Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Allana Lívia Beserra Paulino

Ajuste via modelos lineares generalizados para avaliação do controle biológico de insetos

Campina Grande
Dezembro de 2012

Allana Lívia Beserra Paulino

Ajuste via modelos lineares generalizados para avaliação do controle biológico de insetos

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador:
Ricardo Alves de Olinda

Campina Grande
Dezembro de 2012

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL – UEPB

P328a Paulino, Allana Lívia Beserra.
Ajuste via modelos lineares generalizados para avaliação do controle biológico de insetos[manuscrito] / Allana Lívia Beserra Paulino. – 2012.
38f..: il. color.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2012.
"Orientação: Prof. Dr. Ricardo Alves de Olinda, Departamento de Estatística".

1. Estatística. 2. Modelos Lineares Generalizados. 3. Controle Biológico de Insetos. I. Título.

21. ed. CDD 519.5

Allana Lívia Beserra Paulino

Ajuste via modelos lineares generalizados para avaliação do controle biológico de insetos

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovado em: 14 / 12 / 2012

Banca Examinadora:

Ricardo Alves de Olinda
Prof. Dr. Ricardo Alves de Olinda
Orientador

Prof. Dr. João Gil de Luna
Universidade Estadual da Paraíba

Prof. Msc. Juarez Fernandes de Oliveira
Universidade Estadual da Paraíba

Dedicatória

Dedico este trabalho aos meus pais Antônio Paulino da Silva e Lúcia de Fátima Beserra Paulino que sempre estiveram ao meu lado nessa caminhada, me apoiando e encorajando a superar as dificuldades.

Agradecimentos

Agradeço a Deus por tudo que tem feito em minha vida, pela minha família, amigos e colegas e minha saúde.

Agradeço a Minha Mãe Nossa Senhora de Fátima pelos pedidos concedidos, milagres e curas.

Agradeço as meus pais Antônio Paulino e Lúcia de Fátima pelo apoio, confiança e carinho sempre.

Agradeço aos meus irmãos, Lúcio Flávio, Alisson Lamarque, Allen Luciani e Allan Pedro, por seus incentivos e nunca deixaram que eu disistisse dos meus objetivos, sonhos e aguentaram os meus estresses.

A minha avó Maria por sua alegria contagiosa e sempre muito amorosa e minha tia Fatinha que me apoiou e me transmitiu muita coragem para enfrentar os obstáculos.

Aos meu amigos de graduação André Luiz, Bárbara Camboim que sempre que eu dizia que tinha pago uma cadeira ela dizia "muito bem Allana, ta virando uma mocinha!", Caroline Gonçalves pelos estudos em sua casa, Djair Durand, Janeide Alves que me aguentou durante o curso, Jaiane Silva pelas palavras de apoio, Moisés Moureira, Tamyres Aline que estiveram comigo e que de uma forma ou outra me ajudaram durante o curso e aos meninos Fábio Sandro e Rosendo Chagas pelas momentos de alegrias e descontração.

Agradeço em especial a Sidcleide Barbosa e Priscilla Cabral que estudaram comigo e me incetivaram nos momentos difíceis e que me aperriaram muito para eu estudar, a Diego Alves que sempre acreditou em mim e sempre tem uma palavra de carinho e conforto.

As minhas amigas Dayse Santos, Daniela Sampaio e Jocasta Moura pela compreensão nos momentos que estive ausente por estar estudando e incentivo durante meu curso.

Ao professor e orientador Ricardo Alves de Olinda pela *paciência*, incentivo e dedicação durante o trabalho de conclusão de curso.

E a todos os professores da UEPB que me ajudaram e contribuiram na minha vida acadêmica.

Resumo

Os Modelos Lineares Generalizados (MLG) foram introduzidos no início dos anos 70 como uma maneira de unificar vários modelos estatísticos, tendo um impacto significativo no desenvolvimento da estatística aplicada. Isto permitiu desenvolver um algoritmo geral para a estimativa de máxima verossimilhança em vários modelos. Nos MLG pode-se relacionar a distribuição aleatória da variável dependente no experimento (a função de distribuição) com a parte sistemática (não aleatória) (ou preditor linear) por meio de uma função chamada função de ligação. O uso de modelos lineares clássicos, em alguns casos, não é apropriado para analisar dados de proporções, que são muito frequentes em entomologia, pois as pressuposições do modelo não são atendidas. Uma alternativa para a análise desse tipo de dados é a utilização da teoria de modelos lineares generalizados, sendo a distribuição binomial, um caso particular, indicada para essas situações. O presente trabalho objetivou ajustar uma distribuição de probabilidade aos dados de um ensaio biológico com insetos via modelos lineares generalizados; comparar qual função de ligação melhor se ajusta aos dados por meio do critério de informação de Akaike (AIC) e por fim, verificar a eficiência dos extratos vegetais no controle biológico de insetos. Os dados foram disponibilizados pelo departamento de Plantas e Inseticidas do Departamento de Entomologia e Acarologia, da Escola Superior de Agricultura "Luiz de Queiroz"(ESALQ/USP). Após o ajuste da distribuição de probabilidade, observou-se que a função de ligação complemento log-log foi a mais adequada para se ajustar aos dados em questão, destacando-se alguns níveis do fator em estudo.

Palavras-chaves: Preditor Linear, Função de Ligação, Bioensaios.

Abstract

The Generalized Linear Models (GLM) were introduced at the beginning 70 as a way to unify various statistical models, having a significant impact on the development of statistical applied. This allowed the development of a general algorithm for estimating maximum likelihood in several models. In MLG can relate the random distribution of the dependent variable in the experiment (the function distribution) with the part systematic (non-random) (or predictor linear) through a function call connection function. The use of classical linear models, in general, is not suitable for analyzing data proportions, which are very common in agronomy because the assumptions of the model are not met. An alternative to the analysis of such data is the use of the theory of linear models generalized binomial distribution is a special case, indicated for these situations. This study aimed to set a probability distribution data of a bioassay insects via generalized linear models, which compare function link best fits the data through the information criterion Akaike (AIC) and finally, to verify the efficiency of the plant extracts biological control of insects. The data were provided by the Plants Insecticides department and the Department of Entomology and Acarology, the Escola Superior de Agricultura Luiz de Queiroz " "(ESALQ / USP). After adjustment of the probability distribution, it was observed the complement binding function log-log was most suitable for adjust the data in question, highlighting some factor levels in study.

Key-words:Linear Predictor, Function Liaison, Bioassays

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A importância dos Modelos Lineares Generalizados não é apenas de índole prática. Do ponto de vista teórico a sua importância advém, essencialmente, do fato de a metodologia destes modelos constituir uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações e promover o papel central da verossimilhança na teoria da inferência(TURKMAN; SILVA, 2000). Nelder e Wedderburn (1972) propuseram uma teoria unificadora da modelagem estatística a que deram o nome de Modelos Lineares Generalizados (MLG), como uma extensão dos modelos lineares clássicos. Na realidade, eles mostraram que uma série de técnicas comumente estudadas separadamente podem ser reunidas sob o nome de Modelos Lineares Generalizados. Os desenvolvimentos que levaram a esta visão geral da modelagem estatística, remontam a mais de um século. Um breve histórico (MCCULLAGH; NELDER, 1989; LINDSEY, 1997) pode ser traçado:

i) Regressão linear múltipla, envolvendo distribuição normal (Legendre, Gauss, início do século XIX);

ii) Análise de variância para delineamentos experimentais, envolvendo distribuição normal (FISHER, 1920 a 1935);

iii) Função de verossimilhança, um procedimento geral para inferência a respeito de qualquer modelo estatístico (FISHER, 1922);

iv) Modelo complemento *log-log* para ensaios de diluição, envolvendo distribuição binomial (FISHER, 1922);

v) Família exponencial, uma classe de distribuições com propriedades "ótimas"(estatísticas suficientes) para a estimação dos parâmetros (FISHER, 1934);

vi) Modelo *probit* para proporções, envolvendo distribuição binomial (BLISS, 1935);

vii) Modelo logístico para proporções, envolvendo distribuição binomial (BERKSON, 1944; DYKE;PATTERSON, 1952);

viii) Modelo logístico para análise de itens, envolvendo distribuição Bernoulli (RASCH, 1960);

ix) Modelos log-lineares para contagens, envolvendo distribuição poisson e multinomial (BIRCH, 1963);

x) Modelos de regressão para dados de sobrevivência, envolvendo distribuição exponencial (FEIGL; ZELEN, 1965; ZIPPIN; ARMITAGE, 1966; GASSER, 1967);

