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

Dados entomológicos discretos e alguns modelos para análise estatística

João Vitor Ribeiro Silva

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

Piracicaba
2021

João Vitor Ribeiro Silva
Bacharel em Estatística

Dados entomológicos discretos e alguns modelos para análise estatística

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **IDEMAURO ANTONIO RODRIGUES DE LARA**

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

Piracicaba
2021

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

Silva, João Vitor Ribeiro
Dados entomológicos discretos e alguns modelos para análise estatística / João Vitor Ribeiro Silva. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2021 .
54 p.

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

1. Dados de contagem 2. Modelos lineares generalizados 3. Dados politômicos 4. Modelos de mistura . I. Título.

DEDICATÓRIA

Aos meus pais Maria Aparecida e José Geraldo,
dedico com todo amor do mundo.
As minhas irmãs Débora, Bárbara e Lívia,
dedico pelo exemplo de sempre.
As melhores pessoas que conheci: William, Lia, Julienderson e Daniella,
dedico por todo o apoio.
A todos meus familiares e amigos,
dedico por todo suporte e carinho.

AGRADECIMENTOS

Agradeço imensamente aos meus pais, Maria Aparecida e José Geraldo, que sempre me incentivaram, me apoiaram e fizeram de tudo para que eu pudesse concluir mais essa etapa. Sem vocês, eu não teria chegado até aqui. Amo vocês;

Agradeço às minhas irmãs, Débora, Bárbara e Lívia, pelo exemplo de sempre, pelo apoio e por estarmos sempre juntos;

Aos meus familiares, obrigado por todo o apoio, não somente nesse período, mas sempre;

Agradeço aos meus professores e colegas da Universidade Federal de Uberlândia por serem uma parte especial dessa conquista. Prof. Dr. Lúcio Borges de Araújo, meu orientador de graduação, obrigado pelo exemplo e pelo apoio;

Agradeço ao meu orientador Prof. Dr. Idemauro Antonio Rodrigues de Lara, por todos os conselhos, pela orientação e por ter me ajudado a chegar até aqui;

Aos meus amigos e parte de quem sou hoje, William Costa, Lia Resende, Julienderson Costa e Daniella Barros, agradeço por sempre estarem ao meu lado e por trazerem mais alegria aos meus dias;

A minha colega de mestrado que se tornou uma maravilhosa amiga, Gabriela, agradeço por compartilharmos todos os dias do mestrado e por ter me apoiado durante esses dois anos. Você foi essencial;

Agradeço também aos colegas da pós graduação, Deoclécio, Wellington, Suellen e Denise pelas disciplinas juntos e por todo o conhecimento compartilhado;

Caroline Sakuno e Carolina Reigada, muito obrigado por toda a ajuda com os experimentos e pelas discussões;

Agradeço aos professores da ESALQ/USP, em especial às professoras Renata, Taciana e Clarice, por todo o ensinamento passado e pelo amor em passar esse conhecimento aos alunos;

Agradeço à Solange por todas as dúvidas (e olha que foram muitas), por estar sempre disponível e estendo esse agradecimento à todos os servidores da ESALQ/USP pelo excelente trabalho;

Agradeço aos meus colegas de trabalho, que se tornaram grandes amigos, Raphael Dayan, Daniel Dantas, Luiz Salvador, Rodolfo Conversani, Simone Waissman, Matheus Nascimento, Yohanis Fuks e Cássio Oliveira. Obrigado por todo o conhecimento trocado, pelas risadas, companhia e apoio;

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro e pelo financiamento da estrutura do PPG em Estatística e Experimentação Agronômica da ESALQ/USP.

Enfim, agradeço a todos que fizeram parte dessa conquista e contribuíram, de forma direta ou indireta na realização deste trabalho.

EPÍGRAFE

"First, think. Second, believe. Third, dream. Finally, dare."

Walt Disney

"If you're offered a seat on a rocket ship, don't ask what seat!

Just get on."

Sheryl Sandberg

SUMÁRIO

RESUMO

Dados entomológicos discretos e alguns modelos para análise estatística

O uso de contagens na avaliação de experimentos é uma prática muito comum em diversas áreas do conhecimento, assim como na área de ciências agrárias. Tal prática é devida ao fato de sua versatilidade nas avaliações, sendo possível analisar estes resultados de forma numérica ou categorizada. Distribuições usuais, tais como a Poisson, podem não ser a melhor alternativa para o ajuste de modelos para dados de contagem, pois nem sempre a pressuposição de equidispersão é satisfeita. No primeiro capítulo do presente trabalho são apresentadas distribuições alternativas para a análise de dados de contagem, sendo estas: a distribuição binomial negativa, quase-Poisson, COM Poisson e Poisson-Tweedie. No segundo capítulo, adicionalmente são apresentadas as distribuições multinomial, multinomial negativa e a em dois estágios Dirichlet-multinomial como alternativas para se ajustarem modelos com dados categorizados. Em ambos os casos são apresentados estudos de motivação, em sua maior parte, da entomologia. Os parâmetros dos modelos foram estimados utilizando o estimador de máxima verossimilhança (EMV) e os ajustes avaliados por meio do critério de informação de Akaike (AIC) e do gráfico meio normal de probabilidade com envelope simulado (*half-normal plot*). Verificou-se que, em casos com média e variância distintas, os modelos com distribuição binomial negativa e quase-Poisson possuem melhor ajuste quando comparados ao tradicional Poisson. No caso categorizado, fez-se necessário o uso de uma mistura hierárquica de distribuições para o ajuste do modelo, sendo utilizada a distribuição Dirichlet-multinomial.

