Universidade de Brasília
Departamento de Estatística

Modelo de regressão para dados de peixes da espécie *Notropis Dourado, Crysoleucas de Notemigonus* via técnicas de análise de sobrevivência

Raul Henrique Athayde Braz

Monografia apresentada para obtenção do título de Bacharel em Estatística.

Brasília
2018

Raul Henrique Athayde Braz
Bacharel em Estatística

Modelo de regressão para dados de peixes da espécie *Notropis Dourado, Crysoleucas de Notemigonus* via técnicas de análise de sobrevivência

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

Monografia apresentada para obtenção do título de Bacharel em Estatística.

Brasília
2018

AGRADECIMENTOS

A professora Dra. **Juliana Betini Fachini Gomes**, pela presença constante e o apoio em cada etapa deste trabalho.

Ao **Observatório da Juventude da UnB**, que fez parte do meus melhores momentos dentro da Universidade.

Á minha avó **Elza**, ao meu pai **Damião**, a minha madrinha **Danielle** e a minha namorada **Rafaella**, que fora da Universidade também se fizeram presentes em diversos momentos da vida acadêmica.

Aos colegas **Alex**, **Cauan**, **Flávia** e **Taíssa** que ao longo da graduação partilharam o estudo e a boa companhia.

Aos professores do departamento de estatística e matemática, por todas as aulas e o conhecimento compartilhado ao longo desses anos.

Á Universidade de Brasília como um todo que me recebeu de braços abertos como aluno e me fez sentir em casa.

SUMÁRIO

RESUMO

Modelo de regressão para dados de peixes da espécie *Notropis Dourado, Crysoleucas de Notemigonus* via técnicas de análise de sobrevivência

Neste trabalho foram utilizados os modelos de regressão do valor extremo e logístico a partir de técnicas oriundas da análise de sobrevivência. Os parâmetros dos modelos foram estimados pelo método de máxima verossimilhança. Uma análise de resíduo foi feita para verificar a qualidade do ajuste global dos modelos. Um conjunto de dados reais de peixes da espécie *"Notropis Dourado, Crysoleucas de Notemigonus"* foram utilizados para a análise e a aplicação dos modelos propostos.

Palavras-chave: Modelos de regressão, análise de sobrevivência, dados censurados, análise de resíduo.

1 INTRODUÇÃO

Análise de sobrevivência é uma área da estatística que estuda modelos e técnicas para analisar o tempo até que um evento de interesse ocorra. Evento esse que quando ocorre é denominado como tempo de falha. Estudos de sobrevivência são bastante comuns na área médica, principalmente, onde o tempo de falha poderia por exemplo ser a morte de um paciente. No entanto, é cada vez mais comum que diversas áreas do conhecimento busquem técnicas de análise de sobrevivência afim de compreender o seu estudo.

A resposta do estudo em sobrevivência é bastante explicada pelo tempo de falha e pela censura, que é a principal característica dos dados de sobrevivência. A censura é observada quando o evento de interesse não ocorre em um indivíduo durante o período de estudo, gerando, assim, observações incompletas. É um evento corriqueiro em estudos de longa duração, onde é comum que exista perda no acompanhamento de alguns indivíduos em algum momento do tempo, ou mesmo se o tempo do estudo finalizar e o indivíduo em questão não apresentar a falha. A presença de censura torna difícil uma análise descritiva tradicional dos dados, como média e variância. Entretanto, eliminar os dados censurados acarretariam em conclusões viciadas que prejudicariam qualquer conclusão sobre o tema abordado. Os métodos de sobrevivência permitem incorporar esses dados censurados e tornam o estudo mais consistente.

Ao considerar as duas características acima, tempo de falha e censura, e ainda a presença de variáveis explicativas nos dados, este trabalho admitirá como foco principal propor um modelo de regressão em que os tempos assumem distribuição da família Gumbel, também conhecida como valor extremo, para um conjunto de dados de sobrevivência de peixes da espécie *"Notropis Dourado, Crysoleucas de Notemigonus"*. Toda a análise será realizada no *software* R e em particular através do pacote *Survival*.

2 REVISÃO DE LITERATURA

Com o intuito de fundamentar o presente trabalho que será desenvolvido utilizando técnicas de análise de sobrevivência, serão expostos, a seguir, conceitos e notações utilizadas na literatura.

2.1 Tempo de falha

Uma característica essencial aos dados de análise de sobrevivência é o tempo até que o evento de interesse do estudo ocorra, denominado como tempo de falha. Esse tempo é o intervalo medido entre o início do estudo até a ocorrência da falha.

Segundo Colosimo e Giolo (2006), esse período de tempo precisa ser bem definido, assim como o seu ponto inicial. No entanto, não é necessário que todos os elementos do estudo entrem no mesmo tempo inicial. Eles podem entrar a qualquer momento do intervalo estipulado e serão observados até o tempo final do estudo. É necessário que a escala de medida seja a mesma para todos os indivíduos e é importante que isso seja feito para garantir que os tempos observados no estudo até a falha estejam na mesma escala e possam ser efetivamente comparados.

2.2 Censura

A censura é observada quando o evento de interesse não ocorre em algum indivíduo durante o período de estudo, gerando, assim, observações incompletas. É um evento corriqueiro em estudos de longa duração, onde é comum que exista perda no acompanhamento de alguns indivíduos em algum momento do tempo. Ou mesmo se o tempo do estudo finalizar e o indivíduo em questão não apresentar a falha. A presença de censura torna difícil uma análise descritiva tradicional dos dados, como média e variância. Entretanto, eliminar os dados censurados acarretaria em conclusões viciadas que prejudicariam qualquer conclusão sobre o tema abordado. Os métodos de sobrevivência permitem incorporar esses dados censurados e tornam o estudo mais consistente.

Existem três tipos de censura: Censura à esquerda, censura intervalar e censura à direita, que podem ou não ocorrer dependendo do objeto de estudo. A censura à esquerda ocorre quando o tempo verificado é maior que o tempo de falha, ou seja, o evento de interesse já ocorreu quando o indivíduo em questão foi observado. A censura intervalar ocorre quando o momento no tempo onde ocorre o evento de interesse não é exato, e sim pertence a um intervalo conhecido. A censura à direita ocorre quando o tempo de ocorrência de um evento

está além do tempo observado, em outras palavras, o evento não ocorreu ao fim do limite temporal estipulado.

A seguir serão definidas algumas funções que serão utilizadas para representar a variável resposta do estudo. Bem como algumas técnicas descritivas para dados com censura.

