**Universidade de Brasília**
**Departamento de Estatística**

# Modelo de Regressão Kumaraswamy Log-Logístico Discreto em Análise de Sobrevivência
## Uma aplicação para dados de Ciência Política

**Mateus Felipe Santos Araújo**

Relatório 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**
**2020**

**Mateus Felipe Santos Araújo**

# Modelo de Regressão Kumaraswamy Log-Logístico Discreto em Análise de Sobrevivência
## Uma aplicação para dados de Ciência Política

Orientadora:
Profa. Dra. Juliana Betini Fachini Gomes

Relatório 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**
**2020**

# Resumo

Este trabalho possui o objetivo de estudar a aplicabilidade da distribuição Kumaraswamy Log-Logística discreta para dados de Análise de Sobrevivência, comparando o respectivo modelo de regressão proposto por Simões e Silva (2019) com o modelo de regressão Log-Logístico discreto. O banco de dados utilizado é referente à política de proibição de fumar pelos estados americanos e os modelos verificam a influência de cada covariável no tempo de adesão da política. As análises foram realizadas por meio do *software R*.

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

# Abstract

This paper work has the objective of studying the discrete Kumaraswamy Log-Logistic for data in Survival Analysis aplicability, comparing the respective regression model, proposed by Simões e Silva (2019), with the discrete Log-Logistic regression model. The data base used on this paper refers to the smoke prohibition policy by the american states. The regression models verifies the influence of each variable in policy adherence time. All analysis were done using *software R*.

**Keywords:** Survival Analysis, Kumaraswamy Log-Logistic distribution, Log-Logistic distribution, Regression Model, Censored Data.

# Sumário

**1 Introdução** . . . . . 5
**2 Objetivos** . . . . . 6
2.1 Objetivo Geral . . . . . 6
2.2 Objetivos Específicos . . . . . 6
**3 Revisão de Literatura** . . . . . 7
3.1 Notação e conceitos básicos . . . . . 7
3.2 Funções . . . . . 8
3.3 Técnicas não-paramétricas . . . . . 10
3.3.1 Estimador de Kaplan-Meier . . . . . 10
3.3.2 TTT Plot e Função de Risco . . . . . 12
3.4 Distribuições de Probabilidade . . . . . 14
3.4.1 Distribuição Kumaraswamy . . . . . 14
3.4.2 Distribuição Log-Logística . . . . . 15
3.4.3 Distribuição Kumaraswamy Log-Logística . . . . . 15
3.4.4 Método da Máxima Verossimilhança . . . . . 16
3.4.5 Discretização de Distribuições Contínuas . . . . . 17
**4 Metodologia** . . . . . 19
4.1 Material . . . . . 19
4.2 Distribuição Log-Logística discreta . . . . . 21
4.3 Distribuição Kumaraswamy Log-Logística discreta . . . . . 21
4.4 Modelo de regressão Log-Logístico discreto . . . . . 22
4.5 Modelo de regressão Kumaraswamy Log-Logístico discreto . . . . . 23
**5 Resultados** . . . . . 25
5.1 Análise Descritiva . . . . . 25
5.2 Ajuste da Distribuição Log-Logística Discreta . . . . . 31
5.3 Ajuste da Distribuição Kumaraswamy Log-Logística Discreta . . . . . 33
**6 Conclusões** . . . . . 38
**Referências** . . . . . 41

# 1 Introdução

Com o crescimento da ciência de dados, algumas áreas da estatística vêm crescendo bastante nas últimas décadas, em específico, a análise de sobrevivência, a qual cresceu bastante na década de 80 devido a sua aplicação em medicina. Esta técnica considera como variável resposta o tempo até a ocorrência de um evento de interesse, denominado de tempo de falha, podendo ser o tempo até a morte de um paciente ou até a falha de um produto industrial ou até um cliente se tornar inadimplente.

Os dados de sobrevivência apresentam uma peculiaridade, que é a presença de observações parciais da resposta, denominadas de censura. Estas são observações que não apresentaram o evento de interesse até o final do estudo ou por algum motivo não puderam mais ser observadas. Apesar de serem observações "incompletas", elas devem ser contidas na análise dos dados, pois a omissão delas pode ocasionar em conclusões viesadas. Logo, uma das vantagens da técnica de análise de sobrevivência é a inclusão de dados censurados na análise.

Para descrever o tempo de falha, são utilizadas distribuições de probabilidade. Na literatura as funções mais utilizadas, como Weibull, Exponencial, Log-Logística, Log-Normal e Gamma, apresentam algumas limitações quanto a forma da função de risco dos dados. Nos últimos anos algumas distribuições para tempos contínuos foram apresentadas, entre elas, destacam-se as distribuições Weibull exponenciada (Mudholkar et al., 1995), Weibull modificada generalizada (Carrasco et al., 2008), Beta Weibull modificada (Silva et al., 2010) e a distribuição Kumaraswamy generalizada (Cordeiro e Castro, 2010), estas surgem como distribuições mais flexíveis, logo capazes de modelar formas de risco crescente, decrescente, unimodal e banheira. Dependendo de como o tempo é mensurado no banco de dados, a variável pode ser discreta. Para contornar este problema são adotadas distribuições discretas ou distribuições contínuas são discretizadas. Na literatura são poucas as distribuições para este caso. Com isso, nos últimos anos, pesquisadores vêm descobrindo novas distribuições, como a própria Kumaraswamy Log-Logística discreta, proposta por Simões e Silva (2019). Neste trabalho, essa distribuição será utilizada para a modelagem de dados reais da área de Ciência Política, com objetivo de observar sua aplicabilidade.

# 2 Objetivos

## 2.1 Objetivo Geral

Analisar o tempo até que um estado americano adote a política de proibição de fumar, utilizando a distribuição Kumaraswamy Log-Logística discreta, proposta por Simões e Silva (2019), e seu caso particular, a distribuição Log-Logística discreta e comparar os modelos de regressão para estas distribuições.

## 2.2 Objetivos Específicos

- Estudar as técnicas de Análise de Sobrevivência.
- Aplicar e comparar os modelos de regressão Kumaraswamy Log-Logístico discreto e Log-Logístico discreto para analisar dados de difusão de políticas.
- Implementar computacionalmente por meio do *software* R, as metodologias propostas neste trabalho.

# 3 Revisão de Literatura

"O termo análise de sobrevivência refere-se basicamente a situações médicas envolvendo dados censurados. Entretanto, condições similares ocorrem em outras áreas em que se usam as mesmas técnicas de análise de dados. Em engenharia, são comuns os estudos em que produtos ou componentes são colocados sob teste para se estimar características relacionadas aos seus tempos de vida, tais como o tempo médio ou a probabilidade de um certo produto durar mais do que 5 anos"(Colosimo e Giolo 2006, p. 2). Nesta seção será definido os conceitos e elementos de análise de sobrevivência, abrangendo também a parte de modelagem dos dados de sobrevivência e suas distribuições.

## 3.1 Notação e conceitos básicos

Em análise de sobrevivência, a variável resposta é constituída de dois componentes: o tempo de falha e as censuras, ambos caracterizam os dados de sobrevivência. Segundo Colosimo e Giolo (2006) o tempo de falha é constituído por três elementos (tempo inicial, escala de medida e evento de interesse), estes devem ser claramente definidos. O tempo inicial é quando começa a ser realizado o estudo, utilizado para comparação dos indivíduos na origem do estudo. A escala de medida é o "tempo"que será contabilizado, por exemplo, tempo real, meses, semanas, números de ciclos, medidas de carga e muitas outras. O evento de interesse é a própria falha, na maioria dos casos indesejável, que em análise de sobrevivência deve ser definida de forma clara e precisa. Após definir os três elementos, determina-se a variável tempo de falha.

Geralmente, os estudos de sobrevivência terminam antes que todas as observações tenham apresentado o evento, logo possuem observações incompletas ou parciais da resposta. Elas são denominadas censura, a segunda componente da variável resposta, que podem ocorrem por diversas razões, como perda de acompanhamento ou mesmo não ter apresentado a falha antes do término do estudo. Estas observações incompletas devem ser utilizadas na análise estatística, pois a omissão pode acarretar em conclusões viciadas, além de fornecerem informação sobre o tempo de vida.

Colosimo e Giolo (2006) definem três formas de censura, são elas, censura à esquerda que ocorre quando o tempo registrado é maior que o tempo de falha, quando o indivíduo é observado já aconteceu o evento. Censura intervalar que acontece quando observa-se periodicamente as observações, não sendo conhecido o tempo exato de falha, apenas o intervalo. E censura à direita que é a mais encontrada nos estudos, ocorre quando o tempo de ocorrência da falha está à direita do tempo registrado. A última será a utilizada neste trabalho.

## 3.2 Funções

Hoel, Port e Stone (1978) definem, matematicamente, que uma variável aleatória $X$ é uma função real $X(\omega)$ definida em um espaço de probabilidade $(\Omega, \mathcal{A}, P)$, em que $\omega \in \Omega$ e $\{\omega — X(\omega) \leq x\}$ é um evento para todos $-\infty < x < \infty$. Com isso, a função de distribuição $F$ de uma variável aleatória $X$ é definida como:

$$F(x) = P(X \leq x), \quad x \in \Re, \tag{3.2.1}$$

sendo, $0 \leq F(x) \leq 1$ para todo x e $F$ uma função não decrescente de x. Hoel, Port e Stone (1978) observa que geralmente as funções de distribuição são definidas em termos de funções de densidade. Uma função de densidade (em relação a integração) é uma função não-negativa $f$ tal que

$$\int_{-\infty}^{\infty} f(x)dx = 1, \tag{3.2.2}$$

(Hoel, Port e Stone 1978, p. 117). Em termos da função de distribuição de x, pode ser escrito da forma:

$$F(x) = \int_{-\infty}^{x} f(u)du, \tag{3.2.3}$$

para todo $-\infty \leq x \leq \infty$. Logo, uma densidade é definida desde que $F$ seja sempre contínua e que a derivada de $F$ exista e seja contínua em todos os pontos, exceto por um número finito de pontos, define Hoel, Port e Stone (1978).

Mas em um caso que os valores possíveis da variável sejam números inteiros, temos uma variável discreta. "Uma variável aleatória discreta real $X$, em um espaço de probabilidade $(\Omega, \varsigma, P)$, e uma função $X$ cujo domínio é $\Omega$ e cujo contradomínio e um subconjunto finito ou infinito enumerável $\{x_1, x_2, \dots\}$ dos números reais $\Re$ tal que $\{\omega : X(\omega) = x_i\}$ é um evento para todo $i$."(Hoel, Port e Stone 1978, p.50). Logo a probabilidade de um evento $x_i$ acontecer é denotada por $P(X = x_i)$.

