Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Modelo de regressão *odds*-riscos proporcionais para dados de sobrevivência discretos

Maria Gabriella Figueiredo Vieira

Brasília

2018

Modelo de regressão *odds*-riscos proporcionais para dados de sobrevivência discretos

Maria Gabriella Figueiredo Vieira

Orientador: Prof. Dr. Eduardo Yoshio Nakano

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília para obtenção do título de Mestre em Estatística.

Brasília

2018

À minha mãe que é Luz no meu caminho e à minha Bella irmã que é companhia eterna.

Resumo

Existe uma extensa literatura que desenvolve as técnicas de Análise de Sobrevivência para tempo de falha contínuo, principalmente no que diz respeito ao ajuste de modelos de regressão. O uso de modelos contínuos para analisar uma variável discreta não é aceitável em determinadas situações. Dessa forma, como o referencial teórico é restrito, o desenvolvimento de estudos nessa área torna-se muito relevante. O objetivo desse trabalho, portanto, foi estruturar um modelo de regressão para dados de sobrevivência com tempo de falha discreto. Para isso, tomando como base o modelo de regressão de riscos proporcionais de Cox, foi proposto um modelo análogo que adapta a teoria para que as covariáveis ajam multiplicativamente na *odds* (chance) do risco. Como resultado, foi obtida uma técnica de ajuste do modelo de regressão para dados discretos, propondo uma nova forma para ajustar o modelo, verificar as suposições e analisar a qualidade do ajuste. Usando a base de dados de um estudo feito com pacientes que sentiam dor lombar, o modelo de regressão *odds*-riscos proporcionais foi ajustado e desenvolveu-se toda a análise proposta. Concluiu-se que o modelo de regressão *odds*-riscos proporcionais é uma técnica boa e completa para analisar as variáveis que são originalmente discretas.

Palavras-chave: Análise de Sobrevivência, variável discreta, modelo de regressão, razão de chances, dor lombar, tempo de falha discreto

Abstract

There is an extensive literature that develops the Survival Analysis techniques to continuous failure times, mainly as regards the adjustment of regression models. The use of continuous models to analyze a discrete variable is not acceptable in certain situations. Thus, since the theoretical reference is restricted, the development of studies in this area becomes very relevant. Therefore, the purpose of this work is to structure a regression model for survival data with discrete failure time. For this, based on the Cox proportional hazards regression model, an analog model is proposed that adapts the theory so that the covariables act multiplicatively in the risk odds. As result, a technique of regression model adjustment for discrete data is obtained, proposing a new way to adjust the model, to verify the assumptions and to analyze the adjustment quality. Using the database of a study done with patients who feel low back pain, the proportional odds-risks regression model is adjusted and the whole proposed analysis is carried out. It is concluded that the proportional odds-risks regression model is a good and complete technique to analyze the variables that are originally discrete.

Keywords: Survival analysis, discrete variable, regression model, odds ratio, low back pain, discrete failure time

Sumário

Lista de Tabelas

Lista de Figuras

1. Introdução

A Análise de Sobrevivência é utilizada em aplicações de vários segmentos, tais como, a área da saúde, engenharias, entre outros. Trata-se de uma área da Estatística que estuda o tempo até a ocorrência de um evento de interesse (tempo de falha). O que diferencia esse tipo de análise das técnicas estatísticas convencionais é a presença de censura nos dados.

Na metodologia de Análise de Sobrevivência, o tempo de falha é a variável a ser explicada. Essa variável pode tomar a forma discreta ou contínua. A maioria dos estudos desenvolvidos leva em consideração apenas a situação em que o tempo até a ocorrência do evento de interesse é contínuo. Porém, segundo Nakano e Carrasco (2006), o uso de modelos contínuos para analisar o tempo de sobrevivência discreto não é aceitável em todas as situações.

Algumas distribuições e modelos foram propostos para o caso em que o tempo de sobrevivência é discreto. Como existiam algumas distribuições contínuas que se adequavam muito bem aos dados de sobrevivência, foram desenvolvidos estudos para adaptar as distribuições para as variáveis discretas por meio da discretização de distribuições contínuas. Essa metodologia foi utilizada, por exemplo, no texto de Nakano e Carrasco (2006) e Carrasco et al. (2012), com a discretização da distribuição exponencial, por Brunello e Nakano (2015) e Vila et al. (2018), com a distribuição Weibull discreta e por Vieira e Nakano (2017), com a distribuição log-logística discreta. Foram propostos alguns modelos de regressão que utilizavam essas novas distribuições nos trabalhos de Nobre (2017), Santos (2017) e Cardial (2017).

Na análise dos dados de sobrevivência também é de interesse fazer a inclusão de covariáveis que possam de alguma maneira contribuir na explicação do tempo de sobrevivência. Uma forma comum de modelar dados contínuos na presença de covariáveis é por meio da função de risco. Esses modelos são denominados de modelos de riscos proporcionais e consideram que as covariáveis agem multiplicativamente na função de risco. No entanto, esse tipo de modelagem não pode ser adotado quando os tempos de sobrevivência são discretos, pois neste caso

a função de risco é limitada e assume valores somente no intervalo $(0, 1)$.

Neste contexto, o principal objetivo do trabalho é propor um modelo de regressão para dados discretos de sobrevivência que considera que as covariáveis ajam multiplicativamente na *odds* (chance) do risco.

Com a proposição do modelo, é necessário que algumas suposições sejam verificadas e que a qualidade de seu ajuste seja analisada. Todos esses aspectos são explanados ao longo do texto e ilustrados por meio da aplicação em um conjunto de dados reais com pacientes que têm dor lombar. No Capítulo 2 é apresentada uma revisão bibliográfica que abrange conceitos básicos em Análise de Sobrevivência, definições de inferência clássica e apresentação de distribuições discretas. O Capítulo 3 introduz a definição do modelo de regressão *odds*-riscos proporcionais, propõe uma adaptação para verificação de *odds*-riscos proporcionais e mostra os modelos de regressão *odds*-riscos proporcionais para as distribuições Weibull discreta e log-logística discreta. No Capítulo 4, toda a metodologia proposta é aplicada em um conjunto de dados reais, resultando em análise descritiva, verificação da suposição de *odds*-riscos proporcionais, ajuste do modelo de regressão e análise da qualidade do ajuste do modelo. As inferências dos parâmetros do modelo são realizadas no contexto clássico (frequentista) e a metodologia proposta é implementada através do *software* estatístico R.

2. Revisão bibliográfica

2.1. Conceitos básicos em Análise de Sobrevivência

A Análise de Sobrevivência estuda o tempo até a ocorrência do evento de interesse por meio de um conjunto de técnicas estatísticas. Uma característica importante para utilização dessa ferramenta é a presença de informações incompletas nos dados, ou seja, existe a possibilidade de um indivíduo não ser acompanhado até a observação do evento de interesse. Assim, esses procedimentos têm aplicabilidade em diversas áreas, sendo amplamente utilizados na área médica. Dessa forma, a seguir, os conceitos básicos em Análise de Sobrevivência são apresentados para melhor compreensão das técnicas.

2.1.1. Tempo de falha e censura

O objeto de estudo na Análise de Sobrevivência é denominado tempo de falha (ou tempo de sobrevivência), ou seja, o tempo até a ocorrência do evento de interesse. Alguns exemplos podem ser citados na área da saúde, como é o caso do tempo até o óbito do paciente, ou na Engenharia, em que se observa o tempo até a ocorrência de um problema no equipamento, entre outros.

Em algumas situações, o tempo de falha é observado parcialmente por algum motivo que não está no controle do pesquisador ou até mesmo pelo término da pesquisa. Essa interrupção no acompanhamento do tempo até a ocorrência do evento de interesse é denominada censura. A informação do tempo censurado deve ser considerada na análise, sendo necessária a adaptação das metodologias já estabelecidas.

Existem três tipos de censura: à direita, à esquerda e intervalar. Na censura à direita, o tempo registrado (censurado) é menor que o tempo que seria observado caso não houvesse a interrupção. Na censura à esquerda, o evento de interesse ocorre antes mesmo do tempo ser

registrado, ou seja, o tempo registrado é maior do que o tempo de falha. Por fim, na censura intervalar não é possível identificar qual o tempo exato que aconteceu a falha, se conhece apenas um intervalo de tempo em que ocorreu o evento de interesse.

Além dessa classificação, a censura à direita ainda pode ser dividida em três outros grupos: censura do tipo I, censura do tipo II e censura aleatória. A censura do tipo I ocorre quando chega o término do estudo (determinado previamente) e o evento de interesse não foi observado para determinados indivíduos. Considerando que no início da pesquisa é estabelecido o número de falhas observadas, na censura do tipo II o estudo termina após a ocorrência dessas falhas. Ao final do experimento os indivíduos que não provaram do evento de interesse são considerados censurados. Por fim, na censura aleatória o indivíduo sai do estudo antes do término por algum motivo que não pode ser controlado pelo pesquisador.

Nesse contexto, para a i-ésima observação dos dados, o tempo de falha (ou censura) é representado por $t_i, i = 1, ..., n$, e a variável aleatória que indica se existe censura é definida como

$$\delta_i = \begin{cases} 1, \text{ se } t_i \text{ é tempo de falha} \\ 0, \text{ se } t_i \text{ é tempo de censura.} \end{cases}$$

Além disso, em algumas situações, o tempo de sobrevivência pode sofrer influência de alguma outra característica do indivíduo. Esse tipo de informação pode ser incorporada na estimação das funções que descrevem o comportamento do tempo de sobrevivência. Então, considerando $\mathbf{x}_i = (\mathrm{x}_{1i}, \mathrm{x}_{2i}, ..., \mathrm{x}_{pi})$ um vetor de covariáveis do i-ésimo indivíduo, os dados de sobrevivência podem ser representados por $(t_i, \delta_i, \mathbf{x}_i)$ (Colosimo e Giolo, 2006).

2.1.2. Estimador de Kaplan-Meier

O estimador de Kaplan-Meier é a ferramenta não paramétrica mais utilizada para estimar a função de sobrevivência na presença de censuras. Esse estimador funciona muito bem para uma análise inicial dos dados, visto que as técnicas usuais para cálculo de medidas resumo não funcionam bem nessa situação (Colosimo e Giolo, 2006).

Considere tempos distintos de falha $t_1, t_2, ..., t_k$, em que $t_1 < t_2 < ... < t_k$. Existem n indivíduos com seus respectivos tempos de falhas e dentre esses, k são tempos distintos que não apresentam censuras. Dessa forma, tem-se que $k \leq n$ e cada tempo $t_j, j = 1, ..., k$, pode ser observado mais de uma vez. O estimador de Kaplan-Meier para a função de sobrevivência é definido como (Kaplan e Meier, 1958)

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

sendo que d_j é o número de falhas no tempo t_j e n_j é o número de indivíduos que não provaram do evento de interesse e que não foram censurados até o tempo imediatamente anterior a t_j (Colosimo e Giolo, 2006).

Seja $w > 0$ e $t_1 \leq w < t_2$, onde w é um tempo qualquer e t_1, t_2 são tempos de falha. As estimativas de $S(t)$ respeitam a seguinte relação

$$\hat{S}(w) = P(T > w) = P(T > t_1) = \hat{S}(t_1).$$

Esse é um dos motivos que a representação gráfica da função de sobrevivência estimada por Kaplan-Meier tem a forma de escada.

2.1.3. Descrição do comportamento do tempo discreto de sobrevivência

Seja T uma variável aleatória discreta e não-negativa. A função de probabilidade é definida por $p(t) = P(T = t)$ e satisfaz as condições a seguir (Magalhães, 2006):

1. $0 \leq p(t) \leq 1$;

2. $\sum_{t=0}^{\infty} p(t) = 1$.

Como T representa o tempo de falha, então $p(t)$ é definida como a probabilidade desse indivíduo experimentar o evento de interesse em um tempo t.

Conhecendo a função massa de probabilidade, é possível obter a função de distribuição acumulada $F(t)$ da variável aleatória T através da expressão

$$F(t) = P(T \leq t) = \sum_{k=0}^{t} p(k) = \sum_{k=0}^{t} P(T = k),\ t = 0, 1, 2, \ldots . \tag{2}$$

A expressão (2) mostra a probabilidade de um indivíduo experimentar o evento de interesse em um tempo menor ou igual a determinado valor t.

O tempo de falha (ou de sobrevivência) também pode ser especificado pela função de sobrevivência, $S(t)$, que representa a probabilidade do indivíduo sobreviver ao tempo t. Segundo Colosimo e Giolo (2006), essa função é a mais importante para a análise descritiva do tempo de sobrevivência. Para uma variável aleatória discreta, $S(t)$ é dada por

$$S(t) = P(T > t) = \sum_{k=t+1}^{\infty} p(k) = \sum_{k=t+1}^{\infty} P(T = k),\ t = 0, 1, 2, \ldots . \tag{3}$$

Com as fórmulas (2) e (3), encontra-se a relação

$$F(t) = 1 - S(t). \tag{4}$$

Para variáveis aleatórias discretas, a função de sobrevivência em $t = 0$ não é necessariamente igual a 1. Essa é uma característica importante que diferencia $S(t)$ no caso discreto e no contínuo (Santos, 2017).

Uma outra função usada para caracterizar o comportamento do tempo discreto de sobrevivência é a função de risco (ou taxa de falha). Trata-se de uma função que representa a probabilidade do indivíduo falhar no tempo t, dado que ele sobreviveu até esse tempo t. A taxa de falha é representada por

$$h(t) = P(T = t | T \geq t),\ t = 0, 1, 2, \ldots, \tag{5}$$

e para $t < 0$ ou não-inteiro essa função é dada por $h(t) = 0$. É importante destacar que a função

de risco pode assumir valores somente no intervalo $[0, 1]$, já que se trata de uma probabilidade (Nakano, 2017).

Assim como a função de distribuição acumulada resulta da soma da função de probabilidade nos pontos 0 a t, a função de risco acumulada (ou taxa de falha cumulada) é obtida através da soma da função de risco e é dada por

$$H(t) = \sum_{k=0}^{t} h(k),\ t = 0, 1, 2, \ldots . \tag{6}$$

Utilizando as expressões (3) e (5), é possível encontrar a seguinte relação

$$h(t) = \frac{p(t)}{p(t) + S(t)},\ t = 0, 1, 2, \ldots . \tag{7}$$

Por consequência, a função de probabilidade pode ser expressa por

$$p(t) = \frac{h(t)}{1 - h(t)} S(t),\ t = 0, 1, 2, \ldots . \tag{8}$$

Além das funções apresentadas para descrever o tempo discreto de sobrevivência, algumas relações podem ser demonstradas. Segundo Nakano (2017), $p(t)$ pode ser reescrita usando a função de sobrevivência da seguinte forma

$$p(t) = \begin{cases} 1 - S(0),\ \text{se } t = 0 \\ S(t-1) - S(t),\ \text{se } t = 1, 2, \ldots \end{cases} \tag{9}$$

por outro lado, a função de sobrevivência pode ser expressa em termos da função de risco,

$$S(t) = \prod_{k=0}^{t} [1 - h(k)],\ t = 0, 1, 2, \ldots . \tag{10}$$

Então, usando (8) e (10), encontra-se a função de probabilidade como função de $h(t)$,

expressa por

$$p(t) = \frac{h(t)}{1-h(t)} \prod_{k=0}^{t} [1-h(k)],\ t = 0, 1, 2, \ldots . \tag{11}$$

2.2. Inferência clássica

Como dito anteriormente, os dados de sobrevivência têm em sua estrutura a ocorrência de censuras. Portanto, é preciso incorporar essa informação na estimação dos parâmetros de interesse. Nessa seção serão apresentadas as ferramentas para estimação pontual, estimação intervalar e testes de hipóteses no contexto de Análise de Sobrevivência para tempos de sobrevivência discretos.

2.2.1. Método de máxima verossimilhança

Considere uma variável aleatória discreta T com função de probabilidade $p(t;\boldsymbol{\theta})$, sendo $\boldsymbol{\theta}$ um parâmetro desconhecido ou um vetor de parâmetros. Retirando uma amostra aleatória de T, são observados os valores $t_1, t_2, ..., t_n$. Então, a função de verossimilhança dessa amostra é definida por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(t_i;\boldsymbol{\theta}) \ .$$

O método de estimação de máxima verossimilhança consiste na maximização da função de verossimilhança, ou seja, deseja-se encontrar uma estimativa para $\boldsymbol{\theta}$ que torne máximo o valor de $L(\boldsymbol{\theta})$.

