UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

O Modelo Weibull de Longa Duração Inflacionado de Zero Aplicado à Pacientes Diagnosticados com Doença de Crohn

Patrícia Picardi Morais de Castro

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP
SÃO CARLOS

Patrícia Picardi Morais de Castro

O Modelo Weibull de Longa Duração Inflacionado de Zero Aplicado à Pacientes Diagnosticados com Doença de Crohn

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestra – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *VERSÃO REVISADA*

Área de Concentração: Matemática, Estatística e Computação

Orientadora: Profa. Dra. Gleici da Silva Castro Perdoná

USP – São Carlos
Janeiro de 2023

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

P586m

Picardi Morais de Castro, Patrícia
O Modelo Weibull de Longa Duração Inflacionado de Zero Aplicado à Pacientes Diagnosticados com Doença de Crohn / Patrícia Picardi Morais de Castro; orientadora Gleici da Silva Castro Perdoná. -- São Carlos, 2023.
71 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Sobrevivência. 2. Doença de Crohn. 3. Inflação de Zero. 4. Longa Duração. I. da Silva Castro Perdoná, Gleici, 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

Patrícia Picardi Morais de Castro

The Zero Inflated Long Term Weibull Model Applied to Patients Diagnosed with Crohn's Disease

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *FINAL VERSION*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Profa. Dra. Gleici da Silva Castro Perdoná

USP – São Carlos
January 2023

Este trabalho é dedicado a todas as pessoas que sofrem com a doença de Crohn, espero que meu trabalho possa, de alguma forma, ajudar a melhorar a vida de cada um de vocês.

Dedico também a todos os pesquisadores que estão por vir, vocês são o futuro da ciência.

AGRADECIMENTOS

Aos meus pais, **Saulo Morais de Castro** e **Cláudia Picardi Morais de Castro** e à minha irmã **Carini Picardi Morais de Castro**, por nunca terem desistido de mim e por todo apoio em cada sonho que busco realizar.

Ao meu namorado e amigo, **Bruno Fernandes Bispo**, por toda parceria e amor nos últimos anos, por acreditar que eu era capaz até quando eu mesma não acreditava em mim. Sem o seu apoio eu provavelmente teria desistido no meio do caminho.

A todos meus **amigos** e **familiares**, que comemoram comigo mais esta conquista e que sempre estiveram ao meu lado durante esta etapa da minha vida.

Á minha orientadora, **Profa. Dra. Gleici da Silva Castro Perdoná**, por ter me guiado durante este trabalho e por todo cuidado e humanidade durante o processo de construção desta tese. Foi por causa da sua orientação, que este trabalho pode ser concluído.

Ao **Dr. Sandro da Costa Ferreira (USP - FMRP)** por ter coletado e disponibilizado os dados dos pacientes diagnosticados com Crohn que fazem ou fizeram tratamento no Hospital das Clínicas de Ribeirão Preto.

A **todos os professores e funcionários do ICMC e da USP em geral**, por enfrentarem as adversidades que acompanham a profissão e pela dedicação aos alunos. É através do trabalho de vocês que o pensamento crítico e o estudo científico continuam vivos em nossa sociedade.

“In God we trust,
All others must bring data.”
(Dr. W. Edwards Deming)

RESUMO

PATRÍCIA, P. M. C. **O Modelo Weibull de Longa Duração Inflacionado de Zero Aplicado à Pacientes Diagnosticados com Doença de Crohn**. 2023. 71 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

A doença de Crohn é uma doença inflamatória crônica que se manifesta principalmente na região intestinal, se não tratada, a doença pode evoluir para quadros mais graves como a cirurgia de remoção total ou parcial do cólon ou intestino. Por se tratar de uma doença crônica, estudar os fatores de risco que podem piorar o quadro clínico do paciente, pode ajudar a melhorar a qualidade das pessoas que receberam tais diagnósticos e precisam conviver com Crohn para o resto de suas vidas. Modelos de sobrevida são comumente utilizados para encontrar associação entre informações de hábitos de vida e de saúde dos pacientes e a progressão dessas doenças. Porém, os modelos de sobrevivência convencionais pré-supõe que, para tempos suficientemente longos, todos os indivíduos terão sofrido o evento de interesse, caso contrário serão censurados. Além disso, estes modelos não se ajustam a indivíduos com tempo de sobrevida iguais a zero. Desta forma, pacientes que fazem a cirurgia no mesmo dia em que recebem o diagnóstico ou pacientes que nunca irão realizar cirurgia não são considerados nestes modelos. Em casos de estudos com pacientes diagnosticados com Crohn, ambas características podem estar presentes e, por este motivo, é necessário utilizar um modelo que se ajuste a inflação de zero e aos dados de longa duração. Esta dissertação propõe a utilização do modelo de sobrevivência chamado modelo Weibull Inflacionado de Zero de Longa Duração ou W-ZICR, para conseguir estimar corretamente as proporções de indivíduos curados e as proporções de indivíduos com tempos iguais a zero. O modelo foi aplicado para avaliar o risco e a sobrevida dos pacientes diagnosticados com doença de Crohn que fizeram acompanhamento no Hospital das Clínicas de Ribeirão Preto. De forma geral, concluiu-se que pacientes que fizeram uso de corticoide, pacientes fumantes, pacientes com forma da doença B2 ou B3 e pacientes com localização da doença L3 ou L4 possuem sobrevida menor e risco maior de realizar cirurgia quando comparados aos demais. Além disso, o modelo utilizado se ajustou bem aos dados inflacionados de zero e se mostrou uma excelente ferramenta para estudar pacientes diagnosticados com doença de Crohn.

Palavras-chave: Modelos Estatísticos, Doenças Intestinais, Colectomia, Enterectomia.

ABSTRACT

PATRÍCIA, P. M. C. **The Zero Inflated Long Term Weibull Model Applied to Patients Diagnosed with Crohn's Disease**. 2023. 71 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Crohn's disease is a chronic inflammatory disease that manifests itself mainly in the intestinal region, if left untreated, the disease can progress to more serious conditions such as surgery to remove all or part of the colon or intestine. As it is a chronic disease, studying the risk factors that can worsen the patient's clinical condition can help improve the quality of people who have received such diagnoses and need to live with Crohn's for the rest of their lives. Survival models are commonly used to find an association between information on patients' lifestyle and health habits and the progression of these diseases. However, conventional survival models presuppose that, for sufficiently long times, all individuals will have suffered the event of interest, otherwise they will be censored. In addition, these models do not fit individuals with zero survival times. Thus, patients who have surgery on the same day they receive the diagnosis or patients who will never undergo surgery are not considered in these models. In case studies with patients diagnosed with Crohn's, both characteristics may be present and, for this reason, it is necessary to use a model that fits zero inflation and long-term data. This dissertation proposes a survival model called the Long Term Zero Inflated Weibull model or W-ZICR, which can correctly estimate the proportions of cured individuals and the proportions of individuals with times equal to zero. The model was applied to assess the risk and survival of patients diagnosed with Crohn's disease who were followed up at the Hospital das Clínicas in Ribeirão Preto. In general, it was concluded that patients who used steroids, smokers, patients with B2 or B3 Montreal B-category and patients with L3 or L4 Montreal L-category have a lower survival rate and a higher risk of undergoing surgery when compared to the others. In addition, the proposed model adjusted well to zero-inflated data and proved to be an excellent tool for studying patients diagnosed with Crohn's disease.

Keywords: Statistical Models, Intestinal Diseases, Colectomy, Enterectomy.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A Doença de Crohn é uma doença inflamatória crônica, que afeta principalmente o trato gastrointestinal podendo manifestar complicações extra-intestinais e distúrbios imunológicos associados. Pacientes diagnosticados com a Doença de Crohn podem apresentar dor abdominal, febre, sinais clínicos de obstrução intestinal e diarreia com passagem de sangue e/ou muco. Sua incidência e prevalência vem crescendo em todos os grupos étnicos, além disso, por se tratar de uma doença sistêmica ela pode afetar todo o corpo humano, estes fatores fazem com que cada vez mais médicos se interessem pelo estudo da doença (BAUMGART; SANDBORN, 2012).

De acordo com (PARK *et al.*, 2017), a Doença de Crohn pode apresentar uma evolução, partindo de lesões inflamatórias iniciais a lesões mais graves que podem eventualmente progredir para uma cirurgia de ressecção e, em quadros ainda mais graves, o paciente pode desenvolver câncer de intestino. As cirurgias mais comuns realizadas em pacientes diagnosticados com a doença são as de remoção parcial ou total do cólon e do intestino, chamadas Colectomia e Enterectomia, respectivamente. Além disso, um estudo levantado por (JESS *et al.*, 2004) mostrou que a Doença de Crohn pode aumentar em até 60 vezes o risco de um paciente desenvolver câncer de intestino delgado, quando comparado a um paciente que não possui a doença.

Por ser uma doença crônica, pacientes diagnosticados com Crohn precisam conviver o resto da vida com ela e é extremamente importante que estes pacientes façam o acompanhamento e tratamento necessários para que a doença não evolua para uma eventual cirurgia. Encontrar os precursores da intervenção cirúrgica é primordial para garantir a qualidade de vida dos pacientes e para evitar o desenvolvimento de um quadro mais sério, como o câncer.

Modelos de sobrevivência são utilizados onde se deseja encontrar os fatores de risco que levam à evolução de um quadro clínico ao longo do tempo. A forma mais comum de aplicar esses modelos é quando se deseja estudar o tempo até a ocorrência de um evento, que na área médica normalmente significa a recorrência de uma doença, uma intervenção cirúrgica ou a morte de um paciente. Em estudos envolvendo a doença de Crohn, um modelo de sobrevivência pode ser

aplicado para avaliar o tempo até a ocorrência de uma cirurgia de ressecção, por exemplo.

Os modelos de sobrevida mais utilizados pré-supõem que, para tempos suficientemente longos, todos os pacientes terão apresentado o evento de interesse, caso isto não aconteça estes indivíduos serão censurados, em outras palavras, serão desconsiderados no estudo. Outra pré-suposição é que no tempo 0 (início do estudo ou momento em que o paciente passou a ser observado), a probabilidade do paciente ter sofrido o evento de interesse é igual a zero, o que significa que estes modelos não permitem que indivíduos com tempo de ocorrência do evento igual a zero sejam analisados (LAWLESS, 2011).

Porém, em alguns casos, o paciente diagnosticado com Crohn pode demorar anos até fazer a cirurgia ou até mesmo não sofrer nenhuma intervenção cirúrgica, ou seja, mesmo que se observe estes pacientes por tempos suficientemente longos, alguns deles podem não apresentar o evento de interesse. Em outra situação, o paciente pode chegar ao hospital com uma complicação séria e descobrir que possui a doença no momento da cirurgia, tornando seu tempo de ocorrência igual a zero. Desta forma, os modelos de sobrevida mais usuais foram adaptados para contornar estas limitações, criando-se modelos de sobrevivência com fator de cura e inflacionados de zero.

Os primeiros autores a proporem um modelo de sobrevida com fração de cura foram (BERKSON; GAGE, 1952) e (BOAG, 1949), em seu trabalho Berkson e Gage (1952) sugeriram que a população fosse dividida em dois subgrupos: os "curados" e os "não-curados". O primeiro representa os indivíduos que não irão sofrer o evento de interesse até o final do estudo, em contrapartida, o segundo grupo representa os indivíduos que estão suscetíveis ao desfecho. Modelos como os citados anteriormente são chamados modelos de mistura. Outro grande pioneiro dos modelos com fração de cura foi o trabalho de (CHEN; IBRAHIM; SINHA, 1999) ao criarem um modelo hierárquico baseado em modelos de mistura. O artigo discutiu a promoção de um modelo de sobrevivência de longa duração que permitiu calcular riscos proporcionais quando aplicado na presença de covariáveis, dentre outras vantagens do modelo de mistura. Ao passar dos anos, novos modelos de cura têm surgido se adaptando às mais diversas dificuldades encontradas nos dados, podemos citar (RAMOS; NASCIMENTO; LOUZADA, 2017), (LEÃO *et al.*, 2020) e (CANCHO *et al.*, 2021) como alguns trabalhos atuais envolvendo o tema.

Em relação aos dados inflados de zero, vários métodos já foram propostos para tratar dados com esta característica, podemos citar (LAMBERT, 1992), (HALL, 2000) e (RIDOUT; HINDE; DEMÉTRIO, 2001) como três dos trabalhos mais referenciados envolvendo o tema. Porém, a grande maioria dos artigos publicados na área não abrange dados de sobrevivência, se tornando um campo pouco explorado até o momento. Segundo (SOUZA *et al.*, 2021), atualmente, existem apenas 5 trabalhos publicados envolvendo inflação de zero em dados de sobrevivência, (BRAEKERS; GROUWELS, 2016) desenvolveu um modelo de regressão de Cox semi-paramétrico para tratar censuras à esquerda infladas de zero. Em seus dois trabalhos publicados, (LOUZADA; JR; MOREIRA, 2015) e (JR; MOREIRA; LOUZADA, 2017) propuseram um modelo baseado na distribuição Weibull para tratar dados bancários envolvendo questões de

fraude. No seu artigo, (CALSAVARA *et al.*, 2017) apresentou um modelo de taxa de cura flexível incorporando um termo de fragilidade no risco latente de uma distribuição Binomial Negativa. Por fim, em seu artigo, Souza (2021) trouxe um modelo de regressão de cura Log-Normal inflado de zero para tratar dados de mulheres africanas em trabalho de parto.

O objetivo deste trabalho é ajustar um modelo de cura inflacionado de zero chamado modelo Weibull de Longa Duração Inflacionado de Zero (W-ZICR), para descrever e estimar riscos associados a Doença de Crohn e a fração de cura em pacientes diagnosticados com ela. O modelo W-ZICR foi reduzido da distribuição Gama Generalizada e baseado no modelo de sobrevida com fração de cura de (BERKSON; GAGE, 1952) e no modelo Weibull inflacionado de zero de (LOUZADA; JR; MOREIRA, 2015).

Esta dissertação está divida em 7 capítulos sendo o primeiro a introdução. O segundo capítulo apresenta os objetivos gerais e específicos deste trabalho e a justificativa da importância do mesmo. O terceiro capítulo descreve o referencial teórico da dissertação e discursa sobre os métodos estatísticos aplicados neste trabalho. O quarto capítulo apresenta a especificação e descrição do modelo aplicado aos dados. No quinto capítulo, características do estudo são apresentadas, descrevendo as variáveis de estudo, quais métodos estatísticos foram abordados e descreve um resumo geral do local e população do estudo. Por fim, os capítulos 6 e 7 demonstram os resultados e conclusão deste trabalho, respectivamente.

CAPÍTULO

2

OBJETIVOS E JUSTIFICATIVAS

2.1 Objetivo Geral

Ajustar um modelo de sobrevivência de longa duração inflacionado de zero que consiga descrever e estimar os possíveis riscos associados à Doença de Crohn e seus efeitos em pacientes diagnosticados com ela.

2.2 Objetivos Específicos

- Detectar os possíveis fatores de risco que diminuem o tempo até a ocorrência de cirurgias de Enterectomia e Colectomia nos pacientes diagnosticados com Doença de Crohn;
- Ajustar um modelo que se adapte aos pacientes com tempo de ocorrência igual a zero e que considere que parte deles não sofrerá o desfecho.

2.3 Justificativa

Modelos de sobrevivência usuais pressupõem que, ao final do estudo, todos os indivíduos terão sofrido o evento, além disso, a probabilidade de um paciente não apresentar o evento no tempo zero é igual a 1. Porém, pacientes diagnosticados com doença de Crohn nem sempre sofrem o desfecho ao final do período de estudo e podem descobrir a doença no momento em que são levados para cirurgia.

Desta forma, é necessário utilizar um modelo que consiga se ajustar aos zeros inflacionados e aos indivíduos de longa duração para que não haja super ou subestimação do tempo de sobrevida dos pacientes. Além disso,a melhorar a qualidade de vida de pacientes com uma doença crônica que acomete o cólon e o intestino e possivelmente evitar que esta doença evolua para uma cirurgia de remoção total ou parcial do órgão é a principal motivação deste trabalho.

CAPÍTULO

3

REFERENCIAL TEÓRICO

3.1 Análise de Sobrevivência

A análise de sobrevivência é a área da estatística que estuda o tempo de falha, que pode ser caracterizado como o tempo até a ocorrência de um evento de interesse. Este evento vai desde o diagnóstico de uma doença, a morte de uma paciente, reprodução de uma bactéria ou até mesmo a falha de uma máquina. Nos dados deste estudo a variável resposta é o tempo até a cirurgia de um paciente diagnosticado com a doença de Crohn. A principal característica dos dados de sobrevivência é a presença de tempos incompletos, ou seja, situações onde o acompanhamento do objeto de estudo é interrompido ou que o evento não ocorre antes que se termine o estudo, estas situações são chamadas de censuras. As definições presentes na seção 3.1 foram todas extraídas do livro "*Análise de Sobrevivência Aplicada*", escrito por Colosimo e Giolo em 2006 (COLOSIMO; GIOLO, 2006).

