Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelos de regressão em análise de sobrevivência: uma aplicação na modelagem do tempo de vida de *Micrurus corallinus* em cativeiro

Glória Cristina Vieira de Sousa

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2019

Glória Cristina Vieira de Sousa
Licenciada em Matemática

Modelos de regressão em análise de sobrevivência: uma aplicação na modelagem do tempo de vida de *Micrurus corallinus* em cativeiro

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Profa. Dra. **SÔNIA MARIA DE STEFANO PIEDADE**

Dissertação apresentada para obtenção do título de Mestra em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2019

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Sousa, Glória Cristina Vieira de

Modelos de regressão em análise de sobrevivência: uma aplicação na modelagem do tempo de vida de *Micrurus corallinus* em cativeiro / Glória Cristina Vieira de Sousa. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2019 .

59 p.

Dissertação (Mestrado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Análise de sobrevivência 2. Dados censurados 3. Modelos de regressão 4. Distribuição gama generalizada 5. Distribuição log-logística 6. Distribuição odd log-logística generalizada Weibull . I. Título.

DEDICATÓRIA

À Deus.
Aos meus pais, Alcides e Rosmary.
As minhas avós, Glória (in memorian) e Cristina.

AGRADECIMENTOS

Primeiramente, agradeço a Deus e a Nossa Senhora Aparecida por estarem comigo ao longo dessa jornada.

Agradeço aos meus pais, Rosmary e Alcides, que são sem dúvidas os melhores pais do mundo. Obrigada por cada palavra de apoio, cada conversa, cada abraço e cada gesto durante esse tempo de mestrado, de graduação e de vida. É somente por vocês que tenho forças todos os dias para encarar tantos desafios e estar aqui. Nunca conseguirei mostrar o quão grata eu sou por ser filha de pessoas tão incríveis.

A minha orientadora Prof.ª Dr.ª Sônia Maria De Stefano Piedade que me acolheu com afeto e compreensão, sendo inspiração para meus dias. Obrigada pelos encontros de orientação cheias de ensinamentos, fossem eles acadêmicos ou diversos.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

Ao conselho e a todo o corpo docente do Programa de Pós-Graduação em Estatística e Experimentação Agronômica, em especial ao Prof. Dr. Idemauro Rodrigues de Lima, ao Prof. Dr. Carlos Tadeu dos Santos Dias e ao Prof. Dr. César Golçalves de Lima.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, à secretária Luciane Brajão, ao técnico em informática Jorge Alexandre Wiendl e ao técnico de laboratório Eduardo Bonilha. Em especial à secretária Solange de Assis Paes Sabadin que além de ajudar a todos com questões burocráticas, esteve sempre conosco.

À professora Prof.ª Dr.ª Selene Maria Coelho Loibel pela contribuição feita ao trabalho e ao meu aprendizado desde a graduação.

Aos meus amigos do Programa de Pós-Graduação em Estatística e Experimentação Agronômica, Pollyane, Maria Leticia, Cristiane, Caroline, Eduardo, Welinton, Vivian, Pórtya, Hercílio, Val, César, Rick e todos os outros que viveram comigo muito além da salinha e cafés, dividiram momentos importantes e me mostraram que sempre há motivos para continuar, levantar a cabeça e seguir em frente.

Ao meu namorado, Pedro, por todo o apoio com traduções, correções, por ser um exemplo de pós graduando, um exemplo de pessoa, por cada gesto de amor e de carinho.

Agradeço aos meus amigos de vida, Daniela Luciana, Giane, Monik, Carlos Bonatti, Alex Melges, Célia, Sérgio, Tayna, Marina, Bruna Goudinho e Carol. Obrigada a cada um de vocês por cruzarem meu caminho.

Obrigada a todos que direta ou indiretamente tornaram essa dissertação algo palpável.

EPÍGRAFE

Nadei até onde podia
Até onde minhas forças me deixaram ir.
Não me entreguei.
Lutar pra sempre, resistir.
Sobrevivi.
Rompi barreiras.
Eu venci.
Luciano Garcia - CPM22

SUMÁRIO

.

RESUMO

Modelos de regressão em análise de sobrevivência: uma aplicação na modelagem do tempo de vida de *Micrurus corallinus* em cativeiro

Os dados de sobrevivência possuem peculiaridades que necessitam de uma atenção especial no momento em que se deseja realizar uma análise nos mesmos. Em tais dados é comum a presença de censuras e sua variável resposta é definida como o tempo de vida até a ocorrência de um evento de interesse. Existem distribuições que acolhem dados de sobrevivência, como as distribuições exponencial, Weibull, gama, gama generalizada, entre outras, assim como seus respectivos modelos de regressão adaptados para esse tipo de estudo. Os modelos de regressão exponencial e Weibull são os mais citados na literatura por terem fácil aplicação e se modelarem bem aos dados. O modelo de regressão gama generalizado geralmente se adapta melhor aos dados por ter três parâmetros, assim como o modelo de regressão log-logístico, que é visto como uma alternativa à distribuição Weibull e é muito utilizado por ter formas explícitas para a sua função de sobrevivência e de falha. No entanto, esses modelos ainda possuem restrições e, por conta disso, novas famílias de modelos de regressão estão sendo desenvolvidas na literatura, assim como a família de distribuições odd log-logística generalizada, que pretende oferecer melhores ajustes pois aparenta ter capacidade de modelar diferentes tipos de dados. O objetivo dessa dissertação foi aplicar técnicas de análise de sobrevivência na modelagem dos tempos de vida de *Micrurus corallinus*, ajustando os modelos já presentes na literatura e o modelo proposto odd log-logística generalizada Weibull (OLLG-W). Conclui-se que o modelo de regressão que se mostrou adequado aos dados foi o log-logístico e o modelo de regressão OLLG-W não apresentou nenhuma vantagem em relação aos que já são frequentes na literatura.

Palavras-chave: Análise de sobrevivência; Dados censurados; Modelos de regressão; Distribuição gama generalizada; Distribuição log-logística; Distribuição odd log-logística generalizada Weibull

ABSTRACT

Regression models in survival analysis: a captivity *Micrurus corallinus* lifetime application modeling

Survival data hold special attention-needed peculiarities the moment you intend to realize an analysis on. These data own censorships and their variable responses are defined as lifetime to interest- event occurrence. There are distributions that harbor these data, such as exponential distribution, Weibull, gamma, generalized gamma, among others, just as their respective event-adapted regression models. Exponential regression and Weibull models are the most literature recurrent, in view of their easy application and appropriate data modeling. The generalized gamma regression model usually is a better fit to the data, due to its three-parameter comprise, just as the log-logistic regression model, which is seen as an alternative to Weibull distribution and is heavily utilized for it's explicit shapes to survivability and fail functions. Nonetheless, these models still retain restrictions and, on account of that, new regression model families are being developed, as in the log logistic generalized distribution family, which intends to offer better settings due to its different real data modeling ability. The purpose of this dissertation was to apply survival analysis techniques in *Micrurus corallinus* lifetime modeling, adjusting already existing models and the proposed Weibull generalized odd log logistic model (OLLG-W). We came to the conclusion that the adequate regression model to *Micrurus corallinus* data was the log-logistic model. The OLLG-W model didn't offer any benefits when compared to literature-recurrent ones.

Keywords: Survival analysis; Censored data; Regression models; Generalized Gama distribution; Log-logistic distribution; Generalized odd log-logistic Weibull distribution

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A análise de sobrevivência é uma das áreas da estatística com maior implementação nos últimos anos. Uma das razões para esse crescimento é a sua vasta aplicabilidade na área médica e industrial, além do desenvolvimento e aprimoramento de tais técnicas. Também pode ser aplicada em áreas como engenharia, sociologia, psicologia, educação, dentre outras. A variável resposta neste tipo de estudo é o tempo até a ocorrência de um evento de interesse (morte do paciente, tempo até a cura ou a recidiva de uma doença, tempo de vida de um determinado componente elétrico, entre outros eventos), denominado tempo de falha. Portanto, a análise de sobrevivência pode ser aplicada em toda área que tenha como variável resposta o tempo de vida, seja ele de um paciente, componente elétrico, plantas, insetos, animais, etc.

A resposta em análise de sobrevivência é composta não só pelo tempo de falha, mas também pela censura. A censura ocorre quando se tem observações parciais da resposta. Por exemplo se em algum momento do estudo o acompanhamento do paciente foi interrompido, se o estudo terminou para análise ou se o paciente morreu de causa diferente da estudada.

Se não houver censuras, os dados podem ser analisados por meio dos métodos usuais (modelos de regressão paramétricos, não paramétricos, análise de variância, etc). Se houver censuras, porém, tais métodos não podem ser utilizados pois necessitam de todos os tempos de falha. Assim, faz-se necessário o uso de técnicas adequadas para tal, ou seja, o uso de métodos de análise de sobrevivência. Os métodos usuais de análise de sobrevivência foram desenvolvidos de forma a incorporar as observações censuradas. Diversas distribuições e modelos foram pensados com tal objetivo. Alguns exemplos são: distribuição exponencial, gama, gama generalizada e Weibull, sendo a última a mais utilizada. Há também a distribuição log-logística, que se mostra como uma boa alternativa à distribuição Weibull por ter formas explícitas de sua função de sobrevivência e de falha. No entanto, alguns desses modelos possuem limitações e por esse motivo é possível encontrar na literatura propostas de criação de novas famílias de distribuições.

Recentemente, Cordeiro et al. (2016) propuseram uma nova classe de distribuições contínuas com dois parâmetros de forma adicional denominada odd log-logística generalizada-G (GOLL-G) dada pela integração da função de distribuição acumulada da família proposta por Geaton e Lynch (2006). Sua função densidade de probabilidade pode ser expressa como uma combinação linear das densidades exponenciadas com base na distribuição base. Essa família oferece melhores ajustes do que as outras classes de distribuições já conhecidas, uma vez que tem capacidade de modelar diferentes tipos de dados reais.

Na área biológica a criação de serpentes em cativeiro é bastante complicada. A instalação de um biotério tem custo elevado pois demanda cuidados específicos com cli-

matização, alimentação adequada e profissionais com capacitação para tal função. Mesmo com todos esses cuidados as serpentes podem sofrer estresses que se refletem em perda de apetite, doenças e até a morte. O mesmo acontece com as serpentes da espécie *Micrurus corallinus* (SERAPICOS E MERUSSE, 2002) conhecidas como "cobra coral". Geralmente essas serpentes criadas em cativeiro têm seu veneno utilizado para a produção de soro antiofídico e, por conta disso é desejável que o tempo de vida dessas serpentes seja longo para o melhor aproveitamento e maior produção de soro (MENDES *et al.*, 2015).

O Laboratório de Herpetologia do Instituto Butantan, entre outras espécies, cria serpentes da éspecie *Micrurus corallinus* com este fim. As serpentes dessa éspecie geralmente têm hábitos diurnos (MARQUES, 1992) e sua dieta é composta por presas de corpo alongado como cobras, lagartos e anfisbênios (MARQUES E SAZIMA, 1997) mesmo em cativeiro.

Esse trabalho tem como objetivo a aplicação de técnicas utilizadas em análise de sobrevivência na modelagem dos tempos de vida de 289 serpentes quanto a um novo manejo que foi adotado pelos pesquisadores do Laboratório de Herpetologia do Instituto Butantan (MENDES *et al.*, 2015). O interesse é mostrar que o tempo de vida das serpentes aumentou, proporcionando assim um aumento da produção de soro antiofídico. Para isso, serão ajustados modelos de regressão presentes da literatura (exponencial, Weibull, gama e gama generalizada) e o modelo de regressão proposto odd log-logístico generalizado Weibull (CORDEIRO *et al.*, 2017).

A dissertação está estruturada da seguinte maneira: introdução, revisão de literatura, materiais, métodos, resultados e discussões, conclusão e referências bibliográficas.

2 REVISÃO DE LITERATURA

Nesse capítulo são apresentados conceitos básicos para estudos de sobrevivência, abordando características específicas de tais dados, funções, modelos probabilísticos e métodos inferenciais iniciais.

2.1 Conceitos importantes em análise de sobrevivência

2.1.1 Planejamento de estudos

Os estudos clínicos são investigações científicas realizadas com o objetivo de verificar uma determinada hipótese de interesse. Essas investigações são conduzidas coletando dados e analisando-os por meio de métodos estatísticos (COLOSIMO E GIOLO, 2006). Esses estudos são geralmente compostos por quatro etapas

1) formulação da hipótese de interesse;

2) planejamento e coleta dos dados;

3) análise estatística dos dados;

4) teste da hipótese formulada.

Em análise de sobrevivência, a variável resposta considerada é o tempo até a ocorrência de um evento de interesse pré-determinado e tem natureza longitudinal. O delineamento pode ser observacional (descritivo, caso-controle e coorte) ou experimental, em que há intervenção do pesquisador na aleatorização dos tratamentos (ensaio clínico aleatorizado).

2.1.2 Tempo de falha e censuras

Um estudo de sobrevivência é definido pelo tempo de falha e a presença de censura, que compõem a variável resposta. O tempo de falha é composto por três características: o tempo inicial, a escala de medida e o evento de interesse (falha).

O tempo inicial é escolhido como sendo a data inicial da aleatorização, a data do diagnóstico ou a data do início do tratamento de doenças e, geralmente, a sua escala de medida é dada em tempo real (horas, dias, semanas, etc.). O evento de interesse ou falha deve ser bem definido de forma clara e precisa para evitar problemas futuros. Portanto, o tempo de falha vai desde o tempo inicial até a ocorrência do evento de interesse.

É comum que estudos de sobrevivência sejam de longa duração e algumas vezes terminem antes que todos os indivíduos falhem, ou seja, é comum que contenham observações incompletas. Essas observações são chamadas de censuras e podem ocorrer por diversos motivos, como a perda de seguimento do indivíduo, a morte por uma razão diferente da estudada ou a não ocorrência do evento de interesse. As observações censuradas não devem ser retiradas do estudo, pois contêm informações importantes sobre aquele conjunto de dados e sobre o tempo de vida, podendo acarretar em conclusões erradas.

