UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Modelos de sobrevivência induzidos por fragilidade discreta com fração de cura e riscos proporcionais

Ana Paula Jorge do Espirito Santo

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

ICMC USP SÃO CARLOS

Ana Paula Jorge do Espirito Santo

Modelos de sobrevivência induzidos por fragilidade discreta com fração de cura e riscos proporcionais

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

Área de Concentração: Estatística

Orientador: Prof. Dr. Vicente Garibay Cancho

USP – São Carlos
Dezembro de 2022

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

JE77m

Jorge do Espírito Santo, Ana Paula
Modelos de sobrevivência induzidos por fragilidade discreta com fração de cura e riscos proporcionais / Ana Paula Jorge do Espírito Santo; orientador Vicente Garibay Cancho. -- São Carlos, 2022.
68 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Fragilidade discreta. 2. Fração de cura. 3. Riscos proporcionais. 4. Distribuição Katz. 5. Distribuição Poisson Generalizada. I. Garibay Cancho, Vicente, orient. II. Título.

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

Ana Paula Jorge do Espirito Santo

Survival models induced by discrete frailty with cure rate and proportional hazards

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

Concentration Area: Statistics

Advisor: Prof. Dr. Vicente Garibay Cancho

USP – São Carlos
December 2022

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 Ana Paula Jorge do Espirito Santo, realizada em 21/10/2022.

Comissão Julgadora:

Prof. Dr. Vicente Garibay Cancho (USP)

Prof. Dr. Josemar Rodrigues (USP)

Profa. Dra. Elizabeth Mie Hashimoto (UTFPR)

Prof. Dr. Josmar Mazucheli (UEM)

Profa. Dra. Roseli Aparecida Leandro (ESALQ/USP)

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

AGRADECIMENTOS

Em primeiro lugar, quero agradecer a Deus por me amparar nos momentos difíceis, me dar força interior para superar as dificuldades e guiar para o caminho das conquistas.

Ao meu esposo Gabriel, não tenho palavras para expressar meu amor e gratidão. Durante todo o período de doutorado, ele esteve ao meu lado, me apoiando e fazendo tudo ao seu alcance para me ajudar, especialmente durante este período de conclusão da tese.

Agradeço aos meus pais, Mauro e Solange, por acreditar em mim e proporcionar meus estudos. Aos meus irmãos, cunhados, cunhadas, sogra, sogro, pois mesmo estando longe sempre se preocuparam comigo.

Agradeço aos meus afilhados Valentina, Catharina e Pedro e ao meu sobrinho Jorge que atráves da leveza de ser criança consegui enxergar o mundo de uma forma mais leve.

Ao meu orientador Vicente por todos os ensinamentos, orientações, conversas e conselhos que foram fundamentais para a realização desse trabalho. Muito obrigada!

Aos amigos que fiz durante o Doutorado, em especial a Bruna, Claudia, Jardel, Marina, Oilson e Vitor por todas as horas de estudos, por todas as conversas e amizade. Aos colegas Caio, Diego, Elizbeth, Fabiano e Tais Roberta pelas ajudas e contribuições.

A todos os professores do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs) do Instituto de Ciências Matemáticas e de Computação (ICMC-USP) e do Departamento de Estatística (DEs-UFSCar), que de forma direta ou indireta contribuíram para minha formação acadêmica.

Agradeço aos membros da banca pelas ricas contribuições em minha pesquisa.

À Coordenação de Aperfeiçoamento Pessoal de Nível Superior - CAPES pelo apoio financeiro.

"Ninguém caminha sem aprender a caminhar,
sem aprender a fazer o caminho caminhando,
refazendo e retocando o sonho pelo qual se pôs a caminhar."
(Paulo Freire)

RESUMO

DO ESPIRITO SANTO, A. P. J. **Modelos de sobrevivência induzidos por fragilidade discreta com fração de cura e riscos proporcionais**. 2022. 68 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, 2022.

Este trabalho apresenta dois modelos de sobrevivência induzidos por fragilidade discreta com heterogeneidade não observada e estrutura de riscos proporcionais para dados de tempo de vida. O primeiro modelo considera a variável discreta de fragilidade com distribuição de Katz e o segundo com distribuição de Poisson Generalizada, que possuem propriedades de sobredispersão, equidispersão e subdispersão. Os novos modelos englobam, como caso particular, o modelo de fração de cura de promoção. O modelo proposto com fragilidade discreta Katz ainda contempla o modelo de mistura com fração de cura e o modelo de fração de cura com dispersão. Discutiu-se aspectos de inferência para os modelos propostos, em uma abordagem clássica para distribuição Katz, para qual foram empregadas as ferramentas de máxima verossimilhança e apresentou-se modelos de regressão para avaliar os efeitos das covariáveis na fração de curadas. Além disso, um algoritmo para determinar as estimativas de máxima verossimilhança dos parâmetros do modelo foi apresentado. Para o modelo Poisson Generalizada, empregou-se também uma abordagem bayesiana, através do método de simulação Monte Carlo via Cadeias de Markov, mais especificamente o Algoritmo de Metropolis-Hastings. Finalmente, a modelagem foi totalmente ilustrada em um conjunto de dados de câncer cervical.

Palavras-chave: Fragilidade discreta; Fração de cura; Riscos proporcionais; Distribuição Katz; Distribuição Poisson Generalizada.

ABSTRACT

DO ESPIRITO SANTO, A. P. J. **Survival models induced by discrete frailty with cure rate and proportional hazards**. 2022. 68 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, 2022.

This work presents two new survival models induced by discrete frailty with unobserved heterogeneity and proportional hazards structure, for lifetime data. The first model consider the discrete frailty variable with Katz distribution and the second with Generalized Poisson distribution, which have overdispersion, equidispersion and underdispersion properties. The new models encompasses as particular case the promotion cure rate model. The proposed model with katz discrete frailty also encompasses the mixture cure rate model and cure rate model with dispersion. Inference aspects for proposed models as discussed, in a classical approach for Katz distribution, for which the maximum likelihood tools were used and regression models were presented to evaluate the effects of covariates in the cured fraction. Furthermore, an expectation maximization algorithm for determining the maximum likelihood estimates of the parameters of the model was presented. For Generalized Poisson model, a baysean aprouch was also used, through Markov chain Monte Carlo simulation method, especifically Metropolis–Hastings algorithm. Finally, the modeling was fully illustrated on cervical cancer data sets.

Keywords: Discrete frailty; Cure rate; Proportional hazards; Katz distribution; Generalized Poisson distribution.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AIC	Critério da Informação de Akaike
BCH	*Bounded Cumulative Hazard*
BFGS	Broyden-Fletcher-Goldfarb-Shanno method
BJPS	*Brazilian Journal of Probability and Statistics*
BJPS	*Brazilian Journal of Probability and Statistics*
CPO	Ordenada Preditiva Condicional
DIC	*Deviance Information Criterion*
DP	desvio padrão
EAIC	*Expected Akaike Information Criterion*
EBIC	*Expected Bayesian Information Criterion*
EM	*Expectation Maximization Algorithm*
EMV	Método de Estimação de Máxima Verossimilhança
FOSP	Fundação Oncocentro de São Paulo
GBCH	*Generalized Bounded Cumulative Hazard*
KM	Kaplan-Meier
LI	limite inferior
LPML	*Log Pseudo Marginal Likelihood*
LS	limite superior
M-H	Algoritmo de Metropolis-Hasting
MCMC	Método de Monte Carlo via Cadeia de Markov
PC	probabilidades de cobertura
PVF	*Power Variance Function*
REQM	raiz do erro quadrático médio
RR	razão de risco (ou risco relativo)
TRV	Teste da Razão de Verossimilhanças

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Análise de sobrevivência pode ser definida como um conjunto de procedimentos usados para estudar situações em que a varável resposta é o tempo até a ocorrência de um evento de particular interesse. Este evento (tempo até a falha ou ocorrência) pode ser a morte de um paciente, o aparecimento, recaída ou remissão de uma doença, falha de um componente eletrônico, uso de um cartão de crédito, gravidez, e muitos outros. Dessa forma, o tempo de medida pode ir de segundos até anos, sendo a variável resposta positiva, geralmente univariada e contínua (LAWLESS, 2003).

Dada a natureza destes dados, que por vezes necessitam de anos para serem obtidos é comum que a informação sobre a ocorrência seja incompleta. Dados com observação parcial são chamados de censurados quando se tem informação a respeito do tempo de sobrevivência, mas não se sabe o momento exato, ou seja, o estudo não atingiu o limite de interesse. Por exemplo, em estudos clínicos a interrupção do tratamento sem autorização médica é muito comum, o que ocasiona a presença de observações censuradas. Diferentemente de outras técnicas estatísticas, a análise de sobrevivência incorpora essas informações à análise (KAPLAN; MEIER, 1958). Além disso, as censuras podem ser classificadas de três formas. Quando o tempo de ocorrência do evento é maior que o tempo da análise, tem-se uma censura à direita. Quando o evento ocorreu antes do indivíduo ser observado no estudo considera-se censura à esquerda. Também pode ocorrer a censura intervalar, quando não se tem precisão com relação ao tempo da ocorrência, mas sim um intervalo em que ocorreu.

Por outro lado, estudos ou ensaios controlados buscam analisar e modificar variáveis uma à uma, no entanto, a modelagem de dados reais raramente permite tal simplificação. Para representar fatores que influenciam na taxa de sobrevivência dos indivíduos, os modelos passaram a incorporar variáveis explicativas (COX, 1972). Quando os indivíduos da amostra estudada possuem diferentes características que influenciam na probabilidade de ocorrência do evento, considera-se que esta amostra possui heterogeneidade, podendo ser observada ou não. Algumas

interpretações para essas covariáveis dão base para diferentes metodologias de estudos dentro da análise de sobrevivência.

Modelos de riscos proporcionais partem da hipótese que os riscos de dois indivíduos em um tempo t estão relacionados por uma constante de proporcionalidade que não depende de t (COX; OAKES, 1984). Modelos de risco competitivos supõem causas diferentes afetando a população ao mesmo tempo. Por exemplo, uma análise de sobrevida após um diagnóstico de câncer considera que os organismos possuem diferentes tipos de células e que a recidiva da doença ocorre quando uma determinada célula alterada se manifesta (RODRIGUES *et al.*, 2009a). Modelos de mistura consideram amostras ou populações com perfis diferentes, por exemplo pacientes com e sem possibilidade de cura, em um mesmo modelo.

Os modelos de fragilidade incluem uma variável para expressar a maior ou menor propensão de um indivíduo do grupo sofrer o evento, essa variável aleatória pode receber uma distribuição contínua (VAUPEL; MANTON; STALLARD, 1979). Todavia, a amostra pode conter elementos que não venham a sofrer o evento estudado (BERKSON; GAGE, 1952), esse grupo é frequentemente chamado de taxa de cura ou fração de cura e modelado por meio de modelos de mistura, modelos de longa duração, modelos com fração de cura, e outras variações. Nesse cenário, modelos de fragilidade com distribuição discreta (WIENKE, 2010) incluem a taxa de curados ao estudo, garantindo uma modelagem estatística mais precisa.

Um outro fator que pode aumentar a complexidade da análise está relacionado a maneira como os dados estão distribuídos. Por exemplo, quando a variância dos dados é relativamente maior que a média proposta pelo modelo, apresenta-se um cenário de sobredispersão (MCCULLAGH; NELDER, 1989). Da mesma forma, quando a variância é desproporcionalmente menor que a média, surge a subdispersão. Quando os dados são mais homogêneos, a variância coincide com a média, situação chamada de equidispersão.

Para contribuir com a metodologia na análise de sobrevivência, tem-se como objetivo geral deste trabalho construir modelos de sobrevivência considerando a abordagem de fragilidade, mais especificamente modelos com taxa de cura induzidos por fragilidade discreta.

A primeira distribuição escolhida para a variável de fragilidade foi a Katz (KATZ, 1965). O modelo abrange subdispersão, equidispersão e sobredispersão, com a vantagem de apresentar expressões analíticas para a média e a variância . Também engloba, como casos particulares, o modelo de mistura com fração de cura, modelo de promoção com fração de cura e o modelo de fração de cura com dispersão, além disso, apresenta estrutura de riscos proporcionais. Para avaliar a performance do modelo foram realizados, numa abordagem clássica, estudo de simulação e aplicação em dados reais.

A segunda distribuição empregada foi a Poisson Generalizada (CONSUL; JAIN, 1973). O modelo abrange sobredispersão e pode ser escrito como modelo de mistura com fração de cura. Para avaliar a performance do modelo foram realizados, numa abordagem bayesiana, estudos de

simulação e aplicação em dados reais.

1.1 Objetivos Específicos

- Propor modelos de sobrevivência com fração de cura induzido por uma variável discreta de fragilidade com distribuição Katz e Poisson Generalizada.
- Avaliar o desempenho do modelo Katz na estimação de parâmetros através de um estudo de simulação em uma abordagem clássica, utilizando o Método de Estimação de Máxima Verossimilhança (EMV).
- Avaliar o desempenho do modelo Poisson Generalizado em uma abordagem bayesiana na estimação de parâmetros através do Método de Monte Carlo via Cadeia de Markov (MCMC).
- Avaliar a aplicação dos modelos propostos em um conjunto de dados reais de câncer cervical fornecidos pela Fundação Oncocentro de São Paulo (FOSP).

1.2 Apresentação dos capítulos

No Capítulo 2 tem-se a definição dos conceitos necessários para contextualização da pesquisa, assim como uma revisão bibliográfica da área de modelos de fragilidade discreta. No Capítulo 3 apresenta-se o modelo induzido por fragilidade discreta com heterogeneidade não observada, onde a variável Z tem distribuição Katz. Uma abordagem clássica é empregada para estimação dos parâmetros do modelo, a metodologia é avaliada através de conjunto de dados simulados e uma aplicação em dados reais. Por fim, tem-se as conclusões para este modelo. No Capítulo 4 tem-se o desenvolvimento do modelo de sobrevivência com variável discreta Poisson Generalizada. A abordagem bayesiana é empregada para estimação dos parâmetros e aplicação em conjuntos de dados simulados e reais. Por fim, as conclusões e perspectivas de pesquisa futuras foram discutidas no Capítulo 5.

CAPÍTULO

2

REFERENCIAL TEÓRICO

Na análise de sobrevivência tem-se uma variável aleatória T e a variável indicadora de censura C que assume valor 1 se o tempo de sobrevivência é completamente observado ou 0 se não for. Além disso, associa-se um conjunto de variáveis observáveis, chamadas variáveis explicativas ou covariáveis, que contém informações adicionais para cada indivíduo, como idade, sexo, tipo de tratamento, cirurgia, entre outras. Uma forma de investigar a influência dessas covariáveis no tempo de sobrevivência de um indivíduo é através de modelos de regressão. Cox (1972) desenvolveu o modelo de riscos proporcionais, servindo de base para todo o campo da análise de sobrevivência. Modelos de regressão paramétrica foram apresentados por Lawless (2003) e modelos de regressão semi-paramétricas por Collett (2003).

Em dados de sobrevivência algumas covariáveis relevantes podem não estar sendo levadas em consideração nos estudos. Nesse contexto, modelos de fragilidade (WIENKE, 2010) são geralmente usados para modelar a dependência não observada e a heterogeneidade dos dados de sobrevivência individuais, para explicar a heterogeneidade causada por essas covariáveis não medidas. A ideia de uma heterogeneidade não observada é supor que existem diferentes fragilidades e que pacientes “frágeis” ou “propensos” tendem a experienciar o evento antes dos “não frágeis”, todavia, dado tempo suficiente, todos os índivíduos terão experimentado o evento. Vaupel, Manton e Stallard (1979) introduziram o conceito de fragilidade a partir da bioestatística ao aplicá-lo nos dados de mortalidade populacional em modelos de sobrevivência univariados e Lancaster (1979) introduziu o modelo na economia sob a terminologia de modelo de mistura com riscos proporcionais.

O conceito de fragilidade é geralmente modelado como uma variável aleatória não observada agindo multiplicativamente na função de risco base. Os primeiros trabalhos na área utilizaram a distribuição gama (VAUPEL; MANTON; STALLARD, 1979) por sua simplicidade e conveniência. Hougaard (1984) começou a aplicar diferentes distribuições, mostrando que toda classe de distribuições não negativas exponenciais também compartilhavam propriedades da

gama. Com o avanço das pesquisas, diversas distribuições e suas respectivas propriedades foram utilizadas, por exemplo, a Gaussiana (HOUGAARD, 1984), lognormal (SANTOS; DAVIES; FRANCIS, 1995), positiva estável (HOUGAARD, 1986a) e gama generalizada (BALAKRISHNAN; PENG, 2006). Wienke (2010) menciona a utilização da família *Power Variance Function* (PVF), distribuição sugerida por Tweedie (1984) e estudada independentemente por Hougaard (1986b), uma família generalizada de distribuições de fragilidade que inclui as distribuições gama, Gaussiana inversa e positiva estável.

Uma desvantagem das funções contínuas nos modelos de fragilidade é que não permitem fragilidade zero, o que pode ser necessário em situações específicas (ATA; ÖZEL, 2013). Por exemplo, há alguns anos um indivíduo com diagnóstico de câncer tinha baixíssima chance de cura, com o avanço dos tratamentos médicos, hoje uma parte desses pacientes são curados, dessa forma mantém-se sem a doença depois de prolongados acompanhamentos. Nesses casos tem-se a fragilidade zero, implicando em uma fração de cura. Para solucionar esse problema, distribuições discretas foram introduzidas na variável de fragilidade (WIENKE, 2010; CARONI; CROWDER; KIMBER, 2010a). Essa abordagem levou a uma ligação aos modelos de longa duração (RODRIGUES *et al.*, 2009a).

Modelos com fração de cura compõem uma vasta literatura na análise de sobrevivência, também chamados de modelos de sobrevivência com taxa de cura ou modelos de sobrevivência de longa duração. Muitos desses modelos são obtidos em cenários de riscos competitivos como em Tsodikov, Ibrahim e Yakovlev (2003), Yin e Ibrahim (2005), Cooner *et al.* (2007), Rodrigues *et al.* (2009b). Em um trabalho de grande impacto na área, Rodrigues *et al.* (2009a) apresentam uma unificação de modelos de longa duração que apresenta propriedades de risco proporcional, se e somente se, o número de causas competindo para o evento segue uma distribuição Poisson.

Cancho, Rodrigues e Castro (2011) empregaram a distribuição binomial negativa e uma abordagem bayesiana para modelar dados com taxa de cura. Mais recentemente, Ortega *et al.* (2015) apresentaram um modelo com fração de cura sob o cenário de causas concorrentes com distribuição de séries de potência para prever a sobrevida do carcinoma de mama em mulheres submetidas a mastectomia. Cordeiro *et al.* (2016) propuseram um modelo de sobrevivência com taxa de cura assumindo que o número de causas concorrentes segue uma distribuição binomial negativa e o tempo para o evento de interesse tenha a distribuição Birnbaum-Saunders.