Na situação em que os dados não são censurados, a parcela de contribuição em $L(\boldsymbol{\theta})$ é equivalente a função de probabilidade $p(t;\boldsymbol{\theta})$. O mesmo não acontece para dados censurados. Nos dados de sobrevivência, é necessário ter uma função que explique as observações que não são censuradas e outra função para as censuras. A função de sobrevivência consegue expressar essa contribuição na função de verossimilhança com a presença de informação censurada. Assim, divide-se as observações em uma parte com r elementos não-censurados e outra com $n-r$ censurados (Colosimo e Giolo, 2006).

Considerando $t_i's$ independentes, a função de verossimilhança é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{r} p(t_i;\boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i;\boldsymbol{\theta}) \, ,$$

que é equivalente a

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n} [p(t_i;\boldsymbol{\theta})]^{\delta_i} \, [S(t_i;\boldsymbol{\theta})]^{1-\delta_i} \, , \tag{12}$$

sendo que δ_i é a variável indicadora de censura, que assume o valor 1, se t_i é um tempo de falha ou 0, se t_i é um tempo censurado.

Definida a função de verossimilhança, aplica-se o logaritmo, cuja notação é $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$, para obter os estimadores de máxima verossimilhança (estimadores pontuais). Esses estimadores são representados por $\hat{\boldsymbol{\theta}}$ e resultam da resolução do sistema de equações gerados por

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \, .$$

2.2.2. Intervalo de confiança e teste de significância dos parâmetros

A estimação intervalar é caracterizada pelos intervalos de confiança que são construídos com base na distribuição assintótica do estimador de máxima verossimilhança $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)$ para grandes amostras. Segundo Colosimo e Giolo (2006), $\hat{\boldsymbol{\theta}}$ segue uma normal multivariada representada por

$$\hat{\boldsymbol{\theta}} \sim N_k(\theta, Var(\hat{\boldsymbol{\theta}})).$$

Sob condições de regularidade, a matriz de variância-covariância é dada por

$$Var(\hat{\boldsymbol{\theta}}) \approx \left[I(\hat{\boldsymbol{\theta}}) \right]^{-1} ,$$

sendo que

$$I(\hat{\boldsymbol{\theta}}) = - \left(\frac{\partial^2 l(\boldsymbol{\theta})}{(\partial \boldsymbol{\theta})^2} \right) \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

é a informação de Fisher observada.

Usando o método de máxima verossimilhança para estimar os parâmetros de interesse, encontra-se $\hat{\boldsymbol{\theta}}$ e fica simples obter $Var(\hat{\boldsymbol{\theta}})$. Com isso, o intervalo de $100\times(1-\alpha)\%$ de confiança para cada $\theta_j, j = 1, 2, ..., m$ é expresso por

$$IC_{100\times(1-\alpha)\%}(\theta_j) = \left[\hat{\theta_j} - Z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{Var(\hat{\theta_j})};\ \hat{\theta_j} + Z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{Var(\hat{\theta_j})}\right],$$

com $Z_{\left(1-\frac{\alpha}{2}\right)}$ representando o quantil $\left(1-\frac{\alpha}{2}\right)$ de uma distribuição normal padrão.

Por outro lado, quando o interesse é encontrar um intervalo de confiança para uma função dos parâmetros $g(\hat{\boldsymbol{\theta}})$, é preciso calcular $Var[g(\hat{\boldsymbol{\theta}})]$. Assim, pelo do método delta, tem-se que

$$g(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} N\left(g(\boldsymbol{\theta}),\ \left[g'(\hat{\boldsymbol{\theta}})\right]^2 Var(\hat{\boldsymbol{\theta}})\right),$$

sendo $g(\hat{\boldsymbol{\theta}})$ o estimador de máxima verossimilhança de $g(\boldsymbol{\theta})$ e $g'(\hat{\boldsymbol{\theta}})$ a primeira derivada de $g(\hat{\boldsymbol{\theta}})$.

Outra ferramenta usada na inferência clássica é o teste de hipóteses. Com o objetivo de testar a significância das estimativas dos parâmetros de um modelo, são definidas as hipóteses

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \ versus\ H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

A estatística do teste é dada por

$$Z_0 = \frac{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0}{\sqrt{Var(\hat{\boldsymbol{\theta}})}} \sim N(0, 1).$$

O *p*-valor associado ao teste pode ser escrito da seguinte forma,

$$p - valor = 2P(Z > |Z_0|),$$

em que Z é uma variável aleatória com distribuição normal padrão.

Dessa forma, o vetor de parâmetros $\hat{\boldsymbol{\theta}}$ é considerado significativamente diferente de $\boldsymbol{\theta}_0$

se $p - valor < \alpha$, sendo α um valor estabelecido previamente.

2.3. Distribuições discretas em Análise de Sobrevivência

2.3.1. Método para obter distribuições discretas

Algumas distribuições discretas podem resultar de distribuições usadas para variáveis aleatórias contínuas. Considere X uma variável aleatória contínua que assume somente valores maiores ou iguais a zero. A variável aleatória discreta T pode ser pensada como a "parte inteira de X". Como $F(x)$ é conhecido para X, a função de probabilidade de T pode ser definida como (Nakano e Carrasco, 2006)

$$\begin{aligned} p(t) &= P(T = t) = P(t \leq X < t + 1) \\ &= F_X(t + 1) - F_X(t),\ t = 0, 1, 2, \ldots . \end{aligned} \tag{13}$$

As demais funções usadas para descrever o tempo discreto de sobrevivência decorrem das relações enunciadas na seção 2.1.3.

2.3.2. Distribuição geométrica

Seja T uma variável aleatória discreta que tem distribuição geométrica com parâmetro p, $0 < p < 1$. Sua função de probabilidade é dada por

$$p(t) = P(T = t) = p(1 - p)^t,\ t = 0, 1, 2, \ldots . \tag{14}$$

Além de $p(t)$, é possível definir a função de sobrevivência e a função de risco. A primeira pode ser escrita como

$$S(t) = P(T > t) = (1 - p)^{t+1},\ t = 0, 1, 2, \ldots \tag{15}$$

e a segunda é definida por

$$h(t) = p,\ t = 0, 1, 2, \ldots . \tag{16}$$

Percebe-se que a função de risco de uma variável aleatória T com distribuição geométrica é constante para todo $t = 0, 1, 2, ...$ e somente assume valores no intervalo $[0, 1]$.

2.3.3. Distribuição Weibull discreta

A distribuição Weibull é muito utilizada na modelagem do tempo de sobrevivência por ser flexível (Cardial, 2017). Trata-se de uma distribuição popular pelo fato de ter várias formas dependendo da escolha dos parâmetros.

Proposta por Nakagawa e Osaki (1975), a distribuição Weibull discreta resulta da expressão (13). Sua função de probabilidade é dada por

$$p(t) = q^{t^\gamma} - q^{(t+1)^\gamma},\ t = 0, 1, 2, \ldots , \tag{17}$$

sendo $q = \exp\left(-\frac{1}{\alpha^\gamma}\right)$, $0 < q < 1$, $\alpha > 0$ e $\gamma > 0$. A partir da expressão (17), encontra-se a função de sobrevivência

$$S(t) = q^{(t+1)^\gamma},\ t = 0, 1, 2, ... \tag{18}$$

e a função de risco

$$h(t) = \frac{q^{t^\gamma} - q^{(t+1)^\gamma}}{q^{t^\gamma}},\ t = 0, 1, 2, \ldots . \tag{19}$$

Segundo Cardial (2017), a função de risco assume as seguintes formas: estritamente crescente, se $\gamma > 1$; constante, se $\gamma = 1$, reduzindo a distribuição Weibull discreta a uma distribuição geométrica; e estritamente decrescente, se $\gamma < 1$.

2.3.4. Distribuição log-logística discreta

A distribuição log-logística discreta é obtida através da discretização da distribuição log-logística contínua.

Segundo Santos (2017), seja $\alpha > 0$ o parâmetro de escala e $\gamma > 0$ o parâmetro de forma da distribuição log-logística contínua, então, a partir da expressão (13), a função de probabilidade da distribuição log-logística discreta é dada por

$$p(t) = \frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} - \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}},\ t = 0, 1, 2, \ldots . \tag{20}$$

Dessa forma, as funções de sobrevivência e de risco são dadas, respectivamente, por

$$S(t) = \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}},\ t = 0, 1, 2, \ldots \tag{21}$$

e

$$h(t) = \frac{\frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} - \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}}}{\frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}}},\ t = 0, 1, 2, \ldots . \tag{22}$$

3. Modelo de regressão *odds*-riscos proporcionais

No capítulo anterior foram apresentadas ferramentas para descrever o tempo de sobrevivência sem considerar a influência das covariáveis na variável resposta T. Neste capítulo será abordada uma metodologia para incluir covariáveis no modelo por meio da função de risco de variáveis discretas.

3.1. Método para obtenção do modelo de regressão *odds*-riscos proporcionais

Seja T uma variável aleatória contínua e $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_p)'$ um vetor de p covariáveis. É possível ajustar um modelo de riscos proporcionais com a finalidade de incluir essas covariáveis no modelo. Segundo Colosimo e Giolo (2006) esse modelo é expresso por

$$h(t|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})\, h_0(t), \tag{23}$$

em que $g(\mathbf{x}'\boldsymbol{\beta})$ é uma função não-negativa que assume o valor 1 quando o seu argumento é nulo e $h_0(t)$ é a função de risco base (função de risco quando todas as covariáveis são iguais a zero).

Como visto em (5), no caso em que T é uma variável aleatória discreta, a função de risco é uma probabilidade, ou seja, $0 < h(t) < 1$. Nesse contexto, não há como usar riscos proporcionais para incluir covariáveis no modelo de regressão, mas é possível usar a chance (*odds*) de $h(t)$, isto é

$$odds\{h(t)\} = \frac{h(t)}{1 - h(t)}. \tag{24}$$

Como $odds\{h(t)\} > 0$, o modelo proposto considera que as covariáveis $\mathbf{x}$ agem mul-

tiplicativamente (proporcionalmente) no *odds* do risco. Isto é,

$$odds\{h(t|\mathbf{x})\} = g(\mathbf{x}'\boldsymbol{\beta})\ odds\{h_0(t)\},$$

que resulta em

$$\frac{h(t|\mathbf{x})}{1-h(t|\mathbf{x})} = g(\mathbf{x}'\boldsymbol{\beta})\ \frac{h_0(t)}{1-h_0(t)}, \tag{25}$$

sendo que $h_0(t)$ é a função de risco base, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ é o vetor de coeficientes associado ao vetor de covariáveis $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_p)'$ e $g(\cdot)$ é uma função de ligação que satisfaz as seguintes condições:

1. $g(a) > 0,\ \forall a \in \mathbb{R}$;

2. $g(0) = 1$.

Note que o intercepto β_0 não aparece no preditor linear $\mathbf{x}'\boldsymbol{\beta}$. Isto porque a função de risco base, $h_0(t)$, absorve este termo constante.

Além disso, a propriedade *odds*-riscos proporcionais permite interpretar os coeficientes estimados. Considerando a função de ligação exponencial $(g(\cdot) = \exp(\cdot))$, note que a *odds* do risco de dois indivíduos (r e s) que apresentam os mesmos valores para as covariáveis, exceto a m-ésima delas, é dada por

$$\frac{odds\{h(t|\mathbf{x_r})\}}{odds\{h(t|\mathbf{x_s})\}} = \frac{\exp\{\beta_m \mathrm{x}_{rm}\}}{\exp\{\beta_m \mathrm{x}_{sm}\}} = \exp\{\beta_m(\mathrm{x}_{rm} - \mathrm{x}_{sm})\}, \tag{26}$$

que não depende de t.

Observe que (26) é uma razão de chances (chances dos riscos) e, assim, se por exemplo x_m é a covariável dicotômica sexo com $\mathrm{x}_{rm} = 1$ (masculino) e $\mathrm{x}_{sm} = 0$ (feminino), tem-se que a chance de falha (*odds* do risco) dos indivíduos do sexo masculino é $\exp(\beta_m)$ vezes a chance de falha dos indivíduos do sexo feminino, mantendo-se fixas as demais covariáveis.

A partir da expressão (25), é fácil ver que a função de risco de um indivíduo com

covariáveis $\mathbf{x}$ é dada por

$$h(t|\mathbf{x}) = \frac{g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}{1 - h_0(t) + g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}. \tag{27}$$

Segundo a expressão (10) e (27), a função de sobrevivência na presença de covariáveis pode ser escrita como

$$S(t|\mathbf{x}) = \prod_{u=0}^{t}[1 - h(u|\mathbf{x})] = \prod_{u=0}^{t}\left[\frac{1 - h_0(u)}{1 - h_0(u) + g(\mathbf{x}'\boldsymbol{\beta})\ h_0(u)}\right],\ t = 0, 1, 2, \ldots . \tag{28}$$

Assim, usando (9) e (10), a função de probabilidade é dada por

$$p(t|\mathbf{x}) = \begin{cases} h(0|\mathbf{x}),\ \text{se } t = 0 \\ h(t|\mathbf{x})\ S(t-1|\mathbf{x}),\ \text{se } t = 1, 2, \ldots . \end{cases} \tag{29}$$

Para construir o modelo de regressão *odds*-riscos proporcionais, é necessário estimar os parâmetros da distribuição e do modelo. Isso é feito ao maximizar a função de verossimilhança dada por

$$L(\boldsymbol{\theta}) \propto \prod_{i=1}^{n}\left\{\left[h(0|\mathbf{x})^{\mathbb{1}_{\{0\}}(t_i)}\ [h(t_i|\mathbf{x})\ S(t_i - 1|\mathbf{x})]^{1-\mathbb{1}_{\{0\}}(t_i)}\right]^{\delta_i}\ \left[S(0|\mathbf{x})^{\mathbb{1}_{\{0\}}(t_i)} S(t_i|\mathbf{x})^{1-\mathbb{1}_{\{0\}}(t_i)}\right]^{1-\delta_i}\right\}, \tag{30}$$

em que

$$\mathbb{1}_{\{0\}}(t_i) = \begin{cases} 1,\ \text{se } t_i = 0 \\ 0,\ \text{se } t_i \neq 0. \end{cases}$$

3.2. Verificação da suposição de *odds*-riscos proporcionais

Na literatura existem formas de verificar a suposição de riscos proporcionais no caso em que a variável aleatória T é contínua. Uma das ferramentas utilizadas é a verificação gráfica das funções de risco. A verificação gráfica consiste em: (1) estimar a função de risco acumu-

lada, $\hat{H}_j(t), j = 1, 2, 3, ...$, para cada categoria j de determinada covariável; (2) plotar as curvas de $log(\hat{H}_j(t))$ *versus* t; (3) verificar se as diferenças entre as curvas são constantes no tempo (se sim, então a suposição de riscos proporcionais é válida); (4) esse procedimento é repetido para todas as covariáveis.

Essa metodologia de verificação gráfica, no entanto, não é adequada para verificação do pressuposto de *odds*-riscos proporcionais para um modelo de regressão para variável discreta, visto que se os riscos são proporcionais, as *odds* dos riscos não serão necessariamente proporcionais. Com a impossibilidade de utilizar as ferramentas já existentes para verificação da suposição de *odds*-riscos proporcionais, buscou-se uma outra forma de mostrar essa proporcionalidade.

O modelo proposto na seção $3.1.$ pressupõe que as *odds* do risco para dois indivíduos são proporcionais. Considerando, por exemplo, uma covariável x dicotômica que assume os valores 0 e 1, o modelo supõe que

$$\frac{h(t|x=1)}{1-h(t|x=1)} = C\frac{h(t|x=0)}{1-h(t|x=0)}, \tag{31}$$

em que $h(\cdot)$ é a função de risco e C é uma constante que não depende do tempo t.

