Universidade de Brasília - UnB
Instituto de Ciências Exatas - IE
Departamento de Estatística - EST

Modelo de Regressão Kumaraswamy Log-Logístico para dados discretos

Beatriz Leal Simões e Silva

Orientadora: Professora Juliana Betini Fachini Gomes

Brasília

2019

Beatriz Leal Simões e Silva

Modelo de Regressão Kumaraswamy Log-Logístico para dados discretos

Relatório apresentado à disciplina de Trabalho de Conclusão de Curso II de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Orientadora: Professora Juliana Betini Fachini Gomes

Brasília

2019

Resumo

Este trabalho propõe um modelo de regressão Kumaraswamy Log-Logístico discreto para dados de Análise de Sobrevivência. Devido à falta de modelos probabilísticos discretos na área, propor uma distribuição pouco estudada que possui um bom ajuste a dados censurados é de grande interesse. O conjunto de dados utilizado é referente à política Zona Especial de Interesse Social (ZEIS) e o modelo proposto verifica a influência de cada covariável no tempo de adesão dos municípios brasileiros à política. Por fim, utiliza-se a análise de resíduos para a validação do modelo. As análises foram feitas com o *software* R.

Palavras-chave: Análise de Sobrevivência, Distribuição Kumaraswamy Log-Logística Discreta, Modelo de Regressão.

Abstract

This paper proposes a discrete Kumaraswamy Log-Logistic regression model for data in Survival Analysis. Due to lack of discrete probabilistic models, proposing a new class of distribution, that could fit well to censored data, is interesting. The dataset used on this paper refers to the Special Social Interest Zone (Zona Especial de Interesse Social - ZEIS). The proposed model verifies the importance of each variable of Brazilian counties adhering to the policy ZEIS along the time. Finally, a residual analysis is used in order to verify the aspect of the fitted model. All analysis were done using *software* R.

Keywords: Survival Analysis,Kumaraswamy Log-Logistic distribution, Regression Model

Sumário

1 Introdução

Métodos Estatísticos são amplamente utilizados para estudar a relação entre variáveis dentre um conjunto de dados em diversas áreas da ciência, como as áreas médica, financeira e engenharias.Os estudos na área de análise de sobrevivência são observacionais ou experimentais, pois a variável resposta dos dados possui natureza longitudinal (COLOSIMO; GIOLO, 2006). Dessa forma, a Análise de Sobrevivência é uma área da Estatística que consiste em estudar dados relacionados ao tempo decorrido até a ocorrência de determinado evento de interesse também identificado como falha. Esses eventos de interesse formam a variável resposta, que muitas vezes é acompanhada da presença de censura.

A variável resposta em Análise de Sobrevivência é chamada de tempo de falha. Para definir o tempo de falha é preciso definir o início do estudo, a escala do tempo a ser utilizada e estabelecer o evento de interesse. Censura refere-se à perda de informação derivada de não se ter observado a data de ocorrência do desfecho, resultando em observações parciais ou incompletas. As observações censuradas devem ser incluídas na análise dos dados pois essas observações fornecem informações sobre o tempo de vida de objetos e indivíduos e a omissão de dados censurados pode acarretar conclusões viciadas na análise estatística. Dessa forma, em Análise de Sobrevivência, a variável resposta é composta pelo tempo de falha e tempo de censura.

O tempo é uma variável de natureza contínua. Entretanto, muitas vezes a informação obtida do tempo é registrada por intervalos, ou seja, observações que teriam naturalmente tempos de falha diferentes, passam a ocupar o mesmo espaço de tempo devido à característica de seu registro. Portanto, estudos que registram o tempo de falha mensalmente ou anualmente são caracterizados como dados discretos.

Em Análise de Sobrevivência a maior parte das distribuições são contínuas, alguns estudos apresentam propostas de funções de sobrevivência para tempos discretos, como pode ser encontrado em Cardial (2017) e Santos (2018). Devido à falta de modelos para dados discretos em Análise de Sobrevivência, o objetivo deste trabalho será propor um modelo para dados discretos através da aplicação do método da discretização de distribuições conhecidas para dados contínuos, mais especificamente da distribuição Kumaraswamy Generalizada Log-Logística.

1.1 Objetivos Gerais

O objetivo geral é propor um modelo para dados discretos através do método de discretização de distribuições conhecidas, mais especificamente da distribuição Kumaraswamy Log-Logística. Bem como, propor a generalização da nova distribuição para incluir covariáveis originando o modelo de regressão Kumaraswamy-Log-Logístico para dados discretos.

1.1.1 Objetivos Específicos

- Entender o processo de discretização de distribuições contínuas
- Calcular e demonstrar o processo de discretização da distribuição escolhida
- Estender a nova distribuição para incluir covariáveis afim de obter um modelo de regressão
- Implementar computacionalmente a nova distribuição
- Analisar dados de estudo em Análise de Sobrevivência

2 Revisão de Literatura

2.1 Análise de Sobrevivência

A análise de sobrevivência é definida como um conjunto de procedimentos estatísticos para análise de dados em que o tempo até a ocorrência de determinado evento é a variável resposta. O termo análise de sobrevivência faz referência às situações médicas, quando indivíduos submetidos a um determinado evento de interesse são estudados a fim de estimar o seu tempo de vida. Porém, em outras áreas de estudo essa técnica pode ser utilizada em que componentes são colocados sob teste para estimar características relacionadas ao seu tempo de vida.

2.1.1 Tempo de Falha e Censura

O início e final do estudo precisam ser bem definidos para compor a origem dos tempos e o término do estudo. O tempo de vida dos indivíduos será contado a partir do tempo de início do estudo. Ao iniciar essa coleta de dados, os indivíduos tenderão a falhar em determinado tempo, o que será classificado como tempo de falha. Se a falha não ocorrer, será chamado de tempo de censura.

A não ocorrência do evento de interesse dentro do tempo de estudo é classificado com o tempo de censura, ou seja, não foi observado a falha. As observações que possuem essa característica não devem ser retiradas do estudo, pois podem conter informações importantes dos indivíduos e retirá-las poderia ocasionar um viés.

Assim, para representar as censuras, uma variável indicadora é introduzida para identificar que o tempo de vida foi censura ou falha. Essa variável é denotada por:

$$\delta = \begin{cases} 1, & \text{se ocorrer falha} \\ 0, & \text{se ocorrer censura.} \end{cases} \tag{2.1}$$

A censura pode ser classificada em três grupos, e a censura à direita pode ser dividida em três subgrupos diferentes:

- Censura à direta: O tempo até a ocorrência do evento de interesse está à direita do tempo registrado.
  - Censura do Tipo I: O tempo do término do estudo é estabelecido previamente.
  - Censura do Tipo II: O estudo termina após observar o evento de interesse em um número preestabelecido de indivíduos.

  - Censura Aleatória: Ocorre quando um indivíduo sai do estudo sem ocorrer o evento de interesse, ou se o indíviduo falhar por uma razão diferente da estudada.

- Censura à esquerda: O tempo até a ocorrência do evento de interesse está à esquerda do tempo registrado.

- Censura intervalar: O evento de interesse ocorre em determinado intervalo, ou seja, o registro do tempo não foi feito de forma contínua e sim por intervalos de tempo determinado pelo pesquisador.

Os mecanismos de censura à direita estão representados pela Figura 1 para um melhor entendimento do mesmo.

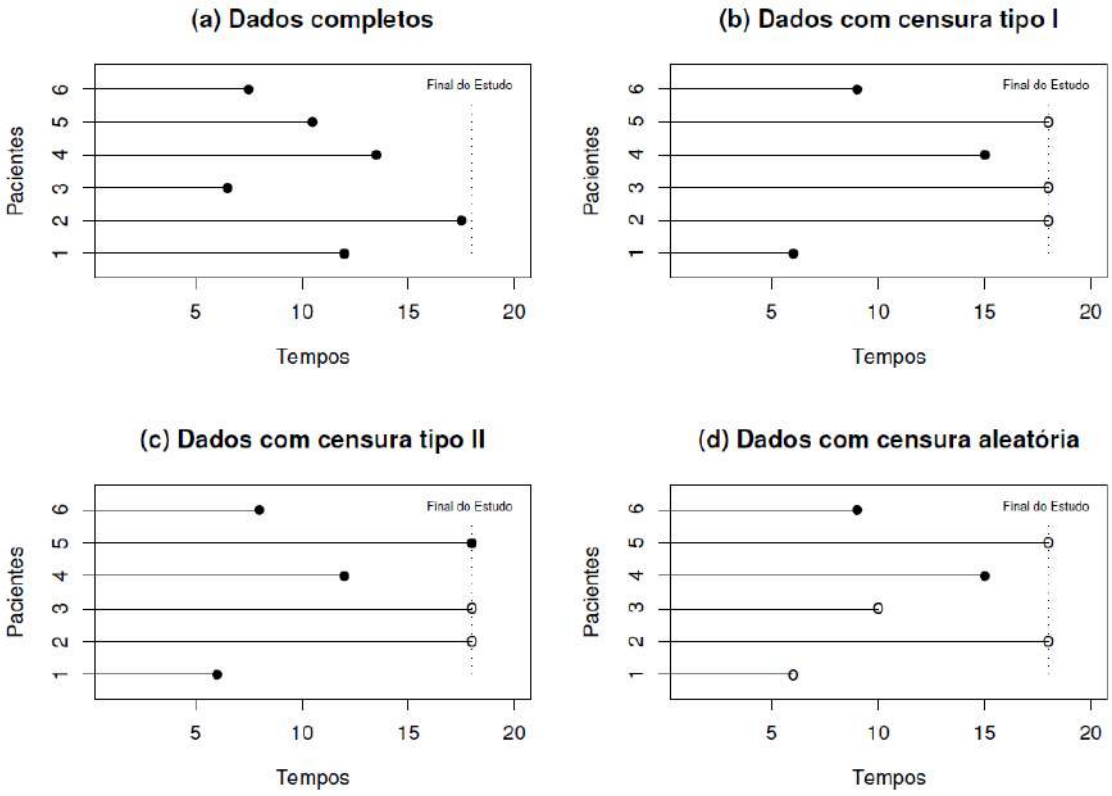

Figura 1 – Ilustração dos tipos de censura à direita. "•" representa falha e "∘" representa a censura. Fonte: Adaptado de Colosimo e Giolo (2006)