Barriga *et al.* (2018) propuseram um novo modelo com taxa de cura, sendo uma extensão do modelo de taxa de cura *Bounded Cumulative Hazard* (BCH), incluindo um parâmetro de dispersão para controlar a heterogeneidade não observada. Em outro trabalho, Barriga *et al.* (2020) apresentaram um modelo que permite diferentes mecanismos de ativação, também com a distribuição Birnbaum-Saunders para o tempo, mas empregando a distribuição geométrica para o número de riscos competindo para o evento de interesse em uma abordagem bayesiana. Silva, Cordeiro e Ortega (2020) propuseram dois modelos de regressão baseados na distribuição beta Weibull modificada, um modelo de mistura que busca estimar os efeitos das covariáveis

na fração de cura e um modelo baseado na distribuição log-beta Weibull modificada buscando estimar os efeitos das covariáveis no tempo de sobrevivência.

Modelos de taxa de cura também podem ser obtidos a partir de modelos de risco proporcional com distribuições de fragilidade discreta como em Caroni, Crowder e Kimber (2010b), que aplicaram, entre outras, as distribuições binomial negativa, Poisson e geométrica na variável fragilidade. Assim como Ata e Özel (2013) derivaram funções de sobrevivência baseados no processo composto de Poisson. Moger *et al.* (2004) apresentaram um modelo de fragilidade usando a distribuição Compound-Poisson para o estudo de incidência de câncer de testículo. Milani *et al.* (2015) estenderam o modelo generalizado logístico dependente no tempo, incluindo a fragilidade e apresentaram um aplicação em estudos de câncer de pulmão. Souza *et al.* (2017) empregaram a Hiper-Poisson como distruibuição flexível para fragilidade, empregando uma abordagem bayesiana, buscaram abranger casos de sobre e subdispersão. Cancho *et al.* (2018) propuseram um modelo de dispersão induzido por fragilidade discreta numa perspectiva bayesiana com distribuição zero-inflacionada power series, que apresenta a propriedade de sobredispersão. Cancho *et al.* (2019) apresentaram o mesmo modelo numa perspectiva clássica.

Em Cancho *et al.* (2022) desenvolveu-se um modelo considerando fatores de risco latentes seguindo uma distribuição Poisson Generalizada, que inclui o modelo de promoção de taxa de cura como caso especial. Para análise do modelo, aplicou-se a abordagem clássica de máxima verossimilhança via algoritmo *Expectation Maximization Algorithm* (EM), enquanto problemas de discriminação foram analisados com a ajuda do teste de razão de verossimilhança. O modelo também foi testado em dados reais de câncer cervical.

CAPÍTULO

3

MODELO DE SOBREVIVÊNCIA COM DISTRIBUIÇÃO DE FRAGILIDADE KATZ

Neste capítulo, tem-se o modelo de sobrevivência com fração de cura induzido por fragilidade discreta com distribuição Katz (KATZ, 1965). Todos os estudos do modelo proposto, incluindo a estimação dos parâmetros, intervalos de confiança, testes de hipóteses, simulações e a aplicação no conjunto de dados reais foram desenvolvidos sob a abordagem clássica.

3.1 Fragilidade

A fragilidade pode ser modelada como uma variável aleatória não observada agindo multiplicativamente na função de risco base. Assim, considere $T > 0$ uma variável aleatória contínua representando o tempo de vida de um indivíduo e Z uma variável aleatória de fragilidade. A função de risco condicional para uma dada variável de fragilidade $Z = z$ no tempo $t > 0$ é dada pelo produto de um fator aleatório z e a função de risco base $h_0(t)$ comum para todos os indivíduos no estudo,

$$h(t|Z = z) = zh_0(t). \tag{3.1}$$

Dessa forma, a correspondente função de sobrevivência definida pela probabilidade de um indivíduo sobreviver ao tempo t condicionada a $Z = z$ é dada por

$$S(t|Z = z) = \exp\{-zH_0(t)\} = S_0(t)^z, \tag{3.2}$$

em que $S_0(t)$ é a função de sobrevivência de base e $H_0(t) = \int_0^t h_0(u)du$ é a função de risco acumulada. Integrando a Equação 3.2 sobre Z tem-se que a função de sobrevivência marginal $S(t)$ é dada por

$$S(t) = \int_0^\infty S(t|Z = z) f_z dz, \tag{3.3}$$

em que f_z é a função densidade de probabilidade da variável de fragilidade contínua Z.

Observe que o modelo de fragilidade na Equação 3.1 apresenta a estrutura de risco proporcional e supondo que z_i é o valor que a variável de fragilidade não observada assume para o paciente i, tem-se que o risco individual cresce se $z_i > 1$, decresce para $z_i < 1$ e para $z_i = 1$ tem-se o modelo de risco proporcional de Cox.

3.1.1 Fragilidade discreta

Como mencionado anteriormente, considere que o interesse seja modelar a fragilidade levando em consideração um grupo de pacientes que não vão experimentar o evento de interesse, seriam os pacientes livres, com risco zero ou fragilidade zero.

Nessa situação, para a Equação 3.2, assumindo que Z é uma variável aleatória discreta com suporte em $\{0,1,2,\dots\}$ e função massa de probabilidade $P(Z=z)=p_z$, tem-se que a função de sobrevivência não condicional, $S(t)$, para um indivíduo na população pode ser obtida por

$$S(t)=\sum_{z=0}^{\infty} S(t|Z=z)p_z = E_Z(S(t|Z)) = G_Z[S_0(t)], \tag{3.4}$$

uma vez que G_Z é a função geradora de probabilidade (f.g.p.) da variável aleatória Z. A função de sobrevivência $S(t)$ em (3.4) possui a mesma expressão matemática que a função de sobrevivência proposta por Tsodikov, Ibrahim e Yakovlev (2003), Rodrigues *et al.* (2009b). Essa fragilidade discreta corresponde a um número aleatório de fatores de risco não observados resultando na heterogeneidade entre os indivíduos.

Se $P(Z=0)>0$, a função de sobrevivência (3.4) é uma função imprópria, isto é, $\lim_{t\to\infty} S(t) = G_Z(0) = P(Z=0) > 0$, fato importante que ajuda a descrever os modelos de longa duração. Quando $P(Z=0)=0$, a função de sobrevivência é própria, isto é, $\lim_{t\to\infty} S(t) = G_Z(0) = P(Z=0) = 0$ e $\lim_{t\to 0} S(t) = G_Z(1) = 1$.

3.2 Modelo com fragilidade Katz

3.2.1 Distribuição Katz

A distribuição Katz foi proposta por Katz (1965). Se Z é uma variável aleatória com distribuição Katz (KATZ, 1965) tem-se que sua função geradora de probabilidade é definida por

$$G_Z(s) = \left[\frac{1-\gamma}{1-\gamma s}\right]^{\frac{\theta}{\gamma}},\ 0 \le s \le 1, \tag{3.5}$$

em que, $\gamma < 1$ mas $\gamma \neq 0$ e $\theta > 0$, com média e a variância $E(Z) = \theta/(1-\gamma)$ e $Var(Z) = \theta/(1-\gamma)^2$, respectivamente. Algumas propriedades da distribuição Katz são apresentadas por Gurland e Tripathi (1975), Tripathi e Gurland (1977), Fang (2003).

Embora a família Katz possua uma estrutura de probabilidade simples, ela carrega as propriedades de equidispersão, subdispersão ou sobredispersão. A sobre e subdispersão podem

ser interpretadas como a falha de algumas suposições subjacentes do modelo, conforme apontado por Mullahy (1997), Bosch e Louise (1998), Luceno (2005), e Kokonendji (2014).

Em relação à razão da variância pela média dada por $\kappa = (1-\gamma)^{-1}$, tem-se a distribuição de Poisson (equidispersão) quando $\kappa = 1$ ($\gamma = 0$). A distribuição binomial com ($\kappa < 1$) ($\gamma < 0$) possui subdispersão e distribuição binomial negativa ($\kappa > 1$) ($0 < \gamma < 1$) possui sobredispersão.

3.2.2 Formulação do Modelo

Um novo modelo de sobrevivência induzido por fragilidade discreta é obtido quando assume-se a fragilidade Katz reparametrizada na média μ, ou seja, para $\theta = (1-\gamma)\mu$, tem-se que a função de sobrevivência (3.4) é dada por

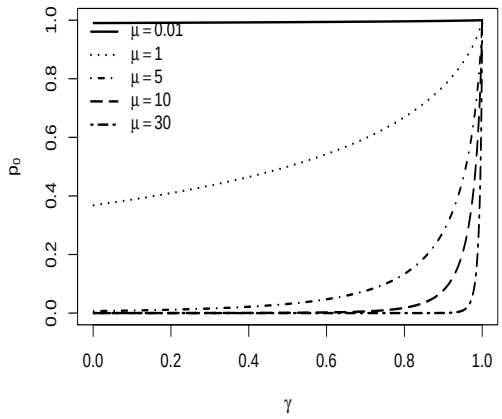
$$S(t) = \left[\frac{1-\gamma}{1-\gamma S_0(t)}\right]^{\frac{(1-\gamma)\mu}{\gamma}}. \tag{3.6}$$

A fração de cura dos indivíduos é dada por $p_0 = \lim_{t\to\infty} S(t)$. Pela Equação 3.6 pode-se escrever

$$p_0 = [1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}} > 0, \tag{3.7}$$

implicando que (3.6) é uma função de sobrevivência imprópria. Observe que $\lim_{\mu\to\infty} p_0 = 0$ e $0 < \gamma < 1$ enquanto que $\lim_{\mu\to 0} p_0 = 1$ como mostra a Figura 1.

Figura 1 – Comportamento da função da fração de cura (p_0) para diferentes valores da média μ.



Fonte: Elaborada pelo autor.

A função de probabilidade

$$\pi(t) = P(Z = 0|T > t) = [1-\gamma S_0(t)]^{\frac{(1-\gamma)\mu}{\gamma}}, \tag{3.8}$$

denota a probabilidade de um indivíduo estar imune ou curado da doença dado que sobreviveu por um tempo $t > 0$ após o tratamento. A probabilidade em (3.8) é uma função crescente em t e para $t = 0$ corresponde a nenhuma informação sobre a imunidade de um indivíduo, exceto a probabilidade geral de ser imune, isto é, a probabilidade é igual a $\pi(0) = p_0$. Tem-se que $\lim_{t\to\infty} \pi(t) = 1$ e certamente o indivíduo estará imune após um longo período de tempo.

Outras funções associadas como a função densidade de probabilidade, função de risco e função de distribuição acumulada podem ser descritas para a variável $t > 0$ e são dadas respectivamente por

$$f(t) = \frac{(1-\gamma)\mu f_0(t)}{1-\gamma S_0(t)} \left[\frac{1-\gamma}{1-\gamma S_0(t)}\right]^{\frac{(1-\gamma)\mu}{\gamma}}, \tag{3.9}$$

em que $f_0(t) = -dS_0(t)/dt$. A função de risco

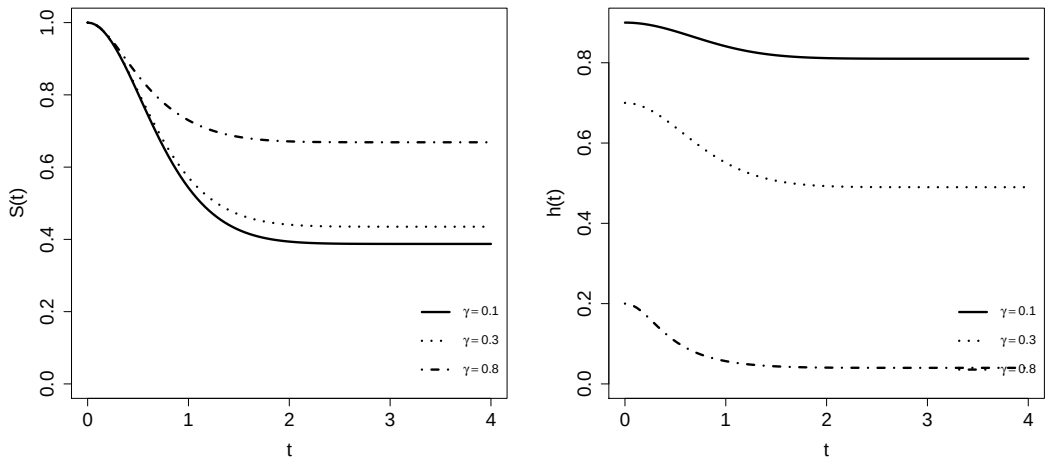
$$h(t) = \mu \left[\frac{(1-\gamma) f_0(t)}{1-\gamma S_0(t)}\right], \tag{3.10}$$

em que $\lim_{t\to\infty} h(t) = 0$ e $\int_0^\infty h(t)dt < \infty$. E a função de risco acumulada é dada por

$$H(t) = \frac{(1-\gamma)\mu}{\gamma} \log\left(\frac{1-\gamma S_0(t)}{1-\gamma}\right), \; t > 0. \tag{3.11}$$

Note que $\lim_{t\to 0} H(t) = 0$ e $\lim_{t\to\infty} H(t) = -\log(p_0)$ implicando que a função de risco acumulada do modelo proposto é limitada por $-\log(p_0)$, isto é, $H(t) \leq -\log(p_0)$. A função de sobrevivência (3.6) e a função de risco (3.11) do modelo considerando diferentes valores do parâmetro γ são apresentadas na Figura 2. Tem-se que a função de risco é descrescente para diferentes valores do parâmetro de γ.

Figura 2 – Função de sobrevivência (gráfico à esquerda) e função de risco (gráfico à direita) com função de risco base $h_0(t) = 2t$ e $\mu = 1$.



Fonte: Elaborada pelo autor.

3.2.3 Casos especiais

Este modelo proposto engloba alguns modelos da literatura como casos especiais. Quando $\gamma \longrightarrow 0$, o modelo fica reduzido ao modelo de fração de cura investigado por Yakovlev e Tsodikov (1996). Outros dois casos particulares são: (i) quando $\gamma < 0$ e $\mu = -\gamma(1-\gamma)$ tem-se o modelo de mistura com fração de cura apresentado por Boag (1949) e Berkson e Gage (1952); (ii) quando $0 < \gamma < 1$ e $\mu > 0$ tem-se o modelo proposto na estrutura do modelo de fração de cura com dispersão, investigado por Cancho, Rodrigues e Castro (2011). Na Tabela 1 apresentam-se a função de sobrevivência $S(\cdot)$, a função de risco imprópria $h(\cdot)$ e a correspondente fração de cura p_0 para os casos especiais do modelo proposto neste trabalho.

Tabela 1 – Função de sobrevivência, função de risco e fração de cura para os casos especiais do Modelo proposto.

Modelo	$S(t)$	$h(t)$	p_0
Promoção ($\gamma \longrightarrow 0$)	$e^{-\mu(1-S_0(t))}$	$\mu f_0(t)$	$e^{-\mu}$
Mistura ($\gamma<0$ e $\mu=-\gamma(1-\gamma)$)	$\frac{1}{1+(1-\gamma)\mu}+\frac{(1-\gamma)\mu}{1+(1-\gamma)\mu}S_0(t)$	$\frac{(1-\gamma)\mu}{1+(1-\gamma)\mu}f_0(t)$	$\frac{1}{1+(1-\gamma)\mu}$
Dispersão ($0<\gamma<1$ e $\mu>0$)	$\left[1+\frac{\gamma(1-S_0(t))}{1-\gamma}\right]^{-(1-\gamma)\mu/\gamma}$	$\frac{(1-\gamma)\mu f(t))}{1-\gamma S_0(t)}$	$[1-\gamma]^{-(1-\gamma)\mu/\gamma}$

Fonte: Elaborada pelo autor.

3.2.4 Propriedade de riscos proporcionais

O modelo com fragilidade discreta proposto em (3.6) fornece uma propriedade para a função de risco em (3.10). Especificamente, observa-se que a função $h(t)$ é multiplicativa em μ e $(1-\gamma)f_0(t)/(1-\gamma S_0(t))$, portanto carrega a estrutura de riscos proporcionais quando $\gamma < 1$ e as covariáveis são modeladas através de μ.

A estrutura de riscos proporcionais considera que a razão das taxas de falha de dois indivíduos distintos j e k independem do tempo t. A partir das estimativas de máxima verossimilhança é possível analisar a razão de risco (ou risco relativo) (RR) entre os indivíduos que receberam diferentes tratamentos. O risco relativo é usado para estimar a força da associação entre as condições dos indivíduos e os tratamentos. Tomando a razão das funções de taxa de risco dadas em (3.10), de dois indivíduos j e k, com função de ligação logarítmica para μ_j e μ_k tem-se

$$RR = \frac{h_j(t)}{h_k(t)} = \frac{\mu_j\left[\frac{(1-\gamma)f_0(t)}{1-\gamma S_0(t)}\right]}{\mu_k\left[\frac{(1-\gamma)f_0(t)}{1-\gamma S_0(t)}\right]} = \frac{\exp\{\boldsymbol{x}_j^\top\boldsymbol{\beta}\}}{\exp\{\boldsymbol{x}_k^\top\boldsymbol{\beta}\}} = \exp\{\boldsymbol{\beta}(\boldsymbol{x}_j^\top - \boldsymbol{x}_k^\top)\}, \tag{3.12}$$

em que $\exp\{\boldsymbol{\beta}\}$ representa o efeito multiplicativo da diferença $(\mathbf{x}_j^\top - \boldsymbol{x}_k^\top)$ no risco de morte, $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ denota o vetor de covariáveis e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ denota o correspondente vetor dos coeficientes de regressão, em que as covariáveis são incluídas através do parâmetro μ.

Relaciona-se μ com as covariáveis por $g(\mu_i) = \eta(\boldsymbol{x}_i;\boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta},\ i = 1,\ldots,n$, em que a função de ligação $g(\cdot)$ é uma função monótona e duas vezes diferenciável. Para este modelo, tem-se a função de ligação logarítmica dada por $g(\mu_i) = \log(\mu_i)$. Outras funções de ligação são discutidas por McCullagh e Nelder (1989) e Fox (2015).

Assim, $\exp\{\boldsymbol{\beta}\}$ pode ser facilmente interpretado em termos biológicos pois é o logaritmo do risco relativo das variáveis explicativas. Por exemplo, β_2 representa a mudança no logaritmo esperado da razão de risco para a mudança de uma unidade em x_2, mantendo todos os outros preditores constantes. Para uma discussão mais aprofundada das vantagens da estrutura de riscos proporcionais, ver Cox e Oakes (1996), Kalbfleisch e Prentice (1980). A importância da medida do efeito da razão de risco é apresentada por Ressing, Blettner e Klug (2010) em um estudo epidemiológico entre outras medidas.