Seja $g_i(\cdot)$ a função *odds* do risco de um indivíduo com covariável $x = i, i = 0, 1$,

$$g_i(t) = \frac{h(t|x=i)}{1-h(t|x=i)}, i = 0, 1, \tag{32}$$

e $G_i(\cdot)$ sua respectiva função *odds*-risco acumulada. Isto é,

$$G_i(t) = \sum_{u=0}^{t} g_i(u) = \sum_{u=0}^{t} \frac{h(t|u=i)}{1-h(u|x=i)}, i = 0, 1. \tag{33}$$

Note que sob a suposição de *odds*-riscos proporcionais escrita em (31), tem-se a partir das expressões (32) e (33) que

$$G_1(t) = CG_0(t). \tag{34}$$

Aplicando-se o logaritmo em (34), encontra-se a seguinte equação

$$\log\left[G_1(t)\right] = \log[C] + \log\left[G_0(t)\right]. \tag{35}$$

Logo, a relação entre $\log\left[G_1(t)\right]$ e $\log\left[G_0(t)\right]$ é uma reta com coeficiente angular igual a 1.

Desta forma, a suposição de *odds*-riscos proporcionais pode ser verificada graficamente ajustando-se uma reta de regressão simples com coeficiente angular $b = 1$ (fixo). Se os pontos estiverem próximos da reta de regressão ajustada, isso indica que as *odds* dos riscos são proporcionais.

O intercepto da reta de regressão (com $b = 1$) pode ser estimado pelo método de mínimos quadrados, resultando em

$$\tilde{a} = \frac{1}{K}\sum_{k=1}^{K}\left(\log\left[G_1(t_k)\right] - \log\left[G_0(t_k)\right]\right), \tag{36}$$

em que $t_k, k = 1, 2, ..., K$, é o k-ésimo tempo distinto observado (censurado ou não censurado).

Nota 1. Para covariáveis categóricas com três ou mais níveis, o mesmo procedimento pode ser realizado comparando cada nível da covariável dois-a-dois.

Nota 2. Para covariáveis numéricas, o mesmo procedimento pode ser realizado categorizando os valores das covariáveis e comparando-os dois-a-dois.

Nota 3. Alternativamente, é possível fazer gráficos de t *versus* $\log\left[G_i(t)\right]$. Curvas paralelas (curvas com mesma distância vertical) indicam *odds*-riscos proporcionais. Este procedimento pode ser utilizado para verificação da suposição de *odds*-riscos proporcionais para variáveis categóricas com três ou mais níveis.

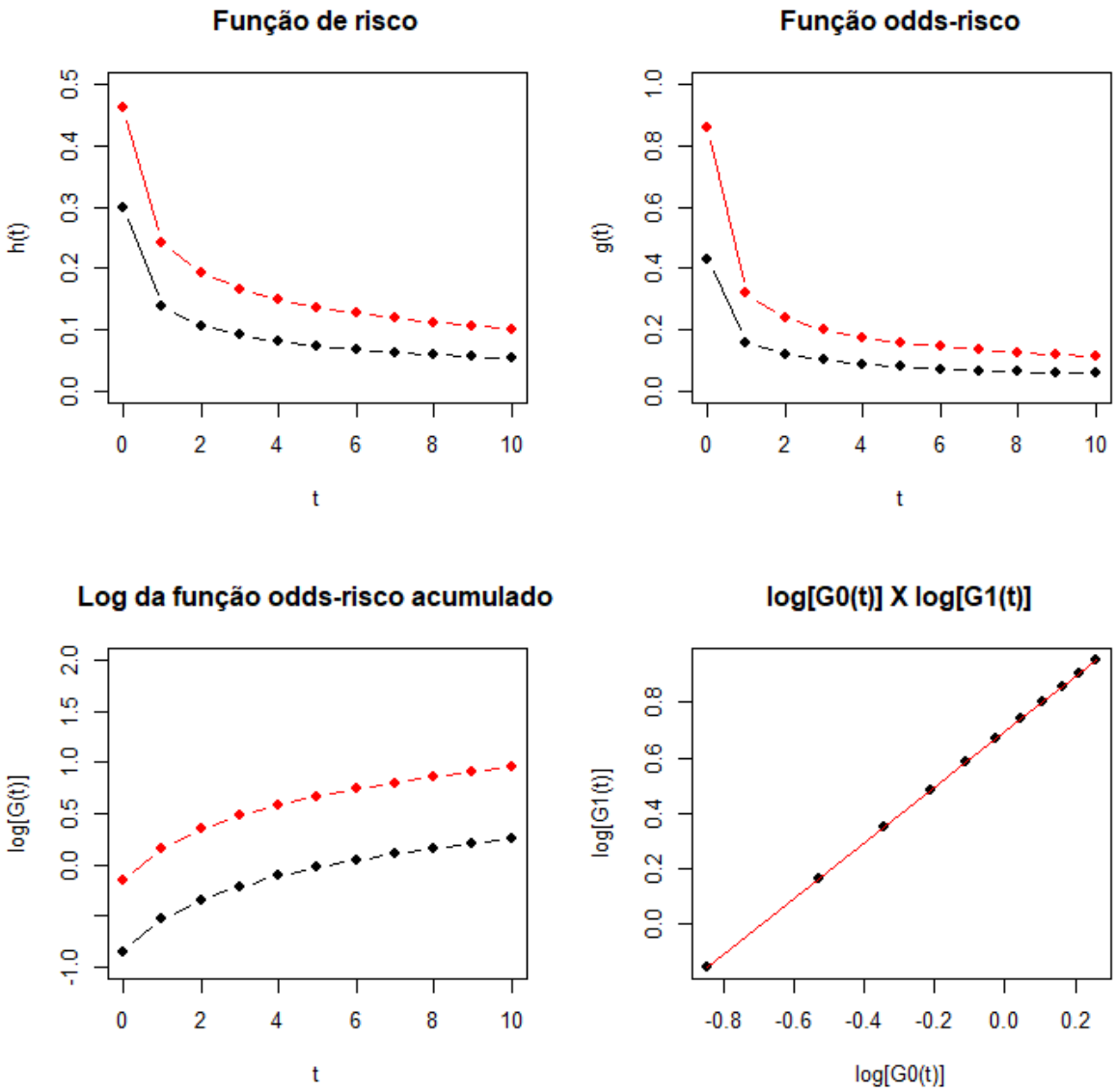

Figura 1: Funções para distribuições com *odds*-riscos proporcionais.

A Figura 1 ilustra uma situação em que as *odds*-riscos são proporcionais. Dispondo de uma covariável com 2 categorias, foi considerado que o risco de um indivíduo com covariável $x = 0$ (curva preta) comporta-se segundo uma distribuição Weibull discreta com parâmetros $q = 0,7$ e $\gamma = 0,5$. Além disso, a *odds*-risco de um indivíduo com covariável $x = 1$ (curva vermelha) é duas vezes a *odds*-risco do indivíduo com $x = 0$. Nesse caso, percebe-se que o gráfico que ilustra $log[G_0(t)]$ *versus* $log[G_1(t)]$ mostra os pontos muito próximos da reta de regressão ajustada quando as *odds*-riscos são proporcionais. Portanto, se as *odds*-riscos são

proporcionais, os riscos não serão necessariamente proporcionais (veja a não proporcionalidade dos riscos no gráfico da Figura 1 que ilustra t *versus* $h(t)$).

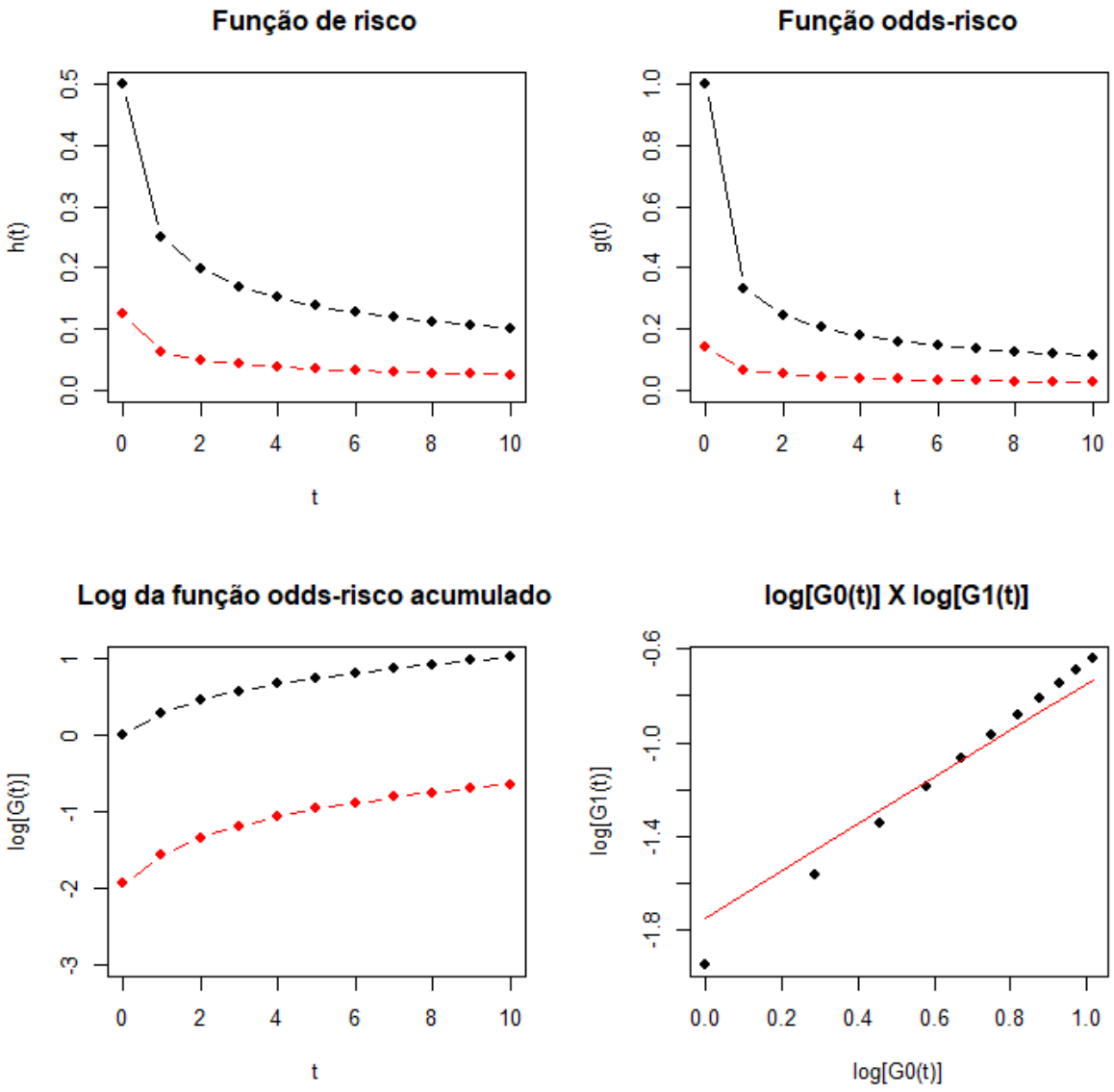

Figura 2: Funções para distribuições com riscos proporcionais.

Na Figura 2 é ilustrada a situação em que os riscos são proporcionais. Nesse caso, foi considerado que o risco de um indivíduo com covariável $x = 0$ (curva preta) comporta-se segundo uma distribuição Weibull discreta com parâmetros $q = 0,5$ e $\gamma = 0,5$ e o risco de um indivíduo com covariável $x = 1$ (curva vermelha) é 25% o risco do indivíduo com $x = 0$.

Ao plotar os pontos de $log[G_0(t)]$ *versus* $log[G_1(t)]$ junto com a reta de regressão ajustada, percebe-se que não há grande desvios, mas a reta não fica completamente em cima dos pontos. Portanto, a verificação de riscos proporcionais não implica que as *odds*-riscos também sejam proporcionais.

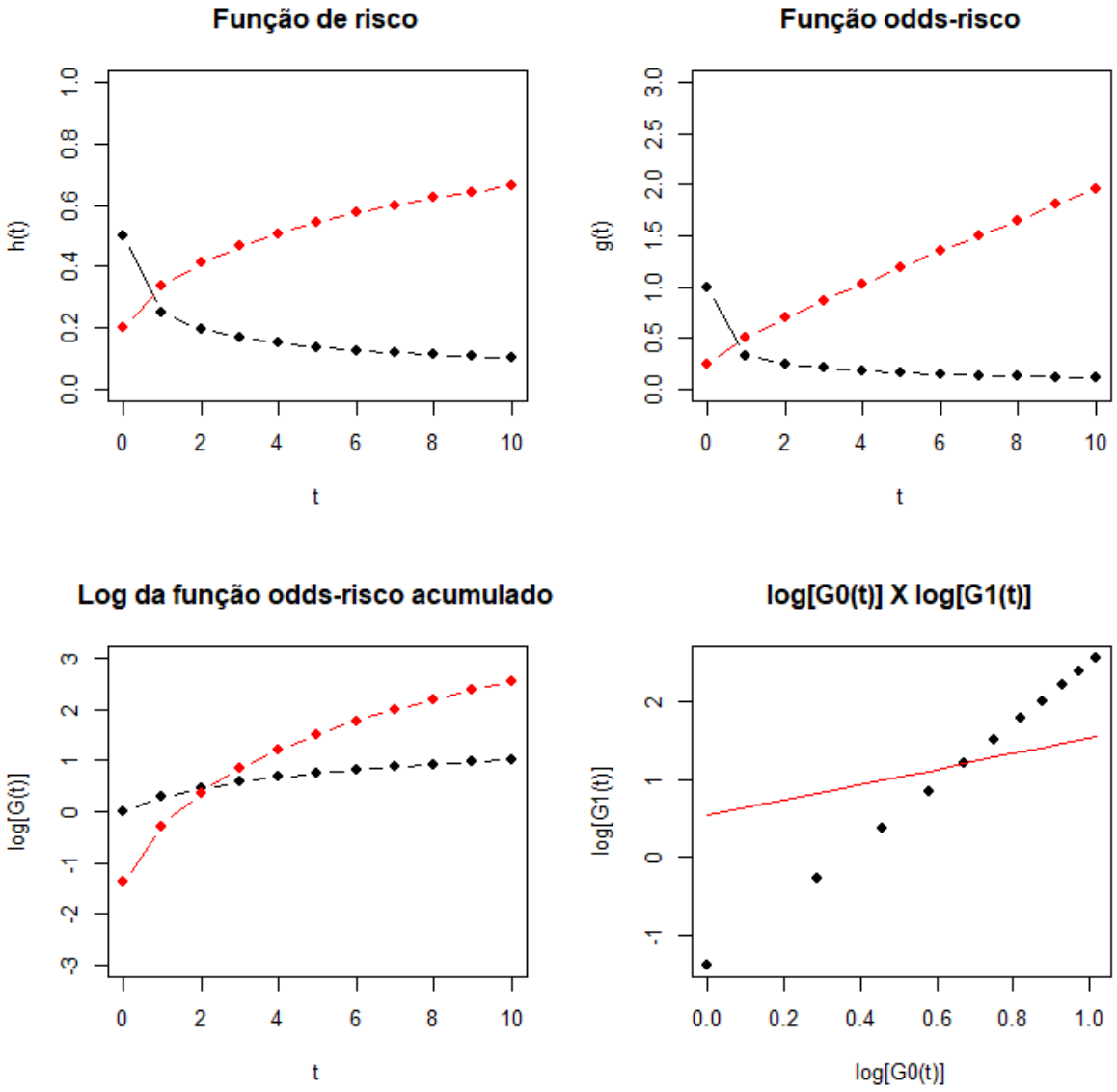

Figura 3: Funções para distribuições com riscos que se cruzam.

Na Figura 3 foi ilustrada a situação em que os riscos se cruzam e, portanto, os riscos não são proporcionais. Nesse cenário, foi considerado que o risco de um indivíduo com

covariável $x = 0$ (curva preta) comporta-se segundo uma distribuição Weibull discreta com parâmetros $q = 0,5$ e $\gamma = 0,5$ e o risco de um indivíduo com covariável $x = 1$ (curva vermelha) comporta-se segundo uma distribuição Weibull discreta com parâmetros $q = 0,8$ e $\gamma = 1,5$. Observa-se que existe uma discrepância muito grande da reta de regressão ajustada com relação ao pontos plotados no gráfico de $log[G_0(t)]$ *versus* $log[G_1(t)]$, comprovando que as *odds* dos riscos também não são proporcionais.