Segundo Hoel, Port e Stone (1978), é denominado função discreta de probabilidade $X$ a uma função real $p$, definida por:

$$p(x) = P(X \ = \ x), \tag{3.2.4}$$

onde um número real $x$ é valor possível de $X$ se $p(x)>0$.

Segundo Colosimo e Giolo (2006) geralmente representa-se o tempo de falha por uma variável aleatória não-negativa $T$, usualmente contínua, com uma função de densidade e

distribuição. Logo, ao aplicar a definição de Hoel, Port e Stone (1978) para a variável tempo de falha, defini-se a função de distribuição, $F(t) = P(T \leq t)$, sendo a probabilidade de uma observação falhar até o tempo $t$ e $f(t)$ é a função de densidade de $T$.

Em análise de sobrevivência, o tempo de falha também é especificado pela função de sobrevivência e pela função de taxa de falha. Colosimo e Giolo (2006) define a função de sobrevivência sendo a probabilidade de uma observação sobreviver até o tempo "t". Logo, é definida por:

$$S(t) = P(T > t) = 1 - F(t). \tag{3.2.5}$$

Em termos probabilísticos, para uma variável aleatória contínua positiva, define-se a função de sobrevivência como:

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du, \quad t>0, \tag{3.2.6}$$

onde $f(.)$ é a função de densidade da variável $T$.

Segundo Fernandes (2013), no caso de uma variável aleatória discreta não negativa, assumindo apenas valores inteiros, $S(t)$ é definida como:

$$S(t) = P(T > t) = \sum_{k=t+1}^{\infty} P(T = k), \quad t = 0, 1, 2, 3, \ldots. \tag{3.2.7}$$

Como definido anteriormente, a função de distribuição está no intervalo [0,1] e é uma função não decrescente. Então, como $S(t) = 1 - F(t)$, tem-se que a função de sobrevivência é decrescente e também definida no intervalo [0,1]. A outra função que especifica o tempo de falha, é a função de taxa de falha que segundo Colosimo e Giolo (2006) é a probabilidade da falha ocorrer em um intervalo de tempo $[t, t+\Delta t]$, assumindo um $\Delta t$ muito pequeno representa a taxa de falha instantânea. A função de taxa de falha $T$ é definida como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{3.2.8}$$

Outra função importante é a função de taxa de falha acumulada, a qual fornece a função de taxa de falha acumulada do indivíduo, sendo definida por:

$$H(t) = \int_0^t \lambda(u)du. \tag{3.2.9}$$

Colosimo e Giolo (2006) evidencia algumas relações importantes entre as funções que

são bastante utilizadas em análise de sobrevivência. Ao conhecer a função de sobrevivência e a função de densidade do tempo de falha, é possível definir a função de taxa de falha e taxa de falha acumulada, respectivamente por:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{3.2.10}$$

e

$$H(t) = -\log S(t). \tag{3.2.11}$$

Fernandes (2013) mostra que no caso de distribuições de probabilidade discretas, a função de taxa de falha é igual a 0, exceto em pontos que pode ocorrer uma falha. Com isso a função de risco, definida por $h_d(t)$ e definida no intervalo $0<h_d(t)<1$, pode ser expressa por:

$$\begin{aligned} h_d(t) &= P(T = t|T \geq t) \\ &= \frac{P(T = t)}{P(T \geq t)} \\ &= \frac{P(T = t)}{P(T>t) + P(T = t)} \\ &= \frac{p(t)}{S_d(t) + p(t)}, \end{aligned} \tag{3.2.12}$$

onde $p(t)$ é a função de probabilidade da variável $T$ e $S_d(t)$ é a função de sobrevivência da variável discreta $T$.

## 3.3 Técnicas não-paramétricas

Em análise de sobrevivência, a presença da censura não permite que a análise descritiva usual seja realizada. Uma alternativa para esse problema é utilizar o estimador de Kaplan-Meier, para estimar a função de sobrevivência e a função de taxa de falha. Com isso, é possível realizar o estudo descritivo dos dados de análise de sobrevivência.

### 3.3.1 Estimador de Kaplan-Meier

O estimador proposto por Kaplan e Meier (1958) é utilizado para estimar a função de sobrevivência e os mesmos provam que este é o estimador de máxima verossimilhança de $S(t)$. Tendo as propriedades de não viciado para amostras grandes, fracamente consistente e converge assintoticamente para um processo gaussiano.

O estimador considera tantos intervalos de tempo quantos forem os tempos de falha. A função $\hat{S}(t)$ é uma função escada sendo que os degraus possuem tamanho $\frac{1}{n}$, em que $n$ é o tamanho da amostra, e estão localizados nos tempos de falha observados. O tamanho do degrau é multiplicado pelo número de empates caso ocorra. Um exemplo gráfico do estimador de Kaplan e Meier para $S(t)$ é apresentado na Figura 1.

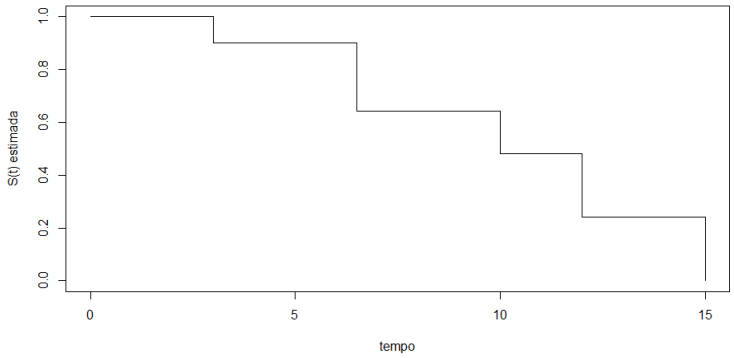


Figura 1: Exemplo de um gráfico do estimador de Kaplan Meier para $S(t)$

A obtenção da estimativa de Kaplan Meier envolve uma sequência de passos, em que o próximo depende do anterior. Logo para se encontrar a estimativa para um tempo $t_j$, devemos considerar a probabilidade de sobrevivência em $t_{j-1}$. A ideia é que para a observação sobreviver a $t_j$, ela primeiro deve sobreviver em $t_{j-1}$ para depois sobreviver ao tempo $t_j$. Em forma matemática, é definida desta maneira:

$$\begin{aligned} S(t_j) &= P(T \geq t_j) = P(T \geq t_{j-1}, T \geq t_j) \\ &= P(T \geq t_{j-1}) P(T \geq t_j | T \geq t_{j-1}). \end{aligned} \tag{3.3.1}$$

Assim, para qualquer $t$, $S(t)$ pode ser escrita em função de probabilidades condicionais. Sabendo que $S(t)$ é uma função com degraus, isto é, com probabilidade maior que zero apenas nos tempos de falha $t_j$, tem-se que:

$$S(t_j) = (1 - q_1)(1 - q_2) \ldots (1 - q_j), \tag{3.3.2}$$

onde:

- $q_j$ é a probabilidade de uma observação falhar no intervalo $[t_{j-1}, t_j)$ sabendo que sobreviveu até $t_{j-1}$.

Para o estimador de Kaplan-Meier, estima-se $q_j$ como sendo a razão entre a quantidade de falhas em $t_{j-1}$ e o número de observações sob risco em $t_{j-1}$. Isto é, indivíduos que

não falharam e não foram censurados até o instante imediatamente anterior a $t_j$. Com isso, a fórmula para as estimativas de Kaplan-Meier é definida por:

$$\hat{S}(t) = \prod_{j:t_j<t} \Big(1 - \frac{d_j}{n_j}\Big), \tag{3.3.3}$$

em que:

- $t_1 < t_2 < \ldots < t_k$, os $k$ tempos de falha ordenados e distintos,
- $d_j$ o número de falhas em $t_j$, j=1,...,k, e
- $n_j$ número de observações sob risco em $t_j$.

Outros dois estimadores não-paramétricos podem ser utilizados. Nelson (1972) propôs e Aalen (1978) estudou as propriedades do estimador chamado de Nelson-Aalen. As estimativas deste são maiores ou iguais as de Kaplan-Meier, mostrou Bohoris (1994). O outro estimador é a Tabela de Vida ou Atuarial, este constrói uma tabela de vida e divide o eixo do tempo em vários intervalos. Os resultados também são bem próximos, inclusive se os intervalos de Kaplan-Meier forem iguais aos da Tabela de Vida, as estimativas apresentarão mesmos valores. Neste estudo será utilizado o estimador de Kaplan-Meier.

### 3.3.2 TTT Plot e Função de Risco

A função de sobrevivência associada a qualquer distribuição de probabilidade, sempre terá forma decrescente. Ao pensar na modelagem das observações, a forma da função de sobrevivência não é tão informativa, então o pesquisador deve voltar-se para as funções de risco.

Como a função de risco pode assumir várias formas, os modelos acabam se tornando concorrentes entre si para ajustar o conjunto de dados. Visto isso, são adotadas duas metodologias para seleção do modelo mais apropriado. O TTT Plot (Tempo Total em Teste) é um método gráfico proposto por Arset (1987), e é construído ao utilizar a seguinte função:

$$G\Big(\frac{r}{n}\Big) = \frac{\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_i}, \tag{3.3.4}$$

em que:

- r = 1,..., n, onde n é o tamanho da amostra,
- $T_{i:n}$, i=1,..., n, são as estatísticas de ordem da amostra em rol crescente,

- $T_i$ é o i-ésimo tempo da amostra, e
- $T_{r:n}$ é o r-ésimo tempo da amostra ordenada em rol crescente.

A Figura 2 apresenta as várias formas que a função TTT pode assumir e a interpretação dessa função é definida por:

- Se os dados apresentam a forma gráfica A, indica que a função de risco dos dados é uma função constante (modelo exponencial).
- A forma gráfica B ou C, indica que a função de risco dos dados é uma função monotonicamente decrescente ou crescente, respectivamente (modelo Weibull).
- A forma gráfica D, indica que a função de risco com forma de "U"(modelos de riscos múltiplos).
- A forma gráfica E, indica que a função de risco é unimodal (modelo Log-Logístico).