As censuras podem ser de vários tipos, como

1. À direita: o evento de interesse está à direita do tempo registrado, isto é, o evento de interesse ainda não ocorreu.

   Tipo I: o estudo de sobrevivência será encerrado após um período de tempo pré estabelecido;

   Tipo I Progressiva: o estudo de sobrevivência será encerrado após um período de tempo pré estabelecido porém os indivíduos não entraram no estudo na mesma data;

   Tipo II: o estudo de sobrevivência será encerrado após o evento de interesse ter ocorrido em um número pré estabelecido de indivíduos;

   Aleatória: o indivíduo sai do estudo de sobrevivência por um motivo diferente do estabelecido. A variável T_i representa o tempo de falha do i-ésimo indivíduo e C_i o tempo de censura associado, ou seja, o tempo da i-ésima observação é dado por $t_i = min(T_i, C_i)$;

2. À esquerda: o tempo registrado é maior do que o tempo de falha, ou seja, o evento de interesse já ocorreu quando o indivíduo entra no estudo.

3. Intervalar: é a mais geral e ocorre quando o evento de interesse ocorre dentro de um intervalo de tempo, isto é, não se conhece o tempo exato de falha mas sabe-se que ocorreu em um intervalo de tempo.

Segundo Colosimo e Giolo (2006), cada observação é representada pelo par (t_i, δ_i) sendo que t_i é o tempo de falha (ou censura) e δ_i é a variável indicadora de falha (ou censura) em que

$$\delta_i = \begin{cases} 1 & \text{se } t \text{ é um tempo de falha,} \\ 0 & \text{se } t \text{ é um tempo de censura.} \end{cases} \tag{2.1}$$

Dessa maneira, a variável resposta em análise de sobrevivência é representada por duas colunas no banco de dados. E, na presença de covariáveis, que podem influenciar o tempo de sobrevivência ou a censura, os dados ficam representados por $(t_i, \delta_i, \mathbf{x}_i)$ em que $\mathbf{x}_i$ é o vetor de covariáveis, sempre verificando se tais covariáveis estão relacionadas com o tempo de sobrevivência, censura ou entre si (FACHINI, 2006).

2.1.3 Algumas funções importantes

Sendo T uma variável aleatória não-negativa e contínua, que representa o tempo de sobrevivência de um indivíduo, seguem algumas funções importantes na caracterização de dados de sobrevivência.

A função de sobrevivência é a probabilidade de um indivíduo não falhar até um tempo t, ou seja, a probabilidade de um indivíduo sobreviver até o tempo t

$$S(T) = P(T \geq t), \tag{2.2}$$

em que $S(T)$ é monótona, decrescente, $S(0) = 1$ e $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

Em decorrência da função de sobrevivência, a função de distribuição acumulada é definida como a probabilidade de um indivíduo não sobreviver ao tempo t

$$F(T) = 1 - S(T) \Rightarrow S(T) = 1 - F(T). \tag{2.3}$$

A função de distribuição de probabilidade é o limite da probabilidade de um indivíduo morrer em um intervalo de tempo $[t + \Delta_t)$

$$f(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \tag{2.4}$$

com isso, pode-se reescrever a função de sobrevivência da seguinte forma

$$S(T) = P(T > t) = 1 - P(T \leq t) = 1 - \int_0^t f(x)dx = 1 - F(t). \tag{2.5}$$

A função de risco de T descreve a forma em que a taxa instantânea de falha muda com o tempo

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{2.6}$$

Sabendo-se que diferentes funções de sobrevivência podem ter formas semelhantes, a função de risco torna-se mais informativa uma vez que diferentes funções de risco podem ter formas totalmente diferentes. Colosimo e Giolo (2006) argumentam que a modelagem da função de risco é um importante método para análise de sobrevivência pois esta função pode ter forma crescente, decrescente ou constante.

A função de risco acumulada, como o próprio nome já sugere, fornece a taxa de risco do indivíduo

$$H(t) = \int_0^t h(u)du. \tag{2.7}$$

Vistas tais funções, pode-se estabelecer algumas relações entre elas

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{logS(t)\}$$

e

$$S(t) = exp\{-H(t)\}.$$

2.1.4 Estimador não paramétrico de Kaplan-Meier

A presença das observações censuradas causa um grande problema quando tenta-se usar técnicas convencionais de análise descritiva. Como alternativa, usa-se o gráfico de dispersão de cada covariável contínua versus a resposta que proporciona uma avaliação, por meio da nuvem de pontos, de uma possível relação linear entre elas ou a adequação de um modelo proposto.

Segundo Colosimo e Giolo (2006), o estimador não paramétrico de Kaplan-Meier, proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência, é o mais utilizado em diversos tipos de estudo e vem ganhando espaço dentre os demais estimadores não paramétricos.

Considerando $t_1 < t_2 < ... < t_n$ diferentes tempos em que ocorreram as censuras em uma amostra de n indivíduos tem-se que o estimador é dado por

$$\widehat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right) \tag{2.8}$$

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

2.2 Modelos probabilísticos

Em estatística, é comum a construção de modelos probabilísticos para melhor entendimento dos dados estudados, não sendo diferente em análise de sobrevivência. As distribuições exponencial, Weibull, gama, gama generalizada e log-logística são as mais comuns entre elas (tratando-se de análise de sobrevivência) e serão discutidas a seguir.

2.2.1 Distribuição exponencial

A distribuição exponencial é considerada a mais simples pois possui apenas um parâmetro e sua função taxa de falha é constante, ou seja, tanto um indivíduo recém adicionado ao experimento quanto um indivíduo que está há mais tempo, têm a mesma probabilidade de falhar (falta de memória da distribuição exponencial). Por muitas vezes é utilizada quando o tempo destinado ao experimento é curto. Sua função densidade de probabilidade é dada por

$$f(t;a) = \frac{1}{a} exp\left[- \left(\frac{t}{a} \right) \right]; t \geq 0 \tag{2.9}$$

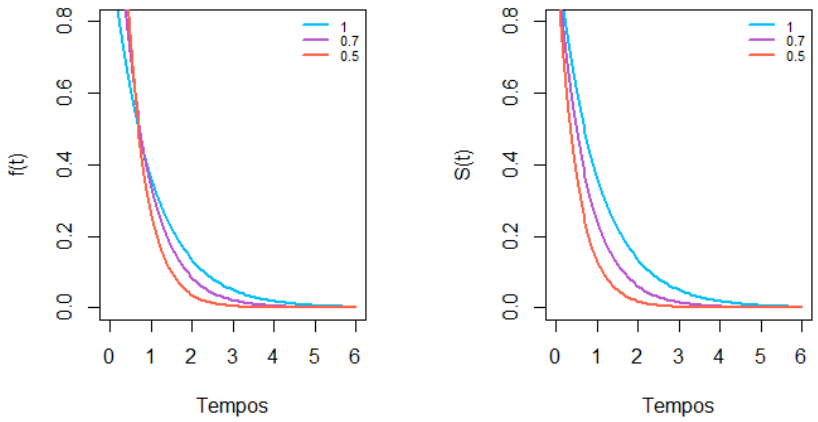
em que a é o tempo médio de vida.

As funções de sobrevivência e de risco são dadas, respectivamente, por

$$S(t;a) = exp\left[-\left(\frac{t}{a}\right)\right] \tag{2.10}$$

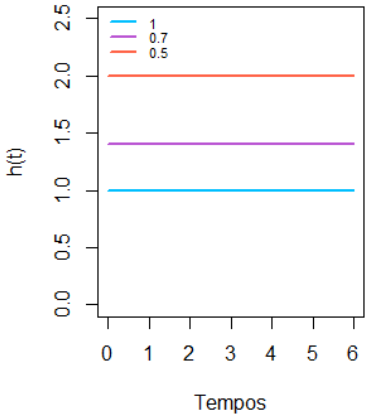
e

$$h(t;a) = \frac{1}{a}. \tag{2.11}$$



(a)

(b)



(c)

Figura 2.1: Forma típica das funções densidade de probabilidade (a), de sobrevivência (b) e de risco (c) da distribuição exponencial para diferentes valores do parâmetro a

A esperança e a variância são, respectivamente:

$$E(T) = a$$

e

$$Var(T) = a^2.$$

Adotando $Y = ln(T)$ temos que a função de distribuição de probabilidade é dada por (VALENÇA, 1994)

$$f_Y(y) = exp(y - \lambda - e^{y-\lambda}),$$

em que $\lambda = ln(a)$. Pode-se representar Y como sendo $a+W$ com W uma variável aleatória seguindo a distribuição Valor Extremo padrão caracterizada pela função distribuição de probabilidade e de sobrevivência dadas por

$$f_W(w) = exp(w - e^w) \qquad \text{e} \qquad S_W(w) = exp(-e^w). \tag{2.12}$$

A distribuição Valor Extremo padrão é muito utilizada em análise de sobrevivência pois caracteriza de forma adequada a distribuição do logaritmo de certos tempos de vida (COLOSIMO E GIOLO, 2006).

2.2.2 Distribuição Weibull

A distribuição Weibull foi proposta em 1939 por Weibull e a sua popularidade em aplicações práticas se deve ao fato de apresentar uma grande variedade de formas, todas com uma propriedade básica, ou seja, a sua função taxa de falha é monótona, crescente, decrescente ou constante. A função densidade de probabilidade é dada por

$$f(t; a, b) = \frac{b}{a^b} t^{b-1} exp\left[-\left(\frac{t}{a}\right)^b\right]; t \geq 0, \tag{2.13}$$

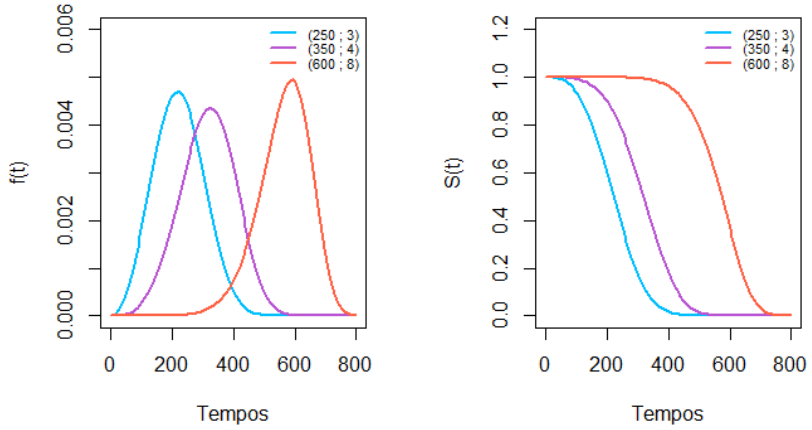
em que $a > 0$ é o parâmetro de escala e $b > 0$ é o parâmetro de forma.

Tem-se que a distribuição exponencial é um caso particular da distribuição Weibull, basta tomar $b = 1$. Sua função de sobrevivência e de risco são dadas por

$$S(t; a, b) = exp\left[-\left(\frac{t}{a}\right)^b\right] \tag{2.14}$$

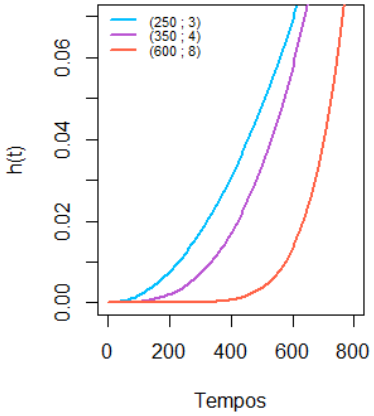
e

$$h(t; a, b) = \frac{b}{a^b} t^{b-1}. \tag{2.15}$$



(a)

(b)



(c)

Figura 2.2: Forma típica das funções densidade de probabilidade (a), de sobrevivência (b) e de risco (c) da distribuição Weibull para diferentes valores dos parâmetros (a, b)

A esperança e variância são

$$E(T) = a\Gamma(1 + (1/b))$$

e

$$Var(T) = a^2[\Gamma(1 + (2/b)) - \Gamma(1 + (1/b)^2)],$$

em que $\Gamma(k)$ é a função gama definida como $\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}dx$.

2.2.3 Distribuição gama

A distribuição gama não é tão utilizada quanto a distribuição Weibull, apesar de se ajustar adequadamente a vários dados de tempo de vida (VALENÇA, 1994) e inclui a disribuição exponencial como seu caso particular ($k = 1$).

A função densidade de probabiblidade é

$$f(t; a, k) = \frac{1}{\Gamma(k)a^k} t^{k-1} exp\left[-\left(\frac{t}{a}\right)\right]; t \geq 0, \tag{2.16}$$

em que $k > 0$ é o parâmetro de forma e $a > 0$ é o parâmetro de escala.

