Universidade de Brasília
Departamento de Estatística

Modelo de Regressão Weibull para Dados Discretos em Análise de Sobrevivência

Ludimila Pereira Nobre

Relatório Final de Monografia apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2016

Ludimila Pereira Nobre

Modelo de Regressão Weibull para Dados Discretos em Análise de Sobrevivência

Orientadora:
Profa. Dra. **Juliana Betini Fachini Gomes**

Relatório Final de Monografia apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2016

Dedicatória

À minha mãe,
***Valdirene**, por me ensinar a lutar pelo o que desejo, além de todo amor, dedicação e apoio.*

Aos meus avós,
***Antonio (in memoriam)** e **Erlite**, pelo grande amor e carinho.*

Agradecimentos

Agradeço a Deus por me presentear com sua boa, perfeita e agradável vontade. Por não desistir de mim e permitir a realização dos meus sonhos.

À Profa. Dra. Juliana Betini Fachini Gomes pela compreensão, apoio, amizade e disposição. Ter te escolhido como orientadora foi uma das melhores decisões que tomei na graduação. Obrigada pelo ótimo ano de trabalho que tivemos e por sempre acreditar em mim.

À minha mãe e à minha avó, por absolutamente tudo. Nada que eu escreva ou diga é capaz de expressar a gratidão que sinto por tudo o que fizeram e continuam fazendo por mim.

Ao meu avô, que apesar de há muito tempo não estar mais aqui, sei que estaria imensamente feliz com esse momento. Gosto de tornar público o quanto te admiro e te amo. Obrigada por ter existido.

À ESTAT Consultoria, principalmente à Diretoria de Projetos, por todo o crescimento pessoal e profissional, pelas experiências, pelas noites mal dormidas e pelas pessoas que conheci e que vou levar pra vida. Não sabia que seria capaz de amar e lutar tanto por essa empresa. ESTAT é muito amor.

Aos meus amigos: de infância, da escola, da graduação, da ESTAT, da vida. Obrigada pelas conversas, risadas, conselhos, lágrimas, apoio, ajuda, paciência, amor e carinho. Vocês são pessoas incríveis e quero que estejam comigo para sempre.

Após terminar os agradecimentos percebi que sou uma pessoa de muita sorte. Há muito prazer em agradecer.

Sumário

Resumo

Modelo de Regressão Weibull para Dados Discretos em Análise de Sobrevivência

Neste trabalho é proposto um modelo de regressão Weibull para dados discretos censurados. A motivação para o desenvolvimento deste modelo diz respeito à inaplicabilidade, em certas situações, de modelos contínuos a dados discretos, que surgem quando a unidade de medida do tempo de sobrevivência está em anos ou meses, por exemplo. Para a construção deste modelo foi utilizada a distribuição Weibull Discreta, que possui flexível função de risco, bem como funções de ligação para que a variável resposta fosse relacionada às covariáveis. Além disso, um algoritmo computacional para a estimação dos coeficientes foi feito. Por fim, aplicou-se um banco de dados real para ilustrar e estudar o modelo de regressão construído.

Palavras-chave: Dados discretos; Distribuição Weibull Discreta; Modelo de regressão; Funções de ligação.

Abstract

Weibull Regression Model for Discrete Data in Survival Analysis

This study proposes a Weibull regression model for censored discrete data. The motivation for the development of it refers to inapplicability in certain situations of continuous models to discrete data, when the unit of measurement of survival time is in years or months, for example. To develop this model were used the Discrete Weibull distribution, which has flexible risk function, both link functions for connect the response variable to the explanatory variables. Beyond that, a computational algorithm for estimating the coefficients was made. Finally, a real database was applied to illustrate and study the regression model built.

Keywords: Discrete data; Discrete Weibull Distribution; Regression model; Link functions.

1 Introdução

A Ciência Estatística tem aplicação nas mais diversas áreas do conhecimento, sendo a Análise de Sobrevivência uma de suas ramificações, que por sua vez é aplicável desde a área médica, passando pelas finanças e chegando às engenharias, motivo pelo qual também é chamada de Análise de Confiabilidade.

O tempo até a ocorrência de determinado evento de interesse é o objeto de estudo, e forma a variável resposta junto à censura, observação parcial da resposta, que por sua vez é a principal característica desses dados. Ao considerar que o tempo é uma variável aleatória aplica-se métodos gráficos e funções não-paramétricas específicas para a definição da distribuição de probabilidade que melhor se adeque às peculiaridades das observações.

Entretanto, quando a unidade de medida do tempo é discretizada, isto é, se o tempo é apurado em anos, por exemplo, adaptações são necessárias nos métodos já conhecidos. Sendo assim, adota-se a distribuição Weibull como base para as modificações a serem feitas, por esta ser uma das distribuições mais empregadas em Análise de Sobrevivência.

O banco de dados utilizado neste trabalho é relativo a um experimento realizado com 137 homens que possuíam câncer de pulmão em estado avançado e inoperável, sendo o tempo de sobrevivência o número de meses desde a entrada do paciente no estudo até a sua morte. Além disso, há uma série de informações sobre os indivíduos, como idade e classificação histológica do tumor.

Para examinar a influência dessas variáveis independentes sobre a variável resposta, utiliza-se a técnica de modelos de regressão por apresentar resultados com maior nível de confiança e por não haver a possibilidade de inclusão direta de covariáveis nos métodos não-paramétricos. Então, após a determinação da distribuição de probabilidade adequada, define-se um modelo de regressão como extensão do modelo probabilístico já proposto.

Por conta do enfoque do trabalho, o principal objetivo se constitui em desenvolver um modelo de regressão baseado na distribuição Weibull Discreta. Os objetivos específicos, por sua vez, são estudar formas de inserir as covariáveis neste modelo e estimar os parâmetros pelo método de máxima verossimilhança através de um algoritmo desenvolvido no software R e interpretá-los.

2 Revisão de Literatura

2.1 Análise de Sobrevivência

A Análise de Sobrevivência é uma área da Ciência Estatística que tem como objeto de estudo o tempo até a ocorrência de determinado evento de interesse. O evento em questão pode ser a morte de um indivíduo, a falha de um equipamento eletrônico, a venda de uma ação na bolsa de valores, dentre outros.

O tempo pode ser medido em dias, meses, anos, ou intervalos pré-determinados. Como a medição da variável é realizada de forma contínua, uma série de acontecimentos pode interromper tal acompanhamento. Cita-se como exemplo a desistência de um paciente em participar de um estudo, ou até a morte do mesmo por causa diferente da esperada. Nestas situações verifica-se a denominada censura. Isto é, a perda de informação decorrente da não observação do evento.

A censura, observação parcial da resposta, compõe a variável resposta em estudo junto ao tempo. A censura pode ser à direita, à esquerda ou intervalar. A censura à direita é definida quando o tempo de ocorrência do evento é superior ao tempo de registro. A censura à esquerda, por sua vez, ocorre quando o evento de interesse aconteceu antes mesmo do indivíduo ser observado. E a censura intervalar é dada em acompanhamentos periódicos.

Existe, ainda, a subdivisão para a censura à direita. Na censura à direita do tipo I o término do estudo é pré-determinado e, ao final, a não ocorrência do evento implica em censura. Já na censura à direita do tipo II, fixa-se a quantidade de falhas no início do estudo. Por fim, na censura à direita aleatória engloba-se os casos em que as observações não falharam por motivos não determinados. Neste trabalho será utilizada a censura à direita aleatória.

A inclusão da censura na análise é importante, pois apesar de incompleta, há informação sobre o tempo dos indivíduos. A omissão da mesma pode introduzir viés nas estimativas, além de não expor de forma verídica a distribuição dos dados.

Os conjuntos de dados de sobrevivência são compostos pela variável resposta, tempo e censura, e covariáveis, as variáveis que possivelmente influenciam o tempo. Esses dados são representados por meio do par (t_i, δ_i), onde t_i representa o tempo observado para o indivíduo i, sendo $i = 1, ..., n$, e δ_i indica se este tempo é de falha ou censura. Logo,

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

Na presença de covariáveis, tem-se então $(t_i, \delta_i, \mathbf{x}_i)$, onde $\mathbf{x}_i$ denota o vetor de covariáveis.

2.2 Funções Específicas

Dada sua aleatoriedade, os tempos formam uma distribuição e podem ser caracterizados por funções específicas. Ao considerar $T = [Y]$, onde $[Y]$ representa "a parte inteira de Y" (maior inteiro menor ou igual a Y), tem-se definida uma variável discreta. Apesar da pouca popularidade das técnicas para dados discretos em Análise de Sobrevivência, a extensão para este caso já está desenvolvida, como visto, por exemplo, em Nakano e Carrasco (2006), Fernandes (2013), entre outros. As funções a seguir são descritas tanto para o caso contínuo como para o caso discreto, que é o foco deste trabalho.

2.2.1 Função Densidade de Probabilidade

Seja T uma variável aleatória contínua, não negativa, que representa o tempo de falha de um elemento. A função de densidade, $f(t)$, é definida como a probabilidade de um indivíduo falhar em um intervalo de tempo, dado que esse intervalo tende a zero. Ou seja, descreve a distribuição de probabilidade dessa variável no intervalo de zero a infinito.

A função de distribuição acumulada, que determina a probabilidade de falha até um determinado tempo, pode ser facilmente obtida ao se conhecer a função de densidade, e vice-versa.

Para o caso discreto, a função de probabilidade é definida como $p(t) = P(T = t)$. A natureza discreta da variável designa valores maiores ou iguais a zero para todo t, principal diferença entre as distribuições discretas e contínuas, onde a probabilidade no ponto é sempre igual a zero no caso contínuo. A função de distribuição acumulada, neste caso, é a soma das probabilidades pontuais até um certo tempo t.

2.2.2 Função de Sobrevivência

A função de sobrevivência é definida como a probabilidade de um indivíduo sobreviver a um determinado tempo t. Portanto, dado que T é uma variável aleatória contínua e que $f(t)$ é sua função densidade de probabilidade:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \int_t^{\infty} f(u)\, du, \end{aligned} \tag{1}$$

em que $f(t) \geq 0$ para todo $t \geq 0$.