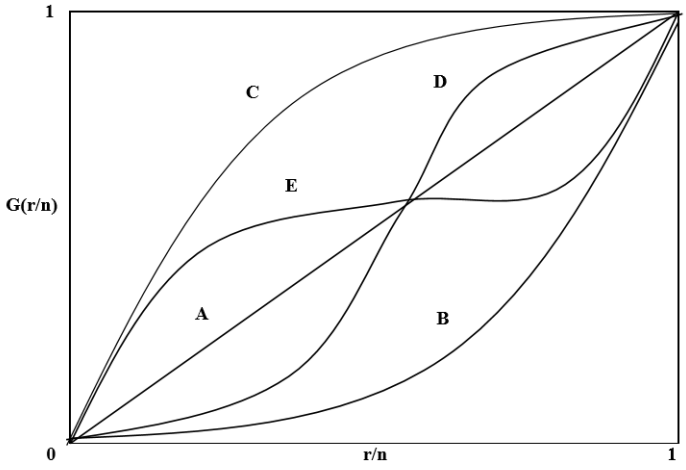


Figura 2: Exemplo das várias formas assumidas pela curva TTT. Fonte: Nakano (2018)

A curva TTT não considera as censuras em sua construção, podendo ter conclusões viesadas. Com isso, $\hat{H}(t)$ (estimativa da função de risco acumulada, definida em 3.2.9) é uma alternativa para o TTT Plot quando o número de censuras é muito grande. A interpretação da estimativa da função de taxa de falha acumulada é inversa a do TTT Plot. Para construir o gráfico da estimativa da função de taxa de falha acumulada, utiliza-se a relação em 3.2.11, logo:

$$\hat{H}(t) = -log\hat{S}(t), \tag{3.3.5}$$

sendo que o estimador de Kaplan-Meier é usado para estimar a função de sobrevivência $S(t)$. Exemplos das formas gráficas da função de risco acumulada são ilustradas na Figura 3 e a interpretação dessa função é definida por:

- Se os dados apresentam a forma gráfica A, indica que a função de risco dos dados é uma função constante (modelo exponencial).
- A forma gráfica B ou C, indica que a função de risco dos dados é uma função monotonicamente crescente ou decrescente, respectivamente (modelo Weibull).
- A forma gráfica D, indica que a função de risco é unimodal (modelo Log-Logístico).
- A forma gráfica E, indica que a função de risco com forma de "U"(modelos de riscos múltiplos).

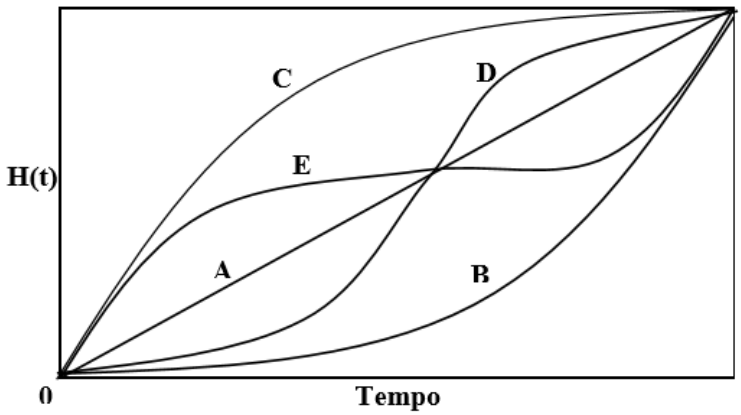


Figura 3: Exemplo das várias formas assumidas pela função de taxa de falha acumulada. Fonte: Nakano (2018)

## 3.4 Distribuições de Probabilidade

Nesta seção serão apresentadas algumas distribuições de probabilidade utilizadas para analisar dados de sobrevivência. Bem como, o método de máxima verossimilhança, para estimar os parâmetros das distribuições. Ainda nesta seção, será descrito o método de discretização de distribuições contínuas.

### 3.4.1 Distribuição Kumaraswamy

Proposta por Kumaraswamy (1980), essa distribuição possui diversas aplicações e é conhecida por sua flexibilidade. A função de distribuição de probabilidade (fdp) é expressa por:

$$f(t;\alpha,\beta) = \alpha\beta t^{\alpha-1}(1 - t^{\alpha})^{\beta-1}, \qquad 0 < t < 1, \tag{3.4.1}$$

e a função de distribuição acumulada (fda) por:

$$F(t,\alpha,\beta) = 1 - (1 - t^{\alpha})^{\beta}, \qquad 0 < t < 1, \tag{3.4.2}$$

sendo $\alpha > 0$ e $\beta > 0$ os parâmetros de forma.

Cordeiro e Castro (2010) apresentam uma nova distribuição, chamada de Kumaraswamy generalizada ($KwG$), que permite a aplicação de outra distribuição à Kumaraswamy, gerando uma nova distribuição. Ela considera as funções de outra distribuição arbitrária $G(t)$ (função de distribuição acumulada) e $g(t)$ (função de distribuição de probabilidade). Cordeiro e Castro (2010) definiram a fda e fdp da $KwG$, respectivamente, por:

$$F(t;a,b) = 1 - [1 - G(t)^{a}]^{b} \tag{3.4.3}$$

e

$$f(t;a,b) = abg(t)G(t)^{a-1}[1 - G(t)^{a}]^{b-1}, \tag{3.4.4}$$

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

### 3.4.2 Distribuição Log-Logística

Uma das distribuições mais utilizadas em análise de sobrevivência é a log-logística, pois tem se apresentado como uma alternativa para das distribuições Weibull e log-normal segundo Colosimo e Giolo (2006). A função de densidade e função de distribuição acumulada da log-Logística são definidas ,respectivamente, por:

$$f(t;\alpha,\gamma) = \frac{\gamma}{\alpha^{\gamma}}\, t^{\gamma-1}\left(1 + \left(\frac{t}{\alpha}\right)^{\gamma}\right)^{-2}, \qquad t > 0 \tag{3.4.5}$$

e

$$F(t;\alpha,\gamma) = \frac{1}{(1 + (t/\alpha)^{-\gamma})} = \frac{t^{\gamma}}{t^{\gamma} + \alpha^{\gamma}}, \qquad t > 0, \tag{3.4.6}$$

sendo $\alpha > 0$ parâmetro de escala e $\gamma > 0$ o de forma.

### 3.4.3 Distribuição Kumaraswamy Log-Logística

Seguindo o conceito da $KwG$, neste trabalho será utilizada como distribuição arbitrária a Log-Logística. Portanto, ao aplicar as funções desta distribuição na Kuma-

raswamy generalizada resulta-se na distribuição Kumaraswamy Log-Logística ($KwLL$).

Dado as equações (3.4.3) e (3.4.4), ao substituir $G(t)$ pela fdp da log-logística (3.4.6) e $g(t)$ pela densidade (3.4.5), obtém-se a funções de distribuição de probabilidade e distribuição acumulada da $KwLL$, expressas respectivamente por:

$$f_{kll}(t;a,b,\alpha,\gamma) = ab\Big(\frac{\gamma}{\alpha^{\gamma}}\ t^{\gamma-1}\Big[1\ +\ \Big(\frac{t}{\alpha}\Big)^{\gamma}\Big]^{-2}\Big)\Big(\frac{t^{\gamma}}{t^{\gamma}\ +\ \alpha^{\gamma}}\Big)^{a-1}\Big[1\ -\ \Big(\frac{t^{\gamma}}{t^{\gamma}\ +\ \alpha^{\gamma}}\Big)^{a}\Big]^{b-1} \tag{3.4.7}$$

e

$$F_{kll}(t;a,b,\alpha,\gamma) = 1\ -\ \Big[1\ -\ \Big(\frac{t^{\gamma}}{t^{\gamma}\ +\ \alpha^{\gamma}}\Big)^{a}\Big]^{b}, \tag{3.4.8}$$

sendo $\alpha > 0$ o parâmetro de escala, $\gamma > 0$, $a > 0$ e $b > 0$ parâmetros de forma da distribuição Kumaraswamy Log-Logística contínua e $t \in \Re$.

### 3.4.4 Método da Máxima Verossimilhança

O método da máxima verossimilhança busca estimar os parâmetros do modelo definido, buscando os melhores valores dos parâmetros que expliquem a amostra observada. Para isso é obtido os valores dos parâmetros que maximizam a probabilidade de ocorrer a amostra de dados observada.

Supondo uma amostra $t_1, t_2, ..., t_n$, a função de verossimilhança para um parâmetro ou vetor de parâmetros genérico $\theta$ desta população, sendo $f(t)$ a fdp da distribuição definida, é expressa por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta), \tag{3.4.9}$$

em que $t_i$ representa o tempo de falha da i-ésima obervação.

Entretanto, neste trabalho será analisado um banco de dados de sobrevivência, logo tem a presença de censuras. A vantagem da estimação por máxima verossimilhança é conseguir incorporar em sua fórmula as observações censuradas. Portanto, considerando $f(t_i;\theta)$ a fdp da distribuição definida, $S(t_i;\theta)$ a função de sobrevivência dessa distribuição, $\theta$ o vetor de parâmetros da mesma e $\delta_i$ uma variável indicadora que assume 1 se a i-ésima observação tiver falhado e 0 caso contrário (censura a direita), a função de verossimilhança $L(\theta)$ é expressa por:

$$L(\theta) \propto \prod_{i=1}^{n} \big[f(t_i;\theta)\big]^{\delta_i}\big[S(t_i;\theta)\big]^{1-\delta_i}, \tag{3.4.10}$$

em que a "contribuição"das falhas é a função de distribuição de probabilidade da distribuição definida e das censuras é dada pela função de sobrevivência da distribuição utilizada.

### 3.4.5 Discretização de Distribuições Contínuas

Na subseção 3.2 foi apresentado a definição de variável aleatória, a qual pode ser discreta, quando assume valores de um conjunto enumerável, e contínua, quando assume valores de um intervalo dos números reais sendo um conjunto não enumerável.

Em análise de sobrevivência a variável estudada é o tempo até a ocorrência de um evento. O tempo é geralmente tratado como uma variável contínua, mas há casos onde o tempo é medido em anos, meses, ou mesmo o estudo é controlado de forma intervalar.

Ao agrupar os valores de tempo em intervalos unitários, é possível obter modelos de variáveis discretas a partir de modelos contínuos. Definindo $X$ como uma variável aleatória contínua, a discreta é dada por $T = [X]$, em $[X]$ representa a "parte inteira de X". Como visto em Nakano e Carrasco (2006), se $F_X(x)$ é a função de distribuição acumulada de $X$, a distribuição de probabilidade de $T$, representada por $p(t)$ pode ser escrita por:

$$\begin{aligned} p(t) &= P(T = t) \\ &= P(t \leq X < t + 1) \\ &= F_X(t + 1) - F_X(t), \qquad \text{t = 0,1,2,...,} \end{aligned} \tag{3.4.11}$$

em que $F_X$ é a fda da variável aleatória contínua.