3.2.5 Indivíduos sob risco

Em modelos com fração de cura tem-se uma parte dos indivíduos da população que não vai sofrer o evento de interesse e, por outro lado existe uma parte dessa população que está suscetível à ocorrência desse evento, a esse público específico chamamos de indivíduos sob risco.

Dessa forma, a função de sobrevivência para um indivíduo sob risco na população denotada por S_{R} pode ser obtida como $S_{\mathrm{R}}(t) = P(T > t \mid Z \geq 1)$ e expressa por

$$S_{\mathrm{R}}(t) = \frac{(1-\gamma)^{\frac{(1-\gamma)\mu}{\gamma}}(1-\gamma S_0(t))^{-\frac{(1-\gamma)\mu}{\gamma}} - [1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}}{1-[1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}},\ t > 0. \tag{3.13}$$

Observe que $S_{\mathrm{R}}(0) = 1$ e $S_{\mathrm{R}}(\infty) = 0$ implicando que a função de sobrevivência é própria. A função densidade de probabilidade própria para um indivíduo sob risco na população é apresentada por

$$f_R(t) = \frac{\frac{(1-\gamma)\mu f_0(t)}{1-\gamma S_0(t)}\left[\frac{1-\gamma}{1-\gamma S_0(t)}\right]^{\frac{(1-\gamma)\mu}{\gamma}}}{1-[1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}},\ t > 0. \tag{3.14}$$

A função de risco para os indivíduos sob risco na população é dada por,

$$h_R(t) = \left[\frac{(1-\gamma)\mu(1-\gamma)^{\frac{(1-\gamma)\mu}{\gamma}}(1-\gamma S_0(t))^{-\frac{(1-\gamma)\mu}{\gamma}}}{(1-\gamma)^{\frac{(1-\gamma)\mu}{\gamma}}(1-\gamma S_0(t))^{-\frac{(1-\gamma)\mu}{\gamma}} - [1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}}\right]\left[\frac{f_0(t)}{1-\gamma S_0(t)}\right],\ t > 0. \tag{3.15}$$

Portanto, (3.15) é multiplicado pelo fator $\frac{(1-\gamma)(1-\gamma)^{\frac{(1-\gamma)\mu}{\gamma}}(1-\gamma S_0(t))^{-\frac{(1-\gamma)\mu}{\gamma}}}{(1-\gamma)^{\frac{(1-\gamma)\mu}{\gamma}}(1-\gamma S_0(t))^{-\frac{(1-\gamma)\mu}{\gamma}} - [1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}} > 1$ comparado a função de risco $h(t)$ de toda a população. Além disso, pode-se mostrar que $\lim_{t\to\infty} h_R(t) = \frac{f_0(t)}{S_0(t)}$ e portanto, $h_R(t)$ converge para a função de risco base $h_0(t)$.

A relação entre o modelo proposto na Equação 3.6 e o modelo de mistura com fração de cura apresentado por Boag (1949) e Berkson e Gage (1952) é dada por

$$S(t) = p_0 + (1 - p_0)\, S_{\mathrm{R}}(t), \tag{3.16}$$

tal que a função de sobrevivência $S_{\mathrm{R}}(t)$ é definida pela Equação 3.13. Dessa forma $S(t)$ é um modelo de mistura com fração de cura igual a $p_0 = [1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}$ e função de sobrevivência $S_{\mathrm{R}}(t)$ para os indivíduos sob risco na população.

3.3 Inferência Clássica

Nesta seção apresenta-se uma abordagem clássica para o modelo proposto. Considerou-se o método de estimação de máxima verossimilhança via algoritmo EM. Para avaliar as incertezas dessas estimativas construiu-se os intervalos de confiança assintóticos para os parâmetros do modelo e os testes de hipóteses baseado no teste da razão de verossimilhanças.

3.3.1 Estimação pelo método de máxima verossimilhança

Considere que os tempos $T_1, \ldots, T_n$ podem ser observados ou não e estão sujeitos a censura à direita C_i, denotando os tempos de censura dos n indivíduos. Em seguida, considere $t_i = \min\{T_i, C_i\}$ e $\delta_i = \mathrm{I}(T_i \leq C_i)$, no qual

$$\delta_i = \begin{cases} 1, & \text{se } T_i \text{ (tempo de vida)} \\ 0, & \text{se } C_i \text{ (censura à direita)} \end{cases}$$

para todos os indivíduos $i = 1, \ldots, n$.

Para cada indivíduo i, tem-se o vetor de covariáveis $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ e o correspondente vetor de coeficientes de regressão desconhecidos, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, no qual tem-se as covariáveis através do parâmetro μ. Relaciona-se μ com as covariáveis por $g(\mu_i) = \eta(\boldsymbol{x}_i; \boldsymbol{\mu}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, $i = 1, \ldots, n$, em que a função de ligação $g(\cdot)$ é uma função monótona e duas vezes diferenciável. Nesse trabalho, assumiu-se a função logarítmica dada por $g(\mu_i) = \log(\mu_i)$, isto é, $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$. Considerando μ_i em (3.6) e (3.10), os coeficientes de regressão podem interpretar o papel dos grupos imunes e não-imunes.

A função de verossimilhança para o modelo com fragilidade Katz é dada por

$$L(\boldsymbol{\vartheta}; \boldsymbol{t}, \boldsymbol{\delta}, \boldsymbol{x}) \propto \prod_{i=1}^{n} f(t_i, \boldsymbol{x}_i; \boldsymbol{\vartheta})^{\delta_i} S(t_i, \boldsymbol{x}_i; \boldsymbol{\vartheta})^{1-\delta_i}, \tag{3.17}$$

em que $\boldsymbol{\vartheta} = (\gamma, \boldsymbol{\varphi}^\top, \boldsymbol{\beta}^\top)^\top$ é o conjunto de parâmetros do modelo.

O logaritmo da função de verossimilhança sob censura não-informativa dos n indivíduos independentes tem a forma

$$\begin{aligned}\ell(\boldsymbol{\vartheta}) &= \sum_{i=1}^{n}\delta_i\log(1-\gamma)+\sum_{i=1}^{n}\delta_i\log(\mu_i)+\sum_{i=1}^{n}\delta_i\log[f_0(t_i;\boldsymbol{\varphi})] \\ &- \sum_{i=1}^{n}(\delta_i+\frac{(1-\gamma)\mu_i}{\gamma})\log\left[1-\gamma S_0(t_i;\boldsymbol{\varphi})\right]+\frac{(1-\gamma)\log(1-\gamma)}{\gamma}\sum_{i=1}^{n}\mu_i,\end{aligned} \tag{3.18}$$

em que $\boldsymbol{\varphi}=(\varphi_1,\ldots,\varphi_q)^\top$ são os parâmetros da distribuição de base, $\boldsymbol{t}=(t_1,\ldots,t_n)^\top$ o vetor de tempos, $\boldsymbol{\delta}=(\delta_1,\ldots,\delta_n)^\top$ o vetor de censuras e $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^\top$ é a matriz de covariáveis.

Dada uma distribuição de base específica, o estimador de máxima verossimilhança (EMV) para o vetor de parâmetros $\boldsymbol{\vartheta}$ pode ser obtido pela maximização do logaritmo da função de verossimilhança (3.18) através de um procedimento de otimização.

Implementou-se o algoritmo EM (DEMPSTER; LAIRD; RUBIN, 1977) uma vez que a ideia do algoritmo EM é substituir uma maximização difícil por uma sequência de maximizações mais fáceis, envolvendo duas etapas: (i) Etapa "E" (*Step-E*) que calcula o valor esperado do logaritmo da função de verossimilhança completa e (ii) Etapa "M" (*Step-M*) que encontra seu máximo, sendo esse processo repetido até atingir a convergência. O algoritmo EM possui várias propriedades vantajosas, dentre elas a convergência estável sobre o método de Newton-Raphson como discutido por McLachlan e Krishnan (2007). Nesse trabalho, utilizou-se o método Broyden-Fletcher-Goldfarb-Shanno method (BFGS) para fazer as maximizações na etapa M, uma otimização baseada no método de Newton-Rapson (PRESS *et al.*, 2007)

3.3.1.1 Algoritmo EM

O i-ésimo elemento do conjunto de observações pode ser derivado de dois grupos diferentes, indivíduos em risco (suscetível) ou curado. Suponha uma variável latente Δ que indica este evento. Seja Δ_i a i-ésima variável latente dada como

$$\Delta_i=\begin{cases}1, & \text{se suscetível ,}\\ 0, & \text{se curado .}\end{cases}$$

A função de verossimilhança completa é dada por

$$L_c(\boldsymbol{\vartheta}) = \prod_{i\in\bar{C}}[1-p_0(\boldsymbol{x_i})]\prod_{i\in\bar{C}}f_R(t_i;\boldsymbol{x}_i,\boldsymbol{\vartheta})\prod_{i\in C}[p_0(\boldsymbol{x_i})]^{1-\Delta_i}\prod_{i\in C}[(1-p_0(\boldsymbol{x_i}))S_R(t_i;\boldsymbol{x}_i,\boldsymbol{\vartheta})]^{\Delta_i},$$

em que $S_R(\cdot)$ e $f_R(\cdot)$ são as funções de sobrevivência (3.13) e densidade de probabilidade (3.14) para cada indivíduo sob risco, p_0 é a fração de cura

$$p_0(\boldsymbol{x_i})=[1-\gamma]^{\frac{(1-\gamma)\exp\{\boldsymbol{x}_i^\top\boldsymbol{\beta}\}}{\gamma}}$$

e $C=\{i\in\{1,2,\ldots,n\}:\delta_i=0\}$ e $\bar{C}=\{i\in\{1,2,\ldots,n\}:\delta_i=1\}$ são os conjuntos de observações censuradas e não-censuradas, respectivamente.

Etapa E: Calcula-se a esperança da logarítmo da função de verossimilhança dos dados com relação à distribuição não observada Δ_i, dados os valores atuais do parâmetro e os dados observados $\boldsymbol{O}$. Note que Δ_i's são variáveis aleatórias Bernoulli na função de verossmilhança completa e calcula-se $\pi_i^{(k)} = E[\Delta_i|\boldsymbol{\vartheta}^{(k)}, \mathbf{O}]$, $i = 1\ldots,n$, em que $\boldsymbol{\vartheta}^{(k)}$ denota o valor do parâmetro atual na etapa de iteração k. Agora, para $i \in C$, tem-se

$$\pi_i^{(k)} = E[\Delta_i|\boldsymbol{\vartheta}^{(k)}, \boldsymbol{O}] = P[\Delta_i = 1|T_i > t_i] = \frac{[1-p_0(\boldsymbol{x_i})]S_R(t_i;\boldsymbol{x}_i,\boldsymbol{\vartheta})}{S_p(t_i;\boldsymbol{x}_i,\boldsymbol{\vartheta})}\,|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{(k)}},$$

e a esperança condicional do logarítmo da função de verossimilhança completa é dada por

$$Q(\boldsymbol{\vartheta}, \pi^{(k)}) = \sum_{i\in\bar{C}} \log[1-p_0(\boldsymbol{x_i})] + \sum_{i\in\bar{C}} \log f_R(y_i;\boldsymbol{x}_i,\boldsymbol{\vartheta}) + \sum_{i\in C}(1-\pi_i^{(k)})\log[p_0(\boldsymbol{x_i})] + \sum_{i\in C}\pi_i^{(k)}\log[(1-p_0(\boldsymbol{x_i}))] + \sum_{i\in C}\pi_i^{(k)}\log[S_R(y_i;\boldsymbol{x}_i,\boldsymbol{\vartheta})].$$

Etapa M: Maximiza-se a função $Q(\boldsymbol{\vartheta}, \boldsymbol{\pi}^{(k)})$ com relação à $\boldsymbol{\vartheta}$ sobre o correspondente espaço de parâmetros Θ dado $\boldsymbol{\pi}^{(k)}$, de modo a obter uma estimativa melhorada de $\boldsymbol{\vartheta}$ como

$$\boldsymbol{\vartheta}^{(k+1)} = \arg\max_{\boldsymbol{\vartheta}\in\Theta} Q(\boldsymbol{\vartheta}, \pi^{(k)}).$$

As etapas E e M são então continuadas iterativamente para obter as estimativas do parâmetro $\boldsymbol{\vartheta}$. Neste trabalho, como os EMVs de $\boldsymbol{\beta}$, γ e $\boldsymbol{\varphi}$ não possuem expressões explícitas, a etapa de maximização é realizada usando o algoritmo de gradiente EM (LANGE, 1995), que é um método de Newton–Raphson de uma etapa e é um caso especial do algoritmo EM generalizado (DEMPSTER; LAIRD; RUBIN, 1977).

3.3.2 Intervalo de confiança e teste de hipótese

Uma vez calculadas as estimativas pontuais dos parâmetros, tem-se interesse na incerteza dessas estimativas (intervalos de confiança) e na afirmação sobre um determinado parâmetro de interesse (teste de hipóteses). Sendo assim, a seguir apresenta-se o intervalo de confiança assintótico e o teste de hipótese.

Os erros padrões das estimativas são obtidos pela inversa da matriz de informação de Fisher observada. Sob condições de regularidade adequadas (MALLER; ZHOU, 1996), tem-se que a distribuição assintótica do EMV de $\widehat{\boldsymbol{\vartheta}}$ é uma distribuição normal multivariada com vetor de média $\boldsymbol{\vartheta}$ e matriz de covariância $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\vartheta}})$, que pode ser estimada por

$$\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\vartheta}}) = \left\{-\frac{\partial^2\ell(\boldsymbol{\vartheta})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\top}\right\}^{-1} = \{-J(\boldsymbol{\vartheta})\}^{-1},$$

avaliada em $\boldsymbol{\vartheta} = \widehat{\boldsymbol{\vartheta}}$.

Logo, o intervalo de confiança assintótico com coeficiênte de confiança α para cada parâmetro ϑ_r é dado por

$$\left(\hat{\vartheta}_r - z_{\alpha/2}\sqrt{\widehat{\Sigma}^{r,r}}, \hat{\vartheta}_r + z_{\alpha/2}\sqrt{\widehat{\Sigma}^{r,r}}\right),$$

em que $\widehat{\Sigma}^{r,r}$ é o r-ésimo elemento diagonal da matriz $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\vartheta}})$ estimado em $\widehat{\boldsymbol{\vartheta}}$, para $r = 1, \ldots, p + \dim(\boldsymbol{\varphi}) + 1$, em que $\dim(\cdot)$ é a dimensão do espaço paramétrico e $z_{\alpha/2}$ é o quantil $1 - \alpha/2$ da distribuição normal padrão.

Para se testar hipóteses da forma $H_0 : \boldsymbol{\vartheta}_1 = \boldsymbol{\vartheta}_{01}$ versus $H_1 : \boldsymbol{\vartheta}_1 \neq \boldsymbol{\vartheta}_{01}$, seja $\hat{\boldsymbol{\vartheta}}_0$ a estimativa que maximiza $\ell(\boldsymbol{\vartheta})$ restrita à hipóteste H_0. A estatística do Teste da Razão de Verossimilhanças (TRV) é dada por

$$\Lambda_n = 2\left[\ell(\hat{\boldsymbol{\vartheta}}) - \ell(\hat{\boldsymbol{\vartheta}}_0)\right], \tag{3.19}$$

em que $\ell(\cdot)$ é o logaritmo da função de verossimilhança. Sob H_0 e algumas condições de regularidade, tem-se que Λ_n converge para a distribuição qui-quadrado com $\nu = \dim(\boldsymbol{\vartheta}_1)$ graus de liberdade.

Outra hipótese de interesse está relacionada à adequação do modelo de cura por tempo de promoção ($\mathrm{H}_0 : \gamma = 0$) versus à não adequação ($\mathrm{H}_1 : \gamma > 0$). A distribuição do teste TRV sob H_0 não é padrão (SELF; LIANG, 1987) e pode ser aproximada por uma mistura de 50-50 da distribuição qui-quadrado com 1 grau de liberdade (χ^2_1) e uma distribuição degenerada em zero. Ou seja, a estatística Λ_n converge para a distribuição $0,5 + 0,5P\left[\chi^2_1 \leq x\right]$.

3.4 Estudos de simulação

Nesta seção apresenta-se o estudo de simulação via algoritmo EM para avaliar o desempenho do método de estimação do modelo proposto. Para a implementação computacional utilizou-se a Linguagem R (R Core Team, 2019), as estimativas de máxima verossimilhança foram obtidas via método *L-BFGS-B* atráves da rotina *optim*.

Por simplicidade no processo de simulação, considerou-se a distribuição exponencial com risco de base $h_0(t) = \lambda = 1$. Para o parâmetro γ da distribuição de fragilidade Katz, tem-se os valores ($\gamma = 0,1; 0,3$ e $0,8$) e uma covariável reparametrizada na média via função de ligação $\log(\mu_i) = \beta_0 + \beta_1 x_i$ onde $\beta_0 = -0,5$ e $\beta_1 = 0,7$, $i = 1, 2, ..., n$, aqui a covariável x_i é gerada a partir de uma distribuição de Bernoulli com probabilidade de $0,5$. Os tempos de censura C_i foram amostrados das distribuições uniformes no intervalo $(0, \tau)$, onde τ é definido para controlar a proporção de observações censuradas. Nessa simulação, tem-se $\tau = 5$, levando a uma proporção de observações censuradas em torno de 60%. Para cada tamanho de amostras $n = 100$, $n = 200$, 500 e 1000, simulou-se $Q = 1000$ amostras do modelo proposto. Os tempos observados foram gerados da seguinte forma:

1. Gera-se um valor de uma distribuição uniforme, $u_i \sim U(0; 1)$;
2. Se $u_i < p_{0i} = (1-\gamma)^{\frac{\mu_i(1-\gamma)}{\gamma}}$, então $t_i = \infty$. Senão $t_i = F^{-1}\left(\frac{(\gamma-1)}{\gamma}\left[1 - \frac{1}{u_i^{\gamma/\mu_i(1-\gamma)}}\right]; \lambda\right)$, tal que $F^{-1}(\cdot\,;\,\lambda)$ é o quantil da distribuição Exponencial com parâmetro $\lambda = 1$;

3. Gera-se o tempo de censura C_i de uma uniforme $U(0, \tau_i)$;

4. Se $T_i \leq C_i$, então o tempo $t_i = T_i$ e $\delta i = 1$. Senão $t_i = C_i$ e $\delta_i = 0$.

Para os dados simulados calculou-se a média amostral do parâmetro (Media), o desvio padrão (DP) das estimativas, a estimativa de viés (Vies), a raiz do erro quadrático médio (REQM) e as probabilidades de cobertura (PC) dos intervalos de confiança de 95% para os parâmetros do modelo.

$$Media(\hat{\vartheta}_r) = \bar{\vartheta}_r = \frac{1}{Q}\sum_{q=1}^{Q}\hat{\vartheta}_{rq} \quad DP(\hat{\vartheta}_r) = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}(\hat{\vartheta}_{rq} - \bar{\vartheta}_r)^2},$$

$$Vies(\hat{\vartheta}_r) = \frac{1}{Q}\sum_{q=1}^{Q}(\hat{\vartheta}_{rq} - \vartheta_r) \quad REQM(\hat{\vartheta}_r) = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}(\hat{\vartheta}_{rq} - \vartheta_r)^2},$$

tal que ϑ_r é o r-ésimo elemento do vetor de parâmetros $\boldsymbol{\vartheta}$ e $\hat{\vartheta}_r$ é a respectiva estimativa de máxima verossimilhança.

