentes aos ajustes do modelo HZMGP aos conjuntos de dados de câncer de melanoma e de pulmão. O material está organizado em duas categorias de validação visual:

- Diagnósticos do Processo de Amostragem: São fornecidos os gráficos de traço e de densidade para cada parâmetro, com o objetivo de avaliar a convergência, a mistura e a estacionariedade das cadeias de Markov.

- Diagnósticos da Distribuição a Posteriori Resultante: São apresentados os histogramas de cada distribuição a posteriori, que permitem uma inspeção detalhada de sua forma, simetria e comportamento geral, complementando os sumários numéricos apresentados no corpo da tese.

Em conjunto, estes gráficos oferecem uma validação visual completa tanto da estabilidade do processo computacional quanto da adequação das amostras geradas para a inferência.

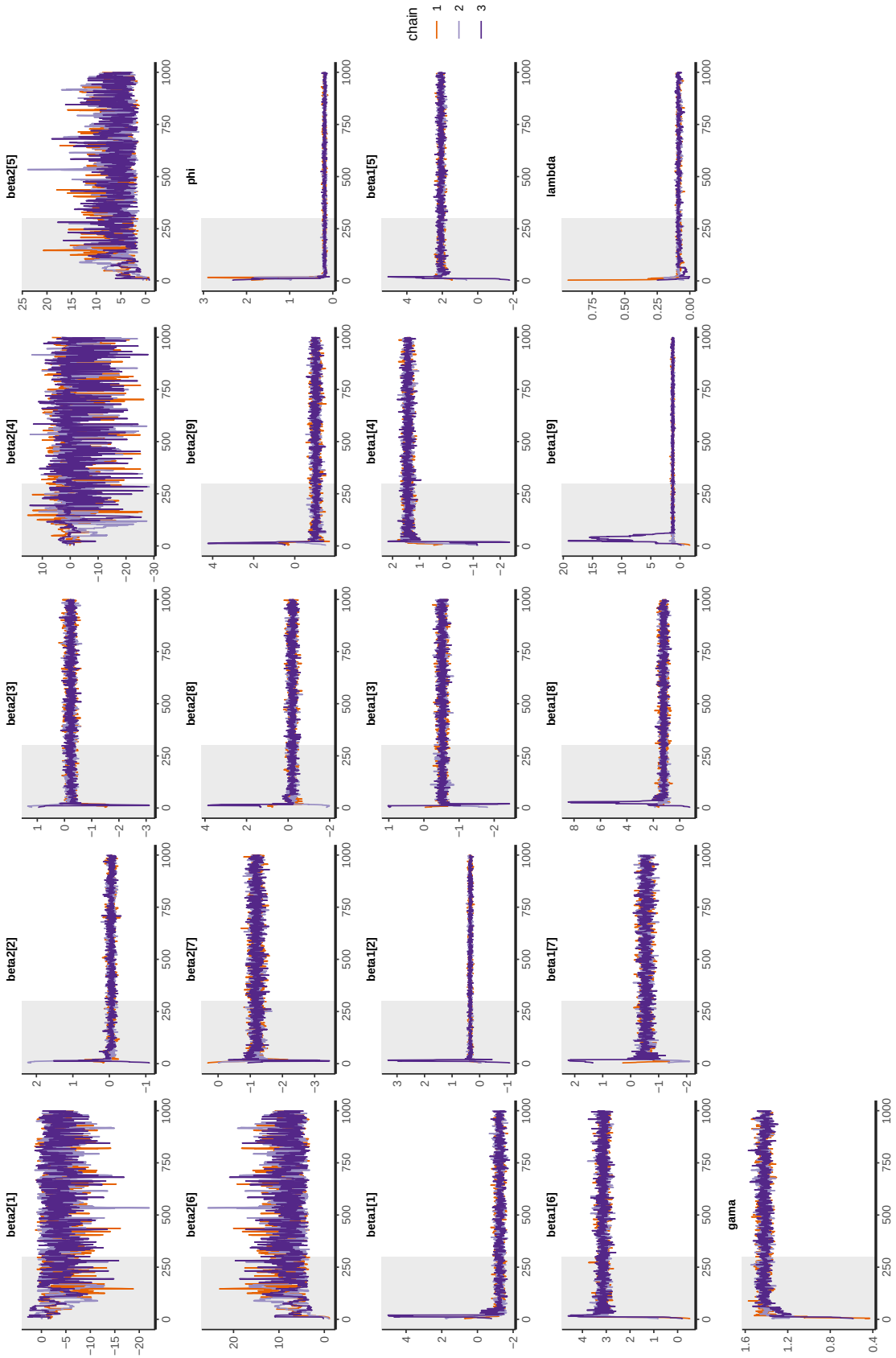

Figura 17 – Gráfico de convergência dos parâmetros do modelo de fragilidade HZMGP ajustado aos dados de câncer de melanoma incluindo todas as covariáveis.