# 4 Metodologia

O objetivo deste trabalho é estudar a aplicabilidade da distribuição $KwLLD$ a dados de sobrevivência com tempos discretos e compará-la ao seu caso particular, a distribuição Log-Logística discreta, por meio do ajuste dos modelos de regressão das distribuições em questão. Logo, nesta seção será definida as distribuições Kumaraswamy Log-Logística discreta, Log-Logística discreta e seus respectivos modelos de regressão. Bem como, os dados que serão utilizados para exemplificar a aplicabilidade da distribuição $KwLLD$.

## 4.1 Material

O banco de dados a ser analisado, possui como evento de interesse a adesão da política de proibição de fumar pelos estados dos Estados Unidos (menos os estados do Havaí e Alasca), com exceção aos locais definidos por cada estado em que é permitido fumar. É definido como variável tempo de falha o tempo até que um estado, do país em questão, tenha adotado esta política.

O tempo inicial é 1995, ano em que o estado da Califórnia lançou a política, contabilizada como ano 0. Por falta de informação (dia e mês), o tempo é medido em anos, sendo assim uma variável discreta. A política busca proteger a saúde da sociedade, visto as mazelas derivadas do fumo e gastos anuais milionários com tratamento de câncer de pulmão.

Buscando entender o que influencia para que o estado adote a política, foram analisadas algumas covariáveis, que contém informação correspondente ao ano de falha ou censura do estado. As variáveis do banco são descritas por:

- Tempo: anos até que o estado adote a política.
- Censura: indica se a informação da variável tempo é falha, sendo definida por 1, ou censura, definida por 0.
- Governo democrata: se o Governador durante o ano de falha ou censura era do partido Democrata, sendo 1 caso fosse.
- Governo republicano: se o Governador durante o ano de falha ou censura era do partido Republicano, sendo 1 caso fosse.
- Margem de vitória: indica a diferença, em pontos percentuais, entre os partidos Republicano e Democrata nas eleições mais próximas ao ano de falha ou censura.

- Governador reeleito: indica se o governador do estado foi reeleito, com 1 informando reeleição.
- Ano eleitoral: diz respeito a existência de processo eleitoral durante o ano de falha ou censura, sendo 1 caso o ano tenha sido eleitoral.
- Ideologia dos cidadãos: medida desenvolvida por Berry et al. (1998) que mede a ideologia política dos cidadãos de cada estado.
- Ideologia do governo: medida desenvolvida por Berry et al, (1998) que mede a ideologia política governamental de cada estado.
- Profissionalismo legislativo: medida desenvolvida por Squire (2007) que mede o profissionalismo legislativo de cada estado.
- Composição da câmara legislativa e senado: descreve se os dois possuíam Democratas ou Republicanos como maioria, sendo codificada com 1 para maioria Democrata nas duas e 0 para maioria Republicana ou nenhum partido tinha maioria nos dois ambientes de cada estado.
- Estados vizinhos adotantes da política: informa o número de estados vizinhos que adotaram a política anteriormente ao estado em questão.
- Políticas de saúde entre 1912-2017: porcentagem de políticas de saúde adotadas no período de 1912 a 2017.
- Políticas de saúde entre 1990-2017: porcentagem de políticas de saúde adotadas no período de 1990 a 2017.
- Porcentagem de fumantes: porcentagem de adultos fumantes no estado.
- População: população de cada estado no ano ou no ano anterior mais próximo da falha ou censura.
- Produção de tabaco: codificado como 1 sendo o estado produz tabaco e 0 caso contrário.
- Restrição em prédios do governo: proporção da população do estado com restrição local para fumar em prédios governamentais.
- Lei estadual para prédios do governo: codificada sendo 1 para estados com restrição e lei estadual para não fumar em prédios do governo e 0 caso contrário.
- Restrição em restaurantes: proporção da população do estado com restrição local para fumar em restaurantes.

- Lei estadual para restaurantes: codificada sendo 1 para estados com restrição e lei estadual para não fumar em restaurantes e 0 caso contrário.

- Restrição de acesso aos jovens: proporção do estado com restrição local de acesso de cigarros aos jovens.

- Lei estadual para acesso aos jovens: codificada sendo 1 para estados com restrição e lei estadual que não permite o acesso dos jovens a cigarros e 0 caso contrário.

- Lobistas na indústria do tabaco: proporção de lobistas que trabalham em indústrias de tabaco.

- Lobistas em organizações de saúde: proporção de lobistas que trabalham em organizações de saúde.

## 4.2 Distribuição Log-Logística discreta

Ao utilizar a metodologia de discretização definida na Seção 3.5.5 e aplicá-la na distribuição Log-Logística definida na Seção 3.5.2, Santos (2017) propõe a distribuição Log-Logística para analisar o tempo até a ocorrência do evento de interesse, quando esse tempo é discreto. Dessa forma, a função de probabilidade da distribuição Log-Logística discreta, bem como, a função de sobrevivência e função de risco são expressas por:

$$p_{lld}(t) = \frac{1}{1+(t/\alpha)^{\gamma}} - \frac{1}{1+[(t+1)/\alpha]^{\gamma}}, \qquad t = 0, 1, 2, ... \tag{4.2.1}$$

$$S_{lld}(t) = \frac{1}{1+[(t+1)/\alpha]^{\gamma}}, \qquad t = 0, 1, 2, ... \tag{4.2.2}$$

e

$$h_{lld}(t) = 1 - \frac{1+(t/\alpha)^{\gamma}}{1+[(t+1)/\alpha]^{\gamma}}, \qquad t = 0, 1, 2, ... \tag{4.2.3}$$

sendo $\alpha$>0 o parâmetro de escala e $\gamma$>0 o prâmetro de forma da distribuição de probabilidade. Segundo Santos (2017) a função de risco da distribuição Log-Logística discreta assume forma decrescente e unimodal.

## 4.3 Distribuição Kumaraswamy Log-Logística discreta

Ao utilizar a metodologia de discretização definida na Seção 3.5.5 e aplicá-la na distribuição Kumaraswamy Log-Logística definida na Seção 3.5.3, Simões e Silva (2019) propõe a distribuição Kumaraswamy Log-Logística para analisar o tempo até a ocorrência

do evento de interesse, quando esse tempo é discreto. De acordo com 3.4.11, a função de probabilidade da distribuição Kumarasamy Log-Logística discreta é expressa por:

$$p_{klld}(t) = \Big[1 \; - \; \Big(\frac{t^{\gamma}}{t^{\gamma} \; + \; \alpha^{\gamma}}\Big)^{a}\Big]^{b} \; - \; \Big[1 \; - \; \Big(\frac{(t+1)^{\gamma}}{(t+1)^{\gamma} \; + \; \alpha^{\gamma}}\Big)^{a}\Big]^{b}, \qquad t = 0, 1, 2, ..., \tag{4.3.1}$$

e a função de densidade acumulada por:

$$F_{klld}(t) = 1 \; - \; \Big[1 \; - \; \Big(\frac{(t+1)^{\gamma}}{(t+1)^{\gamma} \; + \; \alpha^{\gamma}}\Big)^{a}\Big]^{b}, \qquad t = 0, 1, 2, ..., \tag{4.3.2}$$

em que $\alpha > 0$ e $\gamma > 0$ são parâmetros de escala e forma da distribuição Log-Logística e $a > 0$ e $b > 0$ parâmetros de forma da distribuição Kumaraswamy Generalizada.

A função de sobrevivência e de risco da distribuição $KwLLD$ são expressas, respectivamente, por:

$$S_{klld}(t) = \Big[1 \; - \; \Big(\frac{(t+1)^{\gamma}}{(t+1)^{\gamma} + \alpha^{\gamma}}\Big)^{a}\Big]^{b}, \qquad t = 0, 1, 2, ... \tag{4.3.3}$$

e

$$h_{klld}(t) = 1 \; - \; \frac{\Big[1 - \Big(\frac{(t+1)^{\gamma}}{(t+1)^{\gamma}+\alpha^{\gamma}}\Big)^{a}\Big]^{b}}{\Big[1 \; - \; \Big(\frac{t^{\gamma}}{t^{\gamma}+\alpha^{\gamma}}\Big)^{a}\Big]^{b}}, \qquad t = 0, 1, 2, ... \tag{4.3.4}$$

Segundo Simões e Silva (2019) a função de risco da distribuição Kumaraswamy Log-Logística discreta assume forma decrescente, crescente e unimodal.

## 4.4 Modelo de regressão Log-Logístico discreto

Grande parte dos estudos de análise de sobrevivência possuem variáveis que podem influenciar o tempo de falha de um indivíduo. Uma maneira importante de representar a heterogeneidade em uma população é com a utilização dessas covariáveis em um modelo de regressão (Lawless, 2011).

Para estudar esta relação, modelos de regressão paramétricos podem ser formulados, considerando uma reparametrização da distribuição de probabilidade dos tempos. Segundo Lawless (2011), seja $\boldsymbol{x^T} = (1, x_1, \ldots, x_p)$ um vetor constituído pelas informações das $p + 1$ variáveis regressoras, utiliza-se uma função de ligação $g(.)$, sendo possível conectar a variável resposta as variáveis explicativas. Para um conjunto de $p$ covariáveis, $\boldsymbol{\theta}$ é definido por:

$$\boldsymbol{\theta} = g(\boldsymbol{\eta}), \tag{4.4.1}$$

em que $\boldsymbol{\eta} = \boldsymbol{x^T\beta}$ é o preditor linear e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros associados as covariáveis.

Assim, para uma variável aleatória com distribuição Log-Logística discreta, definida em (4.2), Santos (2017) propõe o seguinte modelo de regressão $LLD$:

$$p_{lld}(t) = \frac{1}{1+[t/\exp(\boldsymbol{x^T\beta})]^\gamma} - \frac{1}{1+[(t+1)/\exp(\boldsymbol{x^T\beta})]^\gamma}, \qquad t = 0, 1, 2, ... \tag{4.4.2}$$

As funções de sobrevivência e de risco correspondentes são dadas por:

$$S_{lld}(t) = \frac{1}{1+[(t+1)/\exp(\boldsymbol{x^T\beta})]^\gamma}, \qquad t = 0, 1, 2, ... \tag{4.4.3}$$

e

$$h_{lld}(t) = 1 - \frac{1+[t/\exp(\boldsymbol{x^T\beta})]^\gamma}{1+[(t+1)/\exp(\boldsymbol{x^T\beta})]^\gamma}, \qquad t = 0, 1, 2, ... \tag{4.4.4}$$

em que $\gamma$>0 é o parâmetro de forma da distribuição Log-Logística discreta, $\boldsymbol{x^T}$ o vetor de covariáveis e $\boldsymbol{\beta}$ o vetor dos coeficientes de regressão.