Segundo Demétrio (2002), *apud* Nelder e Wedderburn (1972) mostraram, então, que a maioria dos problemas estatísticos, que surgem nas áreas de agricultura, demografia, ecologia, economia, geografia, geologia, história, medicina, ciência política, psicologia, sociologia, zootecnia etc, podem ser formulados, de uma maneira unificada, como modelos de regressão. Esses modelos envolvem uma variável resposta univariada, variáveis explicativas e uma amostra aleatória de n observações, sendo que:

i) A variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família exponencial na forma canônica (distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; poisson e binomial negativa para contagens);

ii) As variáveis explicativas, entram na forma de um modelo linear (componente sistemático);

iii) A ligação entre os componentes aleatório e sistemático é feita por meio de uma função (por exemplo, logarítmica para os modelos log-lineares).

Conforme Turkman e Silva (2000), devido ao grande número de modelos que englobam e a facilidade de análise associada ao rápido desenvolvimento computacional que se tem verificado nas últimas décadas, os MLG têm vindo a desempenhar um papel cada vez mais importante na análise estatística, apesar das limitações ainda impostas, nomeadamente por manterem a estrutura de linearidade, pelo fato das distribuições se restringirem à família exponencial e por exigirem a independência das respostas. Já existe atualmente, na literatura, muitos desenvolvimentos da teoria no que se refere a modelagem estatística onde estes pressupostos são relaxados, mas, o não acompanhamento dos modelos propostos com software adequado à sua fácil implementação, faz com que se anteveja ainda, por algum tempo, um domínio dos MLG em aplicações de natureza prática.

Diante do exposto, o presente trabalho objetivou ajustar uma distribuição de probabilidade aos dados de um ensaio biológico com insetos via Modelos Lineares Generalizados; comparar qual função de ligação melhor se ajusta aos dados por meio do critério de informação de Akaike e por fim; verificar a eficiência dos extratos vegetais (Tratamentos) no controle biológico de insetos.

2 Fundamentação Teórica

Nessa seção será estudado os principais aspectos teóricos e práticos que servirão de base para o ajuste de distribuições de probabilidade para dados binários e dados de contagem via Modelos Lineares Generalizados.

2.1 Família Exponencial

Conforme Cordeiro e Demétrio (2007) o conceito de família exponencial foi introduzido na estatística por Fisher, mas os modelos da família exponencial apareceram na mecânica estatística no final do século XIX e foram desenvolvidos por Maxwell, Boltzmann e Gibbs. A importância da família exponencial de distribuições teve maior destaque, na área dos modelos de regressão, a partir do trabalho pioneiro de Nelder e Wedderburn (1972) que definiram os MLG. Na década de 80, esses modelos popularizaram-se, inicialmente, no Reino Unido, e, posteriormente, nos Estados Unidos e na Europa.

De acordo com Ehlers (2009) a família exponencial inclui muitas das distribuições de probabilidade mais comumente utilizadas em estatística, tanto continuo quanto discretas. Uma característica essencial desta família é que existe uma estatística suficiente com dimensão fixa.

Conforme Turkmam e Silva (2000) uma variável aleatória Y tem distribuição pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder ser escrita como

$$f(y|\boldsymbol{\theta},\phi) \;=\; \exp\left\{\frac{y\boldsymbol{\theta}-b(\boldsymbol{\theta})}{a(\boldsymbol{\theta})}+c(y,\phi)\right\}, \tag{2.1}$$

em que θ e ϕ são parâmetros, a(.), b(.) e c(. ; .) são funções reais conhecidas.

Varias distribuições importantes podem ser escritas na forma (2.1), tais como: poisson,

binomial, rayleigh, normal, gama e normal inversa (as três últimas com a suposição de que um dos parâmetros é conhecido).

2.1.1 Valor médio e a variância

Conforme Turkmam e Silva (2000) seja o logaritmo da função de verossimilhança $\ell(\theta;\phi,y) = \ln(f(y|\theta,\phi))$, define-se a função Score de derivadas parciais em relação aos seus respectivos parâmetros da seguinte forma,

$$S(\theta) = \frac{\partial \ell(\theta,\phi;y)}{\partial \theta}. \tag{2.2}$$

Sabe-se que para famílias regulares, tem-se que

$$E[S(\theta)] = 0$$

$$E[S^2(\theta)] = E\left[\left(\frac{\partial \ell(\theta,\phi;y)^2}{\partial \theta}\right)\right] = -E\left[\frac{\partial^2 \ell(\theta,\phi;y)}{\partial \theta^2}\right], \tag{2.3}$$

e portanto como, no caso em que $f(y|\theta,\phi)$ é definido por (2.1),

$$\ell(\theta,\phi;y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi),$$

obtém-se

$$S(\theta) = \frac{Y - b'(\theta)}{a(\phi)} \times \frac{\partial S(\theta)}{\partial \theta} = \frac{b''(\theta)}{a(\phi)}, \tag{2.4}$$

em que

$$b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$$

$$e$$

$$b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}.$$

Assim de (2.3) e (2.4)

$$E(Y) = \mu = a(\phi)E[S(\theta) + b'(\theta)] = b'(\theta) \tag{2.5}$$

$$e$$

$$VAR(Y) = a^2(\phi)VAR[S(\theta)] = a^2(\phi)\frac{b''(\theta)}{a(\phi)} = a(\phi)b''(\theta). \tag{2.6}$$

Segundo Turkman e Silva (2000) a variância de Y é o produto de duas funções, uma, $b'(\theta)$, que depende apenas do parâmetro canônico θ (e, portanto, é o valor médio de μ), a que se dá o nome de função de variância e que se costuma representar por $V(\mu)$ e outra, $a(\phi)$, que depende apenas do parâmetro de dispersão ϕ. Em muitas situações de interesse, observa-se que a função $a(\phi)$ toma a forma $a(\phi) = \frac{\phi}{m}$ em que m é uma constante conhecida, obtendo-se, portanto a variância de Y como o produto do parâmetro de dispersão por uma função apenas do valor médio.

Neste caso a função definida em (2.1) pode ser escrita da seguinte forma

$$f(y|\theta, \phi, m) = \exp\left\{\frac{m}{\phi}(y\theta - b(\theta)) + c(y, \phi, m)\right\}. \tag{2.7}$$

2.1.2 Exemplo da distribuição Binomial

Segundo Sounis (1985) a distribuição binomial é uma das distribuições de probabilidade de utilização mais frequente em estatistica aplicada a biologia. É usada sobre tudo quando os dados se apresentam em duas classes (dicotomizados) em duas categorias discretas e a pesquisa se refere a uma amosta.

Segundo Turkmam e Silva (2000), se Y for tal que, mY segue um distribuição binomial e distribuição normal com parâmetros m e π ($Y \sim B(m, \pi)/m$), a sua f.d.p. é definida por

$$f(y|\pi) = \binom{m}{ym}\pi^{ym}(1-\pi)^{m-ym} \quad I^{(y)}_{0,1,\ldots,n}$$

Aplicando-se a propriedade da exponencial $a^{b-c} = \frac{a^b}{a^c}$, tem-se

$$
\begin{aligned}
&= \binom{m}{ym}\pi^{ym}\frac{(1-\pi)^m}{(1-\pi)^{ym}} \\
&= \binom{m}{ym}\pi^{ym}\frac{1}{(1-\pi)^{ym}}(1-\pi)^m \\
&= \binom{m}{ym}\frac{\pi^{ym}}{(1-\pi)^{ym}}(1-\pi)^m \\
&= \binom{m}{ym}(1-\pi)^m\frac{\pi^{ym}}{(1-\pi)^{ym}}.
\end{aligned}
$$

Assim, pela propriedade $\frac{a^c}{b^c} = (a/b)^c$, tem-se que

$$
= \binom{m}{ym}(1-\pi)^m\left(\frac{\pi}{1-\pi}\right)^{ym}.
$$

Organizando-se na forma da família exponencial, tem-se que

$$
f(y|\pi) = \exp\{ym\ln\left(\frac{\pi}{1-\pi}\right) + m\ln(1-\pi) + \ln\binom{m}{ym}\}.
$$

Assim, tem-se que a distribuição de binomial pertence a família exponencial.