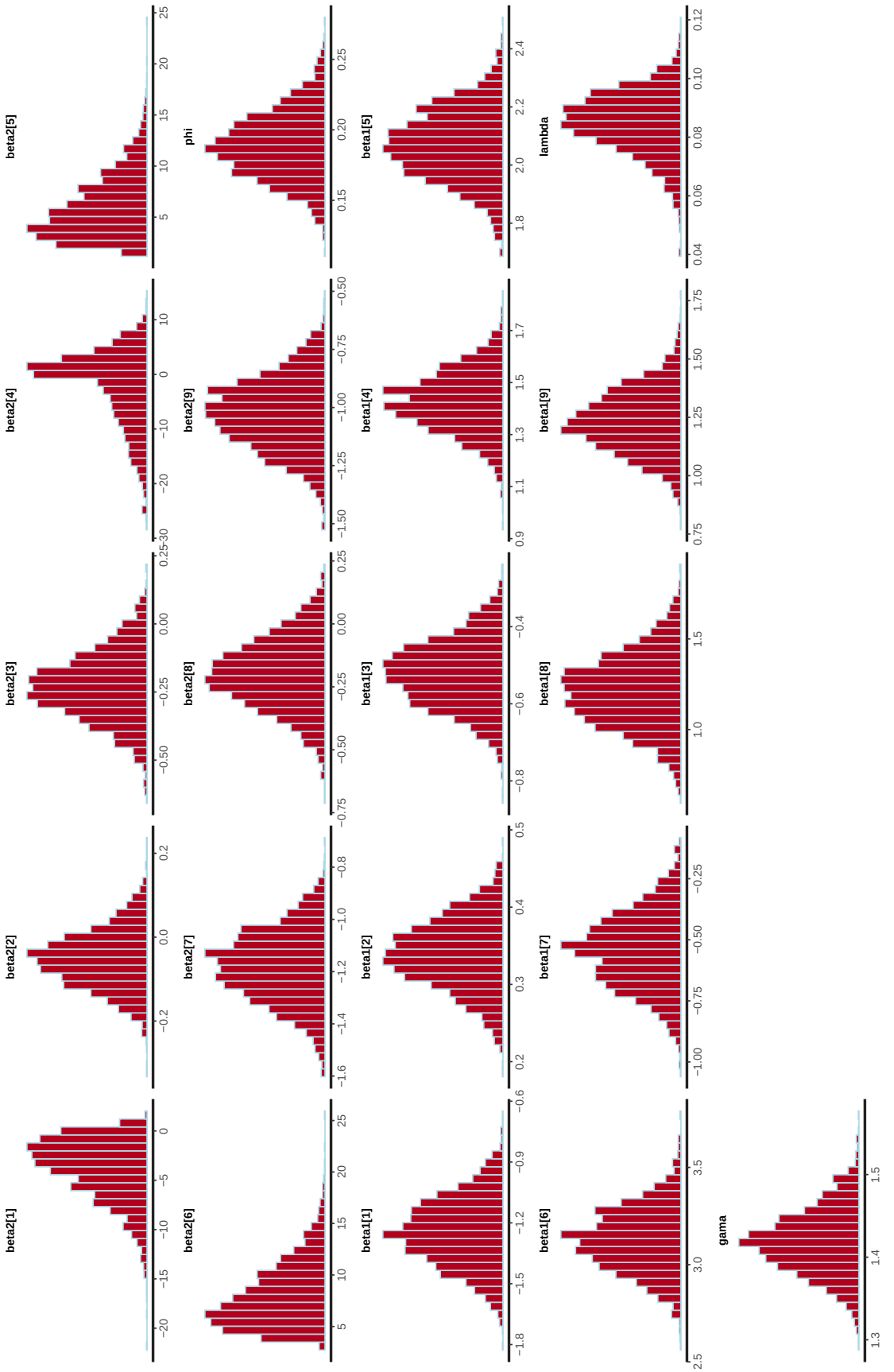

Figura 18 – Histograma dos parâmetros do modelo de fragilidade HZMGP ajustado aos dados de câncer de melanoma incluindo todas as covariáveis.