3.3. Modelo de regressão *odds*-riscos proporcionais Weibull discreta

Seja T uma variável aleatória com distribuição Weibull discreta. A partir das expressões (19) e (27), e considerando que $g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, encontrou-se a seguinte fórmula para a função de risco do modelo *odds*-riscos proporcionais Weibull discreta

$$h(t|\mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{t^\gamma} - q^{(t+1)^\gamma}\right)}{q^{(t+1)^\gamma} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{t^\gamma} - q^{(t+1)^\gamma}\right)},\ t = 0, 1, 2, \ldots . \tag{37}$$

Substituindo $h_0(t)$ e $g(\mathbf{x}'\boldsymbol{\beta})$ em (28), a função de sobrevivência pode ser rescrita como

$$S(t|\mathbf{x}) = \prod_{u=0}^{t}\left[\frac{q^{(u+1)^\gamma}}{q^{(u+1)^\gamma} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{u^\gamma} - q^{(u+1)^\gamma}\right)}\right],\ t = 0, 1, 2, \ldots . \tag{38}$$

Usando (29), (37) e (38), a função de probabilidade é dada por

$$p(t|\mathbf{x}) = \begin{cases} \frac{e^{\mathbf{x}'\boldsymbol{\beta}}(1-q)}{q+e^{\mathbf{x}'\boldsymbol{\beta}}(1-q)},\ \text{se } t = 0 \\ \frac{e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{t^\gamma} - q^{(t+1)^\gamma}\right)}{q^{(t+1)^\gamma}} \cdot \prod_{u=0}^{t}\left[\frac{q^{(u+1)^\gamma}}{q^{(u+1)^\gamma} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{u^\gamma} - q^{(u+1)^\gamma}\right)}\right],\ \text{se } t = 1, 2, \ldots . \end{cases}$$

Portanto, conclui-se que

$$p(t|\mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{t^\gamma} - q^{(t+1)^\gamma}\right)}{q^{(t+1)^\gamma}} \prod_{u=0}^{t}\left[\frac{q^{(u+1)^\gamma}}{q^{(u+1)^\gamma} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(q^{u^\gamma} - q^{(u+1)^\gamma}\right)}\right],\ t = 0, 1, 2, \ldots . \tag{39}$$

Sabe-se que a distribuição geométrica é um caso particular da Weibull discreta. Dessa forma, substituindo $\gamma = 1$ nas expressões (37), (38) e (39), encontra-se, respectivamente, as funções de risco, de sobrevivência e de probabilidade do modelo *odds*-riscos proporcionais geométrico. Com base na definição dessas funções, é possível encontrar a função de verossimilhança a partir da expressão (30) e estimar os parâmetros do modelo.

3.4. Modelo de regressão *odds*-riscos proporcionais log-logística discreta

Com base no modelo de regressão *odds*-riscos proporcionais apresentado na seção 3.1, foi possível definir um modelo de regressão *odds*-riscos proporcionais para a distribuição log-logística discreta.

Seja uma variável aleatória T com distribuição log-logística discreta. Utilizando as expressões (22) e (27), para $g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, chegou-se a fórmula para a função de risco do modelo *odds*-riscos proporcionais log-logística discreta

$$h(t|\mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}\left(\frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} - \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}}\right)}{\frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(\frac{1}{1+\left(\frac{t}{\alpha}\right)^{\gamma}} - \frac{1}{1+\left(\frac{t+1}{\alpha}\right)^{\gamma}}\right)}, \; t = 0, 1, 2, \ldots . \tag{40}$$

De acordo com as expressões (28) e (29), a função de sobrevivência e a função de probabilidade na presença de covariáveis dependem da função de risco e, para distribuição log-logística discreta, podem ser representadas por

$$S(t|\mathbf{x}) = \prod_{u=0}^{t}\left[\frac{\frac{1}{1+(\frac{u+1}{\alpha})^{\gamma}}}{\frac{1}{1+(\frac{u+1}{\alpha})^{\gamma}} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(\frac{1}{1+(\frac{u}{\alpha})^{\gamma}} - \frac{1}{1+(\frac{u+1}{\alpha})^{\gamma}}\right)}\right], \; t = 0, 1, 2, \ldots \tag{41}$$

e

$$p(t|\mathbf{x}) = \frac{e^{\mathbf{x}'\boldsymbol{\beta}}\left(\frac{1}{1+(\frac{t}{\alpha})^{\gamma}} - \frac{1}{1+(\frac{t+1}{\alpha})^{\gamma}}\right)}{\frac{1}{1+(\frac{t+1}{\alpha})^{\gamma}}} \prod_{u=0}^{t}\left[\frac{\frac{1}{1+(\frac{u+1}{\alpha})^{\gamma}}}{\frac{1}{1+(\frac{u+1}{\alpha})^{\gamma}} + e^{\mathbf{x}'\boldsymbol{\beta}}\left(\frac{1}{1+(\frac{u}{\alpha})^{\gamma}} - \frac{1}{1+(\frac{u+1}{\alpha})^{\gamma}}\right)}\right], \; t = 0, 1, 2, \ldots . \tag{42}$$

Conhecendo $h(t|\mathbf{x})$, $S(t|\mathbf{x})$ e $p(t|\mathbf{x})$, é fácil obter a função de verossimilhança. Essa função ao ser maximizada gera as estimativas dos parâmetros para construção do modelo de regressão *odds*-riscos proporcionais.

4. Aplicação em dados reais

4.1. Banco de dados

A aplicação do modelo de regressão *odds*-riscos proporcionais foi realizada no banco de dados de um estudo composto por 150 pacientes com dor lombar divididos em dois grupos: tratamento com corrente interferencial ativa e tratamento com corrente interferencial placebo (Corrêa et al., 2016).

A base de dados era composta por homens e mulheres com idades entre 18 e 80 anos que apresentavam dor lombar sem causa determinada há pelo menos 3 meses. Essas pessoas classificaram a intensidade da dor nos últimos 7 dias em uma escala de 0 (sem dor) a 10 (dor muito forte) e apenas as com classificação maior ou igual a 3 compunham a pesquisa. Não fizeram parte do estudo indivíduos com doenças graves de coluna, doenças neurológicas e/ou cardiorrespiratórias graves, gravidez, infecção no local da dor, câncer, marca-passo cardíaco, lesões de pele ou alergia no local da dor e alterações de sensibilidade (Silva et al., 2017).

Os pacientes que participaram do estudo foram aleatoriamente divididos em dois grupos de tratamento: ativo e placebo. O equipamento utilizado para tratamento da dor lombar, um gerador de correntes alternadas de média frequência via eletrodos (Neurivector), foi utilizado por 30 minutos nos pacientes. Esse procedimento foi repetido em 12 sessões que aconteceram 3 vezes por semana em dias alternados, totalizando 4 semanas de tratamento. De 5 em 5 minutos, a amplitude da corrente era aumentada se houvesse a diminuição da sensação da corrente. Os pacientes que ficaram no grupo placebo passavam pelos mesmos procedimentos, mas não tiveram o aumento da corrente. O alívio ou diminuição da dor lombar foi considerado como a redução da dor em 50% ou mais da escala de dor relatada no início do tratamento (Silva et al., 2017).

A análise dos dados foi feita utilizando técnicas de Análise de Sobrevivência, consi-

derando como evento de interesse o alívio/diminuição da dor lombar. O objeto de estudo foi definido, portanto, como o número de sessões malsucedidas antes da sessão que aliviou ou diminuiu a dor lombar. Neste caso, para $t = 0$, o paciente teria o alívio da dor logo na primeira sessão. As observações eram consideradas censuradas quando o acompanhamento do paciente sofria interrupção por algum motivo estranho ao evento de interesse do estudo ou após a realização de 11 sessões malsucedidas. Tratava-se de uma variável de interesse discreta com censura à direita, fato esse que possibilitou a aplicação do modelo de regressão *odds*-riscos proporcionais.

4.2. Análise descritiva

Na Análise de Sobrevivência é de grande importância conhecer o comportamento da variável a ser explicada e de suas covariáveis. Isso é feito através da análise descritiva dos dados, inicialmente verificando o comportamento das curvas de sobrevivência obtidas pelo estimador de Kaplan-Meier.

Esse estudo foi executado para estudar o efeito das correntes alternadas de média frequência no alívio ou diminuição da dor lombar, então o principal objetivo da análise era comparar o grupo de tratamento e o grupo placebo. Por esse motivo, a análise gráfica ilustrada na Figura 4 apresenta as curvas de sobrevivência para cada grupo de tratamento. Isso é importante para definição de possíveis distribuições que ajustem um modelo de forma correta.

A Figura 4 mostra que a função de sobrevivência tem comportamento decrescente, ou seja, quanto mais sessões o paciente se submeter, menor será a probabilidade de permanecer com dor lombar. Percebe-se que a probabilidade de permanecer com a dor é maior para os indivíduos do grupo placebo do que para os do grupo ativo em todos as situações de número de sessões malsucedidas. Portanto, o grupo ativo precisa de menos sessões para o alívio/diminuição da dor do que o grupo placebo.

Para o grupo ativo, 50% dos pacientes têm o alívio/diminuição da dor na primeira sessão, ou seja, não têm nenhuma sessão malsucedida. Já para o grupo placebo, o número de sessões malsucedidas para o qual 50% dos pacientes têm o alívio/diminuição da dor é igual a 1.

Isso significa que metade dos indivíduos precisam de 2 sessões para a ocorrência do evento de interesse.

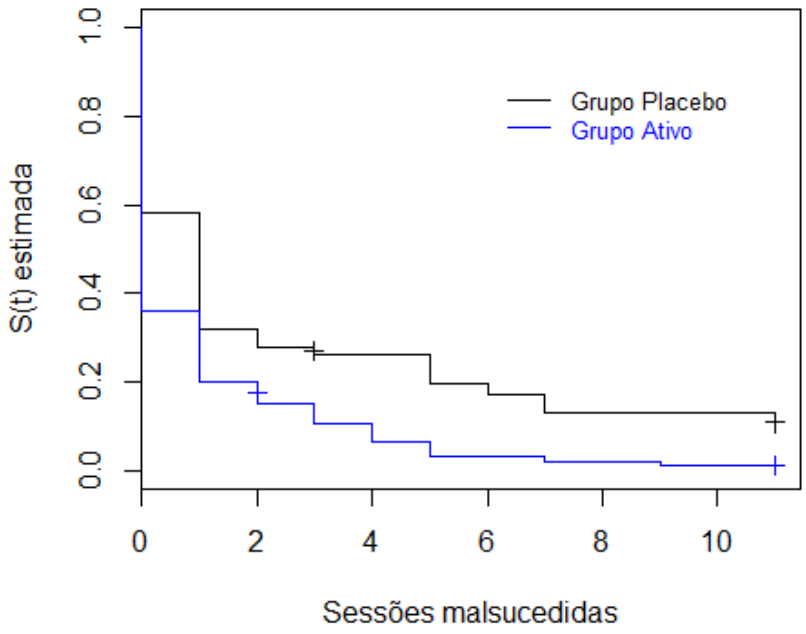

Figura 4: Função de sobrevivência estimada por Kaplan-Meier para dados de pacientes com dor lombar segundo grupo de tratamento.

Na situação em que não é realizada nenhuma sessão malsucedida, para ambos os grupos a probabilidade não é igual a 1, isso significa que alguns indivíduos falharam (tiveram alívio/diminuição da dor) na primeira sessão do tratamento. É importante ressaltar que apenas 9 indivíduos sofreram censura e que elas aconteceram principalmente na 11ª sessão malsucedida.

O grupo de tratamento é apenas uma das covariáveis utilizadas para explicar o número de sessões malsucedidas. As variáveis idade, índice de massa corporal (IMC) e tempo de dor eram originalmente quantitativas e passaram por categorização. Os pacientes foram divididos em dois grupos de idade, um para indivíduos com até 50 anos e outro com 50 anos ou mais; em dois grupos de IMC, não obesos (IMC menor que 30) e obesos (IMC maior ou igual a 30); em dois grupos de tempo de dor, um com menos de 5 anos de dor e outro com 5 anos ou mais de dor. Na Tabela 1 foram dispostas todas as covariáveis que utilizadas para a construção dos modelos de regressão *odds*-riscos proporcionais com suas respectivas frequências e percentuais.

Tabela 1: Frequências e percentuais das covariáveis do estudo.

Covariáveis	Categorias	Frequência	%
Grupo	Placebo	50	33,3
	Ativo	100	66,7
Sexo	Masculino	35	23,3
	Feminino	115	76,7
Idade	Até 50 anos	66	44
	50 anos ou mais	84	56
IMC	Até 30	108	72
	30 ou mais	42	28
Tempo de dor	Menos de 5 anos	57	38
	5 anos ou mais	93	62
Uso de medicamentos	Sim	115	76,7
	Não	35	23,3

A maior parte dos pacientes são do sexo feminino ($76,7\%$), têm 50 anos ou mais (56%), não são obesos (72%), apresentam dor lombar a mais de 5 anos (62%) e usam algum tipo de medicamento ($76,7\%$). A variável resposta, número de sessões malsucedidas antes da sessão que aliviou/diminuiu a dor lombar, já foi analisada segundo o grupo de tratamento na Figura 4.

A relação entre a variável resposta e as outras covariáveis é ilustrada na Figura 5. Os gráficos apresentam a função de sobrevivência estimada pela metodologia de Kaplan-Meier para cada tempo discreto. Com relação ao sexo, percebe-se que a probabilidade de permanecer com dor lombar é maior para mulheres do que para homens com 0 a 5 sessões malsucedidas. Já para variável idade, em qualquer situação os pacientes com até 50 anos têm probabilidade maior de permanecer com dor do que aqueles com mais de 50 anos. A probabilidade de alívio/diminuição da dor lombar em pacientes não obesos é maior do que a de pacientes obesos quando são realizadas de 1 a 4 sessões malsucedidas antes da sessão que aliviou ou diminuiu a dor. Os indivíduos que sentem dor a menos de 5 anos têm probabilidade menor de permanecer com dor do que aqueles que têm dor a 5 anos ou mais. Por outro lado, quem usa algum tipo de medicamento tem probabilidade menor de alívio/diminuição da dor lombar do que os pacientes que não tomam medicamento algum.

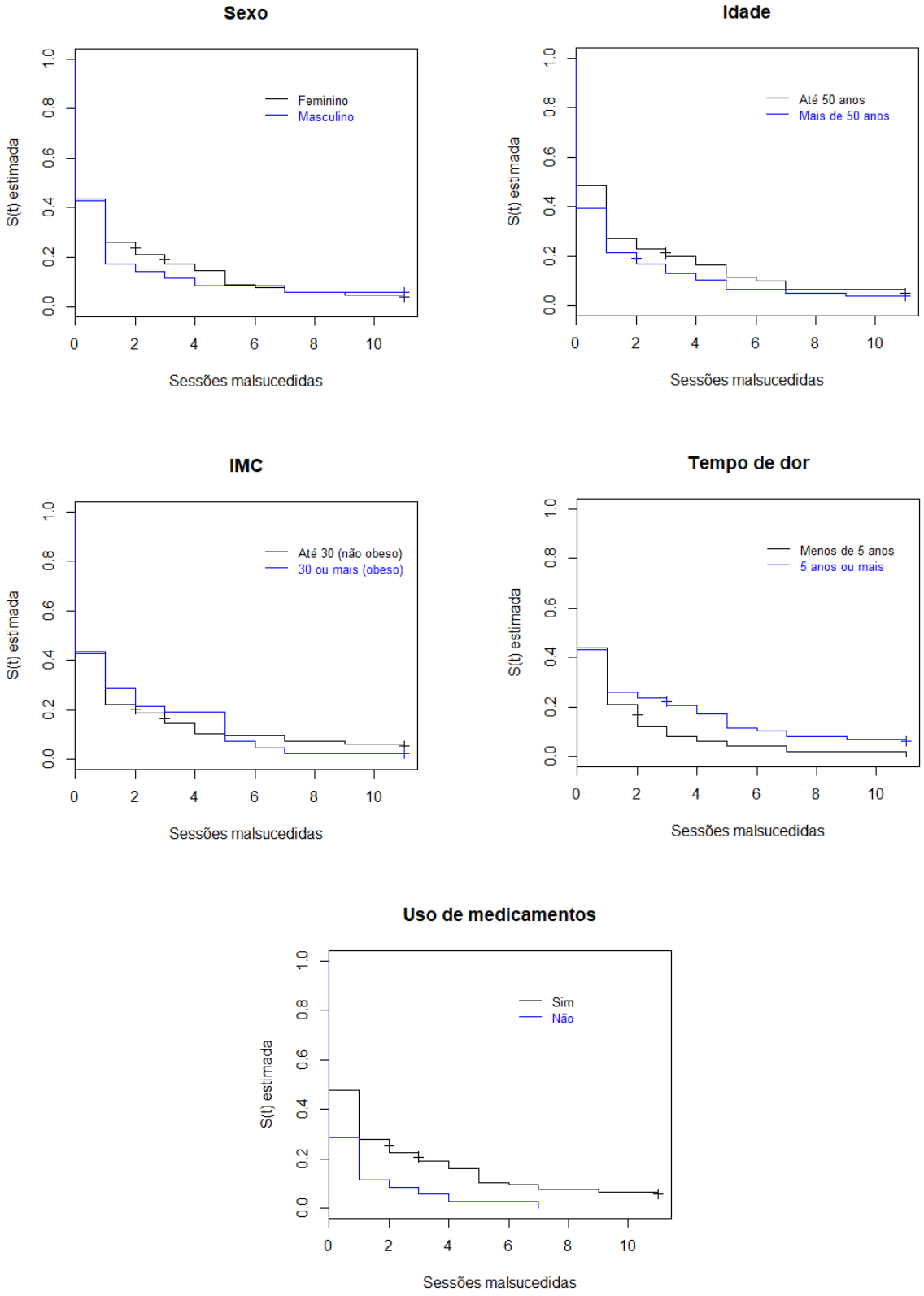

Figura 5: Função de sobrevivência estimada por Kaplan-Meier para dados de pacientes com dor lombar segundo as covariáveis sexo, idade, IMC, tempo de dor e uso de medicamentos.