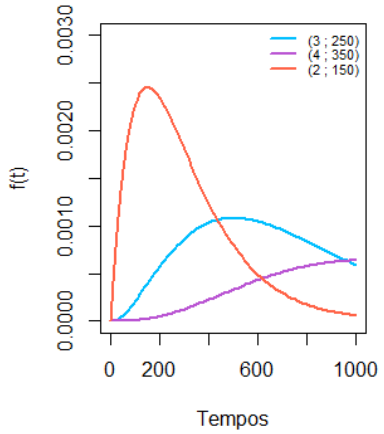
As suas função de sobrevivência e de risco são dadas por (LAWLESS, 1982)

$$S(t; a, k) = \int_t^\infty \frac{1}{\Gamma(k)a^k} u^{k-1} exp\left[-\left(\frac{u}{a}\right)\right] du = 1 - I(k, t/a) \tag{2.17}$$

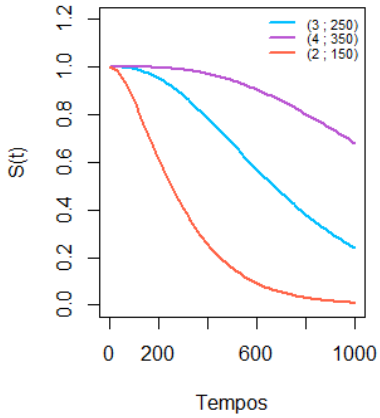
em que $I(k, t/a) = \dfrac{1}{\Gamma(k)} \displaystyle\int_{t/a}^\infty a^{-k} u^{k-1} du$, sendo $\int_{t/a}^\infty a^{-k} u^{k-1} du$ é a função gama incompleta, dada originalmente por $\gamma(k, x) = \int_0^x w^{k-1} e^{-w} dw$ e

$$h(t; a, k) = \frac{f(t)}{S(t)}, \tag{2.18}$$

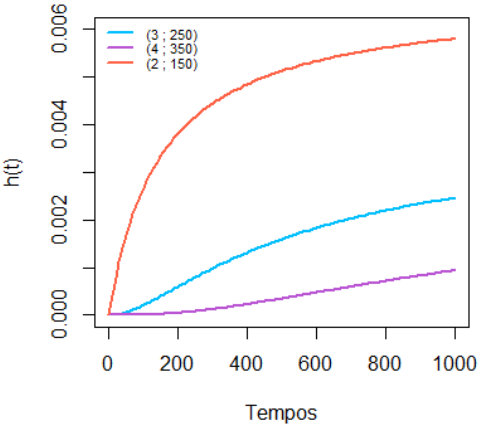
que é monótona crescente para $k > 1$, com $h(0) = 0$ e $\lim_{t\to\infty} h(t) = \dfrac{1}{a}$, é constante em $k = 1$ com $h(t) = \dfrac{1}{a}$ e é monótona descrescente em $0 < k < 1$ com $\lim_{t\to 0} h(t) = \infty$ e $\lim_{t\to\infty} h(t) = \dfrac{1}{a}$.



(a)



(b)



(c)

Figura 2.3: Forma típica das funções densidade de probabilidade (a), de sobrevivência (b) e de risco (c) da distribuição exponencial para diferentes valores do parâmetro a

Sua esperança e variância são, respectivamente

$$E(T) = ka$$

e

$$Var(T) = ka^2.$$

2.2.4 Distribuição gama generalizada

A distribuição gama generalizada foi proposta por Stacy (1962) e é caracterizada por três parâmetros, a, k e b, todos maiores que zero. É considerada muito flexível,

uma vez que possui diversas distribuições como casos particulares e é muito utilizada para discriminar tais distribuições.

A função densidade de probabilidade da distribuição gama generalizada é dada por

$$f(t) = \frac{b}{\Gamma(k)a^{bk}} t^{bk-1} exp\left[-\left(\frac{t}{a}\right)^{b}\right]; t \geq 0, \tag{2.19}$$

com $t > 0$, $\Gamma(k)$ é a função Gama, b o parâmetro de escala e a e k os parâmetros de forma.

De tal maneira que as distribuições exponencial, Weibull e gama são obtidas tomando os parâmetros como na tabela 2.1.

Tabela 2.1: Casos particulares da distribuição gama generalizada

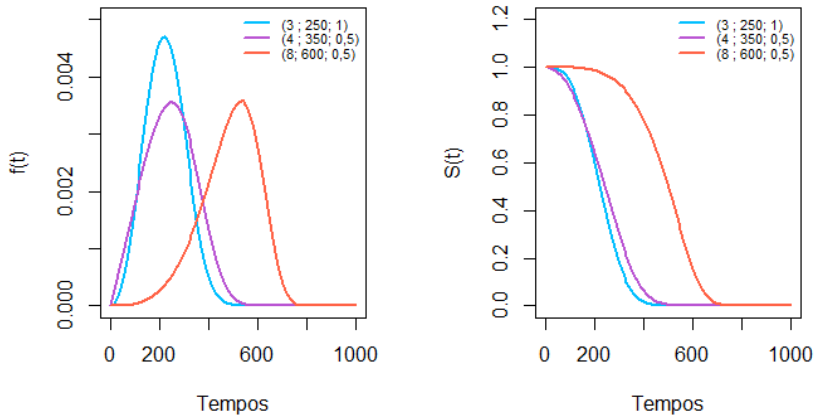
Distribuição	a	b	k
Gama	1	b	k
Weibull	a	b	1
Exponencial	a	1	1

A função de sobrevivência e função de risco são

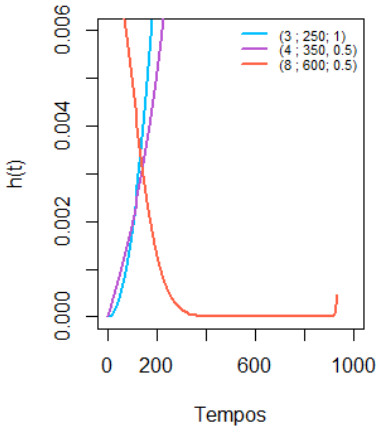
$$S(t) = 1 - \frac{\gamma(k, (t/a)^b)}{\Gamma(k)} \tag{2.20}$$

e

$$h(t) = \frac{t^{bk-1} exp\{-(t/a)^b\}}{\int_0^\infty w^{bk-1} exp\{-(w/a)^b\} dw}. \tag{2.21}$$



(a) (b)



(c)

Figura 2.4: Forma típica das funções densidade de probabilidade (a), de sobrevivência (b) e de risco (c) da distribuição gama generalizada para diferentes valores dos parâmetros (a, b, k)

A esperança e variância são respectivamente dadas por

$$E(T) = \frac{a\Gamma\left(\frac{bk+1}{b}\right)}{\Gamma(k)}$$

e

$$Var(T) = \frac{a^2}{\Gamma(k)}\left[\Gamma\left(\frac{bk+2}{b}\right) - \frac{\Gamma\left(\frac{bk+1}{b}\right)^2}{\Gamma(k)}\right].$$

2.2.5 Distribuição log-logística

Vários autores na literatura consideram a distribuição log-logística como uma alternativa para a distribuição Weibull como Colosimo e Giolo (2006), Santos (2017) e Braguim (2015). Alguns fatores que os levam a essa consideração são o fato de que a distribuição log-logística apresenta formas simples para suas funções de sobrevivência e de risco (SANTOS, 2017; BRAGUIM, 2015), e que as mesmas possuem formas gráficas explícitas, sendo a função de risco considerada unimodal (SANTOS, 2017; GARCIA, 2013).

Sua função densidade de probabilidade é dada por

$$f(t;a,b) = \frac{b}{a^b} t^{b-1}(1+(t/a)^b)^{-2}; t \geq 0 \tag{2.22}$$

sendo $a > 0$ parâmetro de forma e $b > 0$ parâmetro de escala.

Suas funções de sobrevivência e de risco são:

$$S(t;a,b) = \frac{1}{1+(t/a)^b} \tag{2.23}$$

e

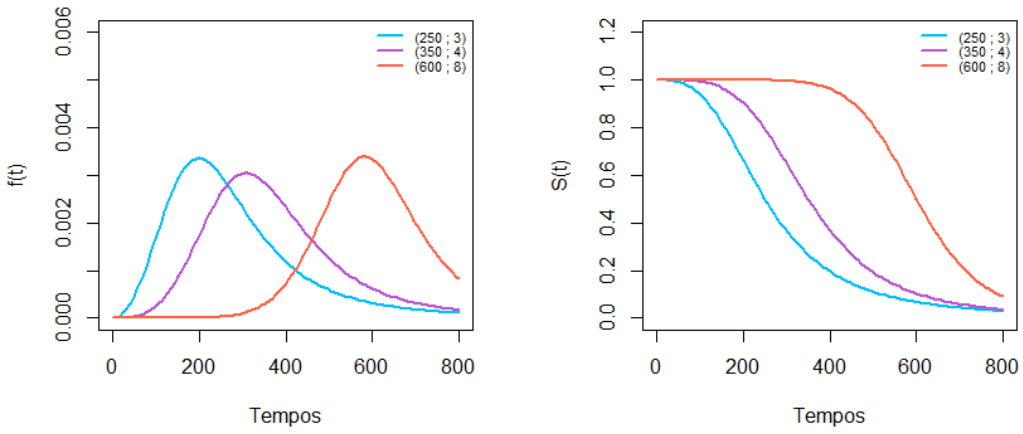
$$h(t;a,b) = \frac{b(t/a)^{b-1}}{a[1+(t/a)^b]}. \tag{2.24}$$

As expressões de esperança e variância são

$$E(T) = [\pi a.cossec(\pi/b)]/b$$

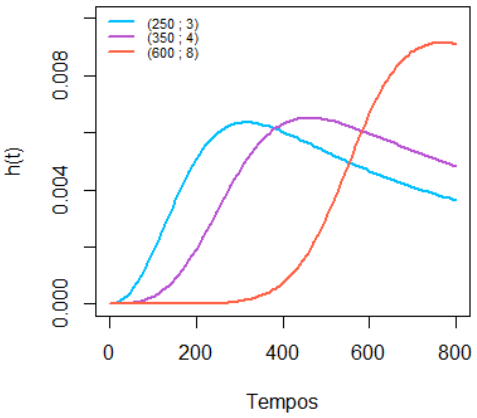
e

$$Var(T) = [(2\pi a^2.cossec(2\pi/b))/b] - E(T^2)$$



(a)

(b)



(c)

Figura 2.5: Forma típica das funções densidade de probabilidade (a), de sobrevivência (b) e de risco (c) da distribuição log logística para diferentes valores dos parâmetros (a, b)

2.3 Estimação dos parâmetros

É possível encontrar na literatura diversos métodos para a estimação dos seus parâmetros. Em análise de sobrevivência utiliza-se o método de máxima verossimilhança que consiste em encontrar o valor do vetor de parâmetros $\boldsymbol{\theta}$ que maximize a função de verossimilhança. Esse estimador é uma ótima opção pois incorpora as censuras, é relativamente simples e possui ótimas propriedades para grandes amostras (COLOSIMO E GIOLO, 2006).

De início, considera-se uma amostra de observações $t_1, t_2, ..., t_n$ da população, em que todas as observações são não censuradas. A função de verossimilhança para o vetor de parâmetros $\boldsymbol{\theta}$ é expressa por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i, \boldsymbol{\theta}), \tag{2.25}$$

em que $f(t_i)$ é a função densidade de probabilidade vista em (2.4).

Agora, é preciso dividir as observações entre as censuradas e as não censuradas, uma vez que os dados considerados são dados de tempos de vida. Ou seja, sejam r observações não censuradas, $(1, 2, ..., r)$ e $n-r$ observações censuradas, $(r+1, r+2, ..., n)$. Assim, a expressão para a função de verossimilhança que abrange todos os tipos de censura é expressa por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{r} f(t_i, \boldsymbol{\theta}) \prod_{i=r+1}^{n} S(t_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} \left[f(t_i, \boldsymbol{\theta})\right]^{\delta_i} \left[S(t_i, \boldsymbol{\theta})\right]^{1-\delta_i}, \tag{2.26}$$

em que $S(t_i)$ é a função de sobrevivência (2.5) e δ_i é a variável indicadora de falha (2.1).

É comum trabalhar com o logaritmo da função de verossimilhança que é dado por

$$\ell(\boldsymbol{\theta}) = \sum_{i \in F} ln[f(y_i)] \sum_{i \in C} ln[S(y_i)], \tag{2.27}$$

em que F denota o conjunto de observações não censuradas e C o conjunto de observações censuradas.

Para maximizar $L(\boldsymbol{\theta})$ ou $\ell(\boldsymbol{\theta})$ e encontrar os valores de $\boldsymbol{\theta}$, basta resolver o sistema de equações

$$U(\boldsymbol{\theta}) = \frac{\partial ln(L(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = 0.$$

Como as equações são, em sua maioria, não lineares em $\boldsymbol{\theta}$ e não apresentam solução análiticа, o sistema deve ser resolvido por um método númerico, como Newton-Raphson.

2.4 Escolha do modelo

É possível encontrar na literatura várias maneiras de se comparar modelos. Alguns desses critérios são bem comuns e levam em conta a complexidade do modelo no critério de seleção e penalizam a verossimilhança utilizando o número de parâmetros a serem estimados e o tamanho da amostra (PRATAVIEIRA, 2017). Neste trabalho foi utilizado, em princípio, o Gráfico TTT-plot (BARLOW E CAMPO, 1975) e posteriormente, o método gráfico (GARCIA, 2013) que consiste em comparar a função de sobrevivência estimada por Kaplan-Meier com a função de sobrevivência estimada da distribuição desejada, o critério de Akaike, o critério de informação bayesiano e o teste da razão de verossimilhanças para a comparação e escolha de modelos.

2.4.1 Gráfico TTT-plot (tempo total em tese)

Existem várias formas que o gráfico da função taxa de falha na variável T pode assumir, portanto é de extrema importância usar o modelo apropriado para tal. Uma técnica capaz de verificar o ajuste do modelo é a construção do TTT-plot, proposto por Barlow e Campo (1975), que tem como objetivo determinar o tipo de função de risco que os dados possuem.

A curva é obtida construindo um gráfico de $G(r/n)$ em relação à r/n, sendo que $G(r/n)$ é definida por:

$$G(r/n) = \frac{\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_{i:n}}, \tag{2.28}$$

em que n é o tamanho da amostra, $r = 1, ..., n$, $i = 1, ...n$ observações censuradas e $T_{i:n}$ são estatísticas de ordem da amostra.

Na figura (2.6) encontram-se exemplos das possíveis curvas que o gráfico TTT-plot pode assumir. A reta (A) indica uma função de risco constante, as curvas convexas como (B) são para funções de risco monotonamente decrescentes e as curvas côncavas como (C) para funções de risco monotonamente crescentes. A curva (D), convexa e depois côncava, a função de risco tem forma de "U" e, se a função de risco é inicialmente côncava e depois convexa, assim como em (E), então a taxa de falha é unimodal.