Os dados de sobrevivência são caracterizados pela presença dos tempos de falha e da censura, que constituem a variável resposta. O tempo de falha é constituído pelo tempo inicial, a escala de medida e o evento de interesse (falha). O primeiro nada mais é que o tempo do início do estudo onde todos os indivíduos precisam ser comparáveis, normalmente definido como a data da aleatorização, data do diagnóstico ou até mesmo do início de um tratamento. Já a escala de medida normalmente é definida como o tempo real que se passou durante o estudo e o evento de interesse é o que chamamos de falha e, na maioria das vezes, é indesejado pelo pesquisador.

Os estudos clínicos de sobrevivência podem durar anos, mas mesmo assim, um estudo pode terminar sem que todos os indivíduos tenham sofrido a falha, nestes casos os indivíduos serão censurados. Para os indivíduos censurados, consideramos que o tempo da ocorrência do evento é maior que o tempo total de observação do estudo.

Na análise de sobrevivência existem três tipos de censura: censura do tipo I, do tipo II e a censura aleatória. Na censura do tipo I o estudo será finalizado após um período pré-estabelecido

de tempo. Já na censura do tipo II, o estudo só termina após um certo número de indivíduos terem falhado. Por fim, a censura aleatória ocorre quando um indivíduo é removido do estudo sem ter sofrido a falha. A censura aleatória pode ser definida por:

$$t = min(T, C), \tag{3.1}$$

onde T é uma variável aleatória que representa o tempo de falha de um indivíduo e C uma variável aleatória independente de T, que representa o tempo de censura associado a este indivíduo. Quando $T \leqslant C$ o indivíduo falhou antes do fim do estudo e quando $T > C$ o indivíduo foi censurado. Desta forma, os dados de sobrevivência para cada indivíduo $i(i = 1, ..., n)$ presente no estudo, pode ser definido como:

$$\delta_i = \begin{cases} 1, & T_i \leqslant C_i \\ 0, & T_i > C_i. \end{cases} \tag{3.2}$$

A variável aleatória T, definida em (3.1), normalmente é especificada em duas funções: a de sobrevivência ou a de risco. A função de sobrevivência é não crescente e pode ser definida como a probabilidade de um indivíduo sobreviver (não falhar) até um tempo t, e pode ser descrita como:

$$S(t) = P(T \geqslant t), \tag{3.3}$$

onde $S(0) = 1$ e o $\lim_{t\to\infty} S(t) = 0$, ou seja, todos os indivíduos estão vivos (ou não falharam) em $t = 0$ e quando o tempo for grande o suficiente, todos os indivíduos terão falhado. Além disso, definindo $F(t) = 1 - S(t)$ temos a função acumulada de sobrevivência, que descreve a probabilidade de um indivíduo não sobreviver ao tempo t. Já a função de risco representa a taxa de falha em um instante t, dado que o paciente sobreviveu até o tempo t e ela pode ser definida por:

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t \leqslant T < t + \Delta t \,|\, T \geqslant t)}{\Delta t}. \tag{3.4}$$

As funções de sobrevida (3.3) e de risco (3.4) definidas anteriormente podem ser relacionadas através da seguinte função:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(log(S(t)). \tag{3.5}$$

Para estimar a função de sobrevida $S(t)$, o método mais utilizado é o estimador de Kaplan-Meier $\hat{S}(t)$, que é uma adaptação da função de sobrevivência empírica. A função $\hat{S}(t)$ é

do tipo escada com degraus nos tempos observados de falha de tamanho $1/n$, onde n é o total de observações na amostra. O estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j<t} (1 - \frac{d_j}{n_j}), \tag{3.6}$$

onde $t_1 < t_2 < ...t_k$, são os k tempos distintos e ordenados de falha, d_j é o número de falhas em t_j, com $j = 1, ..., k$ e n_j é o número de indivíduos que ainda não falharam e não foram censurados até o instante imediatamente anterior a t_j.

O estimador Kaplan-Meier, proposto em 1958 por Edward L. Kaplan e Paul Meier (KAPLAN; MEIER, 1958), é bastante utilizado no meio científico pela sua capacidade de estimar probabilidades de um indivíduo sobreviver ao evento de interesse em um determinado tempo. Por exemplo, é possível calcular a probabilidade de um indivíduo sobreviver a uma doença após 2 meses de tratamento. Além disso, o estimador também pode ser utilizado para obter estimativas do tempo médio e mediano de vida de um paciente, ou até o tempo médio de vida restante daqueles pacientes que ainda não sofreram o evento até um determinado tempo. Dentre as principais propriedades do estimador podemos citar:

- $\hat{S}(t)$ é não-viciado para amostras grandes;
- $\hat{S}(t)$ é fracamente consistente;
- $\hat{S}(t)$ converge assintoticamente para uma gaussiana;
- $\hat{S}(t)$ é o estimador de máxima verossimilhança de $S(t)$.

Apesar de todas as suas características, o estimador de Kaplan-Meier possui algumas limitações. A principal delas é o fato de que o estimador não pode ser utilizado para associar a sobrevida a variáveis contínuas, apenas categóricas, impossibilitando estudos preditivos onde as variáveis preditores são contínuas. Uma solução é recorrer aos modelos probabilísticos paramétricos e não paramétricos, utilizando a função de sobrevida como uma distribuição de probabilidade para prever desfechos, assim como é feito nos modelos de regressão.

Dentre os modelos de regressão paramétricos podemos citar o Exponencial e Weibull, mais comumente utilizados na literatura. Porém, um outro modelo, semi-paramétrico, que também é muito utilizado em estudos clínicos por sua versatilidade é o modelo de riscos proporcionais proposto por Cox em 1972 (COX, 1972). O modelo permite ajustar dados onde a variável de estudo é o tempo de falha, considerando uma ou mais covariáveis.

O modelo proposto por Cox calcula a proporção entre as funções de taxa de falha de cada grupo estratificado pelas covariáveis, esta proporção é a razão das taxas de falha, também chamada de risco relativo. A grande popularidade deste modelo se dá pela presença de um componente não-paramétrico em sua função, o que faz com que ele se torne bastante flexível.

Outros modelos de regressão além do de Cox também foram propostos ao longo dos anos para se ajustar às mais diversas particularidades contidas nos dados. Dentre estas particularidades estão os dados de longa duração e dados inflacionados de zero que não podem ser modelados utilizando um modelo de sobrevida convencional. Como foi descrito nesta seção, o modelo convencional pré supõe que todos os indivíduos estão vivos no tempo 0 e que todos irão falhar se observarmos os dados por tempo suficiente. Porém, em dados de longa duração e/ou inflacionados de zero, o cenário é diferente. Estes dois casos serão discutidos nas seções a seguir.

3.1.1 Dados de Longa Duração

A análise de sobrevivência pré-supõe que os dados sob estudo são, em sua grande maioria, homogêneos, o que significa que todos os indivíduos condicionados às mesmas covariáveis possuem risco igual de sofrer o evento de interesse. Porém, alguns estudos são compostos por dados heterogêneos, onde os indivíduos possuem riscos diferentes de sofrer o desfecho. Além disso, dados de saúde podem apresentar observações com risco iguais a zero, isto se dá pelos avanços recentes da medicina que trouxeram uma melhora significativa nos tratamentos das doenças e em alguns casos até mesmo a cura (CANCHO *et al.*, 2021). Desta forma, espera-se que uma fração dos indivíduos do estudo seja curada e não sofra o evento de interesse, fazendo com a função de sobrevida estacione em um valor e nunca atinja o zero, como demonstrado na figura 1. Perceba que, no gráfico A, a curva de sobrevivência atinge o valor 0, enquanto a curva de sobrevida com fração de cura (B) estaciona no valor 0.2, indicando que aproximadamente 20% dos indivíduos não sofreram o evento de interesse.

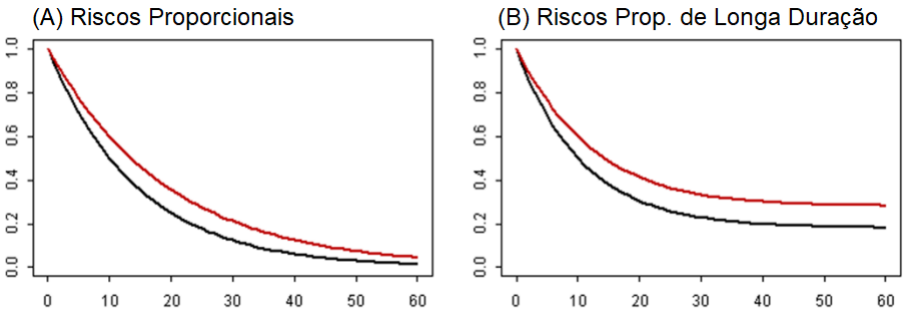

Figura 1 – Comparação das Curvas de Sobrevida com e sem Fração de Cura.

Fonte: Chen (2013).

Essa peculiaridade nos dados deu origem aos chamados modelos de longa duração ou modelos de fração de cura, uma vez que os modelos de sobrevivência convencionais não permitem a existência de indivíduos com risco iguais a zero. Um dos primeiros modelos de longa duração propostos na literatura e também a abordagem adotada neste trabalho, foi o modelo de mistura de Berkson e Gage (1952). Em 1952 os autores propuseram um modelo para estudar pacientes diagnosticados com câncer de mama, onde parte deles é curada pelo tratamento e

não vem a óbito (BERKSON; GAGE, 1952). Assim, a probabilidade do paciente sobreviver ao tratamento foi definida como:

$$S_{pop}(t) = pS_0(t) + (1-p), \tag{3.7}$$

onde p representa a proporção dos pacientes não curados, $1-p$ a proporção de pacientes curados e $S_0(t)$ representa a função de sobrevivência dos pacientes não curados.

No modelo proposto por Berkson e Gage o $\lim_{t\to\infty} S_{pop}(t) = 1-p$, ou seja, quando o tempo vai para infinito a probabilidade do indivíduo sofrer o evento é $1-p$. Diferente da função de sobrevida usual (3.3), onde a probabilidade do indivíduo sofrer o evento, quando tempo vai para infinito, é igual a 0.

Para estimar a função de sobrevida dos pacientes não-curados ($S_0(t)$), Berkson e Gage utilizaram a distribuição Exponencial, porém, ao decorrer dos anos outros autores propuseram outras distribuições, por vezes mais flexíveis que a Exponencial, possibilitando melhores ajustes aos dados reais. Neste trabalho iremos propor a utilização da distribuição Gama Generalizada que será apresentada no capítulo 4.

3.1.2 Dados Inflacionados de Zero

Como visto em (3.3), os modelos de sobrevivência usuais pré-supõem que todos os indivíduos não terão sofrido o desfecho no início do estudo, o que faz sentido, pensando em termos de estudos clínicos onde o evento de interesse, em geral, é o óbito dos pacientes, tornando inválido ou até mesmo insensível considerarmos tempos de sobrevivência iguais a zero (LOUZADA; JR; MOREIRA, 2015).

Apesar de não ser usual, em alguns estudos clínicos onde o desfecho não é o óbito, o indivíduo pode sofrer o evento de interesse assim que o tempo de observação se inicia. Por exemplo, alguns pacientes diagnosticados com a doença de Crohn recebem seu diagnóstico no mesmo dia em que realizam a cirurgia de remoção do cólon ou intestino, caracterizando os dados com tempos iguais a zero.

A figura 2 ilustra uma a função de sobrevivência que caracteriza a problemática dos tempos iguais a zero, nela é possível verificar que a curva inicia-se em torno do valor 0.9, diferente das curvas ilustradas na figura 1 que se iniciam em 1. Isto se dá porque cerca de 10% dos indivíduos tiveram tempos iguais a zero, ou seja, sofreram o desfecho ao início do estudo e, para estes casos, é necessário a utilização de um modelo de sobrevivência que os zeros na sua formulação, portanto um modelo de sobrevivência inflacionado de zero.

Motivados por dados de risco de crédito, onde parte dos clientes nunca vai se tornar inadimplente e outra parte solicita o empréstimo apenas para fraudar a instituição, nunca pagando nenhuma parcela do crédito concedido, (LOUZADA; JR; MOREIRA, 2015) propuseram um

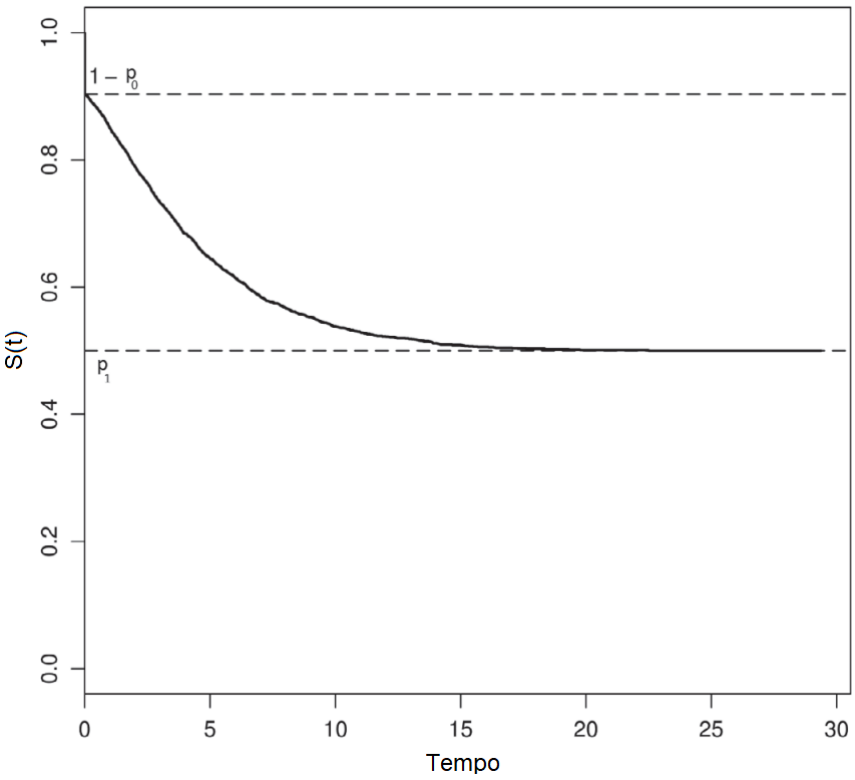

Figura 2 – Função de Sobrevivência do Modelo de Longa Duração com Inflação de Zero.

Fonte: Louzada, Jr e Moreira (2015).

modelo que se ajustasse aos dados inflacionados de zero e que também possuem fração de cura. Inspirados pelo modelo de fração de cura de (BERKSON; GAGE, 1952), Louzada, Oliveira Jr. e Moreira descreveram a probabilidade de um cliente se tornar inadimplente como:

$$S_{pop}(t) = p_1 + (1 - p_0 - p_1)S_0(t), \tag{3.8}$$

onde p_0 é a proporção de indivíduos com tempos iguais a zero, p_1 é a proporção de indivíduos curados e S_0 é a função de sobrevivência relacionada à proporção $(1 - p_0 - p_1)$ de indivíduos suscetíveis a inadimplência.

Este modelo foi chamado de modelo de taxa de cura com inflação de zero e satisfaz a condição de que $\lim_{t\to\infty} S_{pop}(t) = p_1 > 0$ e de que $S_{pop}(0) = 1 - p_0 < 1$. Repare que, se o modelo não possuir inflação de zero, ou seja $p_0 = 0$, voltamos para o modelo de taxa de cura de Berkson e Cage.

3.2 Teste de Log-Rank

Para compararmos grupos ou duas curvas de sobrevida, o teste de log-rank é amplamente utilizado, particularmente em testes clínicos na presença de dados censurados. O teste de log-rank

é um teste de hipóteses não paramétrico que tem como objetivo comparar a função de sobrevida de duas amostras/grupos.

O teste baseia-se em comparar se a distribuição dos acontecimentos observados (dados do estudo) em cada grupo é igual a distribuição dos eventos esperados, se os grupos fossem iguais (JM; ALTMAN, 2004). Ou seja, se baseia na diferença entre as falhas ocorridas nos dois grupos e aquelas falhas que seriam esperadas se os dois grupos fossem iguais. Se essa diferença for próxima de zero há um indício a favor da hipótese de igualdade das curvas.

A diferença entre as falhas observadas e esperadas é avaliada por meio do teste do qui-quadrado.

Segue o o procedimento para compararmos duas curvas de sobrevida:

1. Devemos considerar o tempo de falha de um dos grupos.

2. Para cada tempo t_j, onde $j = 1,...,J$ observa-se quantas ($d_j =$) falhas e quantas ($n_j = n_{1j} + n_{2j}$) pessoas sob risco nesse intervalo de tempo em ambas as amostras, onde n_{ij} representa o número de pessoas sob risco na amostra i .