Será utilizado o método da máxima verossimilhança, definido em (3.5.4), para estimar os parâmetros do modelo. Utilizando a equação 3.4.10, encontra-se o logaritmo da função de verossimilhança:

$$\begin{aligned} log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \Big[ \delta_i log\Big[\frac{1}{1+[t_i/\exp(\boldsymbol{x^T\beta})]^\gamma} - \frac{1}{1+[(t_i+1)/\exp(\boldsymbol{x^T\beta})]^\gamma}\Big] + \\ (1-\delta_i) log\Big[\frac{1}{1+[(t+1)/\exp(\boldsymbol{x^T\beta})]^\gamma}\Big]\Big] + C, \end{aligned} \tag{4.4.5}$$

em que $\boldsymbol{\theta}$=$(\beta, \gamma)$ e $C$ é uma constante que não depende de $\theta$.

## 4.5 Modelo de regressão Kumaraswamy Log-Logístico discreto

Ao utilizar a metodologia de modelos de regressão proposta por Lawless (2011), nesta seção será definido o modelo de regressão Kumaraswamy Log-Logístico discreto. Seja $\boldsymbol{x^T} = (1, x_1, \ldots, x_p)$ um vetor constituído pelas informações das $p$ variáveis explicativas, $g(.)$ uma função de ligação, o vetor $\boldsymbol{\theta}$ é definido por:

$$\boldsymbol{\theta} = g(\boldsymbol{\eta}), \tag{4.5.1}$$

em que $\boldsymbol{\eta} = \boldsymbol{x^T\beta}$ é o preditor linear e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros associados as covariáveis.

Assim, para uma variável aleatória com distribuição Kumaraswamy Log-Logística discreta, definida em (4.3), Simões e Silva (2019) propôs o modelo de regressão $KwLLD$, definido por:

$$p_{klld}(t) = \Big[1 - \Big(\frac{t^\gamma}{t^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b - \Big[1 - \Big(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b, \qquad t = 0, 1, 2, ... \tag{4.5.2}$$

As funções de sobrevivência e de risco correspondentes são dadas por:

$$S_{klld}(t) = \Big[1 - \Big(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b, \qquad t = 0, 1, 2, ... \tag{4.5.3}$$

e

$$h_{lld}(t) = 1 \; - \frac{\Big[1 - \Big(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b}{\Big[1 - \Big(\frac{t^\gamma}{t^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b}, \qquad t = 0, 1, 2, ... \tag{4.5.4}$$

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.

Será utilizado o método da máxima verossimilhança, definido em (3.5.4), para estimar os parâmetros do modelo. Utilizando a equação 3.4.10, encontra-se o logaritmo da função de verossimilhança:

$$\begin{aligned} log(L(\boldsymbol{\theta})) = \sum_{i=1}^{n} \Big[\delta_i log\Big[1 - \Big(\frac{t_i^\gamma}{t_i^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b - \Big[1 - \Big(\frac{(t_i+1)^\gamma}{(t_i+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b + \\ (1-\delta_i) log\Big[1 - \Big(\frac{(t+1)^\gamma}{(t+1)^\gamma + \exp(\boldsymbol{x^T\beta})^\gamma}\Big)^a\Big]^b + C, \end{aligned} \tag{4.5.5}$$

em que $\boldsymbol{\theta}$=($\beta, \gamma$,$a$,$b$) e $C$ é uma constante que não depende de $\theta$.

# 5 Resultados

Nesta seção será abordado a análise do banco de dados, referente à adesão da política de proibição de fumar nos estados, utilizando a técnica de Análise de Sobrevivência e modelando os dados através de duas distribuições, a Log-Logística para dados discretos ($LLD$) e a Kumaraswamy Log-Logística para dados discretos ($KwLLD$).

## 5.1 Análise Descritiva

Primeiro é realizada a análise descritiva dos dados, como os dados de sobrevivência possuem observações parciais da resposta, as censuras, será utilizado técnicas não paramétricas para descrever o banco. Na Figura 4 foi realizada a estimação da função de sobrevivência pelo método de Kaplan-Meier para os estados que aderiram a política de proibição do fumo. Observa-se que 50% dos 48 estados em análise falharam, aderiram a política. Todas as censuras ocorrem no último ano do estudo (2010), o que era esperado, dado a dificuldade de um estado sair do estudo e a última falha ocorre um ano antes (2009).

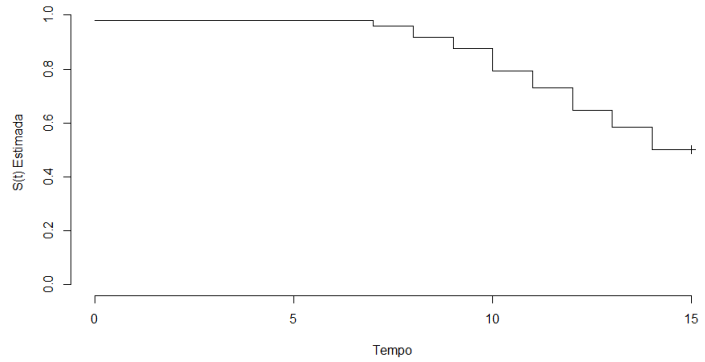


Figura 4: Estimativa da Função de Kaplan Meier para $S(t)$

Como dito anteriormente, o gráfico da estimativa da função de sobrevivência apresenta apenas uma forma, decrescente, o que acaba não sendo tão informativo pensando em modelar uma distribuição de probabilidade nos dados. Com isso, é utilizada a função de risco acumulada, Figura 5, e a curva TTT, Figura 6, nas quais é possível observar um comportamento côncavo e convexo, respectivamente, mostrando evidências que uma distribuição boa para ajustar os dados deva ter função de risco monotonicamente crescente.

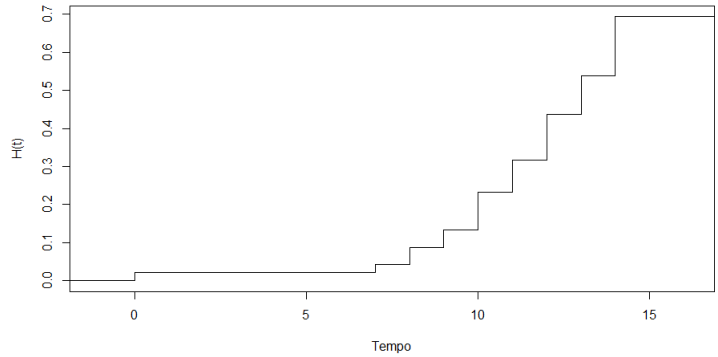


Figura 5: Função de Risco Acumulada

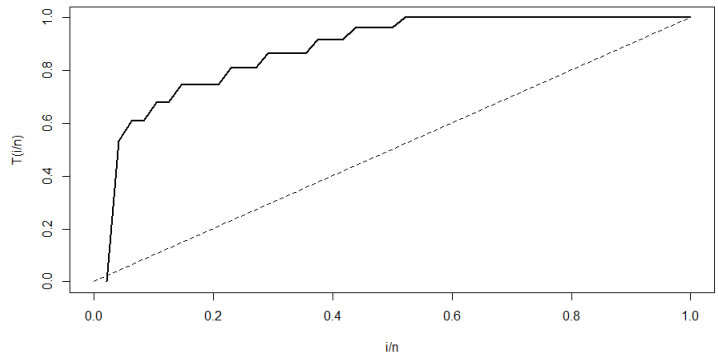


Figura 6: Curva do Tempo Total em Teste (TTT)

Próximo passo é descrever as covariáveis do banco buscando uma melhor compreensão dos dados. Em análise de sobrevivência é estimado a função de sobrevivência para cada categoria da variável e observa-se se as funções se mostram diferentes, indicando que possa ser uma covariável importante para explicar o tempo de falha, ou parecidas.

Para a covariável partido do Governador do estado no ano de falha, os dois partidos observados para os 48 estados foram os mais tradicionais, Partido Democrata e Partido Republicano. Como as covariáveis a serem analisadas são indicadoras se o partido do Governador era Democrata ou não (Figura 7), e Republicano ou não (8), a função de sobrevivência estimada para as duas variáveis do banco são idênticas, trocando apenas suas categorias indicadoras. Observa-se que as curvas se mantém muito próximas durante todo o estudo, dando evidências que estas covariáveis talvez não sejam tão explicativas para o tempo de adesão da política.

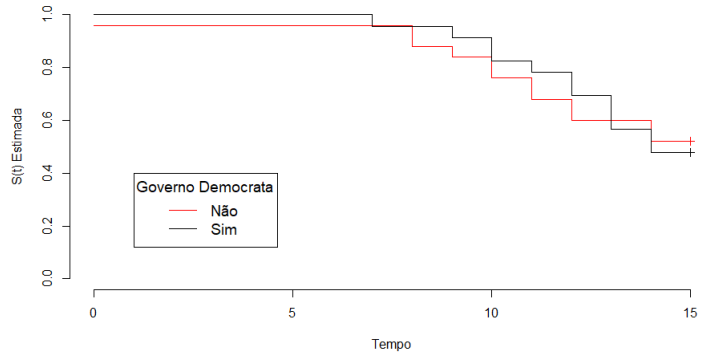


Figura 7: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Governo Democrata

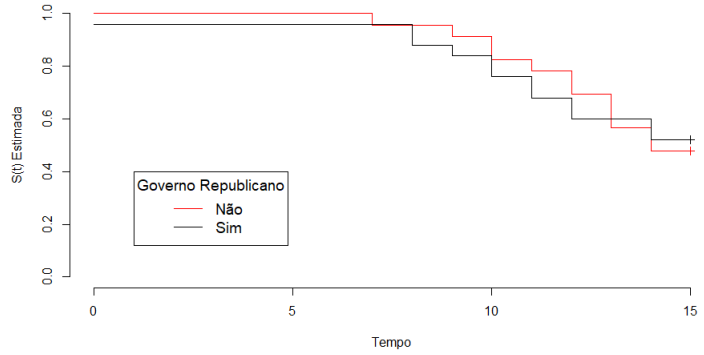


Figura 8: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Governo Republicano

Na Figura 9 observa-se a variável que indica se o Governador em ofício no ano da falha foi reeleito ou não, é possível ver que as curvas de sobrevivência estimadas não se afastam tanto uma da outra, não aparentando descrever bem a adesão à política.