Palavras-chave: Dados de contagem, Modelos lineares generalizados, Dados politômicos, Modelos de mistura

ABSTRACT

Discrete entomological data and some models for statistical analysis

The use of counts in the evaluation of experiments is a very common practice in several areas of knowledge, as well as in the field of agricultural sciences. This practice is due to its versatility in evaluations, making it possible to analyze these results numerically or categorized. Usual distributions such as Poisson may not be the best alternative for fitting models to count data, as the equidispersion assumption is not always satisfied. In the first chapter of this work, alternative distributions for the analysis of count data are presented, which are: the negative binomial, Quasi-Poisson, COM Poisson and Poisson-Tweedie distributions. In the second chapter, the multinomial, the negative multinomial and the Dirichlet-multinomial two-stage distributions are also presented as alternatives to fit models with categorized data. In both cases motivational studies are presented, mostly from entomology. The parameters of the models were estimated using the maximum likelihood estimator (MLE) and the fits evaluated using the Akaike information criterion (AIC) and the half-normal plot with simulated envelopes. It was found that, in cases with different mean and variance, the models with negative Binomial distribution and quasi-Poisson have a better fit when compared to the traditional Poisson distribution. In the categorized case, it was necessary to use a hierarchical mixture of distributions to fit the model, using the Dirichlet-multinomial distribution.

Keywords: Counting data, Generalized linear models, Polytomous data, Mixture models

1 INTRODUÇÃO

Dados de contagem são provenientes do processo de registros numéricos de algum evento em um espaço de tempo contínuo, consistindo em valores inteiros positivos que são aplicados nas mais diversas áreas do conhecimento (Winkelmann, 2008), como em seguradoras (Mccullagh e Nelder, 1989), financeiras (Davutyan, 1989), medicina (Diggle e Zeger, 1995), entomologia (Reigada, 2009) e genética (Cui et al., 2005).

Tais dados exigem o uso de análises e modelos adequados, os quais começaram a ser estudados com os Modelo Lineares Generalizados (MLG) propostos por Nelder e Wedderburn (1972), no qual a distribuição Poisson é a mais conhecida. Entretanto, a distribuição Poisson é uma distribuição clássica que supõe equidispersão (esperança e variância iguais), o que é incomum quando são estudados experimentos da área de ciências agrárias, pois para os cenários de entomologia, genética e outros, nos quais, em geral, ocorre heterogeneidade na variável resposta, resulta em valores elevados para a variância, podendo indicar até cenários de superdispersão (Hinde e Demétrio, 1998). Este fenômeno ainda é alvo de estudos pois, como descrito por Demétrio et al. (2014), a presença da superdispersão pode ser oriunda da variabilidade do material experimental, correlação entre respostas individuais, excessos de zeros ou pela presença de *outliers*. Ainda, Mccullagh e Nelder (1989) afirmaram que tal variabilidade é caracterizada pelo fato de a variabilidade extra ser maior do que a prevista.

A distribuição a ser utilizada para o problema de variabilidade extra depende do tipo da variável resposta e das condições experimentais intrínsecas. Sendo uma variável resposta numérica, a distribuição mais usual para tratar superdispersão é a binomial negativa, porém existem outras distribuições que podem ser utilizadas e podem ser mais adequadas ao cenário em estudo, como as distribuições COM-Poisson (Conway e Maxwell, 1962), quase-verossimilhança (ou quase-Poisson) (Wedderburn, 1974) e Poisson-Tweedie (Jørgensen, 1997). Estas são alternativas para ajustar modelos em cenários de sub ou superdispersão, visto que tais distribuições não exigem equidispersão e possuem outros parâmetros que as tornam mais adequadas. Um resumo destas técnicas pode ser encontrado em Batista (2020).

Tratando-se de uma variável resposta categorizada (com mais de duas categorias), a distribuição multinomial é a mais usual. Outras distribuições como a multinomial negativa (Le Gall, 2006), Dirichlet-multinomial (Mosimann, 1962) ou distribuições bayesianas também podem ser utilizadas, sendo a Dirichlet-multinomial uma boa opção para os casos de superdispersão (Salvador, 2019).

O presente trabalho tem como objetivo principal analisar e comparar a aplicação de diferentes distribuições no ajuste de modelos para dados discretos de contagem e categorizados. Como objetivos específicos, tem-se revisar e comparar os modelos de regressão para dados de contagem como: Poisson, Conway-Maxwell-Poisson, binomial negativo, quase-Poisson, COM Poisson e Poisson-Tweedie; revisar e comparar os modelos de regressão para dados categorizados, como: a distribuição multinomial, a distribuição multinomial negativa e a distribuição em dois estágios Dirichlet-multinomial como alternativas; aplicar os modelos a dois estudos de motivação I e II estendendo as aplicações desses modelos para situações experimentais com estruturas hierárquicas ou com excessos de zeros; comparar as estimativas dos parâmetros geradas pelos modelos; apresentar os pacotes do R com recursos computacionais para o ajuste dos modelos apresentados e respectivos recursos a análise dos resíduos.