3. Sob a hipótese nula (os dois grupos têm sobrevivência idênticos e funções de risco proporcionais) tem a distribuição hipergeométrica com parâmetros n_j,n_{ij} e d_j.

4. A estatística de teste é dada por:

$$T = \frac{\sum_{j=1}^{J}(d_{1j} - E_{1j})^2}{\sum_{j=1}^{J} V_j} \tag{3.9}$$

onde;

$$E_{1j} = \frac{d_j n_j}{n_j} \tag{3.10}$$

e,

$$V_j = \frac{d_j(n_{1j}/n_j)(1 - n_{1j}/n_j)(n_j - d_j)}{n_j - 1} \tag{3.11}$$

5. T possui distribuição qui-quadrado com 1 grau de liberdade para amostras grandes;

6. Rejeitamos a hipótese de igualdade se a estatística de teste T retornar valores grandes, ou seja se $T > \chi_{1-\alpha}$ e $\chi_{1-\alpha}$ é quantil de probabilidade da distribuição Qui-quadrado com 1 grau de liberdade;

7. Analisando pelo p-valor, rejeitamos H_0, se $p - valor < \alpha$;

3.3 Métodos de Seleção de Modelos

3.3.1 Métodos Gráficos

Existem vários métodos gráficos para auxiliar na escolha do modelo probabilístico para o tempo do evento. Um deles é comparar a curva de sobrevida do modelo proposto com o estimador de empírico de Kaplan-Meier. Para construir a comparação gráfica é necessário plotar uma reta $y = x$ e um gráfico de dispersão onde as estimativas de sobrevivência obtidas pelo modelo proposto irão representar o eixo y e as estimativas de Kaplan-Meier o eixo x. Desta forma, quanto mais próximas da reta os pontos do gráfico de dispersão estiverem, mais adequado aos dados será o modelo proposto. A figura 3 expõe um exemplo de como comparar alguns modelos utilizando esta técnica, observe que neste caso, os pontos dos modelos Weibull e log-normal se aproximam mais da reta $y = x$, indicando que estes modelos se ajustaram melhor aos dados do exemplo comparado com a figura do modelo exponencial.

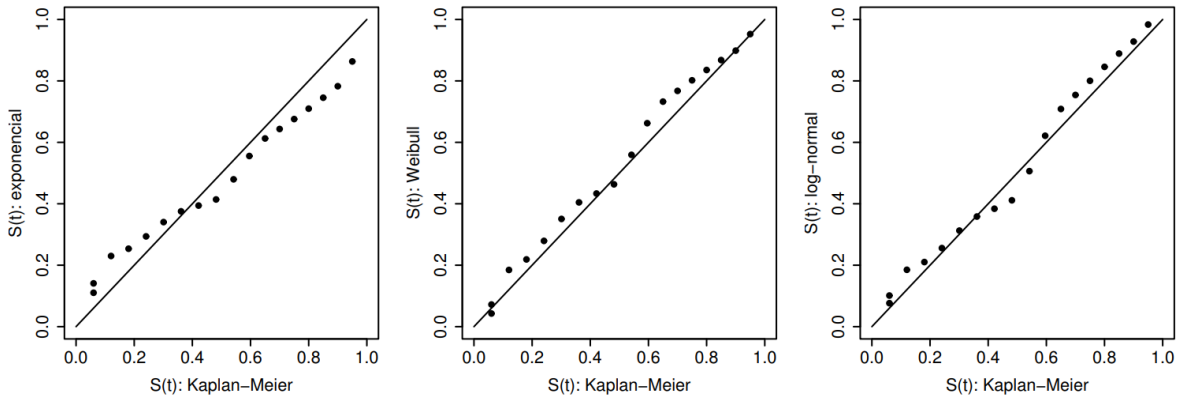

Figura 3 – Gráficos das Sobrevivências Estimadas por Kaplan-Meier versus as Sobrevivências Estimadas Pelos Modelos Exponencial, Weibull e Log-Normal.

Fonte: Colosimo e Giolo (2006).

Outro método descrito por (COLOSIMO; GIOLO, 2006) é o método da linearização. Nesta técnica realiza-se a linearização da função de sobrevivência, o que faz com que, ao ser plotada, a curva de sobrevivência se aproxima de uma reta, crescente ou decrescente, caso o modelo proposto seja adequado aos dados. Com esse método, caso haja violação de linearidade, o gráfico mostrará facilmente. A figura 4 apresenta um exemplo de como comparar modelos utilizando sobrevivências linearizadas, observe que os pontos dos modelos Weibull e log-normal, novamente, se mostraram mais adequados aos dados do exemplo, uma vez que o ajuste exponencial gera um gráfico que se distancia do formato de uma reta.

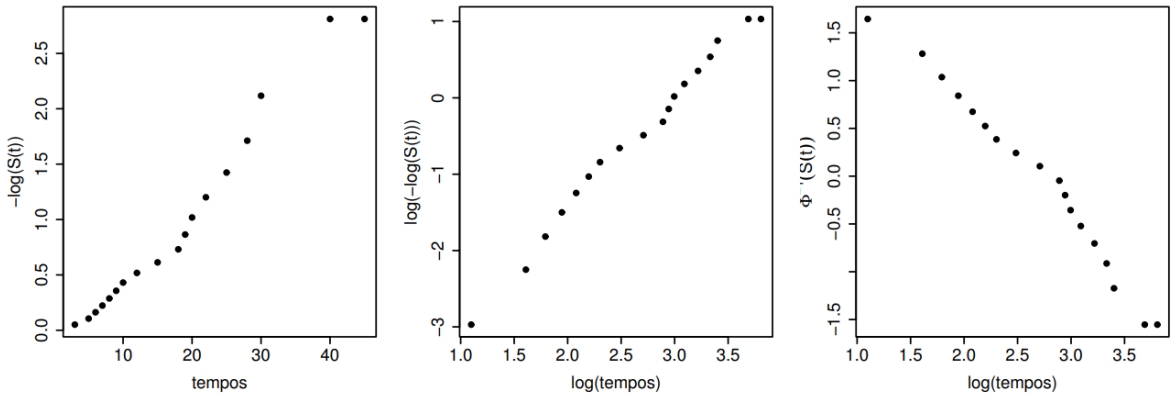

Figura 4 – Gráficos Linearizados dos Modelos Exponencial, Weibull e Log-Normal.

Fonte: Colosimo e Giolo (2006).

As transformações lineares dos modelos exponencial, Weibull e log-normal estão descritas a seguir.

- **Modelo exponencial:**

A linearização da função de sobrevivência do modelo exponencial é definida por:

$$-\log[S(t)] = \frac{t}{\alpha} = (\frac{1}{\alpha})t, \tag{3.12}$$

onde $\hat{S}(t)$ é o estimador de Kaplan-Meier e $-\log[S(t)]$ é uma função linear de t, portanto o gráfico ($x = t$, $y = -\log[\hat{S}(t)]$) deve ser aproximadamente linear e passar pela origem caso o ajuste do modelo seja apropriado aos dados.

- **Modelo Weibull:**

A linearização da função de sobrevivência do modelo Weibull com parâmetros (γ, α) é defina como:

$$-\log[S(t)] = (\frac{t}{\alpha})^{\gamma} \tag{3.13}$$

$$\log[-\log[S(t)]] = -\gamma\log(\alpha) + \gamma\log(t) \tag{3.14}$$

onde $\hat{S}(t)$ é o estimador de Kaplan-Meier e $\log[-\log[S(t)]]$ é uma função linear de t, portanto o gráfico ($x = \log(t)$, $y = \log[-\log[\hat{S}(t)]]$) deve ser aproximadamente linear caso o ajuste do modelo Weibull seja apropriado aos dados.

- **Modelo log-normal:**

Por fim, a linearização da função de sobrevivência do modelo log-normal é dada por:

$$\phi^{-1}[S(t)] = \frac{-\log t + \mu}{\sigma} \tag{3.15}$$

onde $\hat{S}(t)$ é o estimador de Kaplan-Meier e $\phi^{-1}[.]$ representa os percentis da distribuição Normal padrão, portanto o gráfico ($x = \log(t)$, $y = \phi^{-1}[\hat{S}(t)]$) deve ser aproximadamente linear com intercepto μ/σ e inclinação $-1/\sigma$ caso o ajuste do modelo log-normal seja apropriado aos dados.

3.3.2 Critérios de Akaike (AIC)

Outra técnica bastante utilizada na seleção e comparação de modelos é o critério AIC, ou *Akaike Information Criterion* proposta pelo japonês Hirotugu Akaike em 1974 (AKAIKE, 1974). Esta técnica consiste em mensurar a qualidade de um modelo levando em conta o equilíbrio entre a qualidade e a parcimônia do modelo, ou seja, o critério de AIC avalia tanto a qualidade do ajuste quanto a quantidade de variáveis que este modelo possui. O cálculo do valor de AIC é dado por:

$$AIC_i = 2k - 2\log(L_i), \tag{3.16}$$

onde k é o número de parâmetros estimados pelo modelo proposto e L_i é a verossimilhança do $i-simo$ modelo. Quanto melhor o ajuste do modelo menor será o valor do AIC, portanto, ao comparar dois ou mais modelos aquele com menor valor de AIC provavelmente terá o ajuste mais adequado aos dados.

3.3.3 Gráfico de TTT

As comparações gráficas para seleção de modelos que foram discutidas nas seções anteriores, utilizam as curvas de sobrevivência estimadas pelo modelo proposto e as comparam com as curvas estimadas pelo método de Kaplan-Meier. Porém, a curva do risco também pode ser utilizada quando deseja-se comparar modelos ou entender quais ajustes se adaptam melhor aos dados. Em 1975, Barlow e Campo (BARLOW; CAMPO, 1975) propuseram em seu artigo "*Total time on test processes and applications to failure data analysis*"um novo método para analisar dados que utiliza o conceito de Tempo Total em Teste (TTT) para plotar gráficos que auxiliam na escolha de um modelo probabilístico.

A estatística TTT calculada a partir de uma função de distribuição acumulada (FDA) $F(t)$ é definida como:

$$H_F^{-1}(u) = \int_0^{Q(u)} (1 - F(t))dt, \tag{3.17}$$

onde $Q(.)$ é uma função quantílica e $0 < u < 1$. Dado um conjunto de dados não censurados e seja $T_{(1)}, T_{(2)}, ..., T_{(n)}$ os n tempos completos e não censurados, a estatística TTT_i para o i-ésimo tempo completo pode ser definida como:

$$TTT_i = \sum_1^n T_{(i)} + (n-i)T_{(i)}, \tag{3.18}$$

onde $i = 1, 2, ..., n$ e $T_0 = 0$.

Sendo assim, o gráfico do Tempo Total em Teste ou gráfico do TTT é composto por $y = TTT_i/TTT_n$ e $x = i/n$. Em sua tese de doutorado, (SOUZA, 2015) trouxe uma comparação entre as formas que o gráfico de TTT pode assumir, representado aqui pela figura 5. Quando o risco for constante o formato do gráfico será o de uma reta $y = x$ (A), o gráfico assumirá um formato convexo (B) caso o risco seja crescente e côncavo (C) se for decrescente, se o risco tiver forma de banheira seu formato será convexo e depois côncavo (D) e caso tenha forma unimodal terá o formato contrário ao anterior (E).

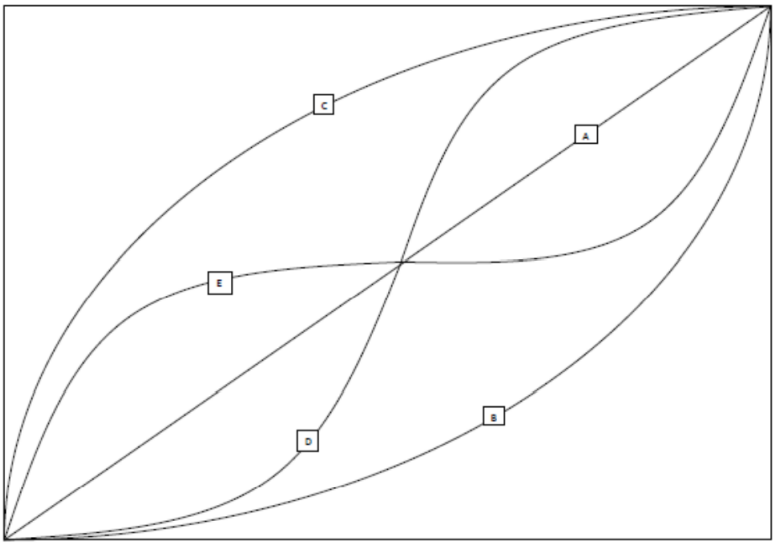

Figura 5 – Exemplos das Curvas do Gráfico de TTT.

Fonte: Souza (2015).

O formato da curva do gráfico de TTT nos auxilia na escolha do modelo adequado, por exemplo se a curva for do tipo convexo, ou seja, possuir risco crescente, é um indicativo de que o modelo Weibull pode ser o modelo indicado para realizar um ajuste aos dados.

CAPÍTULO

4

O MODELO GAMA GENERALIZADO DE LONGA DURAÇÃO INFLACIONADO DE ZERO

4.1 Especificação do Modelo

Para acomodar o excesso de zeros e a taxa de cura nos dados de sobrevivência do problema das doenças intestinais, este trabalho utilizará o modelo de longa duração inflacionado de zero definido em 3.8. A Função (imprópria) de Distribuição Acumulada (FDA) e a Função Densidade de Falha (FDP) associadas ao modelo de longa duração inflacionado de zero estão definidas abaixo nas equações 4.1 e 4.2.

$$F(t) = p_0 + (1 - p_0 - p_1)F^*(t),\ t \geqslant 0, \tag{4.1}$$

$$f(t) = \begin{cases} p_0, & se\ t = 0 \\ (1 - p_0 - p_1)f^*(t), & se\ t > 0. \end{cases} \tag{4.2}$$

Além disto, como visto anteriormente temos por definição que $F(t) = 1 - S(t)$, sendo assim, o modelo de longa duração inflacionado de zero descreve a probabilidade do paciente não sofrer a cirurgia como:

$$S(t) = p_1 + (1 - p_0 - p_1)S^*(t). \tag{4.3}$$

Os parâmetros p_0 e p_1 são os mesmos definidos em 3.8, $S^*(t)$, $F^*(t)$ e $f^*(t)$ são as funções sobrevivência, FDA e FDP, respectivamente, que descrevem a proporção $(1 - p_0 - p_1)$ de pacientes suscetíveis a cirurgia.

4.2 O Modelo Gama Generalizado de Longa Duração Inflacionado de Zero (GG-ZICR)

Uma distribuição bastante versátil e que pode ser utilizada para descrever a probabilidade dos indivíduos suscetíveis a falha é a Gama Generalizada (GG), introduzida inicialmente por (STACY, 1962). Uma nova parametrização das funções densidade de probabilidade (FDP) e de sobrevivência ($S^*_{GG}(t)$) da distribuição GG, definida como:

$$f^*_{GG}(t) = \frac{\lambda}{\sigma t}\frac{exp(\lambda^{-2}(\lambda w + log(\lambda^{-2})) - exp(\lambda w + log(\lambda^{-2}))}{\Gamma(\lambda^{-2})} \tag{4.4}$$

e

$$S^*_{GG}(t) = 1 - \Gamma_1\left[\lambda^{-2}exp(\lambda w); \lambda^{-2}\right]. \tag{4.5}$$

onde $t > 0$, $w = \frac{log(t)-\mu}{\sigma}$, $\Gamma_1(\nu, k)$ é uma função gama incompleta, $-\infty < \mu < \infty$, $\sigma > 0$ e $\lambda > 0$ (MEEKER; ESCOBAR, 1998).

O modelo Gama Generalizado de longa duração inflacionado de zero (GG-ZICR) é obtido quando a distribuição GG é utilizada para descrever o comportamento de sobrevivência da variável aleatória não negativa T, que representa o tempo até o desfecho dos indivíduos suscetíveis a falha. Baseado no modelo proposto por (LOUZADA; JR; MOREIRA, 2015), a função de sobrevivência e a função densidade de probabilidade GG-ZICR são dadas por:

$$S_{pop}(t) = p_1 + (1 - p_0 - p_1)\left(1 - \Gamma_1\left[\lambda^{-2}exp(\lambda w); \lambda^{-2}\right]\right). \tag{4.6}$$

e

$$f_{pop}(t) = p_0 I_{\{0\}}(t) + (1 - p_1 - p_0)\frac{\lambda}{\sigma t}\frac{exp(\lambda^{-2}(\lambda w + log(\lambda^{-2})) - exp(\lambda w + log(\lambda^{-2}))}{\Gamma(\lambda^{-2})} I_{\{\mathbb{R}^+\}}(t), \tag{4.7}$$

onde p_0 em $(0, 1 - p_1)$ é a proporção de indivíduos com tempos iguais a zero e p_1 em $(0, 1)$ é a proporção de indivíduos curados.