2.1.2 Função de Probabilidade

Uma variável aleatória, de forma intuitiva, representa uma quantidade numérica definida a partir do resultado de um experimento aleatório. A função que descreve a

probabilidade relativa de uma variável aleatória se tornar um valor específico é chamada de função de probabilidade $0 \leq p(x) \leq 1$ para variáveis aleatórias discretas, e função de densidade de probabilidade (fdp) $0 \leq f_X(x) \leq 1$ para variáveis aleatórias contínuas.

2.1.3 Função de Sobrevivência

A probabilidade de um indivíduo sobreviver é representada pela função de sobrevivência $S(t)$ e possui relação direta com a função de distribuição acumulada (fda). Quando a variável aleatória T é contínua, sua forma é dada por:

$$S_X(t) = P(T > t) = 1 - P(T \leq t) = 1 - F_X(t) \tag{2.2}$$

No caso em que a variável aleatória T é discreta, a forma da função de sobrevivência se difere da contínua. Isso ocorre porque para os dados discretos o indivíduo pode experimentar a falha no tempo $T = 0$ e assim como na função de probabilidade, o evento de interesse ocorre somente em intervalos consistentes do tempo T. A função de sobrevivência no caso discreto é dada por:

$$\begin{aligned} S(t) &= P(T > t) \\ &= \sum_{k=t+1}^{\infty} p(k) \quad \text{t=0,1,2...,} \end{aligned}$$

em que $p(k)$ é a função de probabilidade de k.

2.1.4 Função de Risco

A função de risco ou taxa de falha, representa o risco de um indivíduo experimentar o evento de interesse estudado no tempo T, dado que ele sobreviveu até o tempo T. A função de risco é representada por $h_X(t)$ e pode assumir diversas formas, como: constante, crescente, decrescente, em forma de banheira e unimodal. A forma da função de risco quando T é uma variável contínua é dada por:

$$h_X(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.3}$$

Além disso, a função de risco $h_X(t)$ possui uma importante relação com a função de densidade $f_X(t)$ e a função de Sobrevivência $S_X(t)$:

$$h_X(t) = \frac{f_X(t)}{S_X(t)}. \tag{2.4}$$

No caso discreto, a função de risco é definida por:

$$h(t) = P(T = t|T \geq t). \tag{2.5}$$

Essa função é definida no intervalo [0,1] e assume valor zero quando t é negativo ou não inteiro. Ao considerar a relação entre as funções definida em (2.5), no caso discreto a relação é definida por:

$$h(t) = \frac{P(T = t)}{P(T > t) + P(T = t)} = \frac{p(t)}{S(t) + p(t)}. \tag{2.6}$$

2.2 Distribuições de probabilidade

2.2.1 Distribuição Kumaraswamy

A distribuição escolhida para este trabalho foi proposta por Kumaraswamy (1980). Levando seu próprio nome, a distribuição Kumaraswamy possui aplicações em diversas áreas devido a sua flexibilidade. A função de densidade de probabilidade Kumaraswamy é definida por:

$$g(t; a, b) = abt^{a-1}(1 - t^a)^{b-1}, \quad 0<t<1, \tag{2.7}$$

e sua respectiva função de distribuição acumulada é:

$$G(t; a, b) = 1 - (1 - t^a)^b, \quad 0<t<1, \tag{2.8}$$

em que $a > 0$ e $b > 0$ são os parâmetros de forma.

Uma nova classe de distribuições é proposta por Cordeiro e Castro (2010), chamada de Kumaraswamy generalizada. A formulação da distribuição K-G considera uma função de densidade de probabilidade (fda) $G(t)$ arbitrária, de forma que sua fda $F(t)$ e fdp $f(t)$ são, respectivamente:

$$F(t; a, b) = 1 - [1 - G(t)^a]^b, \tag{2.9}$$

$$f(t; a, b) = abg(t)G(t)^{a-1}[1 - G(t)^a]^{b-1}, \tag{2.10}$$

em que a>0 e b>0 são dois parâmetros de forma que introduzem a assimetria e flexibilização dos pesos das caudas, (FACHINI-GOMES et al., 2018).

2.2.2 Distribuição Log-Logística

A distribuição Log-Logística é um modelo de probabilidade importante para a Análise de Sobrevivência. Devido a sua função de densidade de probabilidade apresentar caudas pesadas, o comportamento de sua função de sobrevivência se adéqua a uma variabilidade de dados. A função de densidade da distribuição Log-Logística é dada por:

$$g(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left[1 + \left(\frac{t}{\alpha}\right)^\gamma\right]^{-2}, \tag{2.11}$$

e sua respectiva função de distribuição acumulada é:

$$G(t) = \frac{1}{(1 + (t/\alpha)^{-\gamma})} = \frac{t^\gamma}{t^\gamma + \alpha^\gamma}, \tag{2.12}$$

em que os parâmetros $\alpha > 0$ e $\gamma > 0$ são os parâmetros de escala e de forma respectivamente.

2.2.3 Distribuição Kumaraswamy Log-Logística

A distribuição Kumaraswamy generalizada considera uma função de distribuição de probabilidade arbitrária $G(t)$. Essa função pode ser qualquer função de distribuição estatística, preferencialmente distribuições aplicadas à Análise de Sobrevivência. Neste trabalho, a distribuição utilizada é a distribuição Log-Logística. Assim, a aplicação na função Kumaraswamy generalizada tem como resultado a distribuição Kumaraswamy Log-Logística (KLL).

Dada as distribuições obtidas anteriormente nas subseções 2.2.1 e 2.2.2, a função de densidade Kumaraswamy-Log-Logística é dada por:

$$f_{kll}(t) = ab \left(\frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \left[1 + \left(\frac{t}{\alpha}\right)^\gamma\right]^{-2}\right) \left(\frac{t^\gamma}{t^\gamma + \alpha^\gamma}\right)^{a-1} \left[1 - \left(\frac{t^\gamma}{t^\gamma + \alpha^\gamma}\right)^a\right]^{b-1}, \tag{2.13}$$

e sua respectiva função de distribuição acumulada é:

$$F_{kll}(t) = 1 - \left[1 - \left(\frac{t^\gamma}{t^\gamma + \alpha^\gamma}\right)^a\right]^b, \tag{2.14}$$

em que $\alpha > 0$ é o parâmetro de escala, $\gamma > 0$, $a > 0$ e $b > 0$ são os parâmetros de forma da distribuição Kumaraswamy-Log-Logística contínua.

2.3 Discretização

Uma variável aleatória é definida como uma quantidade associada a cada possível resultado do espaço amostral. Se a variável aleatória é enumerável, ela é considerada uma

variável aleatória discreta. Se a variável aleatória assume qualquer valor em um intervalo de números reais, ela é denominada uma variável aleatória contínua (CARDIAL, 2017).

A variável aleatória estudada em análise de sobrevivência, o tempo, é uma variável contínua, pois sua escala normalmente é dada em dias e em horas. O tempo de ocorrência de um certo evento de uma observação raramente será igual ao tempo de ocorrência de alguma outra observação. Entretanto, em alguns estudos não é possível obter o registro do tempo de ocorrência do evento de forma precisa. Por exemplo, em casos em que o controle de saúde de um paciente é feito de forma mensal, o registro será feito em intervalos mensais. Assim, indivíduos que na forma contínua teriam tempos de falha distintos, na forma de registro intervalar, passam a ocupar o mesmo espaço de tempo.

Todos os modelos de variáveis contínuas podem ser usados para gerar modelos discretos agrupando os tempos em intervalos unitários. Como visto em (NAKANO; CARRASCO, 2006): a variável discreta é dada por $T = [X]$, em que $[X]$ representa "a parte inteira de X (maior inteiro menor ou igual a X)". A função de probabilidade de T pode ser escrita como :

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq X < t + 1) \\ &= P(X < t + 1) - P(X \leq t) \\ &= F_X(t + 1) - F_X(t) \quad \text{t=0,1,2...,} \end{aligned} \tag{2.15}$$

em que $F_X(.)$ representa a função de distribuição acumulada da distribuição contínua.