2.3 Representando o tempo de sobrevivência

A função de sobrevivência é definida como a probabilidade de uma observação sobreviver ao tempo t ou como a probabilidade de não falhar até um determinado tempo t, e é definida por:

$$S(t) = P(T \geq t).$$

Outra forma de se encontrar a função de sobrevivência é utilizando a função acumulada F(t). Isto é:

$$S(t) = 1 - F(t).$$

É importante destacar que a S(t) é uma função monótona, decrescente e contínua (Lawless, 2003).

2.4 Função da taxa de falha

A função da taxa de falha, também conhecida como função de risco, representa a taxa de falha instantânea no tempo t condicional a sobrevivência até o tempo t. Considerando-se o intervalo $[t, t+\Delta t)$ e assumindo que a falha não tenha acontecido até o instante de tempo t, tem-se que:

$$\lambda(t) = \frac{S(t) - S(t+\Delta t)}{\Delta t \cdot S(t)}.$$

A função taxa de falha é útil para descrever a distribuição do tempo de sobrevivência do estudo, e nessa situação, ela é mais informativa que a função de sobrevivência. Isso ocorre uma vez que diferentes funções de sobrevivência mostram formas semelhantes, enquanto que a as funções de risco se diferem com facilidade. As formais mais usuais da função de risco são: constante, monótona (crescente e decrescente) e não monótonas (unimodal e a de banheira), como mostra a Figura 1 (NAKANO, 2017).

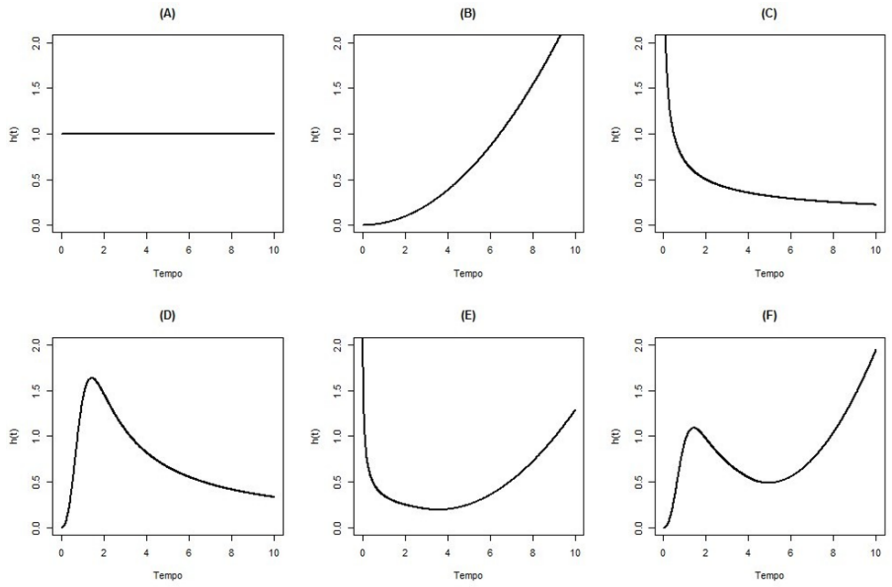

Figura 1 – Funções de rico: constante (A), crescente (B), decrescente (C), unimodal (D), banheira (E) e multimodal (F).

2.5 Estimador de Kaplan-Meier

Em estudos de sobrevivência, a presença de censura torna difícil uma análise descritiva tradicional dos dados, como média e variância. Entretanto, eliminar os dados censurados acarretariam em conclusões viciadas que prejudicariam qualquer conclusão sobre o tema abordado. Os métodos de sobrevivência permitem incorporar esses dados censurados e tornam o estudo mais consistente. Um deles é o estimador de Kaplan-Meier da função de sobrevivência.

A utilização desse estimador necessita de uma sequência de passos, em que a informação do passo anterior é crucial para a obtenção do passo seguinte. Como considerações preliminares que devem ser feitas para a utilização do estimador de Kaplan-Meier, temos (COLOSIMO e GIOLO, 2006):

- $t_1 < t_2 < ... < t_k$ são os k tempos distintos e ordenados de falha;
- d_j é o número de falhas em $t_j, j = 1, ..., k$ e
- n_j é o número de indivíduos sob risco (isto é, que não falharam) em t_j.

Assim, o estimador de Kaplan-Meier é definido como:

$$\widehat{S}(t) = \prod_{j:tj<t} \left(\frac{nj - dj}{nj} \right) = \prod_{j:tj<t} \left(1 - \frac{dj}{nj} \right).$$

Segundo Colosimo e Giolo (2006), o estimador de Kaplan Meier tem como característica ser uma função escada, em que cada degrau ocorre no instante de tempo t que ocorre a sua respectiva falha. Em outras palavras, cada vez que ocorrer uma falha no estudo, existirá um degrau a mais na escada. Esse estimador tem a propriedade de ser um estimador de máxima verossimilhança para a função de sobrevivência, $S(t)$.

2.6 Determinação empírica da forma da função de risco

Após a utilização de técnicas não paramétricas afim de reconhecer o comportamento dos dados, o próximo passo é descobrir qual função de probabilidade consegue modelar melhor os dados. Para esse fim, é importante utilizar uma metodologia adequada para identificar o modelo mais apropriado. Uma técnica reconhecida de verificação gráfica é a curva do tempo total em teste (curva TTT), proposto por Aarset (1987).

A curva TTT é obtida construindo um gráfico de:

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

por r/n, sendo que $r = 1, ..., n$ e $T_{i:n}, i = 1, ..., n$ são estatísticas de ordem da amostra. A curva gerada seguindo esse gráfico é associada a uma forma diferente da função da taxa de falha.

A Figura 2 possibilita observar algumas das possíveis formas da curva TTT, em que cada uma delas significa:

- A: A reta em formato diagonal representado pela letra A apresenta uma função de risco constante.

- B: A curva B apresenta um formato convexo e está relacionado a função de risco decrescente.

- C: A curva C é côncava e está relacionada a uma função de taxa de falha crescente.

- D: A curva começa convexa e depois apresenta um formato côncavo, assim assume-se que a função taxa de falha possui a forma de banheira.

- E: Por fim, a curva E começa côncava e depois assume-se convexo. A função de risco é unimodal.