4.2.1 A Função de Verossimilhança

A estimativa pontual é baseada na máxima verossimilhança, assumindo um processo de censura aleatório independente, onde o i-ésimo indivíduo tenha um tempo de vida T_i e um tempo de censura C_i, com $t_i = min(T_i, C_i)$ e $i = 1, ..., n$. Em seu artigo, (LOUZADA; MOREIRA; OLIVEIRA, 2018) definiram a função de verossimilhança para os dados observados como:

$$\begin{aligned} L(\vartheta, t_i, \delta_i) &= \prod_{i:\, t_i=0} p_0 \prod_{i:\, t_i>0} \left[\{(1-p_0-p_1) f_{GG}(t_i)\}^{\delta_i} \{p_1 + (1-p_0-p_1) S_{GG}(t_i)\}^{1-\delta_i} \right] \\ &= {p_0}^m \prod_{i:\, t_i>0} \left[\{(1-p_0-p_1) f_{GG}(t_i)\}^{\delta_i} \{p_1 + (1-p_0-p_1) S_{GG}(t_i)\}^{1-\delta_i} \right] \end{aligned}$$

onde $m(< n)$ é o número de indivíduos com $t = 0$ e $\vartheta = (p_0, p_1, \mu, \sigma, \lambda)$ o vetor de parâmetros do modelo. O logaritmo da verossimilhança é dado por

$$\begin{aligned} l(\vartheta, t_i, \delta_i) &= \log\{L(\vartheta, t_i, \delta_i)\} \\ &= m \log(p_0) + \sum_{i:\, t_i>0} \delta_i \log(1-p_0-p_1) \\ &\quad + \sum_{i:\, t_i>0} \delta_i \log \left[\frac{\lambda}{\sigma t} \frac{exp(\lambda^{-2}(\lambda w + log(\lambda^{-2})) - exp(\lambda w + log(\lambda^{-2}))}{\Gamma(\lambda^{-2})} \right] \\ &\quad + \sum_{i:\, t_i>0} (1-\delta_i) \log \left\{ p_1 + (1-p_0-p_1) \left(1 - \Gamma_1 \left[\lambda^{-2} exp(\lambda w_i); \lambda^{-2}\right]\right) \right\} \end{aligned}$$

onde $w_i = \frac{log(t_i)-\mu}{\sigma}$, $\delta_i = 0$ se $T_i < C_i$ é um tempo censurado e $\delta_i = 1$ se T_i for um tempo completo. Ao resolver o sistema de equações não-linear $\left(\frac{\partial l(\vartheta)}{\partial p_0}, \frac{\partial l(\vartheta)}{\partial p_1}, \frac{\partial l(\vartheta)}{\partial \mu}, \frac{\partial l(\vartheta)}{\partial \sigma}, \frac{\partial l(\vartheta)}{\partial \lambda}\right) = 0$ para uma amostra, as estimativas de máxima verossimilhança (emv) são alcançadas.

A utilização de algoritmos iterativos são uma alternativa para solucionar as equações já que os problemas aritméticos podem ser um desafio. O procedimento *optim()*, que pode ser encontrado no pacote estatístico R (TEAM, 2013), foi utilizado neste estudo.

4.2.2 O Modelo de Regressão

Uma das finalidades dos estudos clínicos é entender como as particularidades de cada paciente (covariáveis) se associam ao desfecho. Os modelos de regressão são uma ferramenta útil na prática pois permitem aos pesquisadores validar o impacto de uma ou mais características individuais na sobrevivência, o que auxilia na tomada de decisão. O modelo de regressão GG-ZICR descreve o desfecho do evento com base nas variáveis preditoras, desta forma, é possível reescrever a sobrevivência definida em 4.6 a partir de um vetor de k covariáveis:

$$S(t|x = (x_1, x_2, ..., x_k)) = p_1(x) + (1 - p_0(x) - p_1(x)) S_{GG}(t|x),\ t \geqslant 0. \tag{4.8}$$

Assim, o modelo de regressão GG-ZICR obtido a partir da função de sobrevivência 4.6 pode ser definido como

$$S_i(t) = p_{1i} + (1 - p_{0i} - p_{1i}) \left(1 - \Gamma_1 \left[{\lambda_i}^{-2} exp\left(\lambda_i \frac{log(t) - \mu_i}{\sigma_i}\right); {\lambda_i}^{-2}\right]\right). \tag{4.9}$$

As funções link são definidas de acordo com o espaço paramétrico de cada parâmetro,

$$\begin{cases} \mu_i & = & \eta_{1i}, \\ log(\sigma_i) & = & \eta_{2i}, \\ log(\lambda_i) & = & \eta_{3i}, \\ \left(\log\left(\frac{p_{0i}}{1-p_{0i}-p_{1i}}\right), \log\left(\frac{p_{1i}}{1-p_{0i}-p_{1i}}\right)\right) & = & (\eta_{4i}, \eta_{5i}), \end{cases} \tag{4.10}$$

onde o preditor linear para o i-ésimo indivíduo é dado por,

$$\eta_{ji} = \beta_{j0} + \beta_{j1}x_{1i} + \beta_{j2}x_{2i} + ... + \beta_{jk}x_{ki} = \beta_{j0} + \sum_{c=1}^{k} \beta_{jc}x_{ci}, \tag{4.11}$$

em que x_{ci} é o valor da c-ésima variável para o i-ésimo indivíduo, $j = 1,2,3,4,5$, $c = 1,2,...,k$ e $i = 1,2,...,n$ e os coeficientes lineares β_{jc} são baseados no procedimento de máxima verossimilhança. Obtendo as estimativas intervalares de β_{jc}, por procedimento padrão assintótico de intervalo de confiança, considerando a aproximação $(1-\upsilon)$ 100% intervalo de confiança (IC) para β_{jc} dado por, $\hat{\beta_{jc}} \pm z_{1-\frac{\upsilon}{2}}\sqrt{Var(\hat{\beta_{jc}})}$, onde $z_{1-\frac{\upsilon}{2}}$ representa a $(1-\frac{\upsilon}{2})\%$ o quantil da normal padrão e Var($\hat{\beta_{jc}}$) é obtido via a matriz de informação observada.

4.2.3 Casos Particulares da GG-ZICR

A principal característica da distribuição Gama Generalizada é que sua função representa uma família paramétrica de distribuições que, dependendo dos valores assumidos pelos seus parâmetros, podem se tornar casos particulares de outras distribuições. Desta maneira, a distribuição GG-ZICR também pode ser reduzida a outras distribuições, como a log-Normal ZICR quando $\lambda \to 0$, a exponencial ZICR quando $\lambda = \sigma = 1$ e também a Weibull ZICR ($\lambda = 1$) utilizada por (LOUZADA; MOREIRA; OLIVEIRA, 2018) em seu artigo. Este é um benefício significativo, pois o modelo GG-ZICR pode ser ajustado e avaliado para testar se um determinado cenário mostra um ajuste superior, calculando estimativas de parâmetros e utilizando testes de hipóteses ou métodos de comparação.

4.2.3.1 O Caso Weibull-ZICR

Partindo da distribuição GG-ZICR, o modelo escolhido para estudar os dados dos pacientes diagnosticados com Crohn e que será utilizado como base nesta dissertação é o Weibull de longa duração inflacionado de zero (Weibull-ZICR). Ele representará o comportamento aleatório da variável aleatória não negativa T, que descreve o tempo até a cirurgia nos pacientes diagnosticados com a doença. A Função de Distribuição Acumulada (FDA) da distribuição Weibull é dada por:

$$F_w^*(t) = 1 - \exp(-\frac{t}{\sigma})^{\mu},\ t \geqslant 0, \tag{4.12}$$

onde $\mu > 0$ e $\sigma > 0$ são, respectivamente, os parâmetros de forma e escala da função. A Função Densidade de Probabilidade (FDP) derivada da função FDA da distribuição Weibull é definida como:

$$f_w^*(t) = \frac{\mu}{\sigma}(\frac{t}{\sigma})^{\mu-1}\exp(-\frac{t}{\sigma})^{\mu},\ t \geqslant 0. \tag{4.13}$$

Por definição $F(t) = 1 - S(t)$, sendo assim, o modelo Weibull-ZICR descreve a probabilidade do paciente não sofrer a cirurgia como:

$$S_W^*(t) = p_1 + (1 - p_0 - p_1)\exp(-\frac{t}{\sigma})^{\mu}. \tag{4.14}$$

O modelo Weibull-ZICR pretende classificar os pacientes em 3 subgrupos de acordo com o seu tempo de sobrevivência e a sua probabilidade de sofrer a falha: pacientes que não farão a cirurgia durante o período de observação, ou seja, curados ou de longo prazo, pacientes que farão a cirurgia depois de um certo período após o diagnóstico e por fim, pacientes que receberão o diagnóstico no dia em que realizarem a cirurgia. Esta subdivisão é importante para que o modelo consiga estimar de forma correta a proporção de pacientes realmente suscetíveis à realização da cirurgia. Se ignorarmos a fração de indivíduos curados e com tempos iguais a zero, o modelo pode sub ou superestimar a verdadeira taxa de pacientes suscetíveis à falha.

CAPÍTULO

5

DADOS SOBRE A DOENÇA DE CROHN

5.1 Delineamento do Estudo

Este estudo é caracterizado como uma coorte retrospectiva e atende aos pré-requisitos éticos de estudos realizados com seres humanos. O estudo foi aprovado pelo comitê médico de ética da Faculdade de Medicina de Ribeirão Preto *(protocolo Nº 3147/2019/03/23/2019)*. Além disso, todos os pacientes concordaram em participar do estudo e forneceram consentimento informado por escrito.

5.2 Local e População de Estudo

A população alvo deste estudo são pacientes diagnosticados com Doença de Chron que fazem acompanhamento da doença no Hospital das Clínicas de Ribeirão Preto (HCRP) nos últimos 35 anos, e que realizaram ou não cirurgia durante este tempo.

A coleta dos dados utilizada nesta pesquisa foi realizada pelo HCRP, com início no ano de 2015 e permanece ativa até os dias de hoje. Para este estudo, foram selecionados apenas os pacientes que participaram da pesquisa entre os anos de 2015 e 2020.

Os dados deste trabalho fazem parte de um estudo clínico coordenado pelo pesquisador Dr. Sandro da Costa Ferreira (USP - FMRP) que acompanhou a coleta dos dados e o quadro clínico dos pacientes. Além disso, a pesquisa foi financiada pela Fundação de Apoio e Ensino e Pesquisa do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (FAEPA).

5.3 Descrição das variáveis do estudo

A tabela 1 expõe as variáveis preditoras selecionadas. Além delas, temos também o tempo de segmento, que representa o tempo em semanas entre a admissão e a cirurgia ou tempo entre a

Tabela 1 – Descrição das Variáveis de Estudo.

Variável de Estudo	**Tipo**	**Descrição**
Tempo (em semanas)	Numérica	Tempo de segmento do paciente
Sexo	Categórica	Gênero do paciente
Cor	Categórica	Raça auto-declarada do paciente
Tabagismo	Categórica	Se o paciente fuma ou não
Localização	Categórica	Localização da doença
Doença Perianal	Categórica	Se o paciente possui penetração perianal
Corticoides	Categórica	Se o paciente faz uso de corticoides
H. Familiar de DII	Categórica	Se o paciente possui histórico familiar da doença
Forma da Doença	Categórica	Forma identificada da doença
Manif. Extra Int.	Categórica	Se o paciente possui manifestação extra-intestinal da doença
Biológicos	Categórica	Se o paciente faz uso de biológicos
Idade Diag.	Numérica	Idade do paciente quando a doença foi diagnosticada
Idade Atual	Numérica	Idade atual do paciente

admissão e os dias atuais, para os casos onde não houve cirurgia. A variável "Idade Diag."foi categorizada em menor que 17 anos, entre 17 e 40 anos e acima de 40 anos. A figura 6 apresenta a ilustração da forma patológica da doença de Crohn, as variáveis de estudo "Localização", "Doença Perianal", "Forma da Doença"e "Manifestação Extra Intestinal"estão descritas nesta ilustração.

Uma vez que o paciente recebe o diagnóstico da doença, ela deve ser caracterizada de acordo com a classificação de Montreal (figura 6), que dirá onde a doença está localizada (categoria "L") e qual a sua forma (categoria "B"). A localização, que pode assumir valores de L1 a L4, descreve em quais partes do intestino e do cólon a doença foi encontrada. Já a forma da doença, indo de B1 a B3, descreve de qual maneira a doença afetou o cólon e o intestino, neste trabalho a classificação B3p foi atribuída a uma nova variável, "Doença Perianal", indicando se o paciente possui ou não este tipo de penetração da doença. Além disto, o paciente se submete a uma série de exames para entender se a doença manifestou alguma complicação extra-intestinal, os tipos mais comuns de outras manifestações da doença, fora a região intestinal, estão descritos na parte "B" da figura 6.

Como variável resposta, ou *outcome*, utilizou-se a incidência de uma das duas cirurgias observadas no estudo: enterectomia e colectomia. Se o paciente realizar uma das duas ou ambas as cirurgias a variável resposta recebe 1, caso contrário, 0. O tempo foi calculado individualmente para cada paciente, nos casos em que o paciente realizou ambas as cirurgias utilizou-se a data da primeira cirurgia para calcular o tempo de segmento.

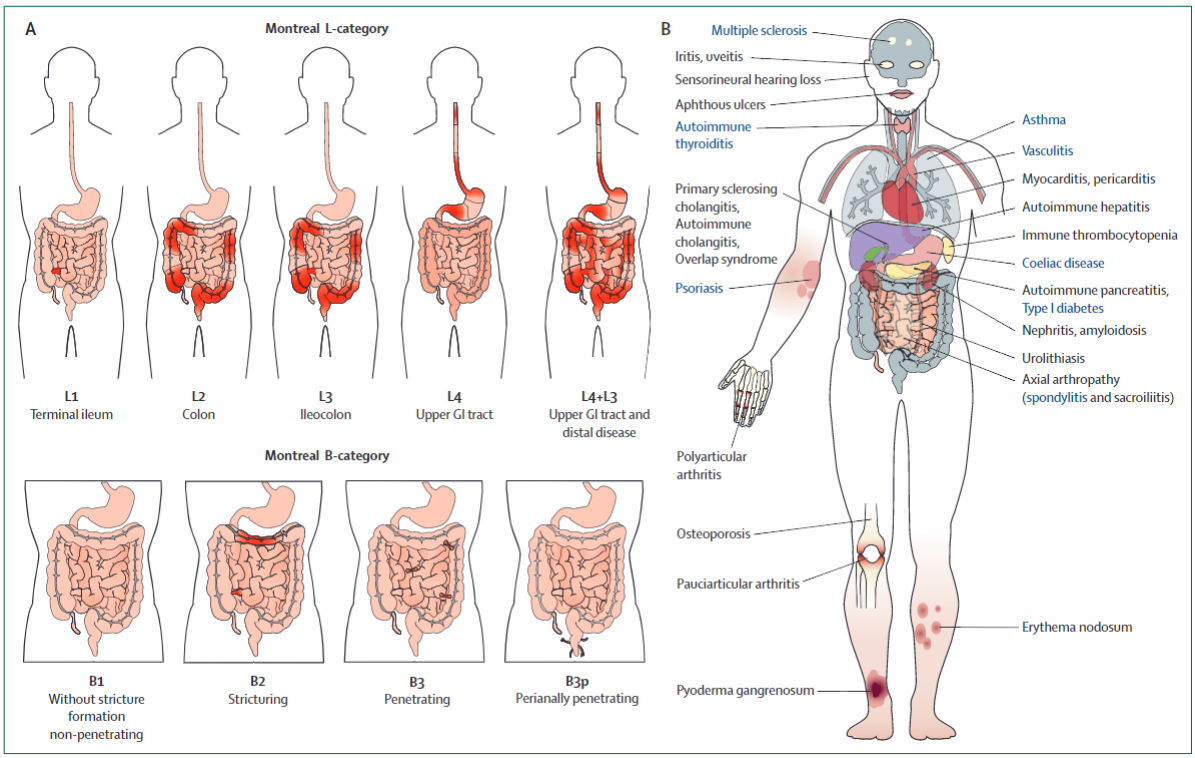

Figura 6 – Fenótipo da Doença de Crohn.

Fonte: Baumgart e Sandborn (2012).

CAPÍTULO

6

RESULTADOS

Neste capítulo estão descritos os resultados obtidos após a aplicação do modelo Weibull-ZICR no banco de dados de pacientes do Hospital das Clínicas de Ribeirão Preto diagnosticados com a doença de Crohn.

Na seção 6.1 apresenta-se a análise descritiva dos dados dos pacientes diagnosticados com a doença, a seção 6.2 demonstra a comparação entre os ajustes dos modelos exponencial, Weibull e log-normal aos dados e, na seção seguinte, tem-se os resultados da aplicação do modelo escolhido. A seção 6.3 foi subdividida em 4 seções (6.3.1 a 6.3.4), onde cada uma delas expõe a aplicação do modelo as quatro variáveis de estudo selecionadas de acordo com sua importância estatística.