2.4 Método de Máxima Verossimilhança

Considerando uma amostra aleatória observada $t_1, t_2, .., t_n$ de uma variável aleatória discreta T, o método de máxima verossimilhança permite estimar os parâmetros das distribuições estudadas. Considerando um vetor de parâmetros θ, a função de verossimilhança para este vetor é dada por (SANTOS, 2018):

$$L(\theta) = \prod_{i=1}^{n} p(t_i; \theta). \tag{2.16}$$

Assim, é possível encontrar o valor de θ que maximiza $L(\theta)$, ou seja, o valor de θ que maximiza a probabilidade da amostra observada ocorrer.

Todavia em Análise de Sobrevivência, existem casos em que alguns indivíduos não venham a experimentar o evento de interesse, as chamadas observações censuradas definidas em 2.1.1. Nesse caso, as observações deverão ser divididas entre observações

censuradas e observações não censuradas. Assim, a função de verossimilhança com a presença de censura é dada por:

$$L(\theta) \propto \prod_{i=1}^{n} p(t_i;\theta)^{\delta_i} S(t_i;\theta)^{1-\delta_i}, \tag{2.17}$$

em que a contribuição das observações censuradas é dada pela função de sobrevivência e a contribuição das observações não censuradas é dada pela função de probabilidade, ambas provenientes de qualquer distribuição de probabilidade.

3 Metodologia

Com o objetivo de propor uma nova classe de distribuição de probabilidade para analisar dados de Análise de Sobrevivência, ao considerar que o tempo é uma variável aleatória discreta, nesta seção será definida a distribuição Kumaraswamy Log-Logística discreta, bem como o banco de dados utilizado neste trabalho.

3.1 Material

De acordo com Ribeiro (2018), "A política Plano Diretor é o instrumento básico para o desenvolvimento de um município, instituído pela Constituição Federal de 1988. O objetivo principal desta política é instruir tanto o poder público quanto a iniciativa privada em relação ao desenvolvimento e crescimento urbano e rural. As Zonas Especiais de Interesse Social (ZEIS) surgiram a partir da década de 1980 e são definidas como a parcela de área urbana instituída pelo Plano Diretor ou definida por outra lei municipal, destinada preponderantemente à população de baixa renda através de urbanização de imóveis públicos, aprovação de loteamentos ou desmembramentos e regularização de núcleos urbanos informais consolidados."

O conjunto de dados selecionado para a análise desses dados foi disponibilizado através de uma parceria com o Instituto de Ciências Políticas da UnB, as Zonas Especiais de Interesse Social (ZEIS). A variável resposta nesse estudo é o tempo até a adesão à política ZEIS do município. A não adesão da política no intervalo de tempo estudado corresponde à censura. Neste conjunto de dados, todas as censuras estão no final do estudo, ou seja, as censuras são à direita do tipo I.

A partir das eleições de 1996, uma variável importante no estudo passou a ser registrada, a variável "Competição Política". Dessa forma, o período analisado nesse estudo será de 1997 a 2015, assim o intervalo de tempo datado é anual, sendo o primeiro tempo 1997($t_i = 0$) e o último tempo 2015 ($t_i = 18$).

Para alguns municípios, o ano de adesão apresentou valores como "Não soube informar" e "recusa" e também alguns valores inferiores à 1997. Optou-se por retirar essas observações do banco, para não interferir na análise. Após essa exclusão, o banco ficou com 5365 observações que serão efetivamente utilizadas na análise.

A fim de verificar a influência de outros fatores no tempo de adesão à política, 7 covariáveis que representam aspectos políticos e geográficos dos municípios foram analisadas. A inclusão é realizada referente ao período imediato para os municípios que falharam, ou seja, as informações das covariáveis correspondem ao período em que ocorreu falha. As

covariáveis, segundo Ribeiro (2018), são descritas por:

- Margem de vitória: apresenta o percentual de vitória do candidato eleito em relação ao segundo colocado, medido a partir de 1996 e com periodicidade de quatro em quatro anos, obrigatoriamente em anos eleitorais;
- NEP: número efetivo de partidos políticos, medido a partir do ano 2000 e com periodicidade de quatro em quatro anos, obrigatoriamente em anos eleitorais;
- Região: representa a região administrativa do Brasil através de 4 variáveis *dummies*, em que a categoria de referência será a região Sul;
- Conselho de Política Urbana: variável binária, que indica se foi criado no município um conselho de políticas urbanas no período anterior à falha do município, com medição a partir do momento da sua criação independente da existência de período eleitoral;
- População: devido ao tamanho dos municípios, a diferença entre os valores correspondentes ao tamanho da sua população e as demais variáveis, será considerado o logaritmo da população de cada município, obtida por meio do censo demográfico de 2000 e 2010;
- Prefeito Reeleito: variável binária que indica se o prefeito foi reeleito no ano interior ou igual ao que o município falhou, medida em anos eleitorais a partir de 1996;
- Ano Eleitoral: variável binária que indica se o ano de falha também é ano eleitoral no município.

3.2 Distribuição Kumaraswamy Log-Logística discreta

A distribuição Kumaraswamy-Log-Logística discreta pode ser obtida através do método da discretização definido na seção 2.3. Dessa forma, ao considerar a equação (2.15) e a distribuição Kumaraswamy Log-Logística definida na equação (2.14), a função de probabilidade de uma variável aleatória Kumaraswamy Log-Logística discreta (KLLD) é definida por:

$$p_{klld}(t) = \left[1 - \left(\frac{t^\gamma}{t^\gamma + \alpha^\gamma}\right)^a\right]^b - \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \quad \text{t=0,1,2...,} \tag{3.1}$$

e sua respectiva função de densidade acumulada (fda) é:

$$F_{klld} = 1 - \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \quad \text{t=0,1,2...,} \tag{3.2}$$

em que os parâmetros $\alpha > 0$ e $\gamma > 0$ são os parâmetros de escala e de forma da distribuição Log-Logística. Os parâmetros $a > 0$ e $b > 0$ são os parâmetros de forma da distribuição Kumaraswamy Generalizada. Quando $a = 1$ e $b = 1$, tem-se a função Log-Logística discreta.

A função de sobrevivência e de risco associadas a distribuição Kumaraswamy Log-Logística discreta são definidas, respectivamente, por:

$$S_{klld}(t) = \left[1 - \left(\frac{(t+1)^{\gamma}}{(t+1)^{\gamma} + \alpha^{\gamma}}\right)^{a}\right]^{b} \quad \text{t=0,1,2...,} \tag{3.3}$$

e

$$h_{klld}(t) = 1 - \frac{\left[1 - \left(\frac{(t+1)^{\gamma}}{(t+1)^{\gamma}+\alpha^{\gamma}}\right)^{a}\right]^{b}}{\left[1 - \left(\frac{t^{\gamma}}{t^{\gamma}+\alpha^{\gamma}}\right)^{a}\right]^{b}} \quad \text{t=0,1,2...} \tag{3.4}$$

A definição completa das funções definidas nas equações (3.1) à (3.4) está no Apêndice A deste trabalho.

Nas Figuras 2 e 3, apresenta-se o comportamento das funções de probabilidade e de risco da Kumaraswamy Log-Logística Discreta (KLLD) para diferentes valores dos parâmetros. Pode-se observar na Figura 3 que a função de risco da distribuição KLLD assume forma decrescente,crescente e unimodal. Um estudo mais aprofundado sobre as formas da função de risco da distribuição KLLD será desenvolvido em trabalhos futuros. Os parâmetros da nova distribuição podem ser estimados pelo método de máxima verossimilhança definido na seção 2.4.

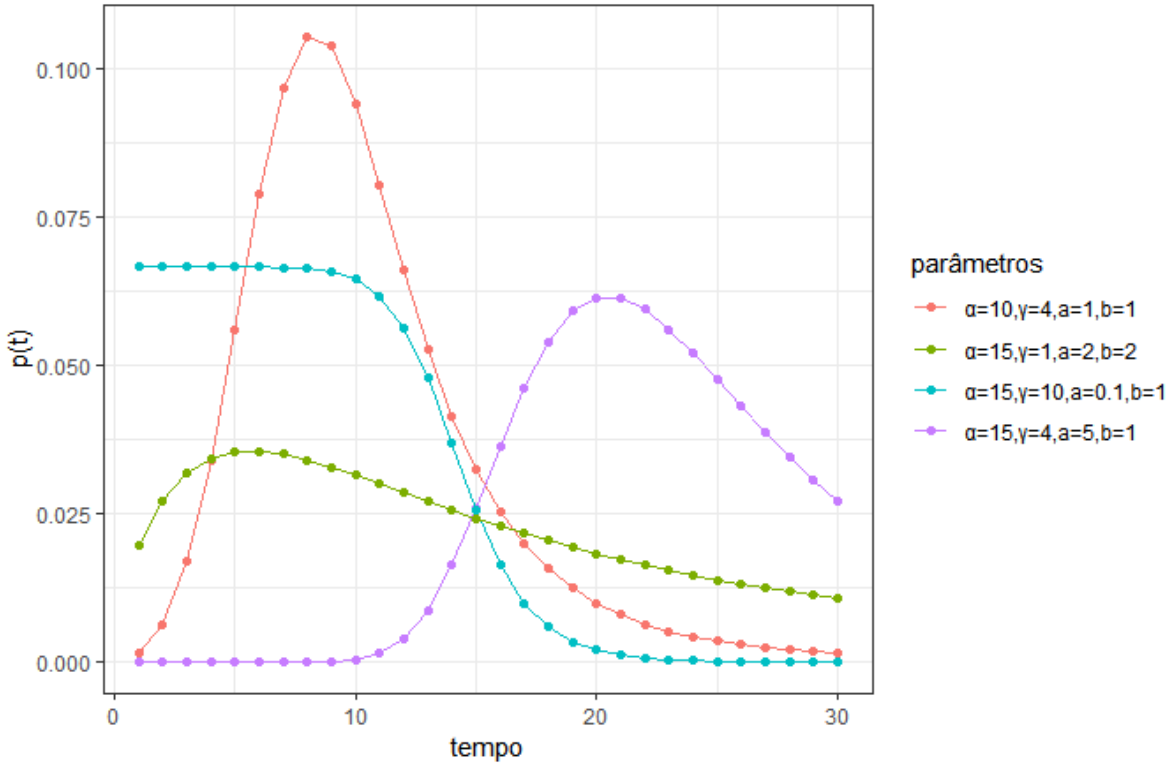

Figura 2 – Função de probabilidade KLLD para diferentes valores de α, γ, a e b.