2.1.3 Exemplo da distribuição de poisson

Segundo Ross (2010) a distribuição de probabilidade de poisson foi introduzida por Siméon Denis Poisson em um livro que escreveu a respeito da aplicação da teoria de probabilidade a processos, julgamentos criminais e similares. A variável aleatória de poisson encontra-se numa consideravel faixa de aplicações em diversas áreas, pois pode ser usada como uma aproximação para variável aleatória binomial com parâmetros (n, p) no caso particular de n grande e p suficientemente pequeno para que np tenha tamanho moderado.

Seja $X_1, ..., X_n \sim$ poisson(λ), então, segundo Ehlers (2009) a distribuição de poisson pretence a família exponencial se,

$$
\begin{aligned}
P(y|\lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \quad I^{(y)}_{0,1,...,} \\
&= \frac{1}{y!}e^{-\lambda}\lambda^y
\end{aligned}
$$

Assim, aplicando-se a exponencial, tem-se que

$$P(y|\lambda) = \frac{1}{y!}\exp\{\ln(e^{-\lambda}\lambda^{y})\}$$

Organizando-se na forma da família exponencial, tem-se que

$$P(y|\lambda) = \exp\left\{\frac{1}{y!}\lambda + y\ln\lambda\right\}.$$

Assim, conclui-se que a distribuição de poisson pertence a família exponencial.

2.2 Descrição do Modelo Linear Generalizado

Segundo Turkman e Silva (2000) os modelos lineares generalizados são uma extensão do modelo linear clássico definado na equação abaixo

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{Z}$ é uma matriz de dimensão $n \times p$ de especificação do modelo (em geral a matriz de covariáveis $\mathbf{X}$ em que a primeira coluna corresponde a um vetor unitário), associada a um vetor $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ de parâmetros, e $\boldsymbol{\varepsilon}$ é um vetor $n \times 1$ de erros aleatórios com distribuição que se supõe $N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

Segundo Turkman e Silva (2000) estas hipóteses implicam obviamente que $E(\mathbf{Y}|\mathbf{Z}) = \boldsymbol{\mu}$ com $\boldsymbol{\mu} = \mathbf{Z\beta}$, ou seja, o valor esperado da variável resposta é uma função linear das covariáveis. A extensão mencionada é realizada em duas direções. Por um lado, a distribuição considerada não tem de ser normal, podendo ser qualquer distribuição da família exponencial; por outro lado, embora se mantenha a estrutura da linearidade, a função que relaciona o valor esperado e o vetor de covariáveis pode ser qualquer função diferencial.

Assim os MLG são caracterizados pela seguinte estrutura:

i) **Componente aleatório**

Dado o vetor de covariáveis $\mathbf{X}$, as variáveis Y_i são (condicionalmente) independentes com distribuição pertencente à família exponencial da forma (2.1) ou (2.7), com

$E(Y_i, X_i) = \mu_i = b'(\theta_i)$ para $i = 1,...,n$, e possivelmente, um parâmetro de dispersão ϕ que não depende de i.

ii)**Componente estrutural ou sistemática**

O valor esperado μ_i está relacionado com o preditor linear $\eta_i = \mathbf{Z}_i^T\boldsymbol{\beta}$ por meio da relação $\mu_i = h(\eta_i) = h(\mathbf{Z}_i^T\boldsymbol{\beta})$, $\eta_i = g(\mu_i)$, em que h é uma função monótona e diferenciável, $g = h^{-1}$ é a função de ligação, $\boldsymbol{\beta}$ é um vetor de parâmetros de dimensão $p\times 1$, $\mathbf{Z}_i$ é um vetor de especificação de dimensão $p \times 1$, ou seja, é uma função das covariáveis x_i, $i = 1, 2, ...$

2.3 Modelos para dados Binários

De acordo com Turkman e Silva (2000), suponha-se que tem-se n variáveis respostas independentes $Y_i \sim B(1, \pi_i)$, então a distribuição de Y_i é,

$$f(y_i|\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \qquad y_i = 0, 1$$

e que a cada indivíduo i ou unidade experimental, está associado um vetor de especificação $\mathbf{Z}_i$, resultante do vetor de covariáveis $\mathbf{x}_i$, $i = 1, ..., n$.

Como $E(Y_i) = \pi_i$ e, de acordo com alguma distribuição exponencial, se tem para este modelo, $\theta_i = \ln(\frac{\pi_i}{1-\pi_i})$, ao fazer

$$\theta_i \;\; = \;\; \eta_i = \mathbf{Z}_i^T\boldsymbol{\beta},$$

concluí-se que a função de ligação canônica é a função *logit*. Assim a probabilidade de sucesso, ou seja, $P(Y_i = 1) = \pi_i$ está relacionada com o vetor $\mathbf{Z}_i$ por meio de

$$\pi_i \;\; = \;\; \frac{\exp(\mathbf{Z}_i^T\boldsymbol{\beta})}{1 + \exp(\mathbf{Z}_i^T\boldsymbol{\beta})}. \tag{2.8}$$

É fácil ver que a função $F : \mathbb{R} \to [0, 1]$, definida por

$$F(x) \;\; = \;\; \frac{\exp(x)}{1 - \exp(x)},$$

ela é, certamente, a função de distribuição logística.

Por esse motivo, o MLG definido pelo modelo binomial com função de ligação canônica (*logit*) é conhecido por modelo de regressão logística, (TURKMAN; SILVA, 2000). Repare-se que devido ao fato de, neste modelo, se ter $E(Y_i) = \mu_i \in [0, 1]$, em princípio, não só

a função de distribuição logística, como qualquer outra função de distribuição, pode ser candidata a função inversa da função de ligação. Nomeadamente pode-se supor que a relação existente entre as probabilidades de sucesso π_i e o vetor de covariáveis da forma

$$\pi_i \quad = \quad \Phi(\eta_i) = \Phi(Z_i^T \boldsymbol{\beta}), \tag{2.9}$$

em que $\Phi(.)$ é a função de distribuição de uma variável aleatória N(0,1). Obtém-se assim uma função de ligação $g(\mu_i) = \Phi^{-1}(\mu_i)$ designada por uma função de ligação *probit*.

Segundo Turkman e Silva (2000) o MLG, obtido pela associação do modelo binomial para as respostas, com a função de ligação *probit* conduz ao modelo de regressão *probit*. Outra função de distribuição que se costuma considerar para candidatar-se a função inversa da função de ligação é a função de distribuição de Gumbel, ou função de distribuição de extremos,

$$F(x) \quad = \quad 1 - \exp(-\exp(x)), \qquad x \in \mathbb{R} \ .$$

Considerando-se então

$$h(\boldsymbol{Z}_i^T \boldsymbol{\beta}) = 1 - \exp(-\exp(\boldsymbol{Z}_i^T \boldsymbol{\beta})) = \pi_i,$$

obtém-se a função complemento $log - log$

$$\ln(-\ln(1 - \pi_i)) \quad = \quad \boldsymbol{Z}_i^T \boldsymbol{\beta} \tag{2.10}$$

para função de ligação.

Segundo Turkman e Silva (2000)o MLG, obtido pela associação do modelo binomial para as respostas, com a função de ligação complemento *log- log* conduz ao modelo de regressão complemento *log- log*. A utilização de uma ou outra função de ligação, e consequentemente, a escolha do modelo de regressão à utilizar depende da situação em causa. Em geral, a adaptabilidade dos modelos *probit* e *logístico* é bastante semelhante, já que as funções correspondentes não se afastam muito uma da outra após um ajustamento adequando dos correspondentes preditores lineares. O modelo complemento *log-log* pode dar respostas diferentes já que a função complemento *log-log*, mesmo após o ajustamento do preditor linear η, se distancia das anteriores, tendo um crescimento mais abrupto (ver, Fahrmeir e Tutz, 1994, pg.27). A função de ligação complemento *log-log* é mais utilizada para análise de dados sobre incidência de doenças.