Além disso, sabe-se que $F(t) = 1 - S(t)$. Isto é, a função de distribuição acumulada é dada pela probabilidade do indivíduo não sobreviver a um certo tempo t.

Entretanto, ao assumir somente valores inteiros positivos, ou seja, $t = 0, 1, 2, \ldots$, a variável assume caráter discreto e suas funções perdem a propriedade de continuidade. Sendo assim, tem-se que:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \sum_{k=t+1}^{\infty} P(T = k), \qquad t = 0, 1, 2, \ldots \end{aligned} \tag{2}$$

A função $p(t) = P(T = t)$ é a função de probabilidade da variável discreta.

2.2.3 Função de Risco

Dado que o indivíduo sobreviveu ao tempo t, a função de risco, $h(t)$, estabelece a probabilidade da falha acontecer em um curto intervalo de tempo.

Esta função é expressa em termos do limite da probabilidade:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{3}$$

ou seja, é a probabilidade do indivíduo falhar no intervalo $(t, t + \Delta t)$, com Δt tendendo a zero. A função de risco pode ser encontrada a partir da sua relação com a função de densidade e função de sobrevivência, como a seguir:

$$h(t) = \frac{f(t)}{S(t)}. \tag{4}$$

A função de risco é sempre maior ou igual a zero e descreve a probabilidade instantânea de falha ao longo do tempo.

Já no caso discreto, a relação (4) sofre uma leve modificação. Nos pontos onde não ocorrem falhas, a função de risco assume valor zero e, é definida como a probabilidade do indivíduo falhar no tempo t dado que está vivo neste mesmo tempo. Deste modo,

$$\begin{aligned} h(t) &= P(T = t|T \geq t) \\ &= \frac{P(T = t)}{P(T \geq t)} \\ &= \frac{P(T = t)}{P(T > t) + P(T = t)} \\ &= \frac{p(t)}{S(t) + p(t)}, \qquad t = 0, 1, 2, \ldots \end{aligned} \tag{5}$$

Assume-se ainda que $0 \leq h(t) \leq 1$.

2.2.4 Relações Importantes

Como visto anteriormente, as funções que caracterizam os tempos de sobrevivência são matematicamente interligadas. Assim, seja $S(t)$ a função de sobrevivência, então, segundo Fernandes (2013):

$$\begin{aligned} S(t) &= P(T > t) \\ &= P(T > t - 1) - P(T = t) \\ &= S(t - 1) - p(t), \qquad t = 0, 1, 2, \ldots, \end{aligned}$$

e $S(0) = 1 - p(0)$. Logo,

$$S(t) = S(t - 1) - p(t) \quad \Leftrightarrow \quad S(t) + p(t) = S(t - 1) \tag{6}$$

Ao manipular a equação (6) algebricamente:

$$\frac{1}{S(t) + p(t)} = \frac{1}{S(t - 1)} \tag{7}$$

A multiplicação de ambos os lados da equação (7) pela função de probabilidade, $p(t)$, implica, pois:

$$\begin{aligned} h(t) &= \frac{p(t)}{S(t - 1)} \\ &= \frac{S(t - 1) - S(t)}{S(t - 1)} \\ &= 1 - \frac{S(t)}{S(t - 1)}, \qquad t = 1, 2, \ldots \end{aligned} \tag{8}$$

E ainda, tem-se que $h(0) = P(T = 0)$.

2.3 Estimador de Kaplan-Meier

O primeiro passo na análise de qualquer banco de dados é a discriminação das observações a partir de gráficos e medidas descritivas. Em Análise de Sobrevivência, entretanto, esse primeiro passo é prejudicado por conta da presença de censuras na amostra. Isto porque o tempo de censura informa somente que o tempo de falha do indivíduo em questão é maior que o tempo registrado (Colosimo e Giolo, 2006).

Dessa forma, o cálculo da média e variância, por exemplo, não retorna valores consistentes, e técnicas específicas, que incorporem a censura, são necessárias. Com base no que já é conhecido, a função de sobrevivência é capaz de informar aspectos básicos do conjunto de dados. Sendo assim, estimar a função de sobrevivência de forma não-paramétrica é a melhor alternativa em questão para se ter uma ideia inicial do que se passa.

O estimador de Kaplan-Meier interpreta a função de sobrevivência em termos de probabilidades condicionais. Ao considerar n indivíduos em observação e $k(\leq n)$ falhas distintas e ordenadas nos tempos $t_1 < t_2 < \ldots < t_k$, tem-se, então:

$$\begin{aligned} \hat{S}(t) &= \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) \\ &= \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), \end{aligned} \tag{9}$$

em que d_j é o número de falhas em t_j, $j = 1, \ldots, k$, e, n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Em Kaplan e Meier (1958), é provado que $\hat{S}(t)$ é estimador de máxima verossimilhança de $S(t)$ e que, então, possui as propriedades dos estimadores dessa classe, como a convergência para uma distribuição Normal quando o tamanho da amostra tende a infinito, por exemplo.

Estimativas, como a do tempo mediano, podem ser encontradas. Além disso, intervalos de confiança e testes de hipóteses também podem ser construídos com base nessa estimação.

2.3.1 Curva do Tempo Total em Teste

A função de risco da variável aleatória T, ou função taxa de falha, definida na Seção 2.2.3 pode assumir diversos comportamentos. Uma das formas de definir o modelo probabilístico mais adequado aos dados é utilizar a curva do tempo total em teste, TTT plot (Aarset, 1987), que é obtida ao construir o gráfico da função abaixo:

$$G(r/n) \quad = \quad \frac{[(\sum_{i=1}^{r} T_{1:n}) + (n-r)T_{1:n})]}{(\sum_{i=1}^{n} T_{1:n})},$$

por r/n, em que $r = 1, \ldots, n$, e $T_{i:n}, i = 1, \ldots, n$ são as estatísticas de ordem da amostra.

A Figura 1 ilustra as possíveis formas do TTT plot.

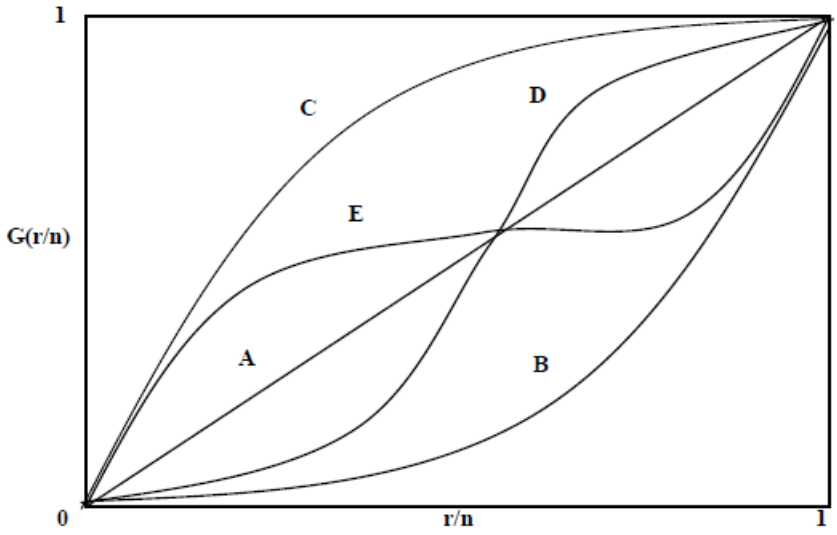


Figura 1: Possíveis formas do TTT plot.

Sendo que:

- Reta diagonal (**A**) ⇒ Função taxa de falha constante é adequada.
- Curva convexa (**B**) ou côncava (**C**) ⇒ Função taxa de falha é monotonicamente decrescente ou crescente, respectivamente.
- Curva convexa e depois côncava (**D**) ⇒ Função taxa de falha tem forma de **U**.
- Curva côncava e depois convexa (**E**) ⇒ Função taxa de falha é unimodal.

2.4 Modelos de Probabilidade

As distribuições de probabilidade assumem importante papel na modelagem dos dados de sobrevivência. A experiência demonstra que algumas distribuições são mais adequadas que outras, pois relatam as características do tempo de forma mais precisa.

A distribuição Weibull, por sua vez, é amplamente utilizada em Análise de Sobrevivência por apresentar flexível função de risco. Uma breve descrição para o caso contínuo e extensão detalhada para o caso discreto são apresentadas a seguir.

2.4.1 Modelo Weibull

Ao considerar a variável aleatória contínua T, com distribuição Weibull, sua função de densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \qquad t \geq 0, \tag{10}$$

em que γ e α são positivos e são, respectivamente, os parâmetros de forma e o de escala. O parâmetro de escala, α, tem a mesma unidade de medida de t e γ não tem unidade.

Pela definição de função de sobrevivência dada na equação (1), tem-se para a distribuição Weibull:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}. \tag{11}$$

E, pela relação (4),

$$h(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}. \tag{12}$$

O parâmetro γ determina a forma da função de risco. Para $\gamma < 1$, a função de risco é monótona decrescente. Se $\gamma > 1$, monótona crescente e, se $\gamma = 1$, tem-se a distribuição Exponencial com função de risco constante.

Os modelos discretos podem ser obtidos através de modelos contínuos. As distribuições Exponencial e Gama (para o caso de variáveis aleatórias contínuas), por exemplo, possuem correspondentes distribuições discretas (para o caso de variáveis aleatórias discretas). Sendo assim, também é possível encontrar a distribuição discreta correspondente à distribuição Weibull (Nakagawa e Osaki, 1975).

