Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

Uma modelagem estatística aplicada ao controle biológico da
praga que ataca a cultura do algodão

Abraão de Paula Taveira

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

Piracicaba
2017

Abraão de Paula Taveira
Bacharel em Estatística

**Uma modelagem estatística aplicada ao controle biológico da praga que ataca
a cultura do algodão**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:

Prof. Dr. **CRISTIAN MARCELO VILLEGAS LOBOS**

Dissertação apresentada para obtenção do título de Mestre
em Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

**Piracicaba
2017**

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

Taveira, Abraão de Paula

Uma modelagem estatística aplicada ao controle biológico da praga que ataca a cultura do algodão/ Abraão de Paula Taveira – – versão revisada de acordo com a resolução CoPGr 6018 de 2011, – – Piracicaba, 2017.
71 p.

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

1. Modelos lineares generalizados 2. Análise de sobrevivência 3. Seleção de modelos 4. Análise de resíduos I. Título.

DEDICATÓRIA

Aos meus pais, Jeruza de Paula Taveira e José Carlos Cardoso Taveira, pessoas de suma importância para minha vida.

AGRADECIMENTOS

Agradeço a Deus, por sempre me fortalecer e está à frente de minhas vitórias;

Aos meus pais, José Carlos Cardoso Taveira e Jeruza de Paula Taveira, as minhas irmãs, Lidiane de Paula Taveira e Ana Lúcia de Paula Taveira, ao meu padrinho Otávio Cardoso Taveira e ao meu tio Adalberto Cardoso Taveira, por todo apoio e compreensão, nessa fase da minha vida;

Ao meu orientador Prof. Dr. Cristian Marcelo Villegas Lobos, pela disposição e competência na condução deste trabalho;

Aos professores do departamento de Matemática e Estatística da ESALQ-USP, pelos ensinamentos;

À minha namorada Poliana Belo Barbosa, pelo apoio e incentivo para a conclusão deste trabalho;

À minha turma de mestrado, Leomir Ferreira Sousa, Raquel Aline de Oliveira, Fábio Prata Vieira. Assim como também aos amigos, Erasnilson Vieira, Rick Anderson e Douglas Toledo.

Aos alunos Francisco Sales Fernandes, Victor Wilson Botteon e Ana Paula Borges Battel, pelo fornecimento do conjunto de dados;

SUMÁRIO

RESUMO	6
ABSTRACT	7
LISTA DE FIGURAS	8
LISTA DE TABELAS	10
1 INTRODUÇÃO	11
2 OBJETIVOS	15
2.1 Objetivo Geral	15
2.2 Objetivos Específicos	15
3 REVISÃO DE LITERATURA	17
3.1 Introdução	17
3.2 Definição de um modelo linear generalizado	18
3.2.1 Modelo Gama	19
3.2.2 Estimação dos parâmetros	22
3.2.3 Técnicas de diagnóstico	25
3.2.3.1 Técnicas gráficas	27
3.2.3.2 Avaliação da função de ligação	29
3.3 Análise de sobrevivência	29
3.3.1 Modelos Probabilísticos	32
4 MATERIAL E MÉTODOS	41
4.1 Material	41
4.2 Métodos aplicados ao conjunto de dados que não contém censuras	44
4.3 Métodos aplicados ao conjunto de dados que contém censuras	47
5 RESULTADOS E DISCUSSÃO	49
5.1 Resultados referentes ao conjunto de dados que não contém censuras	49
5.2 Análise incluindo as observações censuradas	54
6 CONCLUSÕES	59
REFERÊNCIAS	61

RESUMO

Uma modelagem estatística aplicada ao controle biológico da praga que ataca a cultura do algodão

As distribuições de probabilidade gama, normal inversa, Weibull, log-normal e exponencial são uma boa alternativa para modelar observações associadas ao tempo, pois, em geral, a variável tempo possui assimetria à esquerda ou à direita, o que caracteriza as distribuições citadas anteriormente. O objetivo deste trabalho constitui-se em avaliar o comportamento dos predadores, *Euborellia annulipes* (“Tesourinha”) e *Harmonia axyridis* (“Joaninha”), em relação à praga conhecida como *Aphis gossypii* (“Pulgão”). Outra pretensão deste trabalho é a aplicação da modelagem estatística, dando ênfase as técnicas dos modelos lineares generalizados e análise de sobrevivência, as quais foram aplicadas aos dados provenientes de um experimento, instalado no Laboratório de Ecologia de Insetos da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ). O experimento foi composto por 21 repetições, sendo cada repetição efetuada por meio de uma placa de Petri medido $60 \times 15 \text{ mm}$. Em cada placa foi liberado um pulgão adulto áptero na parte central, tendo três pesquisadores responsáveis por observar a variável definida como tempo de ataque. Inicialmente, foram ajustados os modelos com distribuição gama e diferentes funções de ligação, e o modelo com a distribuição normal inversa com função de ligação canônica. Esses modelos foram ajustados aos dados desconsiderando as censuras, em que por meio do gráfico *half-normal plot* e testes de hipóteses, verificou que o modelo com a distribuição normal inversa com função de ligação canônica, apresentou o melhor ajuste. Posteriormente, foram ajustados os modelos exponencial, Weibull e log-normal para os dados considerando as censuras, os quais foram avaliados mediante o teste de razão de verossimilhança, sendo o modelo log-normal mais apropriado aos dados.

Palavras-chave: Modelos lineares generalizados; Análise de sobrevivência; Seleção de modelos; Análise de resíduos

ABSTRACT

An statistical model applied to the biological control of the pest that attacks the cotton crop

The probability density function of gamma, inverse normal, Weibull, log-normal and exponential distributions are good alternatives for modelling observations related with time, since, in general, the time variable has left or right asymmetry, which characterizes the distributions previously mentioned . The aim of this work is the application of statistical modeling, emphasizing the techniques of generalized linear models and survival analysis, which were applied to data from an experiment, installed in the Laboratory of Insect Ecology of the “Luiz de Queiroz” College of Agriculture (ESALQ), in which the goal of this experiment was to evaluate the behavior of predators, *Euborellia annulipes* (“ring-legged earwig”) and *Harmonia axyridis* (“Ladybird”), in relation to the pest known as *Aphis*. The experiment was composed of 21 replicates, each replicate being done by means of a petri dish measured 60×15 mm. On each plate an adult aphid was released in the central part, with three researchers responsible. The model with distribution was used to determine the variance, which was defined as the attack time. Normal distribution with canonical link function. These models were adjusted to the data disregarding censorship, in which through the half-normal plot and hypothesis tests, verified that the model with the normal inverse distribution with canonical link function, presented the best fit. Subsequently, the exponential, Weibull and log-normal models were adjusted for the data considering the censorship, which were evaluated by the likelihood ratio test, the log-normal model being more appropriate to the data.

Keywords: Generalized linear models; Survival analysis; Models selection; Residual analysis

LISTA DE FIGURAS

Figura 1 - Função densidade de probabilidade da distribuição gama para diferentes situações: (a) μ variando e $v = 1$, (b) $\mu = 2$ e $0 < v < 1$, (c) $\mu = 0.5$ e $v > 1$, (d) $v = k/2$ e $\mu = k$	21
Figura 2 - Relação entre a distribuição gama e algumas distribuições de probabilidades.	22
Figura 3 - Representação gráfica para os tipos de censuras citadas anteriormente. . .	31
Figura 4 - Funções de densidade de probabilidade $f(t)$ (a) , de sobrevivência $S(t)$ (b) e de taxa de falha (c) da distribuição exponencial para $\alpha = 1$ (cor azul), $\alpha = 0.7$ (cor vermelha) e $\alpha = 0.5$ (cor verde).	33
Figura 5 - Funções de densidade de probabilidade $f(t)$ (a) , de sobrevivência $S(t)$ (b) e de taxa de falha (c) da distribuição de Weibull para alguns valores dos parâmetros (γ, α) , cor azul (3, 250), cor vermelha (4, 350), cor verde (8, 600), cor cinza (1, 150) e cor preta (0.5, 50).	35
Figura 6 - Funções de densidade de probabilidade $f(t)$ (a) , de sobrevivência $S(t)$ (b) e de taxa de falha (c) da distribuição log-normal para alguns valores dos parâmetros μ, α , cor azul (0, 0.5), cor vermelha (0, 0.7), cor verde (0, 1.5), cor cinza (1, 0.7) e cor preta (1, 2).	36
Figura 7 - (A) Espécie <i>Harmonia axyridis</i> (“Joaninha Fêmea”) e (B) (“Joaninha Macho”).	42
Figura 8 - (A) Espécie <i>Euborellia annulipes</i> (“Tesourinha Fêmea”) e (B) (“Tesourinha Macho”).	42
Figura 9 - Imagem referente à praga da espécie <i>Aphis gossypii</i> (“Pulgão”), sexo fêmea.	43
Figura 10 - Imagem referente as placas de Petri utilizada para realizar o experimento.	44
Figura 11 - <i>Boxplot</i> referente ao tempo de ataque em minutos dos predadores em relação à praga nos períodos, (a) diurno e (b) noturno. Legenda: Tesourinha Fêmea (TF), Tesourinha Macho (TM), Joaninha Fêmea (JF) e Joaninha Macho (JM).	50
Figura 12 - <i>Half-normal plot</i> para os modelos: distribuição gama com função de ligação identidade (a), distribuição gama com função de ligação logarítmica (b), distribuição gama com função de ligação inversa (c), distribuição normal inversa com função de ligação canônica (d).	51

Figura 13 - <i>Half-normal plot</i> para o modelo normal inversa (a) e gráfico dos componentes do desvio versus os valores ajustados (b).	53
Figura 14 - Estimativas de Kaplan-Meier para curvas de sobrevivência para os predadores tesourinha e joaninha, ambos os sexos no período diurno (a) e no período noturno (b).	55
Figura 15 - Análise gráfica dos resíduos de Cox-Snell para os modelos exponencial, Weibull e log-normal.	57

LISTA DE TABELAS

Tabela 1 - Principais distribuições pertencentes à família exponencial.	19
Tabela 2 - Medidas descritivas do tempo de ataque (em minutos) dos predadores Tesourinha e Joaquina em relação à praga Pulgão nos períodos, diurno e noturno.	49
Tabela 3 - Teste F para o modelos ajustados com todas as combinações possíveis para o preditor linear, utilizando a distribuição normal inversa e função de ligação canônica.	52
Tabela 4 - Estimativas e erros padrão dos parâmetros do modelo normal inversa com função de ligação canônica.	53
Tabela 5 - Logaritmo da função $L(\theta)$ e os resultados dos TRV para os modelos com diferentes distribuições.	57
Tabela 6 - Tabela para estimativas dos parâmetros, erro padrão e p -valor.	58

1 INTRODUÇÃO

Sendo cultivado há mais de 4.000 anos, o algodão surgiu no sul da Arábia e atualmente está, entre as mais importantes culturas de fibras do mundo. Todos os anos, uma média de 35 milhões de hectares de algodão é plantada por todo o planeta, tendo sua demanda mundial, aumentado gradativamente desde a década de 1950, ocorrendo um crescimento anual médio de 2% (ASSOCIAÇÃO MATOGROSSENSE DOS PRODUTORES DE ALGODÃO - AMPA, 2016). O comércio mundial do algodão movimenta anualmente, cerca de U\$\$ 12 bilhões e envolve mais de 350 milhões de pessoas em sua produção, desde as fazendas até o processo final, à embalagem (ASSOCIAÇÃO BRASILEIRA DOS PRODUTORES DE ALGODÃO - ABRAPA, 2016). Em uma lista de 60 países, China, Índia, Estados Unidos, Paquistão e Brasil, estão como os principais produtores de algodão.

No Brasil, a cultura do algodão é de suma importância para expressão sócio-econômica dos setores primário e secundário. Entretanto, as pragas constituem-se um dos fatores limitantes para sua exploração, caso não sejam tomadas medidas eficientes de controle. Entre vários tipos diferentes de pragas que atacam a cultura do algodão, a espécie *Aphis gossypii*, conhecida popularmente como pulgão é uma das pragas que mais acarreta danos para produtividade do algodão. Sendo um inseto de tamanho pequeno, medindo cerca de 1,3 mm de comprimento, apresenta coloração variável do amarelo-claro ao verde escuro, quando está na fase adulta predomina a cor verde escuro (ALVARENGA, 1992). É uma das primeiras pragas a aparecer na cultura do algodão, localizam-se na parte inferior das folhas e brotos, preferencialmente nas do ponteiro, sugando a seiva, provocando a curvatura do limbo foliar para baixo, ocasionando a paralisação temporária do crescimento das plantas, causando uma redução na ordem de 24% no peso do algodão em caroço e atraso na maturação (AYRES; SANTOS, 2007).

Então, com o objetivo de combater esse tipo de praga, tradicionalmente é utilizada aplicação de agrotóxicos, porém, essa forma de combate apresenta algumas consequências consideradas graves, tais como, danos para à saúde do agricultor e influência na qualidade da produção do algodão, além disso, fazem com que as pragas adquiram resistência, exigindo doses mais altas ou produtos mais tóxicos. Como uma forma alternativa de controle de pragas, estudos têm avaliado a eficiência predatória de inimigos naturais, onde esse procedimento é conhecido como Manejo Integrado de Pragas – MIP e tem como

prioridade, reduzir a exposição dos produtores e técnicos aos pesticidas, eliminar o aparecimento de casos de intoxicação por agrotóxicos (no Brasil é um dos principais problemas de saúde pública), reduz a eliminação de organismos benéficos presentes nas lavouras, ou seja, torna o agroecossistema algodoeiro ecologicamente viável (ALMEIDA; DOMINGUES; RAMALHO, 2008). Nesse cenário, a espécie *Euborellia annulipes* (“Tesourinha”) e a espécie *Harmonia axyridis* (“Joaninha”), apresentam-se como umas das principais espécies predadores do pulgão. Com isto, foi instalado no Laboratório de Ecologia de Insetos da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ), um experimento que teve como objetivo, comparar o tempo de ataque dos predadores *Euborellia annulipes* (“Tesourinha”) e *Harmonia axyridis* (“Joaninha”), por sexo (Macho e Fêmea), em relação à praga *Aphis gossypii* (“Pulgão”), nos períodos diurno e noturno.