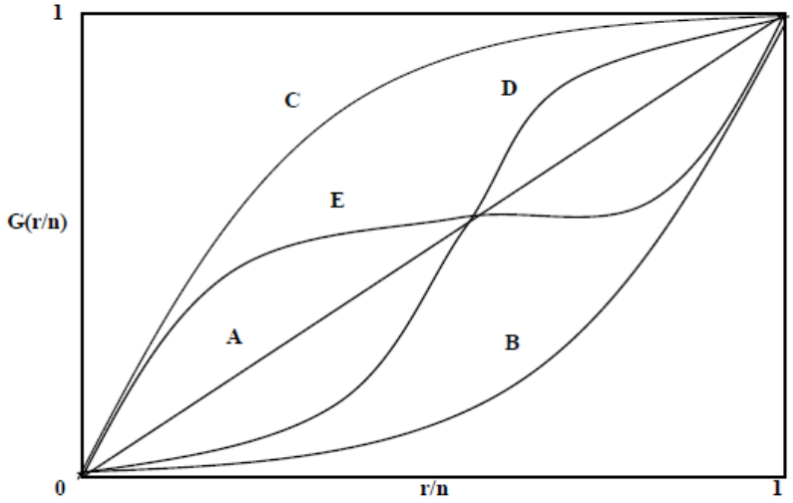

Figura 2 – Gráfico ilustrativo de algumas curvas TTT

Assim, é através da curva TTT que será encontrada a função de risco, que irá restringir quais funções de probabilidade irão melhor refletir os dados. Esse é mais um motivo pelo qual esse tipo de metodologia paramétrica será utilizada para analisar descritivamente os dados.

2.7 Modelos probabilísticos

Conforme dito anteriormente, a utilização de métodos não paramétricos tem como objetivo encontrar uma distribuição de probabilidade que reflita o comportamento dos dados. Uma vez que ela é encontrada, podemos calcular a probabilidade de ocorrência nos intervalos dos valores da variável estudada. Existem diversas distribuições de probabilidade conhecidas com diferentes particularidades, o que faz com que cada distribuição seja mais eficaz para cada tipo de variável aleatória.

Pensando em análise de sobrevivência, existem distribuições que são descartadas imediatamente, uma vez que a variável resposta não assume tempos negativos. A partir dessa particularidade, as possíveis distribuições capazes de modelar os dados de sobrevivência são as distribuições onde a variável aleatória é definida para valores maiores ou iguais a zero.

2.8 Distribuição Exponencial

Essa distribuição é reconhecida como um dos modelos probabilísticos mais simples usado para descrever o tempo de falha. A distribuição exponencial apresenta um único

parâmetro e é particularmente caracterizada por apresentar uma taxa de falha constante, o que a diferencia das demais distribuições. Isso significa que tanto uma unidade velha quanto uma nova, que ainda não apresentaram a falha, possuem a mesma taxa de falha em um intervalo futuro. Em geral, é utilizada com o intuito de descrever o tempo de vida de determinados materiais.

A sua função densidade de probabilidade é dada por:

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

em que o parâmetro $\alpha > 0$ é o tempo médio de vida. O parâmetro α tem a mesma unidade do tempo de falha t. Assim, se t é medido em anos, α também será fornecido em anos.

Já em análise de sobrevivência, suas funções de sobrevivência $S(t)$ e de taxa de falha $\lambda(t)$ são definidas por:

$$S(t) = \exp\left\{-\frac{t}{\alpha}\right\},$$

e

$$\lambda(t) = \frac{1}{\alpha} \text{ para } t \geq 0.$$

Por apresentar uma taxa de falha constante, qualquer unidade nova ou velha que ainda não tenha falhado apresenta a mesma taxa de falha em qualquer intervalo futuro.

2.9 Distribuição de Weibull

Essa distribuição é mais utilizada para modelar dados de sobrevivência, uma vez que ela é adaptável a uma grande variedade de formas. Como característica principal, a distribuição de Weibull é monótona, ou seja, ela decresce, cresce ou mantem-se constante. Nesse último caso ela torna-se um caso particular visto anteriormente: a distribuição exponencial. Se a variável aleatória seguir a distribuição de Weibull, ela terá a seguinte função de densidade de probabilidade:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, t \geq 0, \tag{2}$$

em que $\gamma > 0$ é o parâmetro de forma e $\alpha > 0$, é o parâmetro de escala. O parâmetro α tem a mesma unidade de medida de t e γ não tem unidade. Suas funções de sobrevivência e taxa de falha são, respectivamente:

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

e

$$\lambda(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1},$$

para t ≥ 0, α e $\gamma > 0$.

A função da taxa de falha $\lambda(t)$ é estritamente crescente para $\gamma > 0$, estritamente decrescente para $\gamma < 1$ e constante para $\gamma = 0$. No caso de $\gamma = 1$, tem-se a função de taxa de falha da distribuição exponencial.

2.10 Distribuição do valor extremo

É importante destacar uma distribuição que está relacionada com a distribuição Weibull. Ela é denominada como distribuição do valor extremo ou de Gambel e será utilizada quando tomamos o logaritmo de uma variável com distribuição de Weibull. Ou seja, se uma variável X assume distribuição Weibull com sua respectiva f(t), definida na equação (2), então a variável Y = log(X) assume distribuição valor extremo com função de densidade:

$$f(y) = \frac{1}{\alpha} \exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}, \tag{3}$$

em que y e $\mu \in \mathbb{R}$ e $\sigma > 0$. Se $\mu = 0$ e $\sigma = 1$ temos a distribuição padrão do valor extremo. Os parâmetros μ e σ são denominados parâmetros de locação e escala, respectivamente. A seguir estão as funções de sobrevivência e taxa de falha, associada a função de densidade definida na equação (3), respectivamente:

$$S(y) = \exp\left\{-\exp\left(\frac{y-\mu}{\sigma}\right)\right\}$$

e

$$\lambda(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\}.$$

Na análise de sobrevivência, muitas vezes é conveniente trabalhar com o logaritmo dos tempos de vida observados, por isso, se os dados possuírem uma distribuição de Weibull, a distribuição do valor extremo aparecerá naturalmente na modelagem.

2.11 Distribuição Log-logística

Trata-se de uma distribuição que serve de alternativa à distribuição Weibull e Log-normal. A diferença básica entre essa distribuição e a Log-normal são que suas funções de sobrevivência e taxa de falha apresentam formas explícitas.