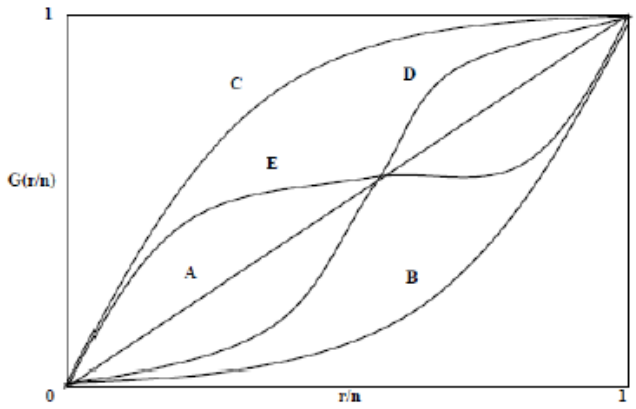


Figura 2.6: Gráfico ilustrativo das curvas TTT-plot (GARCIA, 2013)

2.4.2 Método gráfico

A escolha de modelos pelo método gráfico consiste em comparar a função de sobrevivência estimada dos modelos propostos com a função de sobrevivência estimada pelo Kaplan-Meier, ou seja, é necessário ajustar os modelos ao conjunto de dados para a obtenção das estimativas dos parâmetros e assim construir a função de sobrevivência estimada. O modelo mais adequado é aquele que mais se aproxima da função de sobrevivência estimada por Kaplan-Meier (GARCIA, 2013).

2.4.3 Critérios AIC e BIC

Os mais conhecidos critérios de comparação de modelos são o critério de informação de Akaike - AIC (AKAIKE, 1974) e o critério de informação bayesiano - BIC (SCHUARZ, 1978) e são dados, respectivamente, por

$$AIC = -2\ell(\boldsymbol{\theta}) + 2d \tag{2.29}$$

e

$$BIC = -2\ell(\boldsymbol{\theta}) + dln(n) \tag{2.30}$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros do modelo, d é o número de parâmetros e n o tamanho da amostra e, em ambos os casos, quanto menor o valor, melhor o ajuste do modelo.

2.4.4 Teste da razão de verossimilhanças

O teste da razão de verossimilhanças para modelos encaixados, proposto por Cox e Hinkley (1974), fornece uma conclusão direta sobre qual modelo é mais adequado aos dados. As hipóteses testadas são

$$\begin{cases} H_0: & \text{o modelo de interesse (M) é adequado} \\ H_a: & \text{o modelo geral (G) é adequado} \end{cases} \tag{2.31}$$

e é realizado utilizando a estatística da razão de verossimilhança. Primeiramente, deve-se identificar um modelo geral (G) e calcular o ln de sua função de verossimilhança ($lnL(\widehat{\theta}_G)$). Então, identificar o modelo de interesse (M) e também calcular o ln de sua função de verossimilhança ($lnL(\widehat{\theta}_M)$). A partir desses dados, calcula-se a estatística da razão de verossimilhança dada por

$$TRV = 2[lnL(\widehat{\theta}_G) - lnL(\widehat{\theta}_M)] \tag{2.32}$$

que sob a hipótese nula tem distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros ($\theta_G e \theta_M$) dos modelos a serem comparados (COLOSIMO E GIOLO, 2006).

2.5 Modelo de regressão locação-escala

O modelo de regressão locação-escala emprega famílias paramétricas de distribuições que permitem modelar a variável transformada $Y = ln(T)$, de tal maneira que dado um vetor de covariáveis possua agora um parâmetro de locação μ ($-\infty < \mu < \infty$) que depende das covariáveis e um parâmetro de escala σ ($\sigma > 0$) que é constante. As

distribuições que seguem o modelo de locação-escala tem função de distribuição de probabilidade da seguinte forma

$$f(y;\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{y-\mu}{\sigma}\right), -\infty < y < \infty, \tag{2.33}$$

em que $g(.)$ é a função distribuição de probabilidade considerada. Ou seja, pode-se escrever o modelo log-linear da seguinte forma

$$Y = \mu(\boldsymbol{x}) + \sigma Z, \tag{2.34}$$

em que Y pertence a família de distribuições locação-escala, Z é o resíduo cuja distribuição não depende de $\boldsymbol{x}$, que é o vetor de covariáveis.

2.6 Intervalo de confiança e testes de hípotese sob os parâmetros

Primeiramente, é preciso a apresentação de duas propriedades importantes para a construção de um intervalos de confiança.

Propriedade 1: distribuição assisntôtica do estimador de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$.

Essa propriedade estabelece para grandes amostras e sob condições de regularidade que a distribuição do vetor de parâmetros $\widehat{\boldsymbol{\theta}} = (\theta_1, ..., \theta_n)$ é Normal multivariada com média θ e matriz de variâncias e covariâncias $Var(\widehat{\boldsymbol{\theta}})$ (COLOSIMO E GIOLO, 206).

Propriedade 2: precisão do estimador de $Var(\widehat{\boldsymbol{\theta}})$ que, também sob condições de regularidade, garante que

$$Var(\widehat{\boldsymbol{\theta}}) \approx -[E(\Im)(\theta)]^{-1}$$

em que

$$(\Im)(\theta) = \frac{\partial log(L(\boldsymbol{\theta}))^2}{\partial \boldsymbol{\theta}^2}.$$

em que a diagonal principal da matriz $-[E(\Im)(\theta)]^{-1}$ tem-se as variâncias dos estimadores e nos demais elementos da matriz, encontram-se as convariâncias entre eles.

De posse destas duas propriedades, é necessário, assim, uma estimativa para o erro padrão de $\widehat{\theta}$, ou seja, $\sqrt{Var(t\theta)}$. Então, um intervalo aproximado de $(1-\alpha)100\%$ de confiança para θ é dado por

$$\widehat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\boldsymbol{\theta})} \tag{2.35}$$

em que $z_{\alpha/2}$ é o quantil da distribuição normal com probabilidade $\alpha/2$. Para informações mais detalhadas sobre as pripriedades assintóticas e construção de intervalos de confiança em análise de sobrevivência, recomenda-se a leitura de Colosimo e Giolo (2006).

E o teste de hípotese sob os parâmetros $\widehat{\theta_i}$ que mostra a contribuição de covariáveis no modelo, é dado considerando a hípotese (DEMÉTRIO E ZOCCHI, 2011):

$$\begin{cases} H_0: & \widehat{\theta_i} = 0 \\ H_a: & \widehat{\theta_i} \neq 0 \end{cases}$$

em que a estatística t será dada por:

$$t_{calc} = \frac{\widehat{\beta_i} - \beta_i}{\sqrt{Var(\widehat{\boldsymbol{\theta}})}}, \tag{2.36}$$

e é comparada com o valor da tabela t com $n-p$ graus de liberdade (n= nº de observações e p=nº de parâmetros do modelo considerado). Se $t_{calc} > t_{tab}$ rejeita-se H_0, ou seja, há evidências de que a covariável seja influente no modelo.

3 MATERIAIS

O experimento a ser estudado foi coordenado pela Dr.ª Kathleen F. Grego e Guilherme F. Mendes do Laboratório de Herpetologia do Instituto Butantan em São Paulo-SP (MENDES *et al.*, 2019) e foi composto por 263 serpentes do gênero *Micrurus corallinus* (cobra coral) com a duração de 1554 dias (final do estudo: 15 de julho de 2014). As serpentes foram classificadas de acordo com o sexo (macho e fêmea) e também foram divididas em três grupos, de acordo com o manejo utilizado. Em todos eles as serpentes foram separadas em caixas individuais, receberam um substrato oferecido e um tipo de alimentação, que correspondia a 10% do peso total da serpente. As serpentes que não se alimentaram voluntariamente foram alimentadas via gavagem com um tipo de suplemento que foi modificado de acordo com o manejo. Os grupos são

GRUPO I: pré mudança (1997 à 1999): o substrato era o *Sphagnum* com água a vontade. A alimentação era oferecida uma vez por semana e era composta por animais vivos ou previamente abatidos em CO_2. Geralmente eram oferecidas cobras, lagartos ou anfisbênios (cobra-cega ou cobra de duas cabeças), dependendo da disponibilidade na natureza. Os animais que não se alimentaram receberam uma mistura composta por ração comercial de reptéis e soro fisiológico.

GRUPO II: transição (2010): a frequência de alimentação foi alterada e as serpentes receberam a alimentação por três semanas consecutivas e, após um intervlado de quinze dias, o veneno foi extraído. O substrato segue o mesmo do grupo anterior e a mudança se dá na alimentação da serpentes, que receberam os alimentos provenientes da natureza (cobras, lagartos e anfisbênios) após um período de no mínimo sete dias de congelamento. Também foi iniciada a criação de animais para alimentação das mesmas, especialmente aquelas que não se alimentaram voluntariamente. A mistura era porcionada da seguinte forma: 1L de salina, 15g de ração comercial para serpentes e 500g de serpentes criadas para alimentação.

GRUPO III: pós mudança (2011 à 2014): as mudanças ocorridas no grupo II se mantiveram, exceto o substrato que foi substituído por cascas de árvore.

O objetivo deste trabalho é comprovar que a mudança de manejo aumentou a sobrevivência das serpentes em cativeiro para a obtenção do veneno destinado à produção de soro antiofídico, considerando todas as observações, censuradas e não censuradas, ou seja, comprovar que o grupo III de manejo fornece melhores condições para tal. Pretende-se também verificar se há diferença entre a sobrevivência de machos e fêmeas.

Para as análises desse trabalho, o evento de interesse considerado é a morte da serpente até a data final do estudo. As que sobreviveram depois dessa data são as con-

sideradas censuradas (censura do Tipo I Progressiva (2.2)), ou seja, das 263 observações, 13 são censuradas.

Outros autores aplicaram diferentes técnicas estatísticas sobre variações desse mesmo conjunto de dados. Mendes *et al.* (2019) com um conjunto de 289 serpentes, considerando apenas a covariável grupo (210 no grupo I, 26 no grupo II e 53 no grupo III), utilizaram o estimador não paramétrico de Kaplan-Meier e o teste de Mann-Whitney para mostrar que além de o Grupo III fornecer melhores condições de manejo para a sobrevivência das serpentes, mostram que a troca de substrato (*Sphagnum* para cascas de árvore) aumentou a sobrevivência das mesmas drastricamente e, Silva *et al.* (2016), utilizaram o mesmo conjunto de dados e a mesma covariável, também com o Estimador de Kaplan-Meier, em adição ao teste da razão de verossimilhança e o critério de Akaike para a seleção de modelos, selecionou o modelo de regressão log normal e obtiveram o mesmo resultado, levando em conta de que as censuras nessa análise não foram utilizadas.

4 MÉTODOS

As técnicas não paramétricas, como Kaplan-Meier, são importantes pela simplicidade e facilidade de aplicação e por isso se tornam importantes para descrever dados de sobrevivência. Essas técnicas, no entanto, não permitem a inclusão de covariáveis, o que inviabiliza uma análise mais elaborada dos dados. Portanto, para ver os efeitos de tais covariáveis, é necessário utilizar um modelo de regressão apropriado para dados de sobrevivência. Nesse tópico, serão apresentados os modelos de regressão exponencial, Weibull (que são os mais utilizados pela sua simplicidade e importância), gama, gama generalizada e log-logística. Para informações mais detalhadas destes modelos, recomenda-se a leitura de Colosimo e Giolo (2006), Valença (1994), Lawless (1982), dentre outros.

4.1 Modelo de regressão exponencial

O modelo de regressão exponencial só pode envolver um parâmetro, e por isso é considerado o mais simples. Considerando a variável aleatória T que segue a distribuição exponencial e $Y = ln(T)$, o modelo de regressão log-linear exponencial é dado por

$$Y = \boldsymbol{x}\boldsymbol{\beta} + W \tag{4.1}$$

em que $\boldsymbol{x} = (1, x_1, ..., x_{p-1})$ é o vetor de covariáveis, $\boldsymbol{\beta}$ é o vetor de parâmetros a serem estimados e W tem distribuição do valor extremo padrão. As funções densidade de probabilidade e de sobrevivência de Y dado $\boldsymbol{x}$ são expressas por (VALENÇA, 1994)

$$f(y|\boldsymbol{x}) = exp[y - \boldsymbol{x}\boldsymbol{\beta} - exp(y - \boldsymbol{x}\boldsymbol{\beta})] \tag{4.2}$$

e

$$S(y|\boldsymbol{x}) = exp[-exp(y - \boldsymbol{x}\boldsymbol{\beta})]. \tag{4.3}$$

Para os dados de *Micrurus corallinus* considerando as covariáveis grupo e sexo, tem-se o modelo de regressão exponencial dado por

$$Y = \boldsymbol{g}\boldsymbol{\beta} + W \qquad \text{e} \qquad Y = \boldsymbol{s}\boldsymbol{\beta} + W, \tag{4.4}$$

em que $\boldsymbol{g} = (g_1, g_2, g_3)$ representando o vetor de covariáveis grupo e $\boldsymbol{s} = (s_f, s_m)$ o vetor de covariáveis sexo.

4.2 Modelo de regressão Weibull

Uma forma de generalizar o modelo de regressão exponencial é incluir um parâmetro extra de escala. O modelo então assume a forma

$$Y = ln(T) = \boldsymbol{x\beta} + \sigma W,$$

chamado de modelo de regressão Weibull, pois T deve ter uma distribuição Weibull para que $ln(T)$ tenha a distribuição do valor extremo padrão com parâmetro de escala σ.