Dentre as técnicas analíticas pertinentes para estudos enfatizando pragas agrícolas, destaca-se a modelagem estatística. Os modelos normais têm sido aplicados em diversas áreas do conhecimento, como por exemplo, medicina, agricultura, ecologia, demografia, economia, engenharia, geologia, zootecnia, entre outros (CORDEIRO; DEMÉTRIO, 2007). Esses são utilizados com o objetivo de obter um modelo, que seja o mais simples possível e que descreva bem os dados em estudo. A seleção de modelos é uma das partes mais importantes de toda pesquisa (CORDEIRO; DEMÉTRIO; MORAL, 2014). Sendo utilizados para descrever fenômenos aleatórios, os modelos normais lineares supõem que os resíduos seguem uma distribuição normal, mas quando esta suposição não é satisfeita, algum tipo de transformação pode ser efetuada, sendo a mais conhecida idealizada por Box e Cox (1964).

Tendo como finalidade ampliar o domínio da aplicabilidade dos modelos normais lineares Nelder e Wedderburn (1972), desenvolveram a teoria dos modelos lineares generalizados, integrando modelos cuja variável resposta engloba a família exponencial de distribuições. Podem ser citados como exemplos, os modelos de análise de variância, modelos log-lineares, modelos gama, Weibull para dados contínuos, modelos de Poisson para dados de contagem, dentre outros. As distribuições de probabilidade gama, normal inversa e log-normal são uma boa alternativa para modelar observações associadas ao tempo, pois, em geral a variável tempo possui assimetria a esquerda ou a direita, o que caracteriza as distribuições citadas anteriormente.

Em diversos estudos é comum utilizar-se como variável resposta, dados contínuos associados ao tempo, sejam eles em teste de germinação, avaliação entomológica,

tempos de sobrevivência, entre outros (CALLEGARI-JACQUES, 2009). A aplicação de modelos que não se ajustam a esse tipo de dados pode levar a interpretações equivocadas acerca dos resultados obtidos. Partindo deste princípio, diversas técnicas de modelagem de dados têm sido utilizadas para melhor explicar o real comportamento da variável resposta. Dessa forma, o presente trabalho visa aplicar a teoria dos modelos lineares generalizados para o conjunto de dados que não leva em conta as censuras, bem como a teoria de análise de sobrevivência para o conjunto de dados que contém as censuras, provenientes do experimento citado anteriormente. Todas as análises serão efetuadas por meio do *software* estatístico *R* (R Core Team, 2013). A seguir são descritos os principais objetivos deste trabalho.

2 OBJETIVOS

2.1 Objetivo Geral

O objetivo do presente trabalho é aplicar a teoria dos modelos lineares generalizados e a teoria da análise de sobrevivência para determinar se há diferenças entre o tempo de ataque dos predadores tesourinha e joaninha por sexo, entre os turnos e por inseto, bem como encontrar uma metodologia adequada para resolver uma situação prática.

2.2 Objetivos Específicos

- i) Comparar modelos que possuem suporte nos reais positivos;
- ii) Selecionar o modelo mais adequado ao conjunto de dados em estudo;
- iii) Verificar as suposições do modelo utilizando análise de resíduos.

3 REVISÃO DE LITERATURA

Nesta secção será descrito uma breve revisão sobre os métodos utilizados para conclusão desse trabalho, ou seja, apresenta-se uma fundamentação teórica sobre as técnicas modelos lineares generalizados e análise de sobrevivência.

3.1 Introdução

No decorrer dos anos, pesquisadores identificaram que as pragas efetuam grandes prejuízos ao cenário agroindustrial brasileiro, em consequência desse fato o uso de inseticidas foi cada vez mais frequente na produção agroindustrial brasileira. Com isso, ao longo dos anos várias pesquisas foram desenvolvidas com o objetivo de encontrar uma alternativa de combate a essas pragas, sendo a mais indicada o uso de inimigos naturais como uma forma alternativa de combate.

Diante deste contexto, estudos aprofundados foram intensificados com o intuito de identificar insetos que apresentam grande potencial como predador. Dentre vários predadores estudados, espécies como a tesourinha *Euborellia annulipes* e a joaninha *Harmonia axyridis* deram origem a diversos experimentos, como por exemplo o experimento realizado pelo pesquisador Cristofolletti (2014), o qual desenvolveu o estudo com o objetivo de avaliar a preferência dos predadores *Euborellia annulipes* e a joaninha *Harmonia axyridis* em relação as pragas *Spodoptera frugiperda* para o milho, *Anticarsia gemmatilis* para a soja e *Diatraea saccharalis* na cana de açúcar. Os resultados provenientes desse estudo foram analisados por meio de tabelas de vida, em que, não foram observadas claras preferências por presas. Também pode-se ser citado como exemplo o trabalho escrito por OLIVEIRA. et al. (2010), que teve como objetivo estudar o comportamento do predador, *Euborellia annulipes* em relação ao pulgão, *Apihs gossipii*. Os resultados oriundos desse experimento foram analisados por meio da técnica análise de variância, que se baseia na suposição de normalidade.

Em estudos voltados para área entomológica, são comuns observações de números positivos na escala contínua, como por exemplo o tempo, caracterizando dados assimétricos (PEREIRA, 2009). Dentre as distribuições usuais para modelagem de dados assimétricos, têm-se as distribuições Weibull, gama, log-normal, normal inversa, exponencial, entre outras. Porém, existem análises em que os pesquisadores optam por realizar

algum tipo de transformação aos dados a fim de obter a normalidade e, posteriormente ser aplicado a teoria dos modelos normais lineares, inibindo o uso das distribuições citadas.

Nesse contexto, os modelos lineares generalizados podem ser vistos, como uma classe mais ampla que permite uma generalização dos modelos lineares clássicos e uma extensão da distribuição da variável resposta, que à mesma pertença à família exponencial de distribuições. Com o objetivo de apresentar à temática e explorar os meios pelos quais é possível obter um maior conhecimento a respeito dos modelos lineares generalizados, atualmente dispõe-se de uma ampla literatura, sendo considerado como a principal referência sobre o assunto, o livro de McCullagh e Nelder (1989) que apresenta uma vasta discussão sobre o tema, expondo aplicações práticas e uma série de demonstrações matemáticas. Temas como aplicações e uso de modelos de regressão múltipla, análise de variância para dados contínuos, modelos log-lineares para dados de contagem na forma de tabelas de contingência, entre outros, são descritos em Dobson (2010). Na língua portuguesa, Cordeiro (1986), Cordeiro, Demétrio e Moral (2014), Paula (2013), são referências que apresentam de forma essencial uma introdução ao tema, dispondo aplicações em conjunto de dados de diversas áreas do conhecimento, como medicina, pesca, odontologia, agronomia, biologia e economia.

3.2 Definição de um modelo linear generalizado

Segundo Agresti (2002), os modelos lineares generalizados possuem três características que estão descritas a seguir:

- I) Um componente aleatório, que fundamenta-se em um vetor de observações independentes $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, em que cada Y_i pertence à família exponencial, cuja função densidade de probabilidade é definida por

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{[y_i \theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi) \right\}, \quad i = 1, \dots, n, \quad (1)$$

em que $a(\cdot)$, $b(\cdot)$ e $c(\cdot)$ são funções conhecidas; $\phi > 0$ é denominado parâmetro de dispersão e θ_i é denominado parâmetro canônico que caracteriza a distribuição em (1). Se ϕ for conhecido, a equação (1) representa a família exponencial uniparamétrica indexada por θ_i (CORDEIRO; NETO, 2004), cuja notação é $Y_i \sim FE(\mu_i, \phi)$, então apresenta as seguintes propriedades:

- $E(Y_i) = \mu_i = b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$;
- $V(Y_i) = \phi V_i(\mu_i)$ em que $V_i(\mu_i) = \frac{d\mu}{d\theta} = b''(\theta_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta^2}$ tendo, ϕ é o parâmetro de dispersão. Ver mais detalhes em Cordeiro, Demétrio e Moral (2014). Na Tabela 1, estão descritas algumas distribuições pertencentes à família exponencial.

Tabela 1 - Principais distribuições pertencentes à família exponencial.

Distribuição	ϕ	θ	$b(\theta)$	$c(y, \phi)$
Normal: $N(\mu, \sigma^2)$	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2} \left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$
Poisson: $P(\mu)$	1	$\log(\mu)$	e^θ	$-\log y!$
Binomial: $B(m, \pi)$	1	$\log\left(\frac{\mu}{m-\mu}\right)$	$m \log(1 + e^\theta)$	$\log\binom{m}{y}$
Gama: $G(\mu, v)$	v^{-1}	$-\frac{1}{\mu}$	$-\log(-\theta)$	$v \log(vy) - \log(y) - \log \Gamma(v)$
Normal Inversa: $NI(\mu, \sigma^2)$	σ^2	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\frac{1}{2} \left[\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y} \right]$

Fonte: Paula (2013)

II) Um preditor linear $\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} = \beta_1 X_{i1} + \dots + \beta_p X_{ip} = \sum_{j=1}^p \beta_j X_{ij}$, em que $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ é o vetor de covariáveis associado ao i -ésimo indivíduo, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ é o vetor de parâmetros de dimensão $p \times 1$;

III) O terceiro componente dos modelos lineares generalizados é uma função de ligação que relaciona uma função da média da variável resposta e o preditor linear, isto é,

$$g(\mu_i) = \eta_i,$$

com $\mu_i = E(Y_i)$. A função de ligação $g(\cdot)$ é uma função monótona e diferenciável.

Dentre das distribuições apresentadas na Tabela 1, tem-se na próxima secção uma descrição sobre o modelo gama, o qual será aplicado neste trabalho.

3.2.1 Modelo Gama

A distribuição gama é uma das mais gerais distribuições, pois diversas distribuições são casos particulares dela, por exemplo, a distribuição exponencial, distribuição qui-quadrado, distribuição Weibull, entre outras (veja Figura 2). Essa distribuição é usada em estudos voltados para dados contínuos não negativos. Uma variável aleatória Y é dita ter

distribuição gama de média μ e coeficiente de variação $\phi^{-1/2}$, quando sua função densidade é dada por

$$f(y; \mu, v) = \frac{1}{\Gamma(v)} \left(\frac{vy}{\mu}\right)^v \exp\left(-\frac{vy}{\mu}\right) \frac{1}{y}, \quad \text{em que } v > 0 \text{ e } \mu > 0.$$

Com base na Eq. (1) mostra-se que, a distribuição gama pertencente à família exponencial, isto é

$$f(y; \mu, v) = \exp\left[v\left\{-\frac{y}{\mu} + \log\left(\frac{1}{\mu}\right)\right\} - \log \Gamma(v) + v \log(vy) - \log y\right],$$

$y > 0$, $\mu > 0$, $v > 0$, em que $\Gamma(v) = \int_0^\infty t^{v-1} e^{-t} dt$ é a função gama. Logo, fazendo $\phi = \frac{1}{v}$, $\theta = -1/\mu$, $b(\theta) = -\log(-\theta)$ e $c(y, v) = v \log(vy) - \log(y) - \log \Gamma(v)$ tem-se que a distribuição gama pertence à família exponencial de distribuições, com $E(Y) = b'(\theta) = \mu$ e $V(Y) = \phi^{-1}b''(\theta) = \frac{\mu^2}{v}$. Para $0 < v < 1$ a densidade gama decresce monotonicamente quando $y \rightarrow \infty$ (Figura 1(b)). A distribuição exponencial é um caso especial quando $v = 1$ (Figura 1(a)). Para $v > 1$ a densidade tem um máximo em $y = \mu - \mu/v$ e depois decresce para $y \rightarrow \infty$ (Figura 1(c)). A distribuição χ_k^2 é um outro caso especial quando $v = k/2$ e $\mu = k$ (Figura 1(d)). A distribuição normal é obtida fazendo $v \rightarrow \infty$. Isto é, quando v é grande $Y \sim N(\mu, v^{-1}V(\mu))$. Sendo que $\phi = (E(Y)/Var(Y))^2$ é o inverso do coeficiente de variação de Y ao quadrado ($\phi = 1/(CV)^2$). A função de variância da distribuição gama é dada por $V(\mu) = \mu^2$ (PAULA, 2013). Por meio da Figura 1, apresenta-se alguns gráficos para exemplificar os casos citados anteriormente.

Figura 1 - Função densidade de probabilidade da distribuição gama para diferentes situações: (a) μ variando e $v = 1$, (b) $\mu = 2$ e $0 < v < 1$, (c) $\mu = 0.5$ e $v > 1$, (d) $v = k/2$ e $\mu = k$.

A seguir é apresentado um diagrama, constando algumas distribuições que a partir de casos especiais relacionam-se com a distribuição gama (LEEMIS; MCQUESTON, 2008).

Figura 2 - Relação entre a distribuição gama e algumas distribuições de probabilidades.

3.2.2 Estimação dos parâmetros

Uma das formas de realiza-se a estimação dos parâmetros lineares β' s na teoria dos modelos lineares generalizados é com a aplicação do método de máxima verossimilhança, uma vez que, esse método dispõe propriedades como consistência e eficiência assintótica. Segundo Collett (2002), tendo uma amostra aleatória $\mathbf{y} = (y_1, y_2, \dots, y_n)$ com n observações de uma distribuição pertencente à família exponencial (1), a função de verossimilhança é dada por:

$$L = L(\boldsymbol{\beta}, \phi; \mathbf{y}) = \prod_{i=1}^n f(y_i; \boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T, \phi) = \exp \left\{ \sum_{i=1}^n [\phi^{-1} [y_i \theta_i - b(\theta_i)] + c(y_i; \phi)] \right\},$$

aplicando-se o logaritmo na função de verossimilhança tem-se

$$l = l(\boldsymbol{\beta}; \phi, \mathbf{y}) = \log L(\boldsymbol{\beta}; \phi, \mathbf{y}) = \sum_{i=1}^n \{ \phi^{-1} [y_i \beta_i - b(\beta_i)] + c(y_i; \phi) \}.$$

Derivando a expressão anterior, obtêm-se o vetor $\boldsymbol{\beta}$ de parâmetros que maximizam $l(\boldsymbol{\beta}; \phi, \mathbf{y})$, utiliza-se regra da cadeia, conforme apresentado abaixo.

$$U_j = \frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \phi^{-1} \sum_{i=1}^n (y_i - \mu_i) \frac{1}{V_i} \frac{d\mu_i}{d\eta_i} x_{ij}, \quad (2)$$

igualando-se U_j a zero para $j = 1, \dots, p$ calcula-se a estimativa de máxima verossimilhança, entretanto, existem casos que as equações são não lineares, então para alcançar os devidos resultados é preciso recorrer aos processos iterativos, (Newton-Raphson ou método escore de Fisher).