4.3. Verificação da suposição de *odds*-riscos proporcionais

Para usar o modelo de regressão *odds*-riscos proporcionais, é adequado que a suposição de *odds*-riscos proporcionais seja atendida. A suposição foi verificada para as covariáveis da base de dados através do método gráfico proposto na seção 3.2. As Figuras 6 e 7 foram utilizadas como base para analisar a existência (ou inexistência) de proporcionalidade entre as *odds*-riscos de cada covariável. Para que a suposição fosse válida, os gráficos da Figura 6 deveriam apresentar curvas com diferenças verticais aproximadamente constantes no tempo e, em cada gráfico da Figura 7, os pontos deveriam estar próximos a uma reta com inclinação 1.

Analisando os gráficos da Figura 6, é visível que as covariáveis de tratamento e uso de medicamentos têm *odds* dos riscos aproximadamente proporcionais (curvas aproximadamente paralelas). As curvas das covariáveis idade e tempo de dor têm comportamento paralelo na maioria dos pontos, exceto no ponto final, para variável idade, e no ponto inicial, para variável tempo de dor. Com relação às variáveis sexo e IMC, percebe-se que as curvas se cruzam em algum momento no tempo, mais expressivamente para o IMC do que para sexo.

Para gerar os gráficos da Figura 7, foram plotados os dados com relação a $log[G_0(t)]$ e $log[G_1(t)]$, ajustando uma reta de regressão. Os gráficos mostram que, para as variáveis tratamento e uso de medicamentos, a reta ajustada fica bem próxima aos pontos. Também se percebe que, para idade e tempo de dor, apesar de haver alguns desvios, a reta é bem ajustada aos pontos. Por outro lado, os gráficos das covariáveis sexo e IMC mostram que a distância entre a reta ajustada e os pontos é maior do que a situação observada para outras covariáveis.

Analisando conjuntamente os gráficos presentes nas duas figuras, percebe-se que ambas levam a mesma conclusão sobre a existência de proporcionalidade entre a *odds* dos riscos das categorias de cada covariável. Esses gráficos podem, em um primeiro momento, sugerir indícios de que as variáveis sexo e IMC estariam violando a suposição de *odds*-riscos proporcionais. No entanto, como pôde ser visto pela Figura 5 e também pelas Tabelas 3 a 5, essas covariáveis não apresentam efeito significativo. Desta forma, é possível concluir que os supostos desvios observados nas Figuras 6 e 7 para as variáveis sexo e IMC não são significativos.

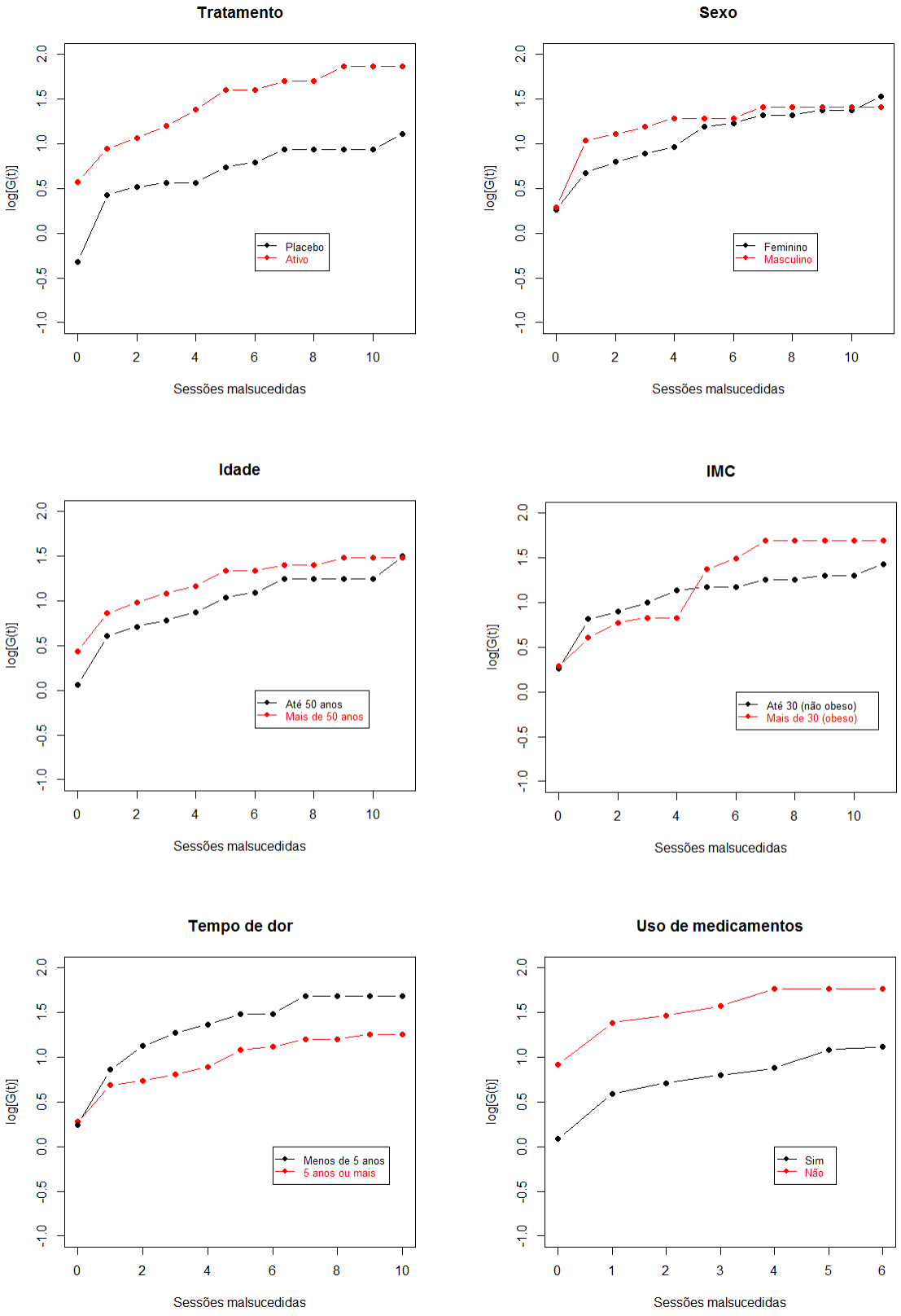

Figura 6: Logaritmo da função *odds*-risco acumulado para dados de pacientes com dor lombar segundo grupo de tratamento, sexo, idade, IMC, tempo de dor e uso de medicamentos.

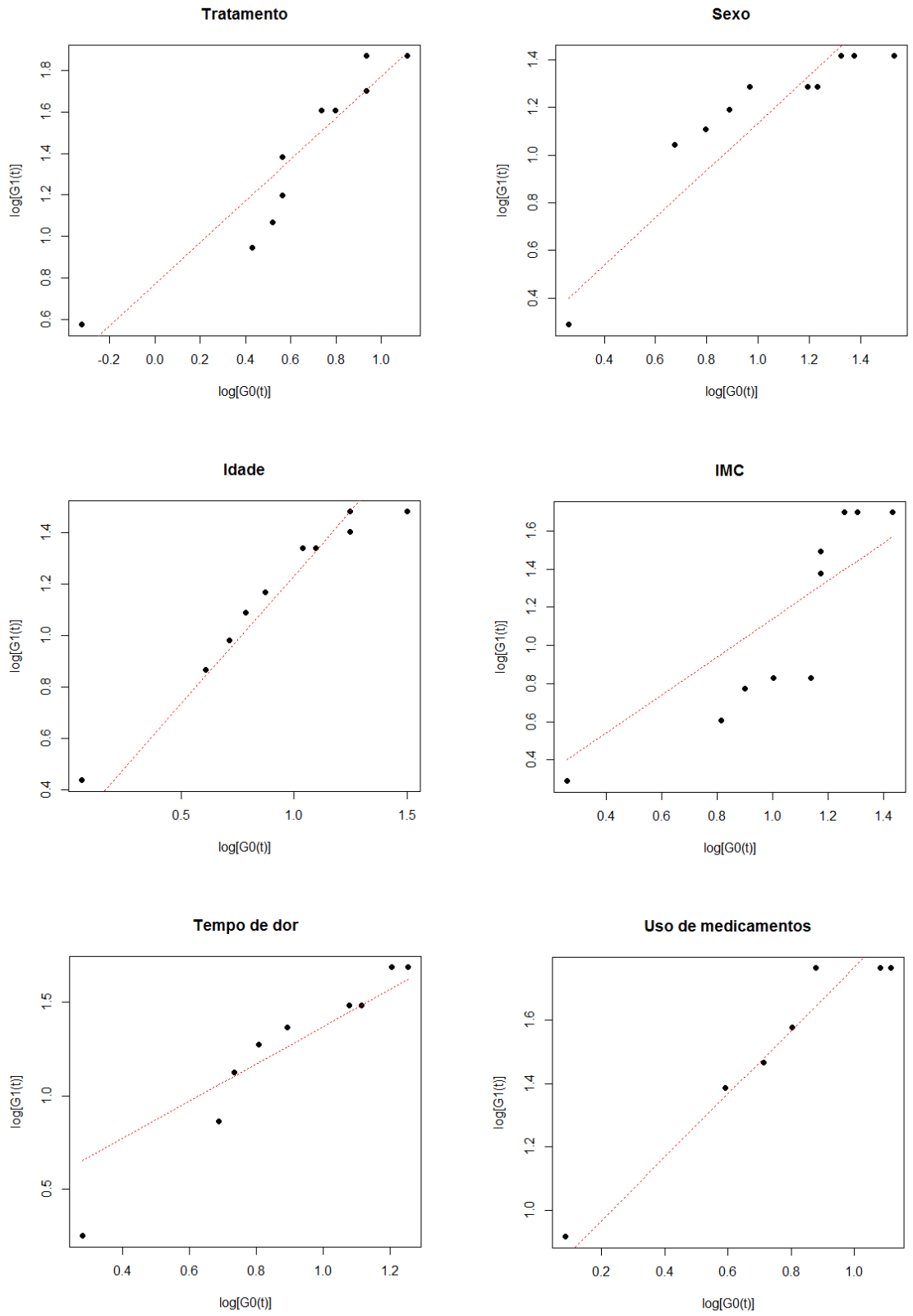

Figura 7: $\log[G_0(t)]$ *versus* $\log[G_1(t)]$ para dados de pacientes com dor lombar segundo grupo de tratamento, sexo, idade, IMC, tempo de dor e uso de medicamentos.

4.4. Modelo de regressão *odds*-riscos proporcionais

Utilizando os dados dos pacientes com dor lombar, foi aplicada a metodologia proposta para ajustar o modelo de regressão *odds*-riscos proporcionais. Ao fazer a análise exploratória dos dados, observou-se que a função de sobrevivência possuía comportamento decrescente. Como era de interesse testar o comportamento do modelo para mais de uma distribuição, toda a análise foi realizada para as distribuições geométrica, Weibull discreta e log-logística discreta.

Inicialmente, três modelos foram ajustados com base nas três distribuições para explicar o número de sessões malsucedidas antes da sessão que aliviou ou diminuiu a dor lombar a partir da covariável de tratamento. Definida a função de verossimilhança com base nas expressões (30), (37), (38), (40) e (41), possibilitou-se a estimação dos parâmetros da distribuição e do modelo através da maximização da função de verossimilhança. Com as estimativas dos parâmetros, foram obtidos valores das funções de risco, sobrevivência e probabilidade.

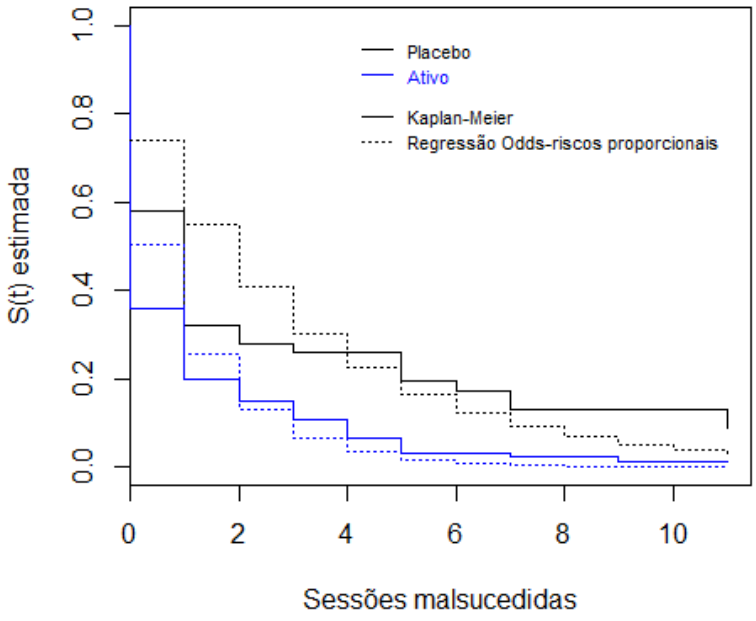

Figura 8: Função de sobrevivência estimada pelo modelo de regressão *odds*-riscos proporcionais geométrico segundo grupo de tratamento.

Nas primeiras linhas das Tabelas 3, 4 e 5, é possível visualizar alguns resultados importantes acerca do ajuste do modelo de regressão *odds*-riscos proporcionais. Percebe-se que para todas as distribuições estudadas (geométrica, Weibull discreta e log-logística discreta), a variável tratamento é significativa (p-valor $< 0,1$) para explicação do número de sessões malsucedidas. Isso se confirma ao analisar os gráficos que fazem um comparativo das curvas de sobrevivência estimadas por Kaplan-Meier e pela regressão *odds*-riscos proporcionais.

A Figura 8 mostra as curvas de sobrevivência ajustadas pelo modelo de regressão *odds*-riscos proporcionais geométrico. Apesar do coeficiente relacionado à covariável de tratamento ser significativo, nota-se que essas curvas não estão muito próximas das curvas estimadas por Kaplan-Meier.