Se $Y \sim Weibull(\gamma, \alpha)$, a variável aleatória discreta T pode ser obtida ao considerar

$T = [Y]$, como definido na Seção 2.2. Segundo Fernandes (2013):

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq Y < t+1) \\ &= P(Y < t+1) - P(Y \leq t) \\ &= F_Y(t+1) - F_Y(t) \\ &= 1 - S_Y(t+1) - 1 + S_Y(t) \\ &= S_Y(t) - S_Y(t+1). \end{aligned}$$

Ao considerar a função de sobrevivência para o caso contínuo como definida na equação (11), segue que:

$$\begin{aligned} p(t) &= \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} - \exp\left\{-\left(\frac{t+1}{\alpha}\right)^{\gamma}\right\} \\ &= q^{t^{\gamma}} - q^{(t+1)^{\gamma}}, \qquad t = 0, 1, 2, \ldots, \end{aligned} \tag{13}$$

em que $q = \exp\left\{-\frac{1}{\alpha^{\gamma}}\right\}$ é interpretado como uma probabilidade, pois $0 < q < 1$ para qualquer γ e α maiores que zero,

A função de sobrevivência, por sua vez:

$$\begin{aligned} S(t) &= \sum_{k=t+1}^{\infty} P(T = k) \\ &= \sum_{k=t+1}^{\infty} \left(q^{k^{\gamma}} - q^{(k+1)^{\gamma}}\right) \\ &= \left(q^{(t+1)^{\gamma}} - q^{(t+2)^{\gamma}}\right) + \left(q^{(t+2)^{\gamma}} - q^{(t+3)^{\gamma}}\right) + \ldots \\ &= q^{(t+1)^{\gamma}}, \qquad t = 0, 1, 2, \ldots. \end{aligned} \tag{14}$$

Enquanto que, a função de risco:

$$\begin{aligned} h(t) &= \frac{p(t)}{S(t) + p(t)} \\ &= \frac{q^{t^{\gamma}} - q^{(t+1)^{\gamma}}}{q^{(t+1)^{\gamma}} + \left(q^{t^{\gamma}} - q^{(t+1)^{\gamma}}\right)} \\ &= \frac{q^{t^{\gamma}}}{q^{t^{\gamma}}} - \frac{q^{(t+1)^{\gamma}}}{q^{t^{\gamma}}} \\ &= 1 - q^{(t+1)^{\gamma} - t^{\gamma}}, \qquad t = 0, 1, 2, \ldots \end{aligned} \tag{15}$$

Assim como no caso contínuo, a função de risco da distribuição Weibull discreta é crescente se $\gamma > 1$ e decrescente se $\gamma < 1$. Se $\gamma = 1$, o modelo é simplificado à distribuição

Geométrica, que, por sua vez, possui função de risco constante.

2.5 Método de Máxima Verossimilhança

Após a determinação da distribuição de probabilidade, a estimação dos parâmetros é necessária. Há dois métodos de estimação muito conhecidos e amplamente adotados na literatura, o Método de Mínimos Quadrados e o Método de Máxima Verossimilhança.

Dada a natureza dos dados de sobrevivência, mais precisamente por conta das censuras, necessita-se de um método que absorva todas as informações disponíveis. Sendo assim, o Método de Mínimos Quadrados se mostra impróprio, uma vez que não é possível incorporar a censura na função que será minimizada.

O Método de Máxima Verossimilhança, que tem como objetivo encontrar o valor do parâmetro que maximiza a probabilidade da amostra observada acontecer, é adequado, pois permite a inclusão da função de sobrevivência, contribuição das censuras, na função de verossimilhança. A função de densidade corresponde aos tempos de falha.

Independente do mecanismo de censura à direita adotado, a expressão para a função de verossimilhança é a mesma, a menos de constantes, e é dada por (Colosimo e Giolo, 2006):

$$L(\theta) \propto \prod_{i=1}^{n} [f(t_i;\theta)]^{\delta_i} [S(t_i;\theta)]^{1-\delta_i}, \tag{16}$$

em que δ_i é a variável indicadora de falha. Para o caso discreto, a equação (16) é atualizada para:

$$L(\theta) \propto \prod_{i=1}^{n} [p(t_i;\theta)]^{\delta_i} [S(t_i;\theta)]^{1-\delta_i}. \tag{17}$$

Ao aplicar o modelo Weibull discreto na função de verossimilhança definida em (17), tem-se (Brunello e Nakano, 2015):

$$L(q,\gamma) = \prod_{i=1}^{n} \left(q^{t_i^{\gamma}} - q^{(t_i+1)^{\gamma}}\right)^{\delta_i} \left(q^{(t_i+1)^{\gamma}}\right)^{(1-\delta_i)}. \tag{18}$$

Por conveniência, ao utilizar o logaritmo na equação (18):

$$\log(L(q,\gamma)) = \sum_{i=1}^{n} \left(\delta_i \log\left(q^{t_i^{\gamma}} - q^{(t_i+1)^{\gamma}}\right) + (1-\delta_i)\log\left(q^{(t_i+1)^{\gamma}}\right)\right). \tag{19}$$

Os estimadores de máxima verossimilhança para q e γ são encontrados ao derivar a função $\log(L(q, \gamma))$ em relação a cada parâmetro, igualando as equações obtidas a zero e resolvendo o sistema de equações. Métodos computacionais de otimização são indicados pois tais cálculos podem ser extensos.

3 Metodologia

3.1 Material

O conjunto de dados em questão se refere ao tempo até a morte de 137 homens com câncer de pulmão avançado e inoperável. Os pacientes foram aleatoriamente distribuídos entre dois tipos de tratamento quimioterápico: o padrão e o teste. Dessa forma, o objetivo principal do estudo é avaliar a influência do tratamento no tempo de sobrevivência (Cordeiro et al. (2011) e Kalbfleisch and Prentice (2002)).

Nas referências citadas anteriormente, a análise é feita com o tempo em dias, porém, para que as propriedades dos dados discretos fossem melhor observadas, o tempo, neste trabalho, é representado pelo número de meses desde o início do tratamento até a morte do indivíduo. Se $t = 0$, o indivíduo morreu antes de completar um mês (1 mês = 30 dias) de tratamento.

Como o tempo de sobrevivência pode ser influenciado por demais aspectos, além do tipo de tratamento (0 = padrão, 1 = teste), outras cinco covariáveis foram analisadas. A primeira covariável é a *performance status*, uma medida que tenta quantificar o estado de bem-estar geral do paciente, neste caso com base na classificação de Karnofsky, com escala entre 0 e 100.

Nos estudos já mencionados essa variável foi utilizada de forma quantitativa, entretanto, a sua descrição sugere certa categorização, e esta foi a configuração aqui adotada. Ao considerar os possíveis estados de bem-estar geral do paciente, tem-se que:

- O paciente é considerado completamente hospitalizado se a *performance status* está entre 0 e 30;
- O paciente é considerado parcialmente hospitalizado se o valor observado está entre 31 e 60;
- O paciente é capaz de cuidar de si mesmo se a medida está entre 61 e 100.

As demais covariáveis observadas são: idade (em anos) do paciente; tempo, em meses, do diagnóstico da doença até a entrada no estudo; terapia prévia (0 = não, 1 = sim); e, por fim, classificação histológica do tumor: 1 se escamoso, 2 se pequeno, 3 se adeno e 4 se grande.

3.2 Métodos

3.2.1 Modelo de Regressão Weibull Discreto

A Análise Descritiva, um dos métodos estatísticos de obtenção de informação, tem papel importante na sumarização de dados. Através de tabelas, gráficos e medidas, é possível conhecer e investigar o comportamento de um conjunto de variáveis, tanto de forma univariada quanto multivariada. Porém, todos os resultados obtidos nestes termos são somente amostrais, isto é, a generalização para a população não ocorre.

Sendo assim, procedimentos que permitem estimar as características da população foram desenvolvidos e são amplamente utilizados. Dependendo do objetivo do estudo, uma possibilidade é recorrer ao método de Análise de Regressão. Os modelos de regressão expressam a relação entre uma variável resposta e uma ou mais variáveis explicativas por meio de um modelo matemático, facilitando, ainda, saber a significância dessa relação.

Na Análise de Sobrevivência, o tempo até a ocorrência do evento de interesse e a censura constituem a variável resposta. O estimador não-paramétrico de Kaplan-Meier, por exemplo, é uma das técnicas para analisar descritivamente o tempo em relação às demais variáveis. Além disso, ao construir o modelo de regressão, o que se deseja é saber como este mesmo tempo é afetado por covariáveis.

O tempo de sobrevivência, por sua vez, pode ser contínuo ou discreto. Os tempos discretos ganham espaço quando a unidade de medida do mesmo é em meses ou intervalos. Para analisar esses dados, uma abordagem inicial é a aplicação de modelos contínuos. Porém, em certas condições, esse tratamento pode não ser o mais indicado (Nakano e Carrasco, 2006).

Compreender a natureza do tempo e a forma adequada de tratá-lo influi diretamente na qualidade das estimativas e consistência das inferências realizadas. Logo, como ajustar um modelo contínuo à tempos discretos não é pertinente, propõe-se um modelo de regressão discreto fundamentado na distribuição de probabilidade Weibull.

Entretanto, faz-se necessário definir como será estabelecida a relação entre as covariáveis e a variável resposta no modelo. Há uma extensa literatura em Análise de Sobrevivência ilustrando a construção de modelos com a característica de continuidade, sendo que uma das maneiras é inserir os preditores em um dos parâmetros da distribuição de probabilidade. Mas, para o caso discreto o mesmo não é encontrado, o que abre o leque para estudar como inserir covariáveis nos modelos discretos, objetivo principal deste trabalho.

Para os dados grupados, que são um caso particular de dados discretos, a probabi-

lidade do indivíduo falhar em um intervalo qualquer dado seu vetor de covariáveis, e que sobreviveu ao início deste mesmo intervalo, é modelada por meio de função de ligação (Hashimoto, 2008; e, Rocha, 2013). Conforme desenvolvido na Seção 2.4.1, a função de probabilidade para a distribuição Weibull Discreta é dada por:

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

sendo o parâmetro $q = \exp\left\{-\frac{1}{\alpha^\gamma}\right\}$ interpretado como uma probabilidade. Portanto, ao unir essas duas informações, este trabalho propõe que o modelo de regressão Weibull Discreto será estruturado pela inserção das covariáveis no parâmetro q mediante função de ligação.

3.2.2 Função de Ligação

A função de ligação denota uma função $g(.)$ que conecta a variável resposta às variáveis explicativas (Agresti, 2007). Para um conjunto de p covariáveis, o parâmetro q passa a ser definido como na equação (20):

$$q_i \;=\; g(\eta_i), \tag{20}$$

em que $\eta_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}$ é o preditor linear e β_0 e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ é o que se deseja estimar.