De acordo com Cordeiro, Demétrio e Moral (2014), o método iterativo de

Newton-Raphson é baseado na aproximação de Taylor de segunda ordem e procede-se da seguinte forma

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + [\mathbf{K}^{(m)}]^{-1} \mathbf{U}^{(m)},$$

sendo $\boldsymbol{\beta}^{(m+1)}$, $\boldsymbol{\beta}^{(m)}$ os vetores de parâmetros estimados nos passos m e $(m+1)$. Tendo $\mathbf{U}^{(m)}$ o vetor escore, com elementos $\left(\frac{\partial l}{\partial \beta_j}\right)$ e $\mathbf{K}^{(m)}$ em que a matriz de informação esperada com os elementos $\left(-\frac{\partial^2 l}{\partial \beta_j \partial \beta_j}\right)$.

Ao substituir a matriz de informação observada $\mathbf{I}$ pela matriz de informação esperada de Fisher obtêm-se

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)},$$

sendo,

- i) $\mathbf{X}$ é a matriz de planejamento de dimensão $n \times p$ do modelo;
- ii) $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ é uma matriz diagonal $n \times n$ de pesos que capta a informação sobre a distribuição, em que $w_i = V^{-1} (d\mu_i/d\eta_i)^2$ denominada função peso;
- iii) $\mathbf{z}^{(m)} = \mathbf{X}\boldsymbol{\beta}^{(m)} + \Delta^{(m)}(\mathbf{y} - \boldsymbol{\mu})^{(m)} = \boldsymbol{\eta}^{(m)} + \Delta^{(m)}(\mathbf{y} - \boldsymbol{\mu})^{(m)}$ é o vetor da variável dependente ajustada no passo m com $\Delta = \text{diag}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$. No passo inicial pode-se tomar $\hat{\boldsymbol{\beta}}^{(0)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\eta}}$, em que $\hat{\eta}_i = (\hat{\mu}_i) = g(y_i)$.

Vale ressaltar que o método de Newton-Raphson é utilizado quando as derivadas parciais de segunda ordem possuem resolução simples, porém, na teoria dos modelos lineares generalizados isso nem sempre ocorre, tendo que recorrer ao método escore de Fisher que, em geral, é mais simples sendo semelhante ao método de Newton-Raphson no caso das funções de ligações canônicas (CORDEIRO; DEMÉTRIO; MORAL, 2014). Esse método tem como característica substituir a matriz de derivadas parciais de segunda ordem pela matriz de valores esperados das derivadas parciais, ou seja, substitui a matriz de informação observada, $\mathbf{I}$, pela matriz de informação esperada de Fisher, $\mathbf{K}$. Sendo assim, faz-se

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + (\mathbf{K}_0^{-1})^{(r)} \mathbf{U}^{(r)}, \quad (3)$$

em que $\mathbf{K}$ tem elementos dados por

$$\kappa_{t,s} = -E \left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_t \partial \boldsymbol{\beta}_s} \right] = E \left[\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_t} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_s} \right],$$

denominada matriz de covariância dos U_j' s. Como resultado da multiplicação de ambos os membros de (3) por $\mathbf{K}^{(r)}$, tem-se

$$\mathbf{K}^{(r)} \boldsymbol{\beta}^{(r+1)} = \mathbf{K}^{(r)} \boldsymbol{\beta}^{(r)} + \mathbf{U}^{(r)}. \quad (4)$$

Desta forma, usando-se (2), resulta em

$$\begin{aligned} \kappa_{t,s} &= E(U_t U_s) = \sum_{i=1}^n \frac{1}{\phi^2} E(Y_i - \mu_i)^2 \frac{1}{[V(\mu_i)]^2} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{it} x_{is} \\ &= \sum_{i=1}^n \frac{1}{\phi^2} \phi V(\mu_i) \frac{1}{[V(\mu_i)]^2} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{it} x_{is} \\ &= \sum_{i=1}^n \frac{1}{\phi} \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 x_{it} x_{is}. \end{aligned}$$

Portanto, a matriz de informação de Fisher para $\boldsymbol{\beta}$ tem a forma

$$\mathbf{K} = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X},$$

em que $\mathbf{W} = \text{diag}\{w_1, \dots, w_n\}$ uma matriz diagonal de pesos com elementos $w_i = \frac{1}{V(\mu_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2$.

Logo, o vetor escore $\mathbf{U} = \mathbf{U}(\boldsymbol{\beta})$ com componentes em (2) pode ser expresso na forma

$$\mathbf{U} = \phi^{-1} \mathbf{X}^T \mathbf{W} \mathbf{G}(\mathbf{y} - \boldsymbol{\mu}),$$

com $\mathbf{G} = \text{diag}\left(\frac{d\eta_1}{d\mu_1}, \dots, \frac{d\eta_n}{d\mu_n}\right) = \text{diag}[g'(\mu_1), \dots, g'(\mu_n)]$.

Ao substituir $\mathbf{K}$ e $\mathbf{U}$ em (4), tem-se

$$\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X} \boldsymbol{\beta}^{(r+1)} = \mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X} \boldsymbol{\beta}^{(r)} + \mathbf{X}^T \mathbf{W} \mathbf{G}^{(r)} (\mathbf{y} - \boldsymbol{\mu}^{(r)})$$

ou

$$\mathbf{X}^T \mathbf{W}^{(r)} \mathbf{X} \boldsymbol{\beta}^{(r+1)} = \mathbf{X}^T \mathbf{W}^{(r)} [\boldsymbol{\eta}^{(r)} + \mathbf{G}^{(r)} (\mathbf{y} - \boldsymbol{\mu}^{(r)})].$$

Ao definir a variável dependente ajustada $\mathbf{z} = \boldsymbol{\eta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$, tem-se

$$\mathbf{X}^T \mathbf{W}^{(m)} \mathbf{X} \boldsymbol{\beta}^{(m+1)} = \mathbf{X}^T \mathbf{W}^{(m)} \mathbf{z}^{(m)}.$$

O método de estimação apresentado anteriormente pode ser implementado no *software* estatístico *R* (R Core Team, 2013), o qual disponibiliza pacotes como **VGAM**, **MASS**, **stats**, dentre outros. No pacote **stats** encontra-se a função **glm**, em que deve-se especificar a definição do modelo e a distribuição assumida pela variável resposta com a função de ligação a ser usada. Outra função importante que também pertence ao pacote **stats**, é a função **optim**, utilizada quando o estimador não apresenta uma expressão fechada, essa função agrupa seis diferentes métodos de otimização.

3.2.3 Técnicas de diagnóstico

A análise de diagnóstico é considerada uma das etapas mais importantes na análise de um ajuste de modelo, sendo essa análise feita por meio das técnicas de diagnóstico, as quais são classificadas em formais e informais. Para o modelo linear clássico as técnicas formais consistem em testes de hipóteses, em que verifica-se a normalidade dos resíduos, a homocedasticidade e independência dos resíduos.

Na teoria dos modelos lineares generalizados tem-se a análise de resíduos equivalente a aplicada nos modelos clássicos, com apenas algumas adaptações (CORDEIRO; DEMÉTRIO; MORAL, 2014), ou seja, nos modelos clássicos a pressuposição de linearidade é verificada por meio dos vetores $\mathbf{Y}$ e $\hat{\boldsymbol{\mu}}$, mas quando trata-se dos modelos lineares generalizados, essa pressuposição é verificada por meio da variável dependente ajustada $\mathbf{z}$ e o preditor linear $\hat{\boldsymbol{\eta}}$. Os objetivos principais desse procedimento é verificar se há afastamento sérios das suposições feitas para o modelo em questão, ou seja, há afastamento da suposição da distribuição da variável resposta; ausência de alguma variável explicativa ou termos (quadrático, cúbico) de variáveis incluídas no modelo; se há indícios de correlação entre as observações.

Pode-se também detectar observações atípicas que destoam do conjunto de

dados, sendo essas classificadas em três grupos. Pontos aberrantes (*outlier*), quando omissão do conjunto de dados resulta em mudanças substanciais nas estatísticas de diagnósticos do modelo; pontos de alavanca, posicionadas em regiões remotas com alta influência no próprio valor ajustado; pontos influentes desproporcional nas estimativas dos coeficientes.

Segundo McCullagh e Nelder (1989), na teoria dos modelos lineares generalizados a variância residual é substituída por uma estimativa de ϕ e a matriz de projeção $\mathbf{H}$ ou matriz “hat”, define-se por $\mathbf{H} = \mathbf{W}^{\frac{1}{2}} \mathbf{X} (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{\frac{1}{2}}$. Cordeiro, Demétrio e Moral (2014) apresentam que os tipos de resíduos mais comuns são:

i) Resíduos ordinários, pelo fato de entrar na definição de outros resíduos, defini-se por

$r_i = y_i - \hat{\mu}_i$, sendo y_i a variável resposta e $\hat{\mu}_i$ representa sua estimativa correspondente;

ii) Resíduo de Pearson estudentizado internamente, definido como

$$r_i^{p'} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V_i(\hat{\mu}_i) (1 - h_{ii})}},$$

sendo h_{ii} o i -ésimo elemento da diagonal da matriz de projeção $\mathbf{H}$, $\hat{\phi}$ é a estimativa consistente do parâmetro ϕ ;

iii) Resíduo de Pearson generalizado definido da seguinte forma,

$$r_{Pi} = (y_i - \hat{\mu}_i) \sqrt{\frac{w_i}{V_i(\hat{\mu}_i)}},$$

em que $V_i(\hat{\mu})$ é a função de variância e w_i é um peso a priori;

iv) Componentes de desvio é outro tipo de resíduo usado dentro da classe dos MLGs, sendo as raízes quadradas dos componentes dos desvio com o sinal igual de $y_i - \hat{\mu}_i$, ou seja,

$$r_i^D = \text{sinal}(y_i - \hat{\mu}_i) \sqrt{2[v(y_i) - v(\hat{\mu}_i) + q(\hat{\mu}_i)(\hat{\mu}_i - y_i)]^{1/2}}.$$

Quando o resíduo r_i^D representa a distância da observação y_i ao seu valor ajustado $\hat{\mu}_i$, medida na escala do logaritmo da função de verossimilhança, obtêm-se $D_p = \sum_{i=1}^n (r_i^D)^2$. Caso r_i^D apresente um valor alto, implica-se dizer, que a i -ésima observação está mal ajustada pelo modelo. Para esse tipo de resíduo pode-se citar algumas vantagens,

como, não requerem o conhecimento da função normalizadora, computação simples após o ajuste do MLG e são definidos para todas observações.

Vale ressaltar, que existe outros tipos de resíduos não menos importantes que os citados anteriormente, tendo-se todas essas medidas grande importância para a escolha do modelo mais apropriado. Para melhor detalhamento ver (CORDEIRO; DEMÉTRIO; MORAL, 2014; MCCULLAGH; NELDER, 1989; PAULA, 2013).

3.2.3.1 Técnicas gráficas

As técnicas gráficas são classificadas como técnicas informais que auxiliam na análise de resíduos, posteriormente serão apresentados os gráficos mais utilizados para os MLGs.

- i) Por meio do gráfico dos resíduos versus valores ajustados, pode-se verificar a presença de valores discrepantes e heterogeneidade de variância. Esse gráfico apresentará a distribuição dos resíduos em torno de zero com amplitude constante;
- ii) Gráfico dos resíduos versus as covariáveis, mediante visualização desse gráfico é possível identificar existência de uma relação sistemática entre os resíduos e uma covariável. Esse gráfico apresenta uma distribuição aleatória dos resíduos em torno de zero com amplitude constante;
- iii) Gráfico dos resíduos contra a ordem das observações, por meio desse gráfico verifica-se se há alguma variável altamente correlacionada com a sequência do tempo que as observações foram coletadas;
- iv) Com o gráfico dos resíduos absolutos em função valores ajustados é possível detectar se a função de variância adotada é adequada, espera-se que os pontos apresentem o comportamento de uma distribuição aleatória em torno de zero e amplitude contante;
- v) Para verificar se a função de variância foi corretamente especificada e detectar a presença de *outlier*, esse tipo de análise pode ser feito mediante a construção do gráfico (*normal plot e half-normal plot*) (ATKINSON, 1985). O comportamento esperado para os resíduos para um modelo adequado é aproximadamente uma reta.

Para verificar a qualidade do ajuste do modelo HINDE e DEMÉTRIO (1998), enfatizam a construção do gráfico *half-normal plot* com envelope simulado. Dado que todos os pontos estão contidos no envelope simulado, isto implica em dizer, que há indícios para o modelo está bem ajustado. Com esse gráfico também é possível identificar a existência de observações discrepantes, tal como verificar se a distribuição dos resíduos foi especificada corretamente. De acordo com ATKINSON (1985), para construir o **hnp** é preciso plotar os valores absolutos ordenados de uma determinada medida de diagnóstico apropriada (diferentes tipos de resíduos, distância de Cook, entre outras) contra as estatísticas de ordem esperadas da distribuição meio-normal, em que calculam-se da seguinte forma

$$\Phi^{-1} \left[\frac{(i + n - \frac{1}{8})}{2n + \frac{1}{2}} \right]$$

tendo Φ^{-1} como a função acumulada inversa da distribuição normal padrão, em que $i = 1, \dots, n$, sendo n a dimensão da amostra. Com o objetivo de tornar a análise gráfica mais simples, ATKINSON (1985) desenvolveu o envelope simulado, o qual é elaborado a partir do procedimento descrito abaixo.

- i) Ajustar o modelo e calcular o resíduo pertinente representado por r_i , em valor absoluto e colocá-los em ordem crescente;
- ii) Retornar aos modelos ajustados e para cada amostra simulada calcular os novos $r_{j(i)}$, em valores absolutos, $j = 1, \dots, 99$, $i = 1, \dots, n$ dispor esses valores em ordem crescente;
- iii) Para cada modelo ajustado calcular os percentis 5%, 50% e 95%;
- vi) Plotar os valores desses percentis r_i observados contra as estatísticas esperadas da distribuição meio-normal.

É importante frisar, que foi implementada no *software* livre *R* (R Core Team, 2013) uma função chamada **hnp** (MORAL, 2013), a qual permite gerar os gráficos meio-normais simulado para diferentes tipos de modelos. Para a utilização dessa função é preciso de estarem disponíveis os pacotes **MASS** e **nnet** (VENABLES; RIPLEY, 2002), **psc1** (ZEILEIS; KLEIBER; JACKMAN, 2008), **lme4** (BATES et al., 2015), **VGAM** (YEE, 2013).