Os resultados da simulação são apresentados na Tabela 2. Observa-se que as estimativas médias estão próximas dos verdadeiros valores dos parâmetros. O viés, o desvio padrão e o REQM diminuem à medida que o tamanho da amostra aumenta. Os níveis de cobertura nominal aumentam a medida que n aumenta. Esses resultados são esperados se o esquema de estimativa subjacente estiver funcionando corretamente para produzir estimativas consistentes e assintoticamente normais.

Tabela 2 – Resultados da simulação para o modelo proposto considerando diferentes tamanhos de amostras e diferentes valores para o parâmetro γ da distribuição de fragilidade Katz.

		$\gamma = 0,1$				$\gamma = 0,3$				$\gamma = 0,8$			
n		γ	λ	β_0	β_1	γ	λ	β_0	β_1	γ	λ	β_0	β_1
	Media	0,280	0,885	-0,165	0,711	0,321	0,950	-0,241	0,711	0,555	1,263	-0,580	0,723
	DP	0,349	0,371	0,738	0,301	0,364	0,407	0,795	0,316	0,377	0,714	0,963	0,424
100	VIES	0,180	-0,115	0,335	0,011	0,021	-0,050	0,259	0,011	-0,245	0,263	-0,080	0,023
	REQM	0,392	0,388	0,810	0,301	0,364	0,410	0,836	0,316	0,450	0,760	0,966	0,424
	PC	0,881	0,954	0,996	0,963	0,881	0,954	0,992	0,956	0,892	0,952	0,965	0,960
	Media	0,241	0,907	-0,287	0,707	0,306	0,950	-0,335	0,707	0,654	1,133	-0,523	0,713
	DP	0,317	0,276	0,519	0,211	0,337	0,318	0,576	0,219	0,309	0,520	0,754	0,292
200	VIES	0.141	-0,093	0,213	0,007	0,006	-0,050	0,165	0,007	-0,146	0,133	-0,023	0,013
	REQM	0,347	0,292	0,561	0,211	0,336	0,321	0,599	0,219	0,341	0,536	0,754	0,292
	PC	0,900	0,981	0,999	0,955	0,901	0,967	0,998	0,951	0,914	0,966	0,977	0,951
	Media	0,208	0,939	-0,378	0,700	0,305	0,973	-0,419	0,698	0,729	1,067	-0,518	0,700
	DP	0,258	0,193	0,292	0,138	0,279	0,217	0,339	0,143	0,197	0,339	0,496	0,186
500	VIES	0,108	-0,061	0,122	-0,000	0,005	-0,027	0,081	-0,002	-0,071	0,067	-0,018	0,000
	REQM	0,279	0,202	0,316	0,138	0,279	0,219	0,348	0,143	0,209	0,346	0,496	0,186
	PC	0,912	0,973	0,998	0,945	0,901	0,949	0,949	0,940	0,904	0,953	0,953	0,946
	Media	0,181	0,959	-0,423	0,698	0,292	0,990	-0,461	0,698	0,772	1,027	-0,496	0,702
	DP	0,207	0,142	0,199	0,091	0,236	0,168	0,239	0,093	0,119	0,238	0,349	0,125
1000	VIES	0,081	-0,041	0,077	-0,002	-0,008	-0,010	0,039	-0,002	-0,028	0,027	0,004	0,002
	REQM	0,222	0,148	0,213	0,091	0,236	0,168	0,242	0,093	0,122	0,240	0,349	0,125
	PC	0,919	0,978	0,993	0,963	0,918	0,970	0,986	0,959	0,916	0,954	0,948	0,953

Fonte: Elaborada pelo autor.

3.5 Aplicação - Câncer Cervical

Nesta seção, apresenta-se uma aplicação do modelo proposto, conforme detalhado na Seção 3.1. O conjunto de dados foi fornecido pela Fundação Oncocentro de São Paulo (FOSP), que coordena o Registro Hospitalar de Câncer do Estado de São Paulo. Os dados foram coletados a partir de um levantamento de prontuários de pacientes diagnosticadas com câncer do colo do útero, também denominado câncer cervival, no estado de São Paulo, Brasil, entre 2000 e 2005, com acompanhamento realizado até 2019.

Este tipo de câncer é o quarto mais frequente em mulheres no mundo, com uma estimativa de 570 mil novos casos em 2018 (WHO, 2020). Mais de 95% desses cânceres estão relacionados ao HPV (sigla em inglês para Papilomavírus Humano) um vírus que ataca o sistema reprodutivo e é sexualmente transmissível. Embora a maioria das infecções possa ser tratada, e por vezes até eliminada pelo próprio sistema imunológico, o HPV é capaz de causar mutações que induzem à formação de tumores malignos. Entre os tratamentos para esse tipo de câncer estão a cirurgia, quimioterapia e radioterapia, dependendo do estadiamento da doença e de fatores pessoais.

O conjunto de dados é referente a um teste da eficiência do tratamento por quimioterapia, radioterapia e quimioterapia + radioterapia para prevenir a recorrência da doença. O conjunto de dados inclui 2489 pacientes com diagnóstico de câncer do colo do útero, com aproximadamente 83% de censura (pacientes que não apresentaram a doença ou abandonaram o tratamento durante o período de acompanhamento).

A variável resposta T é o tempo até a ocorrência da recidiva da doença. Os efeitos dos tratamentos na proporção de cura e no tempo de recidiva foram examinados. Para cada paciente foram coletadas as seguintes variáveis:

t_i : tempo de recidiva (em anos) ou censurado;

x_{i1} : idade em anos (≤ 60 anos e > 60 anos);

x_{i2} : cirurgia (0=não, 1=sim);

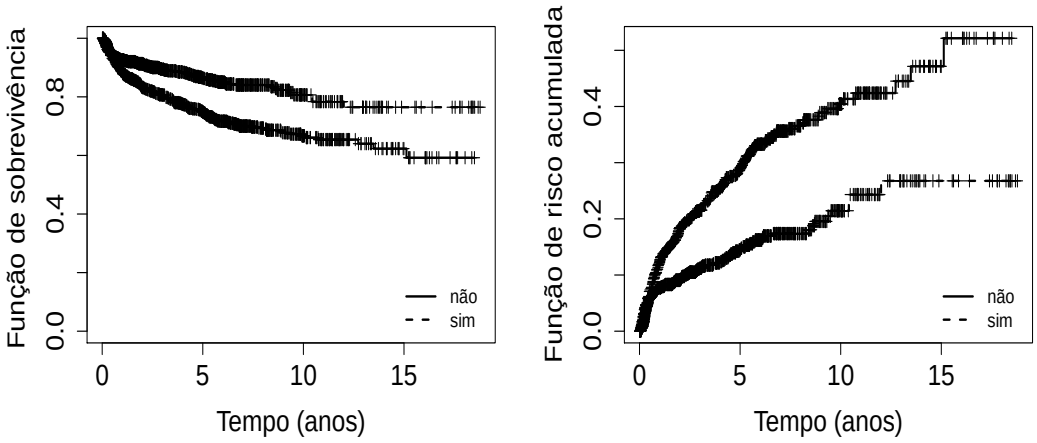
x_{i3} : tratamento (Quimioterapia (Q), Radioterapia (R) e Quimioterapia+Radioterapia (Q+R));

x_{i4} : grupo de estadiamento clínico: (Estágio I, Estágio II+III e Estágio IV).

Para os dados coletados tem-se um tempo médio de recidiva da doença de $2,13$ anos com um desvio padrão de $2,54$ anos. Além disso, $82,7\%$ dos pacientes possuem idade ≤ 60 e apenas $17,3\%$ têm idade superior a 60 anos. Quanto a cirurgia, 56% dos pacientes passaram por procedimento cirúrgico contra 44% que não fizeram. Em relação ao tipo de tratamento tem-se 83% que receberam apenas radioterapia, $12,5\%$ receberam ambos os tratamentos, quimioterapia e radioterapia, e apenas $4,5\%$ dos pacientes receberam quimioterapia. E por fim, tem-se 59% dos pacientes no Estágio I da doença, $34,5\%$ no Estágio II+III e $6,5\%$ no Estágio IV.

As estimativas de Kaplan-Meier (KM) da função de sobrevivência na Figura 3 (gráfico à esquerda) indicam a existência de pacientes livres de doença. As estimativas de KM da função de risco acumulada na Figura 3 (gráfico à direita) são limitadas e côncavas e, portanto, uma distribuição com função de risco monótona pode ser adequada para modelar esses dados.

Figura 3 – Estimativa Kaplan-Meier da função de sobrevivência (gráfico à esquerda) e função de risco acumulada (gráfico à direita) para os dados de câncer do colo do útero estratificados por cirurgia.



Fonte: Elaborada pelo autor.

Assim, considerou-se a distribuição Weibull com função de sobrevivência, $S_0(t;\boldsymbol{\varphi}) = \exp\{-\exp\{\varphi_2\}t^{\varphi_1}\}$, $\varphi_1 > 0$ e $\varphi_2 \in R$ para a função de sobrevivência de base na Equação 3.6. E para o tempo de recidiva do câncer cervical, as covariáveis do modelo, descrito na Seção 3.1, foram reparametrizadas na média μ:

$$\mu_i = \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\} = \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{3_1} x_{i3_1} + \beta_{3_2} x_{i3_2} + \beta_{4_1} x_{i4_1} + \beta_{4_2} x_{i4_2}\}, \qquad (3.20)$$

em que $i = 1, \ldots, 2489$.

Para as covariáveis categóricas com mais de dois níveis (tratamento e estadiamento clínico), definiu-se variáveis *dummies*. Para a covariável tratamento (x_{i3}), tem-se:

$$x_{i3_1} = \begin{cases} 1, & \text{se tratamento Q+R} \\ 0, & \text{se caso contrário} \end{cases} \quad \text{e} \quad x_{i3_2} = \begin{cases} 1, & \text{se tratamento R} \\ 0, & \text{se caso contrário} \end{cases}$$

e para o estadiamento clínico (x_{i4}), tem-se:

$$x_{i4_1} = \begin{cases} 1, & \text{se Estágio II + III} \\ 0, & \text{se caso contrário} \end{cases} \quad \text{e} \quad x_{i4_2} = \begin{cases} 1, & \text{se Estágio IV} \\ 0, & \text{se caso contrário} \end{cases}$$

As estimativas de máxima verossimilhança dos parâmetros do modelo estão listadas na Tabela 3, em que, LI e LS são os limites, limite inferior (LI) e limite superior (LS) do intervalo de confiança, respectivamente.

Tabela 3 – EMV e intervalo de confiança (IC) de 95% para o modelo de fração de cura com dispersão.

Parâmetro	Estimativa	IC (95%)		Erro Padrão	P-valor
		LI	LS		
γ (dispersão - fragilidade)	0,989	0,979	0,999	0,005	—
φ_1 (forma - Weibull)	1,214	1,019	1,409	0,099	—
φ_2 (escala - Weibull)	-3,820	-4,534	-3,107	0,364	—
β_0 (intercepto)	2,328	1,493	3,164	0,426	0,000
β_1 (idade >60)	0,240	0,005	0,474	0,120	0,045
β_2 (cirurgia - sim)	-0,364	-0,614	-0,114	0,128	0,004
β_{3_1} (tratamento - Q+R)	-0,400	-0,727	-0,074	0,167	0,016
β_{3_2} (tratamento - R)	-1,143	-1,589	-0,697	0,228	0,000
β_{4_1} (estágio - II+III)	0,348	0,080	0,616	0,137	0,011
β_{4_2} (estágio - IV)	1,170	0,848	1,491	0,164	0,000

Fonte: Elaborada pelo autor.

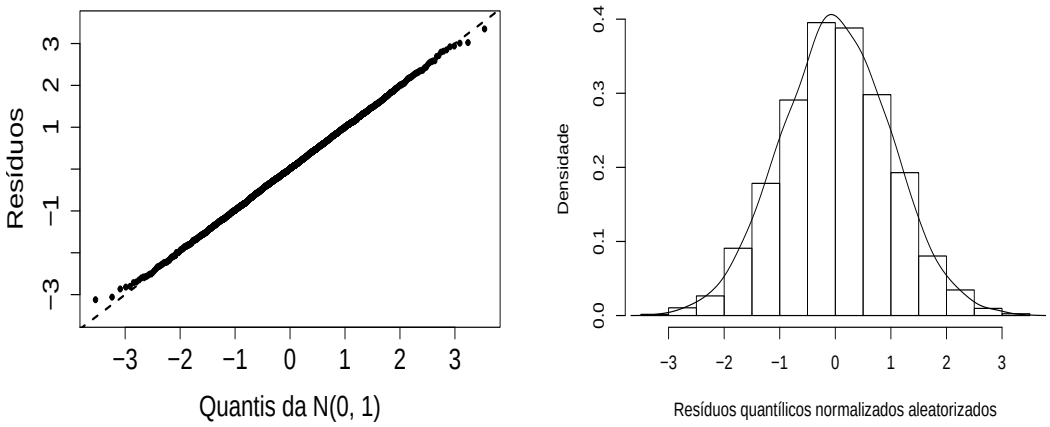
Todas as covariáveis são estatisticamente significativas a um coeficiente de confiança de 5%. A partir da estimativa do parâmetro de dispersão γ tem-se a estimativa da razão da variância com a média ($\kappa = (1-\gamma)^{-1}$) dos riscos latentes igual à $\kappa = 90,91$, implicando na sobredispersão da variável de fragilidade.

Além disso, calculou-se a estatística LR (Λ_n) para testar a adequação do modelo proposto para este conjunto de dados, ou seja, $H_0 : \gamma = 0$ versus $H_1 : \gamma > 0$. Uma vez que Λ_n é igual a $17,361$, com um p-valor $< 1,55 \times 10^{-5}$, o que fornece fortes evidências a favor de H_1, indicando que este modelo é adequado para este conjunto de dados. O gráfico QQ dos resíduos quantílicos normalizados aleatorizados (DUNN; SMYTH, 1996; RIGBY; STASINOPOULOS, 2005) na Figura 4 sugere que o este modelo com fragilidade discreta para a heterogeneidade não observada tem ajuste aceitável.

Observe que as quatro covariáveis (x_1, x_2, x_3 e x_4) são estatisticamente significativas, veja Tabela 3. Consequentemente, há uma diferença significativa entre os níveis dessas covariáveis quanto à proporção de cura (p_0) dos indivíduos e quanto ao risco de recidiva da doença. Assim, pacientes com idade superior a 60 anos tiveram pior prognóstico da doença ($\beta_1 = 0,240 > 0$) o que indica que o risco de recidiva da doença desses pacientes é $\exp\{0,24\} = 1,27$ vezes maior que para pacientes com idade inferior a 60 anos. Pacientes com cirurgia têm melhor prognóstico ($\beta_2 = -0,364 < 0$), ou seja, têm maior tempo de sobrevida. As estimativas dos coeficientes de regressão associados ao tratamento são todas negativas ($\beta_{3_1} = -0,400$ e $\beta_{3_2} = -1,143$) e indicam que os pacientes cujo protocolo de tratamento incluiu Q+R combinado e R sozinho têm menor risco de recidiva da doença do que pacientes que receberam apenas Q. Pacientes com estágio II+III ou estágio IV da doença ($\beta_{4_1} = 0,343$ e $\beta_{4_2} = 1,170$, respectivamente) têm maior risco de recidiva da doença do que os pacientes no estágio I.

A Figura 5 mostra as curvas de Kaplan-Meier para cada covariável estratificada pelos níveis, e todas mostraram ter diferenças significativas na proporção de curados. Para o tratamento,

Figura 4 – QQ-plot dos resíduos quantílicos normalizados aleatorizados (gráfico à esquerda) e histrograma (gráfico à direita) para o ajuste do modelo proposto nos dados de câncer cervical.



Fonte: Elaborada pelo autor.

tem-se bastante diferença na taxa de curados de quem recebeu apenas Q contra quem recebeu R ou Q+R, a taxa de curados para quem recebeu apenas R é bem maior uma vez que esse tipo de tratamente é menos agressivo para o paciente. Os estágios também mostraram ter bastante diferença na proporção de curados.

A partir das EMVs, também pode-se estimar a probabilidade dos indivíduos ficarem livres da doença (proporção de curados) após o acompanhamento, $t > 0$, dado por:

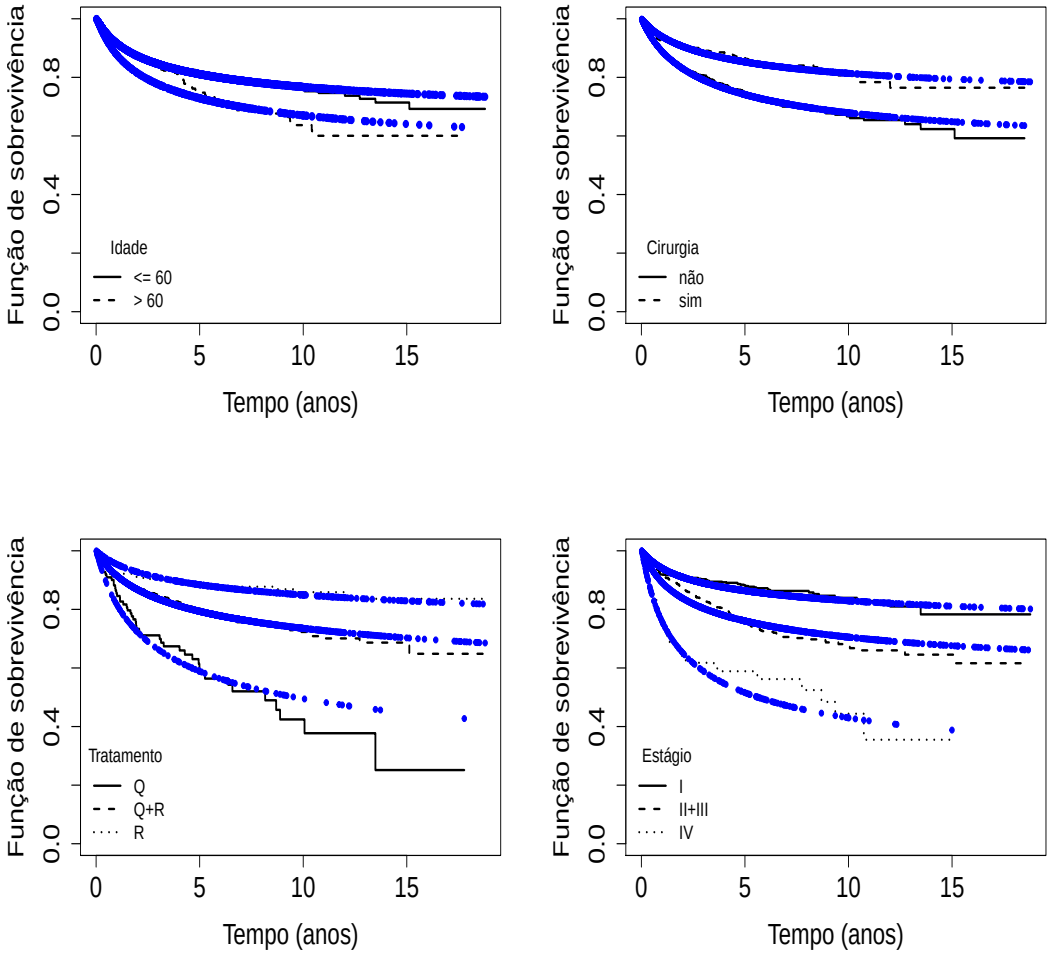
$$\pi(t) = [1 - \gamma \exp\{-\exp\{\varphi_2\} t^{\varphi_1}\}]^{\frac{\mu(1-\gamma)}{\gamma}},$$

em que μ é dado em (3.20). Assim, estimou-se a proporção de pacientes livres de doença ($\pi(0) = p_0$), para todas as combinações de tratamentos. Na Tabela 4 apresentam-se as estimativas e os intervalos de confiança assintóticos de 95% para a proporção de pacientes livres de doença (taxa de cura) após o tratamento ($t = 0$) e após dois anos $t = 2$, $\pi(2)$.