As formas de inserção de covariáveis no modelo serão estudadas por meio de diferentes funções de ligação e, apesar de várias existirem, as mais aplicadas em parâmetros limitados no intervalo de zero a um são a logito, complemento log-log e log-log (Hashimoto, 2008), que serão apresentadas a seguir.

3.2.3 Ligação Logito

A ligação logito é especificada abaixo:

$$\begin{aligned} g(\eta_i) \;&=\; \frac{\exp(-\eta_i)}{1+\exp(-\eta_i)} \\ &=\; \frac{\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}{1+\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}. \end{aligned} \tag{21}$$

Sendo assim, ao aplicar as equações (20) e (21) na função de probabilidade da distribuição Weibull Discreta, encontra-se:

$$p(t) = \left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}{1+\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}\right)^{t^\gamma} - \left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}{1+\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}\right)^{(t+1)^\gamma}, \quad t = 0, 1, 2, \ldots. \tag{22}$$

A função de sobrevivência especificada pela equação (14) é atualizada para:

$$S(t) \quad = \quad \left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))}{1 + \exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))} \right)^{(t+1)^\gamma}. \tag{23}$$

E a função de risco dada pela equação (15), por sua vez:

$$h(t) \quad = \quad 1 - \left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))}{1 + \exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))} \right)^{(t+1)^\gamma - t^\gamma}. \tag{24}$$

Ao considerar as equações (22) e (23), a equação (19), apresentada na Seção 2.5, passa a ser, então:

$$\begin{aligned} \log(L(\beta_0, \boldsymbol{\beta}, \gamma)) \quad = \quad & \sum_{i=1}^{n} \left(\delta_i \log \left(\left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))}{1 + \exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))} \right)^{t_i^\gamma} - \left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))}{1 + \exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))} \right)^{(t_i+1)^\gamma} \right) \right) + \\ & \sum_{i=1}^{n} \left((1 - \delta_i) \log \left(\left(\frac{\exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))}{1 + \exp(-(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))} \right)^{(t_i+1)^\gamma} \right) \right), \end{aligned} \tag{25}$$

em que γ é o parâmetro de forma, β_0 é o intercepto e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ é o vetor de coeficientes das covariáveis do modelo.

Após a construção e estimação do modelo de regressão, é necessário interpretar os coeficientes obtidos.

Para o caso Weibull contínuo, uma alternativa para a interpretação dos coeficientes estimados de modelos de locação e escala, por exemplo, é o uso da razão de tempos medianos. Seja x uma covariável binária, tem-se que a razão dos tempos medianos é dada por (Colosimo e Giolo, 2006):

$$\frac{t_{0,5}(\hat{\beta}|x=1)}{t_{0,5}(\hat{\beta}|x=0)} \quad = \quad e^{\hat{\beta}}. \tag{26}$$

Logo, se $\hat{\beta}$ for positivo, interpreta-se que o tempo mediano de um indivíduo do grupo $x = 1$ é $e^{\hat{\beta}}$ o tempo mediano de um indivíduo do grupo $x = 0$. Entretanto, se $\hat{\beta}$ for negativo, conclui-se que o tempo mediano de um indivíduo do grupo $x = 0$ é $e^{-\hat{\beta}}$ o tempo mediano de um indivíduo do grupo $x = 1$.

A razão dos tempos medianos foi feita para o modelo Weibull Discreto que utiliza a ligação logito, porém, não se obteve o resultado como em (26). Sendo assim, outra opção é considerar, então, a equação (24) em que $t = 0$:

$$h(0) \quad = \quad 1 - q, \tag{27}$$

e, ao fazer uso de (21):

$$\begin{aligned} q &= \frac{\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))}{1+\exp(-(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}))} \\ &= \frac{1}{1+\exp(\beta_0+\beta_1 x_i)} \\ &= 1 - h(0). \end{aligned} \tag{28}$$

Dado que o parâmetro q é interpretado como uma probabilidade e pode ser escrito em termos da função de risco em $t = 0$, associa-se à esta ideia o cálculo da *odds* para o indivíduo i, isto é, a sua chance de falhar em $t = 0$ dado que sobreviveu a este mesmo tempo, como possível modo de obter $e^{\hat{\beta}}$. Logo, ao realizar esse cálculo:

$$\begin{aligned} odds &= \frac{h(0)}{1-h(0)} \\ &= \frac{1-q}{q} \\ &= \frac{1-\dfrac{1}{1+\exp(\beta_0+\beta_1 x_i)}}{\dfrac{1}{1+\exp(\beta_0+\beta_1 x_i)}} \\ &= \exp(\beta_0+\beta_1 x_i). \end{aligned} \tag{29}$$

Ao considerar dois indivíduos diferentes, a razão de *odds*, ou razão de chances, será:

$$\begin{aligned} \text{RO} &= \frac{odds_j}{odds_i} \\ &= \frac{\exp(\beta_0+\beta_1 x_j)}{\exp(\beta_0+\beta_1 x_i)} \\ &= \exp(\beta_1 x_j - \beta_1 x_i) \\ &= e^{\beta_1}, \end{aligned} \tag{30}$$

ao tomar $x_j - x_i = 1$ unidade.

Segue que a interpretação para os coeficientes estimados do modelo de regressão Weibull Discreto utilizando a função de ligação logito, considerando $t = 0$ e tudo o mais constante, é como a interpretação da razão de chances. Ou seja, a chance de um indivíduo em que $x = 1$ falhar no tempo zero é $e^{\hat{\beta}}$ vezes a chance de um indivíduo em que $x = 0$ falhar no tempo zero, tudo o mais constante.

Isto implica que, se $\hat{\beta}$ for positivo, a probabilidade de um indivíduo em que $x = 1$ falhar no tempo zero é maior do que a probabilidade de um indivíduo em que $x = 0$ falhar no tempo zero. Caso $\hat{\beta}$ seja negativo, a probabilidade de um indivíduo em que $x = 1$

falhar no tempo zero é menor do que a probabilidade de um indivíduo em que $x = 0$ falhar no tempo zero.

Se houver covariáveis categorizadas, considera-se variáveis indicadoras e uma das categorias como grupo controle para o cálculo da razão de chances. A interpretação citada acima também é adequada para covariáveis contínuas.

3.2.4 Ligação Complemento Log-Log

A ligação complemento log-log é expressa por:

$$\begin{aligned} g(\eta_i) &= 1 - \exp[-\exp(\eta_i)] \\ &= 1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]. \end{aligned} \tag{31}$$

Assim como na seção anterior, ao aplicar a função de ligação em questão na função de probabilidade, tem-se:

$$p(t) = \left(1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]\right)^{t^\gamma} - \left(1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]\right)^{(t+1)^\gamma}. \tag{32}$$

Já a função de sobrevivência:

$$S(t) = \left(1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]\right)^{(t+1)^\gamma}. \tag{33}$$

E, por fim, a função de risco é dada por:

$$h(t) = \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{(t+1)^\gamma + t^\gamma}. \tag{34}$$

Logo, o logaritmo da função de máxima verossimilhança utilizando a ligação complemento log-log é denotado por:

$$\begin{aligned} \log(L(\beta_0, \boldsymbol{\beta}, \gamma)) = &\sum_{i=1}^{n} \left(\delta_i \log\left(\left(1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]\right)^{t_i^\gamma} - \left(1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]\right)^{(t_i+1)^\gamma}\right)\right) + \\ &\sum_{i=1}^{n}((1-\delta_i)\log((1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})])^{(t_i+1)^\gamma}). \end{aligned} \tag{35}$$

O resultado da razão de tempos medianos e sua interpretação para o caso contínuo não podem ser estendidos para o caso discreto segundo a função de ligação logito, como visto na seção anterior, e ao se testar, nem para a função de ligação complemento log-log.

Ao aplicar a ideia desenvolvida em (29) e (30) para a ligação complemento log-log, também não se obtém $e^{\hat{\beta}}$ como resultado da razão de chances. Sendo assim, a discussão

sobre a interpretação dos coeficientes estimados é retomada para a função de ligação em questão.

O modelo de regressão de Cox utiliza a propriedade de taxas de falha proporcionais do modelo para a interpretação de seus coeficientes estimados (Colosimo e Giolo, 2006). A razão dessas taxas de falha para dois indivíduos distintos, ao considerar somente uma covariável x é dada por:

$$\frac{h(t|x_j)}{h(t|x_i)} = \exp[\beta_1(x_j - x_i)]. \tag{36}$$

Ao assumir a diferença entre x_j e x_i igual a 1 unidade, o que se tem é a razão de tempos medianos para modelos de locação e escala. Pela equação (34), $h(0)$ é dada por:

$$h(0) = \exp[-\exp(\beta_0 + \beta_1 x)]. \tag{37}$$

O valor resultante em (26) é encontrado ao considerar a razão dos logaritmos da função de risco em $t = 0$ para dois indivíduos distintos, sendo x uma covariável binária:

$$\begin{aligned} \frac{\log[h(0|x=1)]}{\log[h(0|x=0)]} &= \frac{-\exp(\beta_0 + \beta_1)}{-\exp(\beta_0)} \\ &= e^{\beta_1}. \end{aligned} \tag{38}$$

Portanto, o logaritmo da probabilidade de falha em $t = 0$ para os indivíduos em que $x = 1$ é $e^{\hat{\beta}_1}$ vezes o logaritmo da probabilidade de falha em $t = 0$ para os indivíduos em que $x = 0$. Isto implica que, se $\hat{\beta}$ for positivo, a probabilidade de um indivíduo em que $x = 1$ falhar no tempo zero é menor do que a probabilidade de um indivíduo em que $x = 0$ falhar no tempo zero. Se $\hat{\beta}$ for negativo, a probabilidade de um indivíduo em que $x = 1$ falhar no tempo zero é maior do que a probabilidade de um indivíduo em que $x = 0$ falhar no tempo zero.

Deste modo, a interpretação para os coeficientes estimados do modelo de regressão Weibull Discreto utilizando a função de ligação complemento log-log, considerando $t = 0$ e tudo o mais constante, é dada em termos da razão dos logaritmos da função de risco de dois indivíduos distintos, tanto para covariáveis binárias quanto para covariáveis categorizadas, em que variáveis indicadoras e uma categoria como grupo controle são considerados para o cálculo em (38). A interpretação também é válida para covariáveis contínuas.