Para o primeiro capítulo tem-se dois estudos de motivação com dados de contagem. O primeiro estudo provém de um experimento conduzido por Sakuno (2021) realizado na Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ) relacionado à oviposição de *Diatraea saccharalis* em culturas de cana-de-açúcar, com o objetivo de avaliar a quantidade de ovos postos no período de 72 horas. Para tal, utilizaram-se as distribuições Poisson, binomial negativa e quase-Poisson para o ajuste do modelo, sendo estes comparados pelo AIC e validados utilizando o gráfico de resíduos com envelope simulado (*half-normal plot*) por meio do pacote *hnp* (Moral et al., 2017). O segundo estudo de motivação provém

de um experimento conduzido por Reigada (2009) envolvendo quatro espécies de vespas parasitóides, cujo objetivo era observar a quantidade de pupas parasitadas no período de 24 horas. Para tal, foram propostos modelos com as distribuições Poisson, COM Poisson e Poisson-Tweedie e comparados os ajustes pelo AIC e, adicionalmente, realizadas as predições para cada caso por meio do modelo selecionado.

Para o segundo capítulo, também foram apresentados dois estudos de motivação, porém as variáveis respostas apresentadas são categorizadas nominais, com três níveis. O primeiro estudo, conduzido pelo Laboratório de Ecologia de Interações do Departamento de Ecologia e Biologia Evolutiva da Universidade Federal de São Carlos (UFSCar) no ano de 2020, visa compreender a preferência entre três categorias de ovos para o parasitismo de vespas parasitóides, no qual foram ajustados dois modelos com a distribuição multinomial, comparados pelo teste de razão de verossimilhanças visando obter o modelo com melhor ajuste. O segundo estudo de motivação provém de um experimento com suínos, primeiramente apresentado por Castro (2016), que objetiva entender o comportamento dos animais em diferentes cenários de enriquecimento ambiental, para o qual foram ajustados modelos com as distribuições multinomial, multinomial negativa e Dirichlet-multinomial e, posteriormente, comparados pelo AIC e validados por meio do gráfico meio normal de probabilidade com envelope simulado (*half-normal plot*).

Referências

Batista, D. T. (2020). *Modelos para dados de contagem não equidispersos com aplicação à ecologia e em estudos longitudinais*. PhD thesis, Escola Superior de Agricultura Luiz de Queiroz.

Castro, A. C. D. (2016). *Comportamento e desempenho sexual de suínos reprodutores criados em ambientes enriquecidos*. PhD thesis, Universidade de São Paulo / Escola Superior de Agricultura "Luiz de Queiroz".

Conway, R. e Maxwell, W. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136.

Cui, Y., Kim, D., e Zhu, J. (2005). On the Generalized Poisson Regression Mixture Model for Mapping Quantitative Trait Loci With Count Data. *Genetics Society of America*, 174:2159–2172.

Davutyan, N. (1989). Bank failures as Poisson variates. *Economics Letters*, 29:333–338.

Demétrio, C., Hinde, J., e Moral, R. (2014). *Models for Overdispersed Data in Entomology. In Ecological Modelling Applied to Entomology*. Springer, Switzerland.

Diggle, P., L. K.-Y. e Zeger, S. L. (1995). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Hinde, J. e Demétrio, C. (1998). Overdispersion : Models and Estimation. *SINAPE*, page 73.

Jørgensen, B. (1997). The Theory of Dispersion Models. *Monographs on Statistics and Applied Probability. London: Chapman & Hall*.

Le Gall, F. (2006). The modes of a negative multinomial distribution. *Statistics and Probability Letters*, 76:619–624.

Mccullagh, P. e Nelder, J. (1989). *Generalized linear models*. Chapman and Hall, London, fifth edition.

Moral, R. A., Hinde, J., e Demétrio, C. G. B. (2017). Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software*, 81:23p.

Mosimann, J. E. (1962). On the Compound Multinomial Distribution , the Multivariate β- Distribution , and Correlations Among Proportions. *Biometrika Trust, Oxford University Press*, 49:65–82.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistics Society*, 135:370–384.

Reigada, C. (2009). *Dinâmica tritrófica experimental em populações de moscas varejeiras*. PhD thesis, Universidade Estadual Paulista.

Sakuno, C. I. R. (2021). Efeitos da movimentação de *Diatraea saccharalis* (fabricius) (lepidoptera: Crambidae) em sistemas compostos por cana-de-açúcar geneticamente modificada e refúgio. Master's thesis, Tese de Mestrado em Entomologia. Escola Superior de Agricultura "Luiz de Queiroz".

Salvador, M. (2019). O problema da superdispersão em dados categorizados politômicos nominais em estudos agrários. Master's thesis, Tese de Mestrado em Estatística e Experimentação Agronômica. Escola Superior de Agricultura "Luiz de Queiroz".

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer, Switzerland, fifth edition.