Como exemplo de interpretação para a Tabela 4, considerou-se a proporção de pacientes livres da doença na mesma faixa etária (≤ 60), e que não passaram por processo cirúrgico (cirurgia = não) e estão nos estágios I e IV da doença foram de $0,591$ e $0,184$, respectivamente. As probabilidades desses pacientes estarem livres da doença após dois anos, $\pi(2)$, são $0,719$ e $0,346$, implicando que em pacientes com doença em estágio I, a fração de cura é maior do que em pacientes nos estágios II+III e IV, assim como a probabilidade de estarem livres da doença após dois anos.

Analisando a Tabela 4 de forma geral, as estimativas de p_0 e $\pi(2)$ revelam que os pacientes tratados com quimioterapia+radioterapia apresentaram uma probabilidade de estarem livres da doença após dois anos maior do que os pacientes tratados com quimioterapia, indicando a eficácia do tratamento conjunto. Também diferem na proporção de pacientes sem doença, pois

Figura 5 – Curvas de Kaplan–Meier para cada covariável estratificada pelos níveis (linhas) e a estimativa da função de sobrevivência (pontos) para o modelo proposto ajustada aos dados de câncer cervical.



Fonte: Elaborada pelo autor.

em pacientes com doença em estágio I, a fração de cura é maior do que em pacientes nos estágios II+III e IV.

Para avaliar o ajuste do modelo Katz considerou-se estatísticas Log-verossimilhança e Critério da Informação de Akaike (AIC) para comparação com outros modelos da literatura apresentadas. Na Tabela 5 tem-se que o este modelo mostrou melhores estatísticas quando comparado aos modelos BCH, BCH-Gama, BCH-Inversa Gaussiana.

Tabela 4 – Estimativas EMVs da proporção de curados e intervalo de confiança de 95% (IC), para pacientes com câncer cervical e após dois anos de acompanhamento, estratificado pelas covariáveis: estágio, tramento, cirurgia e idade.

Estágio	Tratamento	Cirurgia	Idade	p_0		$\pi(2)$	
				Média	IC(95%)	Média	IC(95%)
I	Q	não	≤ 60	**0,591**	(0,447; 0,735)	**0,719**	(0,586; 0,852)
I	Q	não	> 60	0,513	(0,344; 0,681)	0,658	(0,497; 0,819)
I	Q	sim	≤ 60	0,694	(0,584; 0,804)	0,795	(0,698; 0,893)
I	Q	sim	> 60	0,629	(0,491; 0,766)	0,748	(0,624; 0,871)
I	Q+R	não	≤ 60	0,703	(0,616; 0,790)	0,802	(0,719; 0,885)
I	Q+R	não	> 60	0,639	(0,525; 0,753)	0,755	(0,648; 0,863)
I	Q+R	sim	≤ 60	0,783	(0,726; 0,840)	0,858	(0,801; 0,915)
I	Q+R	sim	> 60	0,733	(0,650; 0,815)	0,823	(0,746; 0,900)
I	R	não	≤ 60	0,846	(0,783; 0,909)	0,900	(0,849; 0,951)
I	R	não	> 60	0,808	(0,729; 0,887)	0,875	(0,811; 0,939)
I	R	sim	≤ 60	0,890	(0,846; 0,934)	0,930	(0,894; 0,965)
I	R	sim	> 60	0,862	(0,805; 0,920)	0,911	(0,866; 0,957)
II+III	Q	não	≤ 60	0,475	(0,333; 0,617)	0,627	(0,478; 0,777)
II+III	Q	não	> 60	0,388	(0,233; 0,543)	0,553	(0,381; 0,725)
II+III	Q	sim	≤ 60	0,596	(0,454; 0,738)	0,723	(0,593; 0,854)
II+III	Q	sim	> 60	0,518	(0,353; 0,683)	0,662	(0,505; 0,820)
II+III	Q+R	não	≤ 60	0,607	(0,523; 0,692)	0,732	(0,636; 0,827)
II+III	Q+R	não	> 60	0,530	(0,423; 0,638)	0,672	(0,553; 0,791)
II+III	Q+R	sim	≤ 60	0,707	(0,622; 0,792)	0,805	(0,723; 0,886)
II+III	Q+R	sim	> 60	0,644	(0,534; 0,754)	0,759	(0,654; 0,863)
II+III	R	não	≤ 60	0,789	(0,712; 0,865)	0,862	(0,797; 0,927)
II+III	R	não	> 60	0,740	(0,648; 0,831)	0,828	(0,748; 0,908)
II+III	R	sim	≤ 60	0,848	(0,782; 0,914)	0,902	(0,849; 0,954)
II+III	R	sim	> 60	0,811	(0,729; 0,893)	0,877	(0,811; 0,943)
IV	Q	não	≤ 60	**0,184**	(0,049; 0,319)	**0,346**	(0,149; 0,544)
IV	Q	não	> 60	0,116	(0,001; 0,231)	0,260	(0,064; 0,456)
IV	Q	sim	≤ 60	0,308	(0,138; 0,479)	0,478	(0,278; 0,679)
IV	Q	sim	> 60	0,224	(0,056; 0,393)	0,392	(0,173; 0,611)
IV	Q+R	não	≤ 60	0,322	(0,192; 0,451)	0,491	(0,324; 0,658)
IV	Q+R	não	> 60	0,236	(0,101; 0,372)	0,405	(0,218; 0,592)
IV	Q+R	sim	≤ 60	0,454	(0,315; 0,594)	0,610	(0,458; 0,762)
IV	Q+R	sim	> 60	0,367	(0,206; 0,528)	0,534	(0,352; 0,716)
IV	R	não	≤ 60	0,583	(0,438; 0,728)	0,713	(0,579; 0,847)
IV	R	não	> 60	0,503	(0,340; 0,667)	0,651	(0,493; 0,808)
IV	R	sim	≤ 60	0,687	(0,557; 0,817)	0,791	(0,681; 0,900)
IV	R	sim	> 60	0,621	(0,466; 0,775)	0,742	(0,607; 0,876)

Fonte: Elaborada pelo autor.

Tabela 5 – Estatísticas para comparar o modelo Katz com os outros modelos propostos na literatura.

Modelo	Log-verossimilhança	AIC
Modelo Katz	-1553,448	3126,896
BCH	-1562,120 8	3142,256
BCH-Gama	1562,130	3144,259
BCH-Inversa Gaussiana	-1562,143	3144,286

Fonte: Elaborada pelo autor.

3.6 Comentários Finais

Neste capítulo apresentou-se um novo modelo de sobrevivência induzido por fragilidade discreta Katz com heterogeneidade não observada. O procedimento inferencial foi baseado na abordagem de máxima verossimilhança via algoritmo EM e os estudos de simulação mostraram estimativas de parâmetros eficientes, uma vez que este procedimento não apresentou problemas numéricos, como problemas de convergência. Assim, o modelo proposto pode ser uma alternativa aos modelos existentes, uma vez que engloba alguns modelos como casos particulares, além da vantagem de ter a estrutura de riscos proporcionais e levar em conta a sobredispersão, equidispersão e subdispersão. Além disso, a importância do modelo foi validada na aplicação em dados de câncer cervical, uma vez que o modelo apresentou um valor de $AIC = 3126,896$ enquanto que o modelo BCH apresentou um $AIC = 3142,256$, dando indícios que o modelo proposto conseguiu modelar a heterogeneidade não observada dos dados.

Como resultado desse Capítulo, um artigo intitulado "*A Survival Model for Lifetime with Long-Term Survivors and Unobserved Heteregeneity*" foi aceito para publicação na revista *Brazilian Journal of Probability and Statistics* (BJPS).

CAPÍTULO

4

MODELO DE SOBREVIVÊNCIA COM DISTRIBUIÇÃO DE FRAGILIDADE POISSON GENERALIZADA

Neste capítulo, tem-se o modelo de sobrevivência induzido por fragilidade discreta com distribuição Poisson Generalizada (CONSUL; JAIN, 1973). Os estudos do modelo proposto neste Capítulo, incluindo a estimação dos parâmetros, intervalos de confiança, testes de hipóteses, simulações e a aplicação no conjunto de dados reais foram desenvolvidos sob a abordagem clássica e bayesiana.

4.1 Modelo com fragilidade Poisson Generalizada

4.1.1 Distribuição Poisson Generalizada

Segundo Consul e Jain (1973), uma variável aleatória Z com distribuição Poisson Generalizada possui função massa de probabilidade (f.d.p.) definida por

$$P(Z=z)=\begin{cases}\frac{\theta(\theta+z\gamma)^{z-1}}{z!}\exp\left(-\theta-z\gamma\right), & \text{para z = 0,1,2,...}\\ 0, & \text{caso contrário}\end{cases} \tag{4.1}$$

em que $\theta>0$, $\max(-1,-\frac{\theta}{m})\leq\gamma\leq 1$ e $m\,(\geq 4)$ é o maior inteiro positivo para o qual $\theta+\gamma m>0$ se $\gamma<0$. A média é a variância de Z são $E(Z)=\mu=\theta/(1-\gamma)$ e $Var(Z)=\theta/(1-\gamma)^3$, respectivamente.

A distribuição Poisson Generalizada foi introduzida como um elemento da classe Lagrangiana de distribuições (CONSUL; SHENTON, 1972). Algumas propriedades, inferências e várias aplicações deste modelo em biologia, ecologia e outras áreas foram sugeridos por Consul (1989) e em seguida Consul (1990).

Pela relação entre a média e variância tem-se a sobredispersão para o parâmetro $\gamma > 0$ e para $\gamma < 0$ tem-se subdispersão. Quando o parâmetro $\gamma = 0$ a distribuição Poisson Generalizada se reduz a distribuição de Poisson e, portanto, apresenta equidispersão. Logo o parâmetro γ pode ser interpretado como um parâmetro que controla a dispersão.

4.1.2 Formulação do modelo

Considere a distribuição Poisson Generalizada reparametrizada na média μ, ou seja, $E(Z) = \mu$ e $Var(Z) = \mu/(1-\gamma)^2$. Sob essa reparametrização tem-se que a função geradora de probabilidade de Z (AMBAGASPITIYA; BALAKRISHNAN, 1994) é definida por:

$$G_Z(s) = \exp\left\{-\frac{(1-\gamma)}{\gamma}\mu\left[W(-\gamma e^{-\gamma}s)+\gamma\right]\right\}, \quad 0 \leq s \leq 1, \tag{4.2}$$

em que $\mu > 0$, $0 < \gamma < 1$ e W é uma função Lambert's (Corless *et al.* (1996) ; Jodrá (2010) definida por $W(x)\,e^{W(x)} = x$.

Dessa forma, o novo modelo de sobrevivência induzido por fragilidade discreta é obtido quando assume-se que a variável de fragilidade Z descrita na Equação 3.4 segue uma distribuição Poisson Generalizada dada em (4.1) com respectiva função geradora de probabilidade dada em (4.2). Assim, a função de sobrevivência é dada por

$$S(t) = \exp\left\{-\frac{(1-\gamma)\mu}{\gamma}\left[W(-\gamma e^{-\gamma}S_0(t))+\gamma\right]\right\}. \tag{4.3}$$

e a fração de curados é dada por $p_0 = \lim_{t\to\infty} S(t)$, tal que

$$p_0 = e^{-(1-\gamma)\mu} > 0, \tag{4.4}$$

implicando que a Equação 4.3 é uma função de sobrevivência imprópria.

A função de probabilidade

$$\pi(t) = P(Z = 0|T > t) = \exp\left\{\frac{(1-\gamma)\mu}{\gamma}\left[W(-\gamma e^{-\gamma}S_0(t))+\gamma\right] - (1-\gamma)\mu\right\}, \tag{4.5}$$

denota a probabilidade de um indivíduo estar imune ou curado da doença dado que ele sobreviveu por um tempo $t > 0$ após o tratamento. A Equação 4.5 é uma função crescente em t e, claramente, $\pi(0) = p_0$ e $\lim_{t\to\infty}\pi(0) = 1$.

A correspondente função densidade de probabilidade é dada por

$$f(t) = \frac{\mu(\gamma-1)h_0(t)}{\gamma}\left[\frac{W(-\gamma e^{-\gamma}S_0(t))}{1+W(-\gamma e^{-\gamma}S_0(t))}\right]\exp\left\{-\frac{(1-\gamma)\mu}{\gamma}\left[W(-\gamma e^{-\gamma}S_0(t))+\gamma\right]\right\}. \tag{4.6}$$

em que $f_0(t) = -dS_0(t)/dt$. A função de risco de T é dada por

$$h(t) = \frac{\mu(\gamma-1)}{\gamma}h_0(t)\left[\frac{W(-\gamma e^{-\gamma}S_0(t))}{1+W(-\gamma e^{-\gamma}S_0(t))}\right]. \tag{4.7}$$

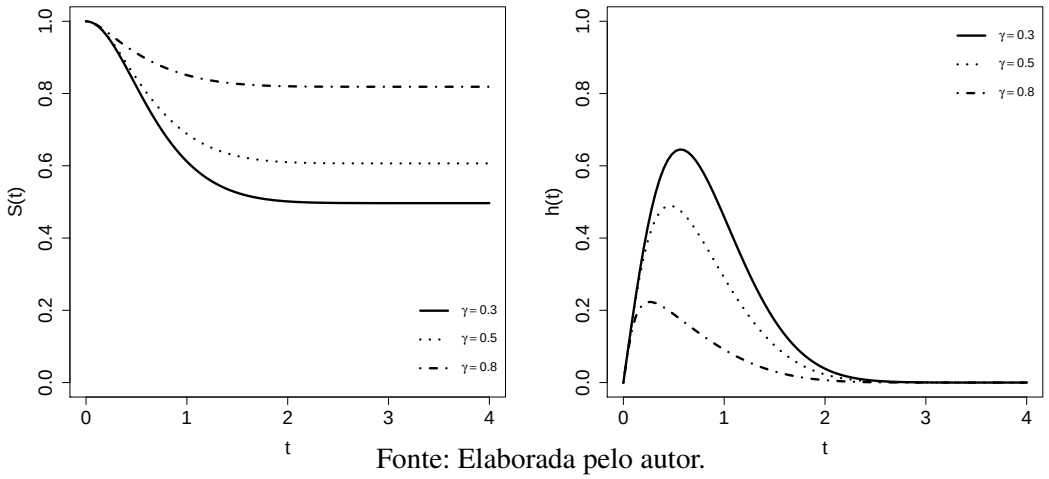
Note que $\lim_{t\to\infty} h(t) = 0$ e $\int_0^\infty h(t)dt < \infty$.

A função de risco acumulada de T é

$$H(t) = \frac{(1-\gamma)\mu}{\gamma}\left[W(-\gamma e^{-\gamma}S_0(t)) + \gamma\right], t > 0,$$

em que, a função $\lim_{t\to 0} H(t) = 0$ e $\lim_{t\to\infty} H(t) = -\log(p_0)$, implicando que $H(t) \leq (1-\gamma)\mu$. A função de sobrevivência (3.6) e a função de risco (4.7) do modelo considerando diferentes valores de γ são apresentadas na Figura 6.

Figura 6 – Função de sobrevivência (gráfico à esquerda) e função de risco (gráfico à direita) com função de risco base $h(t) = 2t$ e $\mu = 1$.



Fonte: Elaborada pelo autor.

4.1.3 Propriedades dos modelos com fragilidade Katz versus Poisson Generalizada

A seguir apresentam-se um estudo comparativo entre os modelos induzidos por fragilidade discreta. A Tabela 6 sumariza as estatísticas da média $E(Z)$, variância $V(Z)$, fração de cura p_0 e relação da média-variância κ para ambos as distribuição de fragilidade Z.

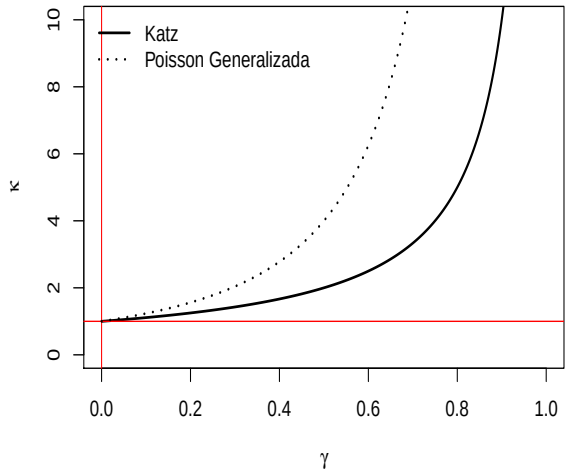
Tabela 6 – Média $E(Z)$, variância $V(Z)$, fração de cura p_0 e relação da média-variância κ para ambos as distribuições de fragilidade Z.

Fragilidade	$E(Z)$	$V(Z)$	κ	p_0
Katz	μ	$\frac{\mu}{(1-\gamma)}$	$(1-\gamma)^{-1}$	$[1-\gamma]^{\frac{(1-\gamma)\mu}{\gamma}}$
Poisson Generalizada	μ	$\frac{\mu}{(1-\gamma)^2}$	$(1-\gamma)^{-2}$	$e^{-(1-\gamma)\mu}$

Fonte: Elaborada pelo autor.