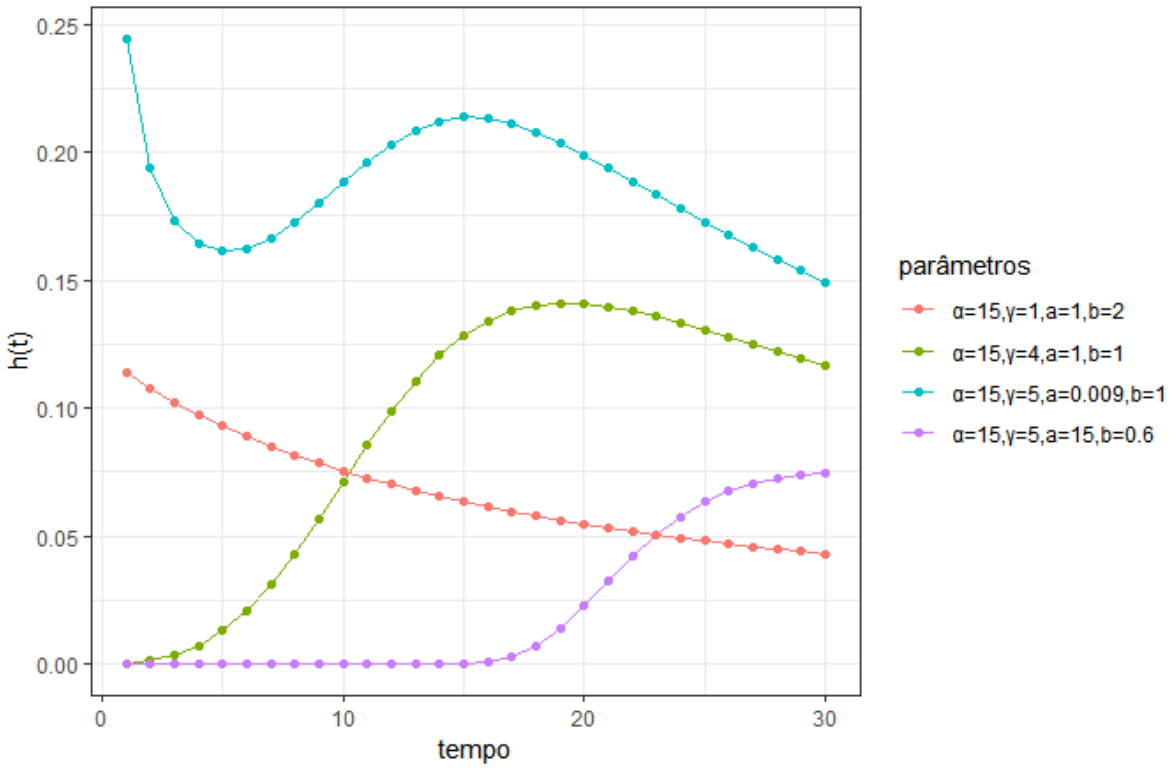

Figura 3 – Função de risco da distribuição KLLD para diferentes valores de α, γ, a e b.

3.3 Modelo de Regressão Kumaraswamy Log-Logístico Discreto

Na maior parte dos estudos em Análise de Sobrevivência, existem variáveis que podem influenciar o tempo de falha de um indivíduo. Essas variáveis, se correlacionadas com o tempo de falha, são chamadas de covariáveis e essa relação pode ser melhor analisada através do modelo de regressão.

Para o estudo em questão, a relação das covariáveis do modelo de regressão será feita através da reparametrização de um dos parâmetros da distribuição de probabilidade.

Segundo Santos (2018) e Lawless (2011), ao considerar um vetor de covariáveis, $\mathbf{x^T} = (1, x_1, ..., x_p)$ e seu respectivo vetor de coeficientes de regressão, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$, é possível conectar as variáveis explicativas à variável resposta ao utilizar uma função de ligação $g(.)$.

Para um conjunto p de variáveis explicativas, ao considerar pelo menos um parâmetro, θ_1, pertencente ao vetor de parâmetros $\boldsymbol{\theta}$ do modelo, a reparametrização é definida por:

$$\boldsymbol{\theta} = g(\boldsymbol{\eta}), \tag{3.5}$$

em que $\boldsymbol{\eta} = \mathbf{x^T}\boldsymbol{\beta}$ é o preditor linear. A reparametrização para o modelo de regressão em questão será feita pelo parâmetro de escala α , positivo, e a função$g(.)$. Sendo assim, tem-se a reparametrização: $\boldsymbol{\alpha} = g(\boldsymbol{\eta}) = exp(\boldsymbol{x^T\beta})$.

Considerando a distribuição Kumaraswamy Log-Logística Discreta (KLLD) definida na seção 3.2, o modelo de regressão KLLD pode ser expresso por:

$$p_{klld}(t) = \left[1 - \left(\frac{t^\gamma}{t^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\right)^a\right]^b - \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\right)^a\right]^b \quad \text{t=0,1,2....} \tag{3.6}$$

Assim, as funções de Sobrevivência e de risco são definidas respectivamente por:

$$S_{klld}(t) = \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\right)^a\right]^b \quad \text{t=0,1,2...,} \tag{3.7}$$

e

$$h_{klld}(t) = 1 - \frac{\left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\right)^a\right]^b}{\left[1 - \left(\frac{t^\gamma}{t^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\right)^a\right]^b} \quad \text{t=0,1,2...,} \tag{3.8}$$

em que $\gamma > 0$, $a > 0$ e $b > 0$ são os parâmetros de forma da distribuição Kumaraswamy-Log-Logística discreta, $\boldsymbol{x^T}$ o vetor de covariáveis e $\boldsymbol{\beta}$ o vetor dos coeficientes de regressão.

3.4 Resíduos de Cox-Snell

Os resíduos de Cox-Snell representam um método para examinar a qualidade do ajuste global do modelo. Este tipo de resíduo está entre os mais utilizados em Análise de Sobrevivência e é determinado por:

$$\hat{e_i} = \hat{H}(t_i/\boldsymbol{x_i}), \tag{3.9}$$

em que $\hat{H}(.)$ é a função de taxa de falha acumulada obtida a partir do modelo ajustado.

Segundo Colosimo e Giolo (2006), pode-se utilizar a análise gráfica para verificar a qualidade de ajuste do modelo através da curva de sobrevivência do resíduo do modelo estimado pelo método de Kaplan-Meier, $\hat{S}_{km}(e_i)$, e a curva de sobrevivência da distribuição exponencial padrão dos resíduos, $\hat{S}(e_i) = \exp(-e_i)$.

Ao considerar o modelo de regressão proposto definido na seção 3.3, o resíduo de Cox-Snell é representado por:

$$\hat{e_i} = \hat{H}(t_i/\boldsymbol{x_i}) = -\log(\hat{S}(t_i/\boldsymbol{x_i})) = -\log\left(\left[1 - \left(\frac{(t_i+1)^\gamma}{(t_i+1)^\gamma + \exp(\boldsymbol{x^T\beta}+)^\gamma}\right)^a\right]^b\right). \tag{3.10}$$

4 Resultados

Neste capítulo será descrito a análise realizada no banco de dados referentes à adesão à política ZEIS dos municípios brasileiros através da distribuição Kumaraswamy Log-Logística para dados discretos.

4.1 Análise descritiva

A estimação da função de sobrevivência de Kaplan-Meier é representada na Figura 4. Observa-se pela figura que aproximadamente 40% dos municípios falharam, ou seja, que aderiram à política ZEIS nos anos observados. Nota-se também que as censuras ocorrem apenas no final do estudo, o que é esperado, sabendo da dificuldade para um município sair do estudo.

A Figura 5 representa a Função de Risco Acumulada para os dados em estudo e a Figura 6 apresenta a função de taxa de falha empírica através do gráfico de tempo total em teste (gráfico TTT). Por meio dessas Figuras, verifica-se que a função de risco para o tempo em estudo tem forma concava, ou seja, a função taxa de falha é monotonicamente crescente.