6.1 Análise Descritiva

A base de dados contém informações de 295 pacientes do Hospital das Clínicas de Ribeirão Preto, sendo eles, em sua maioria (53%), do sexo masculino. Metade das pessoas tinha até 28 anos de idade no momento do diagnóstico da doença e a pessoa mais velha tinha 66 anos de idade ao receber o diagnóstico. Dos 295 pacientes, 73 realizaram a cirurgia de colectomia, 115 realizaram a cirurgia de enterectomia e 157 realizaram pelo menos uma das duas cirurgias, como descrito na tabela 2. Apenas 10,5% dos pacientes chegaram a realizar ambas as cirurgias.

Tabela 2 – Prevalência das Cirurgias de Colectomia e Enterectomia.

Cirurgia	Sim	Não
Colectomia	73 (25%)	222 (75 %)
Enterectomia	115 (39%)	180 (61%)
Pelo menos uma das duas cirurgias	157 (53%)	138 (47%)

A figura 7 apresenta as curvas de Kaplan-Meier para a sobrevivência e o risco acumulado do tempo até a cirurgia dos pacientes diagnosticados com a doença e, na tabela 3, tem-se os

dados de sobrevida para os tempos 0, 5, 10 e 15 anos. Observe que em ambos tabela e gráfico é possível notar a presença de pacientes com tempos iguais a zero, na figura 7 a curva inicia em torno de 0,80 e a tabela 3 mostra que para o tempo igual a zero, 56 pacientes já haviam realizado a cirurgia. Passados 15 anos desde o diagnóstico, a probabilidade de não realizar a cirurgia cai de 81% para 44,7% e 131 pacientes já realizaram pelo menos uma das duas cirurgias. Por fim, observa-se também a presença dos dados de longa duração, já que a curva de sobrevivência não atinge a probabilidade igual a zero ao final do tempo observado.

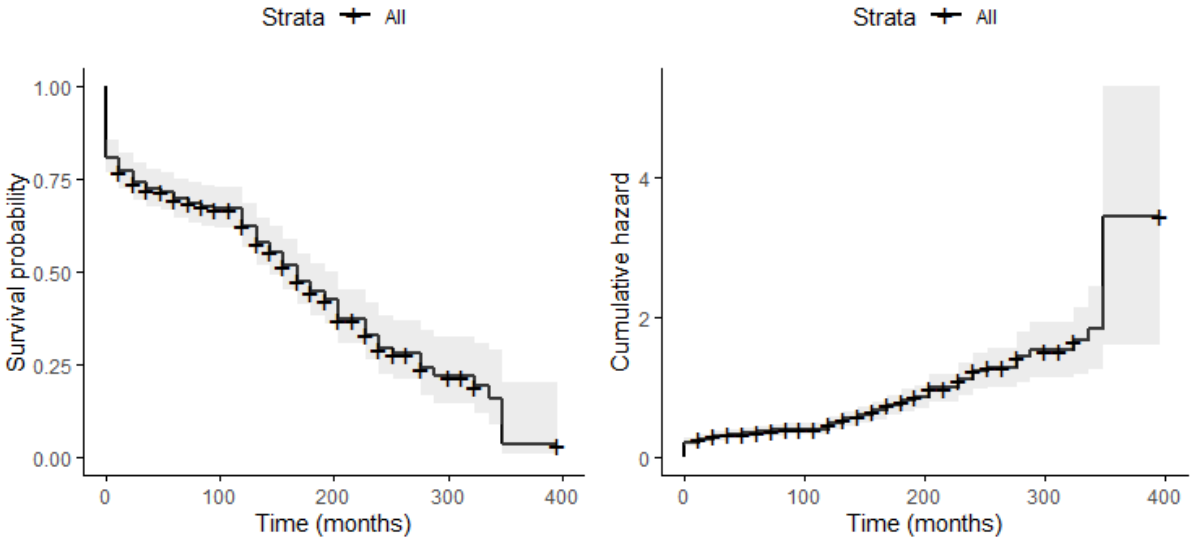

Figura 7 – Curva de Sobrevivência e Risco Acumulado dos Pacientes Diagnosticados com Crohn.

Fonte: Próprio Autor.

Tabela 3 – Estimativa de KM para a Sobrevida dos Tempos 0, 5, 10 e 15 Anos.

Anos	Num. Risco	Evento	Survival	95% IC
0	295	56	0,810	[0,767 - 0,856]
5	183	32	0,697	[0,646 - 0,752]
10	126	15	0,624	[0,567 - 0,686]
15	67	28	0,447	[0,382 - 0,523]

Inicialmente, o recorte da base de dados continha 12 variáveis preditoras, 2 variáveis desfecho indicando as cirurgias de colectomia e enterectomia e uma variável de tempo entre o diagnóstico e a cirurgia, quando houvesse. As variáveis desfecho foram unificadas em uma única variável marcando a realização de pelo menos uma das duas cirurgias (desfecho = 1) e a não realização de cirurgia (desfecho = 0). Já as variáveis preditores foram selecionadas de acordo com sua relevância para o estudo, conforme apresentado na tabela 4, utilizando o teste de log-rank para esta seleção. O tempo até a cirurgia será apresentado em meses.

Tabela 4 – Resultados do Teste de Log-Rank.

Variáveis	X^2	**p-valor**
Sexo	<0	0,9
Tabagismo	12,3	**<0,001**
Hist. Familiar	<0	0,9
Forma da Doença	41,2	**<0,001**
Doença Perianal	0,3	0,6
Manif. Extra Intestinal	0,5	0,5
Biológicos	1	0,3
Corticoides	4	**0,05**
Localização da Doença	22,8	**<0,001**
Cor/Raça	2,3	0,1
Idade Atual	0,7	0,7
Idade Diagnóstico	3,2	0,2

A tabela 4 expõe os resultados do teste, os valores de p-valor que estão em negrito representam as variáveis consideradas significativas pelo teste. Sendo assim, as variáveis tabagismo, corticoides, forma da doença e a sua localização foram selecionadas para compor o modelo.

Tabela 5 – Distribuição de Frequências Absolutas e Relativas das Variáveis de Acordo com a Cirurgia.

		Cirurgia			
		Não		**Sim**	
		N	%	N	%
Tabagismo	Não	112	81,2%	94	59,9%
	Sim	23	18,8%	63	40,1%
Localização	L1	46	33,3%	34	21,6%
	L2	29	21,0%	5	3,2%
	L3	62	44,9%	112	71,3%
	L4	1	0,8%	6	3,9%
Corticoides	Não	90	65,2%	79	50,3%
	Sim	48	34,8%	78	49,7%
Forma da Doença	B1	56	40,6%	5	3,2%
	B2	29	21,0%	81	51,6%
	B3	53	38,4%	71	45,2%

A tabela 5 ilustra as frequências relativas e absolutas por grupo das variáveis categóricas selecionadas para o estudo de acordo com a realização ou não da cirurgia. Observe que, dentro do grupo que realizou cirurgia a proporção dos fatores de risco são maiores, por exemplo, no grupo que não realizou a cirurgia apenas 18,8% dos pacientes afirmou ser fumante, já no grupo de pacientes que fez a cirurgia esse número cresce para 40%. Isto também ocorre para as demais variáveis selecionadas, as localizações L3 e L4 possuem maior prevalência no grupo que realizou a cirurgia, já o uso de corticoides é maior dentro do grupo que não realizou cirurgia e a forma da doença B1, que representa menos perigo, é mais frequente também no grupo que não realizou

cirurgia. Este comportamento é um indicativo de que as variáveis selecionadas para o modelo podem estar associadas à realização da cirurgia.

A figura 8 apresenta as curvas de Kaplan-Meier para o uso de corticoides, tabagismo, localização e forma da doença. Como visto na tabela 4, rejeitou-se a hipótese de igualdade entre as curvas das variáveis em questão. Os dados de sobrevida para os tempos 0, 5, 10 e 15 anos estão descritos na tabela 6. Note que, em todos os grupos apresentados, a sobrevida no tempo = 0 é menor que 100%, justificando o ajuste de um modelo inflacionado de zero aos dados.

Tabela 6 – Estimativa de KM para a Sobrevida das Variáveis Selecionadas nos Tempos 0, 5, 10 e 15 Anos.

		Tempo (anos)			
		0	5	10	15
Corticoides	Não	81,1%	71,9%	69,1%	51,1%
	Sim	81,0%	66,8%	54%	36,7%
Forma da Doença	B1	96,7%	96,7%	92,3%	92,3%
	B2	67,3%	55,2%	47%	29,5%
	B3	85,5%	69,5%	62,7%	42,4%
Localização	L1	77,5%	73,5%	69,4%	58,1%
	L2	97,1%	93,9%	93,9%	81,8%
	L3	79,9%	64,6%	55,5%	36,1%
	L4	71,4%	42,9%	21,4%	0%
Tabagismo	Não	85,0%	75,0%	69,1%	51,1%
	Sim	71,9%	57,7%	47,6%	31,4%

Os pacientes que não fizeram uso de corticoides possuem sobrevida maior a longo prazo, em 15 anos, a sobrevida das pessoas que não usaram corticoides é de 51,1%, no grupo de pessoas que fizeram o uso este número cai para 36,7%. Em relação a forma da doença, a curva de sobrevida do grupo B2 tem caimento maior em relação aos outros grupos, em 15 anos aproximadamente 70% dos pacientes com forma B2 já havia realizado cirurgia, em B3 este número cai para 57% e em B1 apenas 7% realizou a cirurgia após os primeiros 15 anos. Em relação a localização da doença, em 15 anos a chance do paciente ter realizado a cirurgia é baixa nos grupos L1 e L2, em contrapartida, todos os pacientes com localização L4 realizaram cirurgia após 15 anos de diagnóstico. Por fim, pacientes fumantes apresentam sobrevida de 31,4% contra 51,1% dos não fumantes em 15 anos.

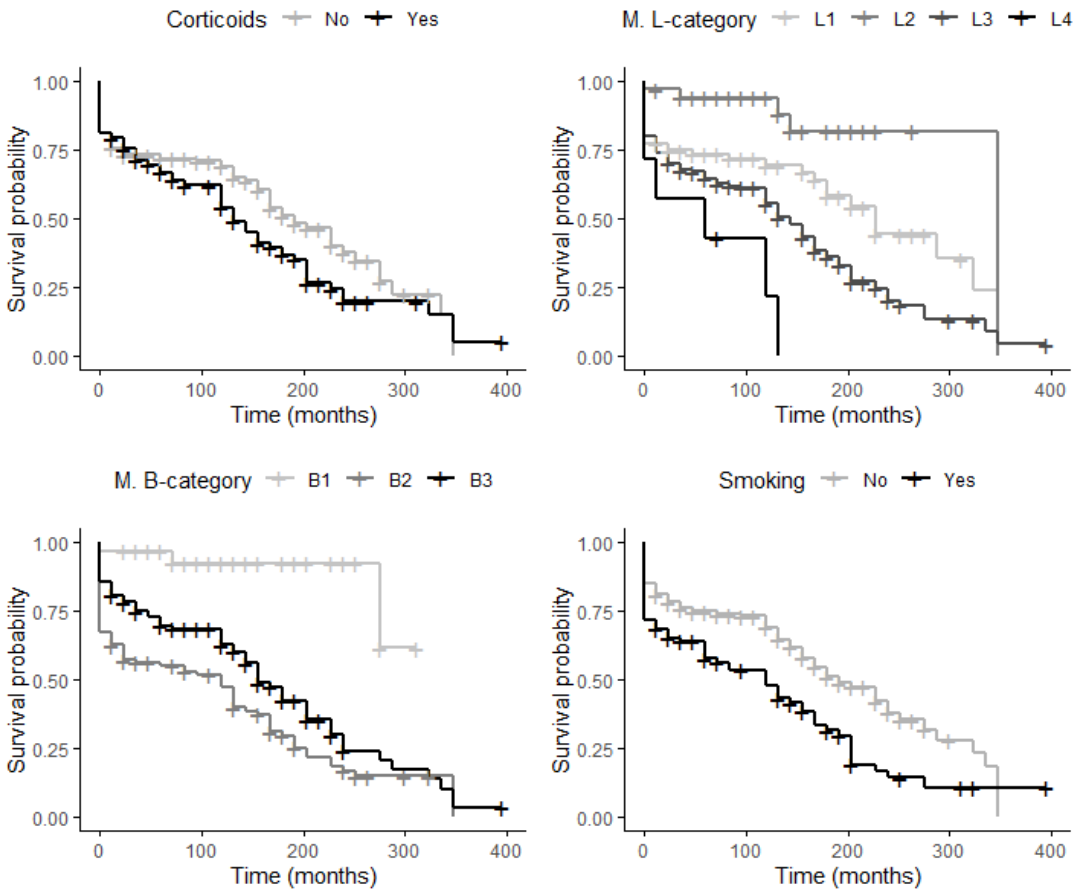

Figura 8 – Gráficos de Kaplan-Meier dos Pacientes Diagnosticados com Crohn para as Variáveis Selecionadas.

Fonte: Próprio Autor.

Voltando a tabela 5, algumas categorias da variável localização possuem pouca incidência, por exemplo apenas 5 pacientes que realizaram a cirurgia possuem localização da doença igual a L2 e somente 1 paciente com localização L4 não realizou a cirurgia. Estas incidências de pouco volume fazem com que o modelo não descrimine bem as curvas de sobrevivência ocasionando numa super ou subestimação das classes. Para contornar este problema, esta variável foi recategorizada em dois novos grupos sendo eles localização (L1+L2) e (L3+L4). Além disto, se observarmos as curvas de sobrevivência da forma da doença na figura 8, as sobrevidas das categorias B2 e B3 possuem caimento parecido o que também pode dificultar a modelagem em termos de discriminação entre as curvas. Por este motivo, estas categorias também foram reagrupadas resultando nos grupos B1 e (B2+B3). Importante ressaltar que, mesmo com a reclassificação, as variáveis localização e forma da doença se mantiveram significativas sob o teste de log-rank, apresentando p-valor $< 0{,}001$ em ambas situações.

6.2 Selecão do Modelo

Como visto anteriormente no capítulo 4, a distribuição Gama Generalizada tem como principal característica a possibilidade de se transformar em outras distribuições de acordo com os valores atribuídos aos seus parâmetros. Neste estudo, ajustou-se três casos particulares da distribuição Gama Generalizada aos dados do estudo: a distribuição exponencial, Weibull e a log-normal. O intuito desse ajuste foi descobrir qual das três distribuições melhor se adaptou aos dados e, para auxiliar na escolha do modelo final, alguns métodos de seleção foram utilizados. Esta seção demonstra a comparação dos ajustes e a justificativa do modelo escolhido.

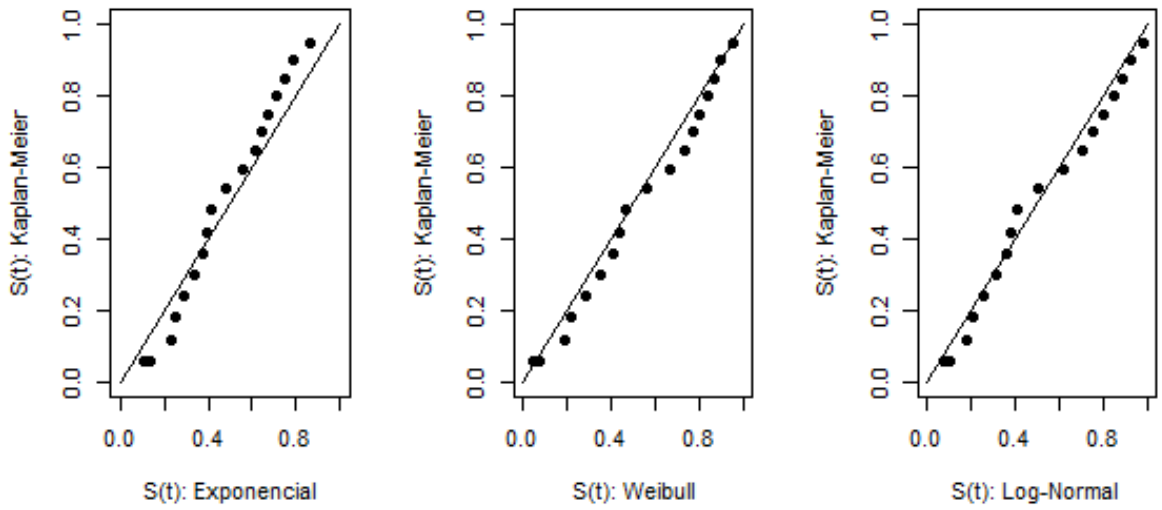

Figura 9 – Gráficos das Sobrevivências Estimadas por Kaplan-Meier Versus as Sobrevivências Estimadas pelos Modelos Exponencial, Weibull e Log-Normal.

Fonte: Próprio Autor.