2 MODELOS PARA DADOS DE CONTAGEM EM OVIPOSIÇÃO DE *DIATRAEA SACCHARALIS* E PARASITISMO DE VESPAS

Resumo

Em experimentos de entomologia, normalmente as observações possuem como objetivo entender fatores diversos, como, por exemplo, adaptação, proliferação, comportamento, dentre outros. Nessa perspectiva, as avaliações podem ser realizadas como contagens pontuais ou acumuladas, para as quais devem ser ajustados modelos utilizando distribuições discretas apropriadas. Como estudos de motivação, apresentam-se dois experimentos entomológicos relativos à oviposição de *Diatraea saccharalis* e o parasitismo de quatro espécies de vespas em pupas de moscas varejeiras que resultaram em dados de contagem. Nesse contexto, este trabalho teve como objetivo avaliar dados de contagem utilizando-se modelos lineares generalizados (MLG) por meio das distribuições Poisson, binomial negativo, quase-Poisson, COM Poisson e Poisson-Tweedie. Para avaliar a qualidade de ajuste dos modelos, foram realizadas as análises de resíduos e o gráfico meio normal de probabilidade com envelope simulado *half-normal plot*. Com base nos estudos de motivação, os modelos alternativos binomial negativo, quase-Poisson e Poisson-Tweedie possuem melhores ajustes e geram a possibilidade de predizer a variável resposta de maneira satisfatória.

Palavras-chave: Modelos lineares generalizados, Modelos Tweedie, Entomologia, Análise residual

2.1 Introdução

O estudo comportamental, fisiológico e reprodutivo de insetos, conhecido como entomologia, é desenvolvido no Brasil desde o final do século XIX a partir de estudos de pesquisadores como Gustavo Dutra, Hermann von Ihering, Carlos Moreira e Emílio Goeldi (Gallo et al., 1988). A palavra entomologia é de origem grega em que *entom* significa inseto e *logia* significa estudo, sendo assim, estudo dos insetos. Dentre as mais diversas possibilidades de estudos com insetos, pode-se citar estudos que visam entender o comportamento de pragas (Girón Pérez, 2013), estudos fisiológicos para entendimento do funcionamento de orgãos internos dos insetos (Terra et al., 2006), estudos reprodutivos (Chichera et al., 2012) e estudos de pragas relacionadas ao melhoramento genético de plantas (Cesnik, 2007).

As avaliações dos experimentos podem ser realizadas das mais diversas formas, como: medição, pesagem, classificação ou contagem, sendo a última o foco desde trabalho e classificada como longitudinal (ao longo do tempo) ou *cross-section* (pontuais ou acumuladas). As avaliações de contagem necessitam distribuições apropriadas para os ajustes de modelos e, devido à natureza da variável resposta, não é aconselhável utilizar um modelo clássico de regressão linear uma vez que existem classes de modelos que são mais adequadas para explicar uma variável discreta, como os modelos lineares generalizados (MLG). Nessa classe usam-se distribuições discretas (Ramalho, 1996), como a de Poisson, binomial negativa, quase-Poisson, COM Poisson, Poisson-Tweedie, para as quais os modelos ajustados permitem trabalhar padrões de equidispersão, subdispersão, superdispersão, heterocedasticidade, entre outros.

Os MLG propostos por Nelder e Wedderburn (1972), vistos como uma extensão dos modelos de regressão linear tradicionais, normalmente são aplicados em casos que a variável resposta é do tipo contagem, categorizada ou assimétrica. Tais modelos são compostos por três componentes principais, sendo estes: o aleatório, o sistemático e a função de ligação. Essa classe de modelos está sendo cada vez mais utilizada e existem aplicações nas mais diversas áreas do conhecimento, como na agricultura (Rocha et al., 2014), entomologia (Bliss, 1985) e medicina (Teles, 1995).

A distribuição de Poisson é a mais comum quando se deseja avaliar contagens e trata-se de uma distribuição discreta proposta pelo engenheiro e matemático francês Poisson (1837) no trabalho *Recherches sur la probabilité des jugements en matières criminelles et matière civile* (Pesquisa sobre a probabilidade de julgamentos em matéria penal e civil - tradução nossa), cujo objetivo é definir a probabilidade de ocorrência de eventos em um período de tempo, sendo estes eventos independentes entre si e com uma taxa de ocorrência constante conhecida. Muito embora seja uma distribuição antiga e amplamente utilizada em modelagens e análises estatísticas, grande parte dos experimentos não satisfaz o princípio de apresentar equidispersão, ou seja, média e variância iguais. Neste ponto, as distribuições binomial negativa, quase-Poisson, COM Poisson e Poisson-Tweedie possuem vantagens em relação à distribuição de Poisson (Batista, 2020).

As distribuições binomial negativa e COM Poisson possuem um parâmetro a mais que a distribuição Poisson, sendo consideradas distribuições mais flexíveis em relação à igualdade de média e variância. Já a quase-Poisson, também proposta por Wedderburn (1974), propõe uma relação tal que a variância é uma função linear da média a partir de um parâmetro de dispersão ϕ, com uma maior flexibilidade para ajustar um modelo para dados de contagem. A distribuição da família Tweedie, a Poisson-Tweedie, é uma distribuição hierárquica, normalmente utilizada para dados com duas ou mais populações, sendo adequada a dados com padrões de sub ou superdispersão (Jørgensen, 1997).