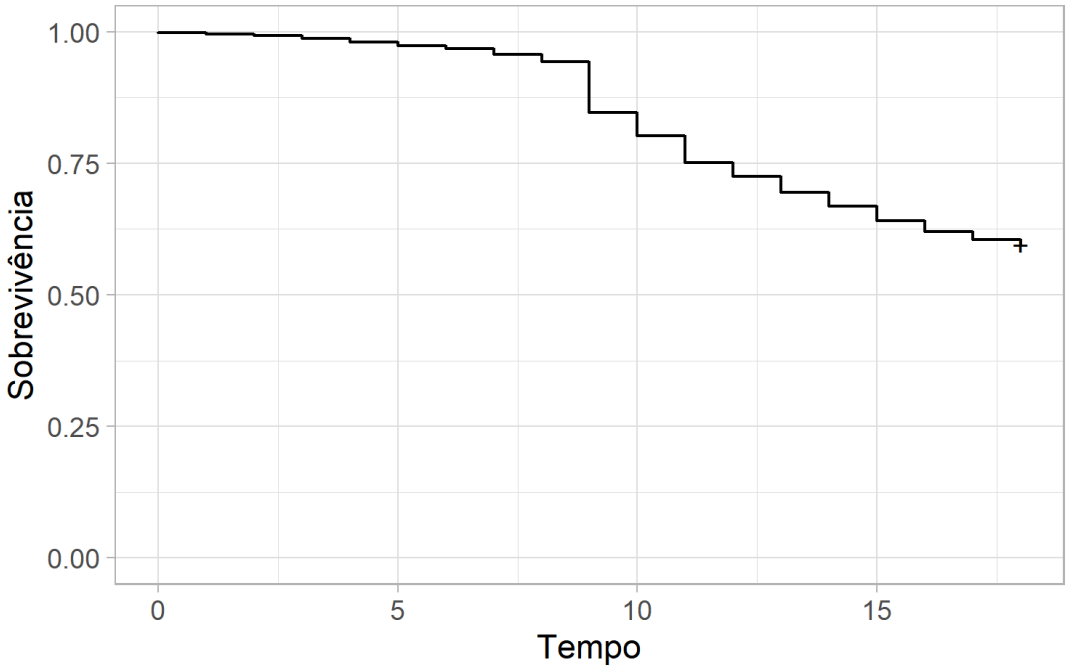

Figura 4 – Estimativa da função de sobrevivência de Kaplan-Meier

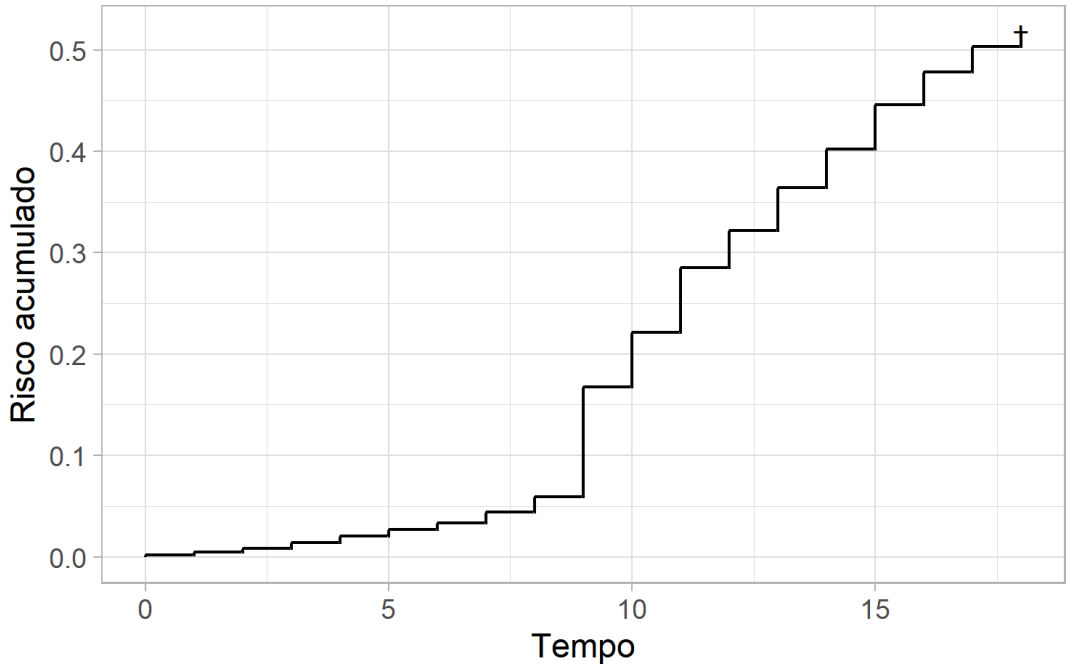

Figura 5 – Função de Risco acumulada

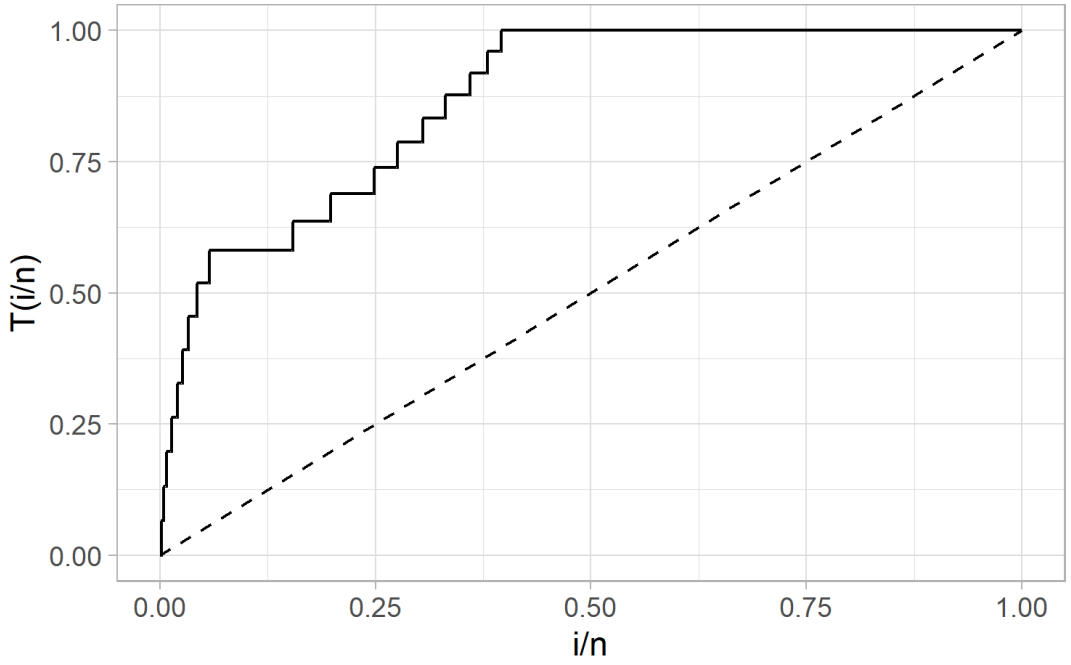

Figura 6 – Tempo Total em Teste (gráfico TTT)

As Figuras 7 a 10 representam graficamente a estimação da função de sobrevivência de Kaplan-Meier para as 4 covariáveis categóricas candidatas ao modelo escolhido. São elas: Região, Conselho de Política Urbana, Prefeito Reeleito e Ano Eleitoral.

Observa-se na Figura 7 que a região Sul possui mais falhas que as outras regiões e está mais distante graficamente. Como definido previamente e confirmado através deste gráfico, a região Sul será usada como a referência na inclusão desta covariável no modelo.

Na Figura 8, todos os municípios que possuem Conselho de Política Urbana falharam, ou seja, não houve nenhuma censura. Isso é um indicativo que esta covariável pode explicar o tempo até a adesão da política ZEIS.

A covariável Prefeito Reeleito possui as curvas de sobrevivência muito próximas. Graficamente representada na Figura 9, esta covariável não aparenta descrever bem a adesão à política.

É fácil ver na Figura 10 que a covariável Ano Eleitoral possui os intervalos de falha do "Sim a cada 4 tempos correspondendo respectivamente as eleições dos anos de 2000, 2004, 2008 e 2012. Assim como a covariável Conselho de Política Urbana, todos os municípios que aderiram em período eleitoral falharam, porém, neste caso a interpretação não pode ser a mesma pois o estudo não acabou em um período eleitoral para apresentar as censuras dos municípios que não falharam neste período.