Inicialmente comparou-se os gráficos de Kaplan-Meier aos tempos de sobrevivência estimados por cada um dos modelos aplicados, que são demonstrados pela figura 9. Neste gráfico o eixo x representa o ajuste pelo modelo e o eixo y representa a curva de Kaplan-Meier, quanto mais distantes os pontos forem da reta $y = x$, pior foi o ajuste do modelo. Observando os gráficos é possível ver que, dentre os modelos testados o exponencial é o que mais se distanciou da reta, o que significa que este modelo pode não ser tão adequado aos dados. Por outro lado, os ajustes Weibull e log-normal se aproximam mais da reta $y = x$, indicando que um destes modelo é, possivelmente, o mais adequado para os dados do estudo.

Para fortalecer as conclusões tiradas a partir da figura 9, plotou-se também os gráficos linearizados para os modelos exponencial, Weibull e log-normal (figura 10). Observe que, os gráficos para os modelos Weibull e log-normal não apresentam formato muito distante de uma reta. Porém, no primeiro gráfico, que representa o ajuste exponencial, os pontos se alinham de

forma um pouco mais disforme, expondo um certo desvio do que seria o formato de uma reta. Isto reafirma as conclusões tiradas anteriormente, apontando as distribuições Weibull e log-normal como igualmente satisfatórios para modelar o risco de cirurgia em pacientes diagnosticados com Crohn.

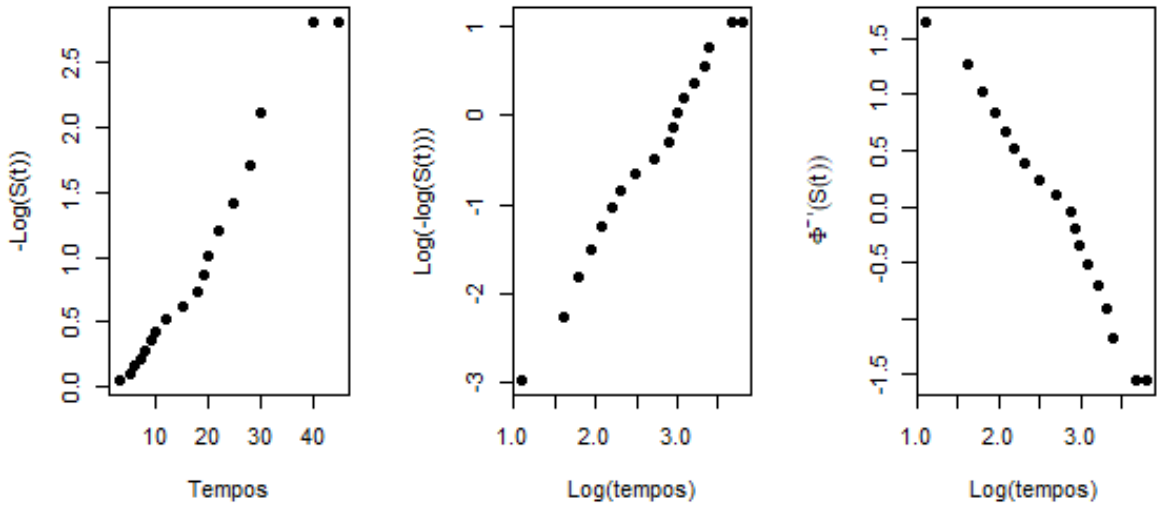

Figura 10 – Gráficos linearizados para os Modelos exponencial, Weibull e Log-Normal.

Fonte: Próprio Autor.

Além das técnicas gráficas, calculou-se o AIC para cada um dos modelos ajustados, a fim de compará-los. A tabela 7 expõe os valores encontrados para as distribuições exponencial, Weibull e log-normal, respectivamente. É possível ver que, os valores de AIC para as distribuições Weibull e log-normal são relativamente menores que o AIC do modelo exponencial, reafirmando as conclusões obtidas através das análises gráficas. Note ainda que, o ajuste log-normal foi o que apresentou menores valores de AIC, porém a diferença deste ajuste para o ajuste o Weibull é relativamente pequena.

Tabela 7 – Valores de AIC para os Ajustes do Modelo Exponencial, Weibull e Log-Normal.

Modelo	AIC
Exponencial	138,55
Weibull	136,27
Log-Normal	135,48

Por fim, comparou-se as curvas de sobrevivência estimadas por meio do ajuste dos modelos com a curva de sobrevivência estimada por Kaplan-Meier, esta comparação está ilustrada na figura 11. Os resultados obtidos até aqui já haviam comprovado que o modelo exponencial não seria uma escolha adequada para estudar os dados e o gráfico da figura 11 só reafirma essa

conclusão dado que é possível ver um distanciamento da curva de Kaplan-Meier e do ajuste exponencial. Observe que, novamente os ajustes Weibull e log-normal se mostraram satisfatórios porém, neste gráfico, o ajuste do modelo Weibull parece ser ligeiramente melhor que o ajuste da log-normal.

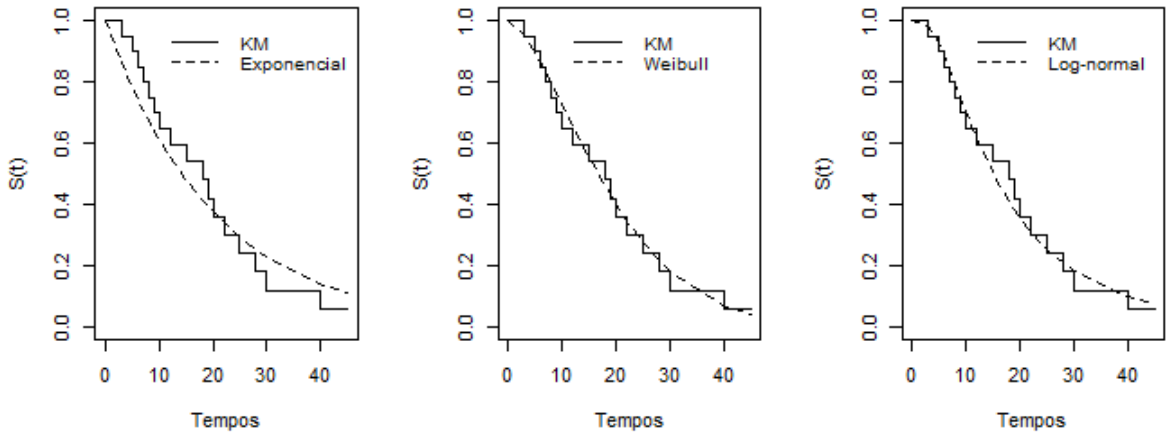

Figura 11 – Curvas de Sobrevivência Estimadas pelos Modelos Exponencial, Weibull e Log-Normal versus a Curva de Sobrevivência Estimada por Kaplan-Meier.

Fonte: Próprio Autor.

Os resultados até aqui mostraram que tanto a distribuição Weibull quanto a log-normal parecem ser adequados para modelar os dados dos pacientes diagnosticados com Crohn. Porém, antes de escolher uma das duas distribuições, plotou-se o gráfico do risco para entender qual o formato do mesmo. A figura 12 apresenta o gráfico do TTT onde o eixo x representa o tempo em anos e o eixo y a probabilidade dos pacientes de sofrer a cirurgia, ou seja, o risco. Observe que o risco cresce conforme os anos passam, reafirmando que ambos ajustes Weibull e log-normal serão boas escolhas para estudar os dados propostos.

As comparações realizadas mostraram que a distribuição exponencial não é adequado para modelar os dados do estudo, enquanto as figuras 9 e 10 mostraram que não há diferença entre os ajustes Weibull e log-normal e que ambos são bons candidatos ao modelo final. Além disso, a AIC apontou o modelo log-normal como melhor ajuste enquanto os gráficos da figura 11 apontaram a distribuição Weibull como melhor escolha. Por fim, o gráfico do TTT reafirmou as conclusões tiradas a partir dos primeiros gráficos, deixando a escolha do modelo final a encargo do pesquisador. Como não houve grandes diferenças entre os ajustes Weibull e log-normal, a escolha final foi a de seguir com o modelo Weibull apenas pela facilidade da interpretação de seus parâmetros. A próxima seção demonstra os resultados do ajuste do modelo final.

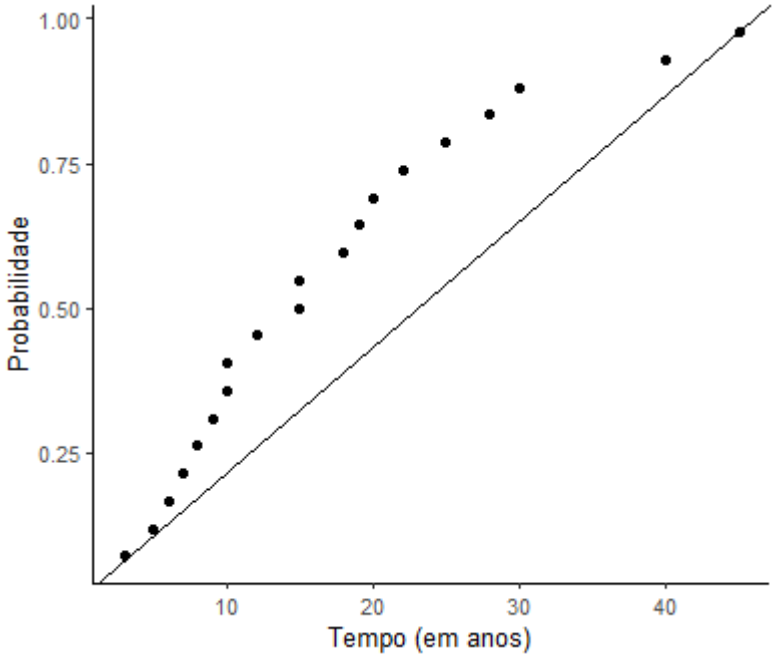

Figura 12 – Gráfico de TTT.

Fonte: Próprio Autor.

6.3 Ajuste do Modelo aos Dados

O modelo Weibull-ZICR foi ajustado separadamente para cada uma das variáveis tornando possível entender qual o efeito individual das co-variáveis no desfecho.

Na seção 4 foram definidas as funções links de acordo com os espaço paramétrico de cada parâmetro, definiu-se também o preditor linear utilizado para estimar cada um dos parâmetros do modelo. Partindo da equação 4.10, podemos definir cada parâmetro estimado pelo modelo Weibul-ZICR como:

$$\begin{cases} \mu_i & = & \beta_{10} + \beta_{11} x_{1i}, \\ \sigma_i & = & exp(\beta_{20} + \beta_{21} x_{1i}), \\ \lambda_i & = & 1, \\ \left(\frac{p_{0i}}{1-p_{0i}-p_{1i}}, \frac{p_{1i}}{1-p_{0i}-p_{1i}}\right) & = & \left(\frac{e^{\beta_{30}+\beta_{31}x_{1i}}}{1+e^{\beta_{30}+\beta_{31}x_{1i}}+e^{\beta_{40}+\beta_{41}x_{1i}}}, \frac{e^{\beta_{40}+\beta_{41}x_{1i}}}{1+e^{\beta_{30}+\beta_{31}x_{1i}}+e^{\beta_{40}+\beta_{41}x_{1i}}}\right), \end{cases} \tag{6.1}$$

onde $\lambda_i = 1$ para que o modelo GG-ZICR possa ser reduzido a distribuição Weibull.

Nas próximas sessões estão descritos os resultados do ajuste do modelo para cada variável de interesse, bem como as estimativas pontuais de máxima verossimilhança dos parâmetros β_{ij}.

6.3.1 W-ZICR para Variável Corticoide

62% dos os pacientes que fizeram uso de corticoides realizaram cirurgia, em contrapartida apenas 47% dos pacientes que não fez uso do medicamento sofreram o desfecho, como demonstrado na tabela 8. Pela figura 13 nota-se que a curva de sobrevivência dos pacientes que fizeram uso do medicamento cai primeiro que a daqueles pacientes que não fizeram uso e se mantém afastada da curva do outro grupo durante todo o tempo, se aproximando apenas nos últimos meses observados.

Tabela 8 – Prevalência da Cirurgia de Acordo com a Variável Corticoide.

Corticoide	Cirurgias	Não-Cirurgias
Não ($x = 0$)	79 (46,7%)	90 (53,3%)
Sim ($x = 1$)	78 (61,9%)	48 (38,1%)
Total	157 (53,2%)	138 (46,8%)

As estimativas pontuais de máxima verossimilhança obtidas para o modelo W-ZICR ajustado a variável corticoide estão descritas na tabela 9. Os gráficos apresentados na figura 13 ilustram as curvas de Kaplan-Meier da variável corticoide sob o ajuste do modelo W-ZICR, bem como a curva do risco relativo para esta variável. Já a tabela 10 expõe os resultados do modelo ajustado a variável de interesse.

Tabela 9 – Estimativas de Máxima Verossimilhança dos Parâmetros do Modelo W-ZICR Ajustado a Variável Corticoide e Estimativas do 95% I.C.

Parâmetro	EMV	95% I.C.
β_{10}	5,706	{5,468; 5,943}
β_{11}	-0,357	{-0,655; -0,058}
β_{20}	-0,337	{-0,573; -0,100}
β_{21}	-0,066	{-0,382; 0,251}
β_{30}	-1,454	{-1,839; -1,069}
β_{31}	0,007	{-0,581; 0,595}
β_{40}	-10,854	{-155,36; 133,66}
β_{41}	-1,828	{-597,29; 593,63}

A partir dos β_s estimados é possível fazer o cálculo dos parâmetros de interesse, como demonstrado em 6.1. Na tabela 10 temos os resultados das estimativas do modelo para os parâmetros separadas por grupos. A estimativa do tempo médio de segmento para o grupo que não fez uso de corticoides e para o grupo que fez foi de 5,7 e 5,3 anos, respectivamente. Note que o tempo médio de sobrevida foi menor no grupo de pacientes que fez o uso de corticoides, indicando que os pacientes deste grupo provavelmente irão realizar a cirurgia antes que os pacientes do outro grupo. Além disso, o tempo estimado que marca a sobrevida 0,5 dos pacientes que usaram e não usaram corticoides durante o tratamento é de 15 e 11 anos, respectivamente.

Como visto anteriormente (3.8), os parâmetros p_0 e p_1 representam, respectivamente, a proporção de indivíduos com tempos iguais a zero e a proporção de indivíduos curados. Os resultados obtidos para $\hat{p}_0$ foram de 0,1893 para o grupo que não fez uso de corticoide e 0,1904 para o grupo que fez uso. Se observarmos bem, as sobrevidas estimadas na tabela 6 para o t = 0 coincidem exatamente com os valores de $1-\hat{p}_0{}^0$ e de $1-\hat{p}_0{}^1$, indicando que o modelo se ajustou bem aos tempos de sobrevivência inflacionados a zero. Porém, se compararmos os valores de indivíduos que não sofreram a cirurgia apresentados na tabela 8 com a proporção de indivíduos curados estimada pelo modelo, podemos ver que o modelo falha ao estimar esta segunda proporção.

Tabela 10 – Resultados do Modelo W-ZICR Ajustado a Variável Corticoide.

Parâmetro	$x=0$	$x=1$
$\hat{\mu}$	5,7055	5,3489
$\hat{\sigma}$	0,7142	0,6687
$\hat{p}_0$	0,1893	0,1904
$\hat{p}_1$	0,0001	0,00001

A proporção de indivíduos suscetíveis a falha se dá através da fórmula $1-p_0-p_1$, resultando em 81,1% de pacientes suscetíveis a cirurgia no grupo que não fez uso de corticoide contra 80,1% no grupo que fez o uso.

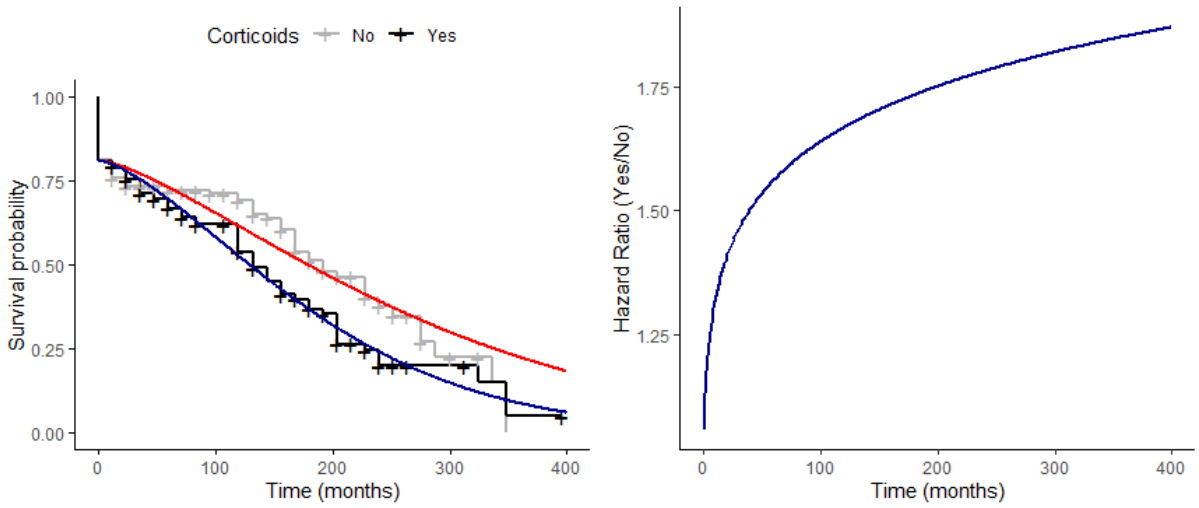

Figura 13 – Curvas de KM sob as Estimativas do Modelo W-ZICR e Gráfico do Risco Relativo para a Variável Corticoide.