Sua função densidade de probabilidade é dada por:

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}(1 + (t/\alpha)^\gamma)^{-2},\ t > 0, \tag{4}$$

em que $\alpha > 0$ é o parâmetro de escala e $\gamma > 0$ o de forma. Suas funções de sobrevivência e taxa de falha são, respectivamente:

$$S(t) = \frac{1}{1 + (t/\alpha)^\gamma}$$

e

$$\lambda(t) = \frac{\gamma(t/\alpha)^{\gamma-1}}{\alpha[1 + (t/\alpha)^\gamma]}.$$

No caso da distribuição log-logística, as funções da taxa de falha não são monótonas como as da distribuição Weibull. Elas crescem até atingir um valor máximo e depois decrescem.

2.12 Distribuição Burr XII

A distribuição Burr XII (SILVA, 2008), com parâmetros s, c e k considera que o tempo de sobrevivência T ($t > 0$) tem função densidade dada por:

$$f(t; s, k, c) = \frac{t^{c-1}}{s^c} ck \left[1 + \left(\frac{t}{s}\right)^c\right]^{(-k-1)} \tag{5}$$

em que $s > 0$ é o parâmetro de escala e $k > 0$ e $c > 0$ são os parâmetros de forma, o que a torna mais flexível. A função de sobrevivência e sua função da taxa de falha são dadas, respectivamente, por:

$$S(t; s, k, c) = P(T \geq t) = \left[1 + \left(\frac{t}{s}\right)^c\right]^{-k}$$

e

$$\lambda(t; s, k, c) = P(T \geq t) = \frac{ck\left(\frac{t}{s}\right)^{(c-1)}}{s\left[1 + \left(\frac{t}{s}\right)^c\right]}.$$

A função da taxa de falha da distribuição Burr XII é decrescente quando $c \leq 1$ e λ é unimodal quando $c > 1$. A partir da função de densidade da Burr XII dada na equação (5), observa-se que quando $\frac{1}{s} = m$ e $k = 1$, a distribuição Burr XII se reduz à distribuição log-logística com função de sobrevivência dada por $S(t; c, s) = \frac{1}{1+(tm)^c}$

2.13 Estimação dos parâmetros

Dentre os diversos métodos existentes na estatística afim de se estimar os valores dos parâmetros dos modelos probabilísticos, o estimador de máxima verossimilhança se destaca em estudos de sobrevivência, uma vez que esse estimador é capaz de incorporar em sua abordagem de cálculo as censuras. A seguir, será feita uma melhor descrição desse método.

2.14 Estimador de máxima verossimilhança

É a partir dos resultados obtidos na amostra que se realiza a estimação dos parâmetros através do método de máxima verossimilhança. O objetivo desse método é encontrar a distribuição de probabilidade que tenha a maior probabilidade de ter gerado a amostra em estudo. A partir disso, o método encontrará os valores de $\boldsymbol{\theta}$ que maximizam a função de verossimilhança $L(\boldsymbol{\theta})$, uma vez que são esses valores que possuem a maior probabilidade de ter gerado os valores da amostra observada. Dessa forma, a função de verossimilhança para uma amostra de dados sem a presença de observações censuradas é definida por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_{\mathrm{i}}; \boldsymbol{\theta}),$$

sendo n o número total de observações e $\boldsymbol{\theta}$ o vetor de parâmetros que pode apresentar um ou mais parâmetros. Ao considerar uma amostra aleatória observada (t_1,δ_1), ... ,($\mathrm{t_n}$,δ_{n}), em que $\mathrm{t_i}$, i = 1,...,n, representa o tempo de falha ou tempo de censura e δ_{i} a respectiva variável indicadora de falha ou censura. A função de verossimilhança é definida por:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [f(t_{\mathrm{i}}; \boldsymbol{\theta})]^{\delta_{\mathrm{i}}} [S(t_{\mathrm{i}}; \boldsymbol{\theta})]^{1-\delta_{\mathrm{i}}},$$

é possível observar que a contribuição de uma observação que falhou é sua função de densidade e as observações com censura é sua função de sobrevivência. No entanto, é indicado trabalhar com o logaritmo da função de verossimilhança, definido por:

$$l(\boldsymbol{\theta}) = \log(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_{\mathrm{i}} \log[f(t_{\mathrm{i}}, \boldsymbol{\theta})] + (1 - \delta_{\mathrm{i}}) \log[S(t_{\mathrm{i}}, \boldsymbol{\theta})]$$

A seguir, deriva-se parcialmente a função $l(\theta)$ afim de se resolver o sistema de equações:

$$U(\boldsymbol{\theta}) = \frac{\partial log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0,$$

para encontrar os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ Por fim, em qualquer área de estudo, seja em análise de sobrevivência ou qualquer outro método estatístico, é comum o desejo de se encontrar uma relação entre a variável resposta e cada uma das demais variáveis presentes no estudo. Umas das formas mais indicadas de prosseguir nesse desejo é formular um modelo de regressão.

2.15 Modelo de regressão

Estudos em estatística, na área de sobrevivência ou não, procuram encontrar uma relação entre a variável resposta e as outras variáveis presentes no estudo. Por exemplo, no setor empregatício, pode ter o interesse em descobrir quanto tempo os funcionários permanecem em um determinado emprego, ou na área da saúde, o interesse pode ser o tempo até a morte de um paciente com câncer no fígado. Em ambas as situações a variável resposta pode ser influenciada por algumas covariáveis como: idade, sexo, salário no caso do emprego ou consumir bebidas alcoólicas na área médica. Uma das formas de relacionar a variável resposta, o tempo, as demais covariáveis é a utilização de um modelo de regressão. Neste trabalho, será utilizado o modelo de locação e escala afim de estudar a influência de algumas covariáveis. Além deste modelo, outros dois bastante utilizados são os modelos de riscos proporcionais e a parametrização da distribuição de probabilidade dos tempos para se obter um modelo de regressão. No entanto, neste trabalho tais métodos não serão abordados.

2.16 Modelo de regressão locação e escala

A classe de modelos de regressão denominada de modelos de locação e escala (Carrasco, 2007) tem uma característica essencial: trabalha com o logaritmo dos tempos de sobrevivência. Ou seja:

$$Y = \log(T)$$

Dessa forma, o modelo de locação e escala possuí a seguinte forma:

$$Y = \log(T) = \mu(x) + \sigma W,$$

em que $\mu(\boldsymbol{x})$ é o parâmetro de locação e $\sigma > 0$ representa o parâmetro de escala e W é o erro aleatório. Geralmente é considerado $\mu(x) = \boldsymbol{x}^T\boldsymbol{\beta}$ em que $\boldsymbol{x}^T$ representa o vetor de covariáveis $\boldsymbol{x}^T = (x_1, x_2, ..., x_p)^T$, sendo que p é a quantidade de covariáveis presentes no modelo de regressão, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ representam o vetor de parâmetros desconhecidos. Portanto, tem-se que o modelo de locação e escala é um modelo log linear para a variável T e as variáveis regressoras agem de forma multiplicativa sobre T.