Pela Figura 7 tem-se que a para $0 < \gamma < 1$ a distribuição Poisson Generalizada apresenta maior variabilidade quando comparada com a distribuição Katz. A fração de cura, para ambos

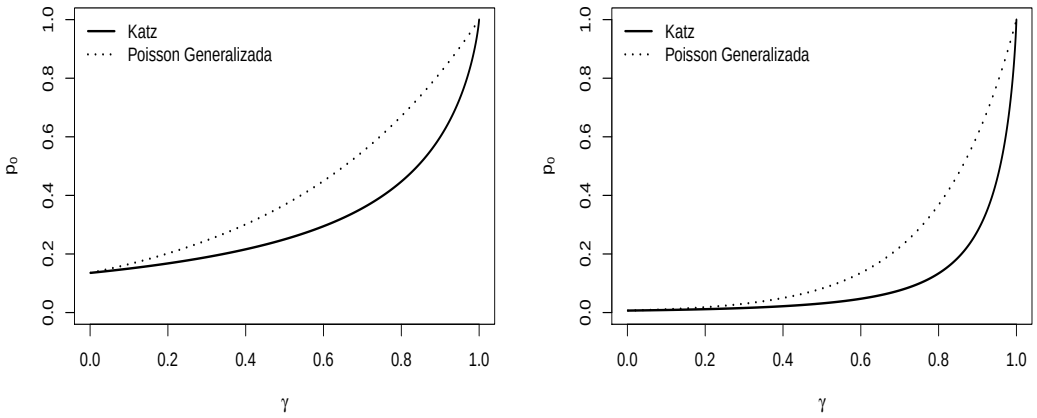
Figura 7 – Comportamento do parâmetro κ para diferentes valores de γ.



Fonte: Elaborada pelo autor.

os modelos, aumenta quando γ vai se aproximando de 1. Além disso, o modelo induzido por fragilidade discreta Poisson Generalizada possui fração de cura superior ao modelo induzido por fragilidade katz, ver Figura 8.

Figura 8 – Comportamento da fração de cura (p_0) para ambos os modelos induzidos por fragilidade discreta, considerando $\mu = 2$ (grafico à esquerda) e $\mu = 5$ (gráfico à direita) e $0 < \gamma < 1$.



Fonte: Elaborada pelo autor.

4.1.4 Modelo induzido por fragilidade discreta x modelo de risco latente

O conhecido modelo de fração de cura BCH proposto por Yakovlev e Tsodikov (1996) assume que existe um processo biológico latente no qual as células cancerígenas evoluem até a observação do evento de interesse (recorrência do câncer). Assim, indivíduos em risco do evento de interesse acontecer devem ser expostos a pelo menos um desses fatores latentes, caso contrário, o indivíduo será considerado completamente curado. O tempo de recorrência é observado quando o primeiro desses fatores latentes é ativado.

Como uma alternativa ao modelo BCH, Cancho *et al.* (2022) propõem um novo modelo com uma distribuição mais flexível para os fatores de riscos latentes, no caso assumindo que o números de riscos latentes possuem uma distribuição Poisson Generalizada, nomeado *Generalized Bounded Cumulative Hazard* (GBCH) que modela a sobredispersão dos fatores de riscos ($Var(M) > E(M)$), essa flexibilidade da variância em relação à média se deve à adição de um parâmetro na distribuição de Poisson.

De acordo com Cancho *et al.* (2022), M denota o número de células alteradas ativadas e podem ocasionar o evento de interesse (por exemplo, a recidiva da doença). Condicionadas a M, tem-se que as causas latentes Z_i são variáveis aleatórias independentes e identicamente distribuidas (iid) com função de sobrevivência $S_0(t)$ e independentes de M. Sob essas condições, tem-se que o tempo observável para toda a população é dado por $T = \min(Z_1, \cdots, Z_M)$ e $P(T > t|M = 0) = 1$ e a função de sobrevivência do modelo GBCH é dada por

$$\begin{aligned} S(t) &= P(M=0) + P(Z_1 > t)P(M=1) + P(Z_1 > t, Z_2 > t)P(M=2) + \ldots \\ &= \sum_{j=0}^{\infty} [S(t)]^j P(M=j) = G_M(S_0(t)). \end{aligned} \tag{4.8}$$

Logo, a função de sobrevivência do modelo GBCH é descrita pela função geradora de probabilidade dada na Equação 4.2 e avaliada na função de risco base $S_0(t)$, ou seja, $S(t) = G_M(S_0(t)$ como mostra a Equação 4.8.

Consequentemente, tem-se que o modelo induzido por fragilidade discreta Poisson Generalizada proposto neste trabalho possui a mesma expressão matemática para a função de sobrevivência do modelo GBCH proposto por Cancho *et al.* (2022). Esse fato define que os modelos com fração de cura podem ser obtidos por ambas as metodologias: fragilidade discreta e riscos latentes.

Em resumo, uma vez que o modelo GBCH apresentou bons resultados na simulação e no ajuste de dados de câncer cervical sob a abordagem clássica (CANCHO *et al.*, 2022). Neste trabalho, apresenta-se a formulação do modelo induzido por fragilidade discreta Poisson Generalizada numa abordagem bayesiana.

4.2 Inferência Bayesiana

Considere que os tempos $T_1,\ldots,T_n$ podem ser observados ou não e estão sujeitos a censura à direita C_i, denotando os tempos de censura dos n indivíduos. Em seguida, considere $t_i = \min\{T_i, C_i\}$ e $\delta_i = \mathrm{I}(T_i \leq C_i)$, no qual

$$\delta_i = \begin{cases} 1, & \text{se } T_i \text{ (tempo de vida)} \\ 0, & \text{se } C_i \text{ (censura à direita)} \end{cases}$$

para todos os indivíduos $i = 1,\ldots,n$. A partir de $(t_1,\delta_1),(t_2,\delta_2),...,(t_n,\delta_n)$ pode-se construir a função de verossimilhança marginal sob censura não informativa dada pela seguinte expressão

$$L(\boldsymbol{\vartheta};\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{X}) \propto \prod_{i=1}^{n} f(t_i,\delta_i,\boldsymbol{x}_i;\boldsymbol{\vartheta}), \tag{4.9}$$

em que $\mathbf{t} = (t_1,...,t_n)^\top$, $\boldsymbol{\delta} = (\delta_1,...,\delta_n)^\top$, $\boldsymbol{\vartheta} = (\gamma,\boldsymbol{\varphi}^\top,\boldsymbol{\beta}^\top)^\top$ e

$$f(t_i,\delta_i,\boldsymbol{x}_i;\boldsymbol{\vartheta}) = \sum_{z_i=0}^{\infty} [S_0(t_i,\boldsymbol{\varphi}]^{z_i-\delta_i}[z_i f_0(t_i;\boldsymbol{\varphi}]^{\delta_i} p(z_i;\gamma,\boldsymbol{\beta}). \tag{4.10}$$

A função de verossimilhança em (4.9) pode ser escrita por

$$L(\boldsymbol{\vartheta};\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{X}) \propto \prod_{i=1}^{n} f(t_i,\boldsymbol{x}_i;\boldsymbol{\vartheta})^{\delta_i} S(t_i,\boldsymbol{x}_i;\boldsymbol{\vartheta})^{1-\delta_i}, \tag{4.11}$$

tal que $\boldsymbol{X} = (\boldsymbol{x}_1,...,\boldsymbol{x}_n)^\top$ é uma matriz de covariáveis, ou seja, para cada indivíduo i tem-se $\boldsymbol{x}_i = (1,x_{i1},\ldots,x_{ip})^\top$ e $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_p)^\top$ o correspondente vetor de coeficientes de regressão desconhecidos, em que relaciona-se μ com as covariáveis através da função logarítmica dada por $\mu_i = \exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})$.

Considerando μ_i em (4.3) e (4.7), os coeficientes de regressão podem representar o papel dos grupos imunes e não-imunes. Assim, substituindo as funções $S(t_i;\boldsymbol{\vartheta})$ e $f(t_i;\boldsymbol{\vartheta})$ do modelo induzido por fragilidade Poisson Generalizado, tem-se que a função de verossimilhança observada é dada por

$$\begin{aligned} L(\boldsymbol{\vartheta};\boldsymbol{t},\boldsymbol{\delta},\boldsymbol{X}) \propto \prod_{i=1}^{n} & \left\{ \frac{\mu_i(\gamma-1)}{\gamma} h_0(t_i;\boldsymbol{\varphi}) \left[\frac{W(-\gamma e^{-\gamma} S_0(t_i;\boldsymbol{\varphi}))}{1+W(-\gamma e^{-\gamma} S_0(t_i;\boldsymbol{\varphi}))} \right] \right\}^{\delta_i} \\ & \exp\left\{ -\frac{(1-\gamma)\mu_i}{\gamma} \left[W(-\gamma e^{-\gamma} S_0(t_i;\boldsymbol{\varphi})) + \gamma \right] \right\} \end{aligned} \tag{4.12}$$

Consequentemente, o logaritmo da função de verossimilhança sob censura não informativa para os n indivíduos independentes tem a forma

$$\begin{aligned} \ell(\boldsymbol{\vartheta}) = & -\sum_{i=1}^{n} \delta_i \log\left(\frac{(1-\gamma)\mu_i}{\gamma}\right) + \sum_{i=1}^{n} \delta_i \log[h(t_i;\boldsymbol{\varphi})] \\ & + \sum_{i=1}^{n} \delta_i \log\left[\frac{W(-\gamma e^{-\gamma} S(t_i;\boldsymbol{\varphi}))}{1+W(-\gamma e^{-\gamma}) S(t_i;\boldsymbol{\varphi})} \right] - \sum_{i=1}^{n} \frac{(1-\gamma)\mu_i}{\gamma} \left[W(-\gamma e^{-\gamma} S(t_i;\boldsymbol{\varphi})) + \gamma \right], \end{aligned} \tag{4.13}$$

em que $\boldsymbol{\varphi} = (\varphi_1,\cdots,\varphi_q)^\top$ é o vetor de parâmetros da distribuição de base.

4.2.1 Distribuições a priori e a posteriori

Sob o enfoque bayesiano, assume-se que os parâmetros γ, $\boldsymbol{\varphi}$, $\boldsymbol{\beta}$ possuem distribuição a *priori* independentes, ou seja,

$$\pi(\boldsymbol{\vartheta}) = \pi(\gamma)\pi(\boldsymbol{\varphi})\pi(\boldsymbol{\beta}). \tag{4.14}$$

Desta forma, tendo como distribuição de base as distribuições exponencial ou Weibull pode-se assumir as seguintes distribuições a *priori*: (i) Exponencial: $\gamma \sim B(a_\gamma, b_\gamma)$, $\lambda \sim G(a_\lambda, b_\lambda)$, $\beta_j \sim N(0, \sigma^2_{\beta_j})$. (ii) Weibull: $\gamma \sim B(a_\gamma, b_\gamma)$, $\varphi_1 \sim G(a_{\varphi_1}, b_{\varphi_1})$, $\varphi_2 \sim N(0, \sigma^2_{\varphi_2})$ e $\beta_j \sim \pi(\beta_j) \propto 1$. Para $B(a,b)$ uma distribuição beta, $N(a,b)$ uma distribuição normal, $G(a,b)$ uma distribuição gama para a como parâmetro de forma, b o parâmetro de taxa.

Combinando a função de verossimilhança da Equação 4.12 com a distribuição a *priori* dada pela Equação 4.14, tem-se que a distribuição a *posteriori* conjunta para $\boldsymbol{\vartheta}$ é dada pela expressão

$$\pi(\boldsymbol{\vartheta}|\mathcal{D}) \propto L(\boldsymbol{\vartheta};\mathcal{D})\pi(\gamma)\pi(\boldsymbol{\varphi})\pi(\boldsymbol{\beta}). \tag{4.15}$$

A densidade a *posteriori* em (4.15) não possui expressão fácil de ser tratada analiticamente e, portanto, as inferências foram baseadas nos métodos de simulação Monte Carlo via Cadeias de Markov (MCMC), mais especificamente o Algoritmo de Metropolis-Hasting (M-H) (GAMERMAN; LOPES, 2006) pois nenhuma das distribuições condicionais completas possui forma fechada.

O algoritmo de M-H segue os mesmos princípios dos métodos de rejeição, em que um valor é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade. Esse processo garante a convergência da cadeia para a distribuição de equilíbrio, no caso, a distribuição a *posteriori*.

4.2.2 Critérios para comparação de modelo

Existem muitos critérios para avaliação dos ajustes dos modelos no enfoque bayesiano, e neste trabalho considerou-se o *Deviance Information Criterion* (DIC), o *Log Pseudo Marginal Likelihood* (LPML), o *Expected Akaike Information Criterion* (EAIC) e o *Expected Bayesian Information Criterion* (EBIC).

O critério DIC (SPIEGELHALTER *et al.*, 2002) é baseado na média a *posteriori* da *deviance* $D(\boldsymbol{\vartheta})$. A partir das amostras MCMC, o DIC pode ser aproximado por $\overline{D} = \sum_{q=1}^{Q} \frac{D(\boldsymbol{\vartheta}_q)}{Q}$ uma vez que $D(\boldsymbol{\vartheta}) = -2\sum_{i=1}^{n} \log[g(t_i;\boldsymbol{\vartheta})]$, para os dados observados tem-se que $g(.)$ representa a função de verosimilhança dos correspondente do modelo e o indíce q indica a q-ésima realização. Portanto, o critério DIC pode ser estimado por

$$\widehat{DIC} = 2\overline{D} - \widehat{D} \tag{4.16}$$

em que os valores de $\widehat{D}$ são dados por: $\widehat{D} = D\left(\frac{1}{Q}\sum_{q=1}^{Q}\gamma^{(q)}, \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\varphi}^{(q)}, \frac{1}{Q}\sum_{q=1}^{Q}\boldsymbol{\beta}^{(q)}\right)$. O modelo que melhor se ajusta é aquele que possui o menor valor do DIC.

O critério LPML (GELFAND; DEY; CHANG, 1992) é derivado da Ordenada Preditiva Condicional (CPO) e é calculado por

$$LPML = \sum_{i=1}^{n} \log(\widehat{CPO}_i). \tag{4.17}$$

Seja D os dados completos e D^{-i} os dados com a i-ésima observação deletada, para a i-ésima observação, tem-se que a densidade a *posteriori* de $\boldsymbol{\vartheta}$ dado D^{-i} é $\pi(\boldsymbol{\vartheta}|\mathcal{D})^{(-i)}$. Logo, a estimativa $\widehat{CPO}_i$ para a i-ésima observação é $\widehat{CPO}_i = \left(\int_{\boldsymbol{\vartheta}} \frac{\pi(\boldsymbol{\vartheta}|\mathcal{D})}{g(t_i;\boldsymbol{\vartheta})} d\boldsymbol{\vartheta}\right)^{-1}$.

Uma estimativa de CPO_i pode ser obtida de uma única amostra MCMC da distribuição a *posteriori* $\pi(\boldsymbol{\vartheta}|\mathcal{D})$. Considere uma amostra de tamanho $\mathcal{Q}$ de $\pi(\boldsymbol{\vartheta}|\mathcal{D})$ depois do *burn-in* dada por $\boldsymbol{\vartheta}^{(1)}, ..., \boldsymbol{\vartheta}^{(Q)}$, uma aproximação Monte Carlo de CPO_i é

$$\widehat{CPO}_i = \left(\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{g(t_i|\boldsymbol{\vartheta}^{(q)})}\right)^{-1}. \tag{4.18}$$

O critério EAIC (BROOKS, 2002) é dado pelo valor esperado do AIC e é obtido por

$$\widehat{EAIC} = \overline{D} + 2m, \tag{4.19}$$

em que m denota o número de parâmetros do modelo que está sendo ajustado, dimensão de $(\boldsymbol{\vartheta})$.

Similarmente, o critério EBIC (CARLIN; LOUIS, 2001) é dado por

$$\widehat{EBIC} = \overline{D} + 2m\log(n), \tag{4.20}$$

em que m denota o número de parâmetros do modelo sendo ajustado, dimensão de $(\boldsymbol{\vartheta})$ e n é o tamanho amostral.

4.3 Estudos de simulação

Nesta seção apresenta-se o estudo de simulação para avaliar o desempenho das estimativas bayesianas do modelo proposto. Para a implementação computacional utilizou-se a Linguagem R (R Core Team, 2019), as estimativas bayesianas foram obtidas via algoritmo de Metropolis-Hasting atráves de uma rotina implementada pelos autores.

Para a distribuição de base tem-se a distribuição exponencial com risco de base $h_0(t) = \lambda = 1$. Para o parâmetro γ da distribuição de fragilidade Poisson Generalizada, considerou-se os valores ($\gamma = 0,3; 0,5$ e $0,8$) e uma covariável reparametrizada na média via função de ligação $\log(\mu_i) = \beta_0 + \beta_1 x_i$ onde $\beta_0 = 0,5$ e $\beta_1 = 2$, $i = 1,2,...,n$, aqui a covariável x_i é gerada a partir de uma distribuição de Bernoulli com probabilidade de $0,5$. Os tempos de censura C_i foram

amostrados das distribuições uniformes no intervalo $(0,\tau)$, onde τ é definido para controlar a proporção de observações censuradas. Nessa simulação, considerou-se $\tau = 5$, levando a uma proporção de observações censuradas em torno de 60%. Para cada tamanho de amostras $n = 200$, 500 e 800, simulou-se $Q = 500$ amostras de dados do modelo proposto, os tempos observados foram gerados seguindo o mesmo algoritmo apresentado na Seção 3.4.

Considerou-se as seguintes distribuições a *priori* independentes para a performance do Algoritmo Metropolis-Hasting: $\gamma \sim B(2;2)$, $\lambda \sim G(1;0,1)$, em que $0,1$ é um parâmetro de taxa, e $\beta_j \sim N(0;10^4)$, $j = 0,1$.

Gerou-se uma cadeia de 55000 iterações MCMC, sendo descartadas as 5000 (*burn-in*) iterações iniciais pois as iterações são influenciadas pelo estado inicial. Para obter uma amostra independente, foram considerados saltos de tamanho 50 iterações (*thin*). Portanto, os resultados a *posteriori* foram baseados em 1000 iterações independentes da cadeia de Markov.

O estudo de simulação foi conduzido com o intuito de examinar o comportamento das estimativas bayesianas baseadas na média amostral do parâmetro, do desvio padrão dos estimativas, da estimativa de viés, da raiz do erro quadrático médio e das probabilidades de cobertura dos intervalos de confiança de 95% para os parâmetros do modelo. Os resultados obtidos estão na Tabela 7, dos quais observou-se que o REQM e o DP descrecem à medida que o tamanho da amostra aumenta.