2.3.1 Sobredispersão ou extra variação binomial

Segundo Turkman e Silva (2000), um fenômeno que ocorre com frequência nas aplicações é as respostas apresentarem uma variância superior à variância explicada pelo modelo binomial. Este fenômeno, denominado de sobredispesão ou extra variação binomial, pode ser devido ao fato de existir heterogeneidade entre os indivíduos não explicado pelas covariáveis, ou pelo fato de haver correlação entre as respostas. Esta última situação acontece quando, por exemplo, as respostas correspondem a indivíduos da mesma família, ou a indivíduos que comungam dos mesmos fatores ambientais, formando-se assim grupos naturais, embora a heterogeneidade não explicada também produza correlação entre as respostas. Este problema pode ser resolvido se introduzir um parâmetro $\phi > 1$ de sobredispersão de tal modo que $VAR[Y_i|x_i] = \phi\frac{\pi_i(1-\pi_i)}{n_i}$, em que $n_i > 1$ é a dimensão do grupo. Nota-se, no entanto, que já não é possível escrever a distribuição de Y_i na forma da família exponencial (2.1). O modelo fica apenas determinado pelo valor médio e variância, (TURKMAN; SILVA, 2000).

2.4 Modelos para resposta na forma de contagem

Conforme Turkman e Silva (2000), dados na forma de contagens aparecem com muita frequência nas aplicações. São exemplos disso número de acidentes, número de chamadas telefônicas, número de elementos numa fila de espera, etc. Também são dados deste tipo as frequências em cada célula de uma tabela de contingência. O modelo de poisson, como se sabe, desempenha um papel fundamental na análise deste tipo de dados. Este é um modelo que pertence à família exponencial que tem a particularidade de o valor médio ser igual à variância. Se considerar que as respostas Y_i são independentes e bem modeladas por uma distribuição de Poisson de valor médio μ_i e que $\ln(\mu_i) = \boldsymbol{Z}_i^T\boldsymbol{\beta}$ com $i = 1, ..., n$, isto é,

$$\begin{aligned} f(y_i|x_i) &= \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \\ &= \frac{1}{y_i!}e^{-\mu_i}\mu_i^{y_i} \\ &= \exp\{\ln\left(\frac{1}{y_i!}e^{-\mu_i}\mu_i^{y_i}\right)\} \\ &= \exp\{-e^{\ln\mu_i} + y_i\ln\mu_i - \ln y_i!\} \\ &= \exp\{-e^{\boldsymbol{Z}_i^T\boldsymbol{\beta}} + y_i\boldsymbol{Z}_i^T\boldsymbol{\beta} - \ln y_i!\}, \quad y_i = 0, 1, ..., \end{aligned}$$

obtém-se um MLG com função de ligação canônica, conhecido por *modelo de regressão de poisson*, ou *modelo log-linear*.

Para o caso do modelo de poisson, a função logarítmica é a função de ligação que geralmente se utiliza, (TURKMAN; SILVA, 2000). Sob condições bastante fracas, pode-se mostrar que a análise de uma tabela de contingência sobre amostragem de poisson, é a mesma que a análise sob amostragem multinomial ou produto-multinomial (CHISTENSEN, 1997). Assim, o modelo de regressão de poisson é também útil na modelagem e estudo de tabelas de contingência multidimensionais, apesar de as observações não serem independentes. A imposição pelo modelo de poisson da variância ser iqual ao valor médio, produz, também com frequência, problemas de sobredispersão idênticos ao referido anteriormente para dados de natureza binaria. O modo mais simples de resolver o problema é, novamente, o de considerar um parâmetro de sobredispersão ϕ de tal modo que $VAR[Y_i|X] = \phi\mu_i$, para $i = 1, ...,$. Há, no entanto, modelos mais complexos que entram em consideração com variação extra nos dados.

2.5 Estimação

Segundo Turkman e Silva (2000) aplicar a metodologia dos modelos lineares generalizados a um conjunto de dados há necessidade, após a formulação do modelo que se pensa adequado, de proceder à realização de inferência sobre esse modelo.

A inferência com MLG é, essencialmente, baseada na verossimilhança. Com efeito, não só o método da máxima verossimilhaça é o método de eleição para estimar os paramêtros de regressão, como também testes de hipóteses sobre os paramêtros do modelo e de qualidade de ajustamento são, em geral, métodos baseados na verossimilhança.

2.5.1 Algoritmo de Estimação

De acordo com Cordeiro e Demétrio (2007) a decisão importante na aplicação dos MLG é a escolha do trinômio: distribuição da variável resposta × matriz modelo × função de ligação. A seleção pode resultar de simples exame dos dados ou de alguma experiência anterior. Inicialmente, considera-se esse trinômio fixo para se obter uma descrição adequada dos dados por meio das estimativas dos parâmetros do modelo. Muitos métodos podem ser usados para estimar os parâmetros $\boldsymbol{\beta}'s$, inclusive o qui-quadrado mínimo, o Bayesiano e a estimação-M. O último inclui o método de máxima verossimilhança (MV) que tem muitas propriedades ótimas, tais como, consistência e eficiência assintótica. As-

sim, considera-se apenas o método de MV para estimar os parâmetros lineares $\beta_1, ..., \beta_p$ do modelo. O vetor escore é formado pelas derivadas parciais de primeira ordem do logaritmo da função de verossimilhança. O logaritmo da função de verossimilhança como função apenas de $\boldsymbol{\beta}$ (considerando-se o parâmetro de dispersão ϕ conhecido) dado o vetor y é definido por $\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}; y)$ e usando-se a expressão (2.7) tem-se

$$\ell(\boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n} c(y_i, \phi), \tag{2.11}$$

em que $\theta_i = q(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$ e $\eta_i = \sum_{i=1}^{n} y_{ir}$. Da expressão (2.11) pode-se calcular, pela regra da cadeia, o vetor escore $U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ de dimensão p, com elemento típico $\boldsymbol{U}_r = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_r} = \sum_{i=1}^{n} \frac{d\mu_i}{d\theta_i} \times \frac{d\theta_i}{d\mu_i} \times \frac{d\mu_i}{d\eta_i} \times \frac{\partial \eta_i}{\partial \beta_r}$, pois

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= f(\theta_1, \theta_2, ..., \theta_i, ..., \theta_n) \\
&\downarrow \\
\theta_i &= \int V_i^{-1} d\mu_i = q(\mu_i) \\
&\downarrow \\
\mu_i &= g^{-1}(\eta_i) = h(\eta_i) \\
&\downarrow \\
\eta_i &= \sum_{r=1}^{p} y_{ir}\beta_r
\end{aligned}$$

e, sabendo-se que $\mu_i = b'(\theta_i)$ e $\frac{d\mu_i}{d\theta_i} = V_i$, tem-se

$$\boldsymbol{U}_r = \frac{1}{\phi}\sum_{i=1}^{n}(y_i - \mu_i)\frac{1}{V_i}\frac{d\mu_i}{d\eta_i}y_{ir}, \tag{2.12}$$

para $r = 1, ..., p$.

Segundo Cordeiro e Demétrio (2007) a estimativa de máxima verossimilhança (EMV) $\widehat{\boldsymbol{\beta}}$ do vetor de parâmetros $\boldsymbol{\beta}$ é obtida igualando-se $\boldsymbol{U}_r$ a zero para $r = 1, ..., p$. Em geral, as equações $\boldsymbol{U}_r = 0$, $r = 1, ..., p$ não são lineares e tem que ser resolvidas numericamente por processos iterativos do tipo Newton-Raphson. O método iterativo de Newton-Raphson para a soluções de uma equação $f(x) = 0$ é baseado na aproximação de Taylor para a

função $f(x)$ na vizinhança do ponto x_0, ou seja,

$$f(x) = f(x_0) + (x - x_0)f'(x_0) = 0,$$

obtendo-se

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

ou de uma forma mais geral

$$x^{(m+1)} = x^{(m)} - \frac{f(x^{(m)})}{f'(x^{(m)})},$$

sendo $x^{(m+1)}$ o valor de x no passo $(m+1)$, $x^{(m)}$ o valor de x no passo m, $f(x^{(m)})$ a função de $f(x)$ avaliada em $x^{(m)}$ e $f'(x^{(m)})$ a derivada da função $f(x)$ avaliada em $x^{(m)}$.