2.17 Resíduos de Cox Snell

É importante avaliar se o modelo proposto está bem ajustado aos dados. Isso pode ser feito por meio de técnicas gráficas. Essas técnicas avaliam a distribuição dos erros e sua principal utilidade é rejeitar modelos inapropriados. Assim, o objetivo dessas técnicas não é aprovar um modelo particular, uma vez que é frequente que mais de um modelo possa ser utilizado. Neste trabalho os Resíduos de Cox Snell serão utilizados. Esses resíduos são quantidades determinadas por:

$$\widehat{e_i} = \widehat{H}(t_i),$$

em que $\widehat{H}(t_i)$ é a função de risco acumulado obtido do modelo ajustado.

Os resíduos $\widehat{e_i}$ são oriundos de uma população homogênea e seguem distribuição exponencial padrão caso o modelo seja adequado (Lawless, 2003). Para que o modelo exponencial seja adequado, o gráfico de $\widehat{e_i}$ versus $\widehat{H}(\widehat{e_i})$ deve ser aproximadamente uma reta. O gráfico das curvas de sobrevivência dos estudos, $\widehat{S}(\widehat{e_i})$, e pelo modelo exponencial padrão, $exp(-\widehat{e_i})$, também auxiliam na verificação da qualidade do modelo ajustado, quanto mais próximas no sentido de que o ajuste aos dados é melhor.

3 METODOLOGIA

3.1 Material

Os dados utilizados são oriundos de um experimento de campo realizado em 2005 no lago Saint Pierre, no Canadá (RODRÍGUEZ, 2010). O estudo incluiu um total de 106 peixes da espécie "*Notropis Dourado, crysoleucas de Notemigonus*", em que cada peixe foi unido através de uma corda a um dispositivo permitindo que o peixe nade em *midwater* e um cronômetro foi inserido nesse dispositivo de forma que a contagem do tempo é iniciada quando o peixe é capturado por um predador. Das 106 observações de tempo no estudo 15 foram censuradas (SILVA, 2008).

O pesquisador observou dois tempos. O primeiro deles é o tempo entre o início do experimento e a recuperação do dispositivo e o segundo é o tempo registrado no cronômetro. A obtenção do tempo de sobrevivência é realizada a partir da diferença entre esses dois tempos e caso o peixe não seja capturado por um predador em um período de 24 horas o dispositivo é então recuperado e a observação é considerada censurada. A resposta de interesse é o tempo em que o peixe levou para ser capturado por um predador.

O objetivo do estudo é identificar quais variáveis influenciam no tempo de captura do peixe. As variáveis explicativas consideradas nesse estudo são:

- y_i: tempo de sobrevivência em horas.
- δ_i: indicador de censura.
- x_{i1}: tamanho do peixe em centímetros.
- x_{i2}: profundidade do rio em centímetros.
- x_{i3}: transparência da água, em que i = 1,2,...,106.

3.2 Métodos: Modelo de Regressão em Análise de Sobrevivência

Ao considerar as distribuições Weibull e log-logística definidas na seção 2.7, este trabalho propõe utilizar os modelos de regressão da família locação e escala. De acordo com a seção 2.10, a distribuição do valor extremo é obtida ao considerar o logaritmo de uma variável aleatória com distribuição Weibull. Ou seja:

$$Y = \log(T)$$

Logo, partimos do pressuposto de que os dados assumem a seguinte função de probabilidade:

$$f(t) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}$$

Então, utilizando o método Jacobiano, a função densidade de probabilidade de Y é dada por:

$$f_Y(y) = f_T\left(\exp\{y\})|\mathbf{J}|\right)$$

Em que $|\mathbf{J}| = \frac{dt}{dy}$, sendo que $t = e^y$.

$$f_Y(y) = \frac{\gamma}{\alpha^{\gamma}} (\exp\{y\})^{\gamma-1} \exp\left\{-\left(\frac{\exp(y)}{\alpha}\right)^{\gamma}\right\} \exp\{y\}$$

$$f_Y(y) = \frac{\gamma}{\alpha^{\gamma}} \exp\left\{y^{\gamma} - \left(\frac{\exp(y)}{\alpha}\right)^{y}\right\}$$

Ao considerar as reparametrizações: $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp(\mu)$, a função densidade de probabilidade de Y é dada por:

$$f_Y(y) = \frac{1}{\sigma} \frac{1}{\exp\left(\frac{\mu}{\sigma}\right)} \exp\left\{\frac{y}{\sigma} - \left(\frac{\exp(y)}{\exp(u)}\right)^{\frac{1}{\sigma}}\right\}$$

$$f_Y(y) = \frac{1}{\sigma} exp\left\{\left(\frac{y-\mu}{\sigma}\right) - exp\left(\frac{y-\mu}{\sigma}\right)\right\}.$$

Ou seja, Y tem distribuição do valor extremo com parâmetros de locação, μ, e de escala, σ. Assim, essa distribuição pertence à família de modelos de locação e escala e Y pode ser reescrito como:

$$Y = \mu + \sigma W,$$

em que $W = \frac{Y-\mu}{\sigma}$, é o erro associado ao modelo e possuí distribuição dada por:

$$f(w) = f_Y\left(y + \mu W)|\mathbf{J}|\right)$$

$$f_W(w) = \frac{1}{\sigma} exp\left\{\left(\frac{\mu + \sigma w}{\sigma}\right) - exp\left(\frac{\mu + \sigma w}{\sigma}\right)\right\} \sigma$$

$$f_W(w) = exp\left\{w + exp\{w\}\right\}.$$

Portanto, W possuí densidade valor extremo.