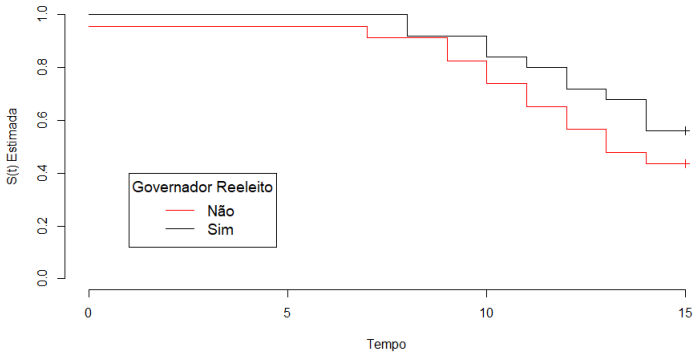


Figura 9: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Governador Reeleito

A partir da Figura 10, observa-se que a variável ano de eleição, que indica se o ano de adesão da política era eleitoral ou não, possui uma queda brusca no ano 11 (2006) para a curva de ano eleitoral, devido a apenas dois estados, dos 48 em análise, estarem em ano eleitoral e ambos falharem no mesmo ano. Portanto, esta variável mostra evidências de que, quando o estado esta em ano eleitoral o risco do mesmo falhar é maior.

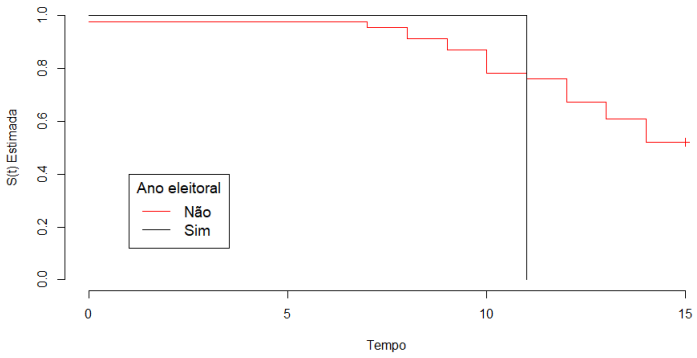


Figura 10: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Ano Eleitoral

Para a Figura 12, é estudado a covariável que indica se a Câmara e o Senado do estado era de maioria Democrata, sendo uma categoria, ou Republicana ou sem maioria partidária sendo a outra categoria. Observa-se que a composição possivelmente impacta na diminuição do tempo da sobrevivência dos estados, uma vez que a partir do tempo 10 as curvas estimadas começam a se distanciar.

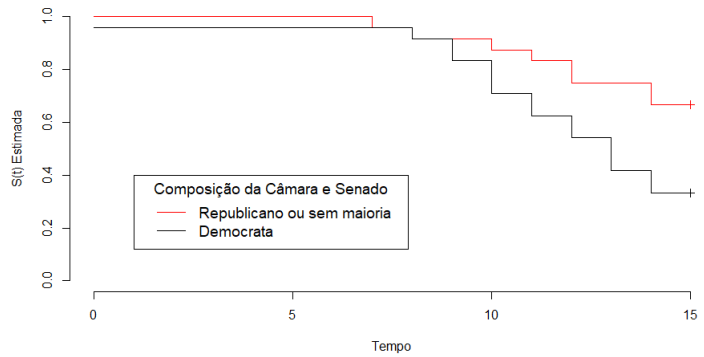


Figura 11: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Composição da Câmara e Senado Americano

Ao estudar a variável que indica se o estado em questão produz tabaco ou não, percebe-se, pela Figura **??**, que a não produção de tabaco influencia na diminuição da probabilidade de sobreviver à adesão da política, visto que a partir do tempo 10 as curvas vão se distanciando.

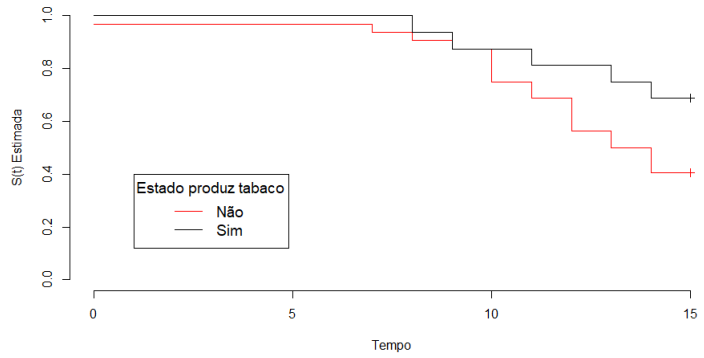


Figura 12: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Produção de Tabaco

Avaliando as covariáveis indicadoras de leis para proibição de fumar em prédios governamentais, Figura 13, proibição em restaurantes, Figura 14, e proibição para jovens, Figura 15, é possível observar que os estados com lei para prédios do governo e restaurantes estão diminuindo a probabilidade de sobrevivência dos estados, sendo possivelmente boas variáveis para explicar o tempo de adesão a política. Para a lei em relação ao acesso dos jovens, as curvas se distanciam no meio do estudo, mas vão se aproximando novamente até o final do estudo, aparentando não ser informativa para a adesão da política de proibição de fumar.

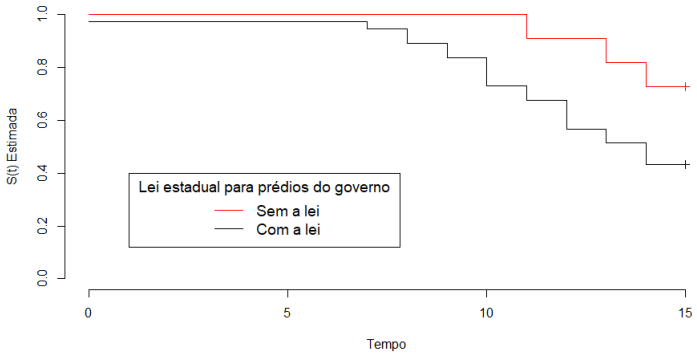


Figura 13: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Lei estadual para restrição em prédios do governo

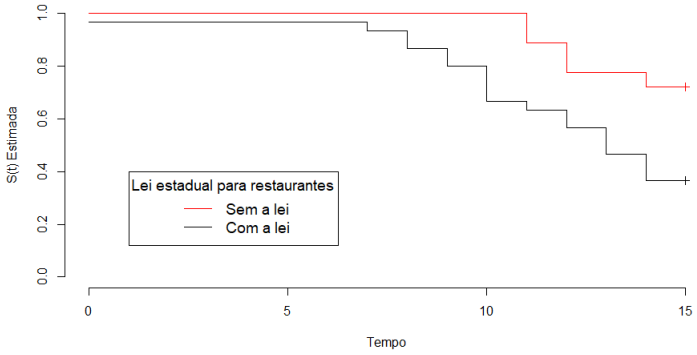


Figura 14: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Lei estadual para restrição em restaurantes

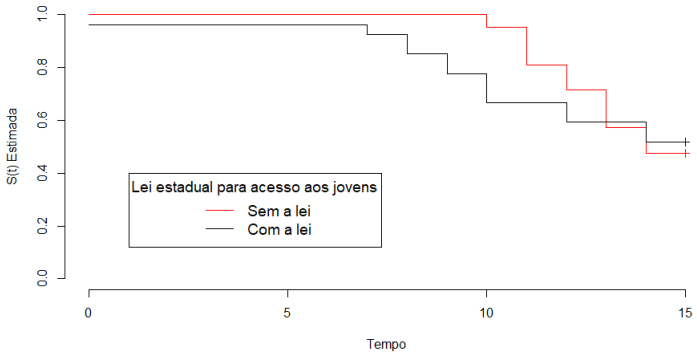


Figura 15: Estimativa da Função de Sobrevivência por Kaplan-Meier para covariável Lei estadual para restrição de acesso aos jovens

Após o estudo inicial, será realizada a modelagem dos dados. Como observado nas

Figuras 5 e 6, o risco dos dados é crescente, neste caso, é possível supor as distribuições Kumaraswamy Log-Logística discreta e Log-Logística discreta para modelar o tempo até a adesão da política de proibição de fumar. Depois do ajuste das distribuições nos dados, será analisada a presença das variáveis explicativas, dando origem aos modelos de regressão $KwLLD$ e $LLD$.

## 5.2 Ajuste da Distribuição Log-Logística Discreta

O ajuste da distribuição Log-Logística Discreta aos dados foi realizado por meio do *software R* utilizando a função *optim*. Ao ajustar a distribuição $LLD$ aos dados e comparando com a curva de Kaplan-Meier, representada pela Figura 16, observa-se, graficamente, que a distribuição escolhida se ajusta bem aos dados em estudo. As estimativas dos parâmetros $\alpha$ e $\gamma$ são apresentadas na Tabela 1. Nota-se que o valor do parâmetro $\alpha$ é elevado, dando evidências que um modelo de regressão possa ser aplicado neste banco de dados, afim de obter um bom ajuste e conseguir explicar o tempo de falha.

Tabela 1: Estimativas dos parâmetros da distribuição $LLD$

| Parâmetro | Estimativa | Erro Padrão |
|---|---|---|
| $\alpha$ | 17,018 | 1,991 |
| $\gamma$ | 2,421 | 0,483 |

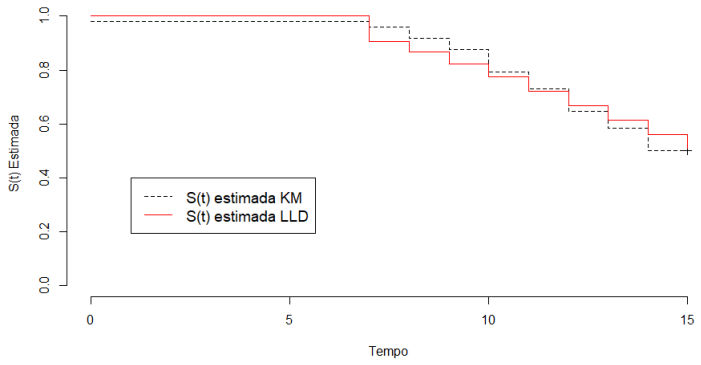


Figura 16: Comparação do ajuste da distribuição Log-Logística Discreta com a função de sobrevivência estimada por Kaplan-Meier

Com o objetivo de propor um modelo de regressão para analisar os dados, primeiramente ajustou-se modelos de regressão Log-Logístico discreto com a presença de uma covariável por vez e observou quais variáveis foram significativas individualmente ao nível de significância de 5%. A Tabela 2 apresenta as estimativas, erro-padrão e p-valor

associado a cada parâmetro de regressão.