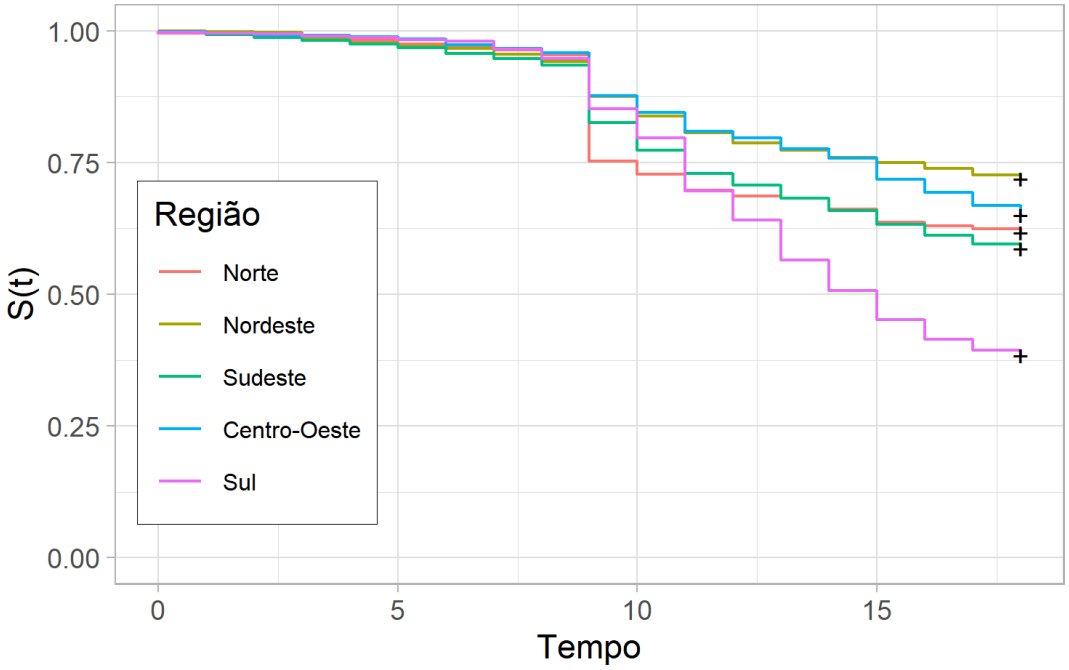

Figura 7 – Estimação da função de sobrevivência de Kaplan-Meier pela covariável Região

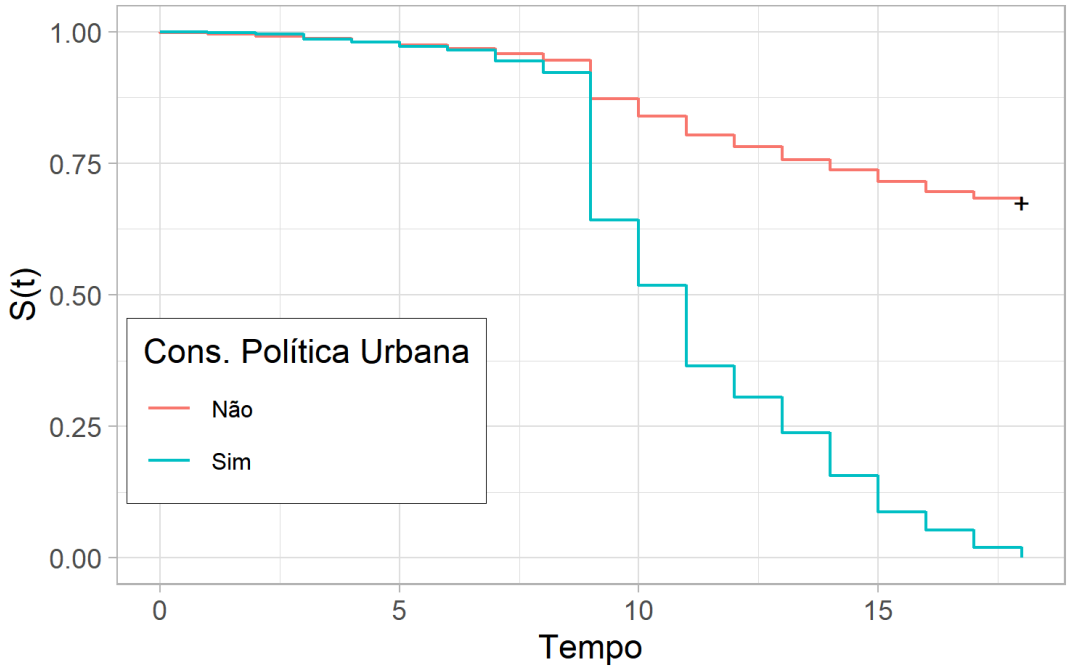

Figura 8 – Estimação da função de sobrevivência de Kaplan-Meier pela covariável Conselho de Política Urbana

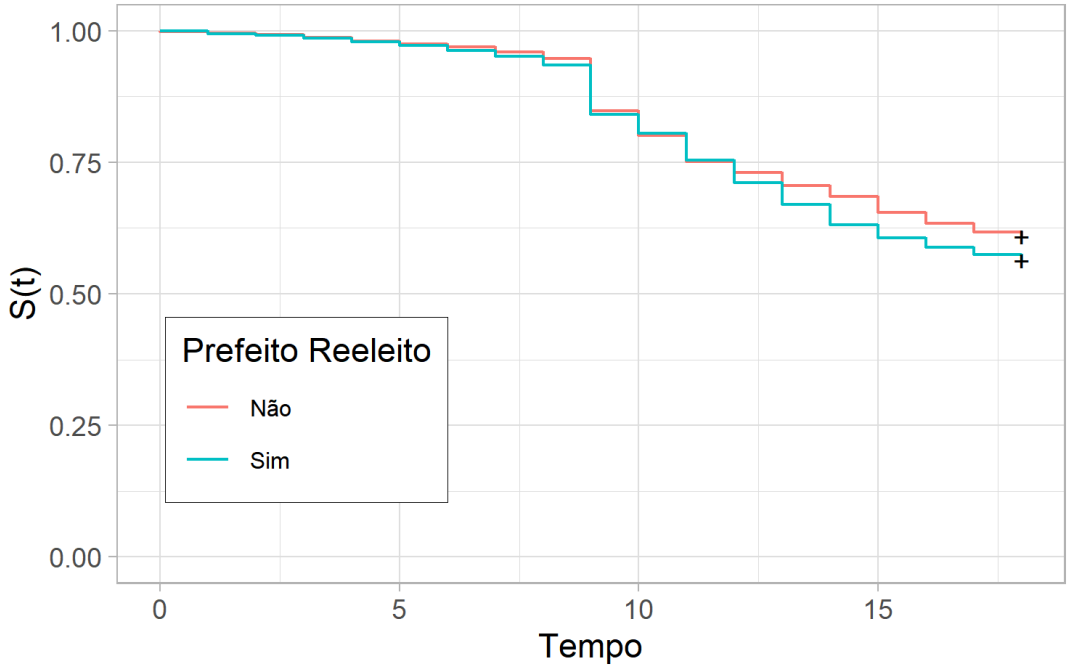

Figura 9 – Estimação da função de sobrevivência de Kaplan-Meier pela covariável Prefeito Reeleito

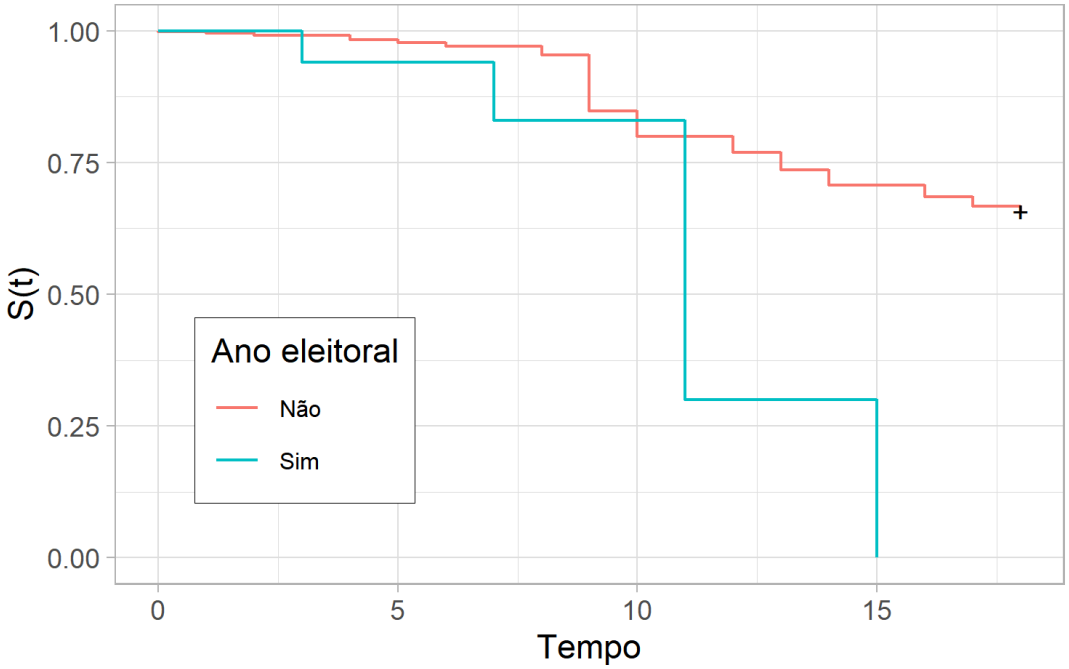

Figura 10 – Estimação da função de sobrevivência de Kaplan-Meier pela covariável Ano eleitoral

4.2 Ajuste da distribuição Kumaraswamy Log-Logística Discreta

Para verificar a qualidade de ajuste da distribuição Kumaraswamy Log-Logística Discreta aos dados, foi realizado com o *software* R o ajuste da distribuição através da função *optim*, pois o *software* R não possui um pacote pronto para esta função específica. A programação referente à implementação desta função de probabilidade e sua otimização está representada no Apêndice B. O ajuste da distribuição KLLD no gráfico de Kaplan-Meier é representado pela Figura 11.

Pode-se observar graficamente que a distribuição escolhida se ajusta bem aos dados em estudo. As estimativas dos parâmetros a, b, α e γ são apresentadas na Tabela 1.