O presente trabalho tem como objetivo estudar e comparar modelos para dados de contagem, pressupondo-se as distribuições de Poisson, binomial negativa, quase-Poisson, COM Poisson e Poisson-Tweedie. Como motivações, tem-se dois estudos experimentais de interesse prático. Este capítulo está dividido em 7 partes. Pode-se encontrar na seção 2.2 uma revisão da classe de modelos para dados de contagem, utilizando as distribuições Poisson, binomial negativa, quase-Poisson, COM Poisson e Poisson-

Tweedie; nas seções 2.3 e 2.4 são apresentados os dois estudos de motivação, ambos entomológicos; na seção 2.5 apresentam-se os métodos utilizados no estudo; a seção 2.6 conta com os resultados e predições para cada estudo por meio do modelo selecionado e, por último, na seção 2.7, são apresentadas as discussões para ambos os experimentos.

2.2 Revisão de modelos para dados de contagem

Nesta seção apresenta-se uma revisão das classes de modelos para dados de contagem utilizando as distribuições Poisson, binomial negativa, quase Poisson, COM Poisson e Poisson Tweedie. Tais modelos, também conhecidos por Modelos Lineares Generalizados, foram propostos por Nelder e Wedderburn (1972) e são compostos por três componentes principais, sendo estas:

(i) **Componente aleatório:** váriáveis aleatórias independentes provenientes de uma mesma distribuição pertencente a família exponencial;

(ii) **Componente sistemático:** é o componente correspondente ao preditor linear, o qual combina as covariáveis e os parâmetros para a estimação dos resultados;

(iii) **Função de ligação:** função monótona e diferenciável que liga o componente aleatório ao sistemático.

2.2.1 Modelo de regressão Poisson

A distribuição Poisson é normalmente utilizada para dados de contagem ou dados em forma de tabelas de contingência (tabelas de dupla entrada). Considerando Y uma variável discreta com $Y \sim Poisson(\theta)$, sendo θ a taxa média de ocorrência de um evento, a função de probabilidade dessa variável é dada por:

$$p(y) = P[Y = y] = \frac{e^{-\theta}\theta^y}{y!}, \;\; y = 0, 1, 2, ... \tag{2.1}$$

Tal distribuição pertence à família exponencial canônica e pressupõe média e variância iguais, sendo ambas coincidentes com a taxa θ, visto que a distribuição de Poisson tem esperança e variância dadas por:

$$E[Y] = \theta = Var[Y] \tag{2.2}$$

Considerando $\boldsymbol{x} = (x_1, x_2, ...x_p)$ um vetor de variáveis explanatórias e $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ o vetor de parâmetros a serem estimados, o modelo de regressão Poisson é da forma:

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.3}$$

Com função de ligação logarítmica, o preditor linear é matricialmente definido por:

$$\eta = log(\mu) = \boldsymbol{x^T \beta} \tag{2.4}$$

2.2.2 Modelo de regressão binomial negativo

A distribuição binomial negativa, assim como a distribuição Poisson, é utilizada em dados de contagem e é resultado de uma mistura das distribuições Poisson e Gama (considerando-se $Y \sim Poisson(\theta)$ e θ uma variável aleatória com $\theta \sim Gama(k, \lambda)$, então $Y \sim BinNeg(\mu, k)$). Se Y uma variável discreta com $Y \sim BinNeg(\mu, k)$, sua função de probabilidade é definida por:

$$p(y) = P[Y = y] = \frac{\Gamma(k + y)}{\Gamma(k)y!} \frac{\mu^y k^k}{(\mu + k)^{K+y}} \tag{2.5}$$

em que: $y = 0, 1, ...$, $k > 0$ e $\mu > 0$.

A esperança e variância desta distribuição são:

$$E[Y] = \mu \tag{2.6}$$

$$Var[Y] = \mu + \frac{\mu^2}{k} \tag{2.7}$$

A vantagem em relação à distribuição Poisson é que a binomial negativa possui um termo adicional para a variância, tornando-a mais flexível para problemas de equidispersão.

Tal com o modelo de regressão Poisson, o modelo utilizando a distribuição binomial negativa pode ser descrito da seguinte forma:

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2.8}$$

em que $\boldsymbol{x} = (x_1, x_2, ...x_p)$ é o vetor de variáveis explanatórias e $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ é o vetor de parâmetros a serem estimados.

Com função de ligação logarítmica, tem-se o preditor linear definido por:

$$\eta = log(\mu) = \boldsymbol{x^T \beta} \tag{2.9}$$

2.2.3 Modelo de regressão quase-Poisson

Como uma alternativa para os casos em que $Var(Y) > E(Y)$ e o modelo Poisson não se ajusta, Wedderburn (1974) propôs os modelos de quase-verossimilhança, o qual é definido como um método de estimação que não necessita ter específica a distribuição da variável resposta no ajuste do modelo.

Tal modelo, conhecido como quase-Poisson, pode ser empregado para dados de contagem com superdispersão pois propõe uma relação entre a média e a variância, sendo a variância uma função linear da média definida por:

$$Var[Y] = \phi\theta \tag{2.10}$$

em que ϕ é um parâmetro adicional de dispersão e $E[Y] = \theta$.