Tabela 2: Estimativas dos coeficientes para o modelo $LLD$ com apenas uma covariável

| Variável | Estimativa | Erro Padrão | P-valor |
|---|---|---|---|
| Ideologia dos cidadãos | -0,0190 | 0,0048 | <0,0001 |
| Porcentagem de políticas de Saúde 1912-2017 | -0,0460 | 0,0118 | <0,0001 |
| Porcentagem de políticas de Saúde 1990-2017 | -0,0229 | 0,0062 | 0,0002 |
| Porcentagem de fumantes | 0,1207 | 0,0362 | 0,0009 |
| Lei de proibição para fumar em restaurantes | -0,4219 | 0,1836 | 0,0216 |
| Profissionalismo Legislativo | -1,8401 | 0,8083 | 0,0228 |
| Restrição para fumar em prédios do governo | -1,1328 | 0,5321 | 0,0333 |
| Composição da Câmara e Senado | -0,3422 | 0,1665 | 0,0399 |
| Restrição de acesso aos jovens | -1,1426 | 0,6047 | 0,0588 |
| Restrição para fumar em restaurantes | -0,9375 | 0,4994 | 0,0605 |
| Margem de vitória | 0,0116 | 0,0065 | 0,0739 |
| Número de estados vizinhos já adotantes | 0,1030 | 0,0576 | 0,0740 |
| Ideologia do Governo | -0,0102 | 0,0057 | 0,0747 |
| Lei de proibição de fumar em prédios do governo | -0,3914 | 0,2209 | 0,0764 |
| Produção de Tabaco | 0,3104 | 0,189 | 0,1007 |
| Lobistas na indústria do Tabaco | -9,6395 | 8,3506 | 0,2484 |
| Ano Eleitoral | -0,3631 | 0,3212 | 0,2583 |
| Governador Reeleito | 0,1751 | 0,1641 | 0,2861 |
| Lei de proibição de fumar para jovens | -0,0717 | 0,1644 | 0,6627 |
| Lobistas em organizações da saúde | 0,5828 | 1,5336 | 0,7039 |
| População | 0,0309 | 0,0892 | 0,7294 |
| Governador Republicano | -0,0161 | 0,1649 | 0,9224 |
| Governador Democrata | 0,016 | 0,1649 | 0,9225 |

Ao nível de significância de 5%, 8 das 23 variáveis possuem evidências estatísticas que são significantes para o tempo de adesão da política, sendo elas: Ideologia dos cidadãos, Porcentagem de políticas de Saúde 1912-2017 e 1990-2017, Porcentagem de fumantes, Lei de proibição para fumar em restaurantes, Profissionalismo Legislativo, Restrição para fumar em prédio do governo e Composição da Câmara e Senado. As duas variáveis categóricas (Lei para restaurantes e Composição) se mostraram capazes de explicar o tempo de falha, como foi observado nas Figuras 14 e 13, respectivamente.

Para encontrar o modelo final, primeiro foi feita uma análise do modelo incluindo todas as 8 variáveis citadas anteriormente e em seguida, foi retirada as variáveis menos significativas até encontrar um modelo. A partir deste modelo, foi testada a entrada e

saída de cada covariável no modelo, mantendo o nível de significância de 5%. A seleção de variáveis terminou quando a entrada de mais nenhuma covariável se mostrou significativa. Após toda análise, foi encontrado o modelo final com 5 variáveis explicativas, sendo elas, Profissionalismo Legislativo, Porcentagem de fumantes, Porcentagem de políticas de Saúde de 1990-2017, Lobistas na indústria do Tabaco e Restrição de acesso aos jovens. As estimativas do modelo final são mostradas na Tabela 3.

Tabela 3: Estimativas dos parâmetros do modelo de regressão $LLD$

| Variável | Estimativa | Erro Padrão | P-valor |
|---|---|---|---|
| $\gamma$ | 5,8301 | 1,0778 | - |
| Intercepto | 3,3916 | 0,6449 | <0,0001 |
| Profissionalismo Legislativo | -2,0404 | 0,6203 | 0,0010 |
| Porcentagem de fumantes | 0,0637 | 0,0219 | 0,0036 |
| Porcentagem de políticas de saúde 1990-2017 | -0,0203 | 0,0050 | <0,0001 |
| Lobistas na indústria do Tabaco | -14,3027 | 5,8205 | 0,0140 |
| Restrição de acesso aos jovens | -0,2836 | 0,1172 | 0,0155 |

O parâmetro $\alpha$ não está descrito na Tabela 3, pois como definido na Seção (4.4) é realizada uma reparametrização neste parâmetro para a inclusão de covariáveis. Dando origem assim, ao modelo de regressão Log-Logístico discreto. Logo, a estimativa de $\alpha$ pode ser obtida por:

$$\hat{\alpha} = \exp(\boldsymbol{x^T\beta}) = exp(\hat{\beta}_0 + \hat{\beta}_1 Prof.Legis. + \hat{\beta}_2 P.Fumantes + \ \ldots \ + \hat{\beta}_5 Rest.Jovens).$$

Todas as 5 variáveis inclusas no modelo completo são quantitativas, logo, pensando em suas interpretações, observa-se que para as covariáveis com coeficientes negativos, Profissionalismo Legislativo, Porcentagem de políticas de saúde, Lobistas e Restrição de acesso aos jovens, quanto maior for o valor destas variáveis, maior a chance do estado "falhar", aderir a política de proibição de fumar. Apenas para Porcentagem de Fumantes que quanto maior o valor, observa-se um aumento no tempo de sobrevivência, logo, menos chance de adesão pelos estados.

## 5.3 Ajuste da Distribuição Kumaraswamy Log-Logística Discreta

O ajuste da distribuição Log-Logística Discreta aos dados foi realizada com o *software R* utilizando a função *optim*. Ao ajustar a distribuição $KwLLD$ ao banco e comparando com a curva de Kaplan-Meier, representada pela Figura 17, observa-se, graficamente, que a distribuição escolhida se ajusta bem aos dados em estudo e comparando

com a curva estimada pela distribuição $LLD$, nota-se que a distribuição de probabilidade $KwLLD$ obteve um ajuste ainda melhor para a probabilidade estimada de sobrevivência dos estados. As estimativas dos parâmetros $\alpha$ e $\gamma$ são apresentadas na Tabela 4. Visto que os valores dos parâmetros apresentados são bem elevados e possuem um erro padrão pequeno, um modelo de regressão se mostra útil para investigar os possíveis fatores que influenciam na adoção da política em estudo.

Tabela 4: Estimativas dos parâmetros da distribuição $KwLLD$

| Variável | Estimativa | Erro Padrão |
|---|---|---|
| $\alpha$ | 9,7630 | 0,0648 |
| $\gamma$ | 50,7514 | 0,0673 |
| $a$ | 0,0056 | 0,0067 |
| $b$ | 0,0226 | 0,0048 |

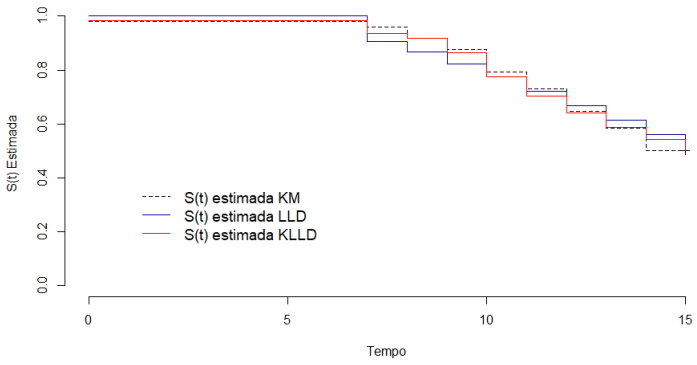


Figura 17: Comparação do ajuste das distribuições Kumaraswamy Log-Logística Discreta e Log-Logística Discreta com a função de sobrevivência estimada por Kaplan-Meier

Com o objetivo de propor um modelo de regressão para analisar os dados, primeiramente será ajustado um modelo com a presença de uma covariável para cada variável explicativa, sendo agora o modelo $KwLLD$, para verificar as covariáveis significantes. Utilizando o nível de significância de 5%, a Tabela 5 apresenta as estimativas, erro-padrão e p-valor associado a cada parâmetro de regressão.

Tabela 5: Estimativas dos coeficientes para o modelo $KwLLD$ com apenas uma covariável

| Variável | Estimativa | Erro Padrão | P-valor |
|---|---|---|---|
| Ideologia dos cidadãos | -0,0145 | 0,0033 | <0,0001 |
| Porcentagem de políticas de Saúde 1912-2017 | -0,0358 | 0,0088 | <0,0001 |
| Porcentagem de políticas de Saúde 1990-2017 | -1,5486 | 0,3938 | <0,0001 |
| Número de estados vizinho já adotantes | 0,0922 | 0,0243 | <0,0001 |
| Ano Eleitoral | -0,4941 | 0,1320 | 0,0002 |
| Porcentagem de Fumantes | -0,1098 | 0,0328 | 0,0008 |
| Profissionalismo Legislativo | -0,6282 | 0,2044 | 0,0021 |
| Margem de vitória | 0,0084 | 0,0028 | 0,0026 |
| Lei de proibição para fumar em restaurantes | -0,2708 | 0,0975 | 0,0055 |
| Restrição de acesso aos jovens | -1,2145 | 0,5475 | 0,0265 |
| Lei de proibição de fumar em prédios do governo | -0,3130 | 0,1447 | 0,0305 |
| Produção de Tabaco | 0,2558 | 0,1275 | 0,0449 |
| Ideologia do Governo | -0,5343 | 0,2875 | 0,0632 |
| Composição da Câmara e Senado | -0,1996 | 0,1088 | 0,0665 |
| Restrição para fumar em prédios do governo | -0,4678 | 0,2987 | 0,1173 |
| Lobistas na indústria do Tabaco | -11,4281 | 8,9513 | 0,2017 |
| Restrição para fumar em restaurantes | -0,6361 | 0,5472 | 0,2451 |
| Lei de proibição de fumar para jovens | -0,1719 | 0,1590 | 0,2798 |
| Governador Reeleito | 0,0904 | 0,0988 | 0,3604 |
| População | 0,0434 | 0,0742 | 0,5589 |
| Lobistas em organizações da saúde | 0,4246 | 0,7869 | 0,5894 |
| Governador Democrata | -0,0292 | 0,1460 | 0,8415 |
| Governador Republicano | 0,0288 | 0,1471 | 0,8448 |

Ao nível de 5%, observa-se que 12 variáveis possuem evidências estatísticas que são significantes para o tempo de adesão da política, sendo elas: Ideologia dos cidadãos, Porcentagem de políticas de Saúde 1912-2017 e 1990-2017, Número de vizinhos já adotantes, Ano eleitoral, Porcentagem de fumantes, Profissionalismo Legislativo, Margem de Vitória, Lei de proibição para fumar em restaurantes, para prédios do governo, Restrição de acesso aos jovens e Produção de Tabaco. As suspeitas obtidas na análise descritiva, Figuras 13, 14, 10, foram confirmadas, dado que as variáveis categóricas, Lei de proibição para fumar em restaurantes, prédios governamentais e Ano eleitoral se mostraram capazes de explicar o tempo de falha.