Tabela 1 – Estimativas dos parâmetros da distribuição KLLD

Parâmetro	Estimativa	Variância	Erro Padrão
γ	42,549	1,774e-07	0,00042
α	8,9908	3,595e-07	0,00059
a	0,0186	3,529e-07	0,00059
b	0,0150	1,078e-07	0,00032

Pela Tabela 1, nota-se que embora o erro padrão seja pequeno, os parâmetros α e γ possuem uma estimativa elevada, portanto, um modelo de regressão se mostra útil afim de se obter estimativas para considerar os possíveis fatores que influenciem na adoção da política.

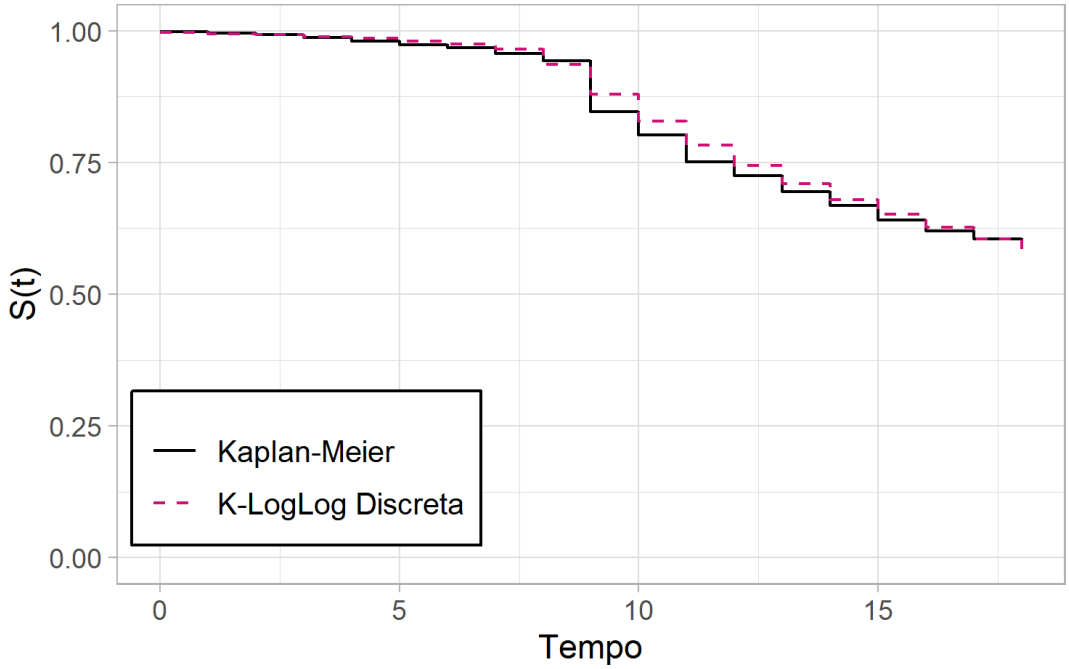

Figura 11 – Ajuste da distribuição KLLD no gráfico de Kaplan Meier dos dados

O primeiro passo para analisar o comportamento das covariáveis é ajustar um modelo de regressão para cada uma das covariáveis separadamente. Dessa forma, as estimativas e os p-valores de cada covariável sozinha permitem a seleção para a inclusão no modelo, uma covariável por vez. A Tabela 2 apresenta as estimativas de cada covariável a ser considerada no modelo.

Tabela 2 – Estimativas dos coeficientes para o modelo de regressão KLLD para apenas uma variável

Variável	Estimativa	Erro Padrão	P-valor
Cons. Política urbana	-0,7818	0,0204	<0,0001
NEP	-0,1716	0,0099	<0,0001
Log(POP)	-0,2555	0,0113	<0,0001
Margem de vitória	-0,0922	0,0537	0,0861
Prefeito reeleito	-0,0067	0,0021	0,0013
Ano eleitoral	-0,7397	0,0214	<0,0001
Sudeste	0,1843	0,0326	<0,0001
Centro-Oeste	0,3176	0,0475	<0,0001
Norte	0,2078	0,0487	<0,0001
Nordeste	0,4243	0,0340	<0,0001

À nível de significância de 5%, apenas a variável ”Margem de vitória” não foi significativa. Dessa forma, o critério de inclusão de cada covariável no modelo foi realizado primeiro pela inclusão das variáveis contínuas, e em seguida com as variáveis discretas, analisando passo a passo a significância de cada covariável e alterações nas demais covariá-

veis presentes. Por fim, o modelo final obtido excluiu apenas a covariável "Prefeito reeleito" do modelo . As estimativas deste modelo estão representadas na Tabela 3.

Tabela 3 – Estimativas dos coeficientes para o modelo de regressão Kumaraswamy Log-Logístico Discreto para o modelo completo

Variável	Estimativa (β_i)	Erro Padrão	P-valor
Intercepto	5,3551	0,1205	<0,0001
Cons. Política urbana	-0,3916	0,0281	<0,0001
NEP	-0,0917	0,0179	<0,0001
Log(POP)	-0,1487	0,0097	<0,0001
Margem de vitória	-0,5482	0,0439	<0,0001
Ano eleitoral	-0,5216	0,0277	<0,0001
Sudeste	0,2337	0,0268	<0,0001
Centro-Oeste	0,3239	0,0426	<0,0001
Norte	0,1136	0,0381	0,0029
Nordeste	0,3768	0,0296	<0,0001
γ	1,4493	0,1041	-
a	1,9471	0,1694	-
b	9,0371	2,4443	-

O parâmetro α corresponde à reparametrização para a inclusão das covariáveis como visto na seção 3.2. Assim, as covariáveis e suas estimativas são incluídas no modelo através do parâmetro α:

$$\alpha = \exp(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 * Cons.Pol.Urbana + ... + \beta_9 * Nordeste) \tag{4.1}$$

As variáveis categóricas "Conselho de Política urbana" e "Ano eleitoral" apresentam as estimativas dos parâmetros negativas, isso confirma a representação gráfica das Figuras 8 e 10 de que os municípios que possuiam o Conselho de Política urbana e estavam em ano eleitoral tem maior probabilidade de falha, ou seja, aderir à política ZEIS.

Para as variáveis contínuas "NEP", "Log(População)" e "Margem de vitória" as estimativas dos seus respectivos parâmetros também tem o sinal negativo. Nestes 3 casos, a interpretação é de quanto maior o valor dessas variáveis, maior será a chance de adesão à política.

Como a região Sul foi a referência por ser a curva mais distante e possuir mais falhas, todas as outras regiões possuem estimativas positivas, como esperado devido ao comportamento da Figura 7. Assim as outras regiões tem menos chance de aderir a política em relação à região sul.

Apesar de a variável "Prefeito Reeleito" ser estatisticamente significativa individualmente, no conjunto com as outras covariáveis ela não se manteve significante no modelo, assim como sua representação gráfica pareceu indicar. Portanto, sua estimativa não está representada e esta variável foi excluída do modelo.

Por fim, o modelo final incluiu seis das sete possíveis variáveis explicativas dos dados em questão. Esse fato é um indicativo de que a distribuição Kumaraswamy Log-Logística se ajusta bem aos dados.

4.3 Análise de Resíduos

Na análise de resíduos, os métodos gráficos são essenciais. Para analisar os resíduos do modelo escolhido, utiliza-se uma medida de qualidade de ajuste, os resíduos de Cox-Snell:

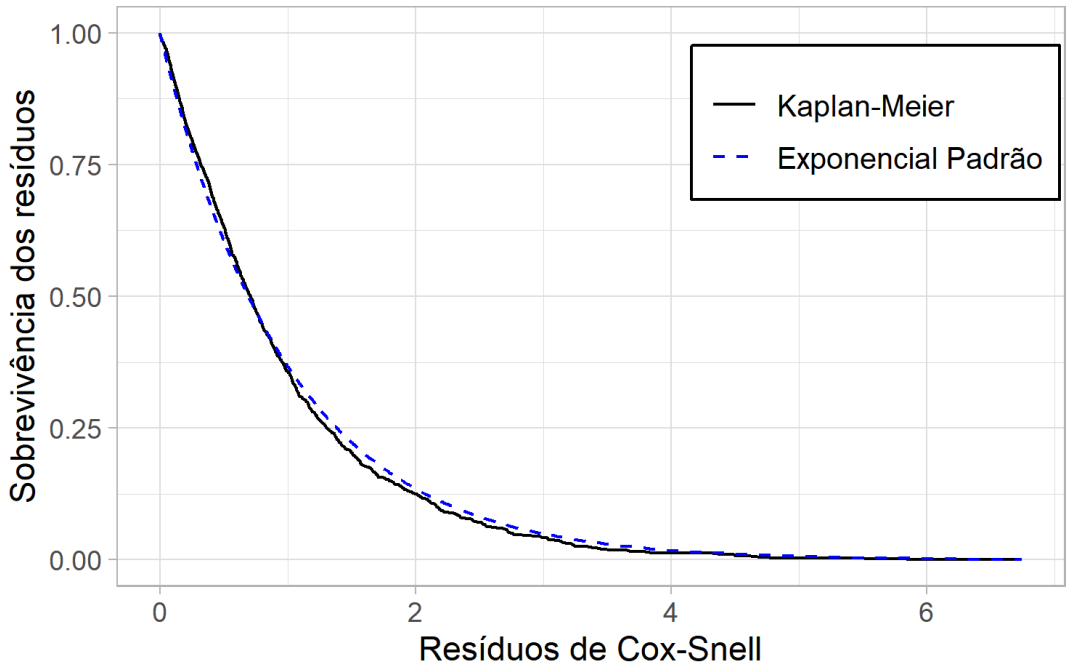

Figura 12 – Resíduos de Cox-Snell e Exponencial Padrão

Como visto na seção 3.4 o gráfico da estimação da função de sobrevivência dos resíduos de Cox-Snell deve se aproximar de uma exponencial padrão. A Figura 12 representa a estimativa da função de Sobrevivência pelo Kaplan-Meier e a estimativa da função de Sobrevivência da distribuição exponencial padrão.