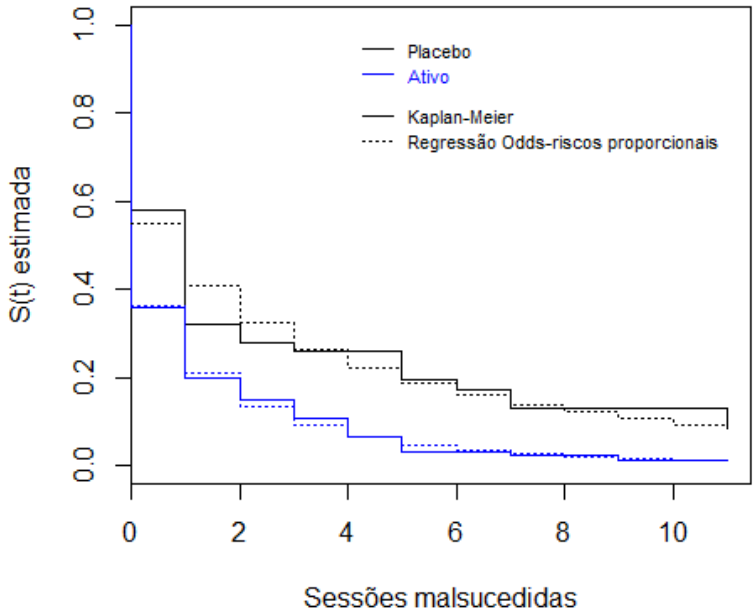

Figura 9: Função de sobrevivência estimada pelo modelo de regressão *odds*-riscos proporcionais Weibull discreta segundo grupo de tratamento.

O ajuste melhora bastante ao analisar o modelo de regressão *odds*-riscos proporcionais Weibull discreta. De acordo com a Figura 9, observa-se que esse ajuste é melhor ao ser comparado com o modelo geométrico. As linhas pontilhadas muitas vezes se confundem com

as linhas contínuas em ambos os grupos de tratamento. O modelo de regressão *odds*-riscos proporcionais log-logística discreta também se mostra adequado quando comparado às outras duas distribuições. Na Figura 10 é possível verificar que as curvas são bem parecidas com as do modelo da Weibull discreta. Portanto, para Weibull discreta e log-logística discreta, percebe-se que o ajuste do modelo é suficientemente adequado.

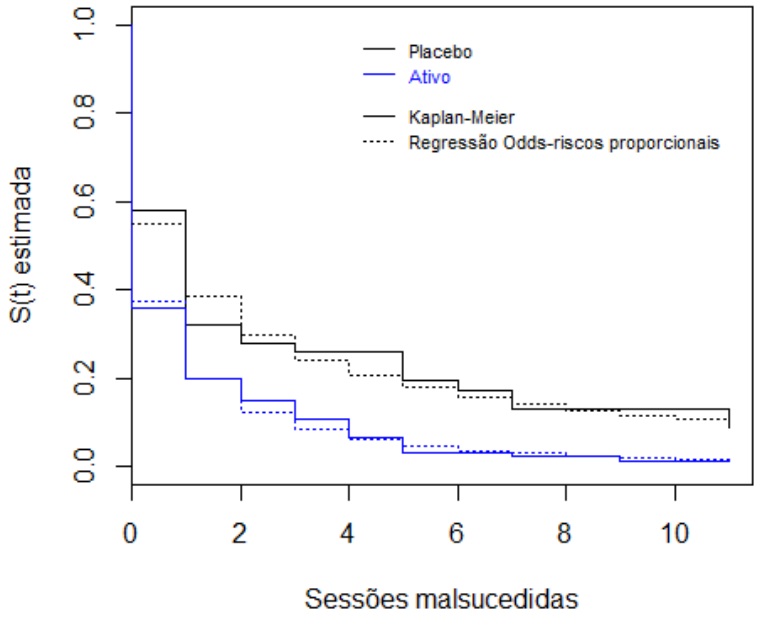

Figura 10: Função de sobrevivência estimada pelo modelo de regressão *odds*-riscos proporcionais log-logística discreta segundo grupo de tratamento.

Visualmente, não é possível dizer qual o modelo mais adequado para a situação em que o número de sessões malsucedidas é explicado apenas pelo grupo de tratamento. Pensando em uma métrica simples, apenas para ter uma ideia inicial da distribuição que melhor se ajusta aos dados, na Tabela 2 é disponibilizado uma espécie de erro quadrático (EQ). Esse erro é calculado para cada distribuição da seguinte forma

$$EQ_D = \sum_{j=1}^{k} \left\{ \sum_{i=1}^{m} \left(S_{KM_j}(t_i) - S_D(t_i|x=j) \right)^2 \right\}, \tag{43}$$

em que t_i é o i-ésimo tempo de sobrevivência distinto observado na amostra, $i = 1, 2, ..., m$ ($m \leq n$) e $j = 1, 2, ..., k$ são as categorias da covariável x.

Para a expressão acima, considera-se $S_{KM_j}(t_i)$ a sobrevivência estimada por Kaplan-Meier no tempo t_i para o grupo j e $S_D(t_i|x = j)$ a sobrevivência estimada pelo modelo de regressão *odds*-riscos proporcionais da distribuição D no tempo t_i para o grupo j.

Nota 4. Para covariáveis numéricas, o EM_D pode ser calculado categorizando os valores das covariáveis.

A partir dos valores do erro médio de ajuste dos modelos, chega-se à conclusão que, segundo essa medida, o ajuste do modelo de regressão *odds*-riscos proporcionais log-logística discreta é o mais adequado quando utiliza-se apenas a covariável de tratamento.

Tabela 2: Erro médio da sobrevivência estimada pelo modelo de regressão *odds*-riscos proporcionais para distribuições geométrica, Weibull discreta e log-logística discreta

Distribuição	Erro médio
Geométrica	0,154
Weibull discreta	0,016
Log-logística discreta	0,012

Além de construir o modelo simples para a variável tratamento, o procedimento foi repetido para as demais covariáveis com o objetivo de entender a significância dessas variáveis na explicação do número de sessões malsucedidas. De acordo com as Tabelas 3, 4 e 5, as variáveis tratamento, tempo de dor e uso de medicamentos são significativas (p-valor $< 0,1$) na explicação da variável resposta para os modelos das três distribuições. As demais covariáveis, sexo, idade e IMC, não apresentaram valores significativos, porém, foram incluídas no modelo completo, já que se acredita que essas variáveis têm influência no número de sessões malsucedidas antes da sessão que aliviou/diminuiu a dor.

Tabela 3: Estimativas dos parâmetros do modelo de regressão simples *odds*-riscos proporcionais geométrico.

Variável *	Parâmetro	Estimativa	EP	IC (95%)	*p*-valor
Tratamento	p	0,259	0,033	(0,194; 0,324)	-
	β_1	1,033	0,224	(0,594; 1,472)	< 0,001
Sexo	p	0,380	0,028	(0,323; 0,435)	-
	β_1	0,097	0,256	(-0,404; 0,598)	0,704
Idade	p	0,346	0,035	(0,277; 0,415)	-
	β_1	0,318	0,214	(-0,102; 0,737)	0,138
IMC	p	0,384	0,030	(0,325; 0,442)	-
	β_1	0,014	0,236	(-0,448; 0,477)	0,952
Tempo de dor	p	0,339	0,030	(0,281; 0,396)	-
	β_1	0,618	0,228	(0,170; 1,065)	0,007
Uso de medicamentos	p	0,343	0,027	(0,290; 0,395)	-
	β_1	1,115	0,297	(0,533; 1,697)	< 0,001

* Nível de referência das variáveis:
Tratamento = Grupo placebo; Sexo = Feminino; Idade = Até 50 anos;
IMC = Até 30; Tempo de dor = 5 anos ou mais; Uso de medicamentos = Sim.

Tabela 4: Estimativas dos parâmetros do modelo de regressão simples *odds*-riscos proporcionais Weibull discreta.

Variável *	Parâmetro	Estimativa	EP	IC (95%)	*p*-valor
Tratamento	q	0,548	0,056	(0,438; 0,658)	-
	γ	0,572	0,064	(0,446; 0,698)	-
	β_1	0,753	0,237	(0,289; 1,217)	0,001
Sexo	q	0,426	0,042	(0,344; 0,509)	-
	γ	0,560	0,058	(0,446; 0,674)	-
	β_1	0,072	0,273	(-0,463; 0,606)	0,792
Idade	q	0,453	0,051	(0,353; 0,552)	-
	γ	0,557	0,059	(0,441; 0,672)	-
	β_1	0,221	0,228	(-0,225; 0,668)	0,331
IMC	q	0,423	0,043	(0,339; 0,507)	-
	γ	0,560	0,058	(0,446; 0,674)	-
	β_1	0,012	0,251	(-0,479; 0,504)	0,960
Tempo de dor	q	0,462	0,046	(0,371; 0,552)	-
	γ	0,563	0,060	(0,446; 0,681)	-
	β_1	0,414	0,242	(-0,061; 0,888)	0,087
Uso de medicamentos	q	0,467	0,043	(0,382; 0,552)	-
	γ	0,575	0,061	(0,455; 0,694)	-
	β_1	0,818	0,312	(0,206; 1,431)	0,009

* Nível de referência das variáveis:
Tratamento = Grupo placebo; Sexo = Feminino; Idade = Até 50 anos;
IMC = Até 30; Tempo de dor = 5 anos ou mais; Uso de medicamentos = Sim.

Tabela 5: Estimativas dos parâmetros do modelo de regressão simples *odds*-riscos proporcionais log-logística discreta.

Variável *	Parâmetro	Estimativa	EP	IC (95%)	*p*-valor
Tratamento	α	1,233	0,292	(0,660; 1,805)	-
	γ	0,969	0,128	(0,718; 1,219)	-
	β_1	0,722	0,239	(0,253; 1,192)	0,003
Sexo	α	0,798	0,129	(0,545; 1,051)	-
	γ	1,124	0,136	(0,858; 1,391)	-
	β_1	0,082	0,274	(-0,456; 0,619)	0,766
Idade	α	0,868	0,165	(0,543; 1,192)	-
	γ	1,075	0,141	(0,799; 1.352)	-
	β_1	0,219	0,228	(-0,228; 0,666)	0,338
IMC	α	0,786	0,127	(0,537; 1,036)	-
	γ	1,134	0,141	(0,858; 1,410)	-
	β_1	0,001	0,251	(-0,491; 0,494)	0,997
Tempo de dor	α	0,885	0,156	(0,579; 1,190)	-
	γ	1,076	0,133	(0,815; 1,338)	-
	β_1	0,371	0,244	(-0,108; 0,850)	0,129
Uso de medicamentos	α	0,914	0,150	(0,620; 1,208)	-
	γ	1,095	0,131	(0,838; 1,352)	-
	β_1	0,792	0,313	(0,179; 1,404)	0,011

* Nível de referência das variáveis:
Tratamento = Grupo placebo; Sexo = Feminino; Idade = Até 50 anos;
IMC = Até 30; Tempo de dor = 5 anos ou mais; Uso de medicamentos = Sim.

Após estudar a significância de cada variável na explicação da variável de interesse, o modelo de regressão completo *odds*-riscos proporcionais foi ajustado para cada distribuição. Resultaram, portanto, as Tabelas 6, 7 e 8 com estimativas, erro padrão, intervalo de confiança e p-valor dos parâmetros da distribuição e do modelo. O mesmo comportamento observado ao ajustar o modelo simples foi observado ao ajustar o modelo completo. As variáveis tratamento, tempo de dor e uso de medicamentos foram consideradas significativas para o modelo, ou seja, conjuntamente essas variáveis são importantes para explicar a variável de interesse.

Assim como no estudo de Silva et al. (2017), as covariáveis sexo, idade e IMC, mesmo não sendo significativas, foram mantidas no modelo por se acreditar que, na prática, essas características podem influenciar no resultado do tratamento.

Tabela 6: Estimativas dos parâmetros do modelo de regressão completo *odds*-riscos proporcionais geométrico.

Parâmetro	Estimativa	Erro padrão	IC (95%)	*p*-valor
q	0,180	0,038	(0,105; 0,255)	-
β_1 (Tratamento)	0,961	0,237	(0,496; 1,426)	< 0,001
β_2 (Sexo)	-0,031	0,278	(-0,576; 0,515)	0,912
β_3 (Idade)	0,254	0,241	(-0,218; 0,727)	0,291
β_4 (IMC)	0,115	0,254	(-0,383; 0,614)	0,650
β_5 (Tempo de dor)	0,633	0,255	(0,134; 1,132)	0,013
β_6 (Uso de medicamentos)	0,878	0,320	(0,252; 1,505)	0,006

Nota: Nível de referência das variáveis:
Tratamento = Grupo placebo; Sexo = Feminino; Idade = Até 50 anos;
IMC = Até 30; Tempo de dor = 5 anos ou mais; Uso de medicamentos = Sim.

Tabela 7: Estimativas dos parâmetros do modelo de regressão completo *odds*-riscos proporcionais Weibull discreta.

Parâmetro	Estimativa	Erro padrão	IC (95%)	*p*-valor
q	0,651	0,071	(0,512; 0,790)	-
γ	0,593	0,071	(0,455; 0,732)	-
β_1 (Tratamento)	0,721	0,247	(0,237; 1,206)	0,004
β_2 (Sexo)	-0,014	0,288	(-0,579; 0,550)	0,960
β_3 (Idade)	0,190	0,249	(-0,297; 0,678)	0,444
β_4 (IMC)	0,092	0,263	(-0,424; 0,607)	0,728
β_5 (Tempo de dor)	0,449	0,263	(-0,067; 0,964)	0,088
β_6 (Uso de medicamentos)	0,684	0,328	(0,041; 1,326)	0,037

Nota: Nível de referência das variáveis:
Tratamento = Grupo placebo; Sexo = Feminino; Idade = Até 50 anos;
IMC = Até 30; Tempo de dor = 5 anos ou mais; Uso de medicamentos = Sim.

Tabela 8: Estimativas dos parâmetros do modelo de regressão completo *odds*-riscos proporcionais log-logística discreta.

Parâmetro	Estimativa	Erro padrão	IC (95%)	*p*-valor
α	2,048	0,804	(0,472; 3,625)	-
γ	0,879	0,127	(0,631; 1,128)	-
β_1 (Tratamento)	0,696	0,250	(0,207; 1,186)	0,005
β_2 (Sexo)	0,004	0,288	(-0,560; 0,568)	0,989
β_3 (Idade)	0,197	0,249	(-0,290; 0,685)	0,427
β_4 (IMC)	0,089	0,264	(-0,428; 0,606)	0,736
β_5 (Tempo de dor)	0,424	0,265	(-0,095; 0,944)	0,109
β_6 (Uso de medicamentos)	0,673	0,327	(0,032; 1,315)	0,040

Nota: Nível de referência das variáveis:
Tratamento = Grupo placebo; Sexo = Feminino; Idade = Até 50 anos;
IMC = Até 30; Tempo de dor = 5 anos ou mais; Uso de medicamentos = Sim.

Na seção 3.1 foi explanado que os coeficientes estimados poderiam ser interpretados. Dessa forma, ajustando o modelo de regressão *odds*-riscos proporcionais e encontrando as estimativas dos coeficientes, através da expressão (26), $\exp\{\beta\}$ adquiriu um significado.

Os valores para $\exp\{\beta\}$, ou razão de chances, foram organizados na Tabela 9. Assim, lembrando que a falha é considerada o alívio ou diminuição da dor lombar, para as covariáveis significativas, tem-se que para o modelo geométrico:

- A chance de falha dos indivíduos do grupo ativo é 2,61 vezes a chance de falha dos indivíduos do grupo placebo, mantendo-se constantes as demais variáveis. Em outras palavras, a chance de falha dos indivíduos do grupo ativo é 1,61 vezes maior do que a chance de falha dos indivíduos do grupo placebo;

- A chance de falha dos pacientes que sentem dor lombar a menos de 5 anos é 1,88 vezes a chance de falha dos pacientes que sentem dor a 5 anos ou mais. Ou seja, a chance de falha dos indivíduos com dor a menos de 5 anos é 0,88 vezes maior do que a chance de falha dos indivíduos que sentem dor a 5 anos ou mais;

- A chance de falha dos indivíduos que não fazem uso de medicamentos é 2,41 vezes a chance de falha dos indivíduos que usam medicamentos. Então, pode-se afirmar que a chance de falha dos pacientes que não usam medicamentos é 1,41 vezes maior do que a chance de falha dos pacientes que fazem uso de medicamentos.

Tabela 9: Razão de chances do modelo de regressão completo *odds*-riscos proporcionais.