Para encontrar o modelo final, primeiro foi feita uma análise do modelo incluindo todas as 12 variáveis citadas anteriormente e em seguida, foi retirada as variáveis menos significativas até encontrar um modelo. A partir deste modelo, foi testada a entrada e

saída de cada covariável no modelo, mantendo o nível de significância de 5%. Por fim, quando a entrada de mais nenhuma covariável foi significativa, encontrou-se o modelo final com 6 variáveis explicativas, sendo elas, Profissionalismo Legislativo, Porcentagem de Fumantes, Ideologia do Governo, Lobistas em organizações da Saúde, Produção de Tabaco e Ano Eleitoral. Tirando a variável Lobistas na Saúde, todas as outras variáveis incluídas no modelo final se mostraram significativas, pela Tabela 5 assumindo 10% de significância. As estimativas dos coeficientes são apresentadas na Tabela 6. O parâmetro $\alpha$ não está descrito na tabela 6, pois como definido na Seção (4.5) é realizada uma reparametrização neste parâmetro para a inclusão de covariáveis. Dando origem assim, ao modelo de regressão Kumaraswamy Log-Logístico discreto. Logo, a estimativa de $\alpha$ pode ser obtida por:

$$\hat{\alpha} = \exp(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}) = exp(\hat{\beta}_0 + \hat{\beta}_1 Prof.Legis. + \hat{\beta}_2 P.Fumantes + \ldots + \hat{\beta}_6 AnoEleitoral).$$

Tabela 6: Estimativas dos parâmetros do modelo de regressão $KwLLD$

| Variável | Estimativa | Erro Padrão | P-valor |
|---|---|---|---|
| $\gamma$ | 55,9225 | 0,1271 | - |
| $a$ | 0,0220 | 0,0121 | - |
| $b$ | 0,0965 | 0,0320 | - |
| Intercepto | 2,0641 | 0,1328 | <0,0001 |
| Profissionalismo Legislativo | -1,4573 | 0,1923 | <0,0001 |
| Porcentagem de Fumantes | 0,0640 | 0,0100 | <0,0001 |
| Ideologia de Governo | -0,0142 | 0,0025 | <0,0001 |
| Lobistas em organizações da Saúde | 1,5224 | 0,1367 | <0,0001 |
| Produção de Tabaco | 0,1571 | 0,0643 | 0,0145 |
| Ano Eleitoral | -0,6955 | 0,0989 | <0,0001 |

Através dos resultados encontrados na Tabela 6, é possível interpretar a influência destas variáveis no tempo de falha do estudo. Neste modelo foram incluídas duas variáveis categóricas (dicotômicas), Ano Eleitoral e Produção de Tabaco. Como a primeira apresenta coeficiente estimado negativo, o fato de ser ano de eleição no estado aumenta o risco do mesmo aderir a política de Saúde. Já para a variável Produção de Tabaco, seu coeficiente estimado assume valor positivo, portanto, estados americanos que produzem tabaco possuem menos chance de adotar a proibição de fumar.

No caso das outras variáveis, as quais são quantitativas, as estimativas de Profissionalismo Legislativo e Ideologia de Governo indicam que o aumento em uma unidade das mesmas reduz o tempo de sobrevivência, ou seja, aumenta a chance do estado aderir a política de proibição de fumar. Já os coeficientes estimados de Porcentagem de Fumantes e Lobistas em organizações da Saúde indicam que o aumento de uma unidade percentual

no valor das variáveis aumenta o tempo de sobrevivência, ou seja, quanto mais fumantes e lobistas em organizações da saúde um estado possuir, menor será a chance de falha do mesmo.

# 6 Conclusões

Este trabalho teve como objetivo estudar a aplicabilidade da distribuição Kumarraswamy Log-Logística discreta em dados de sobrevivência com tempos discretos e compará-la a seu caso particular, a distribuição Log-Logística discreta. Neste caso, foi ajustado os modelos de regressão $KwLLD$ e $LLD$ ao banco de dados proposto para ser possível a comparação entre os dois modelos de regressão.

Para a classificação e seleção de modelos, é possível encontrar o "melhor"modelo utilizando critérios estatísticos definidos na literatura. Os valores dos critérios são informados na Tabela 7. Akaike (1974), Hurvich e Tsai (1989) e Schwarz (1978) concluem que o modelo final escolhido, deve apresentar o menor valor dos critérios propostos por eles (AIC, AICc e BIC, respectivamente). Desta forma, conclui-se que a distribuição Log-Logística discreta obteve um ajuste melhor com seu modelo de regressão do que a distribuição Kumaraswamy Log-Logística discreta.

Tabela 7: Medidas de informação dos modelos de regressão $LLD$ e $KwLLD$

| Modelo | AIC | AICc | BIC |
|---|---|---|---|
| $LLD$ | 152,54 | 155,54 | 165,84 |
| $KwLLD$ | 165,39 | 171,34 | 184,10 |

Entretanto, a distribuição $KwLLD$ apresenta um ganho com seu modelo, em relação a distribuição $LLD$, por apresentar mais variáveis explicativas no modelo final. Outro ganho obtido com o modelo $KwLLD$, seria o acréscimo de dois parâmetros de forma ($a$ e $b$) que permitem uma maior flexibilidade da função de risco.

Devido a distribuição ter sido recentemente proposta, pouco se sabe sobre suas propriedades. Por este motivo, muitas instabilidades computacionais foram encontradas neste trabalho, sendo estes, problemas de convergência e na estimação dos parâmetros dos modelos. Seria de grande interesse trabalhos futuros proporem o estudo mais aprofundado da distribuição Kumaraswamy Log-Logística discreta, com a realização de simulações computacionais afim de validar e justificar a aplicação do modelo em análise de sobrevivência.

# Referências

AALEN, O. Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, v. 6, n. 4, 1978.

AARSET, M. V. How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, R-36, n. 1, April 1987.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974.

BERRY, W. D. et al. Measuring citizen and government ideology in the american states, 1960-93. *American Journal of Political Science*, [Midwest Political Science Association, Wiley], v. 42, n. 1, p. 327–348, 1998.

BOHORIS, G. A. Comparison of the cumulative-hazard and kaplan-meier estimators of the survivor function. *IEEE Transactions on Reliability*, v. 43, n. 2, 1994.

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

CARRASCO, J.; ORTEGA, E.; CORDEIRO, G. A generalized modified weibull distribution for lifetime modeling. *Computational Statistics & Data Analysis*, v. 53, 12 2008.

COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. [S.l.]: Edgar Blücher, São Paulo, 2006. 1-87 p. ABE - Projeto Fisher.

CORDEIRO, G. M.; CASTRO, M. de. A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, v. 81, n. 7, 2011.

DIAS, T. C. M.; FOGO, J. C. *Introdução ao LaTeX*. 2014. Slides de Minicurso do DEs - UFSCar. Utilizado com autorização dos autores.

FERNANDES, L. M. *Inferência Bayesiana em modelos discretos com fração de cura*. Dissertação (Mestrado em Estatística) — Universidade de Brasília - Departamento de Estatística, Brasília, 2013.

GEHAN, E. A. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, v. 52, n. 1-2, 06 1965.

HOEL, P. G.; PORT, S. C.; STONE, C. J. Introdução a teoria da probabilidade. Editora lnterciencia Ltda., Rio de Janeiro, 1978.

HURVICH, C. M.; TSAI, C.-L. Regression and time series model selection in small samples. *Biometrika*, v. 76, n. 2, p. 297–307, 06 1989.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, v. 53, n. 282, 1958.

KUMARASWAMY, P. A generalized probability density function for double-bounded random processes. *Journal of Hidrology*, Março 1980.

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

LIMA, J. G. *Modelo de Regressão para Dados da Política Zoneamento e Uso do Solo na Presença de Observações Censuradas*. Dissertação (Bacharelado em Estatística) — Universidade de Brasília - Departamento de Estatística, Brasília, 2018.

MANTEL, N. Models for complex contingency tables and polychotomous dosage response curves. International Biometric Society, v. 22, 1966.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K.; FREIMER, M. The exponentiated weibull family: A reanalysis of the bus-motor-failure data. *Technometrics*, v. 37, n. 4, 1995.

NAKANO, E. Y. *Um curso de Análise de Sobrevivência*. 2018.

NAKANO, E. Y.; CARRASCO, C. G. Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência. tema - tend. mat. apl. comput. v. 7, n. 1, p. 91–100, 2006.

NELSON, W. Theory and applications of hazard plotting for censored failure data. *Technometrics*, Taylor & Francis, v. 14, n. 4, 1972.

RIZZO, M. L. *Statistical Computing with R*. [S.l.]: Chapman and Hall, 2008.

SANTOS, D. F. dos. *Modelo de Regressão Log-Logístico discreto com fração de cura para dados de sobrevivência*. Dissertação (Mestrado em Estatística) — Universidade de Brasília - Departamento de Estatística, Brasília, 2017.

SCHWARZ, G. Estimating the dimension of a model. *Ann. Statist.*, The Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 03 1978.

SILVA, G. O.; ORTEGA, E. M. M.; CORDEIRO, G. M. The beta modified weibull distribution. *Lifetime Data Analysis*, v. 16, n. 3, 2010.

SIMÕES E SILVA, L. B. *Modelo de Regressão Kumaraswamy Log-Logístico para dados discreto*. Dissertação (Bacharelado em Estatística) — Universidade de Brasília - Departamento de Estatística, Brasília, 2019.

SQUIRE, P. Measuring state legislative professionalism: The squire index revisited. *State Politics & Policy Quarterly*, Sage Publications, Inc., v. 7, n. 2, p. 211–227, 2007.