Pelo gráfico dos resíduos de Cox-Snell o modelo está bem ajustado, a distribuição Kumaraswamy Log-Logística para dados discretos se comporta de forma ótima aos dados de adesão à Política ZEIS.

5 Conclusões

O presente trabalho teve como objetivo propor e estudar uma nova classe de distribuição de probabilidade para analisar dados de Análise de Sobrevivência quando a variável resposta é discreta. Neste caso, foi proposta a distribuição Kumaraswamy Log-Logística devido à alta flexibilidade da distribuição Kumaraswamy Log-Logística para dados contínuos.

Os dados escolhidos para o estudo foram utilizados em um estudo prévio utilizando a distribuição Log-Logística Discreta visto em Ribeiro (2018). Em uma breve comparação, a distribuição Kumaraswamy Log-Logística Discreta se mostrou melhor ajustada do que a distribuição Log-Logística Discreta, devido ao acréscimo dos dois parâmetros de forma.

De forma geral, a distribuição Kumaraswamy Log-Logística Discreta se ajusta de forma ótima aos dados, isso fica claro através dos métodos gráficos utilizados. As covariáveis ajustadas no modelo de regressão também demonstram que além de ser utilizada quase toda a informação disponível pelo banco de dados, as estimativas do modelo correspondem ao que foi visto anteriormente pelos métodos gráficos.

Os resíduos de Cox-Snell confirmaram a qualidade de ajuste do modelo. Dessa forma, a distribuição escolhida e o modelo de regressão são fortes candidatos para utilização em estudos futuros de Análise de Sobrevivência em dados discretos.

No entanto, neste estudo foram encontradas algumas instabilidades computacionais. Por conseguinte, em trabalhos futuros seria de grande interesse estudar a fundo as propriedades matemáticas da distribuição KLLD e realizar simulações computacionais a fim de validar o método e justificar melhor a utilização do modelo.

Referências

CARDIAL, M. R. P. Distribuição weibull discreta exponenciada para dados com presença de censura: uma abordagem clássica e bayesiana. 2017. Citado na página 16.

COLOSIMO, E.; GIOLO, S. Análise de sobrevivência aplicada. 1ª edição. *São Paulo: Editora Edgard Blücher*, 2006. Citado na página 9.

FACHINI-GOMES, J. B. et al. The bivariate kumaraswamy weibull regression model: a complete classical and bayesian analysis. *Communications for Statistical Applications and Methods*, Korean Statistical Society, v. 25, n. 5, p. 523–544, 2018. Citado na página 14.

FACHINI, J. B. *Modelos de regressão com e sem fração de cura para dados bivariados em análise de sobrevivência*. Tese (Doutorado) — Universidade de São Paulo, 2011. Nenhuma citação no texto.

HOEL, P. G.; PORT, S. C.; STONE, C. J. Introdução à teoria da probabilidade. *Rio de janeiro: Interciência*, 1978. Nenhuma citação no texto.

KUMARASWAMY, P. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, Elsevier, v. 46, n. 1-2, p. 79–88, 1980. Nenhuma citação no texto.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. [S.l.]: John Wiley & Sons, 2011. v. 362. Nenhuma citação no texto.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. *Trends in Applied and Computational Mathematics*, v. 7, n. 1, p. 91–100, 2006. Citado na página 16.

RIBEIRO, P. E. d. A. Estudo das zonas especiais de interesse social através da metodologia de análise de sobrevivência. 2018. Nenhuma citação no texto.

SANTOS, D. F. d. Modelo de regressão log-logístico discreto com fração de cura para dados de sobrevivência. 2018. Citado na página 16.

Apêndices

APÊNDICE A – Cálculos

Discretização da distribuição Kumaraswamy Log-Logística Contínua:

$$\begin{aligned} p_(x) &= P(X = x) \\ &= F(x+1) - F(x) \\ &= \left(1 - \left[1 - \left(\frac{(x+1)^\gamma}{(x+1)^\gamma + \alpha^\gamma}\right)^a\right]^b\right) - \left(1 - \left[1 - \left(\frac{x^\gamma}{x^\gamma + \alpha^\gamma}\right)^a\right]^b\right) \\ &= \left[1 - \left(\frac{x^\gamma}{x^\gamma + \alpha^\gamma}\right)^a\right]^b - \left[1 - \left(\frac{(x+1)^\gamma}{(x+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \end{aligned}$$

Função Kumaraswamy-Log-Logística discreta acumulada:

$$\begin{aligned} F_{kllda} &= \sum_{x=0}^{t} p(x) \\ &= \sum_{x=0}^{t} \left[1 - \left(\frac{x^\gamma}{x^\gamma + \alpha^\gamma}\right)^a\right]^b - \left[1 - \left(\frac{(x+1)^\gamma}{(x+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \\ &= 1^b - \cancel{\left[1 - \left(\frac{1^\gamma}{(1^\gamma + \alpha^\gamma}\right)^a\right]^b} + \\ &\cancel{\left[1 - \left(\frac{1^\gamma}{1^\gamma + \alpha^\gamma}\right)^a\right]^b} - \cancel{\left[1 - \left(\frac{2^\gamma}{2^\gamma + \alpha^\gamma}\right)^a\right]^b} + \cdots + \\ &\cancel{\left[1 - \left(\frac{t^\gamma}{t^\gamma + \alpha^\gamma}\right)^a\right]^b} - \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \\ &= 1 - \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \end{aligned}$$

Função de sobrevivência Kumaraswamy-Log-Logística:

$$\begin{aligned} S_{klld} = 1 - F_X(t+1) &= 1 - \left(1 - \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \alpha^\gamma}\right)^a\right]^b\right) \\ &= \left[1 - \left(\frac{(t+1)^\gamma}{(t+1)^\gamma + \alpha^\gamma}\right)^a\right]^b \quad \text{t=0,1,2...} \end{aligned}$$

Função de risco Kumaraswamy-Log-Logística:

$$
\begin{aligned}
h(t) &= \frac{p(t)}{S(t)+p(t)} \\
&= \frac{\left[1-\left(\frac{t^\gamma}{t^\gamma+\alpha^\gamma}\right)^a\right]^b - \left[1-\left(\frac{(t+1)^\gamma}{(t+1)^\gamma+\alpha^\gamma}\right)^a\right]^b}{\left[1-\left(\frac{(t+1)^\gamma}{(t+1)^\gamma+\alpha^\gamma}\right)^a\right]^b + \left[1-\left(\frac{t^\gamma}{t^\gamma+\alpha^\gamma}\right)^a\right]^b - \left[1-\left(\frac{(t+1)^\gamma}{(t+1)^\gamma+\alpha^\gamma}\right)^a\right]^b} \\
&= \frac{\left[1-\left(\frac{t^\gamma}{t^\gamma+\alpha^\gamma}\right)^a\right]^b - \left[1-\left(\frac{(t+1)^\gamma}{(t+1)^\gamma+\alpha^\gamma}\right)^a\right]^b}{\left[1-\left(\frac{t^\gamma}{t^\gamma+\alpha^\gamma}\right)^a\right]^b} \\
&= 1 - \frac{\left[1-\left(\frac{(t+1)^\gamma}{(t+1)^\gamma+\alpha^\gamma}\right)^a\right]^b}{\left[1-\left(\frac{t^\gamma}{t^\gamma+\alpha^\gamma}\right)^a\right]^b}
\end{aligned}
$$

APÊNDICE B – Códigos em R

```
## FUNÇÃO DE PROBABILIDADE KUMARASWAMY LOG-LOGÍSTICA DISCRETA
fklld <- function(t,gamma,alpha,a,b){
        prob <- (1-(((t)^gamma)/((t)^gamma + alpha^gamma))^a)^b -
         (1-(((t+1)^gamma)/((t+1)^gamma + alpha^gamma))^a)^b
        return(prob)
}

## FUNÇÃO DE SOBREVIVÊNCIA KUMARASWAMY LOG-LOGÍSTICA DISCRETA
sklld <- function(t,gamma,alpha,a,b){
        kum <- 1 - (1 - (((t+1)^gamma)/((t+1)^gamma + alpha^gamma))^a)^b
        sob <- 1- kum
        return(sob)
}

## LOG-VEROSSIMILHANÇA KUMARASWAMY LOG-LOGÍSTICA DISCRETA
lvero_klld <- function(par){
        p <- sum(censura*log(fklld(t,par[1],par[2],par[3],par[4])) +
         (1-censura)*log(sklld(t,par[1],par[2],par[3],par[4])))
        return(p)
}

ML_est <- optim(c(1,10,3,1) , lvero_klld,
control=list(fnscale=-1,maxit=1000), hessian=T)
```