Considerando-se que se deseja obter a solução do sistema de equações $\mathbf{U} = \mathbf{U}(\boldsymbol{\beta}) = \partial\ell(\boldsymbol{\beta})\partial\boldsymbol{\beta} = 0$ e, usando-se a versão multivariada do método de Newton-Raphson, tem-se

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

sendo $\boldsymbol{\beta}^{(m)}$ e $\boldsymbol{\beta}^{(m+1)}$ os vetores de parâmetros estimados nos passos m e $(m+1)$, respectivamente, $\mathbf{U}^{(m)}$ o vetor escore avaliado no passo m, e $(\mathbf{J}^{(m)})^{-1}$ a inversa da negativa da matriz de derivadas parciais de segunda ordem de $\ell(\boldsymbol{\beta})$, com elementos $\frac{-\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_r\partial\boldsymbol{\beta}_s}$, avaliada no passo m.

Quando as derivadas parciais de segunda ordem são avaliadas facilmente, o método de Newton-Raphson é bastante útil. Acontece, porém, que isso nem sempre ocorre e no caso dos MLG usa-se o método escore de Fisher que, em geral, é mais simples (coincidindo-se com o método de Newton-Raphson no caso das funções de ligação canônicas). Esse método envolve a substituição da matriz de derivadas parciais de segunda ordem pela matriz de valores esperados das derivadas parciais, isto é, a substituição da matriz de informação observada, $\mathbf{J}$, pela matriz de informação esperada de Fisher, $\mathbf{K}$. Logo,

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} + (\mathbf{K}^{(m)})^{-1}\mathbf{U}^m \tag{2.13}$$

sendo que $\mathbf{K}$ tem elementos típicos definidos por

$$\mathbf{K}_{r,s} = -E[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_r \partial \boldsymbol{\beta}_s}] = E[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_r} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_s}],$$

que é a matriz de covariâncias dos $\mathbf{U}'_{r,s}$.

Multiplicando-se ambos os membros de (2.13) por $\mathbf{K}^{(m)}$, tem-se

$$\mathbf{K}^{(m)}\boldsymbol{\beta}^{(m+1)} = \mathbf{K}^{(m)}\boldsymbol{\beta}^{(m)} + \mathbf{U}^{(m)}. \tag{2.14}$$

O elemento típico κ_{rs} de $\mathbf{K}$ é obtido de (2.5) como

$$k_{r,s} = E(\mathbf{U}_r, \mathbf{U}_s) = \phi^{-2} \sum_{i=1}^{n} E(Y_i - \mu_i)^2 \times \frac{1}{V_i^2} \times (\frac{d\mu_i}{d\eta_i})^2 \times y_{ir} \times y_{is}$$

ou

$$k_{r,s} = \phi^{-1} \sum_{i=1}^{n} \omega_i y_{ir} y_{is}$$

sendo $\omega_i = \frac{1}{V_i} \times (\frac{d\mu_i}{d\eta_i})^2$ denominado peso.

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

$$\mathbf{K} = \phi^{-1}\mathbf{X}^T\boldsymbol{\omega}\mathbf{X}$$

sendo $\boldsymbol{\omega} = diag(\omega_1, ..., \omega_n)$ uma matriz diagonal de pesos que traz a informação sobre a distribuição e a função de ligação usadas e poderia incluir também um termo para peso a *priori*. No caso das funções de ligação canônicas tem-se $\omega_i = V_i$, pois $V_i = V(\mu_i) = d\mu_i/d\eta_i$.

Nota-se que as informações são inversamente proporcionais ao parâmetro de dispersão. O vetor escore $\mathbf{U} = \mathbf{U}(\boldsymbol{\beta})$ com componentes em (2.5) pode, então, ser escrito na forma

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

com $\mathbf{G} = diag\{d\eta_1/d\mu_1, ..., d\eta_n/d\mu_n\} = diag\{g'(\mu_1), ..., g'(\mu_n)\}$.

Assim, a matriz diagonal $\mathbf{G}$ é formada pelas derivadas de primeira ordem da função de ligação segundo Cordeiro e Demétrio, (2007). Substituindo $\mathbf{K}$ e $\mathbf{U}$ em (2.7) e eliminando-se ϕ, tem-se que

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

ou, ainda,

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

Define-se a variável dependente ajustada $\mathbf{Z} = \boldsymbol{\eta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu})$.

Logo,

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

ou

$$\boldsymbol{\beta}^{(m+1)} \;=\; (\mathbf{X}^T\mathbf{W}^{(m)}\mathbf{X})^{-1}\ \mathbf{X}^T\mathbf{W}^{(m)}\mathbf{Z}^{(m)}. \tag{2.15}$$

A equação matricial (2.8) é válida para qualquer MLG e mostra que a solução das equações de MV equivale a calcular repetidamente uma regressão linear ponderada de uma variável dependente ajustada $\mathbf{Z}$ sobre a matriz $\mathbf{X}$ usando-se uma função de peso $\mathbf{W}$ que se modifica no processo iterativo. As funções de variância e de ligação entram no processo iterativo através de $\mathbf{W}$ e $\mathbf{Z}$. Note-se que $COV(\mathbf{Z}) = \mathbf{G}COV(\mathbf{Y})\mathbf{G} = \phi\mathbf{W}^{-1}$, isto é, os $\mathbf{Z}_i$ não são correlacionados. É importante enfatizar que a equação iterativa (2.8) não depende do parâmetro de dispersão ϕ.

2.6 Estimação em modelos especiais

Segundo Cordeiro e Demétrio (2007), para as funções de ligação canônicas $\boldsymbol{\omega} = \mathbf{V} = d\mu/d\eta$ que produzem os modelos denominados canônicos, as equações de MV têm a seguinte forma, facilmente deduzidas de (2.12),

$$\sum_{i=1}^{n} x_{ir}y_i = \sum_{i=1}^{n} x_{ir}\widehat{\mu}_i$$

para $r = 1, ..., p$. Em notação matricial, tem-se

$$\mathbf{X}^T\mathbf{y} \;=\; \mathbf{X}^T\widehat{\boldsymbol{\mu}} \tag{2.16}$$

Nesse caso, as estimativas de MV dos $\boldsymbol{\beta}_s^{'}$ são únicas. Sendo $\mathbf{S} = (s_1, ..., s_p)^T$ o vetor de estatísticas suficientes, definidas por $\mathbf{S}_r = \sum_{i=1}^{n} x_{ir}y_i$, e $\mathbf{S} = (s_1, ..., s_p)^T$ os seus valores

amostrais, as equações (2.16) podem ser expressas por $E(\mathbf{S};\widehat{\boldsymbol{\mu}}) = \mathbf{s}$ significando que as estimativas de MV das médias $\mu_1, ..., \mu_n$ nos modelos canônicos são obtidas igualando-se as estatísticas suficientes minimais aos seus valores esperados. Se a matriz modelo corresponde a uma estrutura fatorial, consistindo somente de zeros e uns, o modelo pode ser especificado pelas margens que são as estatísticas minimais, cujos valores esperados devem igualar aos totais marginais. As Equações (2.16) são válidadas para os seguintes modelos canônicos: *modelo classico de regressão*, *modelo log-linear*, *modelo logistico linear*, *modelo gama* com função de ligação recíproca e *modelo normal inverso* com função de ligação refíproca ao quadrado. Para os modelos canônicos, o ajuste é feito pelo algoritmo (2.15) com $\mathbf{W} = diag\{V_i\}$, $\mathbf{G} = diag\{V_i^{-1}\}$ e variável dependente ajustada com componente tipica expressa por $z_i = \eta_i + (y_i - \mu_i/V_i)$.