Por fim, se considerarmos $\mu = \boldsymbol{x}^T\boldsymbol{\beta}$, temos que:

$$Y = \boldsymbol{x}^T\boldsymbol{\beta} + \sigma W$$

é a função de probabilidade do modelo de regressão do valor extremo ou também conhecida como função densidade de probabilidade do modelo de regressão log-weibull, e é definida por:

$$f(y|x) = \frac{1}{\sigma} exp\left\{\left(\frac{y - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right) - exp\left(\frac{y - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\right\}. \tag{6}$$

E a função de sobrevivência associada a equação (6) é definida por:

$$S(y|x) = exp\left\{-exp\left\{\left(\frac{y - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\right\},$$

em que $\boldsymbol{x}^T = (x_1, x_2, ..., x_p)^T$ é o vetor de covariáveis, e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ é o vetor de parâmetros desconhecidos. Esse modelo é conhecido como modelo de regressão Weibull, pois T deve ter uma distribuição de Weibull para que log(T) tenha uma distribuição de valor extremo com parâmetro de escala σ (Colosimo e Giolo, 2006).

De forma análoga, ao considerar que T tem distribuição log-logística, Y = log(T) terá distribuição logística e podemos encontrar que a função densidade de probabilidade do modelo de regressão logístico é definido por:

$$f(y|x) = \frac{1}{\sigma} \exp\left\{\left(\frac{y - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\left[1 + \exp\left(\frac{y - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\right]^{-2}\right.$$

e a função de sobrevivência associada ao modelo é definida por:

$$S(y|x) = \left[1 + \exp\left(\frac{y - \boldsymbol{x}^T\boldsymbol{\beta}}{\sigma}\right)\right]^{-1}.$$

4 RESULTADOS

4.1 Análise descritiva dos dados

Neste trabalho serão analisados dados do tempo de sobrevivência de 106 peixes que podem ter apresentado censura ou não, considerando três características inerentes a todos eles: o tamanho do peixe, a profundidade do rio e a transparência da água. O objetivo é definir se alguma dessas características garante um tempo maior de sobrevivência ao peixe em questão. Toda a análise será realizada no *software* R, e em particular através do pacote *Survival*.

Uma vez que existem dados censurados não podemos utilizar as técnicas tradicionais para uma análise descritiva dos dados, ou seja, começaremos a análise utilizando o estimador de Kaplan-Meier.

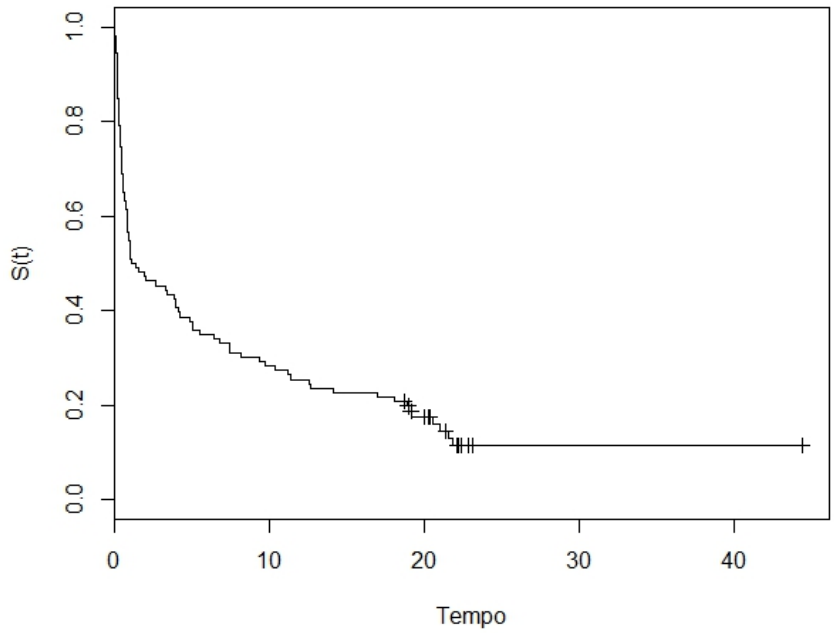

Figura 3 – Curva estimada pelo método não paramétrico de Kaplan-Meier para os tempos de sobrevivência em estudo

Pela Figura 3, observa-se que os peixes apresentam uma alta taxa de falha no primeiro ano do estudo seguido por valores menos repentinos no número de falhas entre os anos 2 e 21. As censuras começam a ocorrer por volta do ano 19, ou seja, a partir dessa faixa

de tempo existem peixes onde há perda de informação sobre sua sobrevivência. A mediana dos dados está no tempo 1,23. Isso quer dizer que em pouco mais de um ano metade dos peixes já não sobreviveram. É interessante observar que por volta do ano 23 a Figura 3 adquire uma reta que poderia ser interpretada como um indicativo de fração de cura para os peixes. Mas além da obviedade disso ser impossível para os dados em questão, por efeitos de mortalidade, nos dados é facilmente verificável que um único peixe foi responsável por esse alongamento no eixo do tempo.

Com isso, é seguro afirmar que a taxa de mortalidade dos peixes da espécie *"Notropis Dourado"* é muito maior no período inicial de suas vidas, pelo menos supondo as características de tamanho, profundidade e transparência observadas no estudo. No entanto, apenas com isso não temos a informação necessária para afirmar qual a distribuição de probabilidade que melhor agregaria essa informação afim de modelar esses dados.

Portanto, a partir dessas informações, o próximo passo natural é seguir em busca de uma distribuição de probabilidade que melhor explique o comportamento desses dados. Para auxiliar nesse objetivo prosseguimos com a curva do tempo total em teste (curva TTT), que irá nos mostrar qual forma que a função da taxa de falha pode assumir.

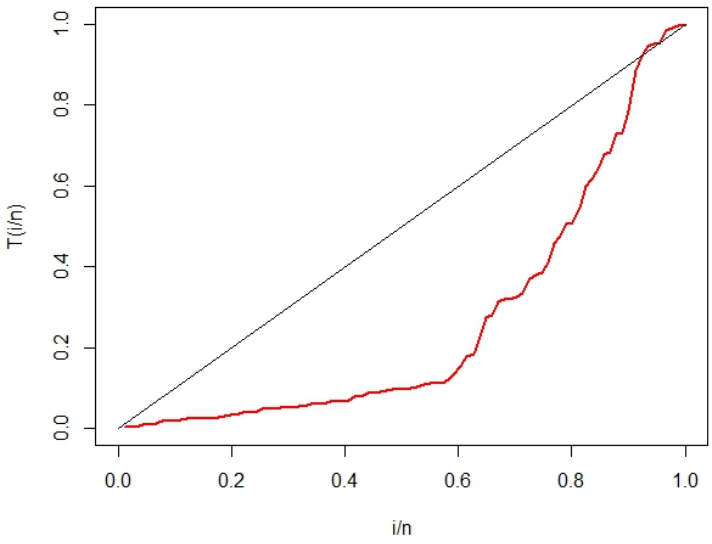

Figura 4 – Curva do tempo total em teste para os dados em estudo

Ao analisar a Figura 4, observamos uma forma convexa que está relacionada

a uma forma de risco decrescente. No entanto, essa forma não permanece assim o tempo inteiro. Há um segmento no final que ultrapassa a linha tracejada que simboliza uma taxa de risco constante.