3.2.3.2 Avaliação da função de ligação

Como citado anteriormente, umas das causas que podem tornar um modelo mal ajustado é a escolha equivocada da função de ligação, sendo essa, muitas vezes de difícil escolha. Na literatura são descritas técnicas para verificar a adequabilidade da função de ligação, as quais são classificadas em técnicas formais e informais.

Apresentada por McCullagh e Nelder (1989) como uma técnica formal para avaliar a adequabilidade da função de ligação, o procedimento de adicionar ao modelo ajustado o preditor linear ao quadrado (η^2), como uma variável explicativa extra e analisar a redução na *deviance*, o que coincide em aplicar o teste da razão de verossimilhança, ou seja, calcular a diferença entre as *deviances*,

$$\xi_{rv} = \left(D(y; \hat{\mu})_{f1} - D(y; \hat{\mu})_{f2} \right)$$

assim sendo, $D(y; \hat{\mu})_{f1}$ é a *deviance* do modelo (sem o preditor linear adicionado no modelo) encaixado em um modelo maior com a *deviance* $D(y; \hat{\mu})_{f2}$ (com η^2 adicionado no modelo). Com hipótese nula assintoticamente tem-se $\xi_{rv} \sim \chi^2_{f2-f1}$, tendo $f1$ e $f2$ como os graus de liberdade associados aos modelos. Logo, caso ocorra uma redução significativa, isto implica em dizer que há indícios da função de ligação predeterminada, seja inadequada.

Outro método que pode ser utilizado para verificar a adequação da função de ligação é classificado por Cordeiro, Demétrio e Moral (2014) como técnica informal, o qual usa o gráfico da variável dependente ajustada estimada $\hat{z} = \hat{\eta} + \hat{G}(y - \hat{\mu})$ contra o preditor linear estimado $\hat{\eta}$. O padrão nulo é uma reta, apresentando indícios que a função de ligação predeterminada seja adequada.

3.3 Análise de sobrevivência

Sendo um dos métodos na área da estatística com mais destaque nas últimas décadas do século passado, a análise de sobrevivência possui como umas das suas características a variável resposta, pois geralmente é o tempo até a ocorrência de um evento de interesse, o qual segundo Colosimo e Giolo (2006) é determinado como tempo de falha, constituindo-se pelos elementos, tempo inicial, a escala de medida e o evento de interesse (falha). Porém, a principal característica para dados de sobrevivência é a presença de censura, ou seja, dados incompletos.

Defini-se como censura, observações incompletas ou parciais, em que essas observações podem ser provenientes de uma série de acontecimentos, isto é, o acompanhamento do indivíduo foi interrompido por algum motivo, o estudo terminou para a análise dos dados, o indivíduo morreu de causa diferente da estudada, entre outras. Baseado nessas situações, compreende-se que toda informação relacionada ao indivíduo estudado aborda o conhecimento de que o tempo de falha é superior àquele observado. Colosimo e Giolo (2006) apresentam a censura sendo classificada em três tipos.

- i) Censura tipo I, aquela em que o estudo será terminado após um período pré-estabelecido de tempo;
- ii) Censura tipo II, aquela em que o estudo será terminado após ter ocorrido o evento de interesse em um número pré-estabelecido de indivíduos;
- iii) Censura aleatória, aquela que acontece em estudos em que os indivíduos são acompanhados em observações periódicas e é conhecido somente que o evento de interesse ocorreu em um certo intervalo de tempo.

Vale ressaltar que existe outros tipos de censura classificadas como intervalar e trucamento. Na Figura 3, apresenta-se uma ilustração de alguns mecanismos de censura em que ● representa a falha e ○ a censura. Para todos os indivíduos o evento ocorreu antes do final do experimento (a), para alguns indivíduos o evento não foi concretizado até o final do experimento (b), o experimento foi finalizado após a ocorrência de um número pré-estabelecido de falhas (c), o acompanhamento de alguns indivíduos foi interrompido por alguma razão e para alguns indivíduos o evento não ocorreu até o termino do experimento (d).

Figura 3 - Representação gráfica para os tipos de censuras citadas anteriormente.

Fonte: Colosimo e Giolo (2006)

Na técnica análise de sobrevivência a variável resposta é comumente especificada pela função de sobrevivência, sendo essa, dada em termos probabilísticos denotada como $S(t) = P(T \geq t)$, em que $S(t)$ representa a função de sobrevivência que é definida como a probabilidade de uma observação não falhar até um certo tempo t , isto é, probabilidade de uma observação prolongar-se por um período de tempo T maior que o tempo t ($P(T \geq t)$). Em decorrência, define-se a função de distribuição acumulada como a probabilidade de uma observação não sobreviver ao tempo t , sendo denotada como $F(t) = 1 - S(t)$. Outra função de suma importância em análise de sobrevivência é a função de taxa de falha acumulada, para um dado conjunto de condições de operação a confiabilidade é a probabilidade de que um sistema um certo período de tempo, sendo essa função definida por:

$$\Lambda(t) = \int_0^t \lambda(u) du,$$

sendo que sua função não possui uma interpretação direta, mas é útil na avaliação da função de maior interesse que é a taxa de falha $\lambda(t)$, a qual tem uma relação matemática importante com as funções densidade de probabilidade e função de sobrevivência, como

descrito a seguir

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

Quando tem-se o interesse em estimar as funções citadas, uma das formas é utilizar o estimador não paramétrico de Kaplan-Meier (KAPLAN; MEIER, 1958), sendo aplicado quando os dados apresentam observações censuradas. Esse estimador possui como principais propriedades a de não ser viciado para grandes amostras, ser fracamente consistente, convergir assintoticamente para um processo gaussiano além de ser, um estimador de máxima verossimilhança de $S(t)$ (BRESLOW; CROWLEY, 1974). Desse modo, o estimador de Kaplan-Meier é apresentado pela expressão

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

sendo $t_1 < t_2 \dots < t_k$, os k tempos distintos e ordenados de falha, d_j o número de falhas em t_j , $j = 1, \dots, k$ e n_j o número de indivíduos sob risco em t_j , ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j .

3.3.1 Modelos Probabilísticos

Entre uma série de modelos probabilísticos existentes na literatura, alguns deles recebem uma posição de destaque pelo fato de apresentarem adaptação em várias situações práticas. Porém, precisa-se ter um amplo conhecimento sobre cada um desses modelos, uma vez que apresentam características diferentes e assim, podem gerar estimadores diferentes para a mesma quantidade desconhecida, logo, a utilização inadequada acarretará erros grosseiros nas estimativas dessas quantidades. Dentre esses modelos, destacam-se os modelos exponencial, Weibull e o log-normal, que são apropriados para descrever o comportamento da variável tempo até a falha, provenientes de estudos voltados para dados assimétricos positivos. A seguir são descritos esses modelos.

Distribuição Exponencial

Desenvolvida por Laplace (1774), a distribuição exponencial é considerada como sendo um dos modelos probabilísticos mais simples usados para descrever o tempo

de falha. Entre suas características, está a sua composição formada por apenas um único parâmetro além de ser, a única distribuição que possui uma função de taxa de falha constante. Sua função de densidade de probabilidade para a variável aleatória tempo de falha T é dada por

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

em que o parâmetro $\alpha > 0$ representa o tempo médio de vida. Desenvolvendo algebricamente a função densidade de probabilidade, obtêm-se as funções de sobrevivência $S(t)$ e de taxa de falha $\lambda(t)$, sendo expressas por

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

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

A forma típica das funções citadas anteriormente são apresentadas na Figura 4.

Figura 4 - Funções de densidade de probabilidade $f(t)$ (a) , de sobrevivência $S(t)$ (b) e de taxa de falha (c) da distribuição exponencial para $\alpha = 1$ (cor azul), $\alpha = 0.7$ (cor vermelha) e $\alpha = 0.5$ (cor verde).

Como mencionado anteriormente, apenas essa distribuição possui taxa de falha constante, isto implica em dizer, que tanto uma unidade velha quanto uma nova, que ainda não falharam, têm o mesmo risco de falhar em um intervalo futuro, em que essa característica é denominada como falta de memória da distribuição exponencial.

Distribuição de Weibull

Uma das distribuições mais populares em aplicações práticas, a distribuição de Weibull foi desenvolvida por Weibull (1939), desde então, vem sendo constantemente aplicada em estudos desenvolvidos pelas áreas biomédicas e industriais. A sua grande aplicabilidade é motivada por apresentar uma variedade de formas, mas com apenas uma propriedade, a sua função de taxa de falha é monótona, ou seja, ela pode ser crescente, decrescente ou constante. Diante disto, uma variável aleatória T com distribuição de Weibull, tem sua função densidade de probabilidade, dada por

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

em que γ representa o parâmetro de forma, e α representa o parâmetro de escala, ambos positivos. A partir do desenvolvimento algébrico obtêm-se as funções de sobrevivência e de risco falha, expressas como

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

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \quad \text{para } t > 0, \quad \alpha \text{ e } \gamma > 0.$$

Destaca-se que quando $\gamma = 1$, tem-se a distribuição exponencial, logo a distribuição exponencial é um caso particular da distribuição Weibull. Conforme apresentado na Figura 5, juntamente com outros esboços da distribuição de Weibull para diferentes valores paramétricos.

Figura 5 - Funções de densidade de probabilidade $f(t)$ (a) , de sobrevivência $S(t)$ (b) e de taxa de falha (c) da distribuição de Weibull para alguns valores dos parâmetros (γ, α) , cor azul (3, 250), cor vermelha (4, 350), cor verde (8, 600), cor cinza (1, 150) e cor preta (0.5, 50).

Distribuição Log-normal

Sendo utilizada em diversas áreas, a distribuição log-normal é uma distribuição flexível fortemente relacionada com a distribuição normal, tal que $\log(T)$ tem distribuição normal com média μ e variância σ^2 , então declara-se que T tem distribuição log-normal, com função densidade de probabilidade dada por

$$f(t) = \frac{1}{\sqrt{2\pi t\sigma}} \exp \left\{ -\frac{1}{2} \left(\frac{\log(t) - \mu}{\sigma} \right)^2 \right\}, \quad t > 0,$$

em que μ é a média do logaritmo do tempo de falha e σ o desvio-padrão. Assim, o logaritmo de uma variável normal com média μ e desvio-padrão σ , esta relação significa que dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, mas considerando o logaritmo dos dados (COLOSIMO; GIOLO, 2006). A seguir apresenta-se as funções de sobrevivência e de taxa de falha para uma variável que tenha distribuição log-normal.

$$S(t) = \Phi \left(\frac{-\log(t) + \mu}{\sigma} \right) \quad \text{e} \quad \lambda(t) = \frac{f(t)}{S(t)},$$

sendo Φ função de distribuição acumulada de uma distribuição normal padrão. Na Figura 6, tem-se a ilustração para cada função citada referente a distribuição log-normal.

Figura 6 - Funções de densidade de probabilidade $f(t)$ (a), de sobrevivência $S(t)$ (b) e de taxa de falha (c) da distribuição log-normal para alguns valores dos parâmetros μ, α , cor azul (0, 0.5), cor vermelha (0, 0.7), cor verde (0, 1.5), cor cinza (1, 0.7) e cor preta (1, 2).

Diante dos modelos citados, procede com uma descrição sobre os métodos utilizados, quando tem o objetivo de selecionar o modelo probabilístico mais adequado aos dados. Esta seleção é realizada por meio de técnicas gráficas e com aplicação de testes de hipóteses com modelos encaixados (COX; HINKLEY, 1974). As técnicas gráficas podem ser aplicadas de diferentes formas, uma delas consiste na comparação da função de sobrevivência do modelo proposto com o estimador de Kaplan-Meier, em que por meio desse estimador obtêm-se a função de sobrevivência empírica, $\hat{S}_{KM}(t)$, posteriormente estima-se a função de sobrevivência considerando uma determinada função densidade de probabilidade, como por exemplo uma distribuição log-normal $\hat{S}_{ln}(t)$.

Em seguida, comparam-se graficamente as funções de sobrevivência estimadas para cada distribuição de probabilidade com $\hat{S}_{KM}(t)$, caso for identificado que o modelo, o qual está sendo averiguado, apresentar sua curva de sobrevivência próxima daquela do estimador de Kaplan-Meier, há indícios de um ajuste satisfatório. Colosimo e Giolo (2006) citam um método que consiste na linearização da função de sobrevivência tendo como ideia

básica a construção de gráficos que sejam aproximadamente lineares, sendo assim, o modelo proposto ajusta-se ao conjunto de dados em estudo.

Outra técnica apresentada na literatura é a técnica de resíduos de COX e SNELL (1968), em que são geradas curvas de relação de resíduos, isto é, relacionando duas distribuições de resíduos, sendo a primeira estimada pela técnica não paramétrica e a segunda por uma distribuição que deseja-se analisar. Segundo Colosimo e Giolo (2006), esses resíduos são quantidades calculadas pela expressão

$$\hat{e}_i = \hat{\Lambda}(t_i | x_i),$$

sendo $\hat{\Lambda}(\cdot)$ a função taxa de falha acumulada obtida do modelo ajustado. para os modelos exponencial, Weibull e log-normal, têm-se os resíduos de Cox-Snell calculados, respectivamente, por

- i) Exponencial: $\hat{e}_i = \left[t_i \exp \left\{ -x_i' \hat{\beta} \right\} \right];$
- ii) Weibull: $\hat{e}_i = \left[t_i \exp \left\{ -x_i' \hat{\beta} \right\} \right]^{\hat{\gamma}};$
- iii) log-normal: $\hat{e}_i = -\log \left[1 - \Phi \left(\frac{\log(t_i) - x_i' \hat{\beta}}{\hat{\sigma}} \right) \right].$

Caso o modelo em averiguação seja adequado, então os resíduos $\hat{e}_i$ devem seguir uma distribuição exponencial padrão (LAWLESS, 1982). A análise dessa suposição é feita por meio do gráfico $\hat{e}_i$ versus $\hat{\Lambda}(\hat{e}_i)$, o qual deve ser aproximadamente uma reta com inclinação 1, caracterizando um ajuste satisfatório do modelo em análise.