Nos modelos com respostas binárias, a variável resposta tem distribuição binomial $B(m_i, \pi_i)$ e o logaritmo da função de verossimilhança em (2.11) é expresso como

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log(\frac{\mu_i}{m_i - \mu_i}) + m_i \log(\frac{m_i - \mu_i}{m_i})\} + \sum_{i=1}^{n} \log \left(\begin{array}{c} m_i \\ y_i \end{array} \right),$$

em que, $\mu_i = m_i\pi_i$. É importante notar que se $y_i = 0$, tem-se $\ell(\boldsymbol{\beta}) = m_i \log[(m_i - \mu_i/m_i])$ e se $y_i = m_i$, tem-se como componente típico da função (3.7) $\ell(\boldsymbol{\beta}) = m_i \log(\mu_i/m_i)$.

No caso especial do modelo logístico linear, obtém-se $\boldsymbol{\eta}_i = g(\mu_i) = \log[\mu_i/(m_i - \mu_i)]$. As iterações em (2.15) são realizadas com matriz de pesos $\mathbf{W} = diag\{\mu_i(m_i - \mu_i)/m_i\}$, $\mathbf{G} = diag\{m_i/[\mu_i(m_i - \mu_i)]\}$ e variável dependente ajustada com componentes iguais a $\mathbf{z}_i = \eta_i + [m_i(y_i - \mu_i)]/[\mu_i(m_i - \mu_i)]$. O algoritmo (2.15), em geral, converge, exceto quando ocorrem médias ajustadas próximas a zero ou ao índice m_i.

Nos modelos *log*-lineares para análise de dados de contagens, a variável resposta tem distribuição de poisson $P(\mu_i)$ com função de ligação logarítmica e, portanto, $\boldsymbol{\eta}_i = \log \mu_i = \mathbf{x}_i^T\boldsymbol{\beta}$, $i = 1, ..., n$. Nesse caso, as iterações em (2.15) são realizadas com matriz de pesos $\mathbf{W} = diag\{\mu_i\}$, $\mathbf{G} = diag\{\mu_i^{-1}\}$ e variável dependente ajustada com componentes iguais a $\mathbf{z}_i = \eta_i + (y_i - \mu_i)/\mu_i$. Esse caso especial do algoritmo (2.15) foi apresentado primeiramente por Haberman (1978).

O algoritmo (2.15) pode ser usado para ajustar inúmeros outros modelos,como aqueles baseados na família exponencial (2.1), bastando identificar as funções de variância e de ligação, (CORDEIRO, 1995).

2.7 Seleção de modelos

De acordo com Cordeiro e Demétrio (2007), é difícil propor uma estratégia geral para o processo de escolha de um MLG a ser ajustado aos dados que se dispõe. Isso está intimamente relacionado ao problema fundamental da estatística que, segundo Fisher, "o que se deve fazer com os dados?". Em geral, o algoritmo de ajuste deve ser aplicado não a um MLG isolado, mas a vários modelos de um conjunto bem amplo que deve ser, realmente, relevante para o tipo de dados que se pretende analisar. Se o processo é aplicado a um único modelo, não levando em conta possíveis modelos alternativos, existe o risco de não se obter um dos modelos mais adequados aos dados. Esse conjunto de modelos pode ser formulado de várias maneiras:

i) Definindo-se uma família de funções de ligação;

ii) Considerando-se diferentes opções para a escala de medição;

iii) Adicionando-se (ou retirando) vetores colunas independentes a partir de uma matriz básica original.

Segundo Cordeiro e Demétrio (2007), pode-se propor um conjunto de modelos para dados estritamente positivos, usando-se a família potência de funções de ligação $\eta = g(\mu;\lambda) = (\mu^{\lambda} - 1)\lambda^{-1}$, em que λ é um parâmetro que indexa o conjunto. Para dados reais positivos ou negativos, outras famílias podem ser definidas como $g(\mu;\lambda) = [\exp(\lambda\mu) - 1]\lambda^{-1}$. A estimativa de MV de λ em geral, define um modelo bastante adequado, porém, muitas vezes, de difícil interpretação. Devem-se analisar não somente os dados brutos mas procurar modelos alternativos aplicados aos dados transformados $z = h(y)$. O problema crucial é a escolha da função de escala $h(.)$. No modelo clássico de regressão, essa escolha visa a combinar, aproximadamente, normalidade e constância da variância do erro aleatório, bem como, aditividade dos efeitos sistemáticos. Entretanto, não existe nenhuma garantia que $h(.)$ exista, nem mesmo que produza algumas das propriedades desejadas.

3 Aplicação

Os dados para a aplicação via modelos lineares generalizados para a avaliação do controle biológico de insetos (gorgulho do milho Sitophilus zeamais Mots. (oleoptera: Curculionidae)), foram disponibilizados pelo departamento de Plantas e Inseticidas do Departamento de Entomologia e Acarologia, da Escola Superior de Agricultura "Luiz de Queiroz"/ Universidade de São Paulo (ESALQ/USP), com o Departamento de Química da Universidade Federal de São Carlos (UFSCar), como parte das atividades do Instituto Nacional de Ciência e Tecnologia de Controle Biorracional de Insetos Pragas, sediado na UFSCar.

Assim, os ensaios biológicos foram conduzidos no Laboratório de Plantas Inseticidas da ESALQ/USP, em Piracicaba, SP, enquanto que as extrações, as análises cromatográficas e os fracionamentos e partições químicas foram desenvolvidas no Laboratório de Produtos Naturais da UFSCar, em São Carlos, SP.

Entre os extratos vegetais testados nos bioensaios anteriores, foram selecionados os dois que apresentaram os resultados mais promissores, refletidos nas menores concentrações letais média. Da mesma forma, selecionou-se a formulação de terra de diatomácea que se mostrou mais eficiente. Visando-se avaliar o efeito interativo de ambas as técnicas, amostras de 50g de milho foram submetidas aos seguintes tratamentos:

T1 - Testemunha (solvente utilizado na ressuspensão dos extratos);

T2 - Extrato 1 (na CL_{30} determinada) dose letal;

T3 - Extrato 1 (na CL_{50} determinada) dose letal;

T4 - Extrato 2 (na CL_{30} determinada) dose letal;

T5 - Extrato 2 (na CL_{50} determinada) dose letal;

T6 - Terra de diatomácea (na $CL30$ determinada);

T7 - Terra de diatomácea (na $CL50$ determinada);

T8 - Terra de diatomácea (na $CL30$ determinada) + Extrato 1 (na CL_{50}) dose letal;

T9 - Terra de diatomácea (na $CL50$ determinada) + Extrato 1 (na CL_{30}) dose letal;

T10 - Terra de diatomácea (na $CL30$ determinada) + Extrato 2 (na CL_{50}) dose letal;

T11 - Terra de diatomácea (na $CL50$ determinada) + Extrato 2 (na CL_{30}) dose letal;

Cada amostra foi infestada entre 30 e 40 insetos (Sitophilus zeamais Mots), não sexados e com idade entre 10 e 20 dias, com seis repetições por tratamento. Representa-se o número de insetos em cada tratamento por n e o número de insetos que morreram pela letra x, sendo assim, a proporção de insetos mortos será representado por $y{=}x/n$, que por sua vez segue uma distribuição binomial. Para obtenção dos resultados utilizou-se o software livre R, versão 2.15.0 e o software SAS Interprise Guide, versão 4.4.

Iniciam-se as análises referentes à associação de extratos vegetais e formulações de terra de diatomácea no controle biológico de insetos por meio da estatística descritiva. Pode-se observar por meio da Figura 1, gráfico de boxplot, a proporção observada de insetos mortos em diferentes concentrações. Ross (2010) ressalta a importância de iniciar uma análise descritiva por meio do boxplot (gráfico de caixa), pois é um gráfico utilizado para avaliar a distribuição empírica dos dados. O boxplot é formado pelo primeiro e terceiro quartil e pela mediana. As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior e do quartil superior até o maior valor não superior ao limite superior. Valores fora desse intervalo são considerados outlier.

Pode-se observar por meio da Figura 1 uma maior proporção de insetos mortos no tratamento 1. Pode-se observar também a presença de um outlier no tratamento 6, nos tratamentos 3, 7 e 8 os dados encontram-se centrados em torno da média, já nos tratamentos 1, 2, 5, 9 e 10 existe uma assimetria maior na sua parte inferior.