Fonte: Próprio Autor.

As curvas de sobrevivência estimadas são bastante próximas às curvas verdadeiras, note também que as curvas estimadas respeitam a inflação de zero de ambos os grupos. O gráfico do risco relativo demonstra o risco que os pacientes que fizeram uso de corticoide têm de sofrer o desfecho em relação àqueles que não fizeram uso. Nota-se que o risco relativo começa pequeno

mas cresce consideravelmente nos primeiros anos de diagnóstico, com o passar do tempo o risco continua crescendo mas vai se estabilizando por volta de 2. Isto mostra que, nos primeiros 100 meses de tratamento, pacientes que fazem uso de corticoide podem ter até 50% mais chances de realizar a cirurgia quando comparado àqueles pacientes que não fizeram uso, em 200 meses este número cresce para 75% e, ao final do tempo de observação, o risco destes pacientes pode chegar ao dobro em comparação ao outro grupo.

6.3.2 W-ZICR para Variável Forma da Doença

Como explicado na seção 6.1, a variável forma da doença foi reagrupada, restante duas categorias: B1 e B2+B3. Dentre os pacientes do primeiro grupo, apenas 5 (8,2%) realizaram a cirurgia, em contrapartida, mais da metade (64,9%) dos pacientes com forma da doença B2 ou B3 realizou alguma intervenção cirúrgica, como demonstrado na tabela 11. Pela figura 14 é fácil perceber que a curva de sobrevivência dos pacientes do segundo grupo cai bem antes que a do primeiro grupo, além disso ela se mantém muito afastada da curva B1 durante todo o tempo de observação. Isto indica que o grupo de pacientes com forma da doença B2 ou B3 possui um risco muito maior de cirurgia que os pacientes com forma B1.

Tabela 11 – Prevalência da Cirurgia de Acordo com a Variável Forma da Doença.

Forma da Doença	Cirurgias	Não-Cirurgias
B1 ($x = 0$)	5 (8,2%)	56 (91,8%)
B2+B3 ($x = 1$)	152 (64,9%)	82 (35,1%)
Total	157 (53,2%)	138 (46,8%)

As estimativas pontuais de máxima verossimilhança obtidas para o modelo W-ZICR ajustado à variável forma da doença estão descritas na tabela 12. As curvas de Kaplan-Meier da variável forma da doença sob o ajuste do modelo W-ZICR e a curva do risco relativo para esta variável podem ser vistas na figura 14. Já os resultados do modelo ajustado à variável de interesse estão descritos na tabela 13.

Tabela 12 – Estimativas de Máxima Verossimilhança dos Parâmetros do Modelo W-ZICR Ajustado a Variável Forma da Doença e Estimativas do 95% I.C.

Parâmetro	EMV	95% I.C.
β_{10}	6,431	{5,162; 7,700}
β_{11}	-1,047	{-2,324; 0,231}
β_{20}	-0,689	{-1,537; 0,159}
β_{21}	0,352	{-0,512; 1,215}
β_{30}	-3,370	{-5,048; -1,692}
β_{31}	2,166	{0,461; 3,872}
β_{40}	-4,145	{-62,456; 54,165}
β_{41}	-8,720	{-264,461; 247,021}

Como demonstrado em 6.1, é possível fazer o cálculo dos parâmetros de interesse a partir dos β_s estimados. A tabela 13 apresenta os resultados das estimativas do modelo para os parâmetros separados por grupos. O tempo médio de sobrevida estimado para a forma da doença B1 foi de 6,4 anos, por outro lado o grupo de pacientes com forma da doença B2+B3 teve uma estimativa de aproximadamente 5,4 anos. Note que o tempo médio até a cirurgia do segundo grupo é 1 ano menor que o do primeiro, o que vai de encontro com as curvas de sobrevivência apresentadas em 14, indicando que o risco de cirurgia é bem maior no segundo grupo quando comparado ao primeiro. Além disso, o tempo estimado que marca a sobrevida 0,5 dos pacientes B1 e B2+B3 é de 33 e 10 anos, respectivamente, ou seja, o primeiro grupo leva mais que o triplo do tempo do segundo para atingir a marca de 50% de probabilidade de sobrevivência.

Tabela 13 – Resultados do Modelo W-ZICR Ajustado a Variável Forma da Doença.

Parâmetro	$x = 0$	$x = 1$
$\hat{\mu}$	6,4305	5,3840
$\hat{\sigma}$	0,5020	0,7136
$\hat{p}_0$	0,0327	0,2307
$\hat{p}_1$	0,0150	0,00001

Os valores estimados de $\hat{p}_0$ e $\hat{p}_0$ foram de 0,0327 e 0,0150 para os pacientes com forma da doença B1 e de 0,2307 e 0,00001 para aqueles com forma B2 ou B3. Se subtrairmos $1-\hat{p}_0{}^0$ e $1-\hat{p}_0{}^1$ teremos valores que coincidem com as sobrevidas estimadas na tabela 6 para o t = 0, indicando que o modelo se ajustou bem aos tempos de sobrevivência inflacionados a zero. Porém, não podemos dizer o mesmo da proporção de indivíduos curados estimada pelo modelo, se compararmos os valores de $\hat{p}_0$ com os valores de não-cirurgia apresentados na tabela 11, é fácil ver que o modelo falha ao estimar esta segunda proporção. A proporção de indivíduos suscetíveis a cirurgia $(1-p_0-p_1)$ após o instante t = 0 foi de 95,2% para o grupo de pacientes com forma da doença B1 e de 76,9% para o grupo com forma da doença B2+B3.

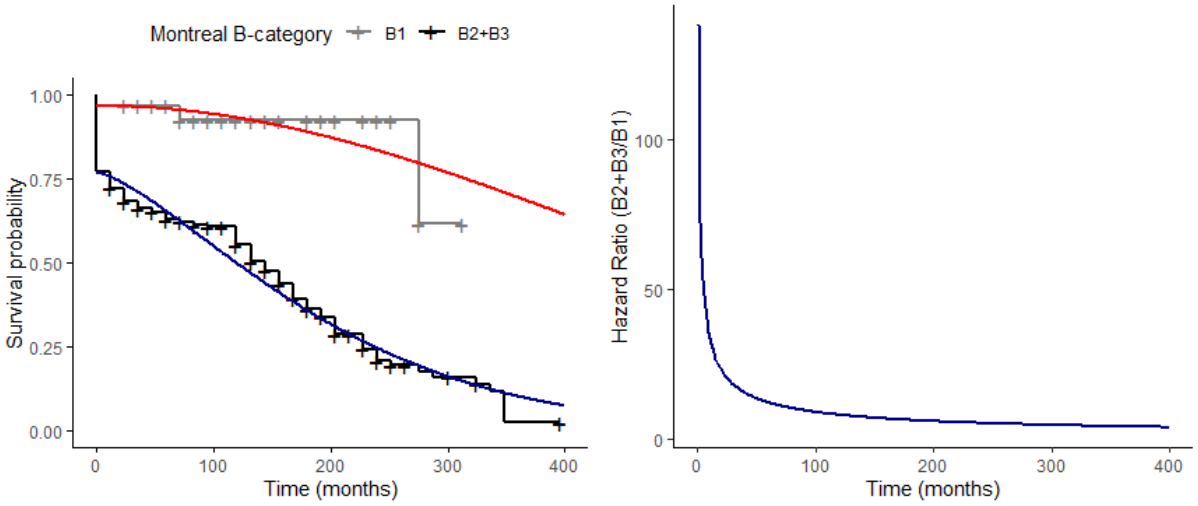

Figura 14 – Curvas de KM sob as Estimativas do Modelo W-ZICR e Gráfico do Risco Relativo para a Variável forma da doença.

Fonte: Próprio Autor.

Note que, as curvas de sobrevivência da figura 14 estimadas pelo modelo se ajustam bem às curvas de Kaplan-Meier, note também que as curvas estimadas respeitam a inflação de zero de ambos os grupos. O gráfico do risco relativo demonstra o risco que os pacientes com forma da doença B2+B3 têm de realizar a cirurgia quando comparado ao grupo B1. Observe que a curva do risco da forma da doença se comporta de forma diferente a do corticoide, enquanto o segundo apresenta uma curva de risco crescente, neste caso o risco relativo vai diminuindo ao passar dos meses. Observe ainda que, nos primeiros meses de diagnóstico o risco da forma da doença B2+B3 em relação a B1 é infinitamente maior e, após aproximadamente 50 meses, este risco se estabiliza próximo a 1 e se mantém assim até o final do tempo de observação. O gráfico do risco relativo termina de concluir o que as curvas de Kaplan-Meier e os tempos médios de sobrevida já tinha sugerido: o risco de realizar a cirurgia é muito maior dentre os pacientes que foram diagnosticados com a forma da doença B2 ou B3, quando comparados aos de forma B1, sendo este risco muito maior nos primeiros anos de diagnóstico.

6.3.3 W-ZICR para Variável Localização

A variável localização foi recategorizada em dois novos grupos: L1+L2 e L3+L4, como mostrado na seção 6.1. Dos pacientes com localização L1 ou L3, apenas 34,2% deles realizou uma das duas cirurgias, por outro lado, 65,2% dos pacientes com localização L3 ou L4 realizou cirurgia, como demonstrado na tabela 14. Através da figura 15 podemos perceber que a curva de sobrevivência dos pacientes com localização L3+L4 decai mais rápido que a do outro grupo, além disso as curvas se mantêm descoladas durante todo o tempo de observação, ficando próximas apenas nos primeiros meses. Isto demonstra que os pacientes diagnosticados com a localização L3 e L4 possuem um risco maior de sofrer a cirurgia quando comparados aos pacientes do

primeiro grupo.

Tabela 14 – Prevalência da Cirurgia de Acordo com a Variável Localização.

Localização	Cirurgias	Não-Cirurgias
L1+L2 ($x = 0$)	39 (34,2%)	75 (65,8%)
L3+L4 ($x = 1$)	118 (65,2%)	63 (34,8%)
Total	157 (53,2%)	138 (46,8%)

A tabela 15 descreve as estimativas pontuais de máxima verossimilhança obtidas para o modelo W-ZICR ajustado a variável localização. As curvas de Kaplan-Meier sob o ajuste do modelo W-ZICR e a curva do risco relativo para esta variável podem ser vistas na figura 15. Já os resultados do modelo ajustado a variável localização estão descritos na tabela 16.

Tabela 15 – Estimativas de Máxima Verossimilhança dos Parâmetros do Modelo W-ZICR Ajustado a Variável Localização e Estimativas do 95% I.C.

Parâmetro	EMV	95% I.C.
β_{10}	5,784	{5,548; 6,020}
β_{11}	-0,460	{-0,749; -0,171}
β_{20}	-0,800	{-1,123; -0,476}
β_{21}	0,513	{0,143; 0,883}
β_{30}	-1,609	{-2,102; -1,117}
β_{31}	0,250	{-0,360; 0,861}
β_{40}	-8,634	{-73,469; 56,201}
β_{41}	-2,388	{-153,725; 148,949}

Os parâmetros de interesse do modelo são calculados a partir dos β_s estimados, como demonstrado em 6.1. Os resultados do cálculo dos parâmetros separados por grupos podem ser vistos na tabela 16. O tempo médio de sobrevida estimado dos pacientes com localização L1+L2 e L3+L4 foi de 5,7 e 5,3, respectivamente. Assim como aconteceu para as variáveis corticoide e forma da doença, o tempo médio de sobrevida do grupo de risco (L3+L4) foi maior, indicando, novamente, que o risco de cirurgia deste grupo é maior que o do primeiro. Além disso, o tempo estimado que marca a sobrevida 0,5 dos pacientes L1+L2 e L3+L4 é de 20 e 10 anos, respectivamente, ou seja, o primeiro grupo leva aproximadamente o dobro do tempo do segundo para atingir a marca de 50% de probabilidade de sobrevivência.

Tabela 16 – Resultados do Modelo W-ZICR Ajustado a Variável Localização.

Parâmetro	$x = 0$	$x = 1$
$\hat{\mu}$	5,7840	5,3243
$\hat{\sigma}$	0,4494	0,7506
$\hat{p}_0$	0,1666	0,2044
$\hat{p}_1$	0,0001	0,0001

A proporção estimada de indivíduos com tempos iguais a zero ($\hat{p}_0$) foi de 0,1666 para L1+L2 e de 0,2044 para L3+L4, já a proporção de indivíduos curados ($\hat{p}_1$) foi a mesma para ambos os grupos (0,0001). Note que, se subtrairmos $1 - \hat{p}_0{}^0$ e $1 - \hat{p}_0{}^1$ teremos valores que são, aproximadamente, as sobrevidas estimadas na tabela 6 para t = 0, indicando que o modelo se ajustou bem aos tempos de sobrevivência inflacionados a zero. Por outro lado, se compararmos os valores de $\hat{p}_0$ com os valores de que não sofreram o desfecho apresentados na tabela 14, podemos ver que o modelo não estima bem a proporção de indivíduos curados. O percentual de indivíduos suscetíveis a cirurgia ($1 - p_0 - p_1$) após o instante t = 0 foi de 81,1% para o grupo de pacientes com localização L1 ou L2 e de 83,3% para o segundo grupo.

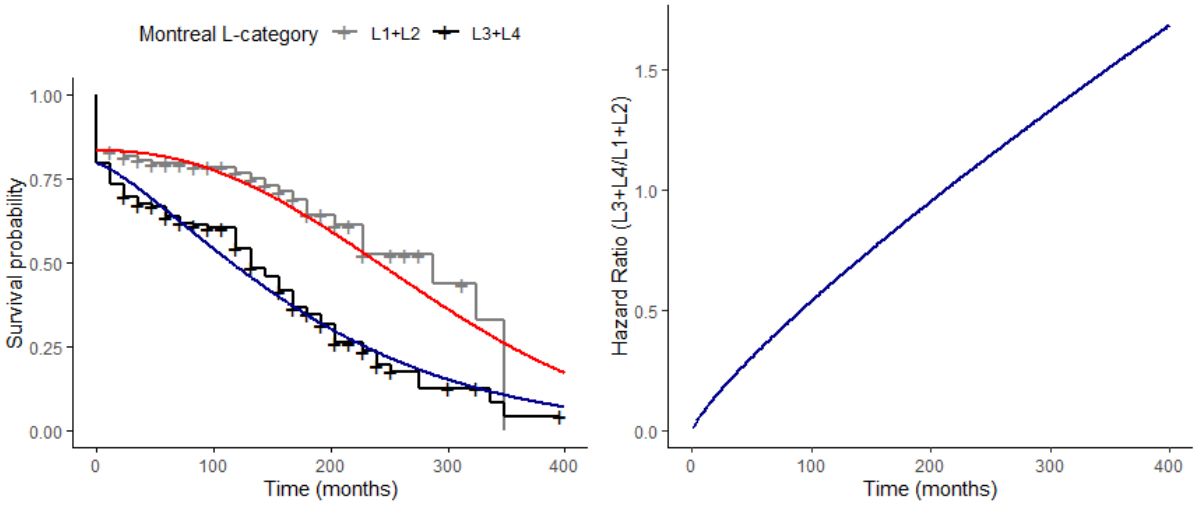

Figura 15 – Curvas de KM sob as Estimativas do Modelo W-ZICR e Gráfico do Risco Relativo para a Variável Localização.

Fonte: Próprio Autor.

Observe que, as curvas de sobrevivência estimadas pelo modelo ilustradas na figura 15 se ajustam bem às curvas de Kaplan-Meier, além disso as curvas estimadas respeitam a inflação de zero de ambos os grupos. O gráfico do risco relativo apresenta o risco que os pacientes com localização da doença L3 ou L4 têm de realizar a cirurgia quando comparado ao grupo L1+L2. Diferente da variável forma da doença, onde o risco era maior nos primeiros anos de diagnóstico, para a localização este cenário se inverte. Note que a curva do risco da localização se comporta de forma crescente durante todo o período de observação, sendo mais baixa nos primeiros meses, o que significa que o risco de cirurgia do segundo grupo em relação ao primeiro vai crescendo ao passar dos meses até atingir um valor próximo a 1,5 ao final do período de observação.