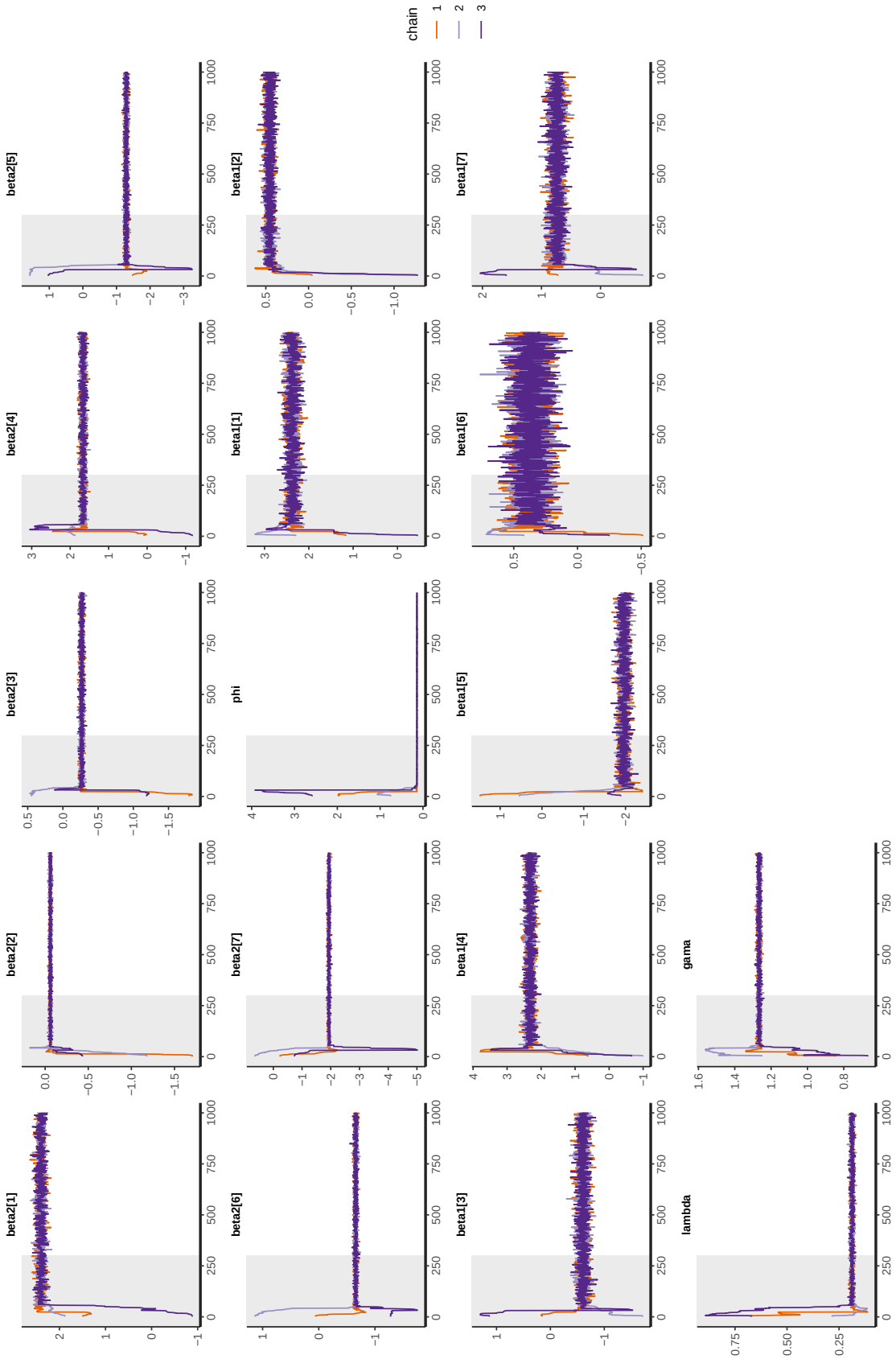

Figura 19 – Gráfico de convergência dos parâmetros do modelo de fragilidade HZMGP ajustado aos dados de câncer de pulmão incluindo todas as covariáveis.