Continuando-se as análises, seleciona-se a função de ligação que melhor se ajusta aos dados por meio do Critério de Informação Akaike(AIC), utiliza-se as funções de ligação *logit*, *probit* e *complemento log-log*. Pode-se observar por meio da Tabela 1 que a função de ligação *complemento log-log* obteve o menor valor AIC, ou seja, a função de ligação *complemento log-log* é a mais adequada para o ajuste dos dados em questão.

Este fato corrobora com Cordeiro (1995), ao afirmar que a função de ligação *complemento log-log* é a mais adequada no ajuste de dados com assimetria à direita e comprova

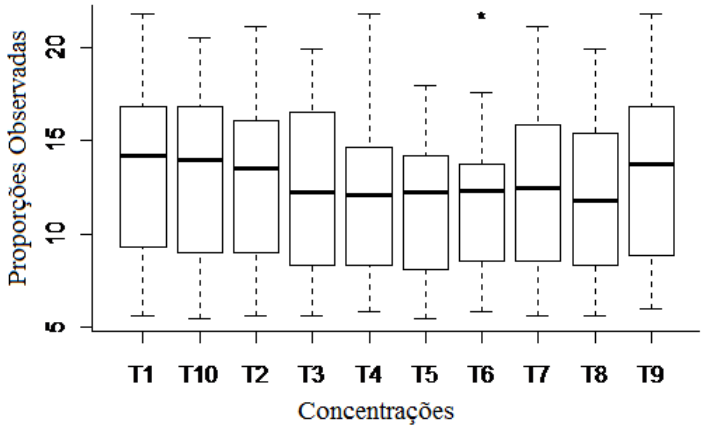

Figura 1: Gráfico de boxplot referente à associação de extratos vegetais e formulações de terra de diatomácea no controle biológico de insetos

Tabela 1: Comparação das funções de ligação por meio do Critério de Informação Akaike(AIC) aos dados de associação de extratos vegetais e formulações de terra de diatomácea no controle biológico de insetos.

Função de ligação	Estimativas	p-valor	AIC
logit	0,16534	$2,17e^{-0,9}$	75,932
probit	0,09131	$3,29e^{-11}$	77,507
complemento log-log	0,05542	$4,7e^{-0,8}$	74,362

que esta função de ligação é apropriada para o modelo binomial. Resende e Beile (2002) concluíram as funções de ligação *logit* e *probit* mostraram-se adequadas na análise dos dados de sobrevivência de plantas de espécies perenes. Os referidos autores afirmaram que as funções de ligação *complemento log-log* e *identidade* mostraram-se inadequadas aos dados de proporções analisados. Pode-se observar por meio da Figura 2 o ajuste das funções de ligação aos dados de proporção de insetos mortos.

O histograma da Figura 3 corrobora com Cordeiro (1995) no que se refere a assimetria dos dados em estudo. Pode-se observar também por meio da Figura 3 que existe uma maior proporção de insetos mortos entre 20 e 30 unidades.

Na sequência, por meio da Figura 4, observa-se a proporção da associação de extratos vegetais e formulações de terra de diatomácea no controle biológico de insetos, ou seja, a proporção de insetos mortos para cada um dos tratamentos avaliados. Pode-se observar que praticamente todos os tratamentos tiveram uma proporção de morte de insetos acima de 0,60, apenas o tratamento 7 obteve uma proporção inferior a 0,35. Os resultados das proporções médias de insetos mortos para cada um dos tratamentos avaliados encontram-se na Tabela2. Com isso, verifica-se a autenticidade das diferenças das proporções médias

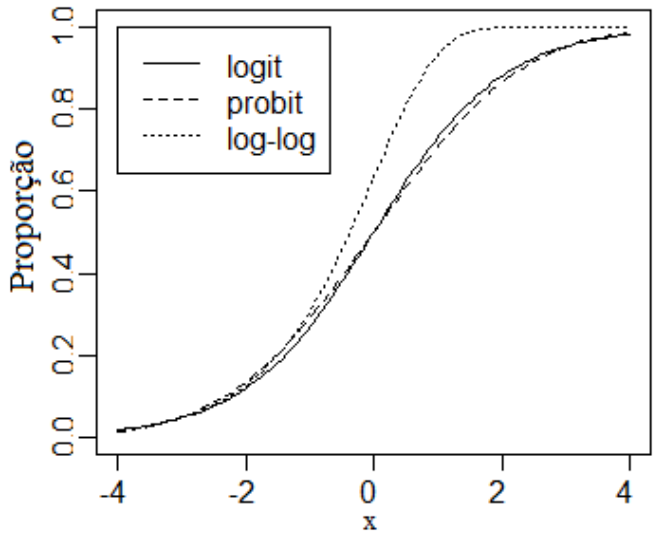

Figura 2: Ajuste das funções de ligação aos dados de proporção de insetos mortos de acordo com os extratos vegetais testados nos bioensaios

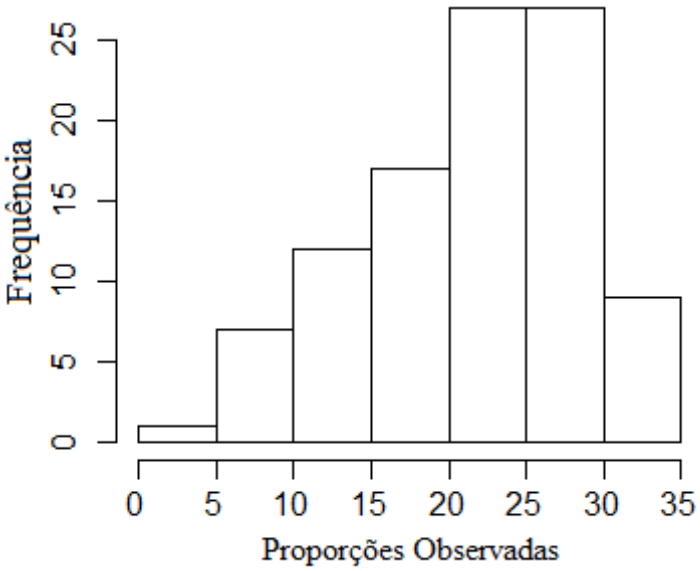

Figura 3: Histograma da proporção de insetos mortos referente a associação de extratos vegetais e formulações de terra de diatomácea

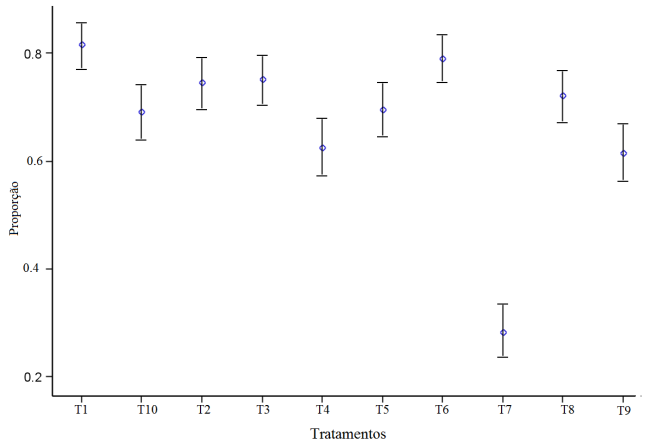

Figura 4: Proporções de insetos mortos para cada um dos níveis avaliados

do fator em estudo, destacando-se os tratamentos 1,2,3 e 6.

Tabela 2: Comparação das proporções médias de insetos mortos para cada um dos tratamentos avaliados.

Tratamentos	Médias
T_1	$0,8159_a$
T_{10}	$0,6914_{cde}$
T_2	$0,7454_{abc}$
T_3	$0,7515_{abc}$
T_4	$0,6254_{def}$
T_5	$0,6963_{cde}$
T_6	$0,7914_{ab}$
T_7	$0,2826_g$
T_8	$0,7211_{bcdf}$
T_9	$0,6156_e$

médias seguidas da mesma letra minúscula nas colunas não diferem(p<0,05).