4.2 Modelagem

Após analisar o comportamento da curva do tempo total em teste (Figura 4), possíveis distribuições candidatas para modelar o tempo de sobrevivência dos peixes são as distribuições Weibull e log-logística, pois elas possuem função de risco de decrescentes. Outra distribuição candidata seria a distribuição Burr XII. Dessa forma, este trabalho irá verificar o ajuste das distribuições Weibull, log-logística e Burr XII aos dados.

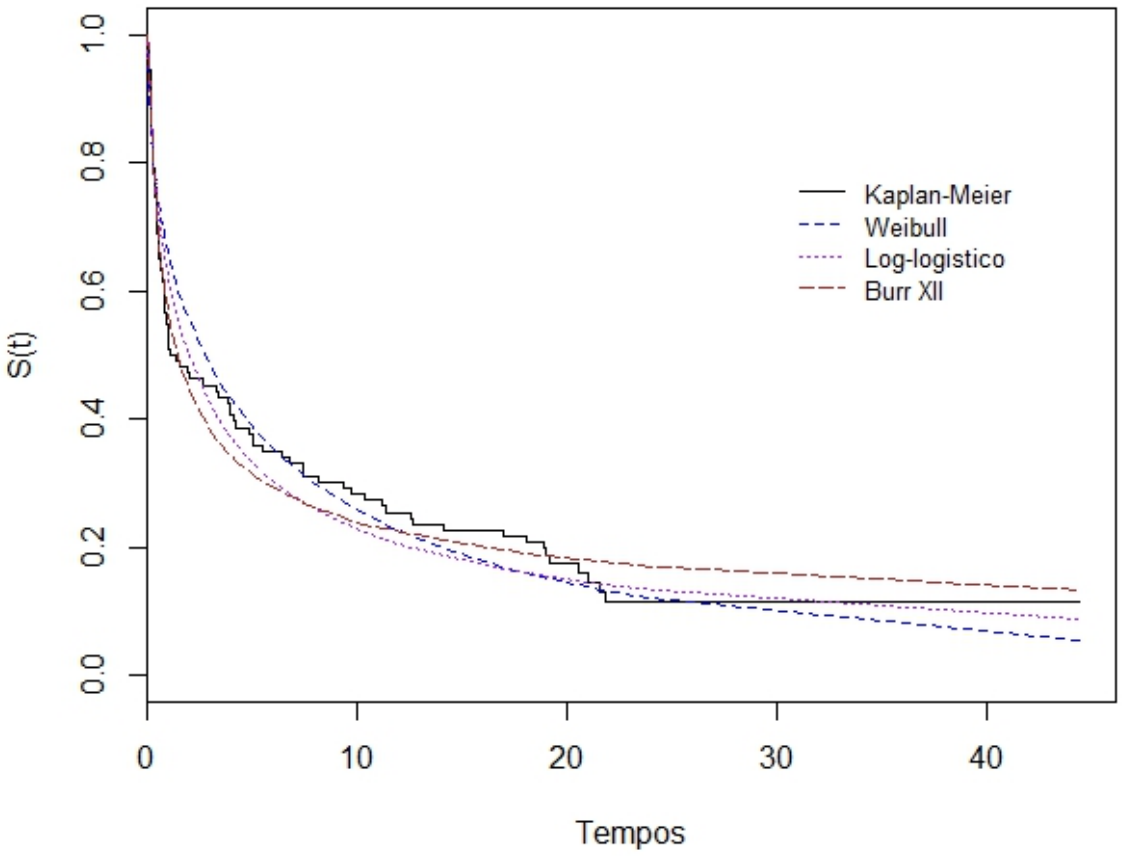

Figura 5 – Estimador de Kaplan Meier e as distribuições de Weibull, Log-logístico e Burr XII ajustadas a ele.

A Figura 5 apresenta a função de sobrevivência estimada pelo método de Kaplan-Meier e as funções de sobrevivência estimadas das distribuições Weibull, log-logística e Burr XII. É possível observar que o ajuste das distribuições Weibull e log-logística se ajustam melhor aos dados do que a Burr XII. Dessa forma, neste trabalho serão considerados as distribuições Weibull e log-logística.

Uma vez que a distribuição de Weibull foi selecionada para avançarmos no objetivo de encontrar um modelo de regressão que melhor explique a sobrevivência dos peixes, na Tabela 1 é apresentado o modelo de regressão log-weibull ou do valor extremo considerando todas as três variáveis contidas no estudo; tamanho, profundidade e transparência:

Tabela 1 – Estimativa dos parâmetros, erro padrão e p-valor para o modelo de regressão do valor extremo para os dados de peixes

Parâmetros	Estimativas	Erro padrão	P-valor
Intercepto	2,957	1,944	0,128
Tamanho	-0,052	0,028	0,064
Profundidade	0,022	0,009	0,014
Transparência	0,510	0,218	0,019

Após o ajuste do modelo aos dados, uma análise de resíduos foi realizada afim de verificar a adequabilidade global do modelo. Para isso foi utilizado o resíduo de Cox-Snell que irá verificar se os dados ajustados pelo modelo vêm de uma população homogênea e seguem uma distribuição exponencial padrão.

Como o modelo se ajusta bem aos dados, pode-se interpretar os resultados apresentados na Tabela 1. Dessa forma, ao considerar o nível de significância de 10%, é possível concluir que peixes maiores possuem menor probabilidade de sobrevivência estimada. Conforme aumenta a profundidade do lago, aumenta a probabilidade de sobrevivência estimada dos peixes. Ou seja, em lugares mais profundos do lago os peixeis vivem mais tempo. E ao aumentar a transparência da água, aumenta a a probabilidade de sobrevivência estimada dos peixes.