Variável	**Distribuição**		
	Geométrica	**Weibull discreta**	**Log-logística discreta**
Tratamento	2,61 *	2,06 *	2,01 *
Sexo	0,97	0,99	1,00
Idade	1,29	1,21	1,22
IMC	1,12	1,10	1,09
Tempo de dor	1,88 *	1,57 *	1,53 *
Uso de medicamentos	2,41 *	1,98 *	1,96 *

* O resultado é significativo ao nível de significância de 10%.

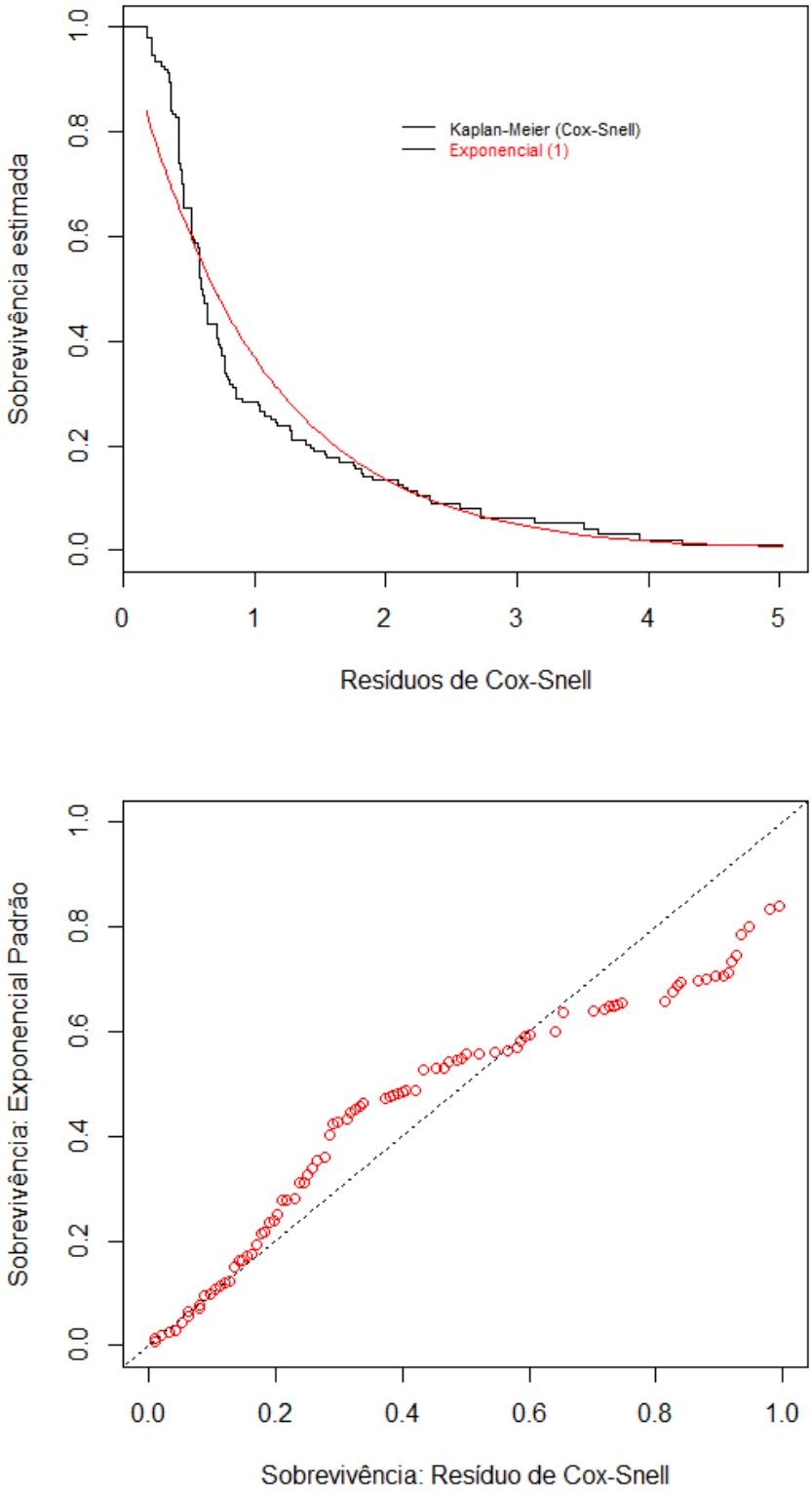

Figura 11: Ajuste do modelo por meio do resíduo de Cox-Snell para o modelo de regressão completo *odds*-riscos proporcionais geométrico.

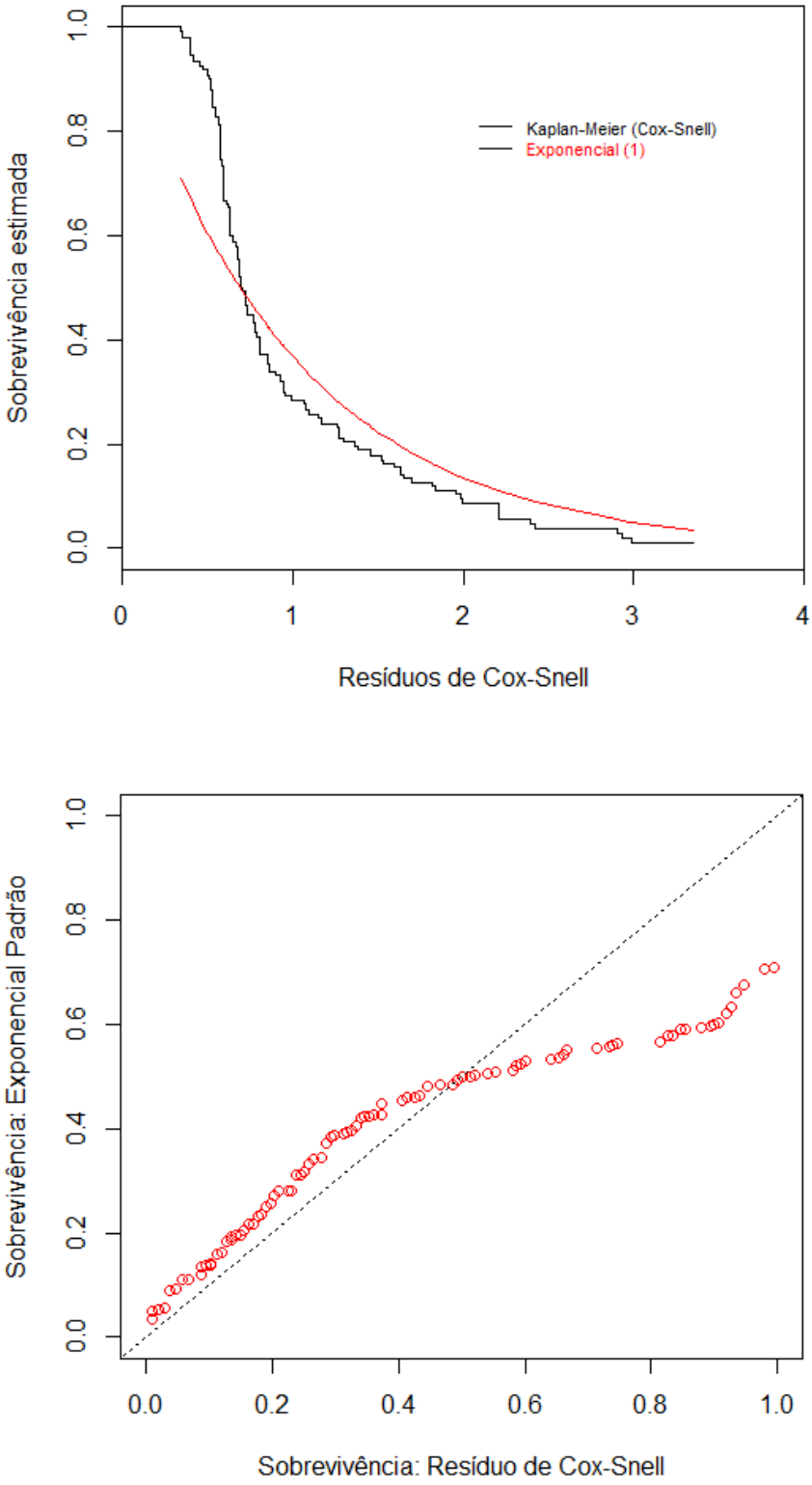

Figura 12: Ajuste do modelo por meio do resíduo de Cox-Snell para o modelo de regressão completo *odds*-riscos proporcionais Weibull discreta.

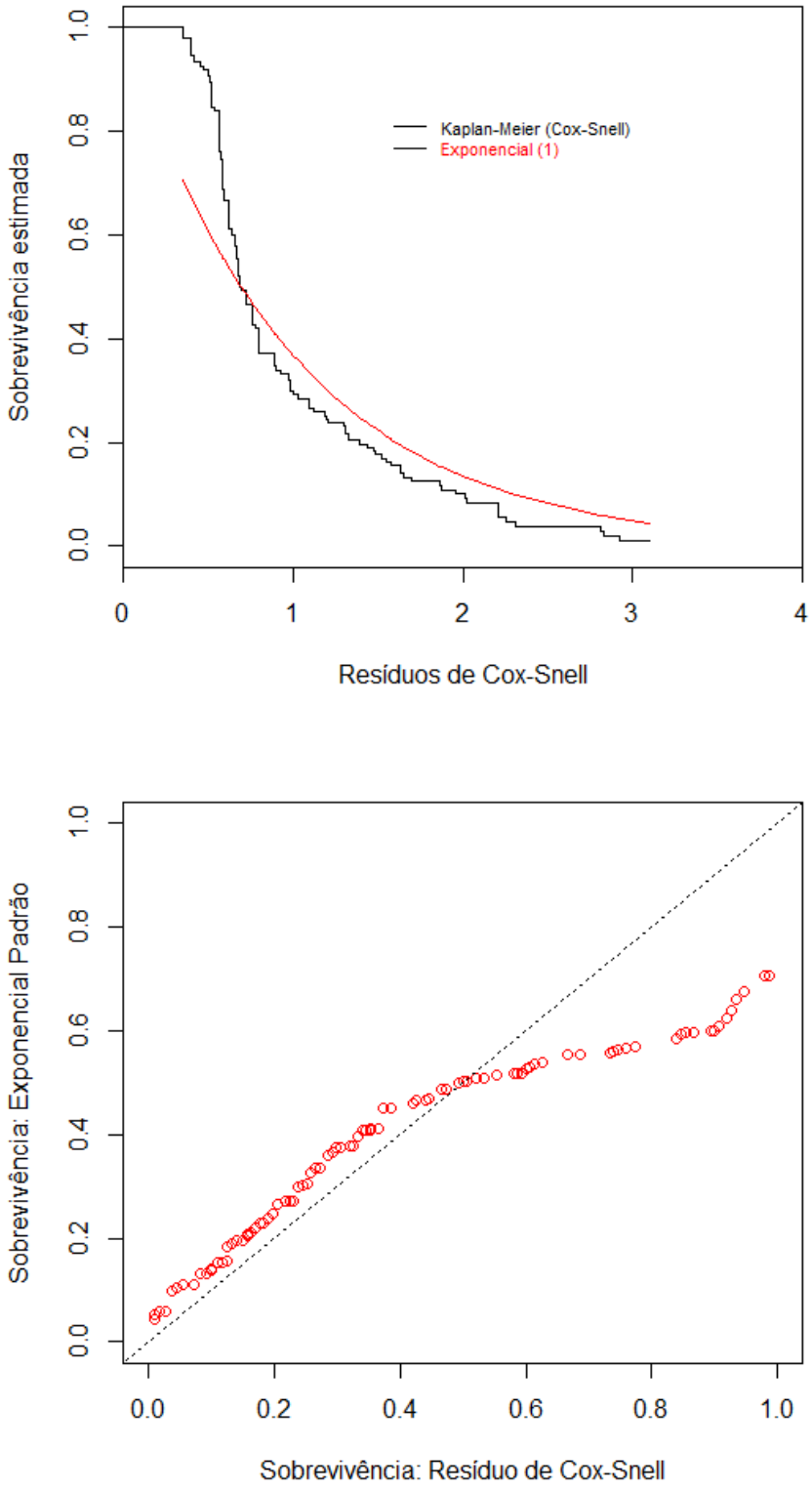

Figura 13: Ajuste do modelo por meio do resíduo de Cox-Snell para o modelo de regressão completo *odds*-riscos proporcionais log-logística discreta.

O último ponto de discussão após o ajuste dos modelos de regressão *odds*-riscos proporcionais é a análise da qualidade desse ajuste. Nas Figuras 11, 12 e 13, estão dispostos dois gráficos para cada modelo de regressão ajustado através das distribuições geométrica, Weibull discreta e log-logística discreta, respectivamente. Os gráficos são construídos com base nos resíduos de Cox-Snell, que por sua vez são definidos como

$$\hat{e}_i = \hat{H}(t_i|\mathbf{x}_i), i = 1, ..., n.$$

No primeiro gráfico de cada figura temos a curva de sobrevivência estimada por Kaplan-Meier dos resíduos de Cox-Snell comparada com a curva de sobrevivência da exponencial padrão. Quanto mais as duas curvas estiverem parecidas, melhor o ajuste do modelo. O segundo gráfico das figuras tem os pontos plotados com relação a sobrevivência estimada na exponencial padrão e a sobrevivência estimada por Kaplan-Meier dos resíduos de Cox-Snell. O ideal é que os pontos estejam próximos de uma reta com inclinação 1. Ambos os gráficos são construídos através de uma comparação entre os resíduos de Cox-Snell e a exponencial padrão.

Analisando conjuntamente os dois gráficos para as três distribuições, percebe-se que a distribuição geométrica é a que melhor se ajustou aos dados. Nota-se também que esses gráficos são similares para os modelos Weibull discreto e log-logístico discreto.

Uma outra forma de verificar o ajuste dos modelos propostos é por meio de uma medida de erro quadrático médio (EQM), que pode ser definida por

$$EQM_D = \frac{\sum_{i=1}^{n}(t_i - \hat{t}_i)^2\delta_i}{\sum_{i=1}^{n}\delta_i},$$

em que t_i é o i-ésimo tempo, δ_i é a variável indicadora de censura para o tempo t_i e $\hat{t}_i$ é o valor predito de t_i que pode ser definido como a esperança da distribuição estimada para o indivíduo i com vetor de covariáveis $\mathbf{x}_i$. De forma alternativa, o EQM_D pode ser calculado definindo o valor preditivo como a mediana da distribuição.

Nota 5: Note que o EQM_D considera somente os valores não censurados.