Tabela 7 – Resultados da simulação para o modelo proposto considerando diferentes tamanhos de amostras e diferentes valores para o parâmetro γ da distribuição de fragilidade Poisson Generalizada.

		$\gamma = 0,3$				$\gamma = 0,5$				$\gamma = 0,8$			
n		γ	λ	β_0	β_1	γ	λ	β_0	β_1	γ	λ	β_0	β_1
	Media	0,372	0,948	0,677	2,003	0,447	1,100	0,474	2,023	0,686	1,322	0,137	2,048
	DP	0,085	0,235	0,245	0,193	0,111	0,292	0,302	0,207	0,110	0,375	0,461	0,232
200	VIES	0,072	-0,052	0,177	0,003	-0,053	0,100	-0,026	0,023	-0,114	0,322	-0,363	0,048
	REQM	0,112	0,240	0,302	0,193	0,123	0,308	0,303	0,208	0,158	0,494	0,586	0,236
	PC	0,988	0,978	0,998	0,954	0,972	0,972	0,982	0,940	0,950	0,930	0,872	0,960
	Media	0,337	0,965	0,595	2,003	0,455	1,063	0,459	2,023	0,759	1,111	0,393	2,001
	DP	0,091	0,174	0,198	0,111	0,105	0,212	0,240	0,124	0,071	0,230	0,333	0,14
500	VIES	0,037	-0,035	0,095	0,003	-0,045	0,063	-0,041	0,023	-0,041	0,111	-0,107	0,001
	REQM	0,098	0,177	0,220	0,111	0,114	0,221	0,244	0,126	0,082	0,255	0,349	0,149
	PC	0,986	0,970	0,992	0,952	0,972	0,982	0,970	0,956	0,952	0,930	0,916	0,950
	Media	0,329	0,968	0,578	1,999	0,453	1,077	0,446	2,004	0,768	1,084	0,389	2,016
	DP	0,086	0,148	0,174	0,100	0,098	0,183	0,215	0,099	0,053	0,173	0,269	0,120
800	VIES	0,029	-0,032	0,078	-0,001	-0,047	0,077	-0,054	0,004	-0,032	0,084	-0,111	0,016
	REQM	0,091	0,151	0,191	0,100	0,109	0,198	0,221	0,099	0,062	0,193	0,291	0,121
	PC	0,984	0,972	0,992	0,940	0,946	0,960	0,954	0,948	0,964	0,938	0,922	0,946

Fonte: Elaborada pelo autor.

4.4 Aplicação - Câncer Cervical

Para avaliar a aplicabilidade do modelo sob a metodologia bayesiana considerou-se o mesmo conjunto de dados de câncer cervival apresentado na Seção 3.5 do Capítulo 3 com as mesmas especificações iniciais, ou seja, considerou-se a distribuição Weibull com função de sobrevivência, $S_0(y;\boldsymbol{\varphi}) = \exp\{-\exp\{\varphi_2\}y^{\varphi_1}\}$, $\varphi_1 > 0$ e $\varphi_2 \in R$ para a função de sobrevivência

de base na Equação 4.3. E para o tempo de recidiva do câncer cervical, as covariáveis do modelo foram reparametrizadas na média μ:

$$\mu_i = \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\} = \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{3_1} x_{i3_1} + \beta_{3_2} x_{i3_2} + \beta_{4_1} x_{i4_1} + \beta_{4_2} x_{i4_2}\}, \quad (4.21)$$

em que $i = 1, \ldots, 2489$.

Para o ajuste bayesiano dos dados de câncer cervical, assumiu-se as *priori* independentes conforme descrito na Subseção 4.2.1. Mais especificamente, $\gamma \sim B(2;2)$, $\varphi_1 \sim G(1;0,1)$, $\varphi_2 \sim N(0;10^2)$ e $\beta_j \sim \pi(\beta_j) \propto 1$, $j = 1,2...,7$.

Implementou-se o algoritmo Metropolis-Hasting na Linguagem (R Core Team, 2019) e considerou-se uma cadeia de 35000 iterações MCMC, sendo descartadas as 5000 (*burn-in*) iterações iniciais pois sabe-se que essas iterações são influenciadas pelo estado inicial. Para obter uma amostra independente, foram considerados saltos de tamanho 5 iterações (*thin*). Portanto, os resultados a *posteriori* foram baseados em 6000 iterações independentes da cadeia de Markov.

As trajetórias (*trace plots*) da cadeia para todos os parâmetros do modelo induzido por fragilidade discreta Poisson Generalizada são apresentadas na Figura 10 e indicam a convergência dos parâmetros. A Figura 11 apresenta as densidades das distribuições a *posteriori* marginais para todos os parâmetros estimados.

Um resumo a *posteriori* dos parâmetros do modelo induzido por fragilidade discreta Poisson Generalizada ajustado aos dados de câncer cervical é apresentado na Tabela 8. Pelos percentis nota-se que todas as covariáveis (x_1, x_2, x_3 e x_4) mostraram-se significativas. Consequentemente, há uma diferença significativa entre os níveis dessas covariáveis quanto à proporção de cura (p_0) dos indivíduos e quanto ao risco de recidiva da doença.

Tabela 8 – Média, desvio padrão e percentil a *posteriori* do modelo induzido por fragilidade discreta Poisson Generalizada ajustado nos dados de câncer cervical com todas as covaráveis.

Parâmetros	Média	Mediana	Desvio Padrão	Percentil	
				2,5%	97,5%
γ (dispersão - fragilidade)	0,946	0,951	0,028	0,877	0,988
ϕ_1 (forma - Weibull)	1,215	1,218	0,106	1,007	1,417
ϕ_2 (escala - Weibull)	-4,011	-3,774	0,996	-6,626	-2,945
β_0 (intercepto)	3,012	2,850	1,003	1,685	5,508
β_1 (idade > 60)	0,249	0,250	0,121	0,007	0,482
β_2 (cirurgia - sim)	-0,414	-0,411	0,127	-0,659	-0,163
β_{31} (tratamento - Q+R)	-0,532	-0,534	0,167	-0,838	-0,199
β_{32} (tratamento - R)	-1,323	-1,322	0,229	-1,771	-0,860
β_{41} (estágio - II+III)	0,258	0,255	0,139	0,005	0,536
β_{42} (estágio - IV)	1,065	1,067	0,162	0,740	1,376

Fonte: Elaborada pelo autor.

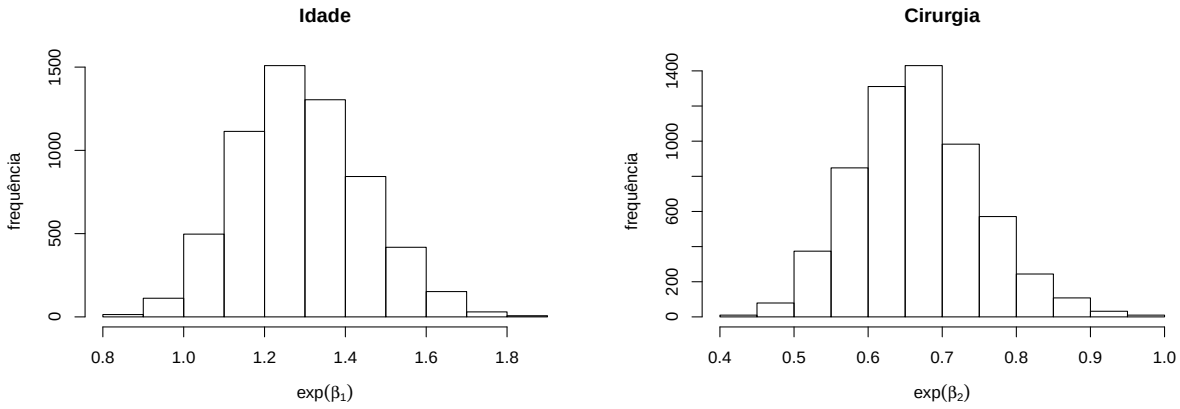
Pela Tabela 8, pacientes com idade superior a 60 anos tiveram pior prognóstico da doença ($\beta_1 = 0.249 > 0$) o que indica que o risco de recidiva da doença desses pacientes é

$\exp\{0,249\} = 1,28$ vezes maior que para pacientes com idade inferior a 60 anos. Pacientes com cirurgia têm melhor prognóstico ($\beta_2 = -0,414 < 0$), ou seja, têm maior tempo de sobrevida. As estimativas dos coeficientes de regressão associados ao tratamento são todas negativas ($\beta_{3_1} = -0,532$ e $\beta_{3_2} = -1,323$) e indicam que os pacientes cujo protocolo de tratamento incluiu Q+R combinado e R sozinho têm menor risco de recidiva da doença do que pacientes que receberam apenas Q. Pacientes com estágio II+III ou estágio IV da doença ($\beta_{4_1} = 0,258$ e $\beta_{4_2} = 1,065$, respectivamente) têm maior risco de recidiva da doença do que os pacientes no estágio I.

Na Figura 9 apresentam-se o comportamento da razão de risco (RR) como descrita na Equação 3.12 para pacientes com idade ≤ 60 anos versus idade > 60 e a razão de risco para pacientes que passaram ou não por cirurgia. $\exp\{\beta(\boldsymbol{x}_j^\top - \boldsymbol{x}_k^\top)$

O risco médio e intervalo de credibilidade HPD de 95% para pacientes com idade > 60 anos é $1,292\ (1,007; 1,619)$ implicando que a fração de cura é maior para pacientes com idade ≤ 60 anos. Para pacientes que passaram por cirurgia tem-se o risco médio e intervalo de credibilidade HPD de 95% dado por $0,666\ (0,517; 0,849)$.

Figura 9 – Razão de risco para pacientes com idade ≤ 60 anos versus idade > 60 anos (gráfico à esquerda) e razão de risco para pacientes que passaram ou não por cirurgia (gráfico à direita).



Fonte: Elaborada pelo autor.

Na Tabela 9 tem-se as estimativas bayesianas e intervalos HPD de 95% do modelo proposto ajustados aos dados de câncer cervical comparado ao modelo BCH. Em ambos os modelos, todas as covariáveis se mostraram significativas, no entanto observou-se que o limite inferior da idade (β_1) e do estágio II+III (β_{41}) estão próximos de zero.

Para termos de avaliação da performance do modelo proposto considerou-se os critérios discutidos na Subseção 4.2.2. Pela Tabela 10 tem-se que o modelo proposto possui menores valores das estatísticas quando comparado ao modelo BCH, portanto o modelo proposto aparenta conseguir modelar a heterogeneidade não observada dos dados.

Tabela 9 – Estimativas bayesianas e intervalos HPD de 95% para o modelo BCH e para o modelo proposto ajustados aos dados de câncer cervical.

	Modelo BCH			Modelo Proposto		
		Intervalo HPD (95%)			Intervalo HPD (95%)	
	Média	LI	LS	Média	LI	LS
γ				0,946	0,877	0,988
ϕ_1	0,797	0,711	0,882	1,215	1,007	1,417
ϕ_2	-1,608	-2,071	-1,276	-4,011	-6,626	-2,945
β_0	-0,327	-0,812	0,185	3,012	1,685	5,508
β_1	0,235	-0,000	0,463	0,249	0,007	0,482
β_2	-0,409	-0,665	-0,165	-0,414	-0,659	-0,163
β_{31}	-0,531	-0,831	-0,189	-0,532	-0,838	-0,199
β_{32}	-1,299	-1,716	-0,860	-1,323	-1,771	-0,860
β_{41}	0,267	0,003	0,522	0,258	0,005	0,536
β_{42}	1,063	0,731	1,381	1,065	0,740	1,376

Fonte: Elaborada pelo autor.

Tabela 10 – Critérios de comparação entre o modelo induzido por fragilidade discreta Poisson Generalizada e o modelo BCH

Modelo	DIC	EAIC	EBIC	LPML
Modelo proposto	3109,431	3129,927	3188,124	-1559,604
BCH	3142,076	3150,948	3203,324	-1571,005

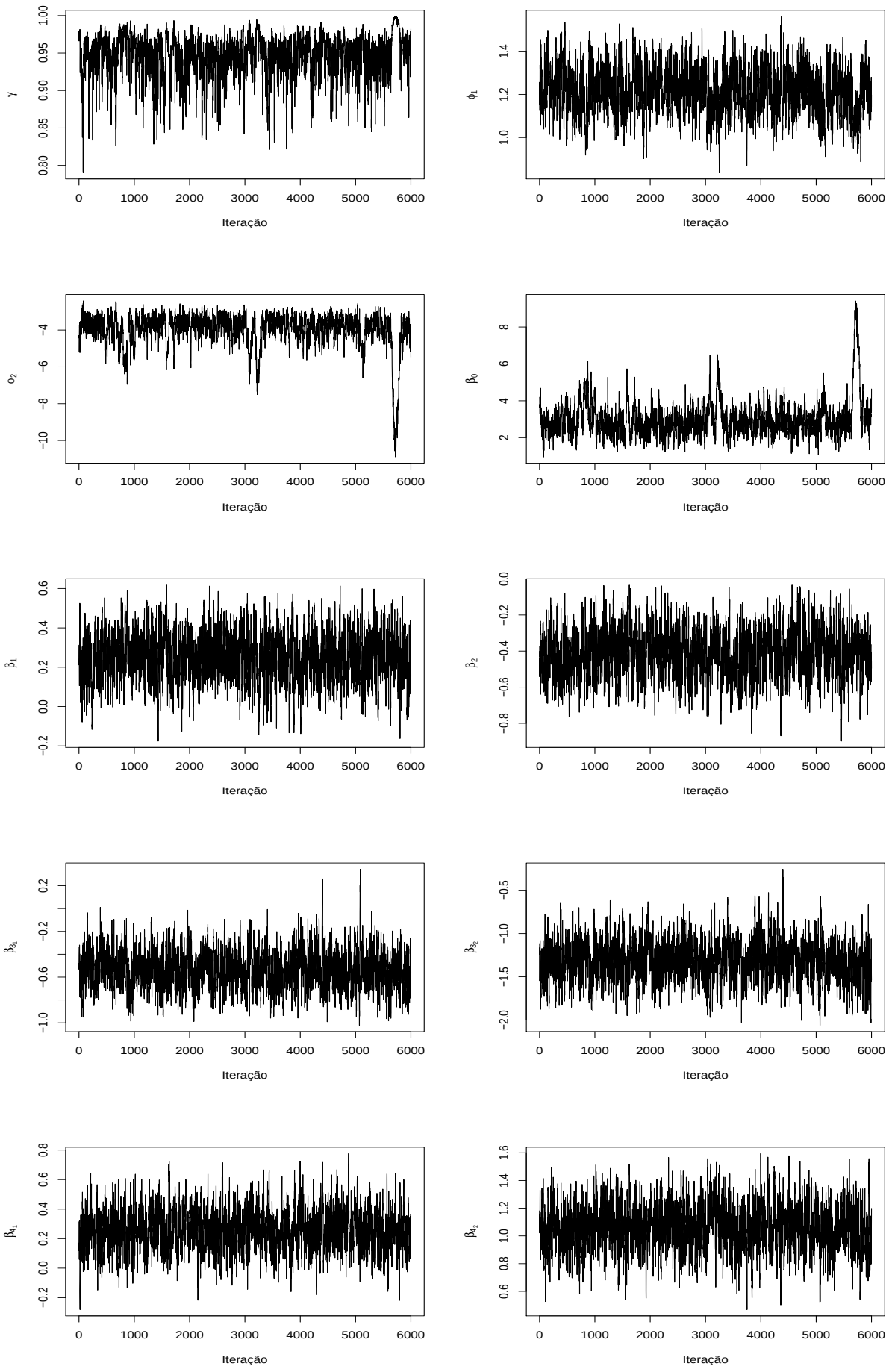
Fonte: Elaborada pelo autor.

4.5 Comentários Finais

Neste capítulo apresentou-se um novo modelo de sobrevivência induzido por fragilidade discreta Poisson Generalizada com heterogeneidade não observada. O procedimento inferencial foi baseado na abordagem bayesiana. O modelo proposto pode ser uma alternativa aos modelos existentes, uma vez que tem a estrutura de riscos proporcionais e leva em conta a sobredispersão, equidispersão. Além disso, a importância do modelo foi validada na aplicação em dados de câncer cervical, uma vez que o modelo apresentou um valor de $DIC = 3109.431$, enquanto que o modelo BCH apresentou um $DIC = 3142.076$, dando indícios que o modelo proposto conseguiu modelar a heterogeneidade não observada dos dados.

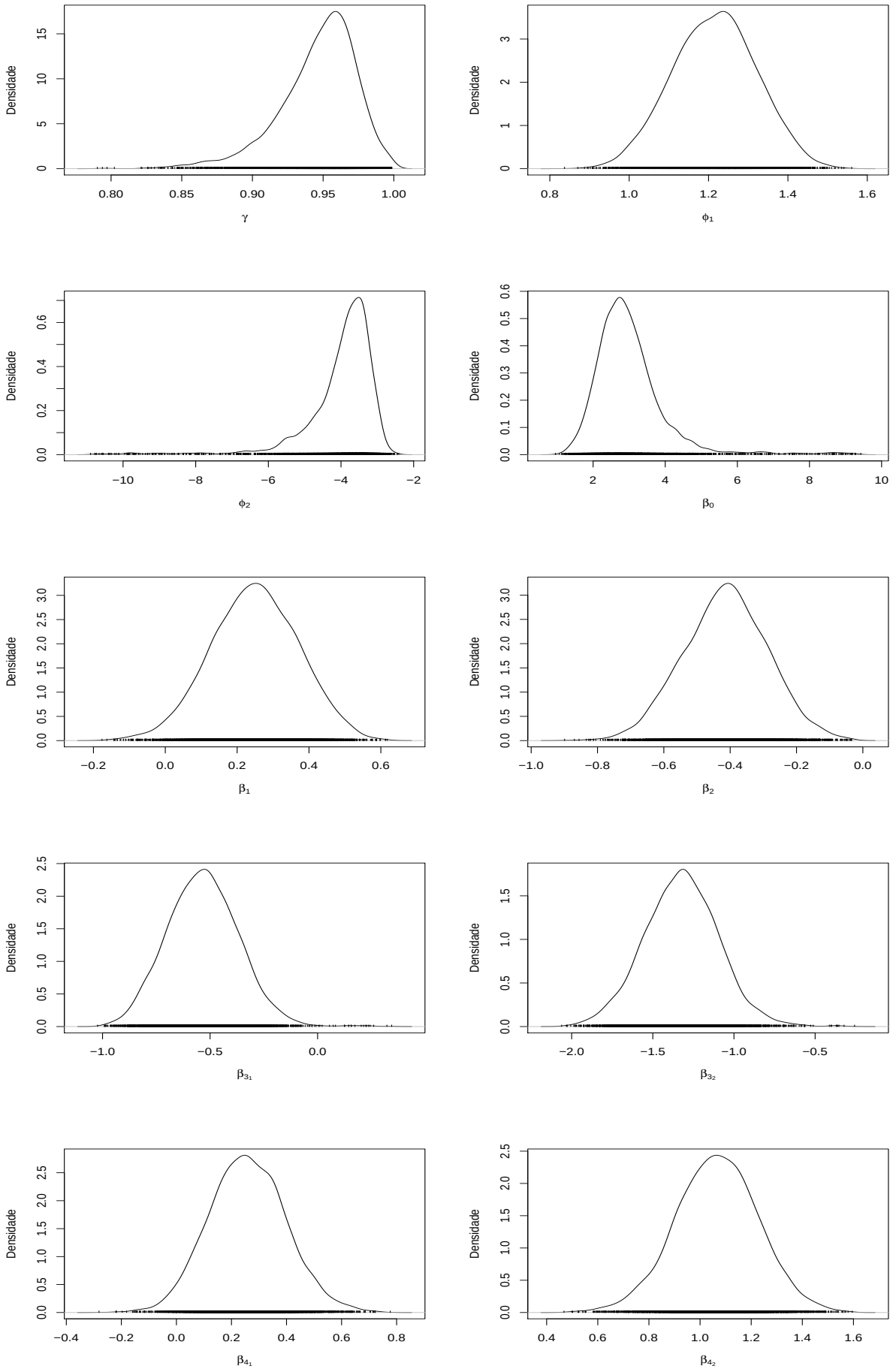
Como parte dos resultados desse Capítulo, sob uma abordagem clássica, tem-se o artigo intitulado "*A survival regression with cure fraction applied to cervical cancer*" publicado na revista *Computational Statistics*.