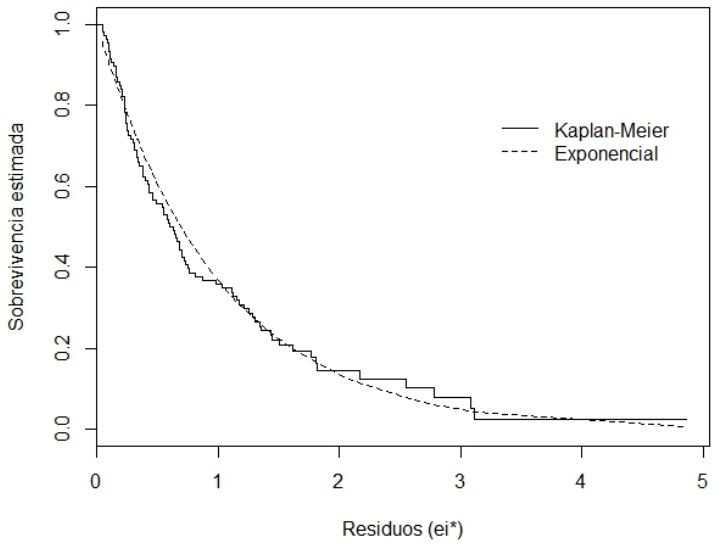

Figura 6 – Análise de resíduos de Cox-Snell para o modelo Weibull

Agora, ao considerar que os tempos seguem uma distribuição log-logística, o modelo de regressão logístico foi estimado e as estimativas dos parâmetros, erro-padrão e p-valor estão na Tabela 2.

Tabela 2 – Estimativa dos parâmetros, erro padrão e p-valor para o modelo de regressão da distribuição log-logístico para os dados de peixes

Parâmetros	Estimativas	Erro padrão	P-valor
Intercepto	0,973	1,981	0,623
Tamanho	-0,042	0,027	0,127
Profundidade	0,025	0,009	0,007
Transparência	0,387	0,217	0,074

Ao nível de significância de 10%, a variável tamanho não é significante. Ao retirar essa variável do modelo e fazer a seleção de variáveis ao nível de significância de 10%, tem-se as estimavas dos parâmetros para o modelo o novo modelo apresentadas na Tabela 3.

Tabela 3 – Estimativa dos parâmetros, erro padrão e p-valor para o modelo de regressão da distribuição log-logístico para os dados de peixes, considerando apenas a profundidade

Parâmetros	Estimativas	Erro padrão	P-valor
Intercepto	-1,957	1,965	0,042
Profundidade	0,027	0,009	0,005

Por meio do resíduo de Cox-Snell, o modelo apresentado na Tabela 3 se ajusta bem ao dados. Ao analisar os resultados da Tabela 3, conclui-se que conforme aumenta a profundidade do lago, aumenta a probabilidade de sobrevivência estimada dos peixes.

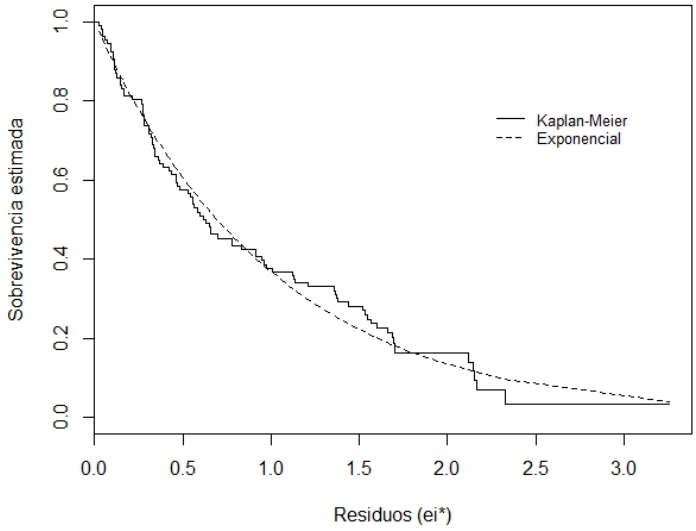

Figura 7 – Análise de resíduos de Cox-Snell para o modelo log logístico

5 CONCLUSÃO

Os resultados obtidos sugerem que ambos os modelos de regressão discutidos, weibull e log-logístico, são adequados aos dados. Porém, no modelo log-logístico a variável tamanho não é relevante a um nível de 10% de significância. Após retirá-la do modelo, a transparência também deixa de ser relevante, restando apenas a profundidade. No modelo weibull isso não ocorre, e as três variáveis: tamanho, profundidade e transparência são relevantes ao nível de significância de 10% no modelo completo. Portanto, por permitir que o estudo seja analisado através de mais variáveis explicativas, o modelo weibull poderia ser considerado como uma boa opção para refletir o comportamento dos dados.

REFERÊNCIAS

CARRASCO, J.M.F. **Modelos de Regressão log-Weibull modificado e a nova distribuição Weibull modificada generalizada.** . 2007. 129p. Dissertação (Mestrado em Estatística e Experimentação Agronômica)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2007.

COLOSIMO, Enrico Antonio; GIOLO, Suely Ruiz. **Análise de sobrevivência aplicada.** São Paulo: E. Blucher, 2006. Xv, 369p.

KALBFLEISH, J. D.; PRENTICE, R. L. **The Statistical Analysis of Failure Time Data.** John Wiley & Sons, New York, 2002

LAWLESS, J. F. **Statistical Methods and Models for Lifetime Data.** John Wiley & Sons, New York, 1982.

NAKANO, E. Y. **Um curso de Análise de Sobrevivência**, Brasília, 2017.

OLIVEIRA, Marcos Lima de. **Análise de dados de transplante de medula óssea; proposta de modelo de regressão Kumaraswamy-Weibull com fração de cura.** 2014. 86 f., il. Trabalho de Conclusão de Curso (Bacharelado em Estatística), Universidade de Brasília, Brasília, 2014.

ROCHA, Tatiana Santos. **Modelos de regressão discretos para dados grupados: uma aplicação em avaliação de risco em produto de crédito parcelado.** 2013. 50 f., il. Monografia (Bacharelado em Estatística), Universidade de Brasília, Brasília, 2013.

RODRÍGUEZ, MARCO, A. **Quantifying habitat-dependent mortality risk in lacutrine fishes by means of tethering trials and survival analyses .** 2010. 12 f., il. Université du Québec à Trois-Riviéres, 2010.

SILVA, G. O. ; Ortega, Edwin M. M. **Modelos de Regressão quando a função de taxa de falha não é monótona. 2011. (Apresentação de Trabalho/Comunicação)**, 2008.