As funções densidade de probabilidade e de sobrevivência de Y dado $\boldsymbol{x}$ são (COLOSIMO E GIOLO, 2006; VALENÇA, 1994)

$$f(y|\boldsymbol{x}) = \frac{1}{\sigma} exp\left[\left(\frac{y - \boldsymbol{x\beta}}{\sigma}\right) - exp\left(\frac{y - \boldsymbol{x\beta}}{\sigma}\right)\right] \tag{4.5}$$

e

$$S(y|\boldsymbol{x}) = exp\left[-exp\left(\frac{y - \boldsymbol{x\beta}}{\sigma}\right)\right]. \tag{4.6}$$

Da mesma forma, considerando as covariáveis grupo e sexo, tem-se

$$Y = \boldsymbol{g\beta} + \sigma W \qquad \text{e} \qquad Y = \boldsymbol{s\beta} + \sigma W. \tag{4.7}$$

4.3 Modelo de regressão gama

O modelo de regressão gama é visto como um modelo útil, mesmo com algumas dificuldades para se obter estimativas como a função de sobrevivência. Encontram-se diferentes opções na literatura para o processo de inferência da distribuição gama. Lawless (1982) diz que basta considerar o modelo de estimação de máxima verossimilhança visto em (2.25) e Colosimo e Giolo (2006) consideram o modelo de regressão apropriado para tal distribuição como sendo o modelo tempo de vida acelerado, quando as covaráveis tem o efeito de acelerar ou desacelerar o tempo de vida, que consiste em

$$Y = ln(T) = \boldsymbol{x\beta} + \sigma \upsilon \tag{4.8}$$

em que υ é o termo de erro com uma distribuição apropriada. A função de sobrevivência nesse caso será dada por

$$S(t|\boldsymbol{x}) = S_0(te^{-\boldsymbol{x'\beta}}) \tag{4.9}$$

em que S_0 é a função de sobrevivência da distribuição desejada, nesse caso, a distribuição gama (2.16), então

$$S(t|\boldsymbol{x}) = 1 - I(k, te^{-\boldsymbol{x'\beta}}/a). \tag{4.10}$$

O modelo de regressão considerando os dados de *Micrurus corallinus* fica

$$Y = \boldsymbol{g\beta} + \sigma \upsilon \qquad \text{e} \qquad Y = \boldsymbol{s\beta} + \sigma \upsilon. \tag{4.11}$$

4.4 Modelo de regressão gama generalizada

Os modelos de regressão exponencial e Weibull são os mais utilizados por sua praticidade. No entanto, o modelo de regressão gama generalizada é considerado mais flexível por conta de seus três parâmetros. Quando iniciados os estudos do processo de inferência para esses três parâmetros, foram observadas dificuldades (LAWLESS, 1982) e para contornar algumas delas, Prentice (1974) trabalhou com uma forma reparametrizada a fim de reduzir tal dificuldade e tornar as propriedades mais flexíveis.

Supondo que T seja a função distribuição de probabilidade gama generalizada e considerando $Y = ln(T)$, o modelo log-linear (2.33) é escrito como

$$Y = ln(a) + b^{-1}Z, \tag{4.12}$$

em que Z tem distribuição log gama com densidade dada por

$$f_Z(z) = \frac{1}{\Gamma(k)} exp(zk - e^z). \tag{4.13}$$

Prentice (1974) considera $q = k^{-c}$ com $c > 0$ e opta por $c = 1/2$ tendo em vista que a distribuição de Y se aproxima da normalidade apenas para esse valor (VALENÇA, 1994). Então:

$$Y = ln(a) + b^{-1}Z = ln(a) + b^{-1}(q\varepsilon + \mu^*) = ln(a) + \mu^* b^{-1} + qb^{-1}\varepsilon \Rightarrow$$

$$\Rightarrow Y = \gamma + \sigma\varepsilon, \tag{4.14}$$

em que $\gamma = ln(a) + \mu^* b^{-1}$ e $\sigma = qb^{-1}$. A função distribuição de probabilidade de Y pode ser escrita como $f_Y = f_Z(z(y)|J)$, $z(y) = \dfrac{y-\gamma}{\sigma}q + \mu^*$ e $J = \dfrac{\partial z(y)}{\partial y} = \dfrac{q}{\sigma}$ e portanto:

$$f_Y(y;\gamma,\sigma,q) = \frac{q}{\sigma\Gamma(q^{-2})} exp\left[\left(\frac{y-\gamma}{\sigma} + \mu^*\right) q^{-2} - exp\left(\frac{y-\gamma}{\sigma}q + \mu^*\right)\right] \tag{4.15}$$

Quando $k \to \infty$ esse modelo é deslocado cada vez mais para a direita e a esperança e a variância tendem a infinito. Prentice *et al.* (1977) e Lawless e Singhal (1980) propõe uma simplificação do modelo (4.15):

$$W = \frac{Z - ln(k)}{k^{-1/2}} = \frac{Z - ln(q^{-2})}{q},$$

em que $f_W = \dfrac{1}{\Gamma(k)} k^{k-1/2} exp\left\{w\sqrt{k} - kexp\left[\dfrac{w}{\sqrt{k}}\right]\right\}$. E então, com $k = q^{-2}$, tem-se

$$f_W(w,q) = \frac{1}{\Gamma(q^{-2})}(q^{-2})^{q^{-2}-1/2} exp\{q^{-1}w - q^{-2}e^{qw}\}. \tag{4.16}$$

Reparametrizando da mesma forma vista em (4.14), o modelo (4.16) é escrito como

$$Y = \mu + \sigma w, \tag{4.17}$$

em que $\mu = ln(a) + b^{-1}ln(q^{-2})$ e $\sigma = qb^{-1}$.

Uma extensão do modelo (acima) é dado considerando $k \in \Re$ e é chamado de modelo log-gama generalizado estendido dado por

$$f_Y(y;\mu,\sigma,q) = \begin{cases} \frac{|q|}{\sigma\Gamma(q^{-2})}(q^{-2})^{q^{-2}}exp\left\{q^{-2}\left[\frac{y-\mu}{\sigma} - exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}, q \neq 0 \\ \\ \frac{1}{\sqrt{2\pi\sigma^2}}exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right], q = 0 \end{cases}, \tag{4.18}$$

e a função de sobrevivência de Y é escrita da seguinte forma

$$S(y) = \begin{cases} Q(q^{-2}, q^{-2}exp(wq)), q > 0 \\ \\ 1 - Q(q^{-2}, q^{-2}exp(wq)), q < 0 \\ \\ 1 - \Phi(w), q = 0 \end{cases}, \tag{4.19}$$

em que $Q(k,x) = \int_0^\infty \frac{x^{k-1}e^{-x}}{\Gamma(k)}dx$ e $\Phi(w) = \int_{-\infty}^{w} = \frac{1}{(2\pi)^{1/2}}exp\left(-\frac{x^2}{2}\right)dx$.

Com essas informações, o modelo de regressão gama generalizado é descrito da seguinte maneira: considere um modelo de regressão que assume uma relação linear entre o $ln(Y)$ e o vetor de covariáveis $\boldsymbol{x}$, então Y dado $\boldsymbol{x}$ tem distribuição log gama generalizada e pode ser escrito como (VALENÇA, 1994)

$$Y = \boldsymbol{x}\boldsymbol{\beta} + \sigma Z, \tag{4.20}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$, σ são parâmetros desconhecidos e Z tem distribuição dada por

$$f_Z(z;q) = \begin{cases} \frac{|q|}{\sigma\Gamma(q^{-2})}(q^{-2})^{q^{-2}}exp\{q^{-1}z - q^{-2}exp(qz)\}, q \neq 0 \\ \\ \frac{1}{\sqrt{2\pi}}exp\left\{-\frac{z^2}{2}\right\}, q = 0 \end{cases}, \tag{4.21}$$

e a função de sobrevivência de Y dado $\boldsymbol{x}$ é

$$S(y|\boldsymbol{x}) = \begin{cases} Q(q^{-2}, q^{-2}exp(q(y - \boldsymbol{x}\beta/\sigma)), q > 0 \\ 1 - Q(q^{-2}, q^{-2}exp(q(y - \boldsymbol{x}\beta/\sigma)), q < 0 \\ 1 - \Phi((y - \boldsymbol{x}\beta/\sigma)), q = 0 \end{cases} . \tag{4.22}$$

Considerando então o modelo (4.17) e os dados de *Micrurus corallinus* tem-se

$$Y = \boldsymbol{g\beta} + \sigma Z \qquad \text{e} \qquad Y = \boldsymbol{s\beta} + \sigma Z. \tag{4.23}$$

4.5 Modelo de regressão log-logístico

O modelo de regressão log-logístico também é baseado no modelo de regressão tempo de vida acelerado, que leva em conta a princípio o modelo

$$Y = ln(T) = \boldsymbol{x\beta} + \sigma\upsilon, \tag{4.24}$$

em que agora υ segue a distribuição log-logística. Portanto, considerando o logaritmo dos dados observados, tem-se

$$f(y) = \frac{1}{\sigma} exp\left(\frac{y-\mu}{\sigma}\right)\left[1 + exp\left(\frac{y-\mu}{\sigma}\right)\right]^{-2}, \tag{4.25}$$

com μ e σ parâmetros de locação e escala, respectivamente, como visto em (2.31). A função de sobrevivência se torna

$$S(y) = \frac{1}{1 + exp\left(\frac{y-\mu}{\sigma}\right)}. \tag{4.26}$$

Considerando o modelo tempo de vida acelerado e as equações (4.25) e (4.26), temos para o modelo de regressão log-logistíco

$$f(y|\boldsymbol{x}) = \frac{1}{\sigma} exp\left(\frac{y-\boldsymbol{x\beta}}{\sigma}\right)\left[1 + exp\left(\frac{y-\boldsymbol{x\beta}}{\sigma}\right)\right]^{-2} \tag{4.27}$$

e

$$S(y|\boldsymbol{x}) = \frac{1}{1 + exp\left(\frac{y-\boldsymbol{x\beta}}{\sigma}\right)}. \tag{4.28}$$

O modelo de regressão considerando as covariáveis grupo e sexo é análogo ao visto em (4.11).

4.6 Odd log-logística generalizada G (OLLG-G)

Cordeiro *et al.* (2017) afirmam que as recentes pesquisas se concentram na extensão de famílias já existentes (ou seja, novas famílias) podendo proporcionar grande flexibilidade na modelagem de dados. Então, seja $G(t,\xi)$ a função de distribuição acumulada em que $\boldsymbol{\xi}$ é um vetor de $p \times 1$ de parâmetros desconhecidos. Gleaton e Lynch (2006) definiram a função de distribuição acumulada da família log-logística generalizada, com um parâmetro de forma adicional $\alpha > 0$, como

$$H = \frac{G(t,\boldsymbol{\xi})^{\alpha}}{G(t,\boldsymbol{\xi})^{\alpha} + \bar{G}(t,\boldsymbol{\xi})^{\alpha}}, \tag{4.29}$$

em que $\bar{G}(t,\boldsymbol{\xi})^{\alpha} = 1 - G(t,\xi)^{\alpha}$.

Portanto, baseados na transformação proposta por Alzaatreh *et al.* (2013), Cordeiro *et al.* (2017) apresentaram uma nova classe de distribuições chamada odd log-logística generalizada G (OLLG-G), dada pela integração da função de densidade log-logística e tem função de distribuição acumulada dada por

$$F(t;\alpha,\lambda,\boldsymbol{\xi}) = \int_0^{G(t,\boldsymbol{\xi})^{\lambda}/(1-G(t,\boldsymbol{\xi})^{\lambda})} \frac{\alpha t^{\alpha-1}}{(1+t^{\alpha})^2} dt = \frac{G(t,\boldsymbol{\xi})^{\alpha\lambda}}{G(t,\boldsymbol{\xi})^{\alpha\lambda} + [1-G(t,\boldsymbol{\xi})]^{\alpha\lambda}}, \tag{4.30}$$

em que $\alpha > 0$ e $\lambda > 0$ são parâmetros de forma adicionais.

A função densidade de probabilidade de (4.30) é dada por

$$f(t;\alpha,\lambda,\boldsymbol{\xi}) = \frac{\alpha\lambda g(t,\boldsymbol{\xi}) G(t,\boldsymbol{\xi})^{\alpha\lambda-1}[1-G(t,\boldsymbol{\xi})^{\lambda}]^{\alpha-1}}{\{G(t,\boldsymbol{\xi})^{\alpha\lambda} + [1-G(t,\boldsymbol{\xi})^{\lambda}]^{\alpha}\}^2}, \tag{4.31}$$

em que $g(t,\boldsymbol{\xi})$ é a função densidade de probabilidade da distribuição desejada.

A equação (4.31) é uma família de distribuições contínuas, incluindo casos especiais

i) Se $\alpha = 1$ tem-se

$$F(t;\alpha,\lambda,\boldsymbol{\xi}) = G(t,\boldsymbol{\xi})^{\lambda} \tag{4.32}$$

e

$$f(t;\alpha,\lambda,\boldsymbol{\xi}) = \lambda G(t,\boldsymbol{\xi})^{\lambda-1} g(t), \tag{4.33}$$

chamada família exponenciada,

ii) Se $\lambda = 1$ tem-se

$$F(t;\alpha,\lambda,\boldsymbol{\xi}) = \frac{G(t)^{\alpha}}{G(t)^{\alpha} + [1-G(t)]^{\alpha}}, \tag{4.34}$$

que é a família odd log-logistica vista em (4.29).

4.6.1 Odd log-logística generalizada Weibull

Tomando agora $G(t;\boldsymbol{\xi})$ e $g(t;\boldsymbol{\xi})$ na equação (4.20) como função densidade de probabilidade da Weibull (2.14) com função de distribuição acumulada $G(t;a,b) = 1 - e^{-(x/b)^a}$ em que $a > 0$ é o parâmetro de forma, $b > 0$ é o parâmetro de escala e $\boldsymbol{\xi} = (a,b)^T$. As funções de distribuição acumulada, de sobrevivência e densidade de probabilidade são dadas por

$$F(t;\alpha,\lambda,a,b) = \frac{[1-e^{-(t/b)^a}]^{\alpha\lambda}}{[1-e^{-(t/b)^a}]^{\alpha\lambda} + [1-([1-e^{-(t/b)^a}]^{\alpha\lambda})]^{\alpha\lambda}} \tag{4.35}$$

$$S(t;\alpha,\lambda,a,b) = 1 - \frac{[1-e^{-(t/b)^a}]^{\alpha\lambda}}{[1-e^{-(t/b)^a}]^{\alpha\lambda} + [1-([1-e^{-(t/b)^a}]^{\alpha\lambda})]^{\alpha\lambda}} \tag{4.36}$$

$$f(t;\alpha,\lambda,a,b) = \frac{\alpha\lambda a t^{a-1} e^{-(t/b)^a}[1-e^{-(t/b)^a}]^{\alpha\lambda-1}\{1-[1-e^{-(t/b)^a}]^{\lambda}\}^{\alpha-1}}{b^a\{[1-e^{-(t/b)^a}]^{\alpha\lambda} + \{1-[1-e^{-(t/b)^a}]^{\lambda}\}^{\alpha}\}^2}, \tag{4.37}$$

e denota-se $T \sim$ OLLG-W(α,λ,a,b).