3.2.5 Ligação Log-Log

A ligação log-log, por sua vez, é definida como:

$$\begin{aligned} g(\eta_i) &= \exp[-\exp(\eta_i)] \\ &= \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]. \end{aligned} \tag{39}$$

A equação (40) define a função de probabilidade para a função de ligação log-log:

$$p(t) = \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{t^\gamma} - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{(t+1)^\gamma}, \tag{40}$$

sua respectiva função de sobrevivência:

$$S(t) = \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{(t+1)^\gamma}, \tag{41}$$

e sua função de risco:

$$h(t) = 1 - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{(t+1)^\gamma + t^\gamma}. \tag{42}$$

Por conseguinte, tem-se como função de máxima verossimilhança para esta ligação:

$$\begin{aligned} \log(L(\beta_0, \boldsymbol{\beta}, \gamma)) = &\sum_{i=1}^{n} \left(\delta_i \log\left(\exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{t_i^\gamma} - \exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{(t_i+1)^\gamma}\right)\right) + \\ &\sum_{i=1}^{n} \left((1-\delta_i) \log\left(\exp[-\exp(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})]^{(t_i+1)^\gamma}\right)\right). \end{aligned} \tag{43}$$

Em relação à interpretação dos coeficientes estimados, novamente a razão de tempos medianos não se adequou ao caso e foi necessário utilizar a abordagem do modelo de regressão de Cox assim como para a ligação complemento log-log.

Seja x uma covariável binária e a função de risco, definida na equação (42), obtida em $t = 0$, tem-se que:

$$h(0) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)]. \tag{44}$$

Para dois indivíduos distintos, x_i e x_j, a razão dos logaritmos de $1 - h(0|x = 1)$ e $1 - h(0|x = 0)$ retorna como resultado $e^{\hat{\beta}_1}$, como pode ser visto na equação abaixo:

$$\begin{aligned} \frac{\log[1 - h(0|x=1)]}{\log[1 - h(0|x=0)]} &= \frac{-\exp(\beta_0 + \beta_1)}{-\exp(\beta_0)} \\ &= e^{\beta_1}. \end{aligned} \tag{45}$$

Desse modo, o logaritmo de 1 menos a probabilidade de falhar no tempo zero de um indivíduo em que $x = 1$ é $e^{\hat{\beta}_1}$ vezes o logaritmo de 1 menos a probabilidade de falhar no tempo zero de um indivíduo em que $x = 0$. Isto implica que, se $\hat{\beta}$ for positivo, a probabilidade de um indivíduo em que $x = 1$ falhar no tempo zero é maior do que a probabilidade de um indivíduo em que $x = 0$ falhar no tempo zero. Se $\hat{\beta}$ for negativo, a probabilidade de um indivíduo em que $x = 1$ falhar no tempo zero é menor do que a probabilidade de um indivíduo em que $x = 0$ falhar no tempo zero.

Por fim, a interpretação dos coeficientes estimados do modelo de regressão Weibull Discreto utilizando a função de ligação log-log, considerando $t = 0$ e tudo o mais constante, se dá em termos da razão dos logaritmos de 1 menos a função de risco, para dois indivíduos distintos. A interpretação é válida para variáveis binárias, categorizadas e contínuas, respeitadas as suas características.

Os estimadores para β_0, $\boldsymbol{\beta}$ e γ são obtidos, em todos os casos, derivando a função $\log(L(\beta_0, \boldsymbol{\beta}, \gamma))$ em relação a cada parâmetro, igualando as equações obtidas a zero e resolvendo o sistema de equações. O algoritmo de otimização desenvolvido para a estimação dos parâmetros, no caso da função de ligação logito, se encontra nos Anexos. Para as demais funções definidas a ideia é a mesma, somente o parâmetro q é atualizado conforme a ligação em questão.

4 Resultados e Discussões

4.1 Análise Descritiva

A análise de resultados é iniciada a partir da análise exploratória das variáveis em estudo.

A Figura 2 apresenta a função de sobrevivência estimada segundo o estimador não-paramétrico de Kaplan-Meier. Como já esperado, a probabilidade de sobrevida diminui gradualmente conforme o tempo de estudo aumenta.

O tempo mediano indica que metade dos pacientes sobrevivem até, aproximadamente, 2 meses após o início do tratamento. Além disso, somente 10 dos 137 homens sobreviveram a um ano ou mais de estudo.

O maior tempo em questão é igual a 33 meses, obtendo este 2 observações.

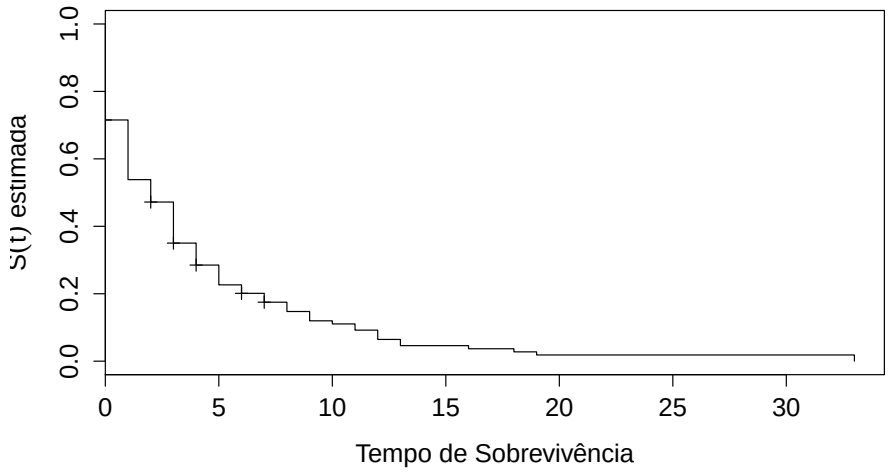


Figura 2: Função de sobrevivência estimada por Kaplan-Meier.

Quando o tempo tende a infinito, a função de sobrevivência acima tende a zero, sendo designada como própria. É interessante citar que somente 9 indivíduos foram censurados.

O TTT Plot, recurso gráfico que auxilia na escolha do modelo probabilístico a ser utilizado com base na função de risco, foi feito. Como pode ser visto na Figura 3, apesar de ser uma função escada, o comportamento convexo é perceptível, indicando que a função taxa de falha é monotonicamente decrescente.

Dado o comportamento observado no TTT Plot, a distribuição Weibull, própria para funções de risco monotonicamente crescentes, decrescentes ou constantes, se adequa

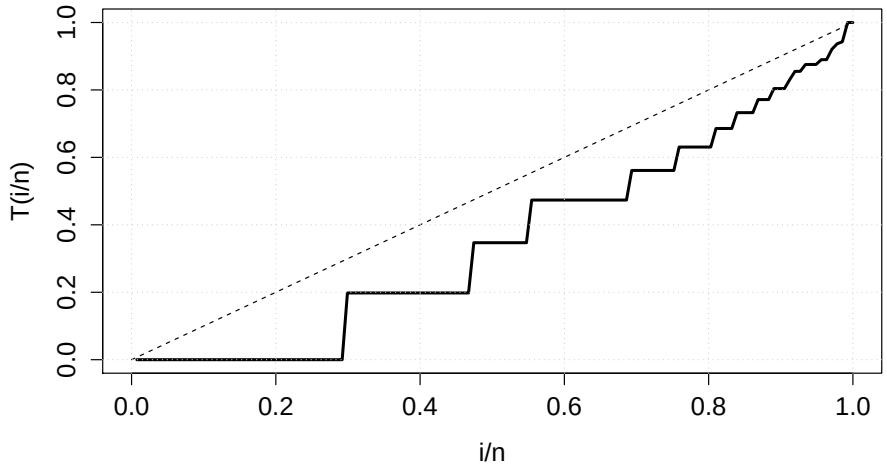


Figura 3: TTT Plot do tempo de sobrevivência.

bem aos dados. A distribuição Weibull discreta, então, será utilizada na seção a seguir para modelar os dados.

A primeira covariável a ser descrita é a *performance status*. Ao observar a Figura 4, verifica-se que, como não há violação da suposição de riscos proporcionais, o Teste de LogRank pode ser aplicado para verificar a igualdade das curvas de sobrevivência dos grupos.

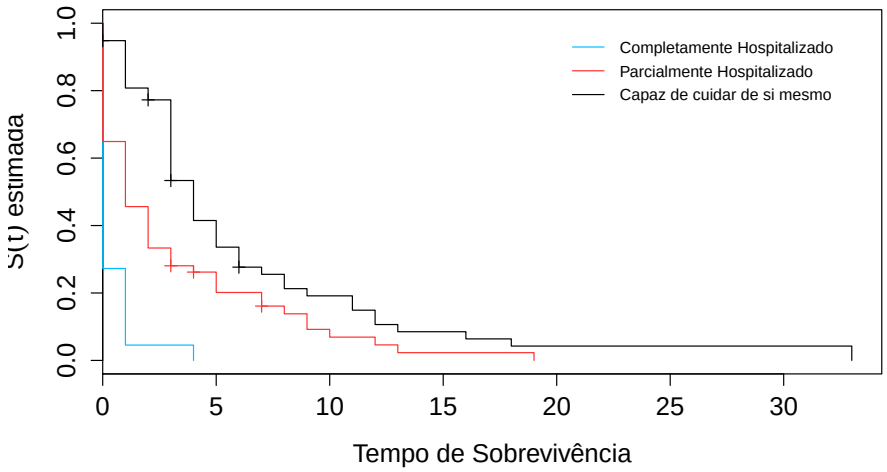


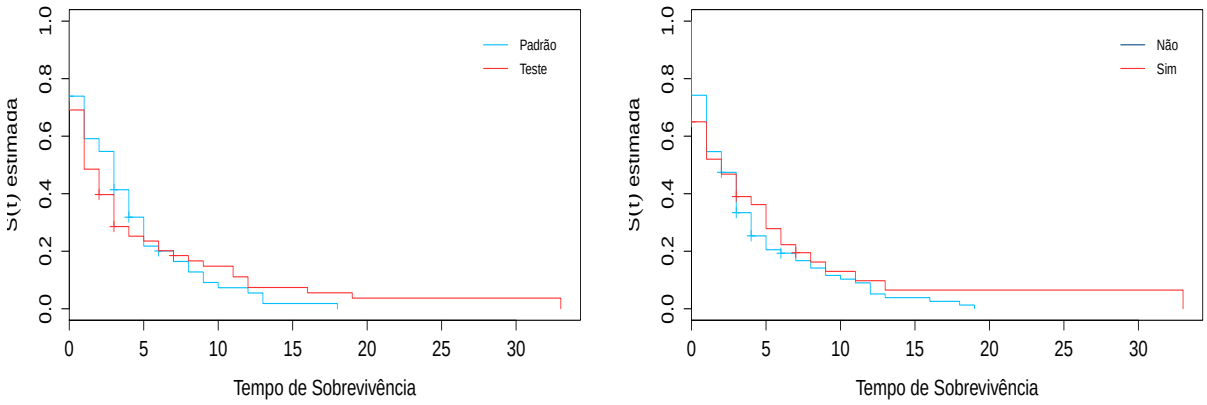
Figura 4: Função de sobrevivência estimada por Kaplan-Meier segundo a covariável Performance Status.