Figura 10 – Gráfico *Traceplot* com as trajetórias da cadeia para todos parâmetros do modelo GBCH.



Fonte: Elaborada pelo autor.

Figura 11 – Densidades a *posteriori* marginais dos parâmetros γ, ϕ_1, ϕ_2, β_0, β_1, β_2, β_{31}, β_{32}, β_{41} e β_{42}.



Fonte: Elaborada pelo autor.

CAPÍTULO

5

CONCLUSÕES E PROPOSTAS FUTURAS

Na primeira parte deste trabalho apresentou-se um novo modelo de sobrevivência induzido por fragilidade discreta Katz com heterogeneidade não observada. O modelo proposto pode ser uma alternativa aos modelos existentes, uma vez que engloba alguns modelos como casos particulares, além da vantagem de ter a estrutura de riscos proporcionais e levar em conta a sobredispersão, equidispersão e subdispersão. Além disso, a importância do modelo foi validada na aplicação em dados de câncer cervical, uma vez que o modelo apresentou um valor de $AIC = 3126.896$, enquanto que o modelo BCH apresentou um $AIC = 3142.256$, dando indícios que o modelo proposto conseguiu modelar a heterogeneidade não observada dos dados. Como resultado do capítulo, um artigo intitulado "*A Survival Model for Lifetime with Long-Term Survivors and Unobserved Heteregeneity*" foi submetido para a revista *Brazilian Journal of Probability and Statistics* (BJPS).

Na segunda parte apresentou-se um novo modelo de sobrevivência induzido por fragilidade discreta Poisson Generalizada sob uma abordagem clássica em que desenvolveu-se o artigo intitulado "*A survival regression with cure fraction applied to cervical cancer*" publicado na revista *Computational Statistics* (CANCHO *et al.*, 2022). Sob uma abordagem bayesiana, a importância do modelo foi validada na aplicação em dados de câncer cervical, uma vez que o modelo apresentou um valor de $DIC = 3109.431$, enquanto que o modelo BCH apresentou um $DIC = 3142.076$, dando indícios que o modelo proposto conseguiu modelar a heterogeneidade não observada dos dados.

Existem muitas metodologias que podem da continuidade à pesquisa realizada até aqui. Entre elas, algumas propostas futuras para esse pesquisa são:

- Comparar os modelos propostos com outros modelos de fragilidade discreta já desenvolvidos;
- Um estudos comparativo das abordagens, clássica e bayesiana, para ambos os modelos.

- Desenvolver um modelo de sobrevência bivariado considerando a distribuição Poisson Generalizada Bivariada na fragilidade sob as perspectivas clássica e bayesiana. Empregando os mesmos procedimentos metodológicos que já utilizamos no artigo "*A multivariate survival model induced by discrete frailty*" (CANCHO *et al.*, 2020).

REFERÊNCIAS

AMBAGASPITIYA, R.; BALAKRISHNAN, N. On the compound generalized poisson distributions. **ASTIN Bulletin**, Cambridge University Press, v. 24, n. 2, p. 255–263, 1994. Citado na página 48.

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 nas páginas 26 e 27.

BALAKRISHNAN, N.; PENG, Y. Generalized gamma frailty model. **Statistics in Medicine**, v. 25, p. 2797–2816, 2006. Citado na página 26.

BARRIGA, G. D.; CANCHO, V. G.; GARIBAY, D. V.; CORDEIRO, G. M.; ORTEGA, E. M. A new survival model with surviving fraction: An application to colorectal cancer data. **Statistical methods in medical research**, SAGE Publications Sage UK: London, England, p. 1–15, 2018. Citado na página 26.

BARRIGA, G. D. C.; DEY, D. K.; CANCHO, V. G.; SUZUKI, A. K. Bayesian analysis of Birnbaum-Saunders survival model with cure fraction under a variety of activation mechanism. **Model Assisted Statistics and Applications**, IOS Press, v. 15, p. 35–51, 2020. ISSN 1875-9068. Citado na página 26.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, v. 42, p. 501–515, 1952. Citado nas páginas 22, 33 e 35.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **J. R. Statist. Soc. B**, v. 11, p. 15–53, 1949. Citado nas páginas 33 e 35.

BOSCH, R. J.; LOUISE, R. M. Generalised poisson models arising from Markov processes. **Statistics & Probability Letters**, v. 39, p. 205–212, 1998. Citado na página 31.

BROOKS, S. P. Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde (2002). **J. R. Statist. Soc. B**, v. 64, p. 616–618, 2002. Citado na página 54.

CANCHO, V. G.; BARRIGA, G.; LEÃO, J.; SAULO, H. Survival model induced by discrete frailty for modeling of lifetime data with long-term survivors and change-point. **Communications in Statistics - Theory and Methods**, Taylor Francis, v. 50, n. 5, p. 1161–1172, 2019. Citado na página 27.

CANCHO, V. G.; BEDIA, E. C.; CORDEIRO, G. M.; PRATAVIERA, F.; ORTEGA, E. M.; SANTO, A. P. A survival regression with cure fraction applied to cervical cancer. **Computational Statistics**, Springer, p. 1–16, 2022. Citado nas páginas 27, 51 e 61.

CANCHO, V. G.; RODRIGUES, J.; CASTRO, M. de. A flexible model for survival data with a cure rate: a bayesian approach. **Journal of Applied Statistics**, Taylor & Francis, v. 38, n. 1, p. 57–70, 2011. Citado nas páginas 26 e 33.

CANCHO, V. G.; SUZUKI, A. K.; BARRIGA, G. D. C.; SANTO, A. P. J. do E. A multivariate survival model induced by discrete frailty. **Communications in Statistics - Simulation and Computation**, Taylor Francis, v. 0, n. 0, p. 1–19, 2020. Disponível em: <https://doi.org/10.1080/03610918.2020.1806323>. Citado na página 62.

CANCHO, V. G.; ZAVALETA, K. E. C.; MACERA, M. A. C.; SUZUKI, A. K.; LOUZADA, F. A bayesian cure rate model with dispersion induced by discrete frailty. **Communications for Statistical Applications and Methods**, The Korean Statistical Society, and Korean International Statistical Society, v. 25, n. 5, p. 471–488, 9 2018. Citado na página 27.

CARLIN, B.; LOUIS, T. A. **Bayes and Empirical Bayes Methods for Data Analysis**. Boca Raton: Chapman Hall/CRC, 2001. Citado na página 54.

CARONI, C.; CROWDER, M.; KIMBER, A. Proportional hazards models with discrete frailty. **Lifetime Data Analysis**, v. 16, n. 3, p. 374–384, 2010. Citado na página 26.

____. Proportional hazards models with discrete frailty. **Lifetime Data Analysis**, v. 16, p. 374–384, 2010. Citado na página 27.

COLLETT, D. **Modelling survival data in medical research**. Florida: [s.n.], 2003. Citado na página 25.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. São Paulo: Edgard Blucher, 2006. Nenhuma citação no texto.

CONSUL, C. P.; SHENTON, R. L. Use of lagrange expansion for generating discrete generalized probability distributions. **Siam Journal on Applied Mathematics - SIAMAM**, v. 23, 1972. Citado na página 47.

CONSUL, P. Generalized poisson distributions: Properties and applications. **Marcel Dekker Inc. New York/Basel**, v. 26, 1989. Citado na página 47.

____. A model for distributions of injuries in auto-accidents. **Atteilungen der Schweiz Vereinigung der Versicherungsmathematiker**, v. 1, p. 161–168, 1990. Citado na página 47.

CONSUL, P.; JAIN, G. On a generalization of the poisson distribution. **Technometrics**, v. 15, p. 791–799, 11 1973. Citado nas páginas 22 e 47.

COONER, F.; BANERJEE, S.; CARLIN, B. P.; SINHA, D. Flexible cure rate modeling under latent activation schemes. **Journal of the American Statistical Association**, v. 102, p. 560–572, 2007. Citado na página 26.

CORDEIRO, G. M.; CANCHO, V. G.; ORTEGA, E. M. M.; BARRIGA, G. D. C. A model with long-term survivors: Negative binomial birnbaum-saunders. **Communications in Statistics-Theory and Methods**, v. 45, p. 1370–1387, 2016. Citado na página 26.

CORLESS, R. M.; GONNET, G. H.; HARE, D. E. G.; JEFFREY, D. J.; KNUTH, D. E. On the Lambert W function. **Advances in Computational Mathematics**, v. 5, p. 329–359, 1996. Citado na página 48.

COX, D.; OAKES, D. **Analysis of Survival Data**. London: Chapman & Hall, 1984. Citado na página 22.

____. **Analysis of Survival Data**. Chapman & Hall, 1996. (Monographs on statistics and applied probability). ISBN 9780412224904. Disponível em: <https://books.google.com.br/books?id=zOhXuAAACAAJ>. Citado na página 34.

COX, D. R. Regression models and life-tables (with discussion). **Journal of the Royal Statistical Society B**, v. 34, p. 187–220, 1972. Citado nas páginas 21 e 25.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the royal statistical society. Series B (methodological)**, JSTOR, p. 1–38, 1977. Citado nas páginas 36 e 37.

DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, v. 5, p. 236–244, 1996. Citado na página 42.

FANG, Y. Gmm tests for the katz family of distributions. **Journal of Statistical Planning and Inference**, v. 110, n. 1, p. 55 – 73, 2003. Citado na página 30.

FOX, J. **Applied regression analysis and generalized linear models**. [S.l.]: Sage Publications, 2015. Citado na página 34.

GAMERMAN, D.; LOPES, H. F. **Markov chain Monte Carlo: stochastic simulation for Bayesian inference**. [S.l.]: CRC press, 2006. Citado na página 53.

GELFAND, A. E.; DEY, D. K.; CHANG, H. Model determination using predictive distributions with implementation via sampling-based methods. In: OXFORD UNIVERSITY PRESS, USA. **Bayesian statistics: proceedings of the Fourth Valencia International Meeting, April 15-20, 1991**. [S.l.], 1992. v. 4, p. 147–167. Citado na página 54.

GURLAND, J.; TRIPATHI, R. Estimation of parameters on some extensions of the katz family of discrete distributions involving hypergeometric functions. **A Modern Course on Statistical Distributions in Scientific Work**, Springer Netherlands, v. 1, p. 59 – 82, 1975. Citado na página 30.

HOUGAARD, P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. **Biometrika**, Oxford University Press, v. 71, n. 1, p. 75–83, 1984. Citado nas páginas 25 e 26.

____. A class of multivanate failure time distributions. **Biometrika**, v. 73, p. 671–678, 1986. Citado na página 26.

____. Survival models for heterogeneous populations derived from stable distributions. **Biometrika**, Oxford University Press, v. 73, n. 2, p. 387–396, 1986. Citado na página 26.

JODRÁ, P. Computer generation of random variables with Lindley or Poisson-Lindley distribution via the Lambert w function. **Mathematics and Computers in Simulation**, v. 81, n. 4, p. 851–859, 2010. Citado na página 48.

KALBFLEISCH, J.; PRENTICE, R. **The statistical analysis of failure time data**. [S.l.]: Wiley, 1980. Citado na página 34.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, p. 457–481, 1958. Citado na página 21.

KATZ, L. Unified treatment of a broad class of discrete probability distributions. **Classical and contagious discrete distributions**, Statistical Publishing Society, Pergamon Press, Oxford, v. 1, p. 175–182, 1965. Citado nas páginas 22, 29 e 30.

KOKONENDJI, C. C. **Over and Underdispersion Models**. [S.l.]: John Wiley & Sons, Ltd, 2014. 506–526 p. Citado na página 31.

LANCASTER, T. Econometric methods for the duration of unemployment. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 939–956, 1979. Citado na página 25.

LANGE, K. A gradient algorithm locally equivalent to the em algorithm. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 425–437, 1995. Citado na página 37.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. Hoboken, NJ: John Wiley & Sons, 2003. Citado nas páginas 21 e 25.

LUCENO, A. Recursive characterization of a large family of discrete probability distributions showing extra-poisson variation. **Statistics**, Taylor & Francis, v. 39, n. 3, p. 261–267, 2005. Citado na página 31.

MALLER, R.; ZHOU, X. **Survival Analysis with Long Term Survivors**. [S.l.]: John Wiley and sons, 1996. Citado na página 37.

MCCULLAGH, P.; NELDER, J. **Generalized linear models**. London: [s.n.], 1989. Citado nas páginas 22 e 34.

MCLACHLAN, G.; KRISHNAN, T. **The EM algorithm and extensions**. [S.l.]: John Wiley & Sons, 2007. v. 382. Citado na página 36.

MILANI, E. A.; TOMAZELLA, V. L. D.; DIAS, T. C. M.; LOUZADA, F. *et al.* The generalized time-dependent logistic frailty model: An application to a population-based prospective study of incident cases of lung cancer diagnosed in northern ireland. **Brazilian Journal of Probability and Statistics**, v. 29, p. 132–144, 2015. Citado na página 27.

MOGER, T. A.; AALEN, O. O.; HALVORSEN, T. O.; STORM, H. H.; TRETLI, S. Frailty modelling of testicular cancer incidence using scandinavian data. **Biostatistics**, v. 5, p. 1–14, 2004. Citado na página 27.

MULLAHY, J. Heterogeneity, excess zeros, and the structure of count data models. **Journal of Applied Econometrics**, v. 12, n. 3, p. 337–350, 1997. Citado na página 31.

ORTEGA, E. M. M.; CORDEIRO, G. M.; CAMPELO, A. K.; KATTAN, M. W.; CANCHO, V. G. A power series beta weibull regression model for predicting breast carcinoma. **Statistics in Medicine**, v. 34, p. 1366–1388, 2015. Citado na página 26.

PRESS, W. H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; FLANNERY, B. P. **Numerical recipes 3rd edition: The art of scientific computing**. [S.l.]: Cambridge university press, 2007. Citado na página 36.

R Core Team. **R: A Language and Environment for Statistical Computing, R version 3.5.3**. Vienna, Austria, 2019. Disponível em: <http://www.R-project.org/>. Citado nas páginas 38, 54 e 56.

RESSING, M.; BLETTNER, M.; KLUG, S. J. Data analysis of epidemiological studies part 11 of a series on evaluation of scientific publications. **Deutsches Ärzteblatt international**, v. 107, p. 187–92, 2010. Citado na página 34.

RIGBY, R. A.; STASINOPOULOS, D. M. Generalized additive models for location, scale and shape (with discussion). **Applied Statistics**, v. 54, p. 507–554, 2005. Citado na página 42.

RODRIGUES, J.; CANCHO, V. G.; CASTRO, M. de; LOUZADA-NETO, F. On the unification of long-term survival models. **Statistics & Probability Letters**, v. 79, p. 753–759, 2009. Citado nas páginas 22 e 26.

RODRIGUES, J.; CASTRO, M. de; CANCHO, V. G.; BALAKRISHNAN, N. COM-Poisson cure rate survival models and an application to a cutaneous melanoma data. **J. Statist. Planng Inf.**, v. 139, p. 3605–3611, 2009. Citado nas páginas 26 e 30.

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 26.

SELF, S.; LIANG, K. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. **Journal of the American Statistical Association**, v. 82, n. 398, p. 605–610, 1987. Citado na página 38.

SILVA, G. O.; CORDEIRO, G. M.; ORTEGA, E. M. M. Surviving and non surviving fraction regression models based on the beta modified Weibull distribution. **Model Assisted Statistics and Applications**, IOS Press, v. 15, p. 111–126, 2020. ISSN 1875-9068. Citado na página 26.

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**, v. 26, n. 5, p. 2011–2028, 2017. Citado na página 27.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, v. 64, n. 4, p. 583–639, 2002. Citado na página 53.

TRIPATHI, R.; GURLAND, J. A general family of discrete distributions with hypergeometric probabilities. **Journal of the Royal Statistical Society: Series B**, v. 39, n. 3, p. 349 – 356, 1977. Citado na página 30.

TSODIKOV, A. D.; IBRAHIM, J. G.; YAKOVLEV, A. Y. Estimating cure rates from survival data: An alternative to two-component mixture models. **Journal of the American Statistical Association**, v. 98, p. 1063–1078, 2003. Citado nas páginas 26 e 30.

TWEEDIE, M. An index which distinguishes between some important exponential families. In: **Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference**. [S.l.: s.n.], 1984. p. 579–604. Citado na página 26.

VAUPEL, J. W.; MANTON, K. G.; STALLARD, E. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, v. 16, p. 439–454, 1979. Citado nas páginas 22 e 25.

WHO, W. H. O. **Global strategy to accelerate the elimination of cervical cancer as a public health problem**. Geneva, 2020. Citado na página 40.

WIENKE, A. **Frailty models in survival analysis**. New York: Chapman and Hall/CRC, 2010. Citado nas páginas 22, 25 e 26.

YAKOVLEV, A. Y.; TSODIKOV, A. D. **Stochastic Models of Tumor Latency and Their Biostatistical Applications**. New Jersey: World Scientific, 1996. Citado nas páginas 33 e 51.

YIN, G.; IBRAHIM, J. G. Cure rate models: a unified approach. **The Canadian Journal of Statistics**, v. 33, p. 559–570, 2005. Citado na página 26.


ICMC
USP
SÃO CARLOS