O modelo de regressão quase-Poisson também utiliza a função de ligação logarítmica e, considerando $\boldsymbol{x} = (x_1, x_2, ...x_p)$ como o vetor de variáveis explanatórias e $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ como o vetor de parâmetros a serem estimados, tal modelo tem o preditor linear definido por:

$$\eta = log(\mu) = \boldsymbol{x^T \beta} \tag{2.11}$$

2.2.4 Modelo de regressão Conway-Maxwell-Poisson

A distribuição Conway-Maxwell-Poisson, também conhecida como COM Poisson (CP) foi proposta inicialmente por Conway e Maxwell (1962), sendo uma generalização da distribuição Poisson com o acréscimo de um parâmetro, sendo um destaque para análises de dados de contagem sub e superdispersos. Apesar de proposta em 1962, tal distribuição se tornou conhecida e amplamente utilizada após as contribuições de Shmueli et al. (2005) em seu estudo sobre as contribuições estatísticas da mesma. A função de probabilidade da distribuição CP é definida por:

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}, \ \ y = 0, 1, 2, ... \tag{2.12}$$

em que $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ é uma constante de normalização, $\lambda > 0$ representa a média e $\nu > 0$ o parâmetro de dispersão.

A esperança e variância para a distribuição CP não possuem formas fechadas e podem ser calculadas por meio de alguns métodos, como o dos momentos ou pela definição de valor esperado em variáveis aleatórias discretas. Sellers et al. (2011) definem a esperança e variância aproximadas para esta distribuição por:

$$E[Y] \approx \lambda^{\frac{1}{\nu}} - \frac{\nu - 1}{2\nu} \tag{2.13}$$

$$Var[Y] \approx (\frac{1}{\nu})^{\lambda^{\frac{1}{\nu}}} \tag{2.14}$$

Para o modelo de regressão COM Poisson, a função de ligação utilizada é a logarítmica e o preditor linear é dado por:

$$\eta = log(\mu) = \boldsymbol{x^T \beta} \tag{2.15}$$

sendo $\boldsymbol{x} = (x_1, x_2, ...x_p)$ o vetor de variáveis explanatórias e $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ o vetor de parâmetros a serem estimados.

2.2.5 Modelo de regressão Poisson-Tweedie

Os modelos hierárquicos foram utilizados pela primeira vez por Karl Pearson, quando foi proposto um modelo de misturas utilizando a função densidade de duas variáveis aleatórias com distribuições normais, contendo médias e variâncias diferentes (Pearson e Erdmann, 1894).

A distribuição Poisson-Tweedie é uma mistura hierárquica, ou também distribuição de dois estágios, proposta por Jørgensen (1997), e muito flexível para modelar dados com sub e superdispersão. Tal hierarquia é definida por:

$$Y|Z \sim Poisson(\theta) \tag{2.16}$$

e

$$Z \sim Tw_v(\mu, \omega) \tag{2.17}$$

em que Tw_v é a distribuição Tweedie com v graus de liberdade.

A distribuição não possui função de densidade de forma fechada, porém pode ser aproximada utilizando o método de Monte Carlo, no qual é possível definir a esperança e variância da distribuição como:

$$E[Y] = \mu \tag{2.18}$$

$$Var[Y] = \mu + \omega\mu^v \tag{2.19}$$

Assim como a COM Poisson, o modelo de regressão Poisson-Tweedie utiliza a função de ligação logarítmica e tem como preditor linear:

$$\eta = log(\mu) = \boldsymbol{x^T \beta} \tag{2.20}$$

em que $\boldsymbol{x} = (x_1, x_2, ...x_p)$ é o vetor de variáveis explanatórias e $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)$ é o vetor de parâmetros a serem estimados.

Levando em consideração as distribuições relatadas acima, tem-se definido na Tabela 2.1 um resumo para tais modelos.

Tabela 2.1: Quadro resumo dos modelos para dados de contagem apresentados na seção 2.2.

Modelo	$P(Y=y)$	$E[Y]$	$Var[Y]$	η
Poisson	$\frac{e^{-\theta}\theta^y}{y!}$	θ	θ	$log(\mu) = \boldsymbol{x^T\beta}$
Binomial negativo	$\frac{\Gamma(k+y)}{\Gamma(k)y!}\frac{\mu^y k^k}{(\mu+k)^{k+y}}$	μ	$\mu + \frac{\mu^2}{k}$	$log(\mu) = \boldsymbol{x^T\beta}$
COM Poisson	$\frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda,\nu)}$	$\approx \lambda^{\frac{1}{\nu}} - \frac{\nu-1}{2\nu}$	$\approx (\frac{1}{\nu})^{\lambda^{\frac{1}{\nu}}}$	$log(\mu) = \boldsymbol{x^T\beta}$
Poisson Tweedie		μ	$\mu + \omega\mu^\nu$	$log(\mu) = \boldsymbol{x^T\beta}$
quase-Poisson		θ	$\phi\theta$	$log(\mu) = \boldsymbol{x^T\beta}$

Pode-se observar que os preditores lineares para os cinco modelos apresentados são iguais e os modelos Poisson Tweedie e quase-Poisson não possuem função de probabilidade pois a distribuição Poisson Tweedie não possui forma fechada (pode-se obter uma aproximação pelo método de Monte Carlo) e quase-Poisson é definido como um método de estimação.