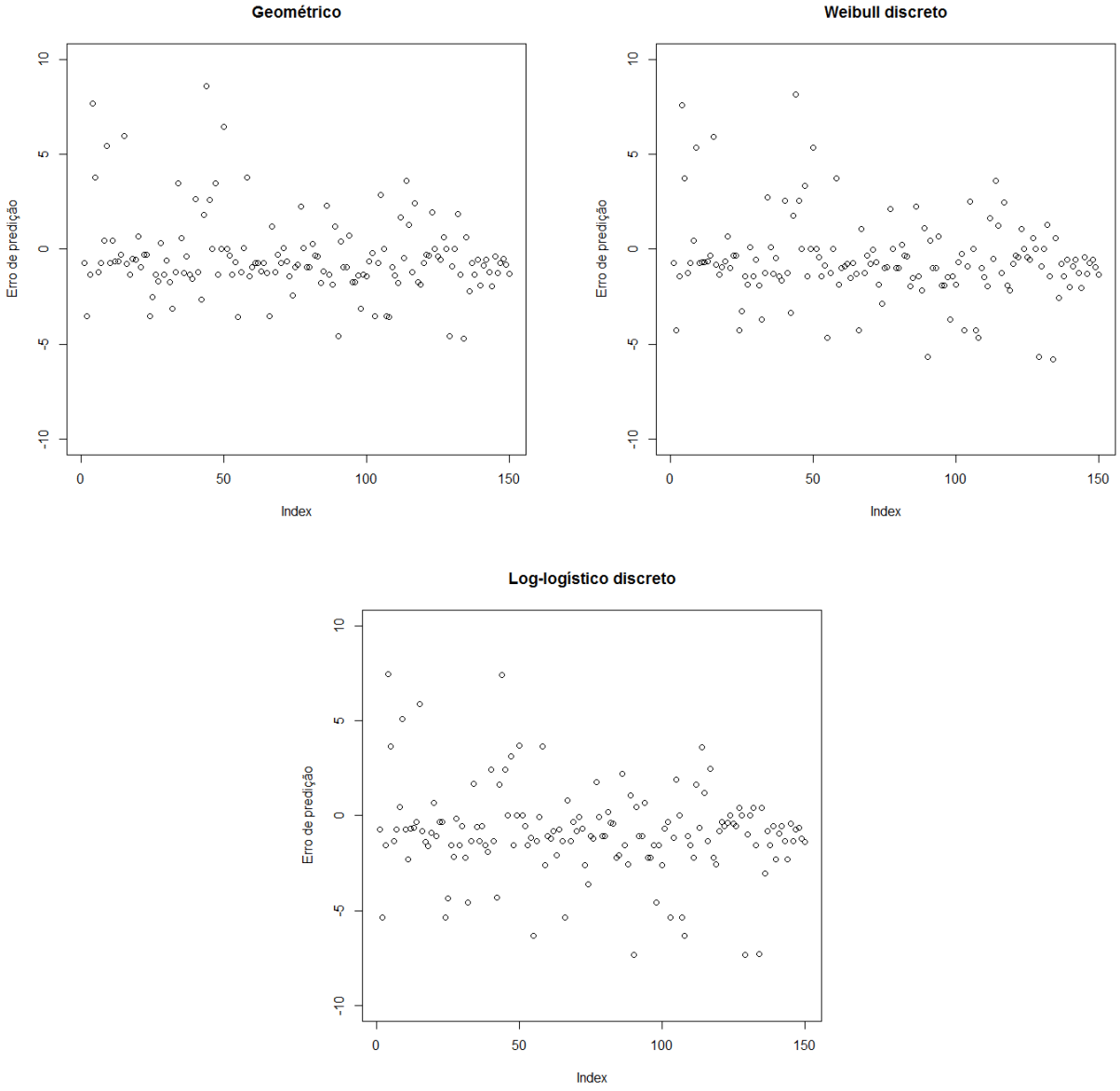

Figura 14: Erro de predição (considerando a esperança como valor preditivo) dos modelos geométrico, Weibull discreto e log-logístico discreto.

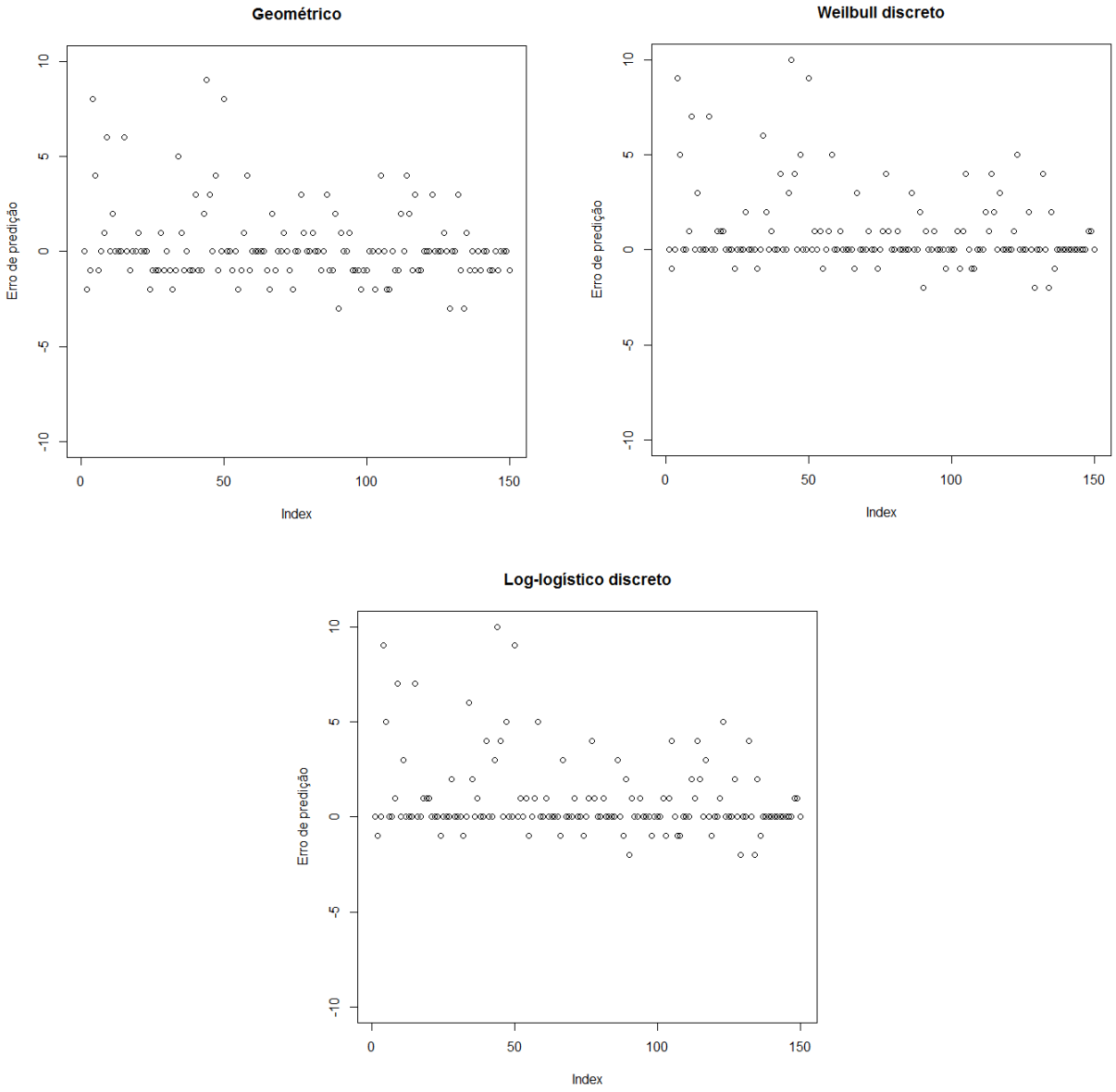

Figura 15: Erro de predição (considerando a mediana como valor preditivo) dos modelos geométrico, Weibull discreto e log-logístico discreto.

A Tabela 10 apresenta os valores do EQM_D para os modelos geométrico, Weibull discreto e log-logístico discreto.

Tabela 10: Erro quadrático médio para os modelo de regressão *odds*-riscos proporcionais para distribuições geométrica, Weibull discreta e log-logística discreta

Distribuição	EQM_D	
	Média	**Mediana**
Geométrica	4,671	4,154
Weibull discreta	5,263	4,958
Log-logística discreta	6,583	4,972

A Figura 14 apresenta os erros de predição, definidos por $t_i - \hat{t}_i$, considerando a média como valor preditivo. Por outro lado, a Figura 15 apresenta esses mesmos erros considerando a mediana como valor preditivo.

Os resultados apresentados pela Tabela 10 e Figuras 14 e 15 indicam que o modelo geométrico foi o que apresentou o melhor ajuste dos dados. Esse resultado obtido pelos $EQM's$ está coerente com àquele encontrado pelo resíduo de Cox-Snell.

5. Considerações finais

Como foi frisado ao longo do texto, na Análise de Sobrevivência, a literatura traz diversas ferramentas para se trabalhar com variáveis contínuas e pouco se fala em adaptações para o uso de variáveis discretas. Por esse motivo, a proposta deste trabalho foi construir uma metodologia de modelo de regressão para o tempo de sobrevivência discreto.

O objetivo proposto foi alcançado com a apresentação do modelo de regressão *odds*-riscos proporcionais. Esse modelo considera que as covariáveis agem multiplicativamente na *odds* (chance) do risco. Assim como é feito no caso contínuo com a suposição de riscos proporcionais no modelo de regressão de Cox, nesse trabalho também foi verificada a suposição de *odds*-riscos proporcionais. Uma adaptação foi realizada para que fosse possível verificar graficamente se as *odds* dos riscos eram proporcionais.

Além de apresentar o modelo de regressão *odds*-riscos proporcionais geral, também foram formulados esses modelos de regressão considerando que o tempo seguia a distribuição geométrica, Weibull discreta ou log-logística discreta. Nesses casos, foram apresentadas as expressões da função de sobrevivência, da função de risco e da função de probabilidade.

Para verificar o funcionamento da metodologia proposta, optou-se por aplicar a técnica em dados reais. A base de dados foi retirada de um estudo com 150 pacientes de 18 a 80 anos que apresentavam dor lombar. Além da variável de interesse, número de sessões malsucedidas antes da sessão que aliviou/diminuiu a dor, também faziam parte do banco de dados as covariáveis, grupo de tratamento, sexo, idade, IMC, tempo de dor e uso de medicamentos.

O modelo de regressão *odds*-riscos proporcionais geométrico, Weibull discreta e log-logística foram ajustados nos dados de dor lombar com todas as covariáveis. Para todos os modelos, somente as covariáveis tratamento, tempo de dor e uso de medicamentos foram significativas. Optou-se por manter sexo, idade e IMC no modelo, pois eram variáveis importantes para explicar o número de sessões malsucedidas.

Também foi verificada a suposição de *odds*-riscos proporcionais para cada uma das covariáveis. Chegou-se à conclusão de que para todas as variáveis era possível considerar as *odds* dos riscos aproximadamente proporcionais.

Por fim, os três modelos foram analisados quanto a qualidade do ajuste. A verificação gráfica mostrou que o ajuste foi consideravelmente bom. Percebeu-se que o modelo de regressão *odds*-riscos proporcionais geométrico foi o que apresentou o melhor ajuste quando comparado com os modelos Weibull discreto e log-logístico discreto.

O resultado ao qual se esperava chegar ao propor um novo modelo de regressão para tempo de sobrevivência discreto foi alcançado. As afirmações feitas nesse trabalho representam uma novidade para a análise de dados discretos e ampliam a literatura sobre o assunto. Ainda existe muito para se discutir no âmbito de variáveis discretas em Análise de Sobrevivência, mas aos poucos mostra-se a importância de criar uma metodologia própria.

Referências bibliográfias

Brunello, G. H. V. e Nakano, E. Y. (2015). "Inferência Bayesiana no modelo Weibull discreto em dados com presença de censura". Em: *TEMA Tend. Mat. Apl. Comput., 16(2):97-110*.

Cardial, M. R. P. (2017). *Distribuição Weibull discreta exponenciada para dados com presença de censura: uma abordagem clássica e bayesiana*. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade de Brasília.

Carrasco, C. G., Tutia, M. H. e Nakano, E. Y. (2012). "Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros". Em: *TEMA Tend. Mat. Apl. Comput., 13(3):247-255*.

Colosimo, E. A. e Giolo, S. R (2006). *Análise de Sobrevivência Aplicada*. Editora Edgard Blucher, São Paulo. ABE - Projeto Fisher.

Corrêa, J. B., Costa, L. O. P., Oliveira, N. T. B., Lima, W. P., Sluka, K. A. e Liebano, R. E. (2016). "Effects of the carrier frequency of interferential current on pain modulation and central hypersensitivity in people with chronic nonspecific low back pain: A randomized placebo-controlled trial". Em: *European Journal of Pain, v.20, p.1653-1666*.

Kaplan, E. L. e Meier, P. (1958). *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, v. 53, p. 457-481.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. 2.ed. New Jersey: John Wiley e Sons.

Magalhães, M. N. (2006). *Probabilidade e variáveis aleatórias*. Edusp, São Paulo.

Nakagawa, T. e Osaki, S. (1975). "The discrete weibull distribution". Em: *IEEE Transactions on Reliability, v.24, p.300-301*.

Nakano, E. Y. (2017). *Um curso de análise de sobrevivência*.

Nakano, E. Y. e Carrasco, C. G. (2006). “Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência”. Em: *TEMA Tend. Mat. Apl. Comput., 7(1):91-100*.

Nobre, L. P. (2017). *Modelo de regressão Weibull para dados discretos em análise de sobrevivência*. Monografia (Graduação em Estatística) - Departamento de Estatística, Universidade de Brasília.

Santos, D. F. (2017). *Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência*. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade de Brasília.

Silva, J. F., Liebano, R. E., Corrêa, J. B., Matsushita, R. Y. e Nakano, E. Y. (2017). “Análise do tempo para o alívio da intensidade da dor em pacientes com dor lombar crônica não específica via modelo de riscos proporcionais de Cox.” Em: *Ciência e Natura, v.39, p.233-243.*

Vieira, I. C. F. e Nakano, E. Y (2017). *Distribuição log-logística discreta para dados de sobrevivência*. Congresso de Iniciação Científica da Unb e Congresso de Iniciação Científica do DF. Anais.

Vila, R., Nakano, E. Y. e Saulo, H. (2018). “Theoretical results on the discrete Weibull distribution of Nakagawa and Osaki”. Em: *Statistics (no plero)*.

Apêndices

A.1. Função de risco para modelo *odds*-riscos proporcionais

$$\frac{h(t|\mathbf{x})}{1-h(t|\mathbf{x})} = g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}$$

$$h(t|\mathbf{x}) = \left[g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}\right]\cdot[1-h(t|\mathbf{x})]$$

$$h(t|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)} - h(t|\mathbf{x})\,g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}$$

$$h(t|\mathbf{x}) + h(t|\mathbf{x})\,g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)} = g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}$$

$$h(t|\mathbf{x})\left[1+g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}\right] = g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}$$

$$h(t|\mathbf{x})\left[\frac{1-h_0(t)+g(\mathbf{x}'\boldsymbol{\beta})\,h_0(t)}{1-h_0(t)}\right] = g(\mathbf{x}'\boldsymbol{\beta})\,\frac{h_0(t)}{1-h_0(t)}$$

$$h(t|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})\cdot\frac{h_0(t)}{1-h_0(t)}\cdot\frac{1-h_0(t)}{1-h_0(t)+g(\mathbf{x}'\boldsymbol{\beta})\,h_0(t)}$$

$$h(t|\mathbf{x}) = \frac{g(\mathbf{x}'\boldsymbol{\beta})\,h_0(t)}{1-h_0(t)+g(\mathbf{x}'\boldsymbol{\beta})\,h_0(t)}.$$

A.2. Função de sobrevivência para modelo *odds*-riscos proporcionais

Substituindo a função de risco, $h(t|\mathbf{x})$, na expressão (10), chega-se ao seguinte resultado

$$S(t|\mathbf{x}) = \prod_{u=0}^{t} [1 - h(t|\mathbf{x})]$$

$$S(t|\mathbf{x}) = \prod_{u=0}^{t} \left[1 - \frac{g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}{1 - h_0(t) + g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}\right]$$

$$S(t|\mathbf{x}) = \prod_{u=0}^{t} \left[\frac{1 - h_0(t) + g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t) - g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}{1 - h_0(t) + g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}\right]$$

$$S(t|\mathbf{x}) = \prod_{u=0}^{t} \left[\frac{1 - h_0(t)}{1 - h_0(t) + g(\mathbf{x}'\boldsymbol{\beta})\ h_0(t)}\right].$$

A.3. Função de probabilidade para modelo *odds*-riscos proporcionais

A partir da expressão (9), para $t = 0$, tem-se que

$$p(0|\mathbf{x}) = 1 - S(0|\mathbf{x}).$$

Usando (29), chega-se ao seguinte resultado

$$p(0|\mathbf{x}) = 1 - [1 - h(0|\mathbf{x})] = h(0|\mathbf{x}).$$

Já para $t = 1, 2, ...$, a função de probabilidade é dada por

$$p(t|\mathbf{x}) = S(t-1|\mathbf{x}) - S(t|\mathbf{x})$$

$$p(t|\mathbf{x}) = \prod_{u=1}^{t-1}[1 - h(u|\mathbf{x})] - \prod_{u=1}^{t}[1 - h(u|\mathbf{x})]$$

$$p(t|\mathbf{x}) = \prod_{u=1}^{t-1}[1 - h(u|\mathbf{x})] \cdot \left[1 - \prod_{u=1}^{t}[1 - h(u|\mathbf{x})]\right]$$

$$p(t|\mathbf{x}) = S(t-1|\mathbf{x}) \cdot [1 - [1 - h(t|\mathbf{x})]]$$

$$p(t|\mathbf{x}) = S(t-1|\mathbf{x}) \cdot h(t|\mathbf{x}).$$