Conforme Motgomery (1997), os graus de liberdade do fator em estudo podem ser desdobrados em contrastes de interesse, ortogonais entre si, cada um com (1) grau de liberdade, no presente estudo analisaram-se apenas sete (7) contrastes ortogonais de interesse prático. os contrastes estão disponiveis na Tabela 3, conforme pode-se observar os contrastes: Testemunha versus Demais (Hopótese $H_0^{(2)}$); Extrato versus Diatomácea (Hipótese $H_0^{(3)}$); Diatomácea CL30 versus Diatomácea CL50 (Hipótese $H_0^{(6)}$) e Extrato1 versus Extrato2 (Hipótese $H_0^{(8)}$) foram significativos ao nível de 0,05 de significância. Os demais contrastes análisados não foram significativos.

Na seguência analisam-se os pressupostos para validação do modelo observando-se por meio da figura 5 os gráficos dos resíduos estudentizados. Pode-se observar que os resíduos estão normalmente distribuidos, ou seja, há indícios para não rejeitar a hipótese nula que os resíduos seguem uma distribuição normal. Isto é confirmado pelo quantil

Tabela 3: Descrição de alguns contrastes ortogonais para a avaliação do controle biológico de insetos

Fator de variação	Grau de Liberdade	Estatística F	p-valor
$H_0^{(2)}$:$\mu_1 = \frac{\mu_2+\mu_3+\mu_4+\mu_5+\mu_6+\mu_7+\mu_8+\mu_9+\mu_{10}+\mu_{11}}{10}$	1	39,75	< 0001
$H_0^{(3)}$:$\frac{\mu_2+\mu_3+\mu_4+\mu_5}{4} = \frac{\mu_6+\mu_7+\mu_8+\mu_9+\mu_{10}+\mu_{11}}{6}$	1	26,98	< 001
$H_0^{(4)}$:$\mu_2 = \mu_3$	1	0,03	08572
$H_0^{(5)}$:$\mu_4 = \mu_5$	1	3,64	00596
$H_0^{(6)}$:$\mu_6 = \mu_7$	1	152,11	< 0001
$H_0^{(7)}$:$\frac{\mu_8+\mu_9}{2} = \frac{\mu_{10}+\mu_{11}}{2}$	1	0,51	04772
$H_0^{(8)}$:$\frac{\mu_2+\mu_3}{2} = \frac{\mu_4+\mu_5}{2}$	1	11,17	00008

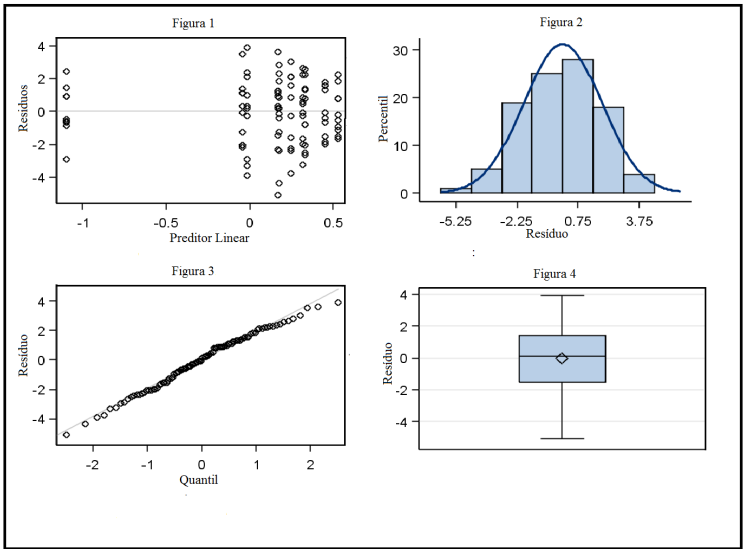

Figura 5: Gráficos dos resíduos estudentizados

amostral versus quantil esperado sob normalidade representado pelo QQ-Plot (gráfico do canto inferior esquerdo da figura 5), percebe-se claramente que os pontos não se desviam do comportamento linear.

4 Conclusão

Este trabalho teve por objetivo a aplicação dos modelos lineares generalizados para a avaliação do controle biologico de insetos, a distribuição de probabilidade seguida foi binomial, sua função de ligação encontrada através do critério de informação de Akaike (AIC), a complemento *log-log*, comparada com as funções *probit* e *logit*, foi a melhor, pois apresentou o menor valor de AIC.

Na comparação das proporções média de insetos mortos para cada uma dos tratamentos, obteve-se que as estemunha versus Demais (Hopótese $H_0^{(2)}$), Extrato versus Diatomácea (Hipótese $H_0^{(3)}$), Diatomácea CL30 versus Diatomácea CL50 (Hipótese $H_0^{(6)}$) e Extrato1 versus Extrato2 (Hipótese $H_0^{(8)}$) foram significativos ao nível de 0,05 de significância.

Na análise gráfico do box plot referente à associação de extratos vegetais e formulação de terra de diatomácea no controle biológico de insetos obteve-se uma maior propoção de insetos mortos no tratamento 1, também a presença de um outlier no tratamento 6, nos tratamentos 3, 7 e 8 os dados encontram-se centrados em torno da média, já nos tratamentos 1, 2, 5, 9 e 10 existe uma assimetria maior na sua parte inferior.

No gráfico de histograma existe uma maior proporção de insetos mortos entre 20 e 30 unidades.

Referências

BIRCH, M.W; Maximum likelihood in three-way contingency tables. **Journal of the Royal Statistical Society**, B52, 1963. 220-233p.

BLISS, C.I; The calculation of the dosage-mortality curve. **Annals of Applied Biology**, 22, 134-167p. 1935.

CORDEIRO, G.M; DEMÉTRIO, C.G.B. **Apostila Modelos Lineares Generalizados**. UFSM, Santa Maria, RS, 2007. 165p.

CORDEIRO, G. M. Performance of a Bartlett-type modification for the deviance. **Journal of Statistical Computation and Simulation**, 51, 1995. 385-403p.

DEMÉTRIO, C.G.B. Modelos Lineares Generalizados em Experimentação Agronômica. Piracicaba, SP,2002. 121p.

DYKE, G.V; Patterson, H.D. Analysis of factorial arrangements when the data are proportions. **Biometrics** 8, 1952. 1-12p.

EHLERS, R.S. **Inferência Estatística**, 2009. 154p.

FEIGL, P. and Zelen, M. Estimation of exponential survival probabilities with concomitant information. **Biometrics** 21, 1965. 826-838p.

FISHER, R.A . On the mathematical foundations of theoretical statistics. **Philosophical Transactions of the Royal Society**, 222, 1922. 309-368p.

GLASSER, M. Exponential survival with covariance. **Journal of the American Statistical Association**, 62, 1967. 561-568p.

Lindsey, J.K. Applying Generalized Linear Models. Springer, New York. 1997.

McCullagh, P; Nelder, J.A. Generalized Linear Models. 2 edition, Chapman and Hall, London. 1989.

MONTGOMERY, D.C.; Desing and Analysis of Experimentos. John Wiley and Sons , New York, 1997. 669p.

NELDER, J.A. and Wedderburn, R.W.M. Generalized linear models. **Journal of the Royal Statistical Society**, A 135, 1972. 370-384p.

RASCH, G; **Probabilistic Models for some Intelligence and Attainment Tests**. Danmarks Paedogogiske Institut, Copenhagen. 1960.

RESENDE, Marcos Deon Vilela de; BIELE Jonathan. **Revista Mat. Estat**., São Paulo, 20:, 2002. 39-65p.

ROSS, S. **Probabilidade: um curso moderno com aplicações/ Sheldon Ross**; tradutor: Alberto Resende de Conti.- 8.ed.- Porto Alegre: Bookman, 2010. 608p.

SOUNIS, E. **Princípios fundamentais, metodologia estatistica aplicação às ciências biológicas Bioestatística**. Rio de Janeiro: Ed. ATHENEU, 1985. 304p.

TURKMAN, M.A.A,; SILVA, **G.L. Modelos Lineares Generalizados da teoria à prática**, Universidade de Lisboa, 2000. 153p.

ZIPPIN, C. and Armitage, P. Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. **Biometrics**, 22, 1966. 665-672p.