Entretanto, as técnicas gráficas não são suficientes para determinar se um modelo está adequado ou não, pois existem um componente subjetivo na interpretação dos gráficos. Então, recorre-se aos testes de hipóteses para obter uma conclusão fidedigna sobre o modelo em análise. As hipóteses a serem testadas são:

$$\begin{cases} \mathbf{H}_0 : \text{O modelo de interesse é adequado;} \\ \mathbf{H}_1 : \text{O modelo de interesse não é adequado.} \end{cases}$$

Entre alguns testes que podem ser utilizados para testar essas hipóteses, destaca-se o teste da razão de verossimilhança em modelos encaixados (COX; HINKLEY,

1974), ou seja, para aplicação desse teste deve ser identificado um modelo geral tal que os modelos de interesse sejam casos particulares. A partir disso, procede-se com o ajuste do modelo geral e obtenção do valor do logaritmo de sua função de verossimilhança $\left(\log L\left(\widehat{\theta}_G\right)\right)$, posteriormente ajusta o modelo de interesse e obtêm o valor do logaritmo de sua função de verossimilhança $\left(\log L\left(\widehat{\theta}_M\right)\right)$, em seguida, calcula-se o valor referente a estatística da razão de verossimilhança, dado pela fórmula

$$TRV = -2 \log \left[\frac{L\left(\widehat{\theta}_M\right)}{L\left(\widehat{\theta}_G\right)} \right] = 2 \left[\log L\left(\widehat{\theta}_G\right) - \log L\left(\widehat{\theta}_M\right) \right],$$

ressalta que, sob H_0 a estatística TRV, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros $\widehat{\theta}_G$ e $\widehat{\theta}_M$ dos modelos sendo comparados. Na teoria análise de sobrevivência, o teste é aplicado utilizando-se a distribuição gama generalizada, a qual possui os modelos exponencial, Weibull, log-normal e gama como modelos encaixados, pois esses modelos são casos particulares da distribuição gama generalizada.

Contudo, quando o modelo adequado é selecionado, as análises continuam e o próximo passo é realizar a estimação dos parâmetros. Em estudos voltados para variável tempo de falha, os parâmetros devem ser estimados a partir das observações amostrais, para que o modelo fique determinado e seja possível responder às perguntas de interesse.

Segundo LAWLESS (2003), o método de estimação mais utilizado em análise de sobrevivência é o método da máxima verossimilhança, uma vez que esse método permite que a informação dos tempos censurados sejam incorporados no processo de estimação, isso implica que a função de verossimilhança considera a contribuição dos indivíduos que apresentam falhas e os indivíduos que foram censurados.

Com isso, sejam $(y_1, x_{i1}, \delta_1), (y_2, x_{i2}, \delta_2), \dots, (y_n, x_{in}, \delta_n)$, n observações independentes em que $y_i = \log(t_i)$, representa o logaritmo do tempo de falha ou censura, $\mathbf{X}_i = (x_{i1}, \dots, x_{ip})^T$ o vetor de covariáveis e δ_i é o indicador de censura, para todo $i = 1, \dots, n$. Então, o logaritmo da função de verossimilhança considerando uma amostra, é dado por:

$$\log (L(\boldsymbol{\theta})) = \log \left(\prod_{i=1}^n [f(t_i; \boldsymbol{\theta})]^{\delta_i} [S(t_i; \boldsymbol{\theta})]^{1-\delta_i} \right)$$

$$l(\boldsymbol{\theta}) = \sum_{i=1}^n \delta_i \log f(t_i; \boldsymbol{\theta}) + (1 - \delta_i) \log S(t_i; \boldsymbol{\theta})$$

em que $f(\bullet)$ representa a função densidade de probabilidade, $S(\bullet)$ representa a função de sobrevivência da variável aleatória e δ_i representa a variável indicadora de falha, o vetor de parâmetros a ser estimado é representado por $\boldsymbol{\theta} = (\beta_1, \dots, \beta_p)^T$, Os estimadores de máxima verossimilhança após derivar $l(\boldsymbol{\theta})$ em relação a cada um dos componentes são obtidos $\boldsymbol{\theta}$.

Teste *log rank*

Quando tem-se o interesse de comparar curvas de sobrevivência, a literatura apresenta o teste de *log-rank*, proposto por Watanabe (1959) para comparar duas populações, mas para o caso de J populações foi desenvolvido por Andersen e Gill (1982), sendo aplicado sob as hipóteses.

$$\begin{cases} H_0 : S_i(t) = S_n(t) \\ H_1 : S_i(t) \neq S_n(t) \end{cases}$$

em que $S_i(t)$ e $S_n(t)$ representam as funções de sobrevivência a serem comparadas. Segundo Colosimo e Giolo (2006) para testar as hipóteses citadas anteriormente, procede-se com os dados em análise sendo arranjados em forma de uma tabela de contingência $2 \times r$ com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i . Condicional à experiência de falha e censura até o tempo t_j e ao número de falhas no tempo t_j , a distribuição conjunta de $d_{2j}, \dots, d_{rj}$ é, então, uma distribuição hipergeométrica multivariada, conforme apresentado a seguir,

$$\frac{\prod_{i=1}^r \binom{n_{ij}}{d_{ij}}}{\binom{n_j}{d_j}},$$

em que a média de d_{ij} é $w_{ij} = n_{ij}d_jn_j^{-1}$, bem como a variância de d_{ij} e a covariância de d_{ij} e d_{ij} são calculados pela seguinte fórmula,

$$(V_j)_{ii} = n_{ij} (n_j - n_{ij}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}$$

e

$$(V_j)_{il} = -n_{ij} n_{il} d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}.$$

Logo, a estatística $v'_j = (d_{2j} - w_{2j}, \dots, d_{rj} - w_{rj})$ tem média zero e matriz de variância-covariância V_j de dimensão $r - 1$, com $(V_j)_{ii}$, $i = 2, \dots, r$, na diagonal principal e os elementos $(V_j)_{il}$, $i, l = 2, \dots, r$ fora da diagonal principal. Assim, forma-se a estatística v , somando sobre todos os tempos distintos de falha,

$$v = \sum_j^k v_j,$$

sendo v um vetor de dimensão $(r - 1) \times 1$, cujos elementos são as diferenças entre os totais observados e esperados de falha. Considerando a suposição de que as k tabelas de contingência são independentes, a variância da estatística v será $V = V_1 + \dots V_k$. Contudo a estatística para o teste *log rank* é dada por,

$$T = v' V^{-1} v.$$

Posteriormente, descreve como procedeu o experimento citado na secção 1 e quais métodos aplicados para obtenção dos respectivos resultados apresentados na secção 5.

4 MATERIAL E MÉTODOS

Nesta secção, apresenta-se uma explanação sobre o experimento citado anteriormente e sobre os métodos que foram utilizados para obtenção dos respectivos objetivos, descritos na secção 3.

4.1 Material

Como uma alternativa para implementações em manejo integrado de pragas, alguns estudos têm avaliado a eficiência de predação de diversas espécies, com o intuito de suprimir, ou pelo menos atenuar o uso de inseticidas. Com isso, o banco de dados que será utilizado neste projeto provém de um experimento que foi desenvolvido durante o segundo semestre do ano de 2013, com o objetivo de avaliar o comportamento dos predadores *Euborellia annulipes* (“Tesourinha”) e *Harmonia axyridis* (“Joaninha”), em relação à praga *Aphis gossypii* (“Pulgão”), a qual acarreta prejuízos para a cultura do algodão. O experimento foi realizado no Laboratório de Ecologia de Insetos da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ). Os bioensaios foram realizados em dois períodos: período diurno com uma luz fluorescente e no período noturno, onde foi utilizada uma luz infravermelha para facilitar a visualização dos isentos, sendo a luz não identificada pelos predadores, com isso o ambiente mantém-se em aspecto noturno. Foi utilizado um delineamento inteiramente casualizado, com quatro tratamentos, isto é, tesourinha fêmea com o pulgão, tesourinha macho com o pulgão, joaninha fêmea com o pulgão e joaninha macho com o pulgão. Nas figuras 7, 8 e 9 encontram-se as imagens ilustrativas para as espécies envolvidas no experimento.

Figura 7 - (A) Espécie *Harmonia axyridis* (“Joaninha Fêmea”) e (B) (“Joaninha Macho”).

Fonte: McCornack *et. al.* (2007).

Na Figura 7 percebe-se que a Joaninha fêmea possui uma pigmentação com cor escura localizada no lábio superior. Enquanto, o macho possui uma pigmentação com cor clara, também localizada no lábio superior.

Figura 8 - (A) Espécie *Euborellia annulipes* (“Tesourinha Fêmea”) e (B) (“Tesourinha Macho”).

Fonte: Lucas *et. al.* (2009).

Na Figura 8 percebe-se que a Tesourinha fêmea é caracterizada por apresentar cercos mais fechados em formato semelhante a uma pinça e, em geral a fêmea é maior.

Enquanto que o macho é caracterizado por apresentar cercos mais abertos.

Figura 9 - Imagem referente à praga da espécie *Aphis gossypii* (“Pulgão”), sexo fêmea.

Fonte: Francisco Sales (2013).

O experimento foi realizado durante 60 minutos, sendo esse tempo determinado após a realização de um experimento piloto. Então, os experimento foi composto por 21 repetições, onde cada repetição foi representada por uma placa de Petri medindo 60×15 mm. Em cada placa foi liberado um pulgão adulto áptero na parte central, três observadores avaliaram as variáveis definidas como tempo de ataque em minutos e tempo de consumo em minutos do predador, utilizando um cronômetro durante o experimento. Vale ressaltar, que antes do experimento ser realizado, os predadores permaneceram 12 horas em jejum. Estão sendo utilizadas técnicas de modelagem para dados contínuos associados ao tempo, com o objetivo determinar se houve diferença ou não, no tempo médio de ataque para cada tratamento. Posteriormente é apresentado uma imagem referente ao croqui utilizado para a instalação do experimento.

Figura 10 - Imagem referente as placas de Petri utilizada para realizar o experimento.

Conforme citado anteriormente, cada placa refere-se a uma repetição para cada tratamento. Dessa forma, os tratamentos foram casualizados por meio de um sorteio, sendo o mesmo, realizado em uma urna contendo 21 papéis representando cada tratamento, totalizando 84 papéis. A cada papel sorteado, uma placa era identificada com o respectivo tratamento sorteado.

4.2 Métodos aplicados ao conjunto de dados que não contém censuras

De acordo com as informações passadas pelos pesquisadores, foi identificado que durante condução do experimento, em algumas placas o predador não efetuou o ataque a praga, evidenciando a presença de censura à direita (23% dos casos), principal característica da técnica análise de sobrevivência. Segundo Colosimo e Giolo (2006), censura à direita é aquela em que o tempo de ocorrência do evento está à direita do tempo de interesse, ou seja, o tempo entre o início do estudo e o evento é maior do que o tempo observado.

Então, diante deste contexto optou-se por realizar dois tipos de análises, sendo uma realizada sem levar em consideração as censuras e outra análise considerando as censuras para mostrar o efeito causado pelo fato de não considerar censuras na análise. Com isso, os resultados apresentados nesse trabalho são baseados em duas metodologias, sendo

elas, modelos lineares generalizados, aplicados aos dados sem considerar as censuras, e a técnica análise de sobrevivência, aplicada aos dados considerando as censuras.

Logo, tendo início com as análises dos dados sem considerar as censuras, primeiramente, foi realizado uma análise descritiva para ter uma ampla compreensão dos dados em estudos. Dessa forma, apresenta-se tabelas e gráficos, em que por meio das tabelas são apresentadas algumas medidas descritivas enquanto, por meio dos gráficos será possível ter uma breve ideia do comportamento dos dados.

Sendo Y_{ijk} o valor observado para a variável resposta obtido para o efeito i -ésimo sexo (fêmea, macho), efeito do j -ésimo inseto (tesourinha, joaninha) e efeito do k -ésimo turno (diurno, noturno). A princípio, o componente sistemático refere-se a um delineamento inteiramente casualizado, com um esquema fatorial ($2 \times 2 \times 2$). Inicialmente considerou-se os modelos em que a variável resposta Y_{ijk} tem distribuição gama com as funções de ligação, logarítmica, identidade e inversa. Também foi ajustado o modelo em que a variável resposta Y_{ijk} tem distribuição normal inversa com função de ligação canônica, ou seja, quando as funções $\eta(\theta)$ e $t(x)$ são iguais a função identidade. Para situação experimental descrita e, considerando os objetivos do estudo, tem-se como estrutura para o preditor linear os modelos:

Modelo 1:

- I) Supõe-se que a variável resposta $Y_{ijk} \sim G(\mu_{ijk}, \phi)$;
- II) O preditor linear é expresso por $\eta_{ijk} = \mu + S_i + I_j + T_k + (SI)_{ij} + (ST)_{ik} + (IT)_{jk} + (SIT)_{ijk}$, sendo μ a média geral, S_i o efeito do i -ésimo sexo, I_j efeito do j -ésimo inseto, T_k o efeito do k -ésimo turno, $(SI)_{ij}$ o efeito da interação entre o i -ésimo sexo e o j -ésimo inseto, $(ST)_{ik}$ o efeito da interação entre o i -ésimo sexo e o k -ésimo turno, $(IT)_{jk}$ o efeito da interação entre o j -ésimo inseto e o k -ésimo turno, $(SIT)_{ijk}$ o efeito da interação entre o i -ésimo sexo e o j -ésimo inseto com o k -ésimo turno;
- III) A função de ligação é a identidade, $g(\mu_{ijk}) = \mu_{ijk} = \eta_{ijk}$.

Modelo 2:

- I) Supõe-se que a variável resposta $Y_{ijk} \sim G(\mu_{ijk}, \phi)$;

- II) O preditor linear será o mesmo descrito no modelo 1;
- III) A função de ligação é a logarítmica, $g(\mu_{ijk}) = \log(\mu_{ijk}) = \eta_{ijk}$.

Modelo 3:

- I) Supõe-se que a variável resposta $Y_{ijk} \sim G(\mu_{ijk}, \phi)$;
- II) O preditor linear será o mesmo descrito no modelo 1;
- III) A função de ligação é a inversa, $g(\mu_{ijk}) = 1/\mu_{ijk}^2 = \eta_{ijk}$.

Modelo 4:

- I) Supõe-se que a variável resposta $Y_{ijk} \sim NI(\mu_{ijk}, \lambda)$;
- II) O preditor linear será o mesmo descrito no modelo 1;
- III) A função de ligação é a inversa, $g(\mu_{ijk}) = 1/\mu_{ijk}^2 = \eta_{ijk}$.