Considerando a função densidade de probabilidade dada em (4.37) é possível obter casos particulares

i) Se $\alpha = \lambda = 1$ tem-se

$$f(t;\alpha,\lambda,a,b) = \frac{a t^{a-1} e^{-(t/b)^a}}{b^a\{[1-e^{-(t/b)^a}] + \{1-[1-e^{-(t/b)^a}]\}}$$

obtem-se a distribuição Weibull.

ii) Se $\lambda = 1$ tem-se

$$f(x;\alpha,\lambda,a,b) = \frac{\alpha a t^{a-1} e^{-(t/b)^a}[1-e^{-(t/b)^a}]^{\alpha-1}\{1-[1-e^{-(t/b)^a}]\}^{\alpha-1}}{b^a\{[1-e^{-(t/b)^a}]^{\alpha} + \{1-[1-e^{-(t/b)^a}]\}^{\alpha}\}^2}$$

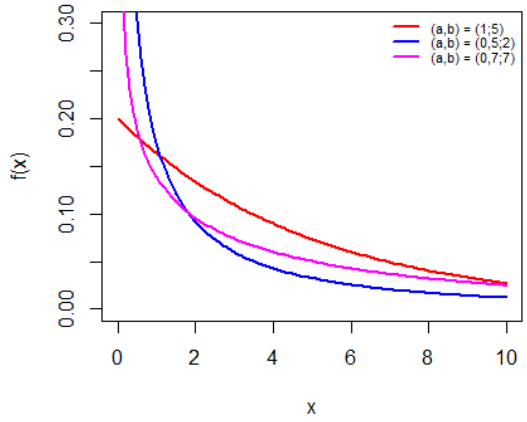
obtem-se a distribuição odd log-logística Weibull dada por Gleaton e Lynch (2006).

iii) Se $\alpha = 1$ tem-se

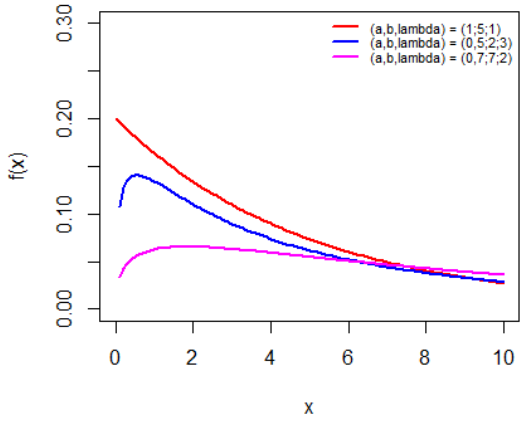
$$f(x;\alpha,\lambda,a,b) = \frac{\lambda a t^{a-1} e^{-(t/b)^a}[1-e^{-(t/b)^a}]^{\lambda-1}}{b^a\{[1-e^{-(t/b)^a}]^{\lambda} + \{1-[1-e^{-(t/b)^a}]^{\lambda}\}\}^2}$$

obtem-se a distribuição Weibull exponenciada.

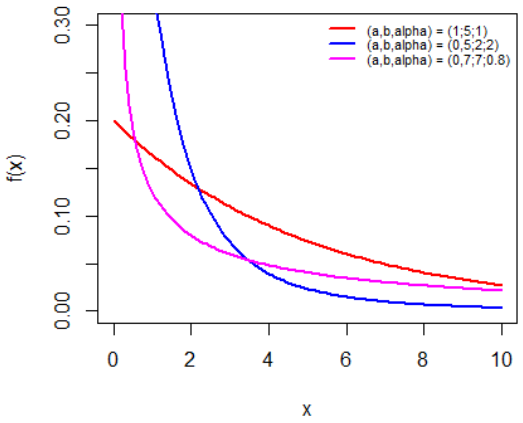
A figura (4.1) apresenta gráficos da distribuição odd log-logística generalizada Weibull para diferentes valores dos parâmetros.



(a)



(b)



(c)

Figura 4.1: (a) distribuição odd log-logística generalizada Weibull para $\alpha = \lambda = 1$ (distribuição Weibull), (b) $\lambda = 1$ (log-logística Weibull), (c) $\alpha = 1$ (Weibull exponenciada).

4.6.2 Modelo de regressão paramétrico LOLLG-W

Na literatura, diversos modelos de regressão em análise de sobrevivência foram propostos, por exemplo, por Ortega *et al.* (2014), que estuda o modelo log-linear para a distribuição log Weibull e Ortega *et al.* (2013), que analisa o modelo de regressão log-beta Weibull para predizer a recorrência de câncer de próstata.

Cordeiro *et al.* (2017) supõe que $\boldsymbol{x}$ denota um conjunto de covariáveis que pode influenciar o tempo de sobrevivência T. Então, pode-se considerar a densidade de T dado $\mathbf{x}$ denotada por $f(t|\mathbf{x})$ com função de sobrevivência $S(t|\mathbf{x})$.

Propõe-se o modelo de regressão com estrutura log-linear, considerando que a variável ateatória T segue a distribuição OLLG-W$(\alpha, \lambda, \boldsymbol{\xi})$ e o modelo considerado nesse

trabalho é o modelo de regressão paramétrico. Em alguns casos, a variável definida por $Y = ln(T)$ pertence aos modelos de locação e escala (2.31), os quais são caracterizados por um parâmetro de locação $-\infty < \mu < \infty$ e um parâmetro de escala $0 < \sigma < \infty$. No caso da distribuição OLLG-W, a função densidade de probabilidade de Y é obtida chamando $\alpha = 1/\sigma$ e $b = e^{\mu}$ e é dada por

$$\begin{aligned} \text{f(y)} = & \frac{\alpha\lambda}{\sigma} exp\left[\left(\frac{y-\mu}{\sigma}\right) - exp\left(\frac{y-\mu}{\sigma}\right)\right]\left\{1 - exp\left[-exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\alpha\lambda-1} \\ & \times \left[1 - \left\{1 - exp\left[-exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\lambda}\right]^{\alpha-1} \\ & \times \left\{\left\{1 - exp\left[-exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\alpha\lambda} + \left[1 - \left\{1 - exp\left[-exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{\lambda}\right]^{\alpha}\right\}^{-2}. \end{aligned} \tag{4.38}$$

A equação (4.38) é chamada de distribuição log odd log-logística generalizada Weibull (LOLLG-W) e diz-se $Y \sim$ LOLLG-W $(\alpha, \lambda, \sigma, \mu)$. A função de sobrevivência de Y é dada por

$$S(y) = \frac{[1 - \{1 - exp[-exp\left(\frac{y-\mu}{\sigma}\right)\}^{\lambda}]^{\alpha}}{\{1 - exp[-exp(\frac{y-\mu}{\sigma})]\}^{\alpha\lambda} + [1 - \{1 - exp[-exp(\frac{y-\mu}{\sigma})]\}^{\lambda}]^{\alpha}}. \tag{4.39}$$

Considerando a variável aleatória padronizada $\mathbf{Z} = \dfrac{Y-\mu}{\sigma}$ com função de densidade

$$\begin{aligned} f(z) = \alpha\lambda exp[(z) - exp(z)]\{1 - exp[-exp(z)]\}^{\alpha\lambda-1}[1 - \{1 - exp[-exp(z)\}^{\lambda}]^{\alpha-1} \\ \times \\ \{\{1 - exp[-exp(z]\}^{\alpha\lambda} + [1 - \{1 - exp[-exp(z)]\}^{\lambda}]^{\alpha}\}^{-2}. \end{aligned} \tag{4.40}$$

Baseados na função de densidade da LOLLG-W, Cordeiro *et al.* (2017) propõe a regressão linear locação-escala para estimar a variável resposta y_i, dada por

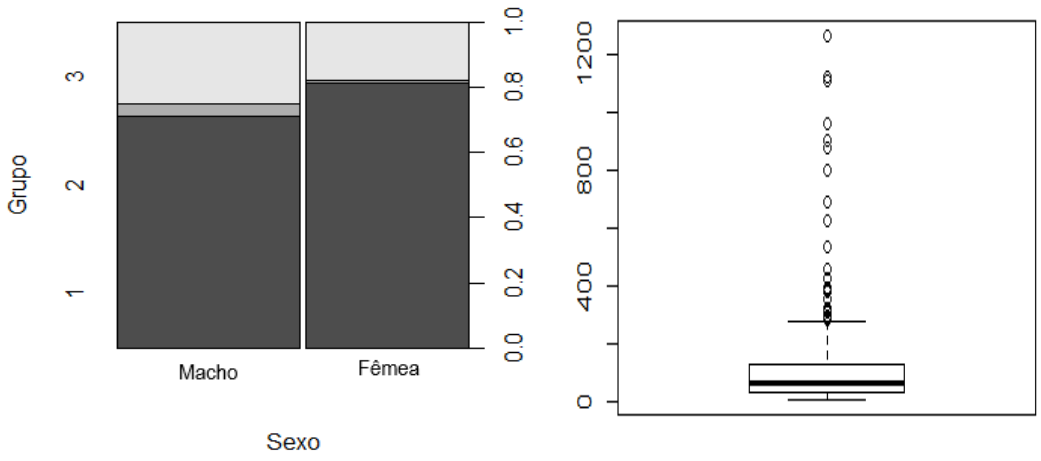
$$y_i = \boldsymbol{v}_i^T \boldsymbol{\beta} + \sigma z_i, \tag{4.41}$$

em que $\boldsymbol{v}_i^T = (v_{i1}, ..., v_{ip})$ é o vetor de covariáveis, z_i é como foi definido em (4.40), $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, $\sigma > 0$ e $\alpha > 0$. O parâmetro $\mu_i = \boldsymbol{v}_i^T \boldsymbol{\beta}$ é de locação de y_i.

Cordeiro *et al.* (2017) afirmam que o estimador de máxima verossimilhança visto na seção (2.3), é um método adequado para a realização de inferência sob os parâmetros do modelo de regressão LOLLG-W.

5 RESULTADOS E DISCUSSÃO

Foi utilizado como auxílio na análise dos dados explanados no capítulo 4 o *software* R (R DEVELOPMENT CORE TEAM, 2014) com o auxílio das bibliotecas *survival*, *flexsurv* e *flexsurvcure*. De acordo com análise descritiva dos dados, tem-se que das 263 observações (sendo 13 censuradas), 139 serpentes são fêmeas e 124 são machos; 200 serpentes pertencem ao grupo I, 6 ao grupo II e 57 ao grupo III. Foram calculadas medidas de interesse tais como mediana (64 dias), média (134,1 dias) e desvio padrão (216,6 dias). O gráfico Boxplot mostra que os dados aparentemente possuem uma distribuição assimétrica e mostra também a presença de *outliers*.



(a) (b)

Figura 5.1: (a) Proporção de tempo de vida das serpentes por sexo e grupo; (b) Boxplot para os tempos de vida de *Micrurus corallinus*

Na figura 5.2 vê-se o gráfico TTT-plot para os tempos de vida que mostra uma curva incialmente côncava e depois convexa, assim como a curva (E) da figura 2.1, indicando que função de falha aparentemente seja unimodal sugerindo o uso das distribuições exponencial, gama, gama generalizada e log-logística. A distribuição Weibull também foi considerada para comparação, uma vez que ela é um caso particular da gama generalizada.

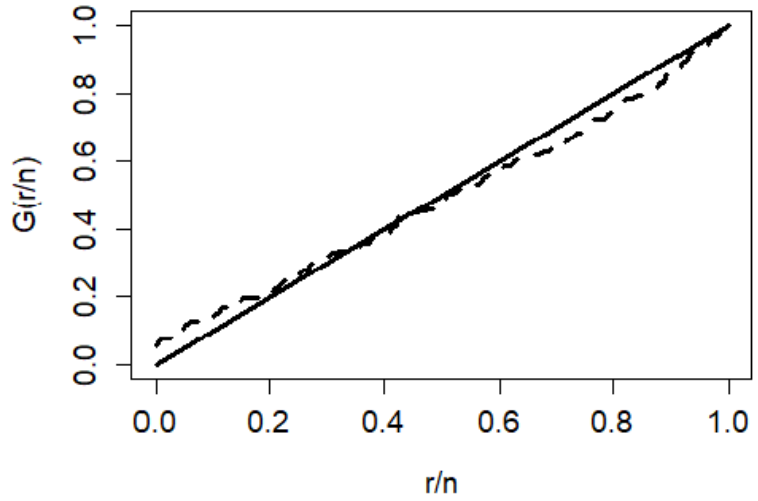


Figura 5.2: Gráfico TTT-plot para os dados de tempo de vida de *Micrurus corallinus*

Para averiguar qual das distribuições melhor se adapta aos dados as covariáveis não foram considareadas em princípio. Também, optou-se por comparar incialmente as distribuições exponencial, Weibull, gama e gama generalizada. Na tabela 5.1 são apresentadas as estimativas dos parâmetros utilizando o método da máxima verossimilhança (seção 2.3) juntamente com os intervalos de confiança de 95% para cada parâmetro. É possível ver que os intevalos de confiança dos parâmetros estimados da distribuição Weibull possuem o valor zero dentro deles, também que os intervalos de confiança para os parâmetros das distribuições gama e gama generalizada são bem proximos ao valor estimado, indicando que aparentemente eles estão bem ajustados.