6.3.4 W-ZICR para Variável Tabagismo

Dentre os pacientes que se declararam fumantes, aproximadamente 70% realizou cirurgia, em contrapartida, apenas 46% dos pacientes que não não fumam chegou a realizar cirurgia, como demonstrado na tabela 17. Pela figura 16 é fácil perceber que a curva de sobrevivência

dos pacientes que fumam cai mais rápido que a dos pacientes não-fumantes, além disso as duas curvas se mantêm distantes durante todo o tempo de observação.

Tabela 17 – Prevalência da Cirurgia de Acordo com a Variável Tabagismo.

Tabagismo	Cirurgias	Não-Cirurgias
Não ($x = 0$)	94 (45,6%)	112 (54,4%)
Sim ($x = 1$)	63 (70,8%)	26 (29,2%)
Total	157 (53,2%)	138 (46,8%)

As estimativas pontuais de máxima verossimilhança obtidas para o modelo W-ZICR ajustado a variável tabagismo estão descritas na tabela 18. Os gráficos apresentados na figura 16 ilustram as curvas de Kaplan-Meier da variável tabagismo sob o ajuste do modelo W-ZICR, bem como a curva do risco relativo para esta variável. Já a tabela 19 expõe os resultados do modelo ajustado a variável de interesse.

Tabela 18 – Estimativas de Máxima Verossimilhança dos Parâmetros do Modelo W-ZICR Ajustado a Variável Tabagismo e Estimativas do 95% I.C.

Par	EMV	95% I.C.
β_{10}	5,636	{5,443; 5,829}
β_{11}	-0,323	{-0,616; -0,030}
β_{20}	-0,367	{-0,567; -0,167}
β_{21}	-0,005	{-0,331; 0,321}
β_{30}	-1,730	{-2,112; -1,349}
β_{31}	0,790	{0,191; 1,390}
β_{40}	-9,413	{-66,880; 48,054}
β_{41}	-4,352	-

Sabe-se que à partir dos β_s estimados é possível fazer o cálculo dos parâmetros de interesse, como demonstrado em 6.1. A tabela 19 apresenta os resultados das estimativas do modelo para os parâmetros separadas pelo grupo de pacientes fumantes e não-fumantes. A estimativa do tempo médio de segmento para o grupo que não fez uso de corticoides e para o grupo que fez foi de 5,6 e 5,3 anos, respectivamente. Note que o tempo médio de sobrevida foi menor no grupo de pacientes fumantes, indicando que os pacientes deste grupo provavelmente terão risco maior de realizar a cirurgia que os pacientes não-fumantes. Além disso, o tempo estimado que marca a sobrevida 0,5 dos pacientes fumantes e não-fumantes durante o tratamento é de 15 e 8 anos, respectivamente.

Tabela 19 – Resultados do Modelo W-ZICR Ajustado a Variável Tabagismo.

Parâmetro	$x = 0$	$x = 1$
$\hat{\mu}$	5,6358	5,3128
$\hat{\sigma}$	0,6927	0,6894
$\hat{p}_0$	0,15	0,2808
$\hat{p}_1$	0,0001	0,000001

Como visto anteriormente (3.8), os parâmetros p_0 e p_1 representam, respectivamente, a proporção de indivíduos com tempos iguais a zero e a proporção de indivíduos curados. Os resultados obtidos para $\hat{p}_0$ foram de 0,15 para o grupo de pacientes não-fumantes e de 0,2808 para o grupo fumante. Se observarmos bem, as sobrevidas estimadas na tabela 6 para t = 0 coincidem exatamente com os valores de $1 - \hat{p}_0{}^0$ e de $1 - \hat{p}_0{}^1$, indicando que o modelo se ajustou bem aos tempos de sobrevivência inflacionados a zero. Porém, se compararmos os valores de não cirurgia apresentados na tabela 17 com a proporção de indivíduos curados estimada pelo modelo, podemos ver que o modelo falha ao estimar esta segunda proporção.

A proporção de indivíduos suscetíveis a falha se dá através da fórmula $1 - p_0 - p_1$, resultando em 84,9% de pacientes suscetíveis a cirurgia no grupo que não fez uso de corticoide contra 71,9% no grupo que fez o uso.

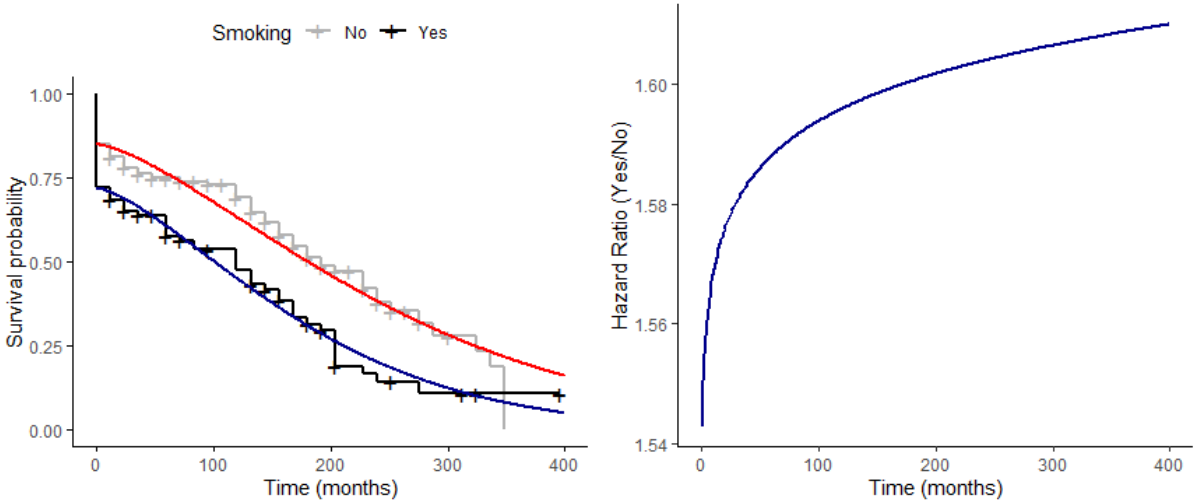

Figura 16 – Curvas de KM sob as Estimativas do Modelo W-ZICR e Gráfico do Risco Relativo para a Variável Tabagismo.

Fonte: Próprio Autor.

As curvas de sobrevivência estimadas são bastante próximas às curvas verdadeiras, note também que as curvas estimadas respeitam a inflação de zero de ambos os grupos. O gráfico do risco relativo demonstra o risco que os pacientes que fumantes têm de sofrer o desfecho em relação àqueles que não fumam. Nota-se que o risco relativo já começa por volta de 1.50 e cresce rapidamente nos primeiros 100 meses de diagnóstico, após esse período o risco continua crescendo mas vai se estabilizando por volta de 1,70. Isto mostra que, nos primeiros 100 meses de tratamento, pacientes fumantes podem ter entre 50% e 60% mais chances de realizar a cirurgia quando comparado aos pacientes que não fumam, à partir de 100 meses este número se estabiliza até o final do período de observação e se mantém por volta de 70%.

CAPÍTULO

7

CONCLUSÃO

Este trabalho propôs um modelo de sobrevivência que se ajustasse aos dados inflacionados de zero e de longa duração, O modelo Weibull inflacionado de zero de longa duração (W-ZICR).

O modelo Weibull-ZICR foi escolhido por ser um modelo prático e de fácil interpretação dos resultados. Na prática, o modelo garante que os pacientes que não realizaram a cirurgia e aqueles que a realizam no momento do diagnóstico sejam considerados na modelagem, já que os modelos de sobrevivência usuais não se ajustam a este tipo de dado. Além disso, o modelo ajuda a descrever e estimar os riscos associados à doença, a fração de cura dos pacientes, bem como a proporção correta dos pacientes suscetíveis à cirurgia.

No Hospital das Clínicas de Ribeirão Preto (HCRP), cerca de 295 pacientes fizeram acompanhamento da doença pelos últimos 35 anos, dentre estes pacientes mais da metade chegou a realizar uma cirurgia de colectomia ou enterectomia. Ao todo, 157 pacientes (53%) realizaram pelo menos uma das duas cirurgias e, apenas 10.5% dos pacientes chegou a realizar ambas. Destes 157 pacientes, cerca de 36% (56) realizou a cirurgia no momento em que foi diagnosticado ou assim que passou a ser observado no estudo, ou seja, sofreu o desfecho no instante t = 0.

As 4 variáveis escolhidas para compor o modelo foram o corticoide, forma da doença, localização e tabagismo. O ajuste do modelo W-ZICR aos dados nos levou a conclusão de que o grupo de pacientes que faz tratamento com corticoide tem até 80% a mais de chances de realizar a cirurgia que o grupo que não fez tratamento com o medicamento. Em relação a forma da doença, nos primeiros meses de diagnóstico o risco do grupo B2+B3 de realizar a cirurgia pode chegar a ser 100 vezes maior que o grupo B1, com o passar dos meses este risco cai e se estabiliza, se tornando praticamente o mesmo para ambos os grupos ao final do período de observação.

Além disso, nos primeiros meses de diagnóstico, o risco dos pacientes com localização

L1+L2 é maior que o do grupo L3+L4, porém, este risco vai diminuindo e chega a 1 em aproximadamente 200 meses, a partir daí o risco passa a ser maior para o segundo grupo até atingir um valor próximo a 1.5 no final do período de observação. Por fim, o risco dos pacientes fumantes de realizar a cirurgia é sempre maior que o daqueles que não fumam, nos primeiros meses de diagnóstico a chance de sofrer cirurgia dos pacientes fumantes é cerca de 56% maior que dos que não fumam e, com os passar dos meses, esse risco aumenta chegando a aproximadamente 70% a mais de chance para o grupo de pacientes que fumam.

Desta forma, concluímos que os pacientes pertencentes aos grupos de risco (faz uso de corticoide, fuma, tem forma da doença B2 ou B3 e localização L3 ou L4) possuem sobrevida menor e risco maior de realizar cirurgia quando comparados aos demais pacientes. Em adicional, o modelo não estimou com excelência a proporção de indivíduos curados, porém as curvas estimadas se ajustaram bem às estimativas de Kaplan-Meier e respeitaram a inflação de zero dos dados. Por fim, conseguimos demonstrar que o modelo W-ZICR pode ser uma excelente ferramenta para estudar pacientes diagnosticados com a doença de Crohn.

Ao final, foi possível encontrar um modelo que pudesse descrever os fatores de risco associados à doença de Crohn, entretanto, estudos adicionais podem ser feitos para corroborar com os resultados encontrados. Por exemplo, outras distribuições podem ser testadas em prol de capturar melhor o percentual de pacientes curados, a distribuição Gama Generalizada pode ser uma excelente candidata dado que a distribuição Weibull é uma particularidade da Gama. É necessário também testar um modelo completo, que englobe todas as 4 co-variáveis para testar o efeito que uma variável pode ter sobre a outra. Além disso, novas variáveis podem ser coletadas e adicionadas ao modelo para tentar estimar ainda melhor a sobrevida dos pacientes.

REFERÊNCIAS

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 36.

BARLOW, R. E.; CAMPO, R. Total time on test processes and applications to failure data analysis. In: **Reliability and fault tree analysis**. [S.l.: s.n.], 1975. Citado na página 36.

BAUMGART, D. C.; SANDBORN, W. J. Crohn's disease. **The Lancet**, Elsevier, v. 380, n. 9853, p. 1590–1605, 2012. Citado nas páginas 21 e 47.

BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. **Journal of the American Statistical Association**, Taylor & Francis, v. 47, n. 259, p. 501–515, 1952. Citado nas páginas 22, 23, 31 e 32.

BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, v. 11, n. 1, p. 15–53, 1949. Citado na página 22.

BRAEKERS, R.; GROUWELS, Y. A semi-parametric cox's regression model for zero-inflated left-censored time to event data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 45, n. 7, p. 1969–1988, 2016. Citado na página 22.

CALSAVARA, V. F.; RODRIGUES, A. S.; TOMAZELLA, V. L. D.; CASTRO, M. de. Frailty models power variance function with cure fraction and latent risk factors negative binomial. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 46, n. 19, p. 9763–9776, 2017. Citado na página 23.

CANCHO, V. G.; BARRIGA, G. D.; CORDEIRO, G. M.; ORTEGA, E. M.; SUZUKI, A. K. Bayesian survival model induced by frailty for lifetime with long-term survivors. **Statistica Neerlandica**, Wiley Online Library, v. 75, n. 3, p. 299–323, 2021. Citado nas páginas 22 e 30.

CHEN, M.-H.; IBRAHIM, J. G.; SINHA, D. A new bayesian model for survival data with a surviving fraction. **Journal of the American Statistical Association**, Taylor & Francis, v. 94, n. 447, p. 909–919, 1999. Citado na página 22.

CHEN, T.-T. Statistical issues and challenges in immuno-oncology. **Journal for immunotherapy of cancer**, Springer, v. 1, n. 1, p. 1–9, 2013. Citado na página 30.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006. Citado nas páginas 27, 34 e 35.

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

HALL, D. B. Zero-inflated poisson and binomial regression with random effects: a case study. **Biometrics**, Wiley Online Library, v. 56, n. 4, p. 1030–1039, 2000. Citado na página 22.

JESS, T.; WINTHER, K.; MUNKHOLM, P.; LANGHOLZ, E.; BINDER, V. Intestinal and extra-intestinal cancer in crohn's disease: follow-up of a population-based cohort in copenhagen county, denmark. **Alimentary pharmacology & therapeutics**, Wiley Online Library, v. 19, n. 3, p. 287–293, 2004. Citado na página 21.

JM, B.; ALTMAN, D. The logrank test. **BmJ**, v. 328, n. 7447, p. 1073, 2004. Citado na página 33.

JR, M. Ribeiro de O.; MOREIRA, F.; LOUZADA, F. The zero-inflated promotion cure rate model applied to financial data on time-to-default. **Cogent Economics & Finance**, Taylor & Francis, v. 5, n. 1, p. 1395950, 2017. Citado na página 22.

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

LAMBERT, D. Zero-inflated poisson regression, with an application to defects in manufacturing. **Technometrics**, Taylor & Francis, v. 34, n. 1, p. 1–14, 1992. Citado na página 22.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. [S.l.]: John Wiley & Sons, 2011. v. 362. Citado na página 22.

LEÃO, J.; BOURGUIGNON, M.; GALLARDO, D. I.; ROCHA, R.; TOMAZELLA, V. A new cure rate model with flexible competing causes with applications to melanoma and transplantation data. **Statistics in Medicine**, Wiley Online Library, v. 39, n. 24, p. 3272–3284, 2020. Citado na página 22.

LOUZADA, F.; JR, M. R. d. O.; MOREIRA, F. F. The zero-inflated cure rate regression model: Applications to fraud detection in bank loan portfolios. **arXiv preprint arXiv:1509.05244**, 2015. Citado nas páginas 22, 23, 31, 32 e 40.

LOUZADA, F.; MOREIRA, F. F.; OLIVEIRA, M. R. de. A zero-inflated non default rate regression model for credit scoring data. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 47, n. 12, p. 3002–3021, 2018. Citado nas páginas 40 e 42.

MEEKER, W. Q.; ESCOBAR, L. A. **Statistical Methods for Reliability Data.** New York,NY: Wiley, 1998. Citado na página 40.

PARK, Y.; CHEON, J. H.; PARK, Y. L.; YE, B. D.; KIM, Y. S.; HAN, D. S.; KIM, J. S.; HONG, S. N.; KIM, Y. H.; JEON, S. R. *et al.* Development of a novel predictive model for the clinical course of crohn's disease: results from the connect study. **Inflammatory bowel diseases**, Oxford University Press Oxford, UK, v. 23, n. 7, p. 1071–1079, 2017. Citado na página 21.

RAMOS, P. L.; NASCIMENTO, D.; LOUZADA, F. The long term fr\'echet distribution: Estimation, properties and its application. **arXiv preprint arXiv:1709.07593**, 2017. Citado na página 22.

RIDOUT, M.; HINDE, J.; DEMÉTRIO, C. G. A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. **Biometrics**, Wiley Online Library, v. 57, n. 1, p. 219–223, 2001. Citado na página 22.

SOUZA, H. C. C. d. **O modelo Weibull Modificado Exponenciado de Longa Duração aplicado à sobrevida do câncer de mama**. Tese (Doutorado) — Universidade de São Paulo, 2015. Citado na página 37.

SOUZA, H. C. Cavenague de; LOUZADA, F.; OLIVEIRA, M. R. de; FAWOLE, B.; AKINTAN, A.; OYENEYIN, L.; SANNI, W.; PERDONÁ, G. d. S. C. The log-normal zero-inflated cure regression model for labor time in an african obstetric population. **Journal of Applied Statistics**, Taylor & Francis, p. 1–14, 2021. Citado na página 22.

STACY, E. W. A generalization of the gamma distribution. **The Annals of mathematical statistics**, JSTOR, p. 1187–1192, 1962. Citado na página 40.

TEAM, R. C. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Disponível em: <http://www.R-project.org/>. Citado na página 41.


ICMC
USP
SÃO CARLOS