2.3 Estudo de motivação I

O primeiro estudo de motivação deste trabalho provém de um experimento entomológico desenvolvido por Sakuno (2021) na Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ) em casas de vegetação objetivando avaliar o comportamento de oviposição de *Diatraea saccharalis* em diferentes variedades de cana de açúcar.

A *Diatraea saccharalis*, também conhecida como broca da cana-de-açúcar, é considerada a principal praga da cana-de-açúcar. Tal inseto é, inicialmente, uma lagarta e em fase adulta é uma mariposa e pode ser encontrada em todo o território nacional. Os danos causados pela lagarta podem ser diretos (quebra de plantas, formação de galerias no interior do colmo, entre outros) ou indiretos (proliferação de fungos nas galerias formadas no interior do colmo) e causam prejuízos milionários (Gallo et al., 1988).

Nesse experimento, após 60 dias do plantio, os vasos com mudas de cana foram condicionados em gaiolas vedadas com tecido voil de $3 \times 2 \times 1$ metros. Nas gaiolas, 10 plantas foram dispostas em duas linhas paralelas (isto é, 5 plantas em cada linha), ilustrado na Figura 2.1. Foram estabelecidos três tratamentos: 1) NBT: apenas cana-de-açúcar não-BT (convencional) na gaiola; 2) BT: apenas cana-de-açúcar BT (transgênica) na gaiola e 3) MIX: cana-de-açúcar BT e NBT combinadas em uma mesma gaiola.

Figura 2.1: Gaiola experimental envolta com tecido voil para o estudo do padrão de distribuição de ovos de *Diatraea saccharalis* em vasos com cana-de-açúcar conduzido por Sakuno em 2021, contendo 10 mudas de cana-de-açúcar NBT.

Nas gaiolas foram feitas 6 marcações equidistantes com $0,5$m uma das outras, possibilitando a avaliação da posição das mariposas ao longo do tempo de observação. Foram liberadas 5 fêmeas adultas coradas e 5 fêmeas adultas sem coloração, totalizando 10 fêmeas no ponto 0 da gaiola. Após a liberação das fêmeas na gaiola, a cada 24 horas, foi observado o número de mariposas presente na área delimitada entre as 6 marcações da gaiola (isto é, quadrantes), bem como o número de mariposas presentes em cada uma das plantas. Ao longo das observações, o número de posturas colocadas pelas fêmeas e a posição

da planta em que os ovos foram encontrados também foram avaliados e quantificados. O experimento obedeceu a um dealineamento inteiramente casualizado e foi replicado 10 vezes para cada tratamento.

Nesse estudo, a variável resposta em análise foi o número de posturas colocadas pelas mariposas em cada uma das variedades e, como variáveis explanatórias, foram observados o vaso (ou a distância percorrida desde o ponto de soltura), a posição (lados direito e esquerdo), os tratamentos (BT, NBT e MIX) e as variedades de cana-de-açúcar (convencional ou BT).

2.4 Estudo de motivação II

O segundo experimento, conduzido por Reigada (2009), teve como objetivo estimar a frequência de parasitismo de espécies parasitóides sobre hospedeiros e, de forma prática, é utilizado para fins de controle biológico, fenômeno que consiste no controle de pragas e insetos indesejados.

O experimento foi realizado com quatro espécies de parasitóides (espécie da família Diapriidae, *Nasonia vitripennis*, *Pachycrepoideus vindemiae*, *Spalangia endius*) e cinco espécies hospedeiras (*Chrysomya albiceps*, *Chrysomya putoria*, *Cochliomyia macellaria*, *Lucilia sericata*, *Chrysomya megacephala*). As fêmeas parasitóides, previamente alimentadas com água e mel, foram inseridas em recipientes plásticos com 20 cm de altura e 15 cm de diâmetro, onde haviam cinco placas de Petri com 5 pupas das espécies hospedeiras (cada placa com apenas uma espécie definidas de forma aleatória, e pupas com cerca de 24 h de idade), conforme ilustrado na Figura 2.2.

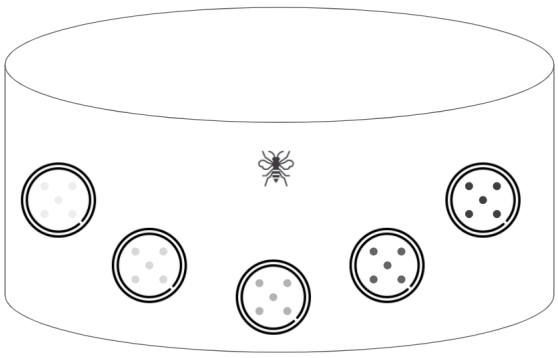


Figura 2.2: Croqui do experimento do parasitismo sobre vespas em pupas de moscas varejeiras desenvolvido por Reigada em 2009.

As fêmeas foram mantidas nos recipientes com as pupas hospedeiras por 24 h em bancadas iluminadadas, com ambiente controlado em $25 \pm 1°C$ e umidade relativa de 70%. Ao final das 24 h, foram registradas as quantidades de pupas parasitadas em cada uma das placas. O experimento foi replicado 10 vezes para cada interação parasitóide/hospedeiro, tendo como variável resposta o número de pupas parasitadas e as variáveis explanatórias se resumem as espécies parasitóides e hospedeiras.