Tabela 5.1: Estimativas e intervalos de confiança dos parâmetros para as distribuições exponencial, Weibull, gama e gama generalizada obtidas pelo método da máxima verossimilhança

Distribuição	$\widehat{a}$	$\widehat{b}$	$\widehat{k}$
Exponencial	141,029	-	-
	(135,398 ; 417,478)	-	-
Weibull	120,865	0,774	-
	(-225,411 ; 467,142)	(-345,502 ; 347,052)	-
Gama	0,005	-	0,755
	(0,004 ; 0,006)	-	(0,649 ; 0,878)
Gama generalizada	$7{,}46 \times 10^{-12}$	0,125	42
	$(3{,}55 \times 10^{-17}; 1{,}57 \times 10^{-6})$	(0,095 ; 0,163)	(24,2 ; 72,8)

De posse das estimativas dos parâmetros escritas na tabela 5.1, pode-se escrever as funções de sobrevivência estimadas das distribuições exponencial, Weibull, gama e gama generalizada.

$$\widehat{S(t)}_E = exp\left[-\left(\frac{t}{141,0291}\right)\right],$$

$$\widehat{S(t)}_W = exp\left[-\left(\frac{t}{120,8658}\right)^{0,7747}\right],$$

$$\widehat{S(t)}_G = \int_t^{\infty} \frac{1}{\Gamma(0,7551)} u^{0,7551-1} exp\left[-\left(\frac{u}{0,0053}\right)\right] du$$

e

$$\widehat{S(t)}_{GG} = 1 - \frac{\gamma(42, (t/7,46 \times 10^{-12})^{0,125})}{\Gamma(42)}.$$

O gráfico 5.3 mostra a curva de Kaplan-Meier e as curvas de sobrevivência estimadas das distribuições. A curva que mais se aproxima da curva estimada por Kaplan-Meier é a pertencente a distribuição gama generalizada, ou seja, aparentemente a distribuição que melhor se ajusta aos dados é a distribuição gama generalizada.

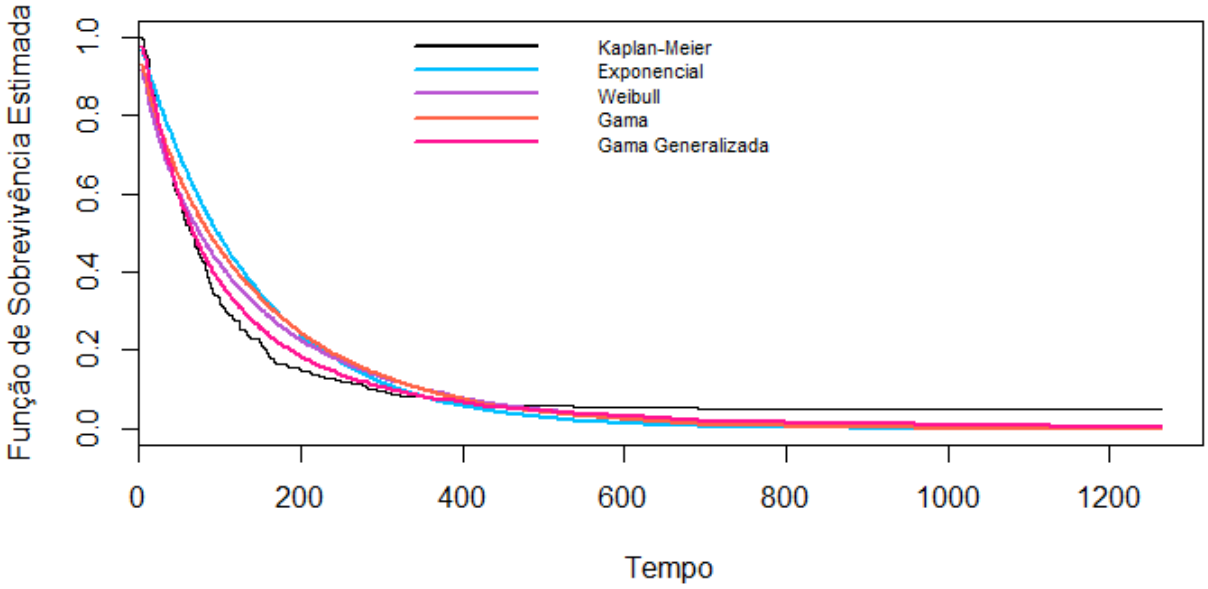


Figura 5.3: Comparação da função de sobrevivência estimada pelo método de Kaplan-Meier e das funções de sobrevivência estimadas das distribuições exponencial, Weibull, gama e gama generalizada

O teste de razão de verossimilhança foi aplicado considerando três hípoteses, em todas considerando o modelo geral como o modelo gama generalizado, obtendo-se as seguintes hipóteses a serem testadas

$$Hip.1 : \begin{cases} H_0 : & \text{o modelo exponencial é adequado} \\ H_a : & \text{o modelo gama generalizado é adequado} \end{cases}$$

$$Hip.2 : \begin{cases} H_0 : & \text{o modelo Weibull é adequado} \\ H_a : & \text{o modelo gama generalizado é adequado} \end{cases}$$

$$Hip.3 : \begin{cases} H_0 : & \text{o modelo gama é adequado} \\ H_a : & \text{o modelo gama generalizado é adequado} \end{cases}$$

e dessas três hipóteses, obtém-se os resultados vistos na tabela a seguir

Tabela 5.2: Resultados do Teste da Razão de Verossimilhanças

Distribuição	-logLik	p-valor
Exponencial	1487,26	$9{,}61\text{x}10^{-24} < 0,01$
Weibull	1469,08	$7{,}11\text{x}10^{-17} < 0,01$
Gama	1480,08	$1{,}03\text{x}10^{-21} < 0,01$
Gama generalizada	1434,26	-

Em todas as hipóteses rejeita-se H_0 ao nível $\alpha = 1\%$ de significância e portanto, o modelo que aparentemente se adequa melhor aos dados é, também, o modelo gama generalizado. Os resultados vistos acima são confirmados aplicando-se os critérios de comparação de modelos AIC e BIC, com estimativas apresentadas na tabela 5.5.

Tabela 5.3: Critérios de comparação de modelos AIC e BIC das distribuições exponencial, Weibull, gama e gama generalizada

Distribuição	AIC	BIC
Exponencial	2976,52	1493,83
Weibull	2942,16	1476,65
Gama	2964,18	1487,66
Gama generalizada	2874,52	1442,83

Na tabela 5.4 são apresentadas as estimativas dos parâmetros utilizando o Método da Máxima Verossimilhança e o intervalo de confiança de 95% para a distribuição log-logística:

Tabela 5.4: Estimativas e intervalos de confiança dos parâmetros para a distribuição log-logística

Distribuição	$\widehat{a}$	$\widehat{b}$
Log-logística	62,6206	1,4459
	(54,1904 ; 72,3622)	(1,3043 ; 1,6030)

E de posse das estimativas dos parâmetros, pode-se escrever a função de sobrevivência estimada para a distribuição log-logística

$$\widehat{S(t)}_{LL} = 1 - \frac{1}{1 + (t/62,6206)^{1,4459}}.$$

O gráfico de comparação 5.4 mostra que a distribuição log-logística aparentemente se adapta melhor aos dados uma vez que sua função de sobrevivência se ajusta melhor à curva de sobrevivência estimada pelo método de Kaplan-Meier. Esse resultado pode ser observado a partir das estimativas de AIC e BIC contidas na tabela 5.5.

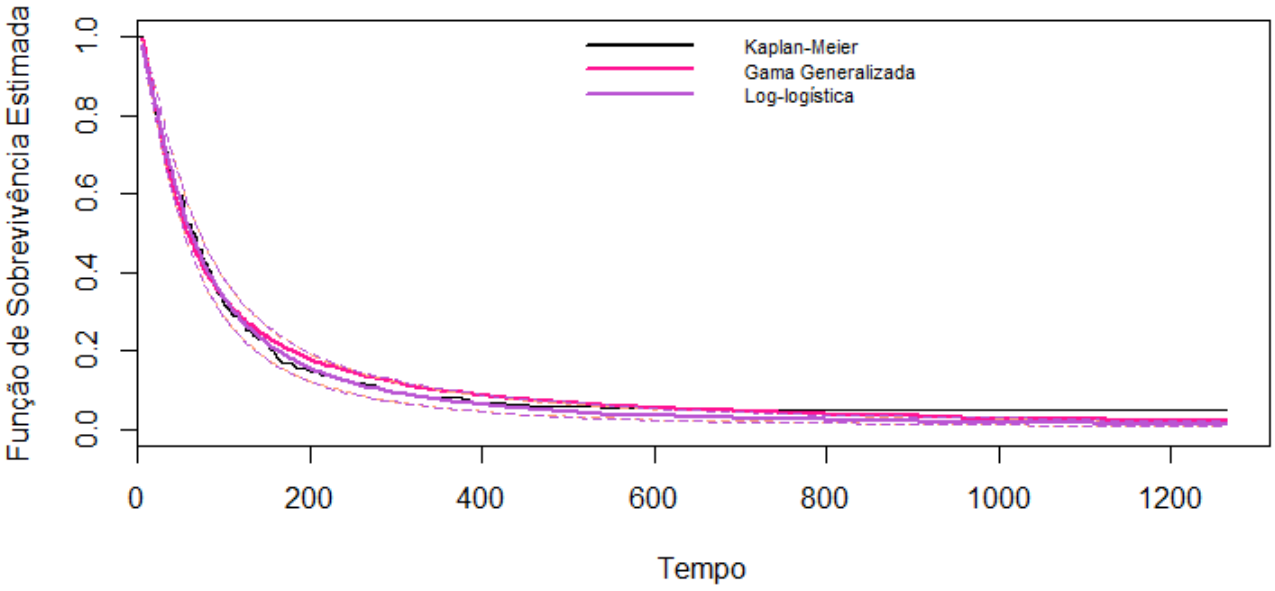


Figura 5.4: Gráfico de comparação da função de sobrevivência estimada pelo método de Kaplan-Meier e das funções de sobrevivência estimadas das distribuições gama generalizada e log-logística

Tabela 5.5: Avaliadores de qualidade AIC e BIC das distribuições gama generalizada e log-logística

Distribuição	AIC	BIC
Gama generalizada	2874,52	2878,09
Log-logística	2858,58	2865,73

A distribuição que mais se ajusta aos tempos de vida de *Micrurus corallinus*, dentre as propostas na literatura, é a distribuição log-logística. Com isso, compara-se a

distribuição odd log-logística generalizada Weibull com a distribuição log-logística a fim de melhorar, ou não, o ajuste dos dados.

Na tabela 5.6, encontram-se as estimativas para os parâmetros da distribuição OLLG-W e seus respectivos intervalos de confiança de 95%. Observa-se que os intervalos de confiança para os parâmetros estimado $\widehat{a}$, $\widehat{\alpha}$ e $\widehat{\lambda}$ incluem o valor zero.

Tabela 5.6: Estimativas e intervalos de confiança dos parâmetros para as distribuição OLLG-W obtidas pelo Método da Máxima Verossimilhança

Distribuição	$\widehat{a}$	$\widehat{b}$	$\widehat{\alpha}$	$\widehat{\lambda}$
OLLG-W	0,1662	118,4668	5,2844	1,3338
	(-242,29 ; 242,32)	(118,24 ; 118,68)	(-1,77 ; 12,34)	(-100,61 ; 103,28)

E de posse das estimativas dos parâmetros, é possível escrever a função de sobrevivência estimada para distribuição OLLG-W

$$\widehat{S(t)}_O = 1 - \frac{[1 - e^{-(t/118{,}47)^{0{,}17}}]^{5{,}28.1{,}33}}{[1 - e^{-(t/118{,}47)^{0{,}17}}]^{5{,}28.1{,}33} + [1 - ([1 - e^{-(t/118{,}47)^{0{,}17}}]^{5{,}28.1{,}33})]^{5{,}28.1{,}33}}$$

O gráfico 5.5 possibilita a comparação da curva de sobrevivência estimada por Kaplan-Meier com as curvas de sobrevivência estimadas a partir dos parâmetros das distribuições log-logística e OLLG-W, que não dá um resultado exato sobre qual distribuição se ajusta melhor aos dados. Portanto, com o cálculo dos critérios de comparação de modelos AIC e BIC (tabela 5.7), é possível afirmar que a distribuição que melhor se adequa aos dados é a distribuição log-logística. Essa falha no ajuste da distribuição OLLG-W pode ter

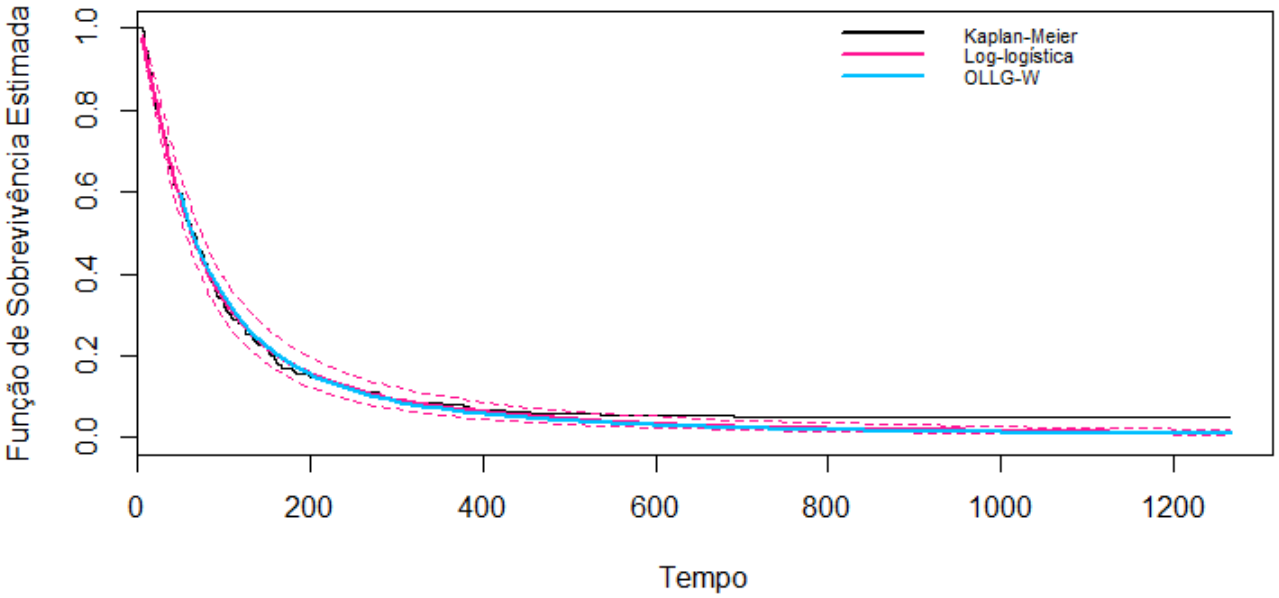


Figura 5.5: Comparação da função de sobrevivência estimada pelo método de Kaplan-Meier e das funções de sobrevivência estimadas das distribuições log-logística e OLLG-W

Tabela 5.7: Avaliadores de qualidade AIC e BIC das distribuições log-logística e OLLG-W

Distribuição	AIC	BIC
Log-logística	2858,58	2865,73
OLLG-W	2863,89	2867,46

Considerando o modelo de regressão log-logístico e adicionando o efeito das covariáveis sexo e grupo separadamente, tem-se

$$Y = \boldsymbol{g}\boldsymbol{\beta} + \sigma\upsilon$$

e

$$Y = \boldsymbol{s}\boldsymbol{\beta} + \sigma\upsilon.$$

em que $\boldsymbol{g} = (g_1, g_2, g_3)$ é o vetor de covariáveis grupo e $\boldsymbol{s} = (s_f, s_m)$ é o vetor de covariáveis sexo.