Todos os modelos propostos foram testados sob as seguintes hipóteses:

$$\begin{cases} H_0 : \text{O modelo ajustado apresenta um ajuste satisfatório;} \\ H_1 : \text{O modelo ajustado não apresenta um ajuste satisfatório.} \end{cases}$$

Para realizar a escolha adequada de uma das hipóteses propostas, foram feitas as comparações da qualidade de ajuste dos modelos com diferentes distribuições de probabilidades e diferentes funções de ligação por meio do *half-normal plot*, considerando o nível de 5% de significância. Vale ressaltar que quanto menor o valor desvio (D_p), melhor será o ajuste do modelo (CORDEIRO; DEMÉTRIO; MORAL, 2014). Sendo assim, encerra-se a metodologia descrita para as análises realizadas sem levar em consideração as censuras, os resultados estão apresentados na secção 5.1, posteriormente é descrita a metodologia referente as análises realizadas levando em consideração as censuras.

4.3 Métodos aplicados ao conjunto de dados que contém censuras

Como citado anteriormente, a análise de sobrevivência é formada por um conjunto de procedimentos utilizados para a análise estatística em dados cuja variável resposta é o tempo até a ocorrência de um evento de interesse, determinado a partir de um tempo inicial pré-estabelecido (HOUGAARD, 1999; COLOSIMO; GIOLO, 2006), tendo como sua principal característica a eficácia de extrair informações de dados que apresentam censuras, isto é, indivíduos não foi observada a ocorrência do evento até o término do experimento. Neste trabalho como já foi descrito, o tempo estudado é o tempo preciso para o predador efetuar o ataque a praga, em que essa variável aleatória ($T > 0$) pode ser expresso pelas funções, densidade de probabilidade $f(t)$, sobrevivência $S(t)$ e a função taxa de falha $\lambda(t)$. Para estimar essas funções aplica-se um estimador não paramétrico conhecido como Kaplan-Meier (KAPLAN; MEIER, 1958), utilizado quando deseja-se estimar as funções citadas considerando a presença de observações censuradas.

Contudo, quando tem-se o objetivo de verificar a influência de covariáveis nos tempos de ataque, a utilização de modelos paramétricos torna-se mais viável, pois assim, pode-se identificar quais covariáveis estão mais associadas com a variável resposta. Nesse sentido, define-se como modelos paramétricos, os modelos que assumem uma distribuição conhecida de probabilidade para a variável resposta, dentre essas distribuições, citam-se exponencial, Weibull, log-normal, entre outras. Para situação experimental descrita e considerando os objetivos do estudo, foram ajustados três modelos de análise de sobrevivência, exponencial, Weibull e log-normal, com o auxílio do pacote **survival** (THERNEU, 2016), sendo utilizado inicialmente o preditor linear

$$\eta_{ijk} = \mu + S_i + I_j + T_k + (SI)_{ij} + (ST)_{ik} + (IT)_{jk} + (SIT)_{ijk},$$

em que, S_i o efeito do i -ésimo sexo, I_j efeito do j -ésimo inseto, T_k o efeito do k -ésimo turno, $(SI)_{ij}$ o efeito da interação entre o i -ésimo sexo e o j -ésimo inseto, $(ST)_{ik}$ o efeito da interação entre o i -ésimo sexo e o k -ésimo turno, $(IT)_{jk}$ o efeito da interação entre o j -ésimo inseto e o k -ésimo turno, $(SIT)_{ijk}$ o efeito da interação entre o i -ésimo sexo e o j -ésimo inseto com o k -ésimo turno.

A seleção do modelo que melhor se ajusta ao conjunto de dados foi feita por meio de técnicas gráficas e testes de hipóteses. Para as técnicas gráficas foram obtidas as funções de sobrevivência empírica a partir do estimador de Kaplan-Meier, denotada por $\hat{S}_{KM}(t)$, em seguida foi estimada a mesma função assumido a distribuição especificada.

Dessa forma, foram comparados graficamente as funções de sobrevivência estimadas para cada distribuição de probabilidade com $\hat{S}_{KM}(t)$, em que o modelo selecionado foi aquele cujas observações estiverem mais próximas da reta $y = x$, ou seja, $x = \hat{S}_{KM}(t)$ e $y = \hat{S}_W(t)$, por exemplo. Também foram utilizados, para selecionar o modelo mais adequado, o critério de informação Akaike (AIC) e o teste da razão de verossimilhanças indicam que o modelo mais adequado foi o qual apresentou os menores valores para esses testes. Cujos resultados estão apresentados na secção 5.2.

Ressalvo, que as metodologias apresentadas nesta secção foram aplicadas com o auxílio do *software* estatístico R (R Core Team, 2013).

5 RESULTADOS E DISCUSSÃO

5.1 Resultados referentes ao conjunto de dados que não contém censuras

Dando início a esta seção é apresentada uma análise descritiva, realizada a fim de ter uma compreensão ampla sobre o conjunto de dados em estudo. Na Tabela 2, apresenta-se algumas medidas descritivas para uma análise inicial.

Tabela 2 - Medidas descritivas do tempo de ataque (em minutos) dos predadores Tesourinha e Joaninha em relação à praga Pulgão nos períodos, diurno e noturno.

Tratamentos	Tempo de Ataque - Período Diurno				
	Mínimo	Média	Mediana	Desvio - Padrão	Máximo
Tesourinha Fêmea	0,93	17,00	8,45	17,58	55,85
Tesourinha Macho	1,03	17,91	4,02	21,80	57,22
Joaninha Fêmea	0,55	10,24	4,39	12,25	48 ,00
Joaninha Macho	0,33	7,89	2,59	10,57	36,25
Tratamentos	Tempo de Ataque - Período Noturno				
	Mínimo	Média	Mediana	Desvio - Padrão	Máximo
Tesourinha Fêmea	0,78	10,69	7,08	10,36	34,55
Tesourinha Macho	0,45	12,86	10,97	15,63	58,55
Joaninha Fêmea	1,27	13,87	13,73	12,88	50,62
Joaninha Macho	2,22	14,74	11,82	10,59	36,10

De modo geral, a Joaninha fêmea apresentou em média menor tempo de ataque no período diurno. Entretanto, no período noturno a Tesourinha fêmea, destacou-se em relação aos demais tratamentos por ter apresentado em média menor tempo de ataque. Pela análise da Figura 11, nota-se a presença de possíveis valores discrepantes, sendo três valores identificados na Figura 11.a e dois valores identificados na Figura 11.b. Por meio

da Figura 11 também é possível notar, indícios de heterogeneidade de variância, além de observar que os dados apresentam uma assimetria positiva, caracterizando uma distribuição que possua suporte em números reais positivos, como por exemplo as distribuições, gama, normal inversa e weibull, entre outras.

Figura 11 - *Boxplot* referente ao tempo de ataque em minutos dos predadores em relação à praga nos períodos, (a) diurno e (b) noturno. Legenda: Tesourinha Fêmea (TF), Tesourinha Macho (TM), Joaninha Fêmea (JF) e Joaninha Macho (JM).

Prosseguindo com as análises, na Figura 12 tem-se o gráfico *half-normal plot* para os modelos propostos na secção 4.2, por meio desse é possível identificar qual o modelo se ajusta melhor aos dados, dentre os modelos propostos. Mediante análise sob as Figuras 12(a), 12(b), 12(c) observa-se que a maioria dos pontos encontram-se fora do envelope simulado, isto implica em dizer, que esses modelos não apresentam um ajuste satisfatório aos dados. Porém, verifica-se por meio da Figura 12(d) que o modelo com a distribuição normal inversa e com função de ligação canônica, apresenta um ajuste adequado, visto que a maioria dos pontos estão localizados dentro do envelope simulado.

Figura 12 - *Half-normal plot* para os modelos: distribuição gama com função de ligação identidade (a), distribuição gama com função de ligação logarítmica (b), distribuição gama com função de ligação inversa (c), distribuição normal inversa com função de ligação canônica (d).

Após a identificação do modelo mais adequado, procedeu-se com as análises referentes ao preditor linear com o intuito de verificar quais fatores deverão ser incorporados nele. Logo, aplicou-se o teste F citado por Paula (2013), como uma medida apropriada para identificar os fatores adequados ao estudo, conforme os resultados apresentados na Tabela 3. De acordo com a Tabela 3, verifica-se que o preditor linear será mais preciso quando

composto com os fatores inseto, turno e a interação entre eles, pois para a estatística do teste F obteve um valor de 4.6219, com um $p - valor = 0.03351$ aproximadamente, sendo significativo ao nível de 5% de significância. A interpretação feita sobre a Tabela 3 é constatada por meio da Figura 13, sendo possível visualizar que todos os pontos se acomodam dentro do envelope simulado Figura 13(a). Além disso, observa-se que os componentes do desvio se concentram entre os valores -2 e 2 Figura 13(b), em que esses indícios mostram um ajuste satisfatório do modelo proposto.

Tabela 3 - Teste F para o modelos ajustados com todas as combinações possíveis para o preditor linear, utilizando a distribuição normal inversa e função de ligação canônica.

Efeito	g.l.	Deviance	F	$p - valor$
inseto	130	36.518	1.0517	0.30711
turno	129	36.518	0.0043	0.94794
sexo	128	36.517	0.0104	0.91894
inseto $\times$ turno	127	36.017	4.6219	0.03351*
inseto $\times$ sexo	126	35.996	0.1980	0.65713
sexo $\times$ turno	125	35.934	0.5747	0.44984
inseto $\times$ sexo $\times$ turno	124	35.915	0.1783	0.67360

Figura 13 - *Half-normal plot* para o modelo normal inversa (a) e gráfico dos componentes do desvio versus os valores ajustados (b).

Posteriormente, apresenta-se a Tabela 4 constando as estimativas e os erros padrão para cada parâmetro que compõe o preditor linear do modelo com a distribuição normal inversa e com função de ligação canônica.

Tabela 4 - Estimativas e erros padrão dos parâmetros do modelo normal inversa com função de ligação canônica.

Parâmetros	Estimativas	Erro - Padrão	<i>p-valor</i>
Intercepto	0.0121	0.0037	0.0015
inseto	-0.0089	0.0041	0.0347
turno	-0.0073	0.0043	0.0952
inseto x turno	0.0114	0.0054	0.0372

Com o intuito de alcançar um dos objetivos deste trabalho, ou seja, verificar se há diferença significativa entre os tratamentos, caso essa diferença seja detectada, identificar quais tratamentos diferem entre si. Então, procedeu-se com o método de modelos encaixados, nos quais foram testados a significância de cada tratamento, com isso, verificou-se que quando compara-se os predadores tesourinha e joaninha no período diurno,

há diferenças significativas, sendo o predador joaninha mais eficaz, pois apresentou o menor tempo médio de ataque. Quando comparados no período noturno, não detecta-se diferenças significativas entre os dois predadores.

5.2 Análise incluindo as observações censuradas

A partir desta secção, apresentam-se os resultados provenientes da análise realizada considerando-se os dados com censura, inicialmente, tem-se na Figura 14 as estimativas de Kaplan-Meier para os predadores tesourinha e joaninha, ambos os sexos no período diurno e noturno. Por meio da Figura 14 (a), nota-se indícios que os predadores tesourinha e joaninha apresentam efeitos diferentes em relação ao ataque a praga, sendo que o predador joaninha apresenta-se ser mais preciso ao efetuar o ataque. Quando analisa ambos os sexos para cada predador, identifica-se que eles não evidenciam divergência, ou seja, o predador tesourinha do sexo fêmea, efetua o ataque de forma similar ao predador tesourinha do sexo macho, o mesmo caso acontece para o predador joaninha. É importante ressaltar, que pela Figura 14 (a) percebe-se que todos os tratamentos tiveram censura, a qual é identificada pelo o sinal de +.

Entretanto, ao observar a Figura 14 (b), visualiza-se que os predadores tesourinha e joaninha possuem características de ataque similar em relação a praga, mas quando considera-se os sexos de cada predador, verifica-se a princípio que o predador tesourinha sexo fêmea tem um comportamento diferente dos demais, sendo esse o único tratamento que não houve censura durante o experimento.

Figura 14 - Estimativas de Kaplan-Meier para curvas de sobrevivência para os predadores tesourinha e joaninha, ambos os sexos no período diurno (a) e no período noturno (b).

Diante disso, com o intuito de alcançar um dos objetivos do trabalho, isto é, analisar quais tratamentos diferem entre si, então procedeu-se com aplicação do teste *log-rank* com a correção de Bonferroni, conforme apresentado por Colosimo e Giolo (2006), em que essa correção é feita com o nível de significância $\frac{\alpha}{c}$, sendo α o nível nominal de significância, nesse caso $\alpha = 0.05$ e c sendo referente ao número de comparações a serem feitas, ou seja, o novo valor que será atribuído ao nível de significância é de $\alpha = 0.03$. Sendo assim, foram testadas as seguintes hipóteses para o predador tesourinha considerando os dois turnos.

$$\begin{cases} \mathbf{H}_0 : & S_{TF}(t) = S_{TM}(t); \\ \mathbf{H}_1 : & S_{TF}(t) \neq S_{TM}(t), \end{cases}$$

e para o predador joaninha considerando os dois turnos as hipóteses foram.

$$\begin{cases} \mathbf{H}_0 : & S_{JF}(t) = S_{JM}(t); \\ \mathbf{H}_1 : & S_{JF}(t) \neq S_{JM}(t). \end{cases}$$

Ao aplicar o teste de log-rank para comparar as curvas de sobrevivência entre os tratamentos tesourinha fêmea e tesourinha macho no período diurno, obteve um p-valor = 0.4, indicando que não há diferença entre essas curvas de sobrevivência. Porém, quando aplica-se o teste para comparar as curvas de sobrevivência entre os tratamentos tesourinha fêmea e tesourinha macho no período noturno, verifica-se que essas curvas apresentam diferenças, pelo fato do teste resultar em um p-valor = 0.01. Para o predador joaninha, ao aplicar o teste *log-rank* para comparar as curvas de sobrevivência entre os tratamentos joaninha fêmea e joaninha macho nos períodos diurno e noturno, verifica-se que as curvas referentes a esses tratamentos não apresentam diferenças, uma vez que, os p-valores obtidos para essas comparações foram 0.5 e 0.6, caracterizando igualdade das curvas.