O p-valor retornado para esse teste é menor que 0,001 indicando que há evidências estatísticas para rejeitar, ao nível de 5% de significância, a hipótese de que as curvas de

sobrevivência são iguais. Este resultado sinaliza que tal covariável pode ser significativa na construção do modelo.

Nota-se que o resultado do teste é intuitivo pois, como é visto na Figura 4, quanto melhor for seu *performance status*, mais tempo o indíviduo sobrevive e mais lentamente sua probabilidade de sobrevivência decresce.

Ao explorar as covariáveis tratamento e terapia prévia, tem-se que a suposição de riscos proporcionais não é atendida, logo o Teste de Wilcoxon é aplicado e a hipótese de igualdade das curvas de sobrevivência não é rejeitada, ao nível de 5% de significância, com p-valor acima de 0,3 em ambos os casos. Inserir a covariável terapia prévia no modelo poderá não ser relevante, porém, por conta do objetivo do estudo, incluir o tipo de tratamento se torna essencial.

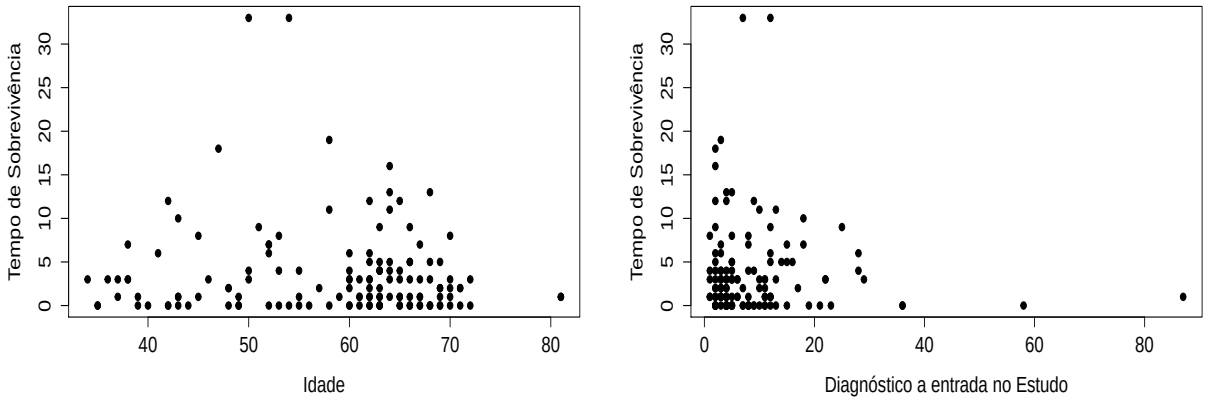


(a) Tratamento (b) Terapia Prévia

Figura 5: Função de sobrevivência estimada por Kaplan-Meier para as covariáveis Tratamento e Terapia Prévia.

A idade do indivíduo e o tempo do diagnóstico da doença até a entrada do paciente no estudo são as duas covariáveis quantitativas do conjunto de dados. Ao apurar a correlação linear das mesmas com o tempo de sobrevivência obteve-se, respectivamente, -0,068 e -0,044. Estes valores e a Figura 6 apontam que há indícios para a não existência de correlação linear entre tais variáveis e o tempo. Como essas relações podem não ser significativas também no modelo de regressão, estas variáveis não serão incluídas no modelo.

A Figura 7, que retrata a função de sobrevivência segundo a classificação do tumor, destaca que o tempo pode ser influenciado pelo tipo de tumor em questão. Há pares de curvas, por exemplo, que não se cruzam ao longo do tempo e apresentam probabilidades de sobrevivência distintas, assinalando que esta pode ser uma importante covariável para o modelo.



(a) Idade (b) Tempo do diagnóstico até a entrada no estudo

Figura 6: Gráfico de dispersão para o tempo de sobrevivência e as covariáveis Idade e Tempo do diagnóstico da doença até a entrada no estudo.

Logo, o Teste de Wilcoxon foi aplicado com o intuito de averiguar se as curvas de sobrevivência dos grupos são iguais. O p-valor do teste, menor que 0,001, indica que há evidências estatísticas para rejeitar a hipótese de que as curvas de sobrevivência são iguais ao nível de significância de 5%. Desta forma, conclui-se que há evidências que o tipo de tumor influencia o tempo de sobrevivência do indivíduo.

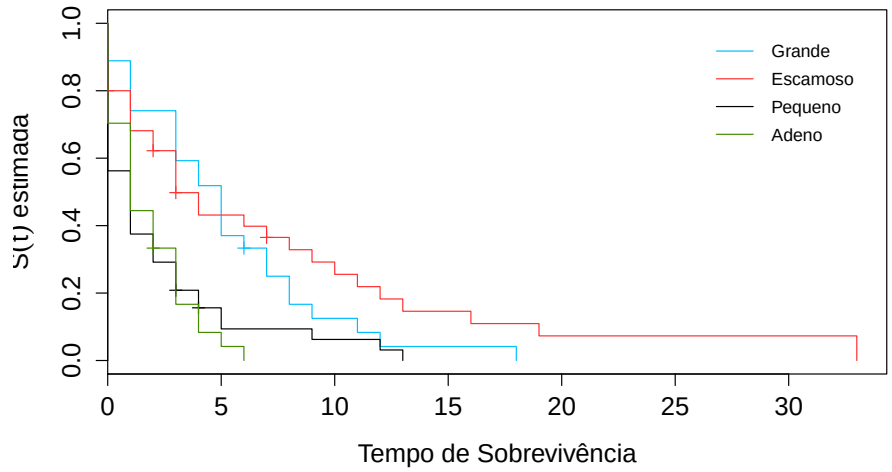


Figura 7: Função de sobrevivência estimada por Kaplan-Meier segundo a covariável Classificação histológica do tumor.

4.2 Estimação

4.2.1 Modelo Weibull Discreto

Conforme a análise exploratória realizada na seção anterior, o TTT Plot indicou que a distribuição Weibull é adequada ao tempo de sobrevivência do conjunto de dados em estudo. O modelo Weibull discreto sem covariável, desenvolvido na seção 2.4.1, é então aplicado, e estimativas iniciais de seus parâmetros são obtidas a partir do algoritmo *Optim* no software R.

A Tabela 1 apresenta tanto os valores estimados dos parâmetros quanto seus respectivos erros padrão.

Tabela 1: Estimativas do modelo Weibull Discreto sem covariável.

Parâmetro	Estimativa	Erro Padrão
q	$0,7106$	$0,0353$
γ	$0,7906$	$0,0630$

Os erros padrão são baixos e indicam que as estimativas são consistentes. Um indício de que o modelo de regressão se adequará bem aos dados é o ajuste satisfatório da função de sobrevivência estimada pelo modelo sem covariável em comparação com as estimativas de Kaplan-Meier. Como pode ser visto na Figura 8, isto ocorre com os dados do câncer de pulmão.

O ajuste geral é aceitável e sugere que o modelo Weibull Discreto pode ser usado para modelar esses dados. Sendo assim, a próxima etapa é incluir covariáveis no modelo e ajustar o modelo de regressão Weibull Discreto proposto neste trabalho.

4.2.2 Modelo de Regressão Weibull Discreto

Por meio da análise descritiva realizada na Seção 4.1 foi possível identificar as covariáveis que, possivelmente, exercem influência na composição do modelo e, consequentemente, explicação da variável resposta.

Sendo assim, dois modelos foram considerados. O primeiro, chamado de modelo generalizado, é constituído pelas variáveis classificação histológica do tumor, performance status e tipo de tratamento. Apesar do Teste de Wilcoxon não rejeitar a hipótese das curvas de sobrevivência dos grupos de tratamento padrão e teste serem iguais, considera-se esta variável por conta do objetivo do estudo.

Como todas as variáveis em questão para o modelo generalizado são categorizadas,

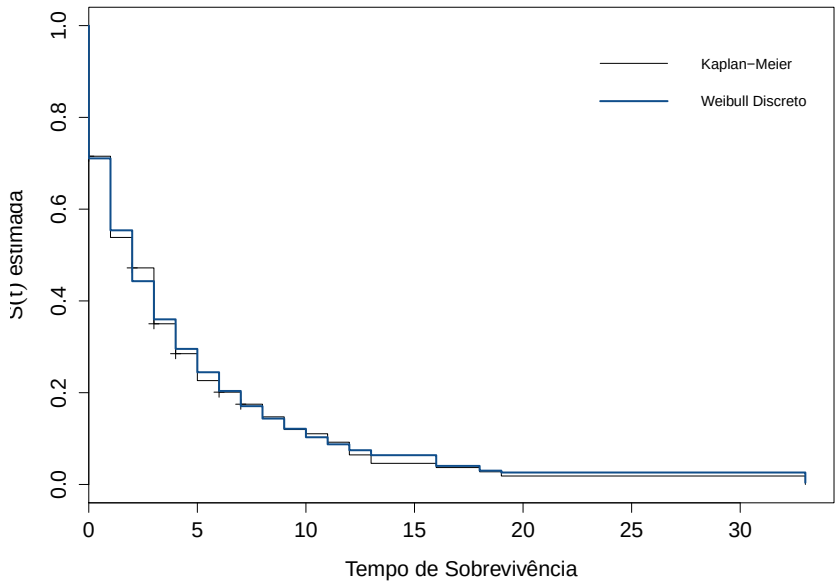


Figura 8: Função de sobrevivência estimada por Kaplan-Meier e pelo modelo Weibull Discreto sem covariável.

necessita-se criar variáveis *dummy* para que a informação qualitativa seja introduzida de forma correta no modelo. A variável dummy indica a presença ou ausência de característica pré-estabelecida em determinada observação.