Avaliando as medidas de ajuste AIC e BIC (tabela 5.8) tem-se que o modelo significativo é aquele em que é considerada a covariável grupo.

As estimativas para os parâmetros $\boldsymbol{\beta}$ são apresentadas na tabela 5.9

Tabela 5.8: Avaliadores de qualidade AIC e BIC da distribuição log-logística considerando as covariáveis sexo e grupo

Distribuição	AIC	BIC
Sem covariável	2858,58	2865,73
Grupo	2743,88	2758,17
Sexo	2859,57	2870,28

Tabela 5.9: Estimativa para o vetor de parâmetros $\boldsymbol{\beta}$

-	Estimativa	Erro Padrão	p-valor
$\widehat{\beta_1}$ (Grupo 1)	3,77		-
$\widehat{\beta_2}$ (Grupo 2)	1,06		$1{,}64 \times 10^{-3}$
$\widehat{\beta_3}$ (Grupo 3)	1,75		$8{,}37 \times 10^{-31}$

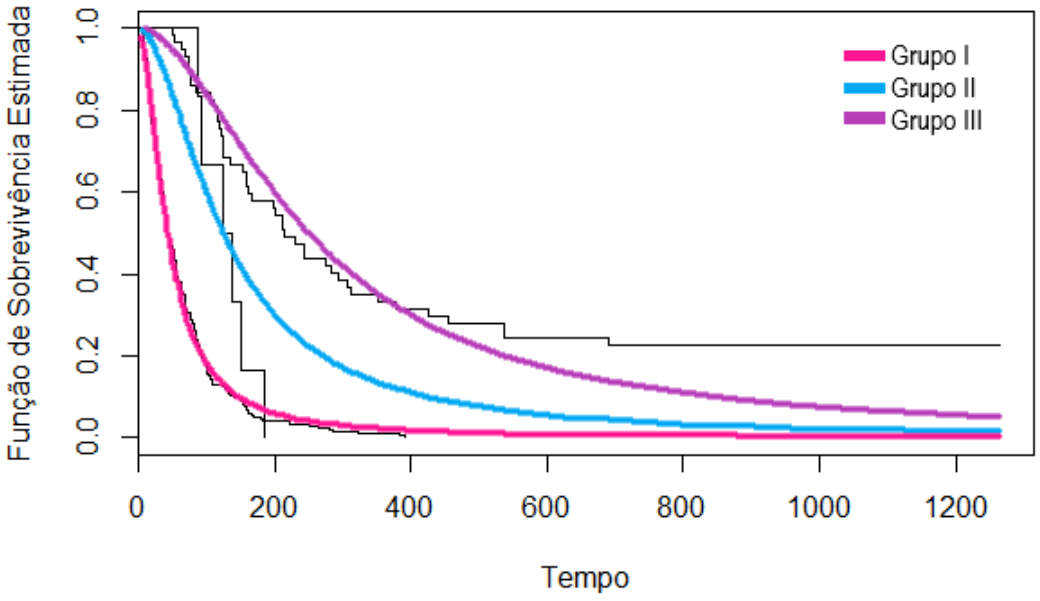


Figura 5.6: Comparação da função de sobrevivência estimada dos tempos de vida de *Micrurus corallinus* pela distribuição log-logística entre grupos

Segundo Jackson (2016), o *software* R fixa a primeira categoria grupo I como categoria de referência, ou seja, os valores de $\widehat{\beta_2}$ e $\widehat{\beta_3}$ são obtidos quando comparados com o primeiro grupo. Sendo assim, considerando as hípoteses $H_0 : \beta_2 = 0$ e $H_0 : \beta_3 = 0$ como em (2.36) e os $p - valores$ contidos na tabela, rejeita-se H_0 em ambos os casos, ou seja, os grupos 2 e 3 aparentemente influenciam o grupo I indicando que o tempo de sobrevivência aumentou do grupo I para o grupo II e III, como mostra o gráfico 5.6 e, a

esse aumento atribui-se a mudança do manejo adotado do grupo I para os demais grupos. Também observa-se no gráfico, o aumento do tempo de sobrevivência entre o grupo II e o grupo III e, atribui-se à esse aumento, a troca de substrato (*Sphagnum* para a casca de árvore).

6 CONCLUSÃO

Após a aplicação dos modelos mais importantes utilizados em análise de sobrevivência, e seus respectivos modelos de regressão na modelagem de dados de *Micrurus corallinus* viu-se que o modelo de regressão que melhor se adequou a esses dados foi o modelo log-logístico. Foi visto que o grupo II (transição) e III (pós mudança) apresentaram um aumento no tempo de sobrevivência em relação ao grupo I (pré mudança), ou seja, o tempo de vida das serpentes da éspecie *Micrurus corallinus* aumentou após a mudança de manejo, como o esperado. Também é válido afirmar que a troca do substrato *Sphagnum* (grupos I e II) para cascas de árvore (grupo III) aumentou o tempo de vida das serpentes do grupo II para o grupo III, isto é, ao olhar do pesquisador a troca de substrato por um mais em conta, seria de fácil aplicação e adaptação.

As distribuições que estão sendo estudadas por diversos autores, assim como as vistas nesse trabalho, são de grande aplicabilidade por se adequarem melhor aos dados por conta de uma maior flexibilidade. Aos dados de *Micrurus corallinus*, no entanto, a distribuição OLLG-W não é adequada, mesmo se encaixando nos diversos fatores que esse conjunto de dados exigiu (função taxa de falha unimodal). E para o objetivo proposto, a distribuição log-logística, além de ter se mostrado através dos avaliadores de qualidade a melhor distribuição, já está implementada no *software* R, facilitando outros ajustes com dados reais.

As distribuições abordadas por neste trabalho também podem ser aplicadas em ciências agrárias. Para pesquisas futuras, pode-se aplicar os modelos de regressão abordados nesse trabalho em dados de engenharia florestal, analisando o tempo de vida de determinadas espécies sob diferentes condições e manejos.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Akaike, H. *A new look at the statistical model identification*. IEEE transactions on automatic control, v. 19, n. 6, p. 716-232, 1975.

[2] Alzaatreh, A.; Lee, C.; Famoye, F. *A new method for generating families of continuous distributions*. Metron, v.71, n.1, p. 63-79, 2013.

[3] Barlow, R. E.; Campo, R. A. *Total time on test processes and applications to failure data analysis*. Society for Industrial and Applied Mathematics, v. 8, p. 451-481, 1975.

[4] Borges, A. I. M. *Análise de sobrevivência com o R*. 78p. Tese (Mestrado em Matemática) - Universidade de Madeira, Funchal, 2014.

[5] Colosimo, E. A.; Giolo, S. R. *Análise de sobrevivência aplicada*. 1 ed. - São Paulo: Edgard Blücher, 2006.

[6] Cordeiro, G.M.; Alizadeh, M.; Ozel, G.; Hosseini, B.; Ortega, E.M.M.; Altun, E. *The generalized odd log-logistic family of distributions: properties, regression models and applications*. Journal of Statistical Computation and Simulation, v. 87, n. 5, p. 908 - 932, 2017.

[7] Cox, D. R.; Hinkley, D. V. *Theoretical statistics*. 1 ed. - Londres: Chapman and Hall, 1974.

[8] Fachini, J. B. *Análise de influência local nos modelos de riscos múltiplos*. 78p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luis de Queiroz", Universidade de São Paulo, Piracicaba, 2006.

[9] Garcia, P. N. A. *Aplicação de técnicas de sobrevivência em pacientes submetidos à intervenção coronária pericutânea*. 64p. Tese (Bacharel em Estatística). Universidade de Brasília, Brasília, 2013.

[10] Gleaton, J.U.; Lynch, J.D. *Properties of generalized log-logistic families of lifetime distributions*. Journal of Probability and Statistical Science, v. 4, n. 1, p. 51-64, 2006.

[11] Jackson, C. H. *flexsurv: A platform for parametric survival modeling in R*. Journal of statistical software, v. 70, n. 80, 2016.

[12] Kaplan, E. L.; Meier, P. *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, v. 53, n. 282, p. 457-481, 1958,

[13] Lawless, J. F. *Statistical models and methods for lifetime data*. 1 ed. - Nova York: Wiley and Sons, 1982.

[14] Lawless, J. F.; Singhal, K. *Analysis of data from life-test experiments under an exponential model*. Naval research logistics, v. 27, n. 2, p. 323-334, 1980.

[15] Marques, O. A. V. *História natural de Micrurus corallinus (Serpentes, Elapidae)*. 80p. Dissertação (Mestrado em Ecologia) - Instituto de Biociências, Universidade de São Paulo, São Paulo, 1992.

[16] Marques, O. A.; Sazima, I. *Diet and feeding behavior of the coral snake, Micrurus corallinus, from the Atlantic forest of Brazil*. Herpetological Natural History, v. 5, p. 88-93, 1997.

[17] Mendes, G. F.; Stuginski, D. R.; Sant'Anna, S. S.; Loibel, S.M.C.; Grego, K. F. *Factors that can influence the survival rates of coral snakes (Micrurus corallinus) for antivenom production*. J Anim Sci, v. 97, n. 2 , p. 972-980, 2019.

[18] Ortega, E.M.M.; Cordeiro, G.M.; Hashimoto, E.M.; Cooray, K. *A log-linear regression model for the odd Weibull distribution with censored data*. Journal of Applied Statistics, v. 41, n. 9, p. 1859-1880, 2014.

[19] Ortega, E.M.M.; Cordeiro, G.M.; Kattan, M.W. *The log-beta Weibull regression model with application to predict recurrence of prostate cancer*. Statistical Papers, v. 54, n. 1, p. 113-132, 2013.

[20] Pratavieira, F. *O modelo de regressão odd log-logística gama generalizada com aplicações em análise de sobrevivência*. 91p. Tese (Mestrado em Estatística e Experimentação Agronômica). Escola Superior de Agricultura "Luis de Queiroz", Universidade de São Paulo, Piracicaba, 2017.

[21] Prentice, R.R. *A log gamma model and its maximum likelihood estimation*. Biometrika, v. 61, p. 539-544, 1974.

[22] Prentice, R. L.; Kalbfleisch, J. D.; Peterson, A. V. ; Flournoy, N.; Farewell, V. T.; Breslow, N. E. *The Analysis of Failure Times in the Presence of Competing Risks*. Biometrics, v. 34, n. 4, p. 541-554, 1977.

[23] R Development Core Team. *R*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-90051-07-0, 2014.

[24] Santos, D. F. *Modelo de Regressãoo log-logístico discreto com fração de cura para dados de sobrevivência*. 99p. Tese (Mestrado em Estatística). Universidade de Brasília, Brasília, 2017.

[25] Schwarz, G. *Estimating the dimension of model* . The annals of statistics, v. 6, n. 2, p. 461-464, 1978.

[26] Serapicos, E. O.; Merusse, J. L. B; *Variação de peso e sobrevida de Micrurus corallinus sob diferentes condições de alimentação em biotério (Serpentes: Elapidae)*. Série Zoologia, v. 92, n. 4, p. 105-109, 2002.

[27] Silva, P.V.; Loibel, S. M. C.; Mendes, G. F.; Grego, K. F. *Comparação de modelos de sobrevivência para avaliação do manejo de serpentes da espécie Micrurus corallinus em cativeiro* In: Simpósio Nacional de Probabilidade e Estatística (SINAPE), 22., Porto Alegre, anais, 239 p, 2016.

[28] Stacy, E.W. *A generalization of gamma distribution*. The Annals of Mathematical Statistics, v. 33, n. 3, p. 1187-1192, 1962.

[29] Valença, D.M. *O modelo de regressão gama generalizada para discriminar entre modelos paramétricos e tempo de vida*. 148p. Tese (Mestrado em Estatística). Universidade Estadual de Campinas, Campinas, 1994.

[30] Weibull, W. *A statistical theory of the strength of materials*. Ingeniors Vatenskaps Akademien Handlingar, n. 151, p. 293-297, 1939.