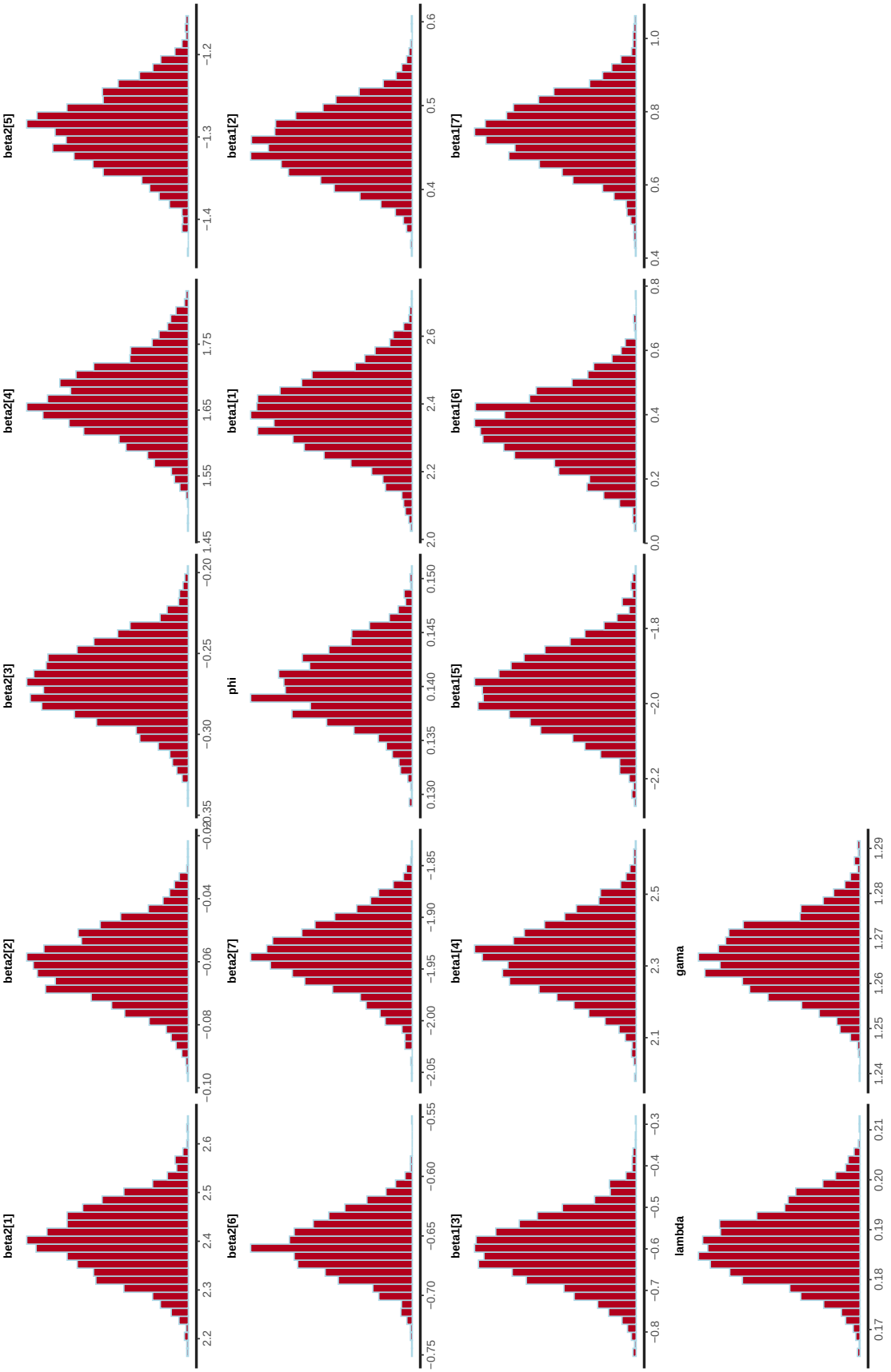

Figura 20 – Histograma dos parâmetros do modelo de fragilidade HZMGP ajustado aos dados de câncer de pulmão incluindo todas as covariáveis.


ICMC
USP
SÃO CARLOS