Os resultados apresentados mediante aplicação do teste de *log-rank*, corroboram para as interpretações provenientes da análise realizada sobre a Figura 14. Continuando-se com as análises, posteriormente são apresentados alguns métodos aplicados com o objetivo de selecionar o modelo mais adequados aos dados, logo, na Figura 15 é exibido a análise dos resíduos de Cox-Snell para os modelos exponencial, Weibull e log-normal. Desta figura, observa-se que o modelo log-normal apresenta um ajuste satisfatório aos dados.

Figura 15 - Análise gráfica dos resíduos de Cox-Snell para os modelos exponencial, Weibull e log-normal.

Na Tabela 5 os resultados referentes a aplicação do teste da razão de verossimilhança utilizado para testar as hipóteses, i) o modelo exponencial é adequado, ii) o modelo Weibull é adequado, iii) o modelo log-normal é adequado, precederam-se utilizando o modelo gama generalizados, pois esse modelo inclui como casos especiais, os modelos exponencial, Weibull e log-normal. Pelos resultados descritos na Tabela 5, verifica-se que o modelo log-normal apresenta ser adequado para análise dos dados desse estudo, resultado semelhante as interpretações feitas sobre a Figura 15.

Tabela 5 - Logaritmo da função $L(\theta)$ e os resultados dos TRV para os modelos com diferentes distribuições.

Modelo	$\log (L(\hat{\theta}))$	TRV	p-valor
Gama Generalizado	-527.84	-	-
Exponencial	-546.92	$2(546.92 - 527.84) = 38.16$	0.000
Weibull	-535.17	$2(535.17 - 527.84) = 14.66$	0.001
Log-normal	-528.89	$2(528.89 - 527.84) = 2.1$	0.142

Considerando, então, o modelo log-normal o próximo passo é identificar quais

fatores prosseguirão na análise. Para isso, aplicou-se o teste t-Student, conforme apresentado na Tabela 6, na qual estão apresentados os valores referentes as estimativas, erro padrão e *p-valor*.

Tabela 6 - Tabela para estimativas dos parâmetros, erro padrão e *p-valor*.

Parâmetro	Estimativas	Erro padrão	<i>p-valor</i>
Intercepto	3.178	0.3759	0.0001
Inseto	-0.515	0.5286	0.3300
Sexo	-0.320	0.5277	0.5450
Turno	-1.864	0.5217	0.0003
Inseto $\times$ Sexo	-0.534	0.7401	0.4700
Inseto $\times$ Turno	2.698	0.7508	0.0003
Sexo $\times$ Turno	0.749	0.7354	0.3000
Inseto $\times$ Sexo $\times$ Turno	-0.215	1.0505	0.8000

Por meio da Tabela 6, verifica-se que dentre as combinações possíveis para o preditor linear, o modelo com o melhor ajuste tem-se o predito linear composto pelos fatores inseto, turno e a interação entre eles, pois além de apresentar um *p-valor* = 0.0003, também apresentou o menor valor para o critério *AIC* (1071.89). Diante desses resultados, constata-se que o fator sexo não apresentou efeito significativo.

6 CONCLUSÕES

Neste trabalho o conjunto de dados analisado foi proveniente de um experimento realizado na área entomológica, tendo como a variável resposta o tempo de ataque dos indivíduos envolvidos no estudo. Tendo esse experimento o objetivo de mensurar esse tipo de variável resposta, logo no estudo originou observações censuradas, portanto, o conjunto de dados foi analisado por duas técnicas, sendo uma aplicada ao conjunto de dados sem considerar as censuras e a outra técnica aplicada aos dados considerando as censuras.

A primeira técnica aplicada foi a teoria dos modelos lineares generalizados, pela qual utilizou quatro modelos. Segundo os critérios de adequabilidade do ajuste dos modelos adotados, ou seja, o gráfico *half-normal plot*, conclui-se que os modelos compostos pela distribuição gama e as funções de ligações identidade, logarítmica e inversa, não apresentaram um ajuste satisfatório aos dados. O modelo composto pela distribuição normal inversa e função de ligação canônica foi o que melhor se ajustou aos dados, tendo como preditor linear os fatores inseto, turno e a interação entre eles.

A segunda técnica aplicada foi análise de sobrevivência, sendo por meio dessa utilizado três modelos, os quais tiveram a sua adequabilidade do ajuste aos dados, sendo verificada pelo teste da razão de verossimilhança, critério AIC e pela análise de resíduos de Cox-Snell. Com isso, conclui-se que os modelos exponencial e Weibull não apresentam um ajuste satisfatório aos dados. O modelo log-normal, por sua vez, foi o melhor que se ajustou aos dados, tendo como preditor linear os fatores inseto, turno e a interação entre eles.

Com relação a verificação de diferenças entre os tratamentos, conclui-se que em ambas técnicas os resultados são semelhantes, ou seja, ambas técnicas demonstraram que os predadores tesourinha e joaninha apresentam diferenças apenas quando comparados no período diurno, caso que não ocorre quando comparados no período noturno. Diante disso, o predador joaninha apresenta-se ser mais eficaz no combate a praga em estudo, pois apresentou em média o menor tempo de ataque. Sobre as duas técnicas aplicadas nesse trabalho, verificou-se que quando o conjunto de dados contém presença de censuras é indicado que seja aplicada a técnica análise de sobrevivência, caso o banco de dados não apresente censuras a teoria dos modelos lineares generalizados é recomendada.

REFERÊNCIAS

- AGRESTI, A. **Categorical data analysis**. 2. ed. Gainesville, Florida: A John Wiley & Sons, 2002. 710 p.
- ALMEIDA, R.P.; DOMINGUES, C. A.; RAMALHO, F. S. Manejo integrado de pragas do algodoeiro no brasil. **O Agronegócio do algodão no Brasil**. Brasília, **EMBRAPA**, p. 1033–1098, 2008.
- ALVARENGA, C. D. Controle integrado do pulgao-verde schizaphis graminum(rondani, 1852) em sorgo atraves de genotipos resistentes e do predador doru luteipes(scudder, 1876). Piracicaba, SP (Brazil), 1992.
- ANDERSEN, Per Kragh; GILL, Richard David. Cox's regression model for counting processes: a large sample study. **The annals of statistics**, JSTOR, p. 1100–1120, 1982.
- ASSOCIAÇÃO BRASILEIRA DOS PRODUTORES DE ALGODÃO - ABRAPA. **Algodão no Mundo**. Disponível em: <<http://www.abrapa.com.br>>, 2016. Acesso: 07 de setembro de 2016.
- ASSOCIAÇÃO MATOGROSSENSE DOS PRODUTORES DE ALGODÃO - AMPA. **História do Alg**. Disponível em: <<http://www.ampa.com.br>>, 2016. Acesso: 20 de setembro de 2016.
- ATKINSON, A. C. **Plots, transformations, and regression: an Introduction to graphical methods of diagnostics regression analysis**. Oxford statistical science series. [S.l.]: Oxford: Clarendon, 1985. 282 p.
- AYRES, D. L.; SANTOS, A. S. Aplicações estatísticas nas áreas das ciências bio-médicas. **Instituto Mamirauá, Belém**, 2007.
- BATES, D.; M., Martin; B., Ben; W., Steve. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015.
- BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 211–252, 1964.
- BRESLOW, N.; CROWLEY, J. A large sample study of the life table and product limit estimates under random censorship. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 2, n. 3, p. 437–453, 1974.
- CALLEGARI-JACQUES, S. M. **Bioestatística: princípios e aplicações**. [S.l.]: Artmed Editora, 2009.
- COLLETT, D. **Modelling binary data**. London: Chapman and Hall/CRC, 2002. 408 p.
- COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. In: **ABE-Projeto Fisher**. [S.l.]: Edgard Blücher, 2006.
- CORDEIRO, G. M. **Modelos lineares generalizados**. [S.l.]: Campinas, VII SINAPE, 1986. 286 p.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. **Departamento de Ciências Exatas, ESALQ, USP**, 2007.

CORDEIRO, G. M.; DEMÉTRIO, C. G.; MORAL, R. A. Modelos lineares generalizados e extensões. **São Paulo**, 2014.

CORDEIRO, G. M.; NETO, E. de A. L. **Modelos paramétricos**. [S.l.: s.n.], 2004.

COX, D.R.; SNELL, E.J. **A general definition of residuals (with discussion)**. Journal of the royal statistical society. [S.l.], 1968. Series B, n. 30.

COX, D. R.; HINKLEY, D. V. **Theoretical Statistics**. Chapman and Hall, London, 1974.

CRISTOFOLETTI, N. **Relações tróficas entre Euborellia annulipes, Harmonia axyridis e três pragas para aplicação em manejo**. 2014. Tese (Doutorado) — Escola Superior de Agricultura "Luiz de Queiroz", 2014.

DOBSON, A. J. **An introduction to generalized linear models**. [S.l.]: CRC press, 2010.

HINDE, J.; DEMÉTRIO, C. G. B. Overdispersion: models and estimation. Computational Statistics & Data Analysis, v. 27, n. 2, p. 151 – 170, 1998.

HOUGAARD, P. Fundamentals of survival data. **Biometrics**, Wiley Online Library, v. 55, n. 1, p. 13–22, 1999.

KAPLAN, E. L; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

LAPLACE, PS de. Mémoire sur les suites récurro-récurrentes et sur leurs usages dans la théorie des hasards. **Mém. Acad. Roy. Sci. Paris**, v. 6, p. 353–371, 1774.

LAWLESS, J.F. **Statistical Models and Methods for Lifetime Data**. John wiley and sons, new york. [S.l.], 1982.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. 2nd ed.. ed. New Jersey: John Wiley & Sons, 2003.

LEEMIS, L. M.; MCQUESTON, J. T. Univariate distribution relationships. **The American Statistician**, Taylor & Francis, v. 62, n. 1, p. 45–53, 2008.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models (Monographs on statistics and applied probability 37)**. [S.l.: s.n.], 1989.

MORAL, R. A. **Modelagem estatística e ecológica de relações tróficas em pragas e inimigos naturais**. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2013. 173 p.

NELDER, J.; WEDDERBURN, R. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, v. 135, n. 3, p. 370–384, 1972.

OLIVEIRA., F. Q.; B., JACINTO L.; MALAQUIAS, J. B.; A., DALVA; O., Robério. Determination of the median lethal concentration lc_{50} of mycoinsecticides for the control of *ceratitis capitata* (diptera: Tephritidae). **Revista Colombiana de Entomología**, Sociedad Colombiana de Entomología, v. 36, n. 2, p. 213–216, 2010.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2013.

PEREIRA, A. C. **Odontologia em saúde coletiva: planejando ações e promovendo saúde**. [S.l.]: Artmed Editora, 2009.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2013. Disponível em: <<http://www.R-project.org/>>.

THERNEU, T. **A package for survival analysis in S**. Disponível em: <<https://CRAN.R-project.org/package=survival>>, Acesso em: 27 abril, 2016.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. Fourth. New York: Springer, 2002. ISBN 0-387-95457-0. Disponível em: <<http://www.stats.ox.ac.uk/pub/MASS4>>.

WATANABE, A. H. **Comparações de populações discretas**. 1959. Tese (Doutorado) — Universidade de São Paulo, 1959.

WEIBULL, W. Statistical theory of strength of materials. **IVB-Handl.**, v. 151, 1939.

YEE, T. W. **Vector Generalized Linear and Additive Models**. R package version 0.9–1., 2013. Disponível em <<http://CRAN.R-project.org/package=VGAM>>.

ZEILEIS, A.; KLEIBER, C.; JACKMAN, S. **Regression models for count data in R**. Journal of Statistical Software, Los Angeles, 2008. 1–25 p.

Apêndice A

Linhas de comando software R

```
rm(list=ls(all=TRUE))
(dados=read.csv2("sem_zero.csv", header = T))
dados$trat=as.factor(dados$trat)

#===== Análise Descritiva =====

Descritiva <- function(x){
  Media <- mean(x,na.rm = TRUE)
  Variancia <- var(x,na.rm = TRUE)
  Minimo <- min(x,na.rm = TRUE)
  Maximo <- max(x,na.rm = TRUE)
  Mediana <- median(x,na.rm = TRUE)
  saida <- data.frame(Minimo=Minimo,Media=Media,
  Mediana=Mediana, Variancia=Variancia,Maximo=Maximo)
  return (saida)
}

tapply(dados$y, dados$trat, Descritiva)
tapply(dados$y, dados$trat, Descritiva)

# Gráfico Boxplot

par(mfrow=c(1,1))
boxplot(y, cex=2, pch=16, main = "(a)", cex.main=1.9,
ylim = c(0,60), ylab = 'Tempo de ataque (minutos)',cex.lab=1.5,cex.axis=1.5, col = 'grey'
, points = T, names=c("TF", "TM", "JF", "JM"),xlab="")
identify(dados1$tempo_ataque_dia_min~dados1$trat, n=3)

boxplot(dados2$tempo_ataque_noite_min~dados2$trat, cex=2, pch=16, main = " (b)", cex.main = 1.9,
ylim = c(0,60), ylab = 'Tempo de ataque (minutos)' ,cex.lab=1.5,cex.axis=1.5, col = 'grey'
, points = T,names=c("TF", "TM", "JF", "JM"),xlab=" ")

identify(dados2$tempo_ataque_noite_min~dados2$trat, n=2)

#===== Modelos_GLM =====

M2 <- glm(y ~ inseto*sexo*turno, family = Gamma(link="identity"), data = dados)
M3 <- glm(y ~ inseto*sexo*turno, family = Gamma(link="log"), data = dados)
M4 <- glm(y ~ inseto*turno, family = Gamma(link="inverse"), data = dados)
M5 <- glm(y ~ inseto*sexo*turno, family = inverse.gaussian(link = "1/mu^2"),
data = dados)

# half normal-plot

require(hnp)
hnp(M2,main = "(a)", cex.main=2.5,ylab = 'Resíduos (deviance)',
xlab='Quantis teóricos', col = 'grey',cex.lab=1.5)

hnp(M3, main = "(b)", cex.main=2.5,ylab = 'Resíduos (deviance)',
```

```

xlab='Quantis teóricos', col = 'grey',cex.lab=1.5)
hnp(M4,main = "(c)", cex.main=2.5,ylab = 'Resíduos (deviance)',
xlab='Quantis teóricos', col = 'grey',cex.lab=1.5)
hnp(M5, main = "(d)", cex.main=2.5,ylab = 'Resíduos (deviance)',
xlab='Quantis teóricos', col = 'grey',cex.lab=1.5)

M6 <- glm(y ~inseto+sexo+turno+inseto*sexo+inseto*turno+sexo*turno+inseto*sexo*turno
, family = inverse.gaussian(link = "1/mu^2"), data = dados)
anova(M6, test = "F")

M8 <- glm(y ~ inseto*turno, family = inverse.gaussian(link = "1/mu^2"),
data = dados)

hnp(M8, cex.main=2.5,ylab = 'Resíduos (deviance)', xlab='Quantis teóricos',
pch=16,cex.lab=1.5)

plot(fitted.values(M8), residuals(M8), ylim = c(-2,2), pch=16,
xlab = "Valores ajustados", ylab = "Componente do desvio",cex.lab=1.5)
abline(h=2,lty = 2, lwd = 2, col="red")
abline(h=-2, lty = 2, lwd = 2, col="red")

summary(M8)

#===== Análise de sobrevivência =====

# Lendo o banco de dados

rm(list=ls(all=TRUE))
(dados=read.csv2("Experimento_4_primeiros_tratamentos.csv" , header = T))
dados$trat=as.factor(dados$trat)

# Pacotes necessários.

require(survival)
require(flexsurv)

# Gráfico Kaplan-Meier.

## Para os fatores inseto e sexo para o turno diurno.

par(mfrow=c(1,2))
ekm_d <-survfit(Surv(tempo_ataque_dia_min, censura)~trat, data= dados)
plot(ekm_d, col = c(1:4), conf.int = F, ylab="S(t) estimada", xlab="Tempos (minutos)",
mark.time = T, main="Diurno")
legend("top",c("TF","TM","JF","JM"), col = c(1:4),
horiz = F, cex=0.5 ,lwd = 2, bty = 'n')

ekm_n <-survfit(Surv(tempo_ataque_noite_min, censura.1)~trat, data= dados)
plot(ekm_n, col = c(1:4), conf.int = F, ylab="S(t) estimada", xlab="Tempos (minutos)",
mark.time = T, main="Noturno")
legend("top",c("TF","TM","JF","JM"), col = c(1:4),
horiz = F, cex=0.5 ,lwd = 2, bty = 'n')

# Aplicação do teste log-rank - Predador tesourinha
# Período diurno.