Para uma variável formada por k categorias, emprega-se $k-1$ variáveis dummy. Como exemplo utiliza-se a variável classificação histológica do tumor que, de acordo com a Seção 3.1, possui as seguintes categorias: escamoso, pequeno, adeno e grande. Então, as variáveis dummy criadas estão abaixo:

$$D_{i1} = \begin{cases} 1, & \text{se o tumor do i-ésimo paciente for escamoso,} \\ 0, & \text{c.c.} \end{cases}$$

$$D_{i2} = \begin{cases} 1, & \text{se o tumor do i-ésimo paciente for pequeno,} \\ 0, & \text{c.c.} \end{cases}$$

$$D_{i3} = \begin{cases} 1, & \text{se o tumor do i-ésimo paciente for adeno,} \\ 0, & \text{c.c.} \end{cases}$$

O mesmo processo foi feito para a variável performance status. A variável tratamento, por sua vez, já se constitui como uma variável indicadora. Logo, tem-se que o parâmetro q, no modelo generalizado, ao utilizar a função de ligação logito, por exemplo, passa a ser definido como na equação (46):

$$q_i = \frac{\exp(\beta_0 + \beta_1 \text{Escamoso}_i + \beta_2 \text{Pequeno}_i + \beta_3 \text{Adeno}_i + \beta_4 \text{Hospital}_i + \beta_5 \text{P. Hospital}_i + \beta_6 \text{Tratamento}_i)}{1 + \exp(\beta_0 + \beta_1 \text{Escamoso}_i + \beta_2 \text{Pequeno}_i + \beta_3 \text{Adeno}_i + \beta_4 \text{Hospital}_i + \beta_5 \text{P. Hospital}_i + \beta_6 \text{Tratamento}_i)}. \quad (46)$$

Como visto acima, as variáveis receberam nomenclatura segundo o que indicam, sendo que o tumor grande e o grupo capaz de cuidar de si mesmo foram considerados como níveis de referência em relação às demais categorias das variáveis a que pertencem.

O segundo modelo, que é um submodelo do modelo generalizado, é composto somente pelas variáveis classificação histológica do tumor e performance status. Deste modo, o parâmetro q deste submodelo, utilizando a função de ligação logito, é definido por:

$$q_i = \frac{\exp(\beta_0 + \beta_1 \text{Escamoso}_i + \beta_2 \text{Pequeno}_i + \beta_3 \text{Adeno}_i + \beta_4 \text{Hospital}_i + \beta_5 \text{P. Hospital}_i)}{1 + \exp(\beta_0 + \beta_1 \text{Escamoso}_i + \beta_2 \text{Pequeno}_i + \beta_3 \text{Adeno}_i + \beta_4 \text{Hospital}_i + \beta_5 \text{P. Hospital}_i)}. \quad (47)$$

Ao utilizar essas covariáveis o parâmetro q também será definido ao considerar as funções de ligação complemento log-log e log-log como definido nas equações (31) e (39). Dessa forma, foram obtidos o modelo de regressão Weibull Discreto com função logito, o modelo de regressão Weibull Discreto com função de ligação complemento log-log e o modelo de regressão Weibull Discreto com função de ligação log-log, tanto para o caso generalizado como para seu submodelo.

A Tabela 2 apresenta as estimativas para os parâmetros do modelo generalizado segundo as funções de ligação consideradas.

Tabela 2: Estimativas do modelo de regressão generalizado Weibull Discreto conforme função de ligação.

Parâmetro	Função de Ligação								
	Logito			Complemento Log-Log			Log-Log		
	Estimativa	Erro Padrão	p-valor	Estimativa	Erro Padrão	p-valor	Estimativa	Erro Padrão	p-valor
β_0	-2,2152	0,3238	<0,0001	0,8216	0,1312	<0,0001	-2,2281	0,2969	<0,0001
β_1	-0,5618	0,3129	0,0726	0,2032	0,1283	0,1133	-0,4929	0,2809	0,0793
β_2	0,6086	0,3024	0,0442	-0,3132	0,1450	0,0307	0,5201	0,2648	0,0495
β_3	0,9953	0,3449	0,0039	-0,5155	0,1785	0,0039	0,8217	0,2965	0,0056
β_4	2,7756	0,4549	<0,0001	-1,6813	0,3335	<0,0001	2,1180	0,3122	<0,0001
β_5	0,6144	0,2331	0,0084	-0,2732	0,1071	0,0108	0,5585	0,2102	0,0079
β_6	0,2420	0,2341	0,3013	-0,0583	0,1067	0,5850	0,2026	0,2006	0,3125
γ	1,0883	0,0888	-	1,0735	0,0866	-	1,0835	0,0881	-

Ao aplicar as funções de ligação logito e log-log observa-se que não há diferença expressiva nas estimativas e erros padrão obtidos. Além disso, a um nível de significância de 5%, as covariáveis "Escamoso" e "Tratamento", esta como já esperado, não são significativas em ambos os casos.

Ademais, os sinais das estimativas são os mesmos para as ligações logito e log-log.

Sendo assim, para a ligação log-log, um indivíduo que possui tumor do tipo Pequeno possui maior probabilidade de falhar no tempo zero, ou seja, morrer antes de completar um mês de tratamento, do que um indivíduo que possui tumor do tipo Grande.

De forma análoga, um indivíduo que foi classificado na *performance status* como hospitalizado possui maior probabilidade de morrer antes de completar um mês de tratamento do que um indivíduo que foi classificado como capaz de cuidar de si mesmo.

As estimativas do submodelo do modelo generalizado, por sua vez, encontram-se na Tabela 3. Na formulação inicial deste modelo a variável Tratamento não é considerada e, como visto na Tabela 2, por não ter sido significativa no modelo generalizado não há problemas em não considerá-la aqui.

Tabela 3: Estimativas do submodelo do modelo generalizado Weibull Discreto conforme função de ligação.

Parâmetro	Função de Ligação								
	Logito			Complemento Log-Log			Log-Log		
	Estimativa	Erro Padrão	p-valor	Estimativa	Erro Padrão	p-valor	Estimativa	Erro Padrão	p-valor
β_0	-2,0944	0,2992	<0,0001	0,7974	0,1244	<0,0001	-2,1273	0,2766	<0,0001
β_1	-0,4721	0,3001	0,1157	0,1826	0,1228	0,1368	-0,4283	0,2731	0,1168
β_2	0,5877	0,3014	0,0512	-0,3116	0,1449	0,0316	0,4983	0,2641	0,0592
β_3	1,0559	0,3390	0,0018	-0,5316	0,1760	0,0025	0,8766	0,2915	0,0026
β_4	2,7537	0,4507	<0,0001	-1,6821	0,3335	<0,0001	2,1190	0,3111	<0,0001
β_5	0,5771	0,2304	0,0123	-0,2641	0,1058	0,0126	0,5269	0,2083	0,0114
γ	1,0796	0,0877	-	1,0725	0,0865	-	1,0769	0,0874	-

O comportamento observado para as estimativas e erros padrão do modelo generalizado se repete para seu submodelo. As funções de ligação logito e log-log apresentam valores próximos e possuem as mesmas variáveis significativas ao nível de 5% de significância.

Para a ligação complemento log-log, as variáveis significativas são as já citadas e, somente neste submodelo, a variável indicadora do tumor pequeno não se encontra no limite da região de rejeição. Apesar disso, discute-se a significância dessa variável nos demais modelos por seu p-valor se apresentar pouco acima de 5%.

Também não se encontra aqui diferenças entre os sinais das estimativas para as ligações logito e log-log. Logo, ao considerar a função de ligação complemento log-log, um indivíduo que possui tumor Pequeno possui maior probabilidade de falhar no tempo zero, isto é, morrer antes de completar um mês de tratamento, do que um indivíduo que possui tumor Grande.

Como observado, há um padrão nos valores das estimativas e erros padrão para este conjunto de dados tanto para o modelo generalizado quanto para seu submodelo considerando as funções de ligação. Porém, esta característica pode não ser observada em outros bancos de dados.

Após a exposição da Tabela 2 e Tabela 3, deve-se selecionar o modelo que melhor se ajusta aos tempos de sobrevivência observados. Alguns dos critérios de seleção de modelos são o Critério de Akaike (AIC), Critério de Akaike Corrigido (AICc) e Critério de Informação Bayesiano (BIC), sendo que o modelo com menor valor na medida em questão é escolhido como o mais adequado.

O uso do AIC é indicado somente quando n/p é maior que 40, porém, como este não é o caso, serão utilizadas as medidas AICc e BIC. A Tabela 4 apresenta as medidas observadas para o modelo generalizado segundo as funções de ligação consideradas.

Tabela 4: Medidas de seleção de modelos para o modelo generalizado segundo função de ligação.

Função de Ligação	AICc	BIC
Logito	575,3547	597,5895
Complemento Log-Log	576,7868	599,0217
Log-Log	575,6750	597,9098

Ao tomar estritamente o menor valor para AICc e BIC, o modelo que utiliza a função de ligação logito é o escolhido por ambas as medidas. Entretanto, o modelo das funções de ligação complemento log-log e log-log também são satisfatórios dado que os seus valores variam pouco em relação aos do logito.

No caso do submodelo do modelo generalizado, mais uma vez o modelo com a função de ligação logito é o que apresenta os menores valores para as medidas, como pode ser visto na Tabela 5. Os demais modelos, assim como no caso generalizado, não devem ser descartados.