```

```

survdif(Surv(tempo_ataque_dia_min[1:42], censura[1:42])~trat[1:42], rho = 0, data = dados)

# Período noturno.

survdif(Surv(tempo_ataque_noite_min[1:42], censura.1[1:42])~trat[1:42], rho = 0, data = dados)

# Aplicação do teste log-rank - Predador joaninha
# Período diurno.

survdif(Surv(c(tempo_ataque_dia_min[43:63], tempo_ataque_dia_min[64:84]),
c(censura[43:63], censura[64:84]))~c(trat[43:63],
trat[64:84]), rho = 0, data = dados)

# Período noturno.

survdif(Surv(c(tempo_ataque_noite_min[43:63], tempo_ataque_noite_min[64:84]),
c(censura.1[43:63], censura.1[64:84]))~c(trat[43:63], trat[64:84]),rho = 0, data = dados)

# Modelos paramétricos : Exponencial, Weibull e Log-normal.

M1 <- survreg(Surv(y,censura)~inseto*sexo*turno, dist = "exponential", data = dados1)
M2 <- survreg(Surv(y,censura)~inseto*sexo*turno, dist = "weibull", data = dados1)
M3 <- survreg(Surv(y,censura)~inseto*sexo*turno, dist = "lognorm", data = dados1)

# Análise dos resíduos de Cox-Snell

## Distribuição exponencial

xb_exp <- M1$coefficients[1]+M1$coefficients[2]*dados1$inseto+
M1$coefficients[3]*dados1$sexo+M1$coefficients[4]*dados1$turno+
M1$coefficients[5]*dados1$inseto*dados1$sexo+M1$coefficients[6]*dados1$inseto*dados1$turno+
M1$coefficients[7]*dados1$sexo*dados1$turno+M1$coefficients[8]*dados1$inseto*dados1$sexo*dados1$turno

ei_exp <- dados1$y*exp(xb_exp)

### Distribuição Weibull

xb_w <- M2$coefficients[1]+M2$coefficients[2]*dados1$inseto+
M2$coefficients[3]*dados1$sexo+M2$coefficients[4]*dados1$turno+
M2$coefficients[5]*dados1$inseto*dados1$sexo+M2$coefficients[6]*dados1$inseto*dados1$turno+
M2$coefficients[7]*dados1$sexo*dados1$turno+M2$coefficients[8]*dados1$inseto*dados1$sexo*dados1$turno
P3 <- 1/M2$scale

ei_w <- (dados1$y*exp(xb_exp))^P3

### Distribuição log-normal

xb_l <- M3$coefficients[1]+M3$coefficients[2]*dados1$inseto+
M3$coefficients[3]*dados1$sexo+M3$coefficients[4]*dados1$turno+
M3$coefficients[5]*dados1$inseto*dados1$sexo+M3$coefficients[6]*dados1$inseto*dados1$turno+
M3$coefficients[7]*dados1$sexo*dados1$turno+M3$coefficients[8]*dados1$inseto*dados1$sexo*dados1$turno

sigma <- M3$scale
res <- (log(dados1$y)-(xb_l))/sigma
ei_l <- -log(1-pnorm(res))

# Teste da razão de verossimilhança

```

```

M4 <- flexsurvreg(Surv(y, censura)~inseto*sexo*turno, dist = "gengamma")
log_LM4 <- M4$loglik      # Valor do logaritmo da função de verossimilhança do modelo M4.
log_LM1 <- M1$loglik[2]   # Valor do logaritmo da função de verossimilhança do modelo M1.
log_LM2 <- M2$loglik[2]   # Valor do logaritmo da função de verossimilhança do modelo M2.
log_LM3 <- M3$loglik[1]   # Valor do logaritmo da função de verossimilhança do modelo M3.

- Cálculo do p-valor

1 - pchisq(38.16, 2)
1 - pchisq(14.66, 1)
1 - pchisq(2.1, 1)

summary(M3)

# Comandos para gerar os gráficos apresentados nas secções 3.3 e 3.2.1.

#===== Gamma =====

set.seed(123); dados<- rexp(100)

densidade <- function(x, mu, v){
fx <- (1/gamma(v)) * ((v*x)/mu)^v * exp(-(v*x)/mu))* (1/x)
return(fx)
}

densidade(x=dados, mu = 0.3,v = 2)

# Para v = 1

par(mfrow=c(1,1), cex=1.5)
curve(densidade(x,mu=0.5,v=1), 0,8,col=1,ylim=c(0,2),main= "(a)",
ylab = expression("f"("y",mu,v)),xlab = "y"
)
curve(densidade(x,mu=1,v =1), 0,8,col=2,add=TRUE)
curve(densidade(x,mu=2,v=1), 0,8,col=3,add=TRUE)
curve(densidade(x,mu=2.5,v=1), 0,8,col=4,add=TRUE)
legend("topright", legend=c(expression(mu==0.5),
expression(mu==1.0),
expression(mu==2.0),
expression(mu==2.5)),col=1:4,lwd=1)

# 0<v<1

mu<- 1;v<- 0.5
curve(densidade(x,mu,v), 0, 3, ylim=c(0,3))
par(mfrow=c(1,1),cex=1.5)

curve(densidade(x,mu=2,v=0.1), 0,4,col=1,ylim=c(0,2),
main= "(b)", cex.main = 1.9,
ylab=expression("f"("y",mu,v)), xlab = "y")
curve(densidade(x,mu=2,v =0.4), col=2,add=TRUE)
curve(densidade(x,mu=2,v=0.6), col=3,add=TRUE)
curve(densidade(x,mu=2,v=0.9), col=4,add=TRUE)
legend("topright", legend=c(expression(v==0.1),
expression(v==0.4),
expression(v==0.6),
expression(v==0.9)),col=1:4,lwd=1)

```

```

# v>1

par(mfrow=c(1,1),cex=1.5)
curve(densidade(x,mu=2,v=100), 0,10,col=1,ylim=c(0,2),main= "(c)",
cex.main = 1.9,ylab=expression("f"("y",mu,v)), xlab = "y")
curve(densidade(x,mu=1,v=1), 0,10,col=2,add=TRUE)
curve(densidade(x,mu=1,v=2.0), 0,10,col=3,add=TRUE)
curve(densidade(x,mu=1,v=2.5), 0,10,col=4,add=TRUE)
legend("topright", legend=c(expression(v==0.5),
expression(v==1.0),
expression(v==2.0),
expression(v==2.5)),col=1:4,lwd=1)

par(mfrow=c(1,1),cex=1.5)
curve(densidade(x,mu=0.5,v=3), 0,1.5,col=1,ylim=c(0,8),
main= "(c)", cex.main = 1.9,
ylab=expression("f"("y",mu,v)), xlab = "y")
curve(densidade(x,mu=0.5,v =6), col=2,add=TRUE)
curve(densidade(x,mu=0.5,v=50), col=3,add=TRUE)
curve(densidade(x,mu=0.5,v=100), col=4,add=TRUE)
legend("topright", legend=c(expression(v==3),
expression(v==6),
expression(v==50),
expression(v==100)),
col=1:4,lwd=1)
#-----

# para v = k/2

par(mfrow=c(1,1),cex=1.5)
curve(densidade(x,mu=1,v=0.5), 0,8,col=1,ylim=c(0,0.5),
main= "(d)", cex.main = 1.9,
ylab=expression("f"("y",mu,v)), xlab = "y")
curve(densidade(x,mu=2,v =1), col=2,add=TRUE)
curve(densidade(x,mu=3,v=1.5), col=3,add=TRUE)
curve(densidade(x,mu=5,v=2.5), col=4,add=TRUE)
legend("topright", legend=c(expression(mu==1),
expression(mu==2),
expression(mu==3),
expression(mu==5)),
col=1:4,lwd=1)

#===== Exponencial =====
# Função de densidade
par(mfrow=c(1,3))
x <- seq(0,6,length.out = 100)

funcao_dens <- function(x,lambda){
funcao <- lambda*exp(-lambda*x)
return(funcao)

}

curve(funcao_dens(x,1), 0,6, col="blue", xlab = "t",ylab = " ", cex.main = 2.5, main="(a)",
cex.lab=1.5,cex.axis=1.5)
curve(funcao_dens(x,0.7), 0,6, col="red", add = T )

```

```

curve(funcao_dens(x,0.5),0,6, col="green", add=T)

# Função sobrevivência

t <- seq(0,6,length.out = 100)

funcao_sobre <- function(t,lambda){
funcao <- exp(-(t/lambda))
return(funcao)
}

plot(function(t) funcao_sobre(t,1),0,6, col= "blue", ylab = " ", xlab="t",
cex.lab=1.5,cex.axis=1.5, cex.main = 2.5, main="(b)")
plot(function(t) funcao_sobre(t,0.7),0,6, col="red",add = T )
plot(function(t) funcao_sobre(t,0.5),0,6, col="green",add = T )

# Função taxa de falha

x <- 1:9
y <- 1:9
plot(x, y, type = "n",xlim=c(0,6),ylim=c(0.5, 2.5),cex.lab=1.5,cex.axis=1.5,
ylab = " ",cex.main = 2.5, main="(c)", xlab="t")
lines(c(0, 6), c(1.0, 1.0), col= "blue") # Para lambda igual a 1.
lines(c(0, 6), c(1.5, 1.5), col= "red") # Para lambda igual a 0.7.
lines(c(0, 6), c(2.0, 2.0), col= "green")

#===== Weibull =====
# Função de densidade

par(mfrow=c(1,3))

curve(dweibull(x, scale = 250, shape = 3), from = 0, to = 800,
col="blue", xlab = "t", ylab = " ",cex.main = 2.5, main="(a)",
cex.lab=1.5,cex.axis=1.5)

curve(dweibull(x, scale = 350, shape = 4), from = 0, to = 800,
col="red", xlab = "t", ylab = "f(t)",
cex.lab=1.5,cex.axis=1.5, add = T)

curve(dweibull(x, scale = 600, shape = 8), from = 0, to = 800,
col="green", xlab = "t", ylab = "f(t)",
cex.lab=1.5,cex.axis=1.5, add = T)

curve(dweibull(x, scale = 150, shape = 1), from = 0, to = 800,
col="gray", xlab = "t", ylab = "f(t)",
cex.lab=1.5,cex.axis=1.5, add = T)

curve(dweibull(x, scale = 050, shape = 0.5), from = 0, to = 800,
col="black", xlab = "t", ylab = "f(t)",
cex.lab=1.5,cex.axis=1.5, add = T)

# Função de Sobrevivência

curve(pweibull(x, shape=3, scale=250, lower.tail=False),0,800,
col= "blue", ylab = " ",cex.main = 2.5, main="(b)", xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(pweibull(x, shape=4, scale=350, lower.tail=False),0,800,add = T,

```

```

col= "red", ylab = "S(t)", xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(pweibull(x, shape=8, scale=600, lower.tail=False),0,800,add = T,
col= "green", ylab = "S(t)", xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(pweibull(x, shape=1, scale=150, lower.tail=False),0,800,add = T,
col= "gray", ylab = "S(t)", xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(pweibull(x, shape=0.5, scale=50, lower.tail=False),0,800,add = T,
col= "black", ylab = "S(t)", xlab="t",cex.lab=1.5,cex.axis=1.5)

# Função taxa de falha

funcao_taxa <- function(x,alpha, gama){
funcao <- (gama/alpha^gama)*x^(gama-1)
return(funcao)
}

curve(funcao_taxa(x, alpha=250,gama=3),0,800,
col= "blue", ylab = " ",cex.main = 2.5, main="(c)", xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(funcao_taxa(x, alpha=350,gama=4),0,800,add = T,
col= "red",ylab = expression(lambda("t")), xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(funcao_taxa(x, alpha=600,gama=8),0,800,add = T,
col= "green", ylab = expression(lambda("t")), xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(funcao_taxa(x, alpha=150,gama=1),0,800,add = T,
col= "gray", ylab = expression(lambda("t")), xlab="t",cex.lab=1.5,cex.axis=1.5)

curve(funcao_taxa(x, alpha=50,gama=0.5),0,800,add = T,
col= "black", ylab = expression(lambda("t")), xlab="t",cex.lab=1.5,cex.axis=1.5)

```