Tabela 5: Medidas de seleção de modelos para o submodelo do modelo generalizado segundo função de ligação.

Função de Ligação	AICc	BIC
Logito	574,1688	593,7405
Complemento Log-Log	574,8283	594,3999
Log-Log	574,4424	594,0140

Como os modelos escolhidos a partir das medidas de seleção de modelos são encaixados, utiliza-se o Teste da Razão de Verossimilhança para que o melhor dentre os dois modelos em questão seja adotado.

A hipótese nula desse teste indica que o modelo com o menor número de parâmetros é o mais adequado. Ao aplicar o TRV, obtém-se como estatística de teste, aproximadamente, -1,0709 e p-valor igual a 1. Ou seja, ao nível de significância de 5%, há evidências estatísticas suficientes para afirmar que o submodelo do modelo generalizado é o mais apropriado aos dados.

Embora o resultado do teste indique a não utilização do modelo generalizado, deve-se levar em consideração o objetivo do estudo, que consiste em avaliar como o tratamento realizado pelo paciente afeta seu tempo de sobrevivência. Para fins puramente estatísticos, adota-se o modelo indicado no TRV, até mesmo em razão do princípio da parcimônia; ao contextualizar o problema deve-se utilizar o modelo generalizado.

Sendo assim, ao considerar o modelo generalizado que utiliza a ligação logito, a chance de um indivíduo que possui tumor Escamoso falhar no tempo zero, ou seja, morrer antes de completar um mês de tratamento, é $e^{-0,5618}$, 0,5702, vezes a chance de um indivíduo que possui tumor Grande falhar no tempo zero, tudo o mais constante. Isto implica que a probabilidade de falhar no tempo zero de um indivíduo que possui tumor Escamoso é menor do que a probabilidade de um indivíduo que possui tumor Grande falhar no tempo zero.

Em relação à *performance status*, a chance de um indivíduo que foi classificado como hospitalizado falhar no tempo zero é $e^{2,7756}$, 16,0483, vezes a chance de um indivíduo que foi classificado como capaz de cuidar de si mesmo falhar no tempo zero, tudo o mais constante. Dessa forma, a probabilidade de falhar no tempo zero de um indivíduo classificado como hospitalizado é maior do que a probabilidade de um indivíduo que foi classificado como capaz de cuidar de si mesmo falhar no tempo zero.

Outro importante ponto a ser analisado é o parâmetro de forma γ. Nos seis modelos apresentados, a estimativa deste parâmetro foi praticamente a mesma, sempre próxima a 1, e seu erro padrão próximo a zero em todos os casos. Dessa forma, tem-se o indicativo de que a distribuição Geométrica também pode ser adequada para os dados.

Ao aplicar o Teste da Razão de Verossimilhança para o modelo generalizado com a função de ligação logito e este mesmo modelo assumindo $\gamma = 1$, que passa a ser o modelo de interesse por ter menos parâmetros a serem estimados. O p-valor retornado é igual a 0,3108, não rejeitando a hipótese de que a distribuição Geométrica associada à ligação complemento log-log se adequa bem aos dados. O mesmo resultado é obtido ao considerar o submodelo do modelo generalizado.

5 Considerações Finais

O presente trabalho buscou colaborar com o desenvolvimento do referencial teórico para a análise de dados discretos em Análise de Sobrevivência, principalmente no que é relativo a modelos de regressão.

Os modelos contínuos são extensamente utilizados e a todo o momento adaptados para que o ajuste aos dados seja o melhor possível. Porém, as características singulares dos tempos de sobrevivência discretos solicitam técnicas específicas.

Sendo assim, foi proposto um modelo de regressão com base na distribuição Weibull Discreta, tomando a ideia de inserção das covariáveis em um dos parâmetros do modelo probabilístico. Ao considerar a teoria de dados grupados, definiu-se as funções de ligação como método capaz de relacionar a variável resposta às explicativas por meio do parâmetro.

O banco de dados empregado para ilustrar o modelo construído retornou resultados e comportamentos próximos tanto para o modelo generalizado como para seu submodelo, para as funções de ligação consideradas. Enfatiza-se o fato de que o mesmo pode não ser obtido para outros conjuntos de dados.

A interpretação para os coeficientes estimados é modificada segundo a função de ligação em questão. A razão de tempos medianos para modelos contínuos de locação e escala foi considerada como alternativa de interpretação, porém, para que este resultado fosse obtido no caso discreto, empregou-se a abordagem de razão de chances para a função de ligação logito e do modelo de regressão de Cox para as funções de ligação complemento log-log e log-log.

Algumas medidas de seleção de modelos e o teste da Razão de Verossimilhança foram calculados, indicando que o submodelo que utiliza a função de ligação complemento log-log foi o melhor dentre todos os estudados. Entretanto, discute-se as variáveis que compõem o modelo com base no objetivo do estudo.

Diante de tudo o que foi apresentado, é de interesse para estudos futuros a análise da adequabilidade dos Resíduos de Cox-Snell para modelos discretos, como forma de averiguar a qualidade do ajuste do modelo.

Referências

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley, Florida.

Brunello, G. H. V. and Nakano, E. Y. (2015). Inferência bayesiana no modelo weibull discreto em dados com presença de censura. *TEMA Tend. Mat. Apl. Comput.*, v. 16:p. 97–110.

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

Cordeiro, G. M., Ortega, E. M. M., and Silva, G. O. (2011). *Modelos de Regressão Estendidos em Análise de Sobrevivência*. ABE - Associação Brasileira de Estatística, Ceará.

Fernandes, L. M. (2013). Inferência bayesiana em modelos discretos com fração de cura. Dissertação (Mestrado em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília.

Hashimoto, E. M. (2008). Modelo de regressão para dados com censura intervalar e dados de sobrevivência grupados. Dissertação (Mestrado em Estatística) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

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

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

Nakano, E. Y. and Carrasco, C. G. (2006). Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *TEMA Tend. Mat. Apl. Comput.*, v. 7:p. 91–100.

Rocha, T. S. (2013). Modelos de regressão discretos para dados grupados: Uma aplicação em avaliação de risco em produto de crédito parcelado. Monografia (Graduação em Estatística) - Departamento de Estatística, Universidade de Brasília, Brasília.

Selvin, S. (2008). *Survival analysis for epidemiologic and medical research: A practical guide*. Cambridge University Press., New York.

Anexos

A.1 Estimação do modelo Weibull Discreto

```
#### FUNÇÃO DE MÁXIMA VEROSSIMILHANÇA ####

vwd <- function(parametros){

  q <- parametros[1]
  gamma <- parametros[2]

  if((q > 0) && (q <1) && (gamma > 0))

    return(-1*(sum(censura*log(q^(tempo^gamma) - q^((tempo + 1)^gamma)) +
    (1 - censura)*log(q^((tempo + 1)^gamma)))))

  else return(-Inf)

}

dados <- read.table('Câncer de Pulmão.csv', sep = ';', header = T)
tempo <- dados$tempo
censura <- dados$censura

e1 <- optim(c(0.5, 1.2), vwd, hessian = T)

h <- e1$hessian
invh <- solve(h)
sqrt(diag(invh))
```

A.2 Estimação do Modelo de Regressão Weibull Discreto

```
#### FUNÇÃO DE LIGAÇÃO LOGITO: MODELO GENERALIZADO ####

vwdm1 <- function(parametros){

beta0 <- parametros[1]
```

```
beta1 <- parametros[2]
beta2 <- parametros[3]
beta3 <- parametros[4]
beta4 <- parametros[5]
beta5 <- parametros[6]
beta6 <- parametros[7]
gamma <- parametros[8]

q <- (exp(-(beta0 + beta1*escamoso + beta2*pequeno + beta3*adeno +
beta4*hospital + beta5*phospital + beta6*tratamento))/(1 + exp(-(beta0 +
beta1*escamoso + beta2*pequeno + beta3*adeno + beta4*hospital +
beta5*phospital + beta6*tratamento))))

if((gamma > 0))

return(-1*(sum(censura*log(q^(tempo^gamma) - q^((tempo + 1)^gamma)) +
(1 - censura)*log(q^((tempo + 1)^gamma)))))

else return(-Inf)

}

m1 <- nlm(vwdm1, c(1.45, 0.26, -0.89, -1, -2.62, -0.7, -0.2, 0.93), hessian = T)

hessianam1 <- m1$hessian
invhm1 <- solve(hessianam1)
epm1 <- sqrt(diag(invhm1))
parm1 <- m1$estimate
sigm1 <- (parm1/epm1)^2

pvalorm1 <- NULL
for(i in 1:length(sigm1)){
pvalorm1[i] <- 1 - pchisq(sigm1[i], 1)
}

#### FUNÇÃO DE LIGAÇÃO LOGITO: SUBMODELO ####

vwdm2 <- function(parametros){
```

```
beta0 <- parametros[1]
beta1 <- parametros[2]
beta2 <- parametros[3]
beta3 <- parametros[4]
beta4 <- parametros[5]
beta5 <- parametros[6]
gamma <- parametros[7]

q <- (exp(-(beta0 + beta1*escamoso + beta2*pequeno + beta3*adeno +
beta4*hospital + beta5*phospital))/(1 + exp(-(beta0 + beta1*escamoso +
beta2*pequeno + beta3*adeno + beta4*hospital + beta5*phospital))))

if((gamma > 0))

return(-1*(sum(censura*log(q^(tempo^gamma) - q^((tempo + 1)^gamma)) +
(1 - censura)*log(q^((tempo + 1)^gamma)))))

else return(-Inf)

}

m2 <- nlm(vwdm2, c(1.45, 0.26, -0.89, -1, -2.62, -0.7, 0.93),
hessian = T)

hessianam2 <- m2$hessian
invhm2 <- solve(hessianam2)
epm2 <- sqrt(diag(invhm2))
parm2 <- m2$estimate
sigm2 <- (parm2/epm2)^2

pvalorm2 <- NULL
for(i in 1:length(sigm2)){
pvalorm2[i] <- 1 - pchisq(sigm2[i], 1)
}

## O mesmo processo foi feito para as funções de ligação complemento
## log-log e log-log. Modificou-se somente a expressão para o parâmetro q na
## função de verossimilhança.
```