2.5 Métodos

Para ambos experimentos, primeiramente, é realizada uma análise exploratória, a fim de compreender a distribuição da variável resposta e a relação média-variância. Posteriormente, ajustam-se alguns modelos utilizando as distribuições para a variável resposta, como: Poisson, binomial negativa, quase-Poisson, COM Poisson e/ou Poisson-Tweedie, a fim de se selecionar uma estrutura que melhor represente a relação média-variância.

Para o primeiro experimento, o preditor linear para o modelo completo é da forma:

$$\eta = \beta_0 + \beta_1 Vaso + \beta_2 Planta + \beta_t Trat_t + \beta_5 Posicao, \tag{2.21}$$

em que β_0 representa uma constante geral; β_1 representa o parâmetro associado ao efeito de Vaso; β_2 representa o parâmetro associado ao efeito de Planta; β_t representa o parâmetro associado ao efeito do t-ésimo Tratamento, com $t = 3, 4$; β_5 representa o parâmetro associado ao efeito de Posição.

No segundo experimento, o preditor linear para o modelo completo é definido por:

$$\eta = \beta_0 + \beta_r Parasitoide_r + \beta_h Hospedeiro_h, \tag{2.22}$$

em que β_0 representa uma constante geral; β_r representa o parâmetro associado ao efeito do r-ésimo parasitóide, com $r = 1, 2, 3$ relacionado aos parasitóides *N. vitripennis*, *P. vindemiae* e *S. endius*; β_h representa o parâmetro associado ao efeito do h-ésimo hospedeiro, com $h = 4, 5, 6, 7$ relacionado aos hospedeiros *C. macellaria*, *C. megacephala*, *C. putoria* e *L. eximia*.

Nos dois estudos, os parâmetros são estimados pelo método da máxima verossimilhança (EMV), introduzida por Fisher (1922). Considere θ o parâmetro de interesse e $X_1, ..., X_n$ n variáveis aleatórias com função de densidade dada por $f(X_1, ..., X_n; \theta)$. O EMV de θ é dado por:

$$\hat{\theta} =_{argmax} L(\theta; X_1, ..., X_n), \tag{2.23}$$

em que $L(\theta; X_1, ..., X_n) = f(X_1, ..., X_n; \theta)$ representa a função de verossimilhança, com $\theta \in \Theta$, com Θ sendo o espaço paramétrico. Nesse estudo os parâmetros são estimados por processo iterativo, uma vez que o EMV não tem forma fechada.

Após a estimação, os modelos ajustados são comparados pelo critério de Akaike (Akaike, 1974), no qual quanto menor o valor, melhor o ajuste. A estatística do AIC é definida por:

$$AIC = 2p_m - 2ln(\widehat{L_m}), \tag{2.24}$$

em que: p_m é o número de parâmetros e $\widehat{L_m}$ é o valor máximo da função de máxima verossimilhança para o m-ésimo modelo.

Adicionalmente, são utilizados os gráficos meio normal de probabilidade com envelope simulado (*half normal plot*) disponíveis no pacote *hnp* (Moral et al., 2017) para verificar os ajustes.

Após a escolha do(s) modelo(s) mais adequado(s), são feitas as predições da variável resposta do experimento e constroem-se intervalos de confiança de 95% (IC_{95}) para o parâmetro μ, estimados por:

$$P(\hat{\mu} - 1,96\sqrt{V(\hat{\mu})} \leq \mu \leq \hat{\mu} + 1,96\sqrt{V(\hat{\mu})}) = 0.95. \tag{2.25}$$

Todas as análises são realizadas utilizando o software estatístico R Core Team (2019), aplicando os pacotes base, *MASS* (Venables e Ripley, 2002), *COMPoissonReg* (Sellers e Lotze, 2011) e *tweedie* (Dunn, 2017), com um nível de significância de $\alpha = 0,05$.

2.6 Resultados

2.6.1 Estudo de motivação I

Com o objetivo de entender a distribuição de ovos por *Diatraea saccharalis*, realizou-se a análise exploratória e as medidas descritivas para os tratamentos e tipos de planta estão exibidos na Tabela 2.2.

Tabela 2.2: Medidas descritivas para o número de posturas colocadas por *Diatraea saccharalis* em relação aos tratamentos no estudo de oviposição.

Tratamento	Mínimo	Média	Mediana	Máximo	Desvio Padrão	Total
NBT	0	8,47	5,0	43	9,54	847
MIX	0	5,60	3,0	34	6,77	560
BT	0	7,11	3,5	55	10,07	711

A média, mediana e a soma total de ovos são maiores para o tratamento NBT, quando comparado aos outros tratamentos. A média de oviposição neste tratamento é, no mínimo, 20% acima dos demais. Quando observados somente os tipos de plantas, o tipo convencional também possui número de ovos postos superiores (totalizando 1.121 posturas) e um desvio padrão menor quando comparados ao tratamento BT (desvio padrão de $8,88$ para as plantas convencionais contra $9,05$ para as plantas BT). Pela distribuição de ovos postos durante o período de 72h observa-se que, assim como a média, a maioria das posturas está em até 10 ovos, conforme apresentado na Figura 2.3, havendo excessões para posturas com mais de 30 ovos.

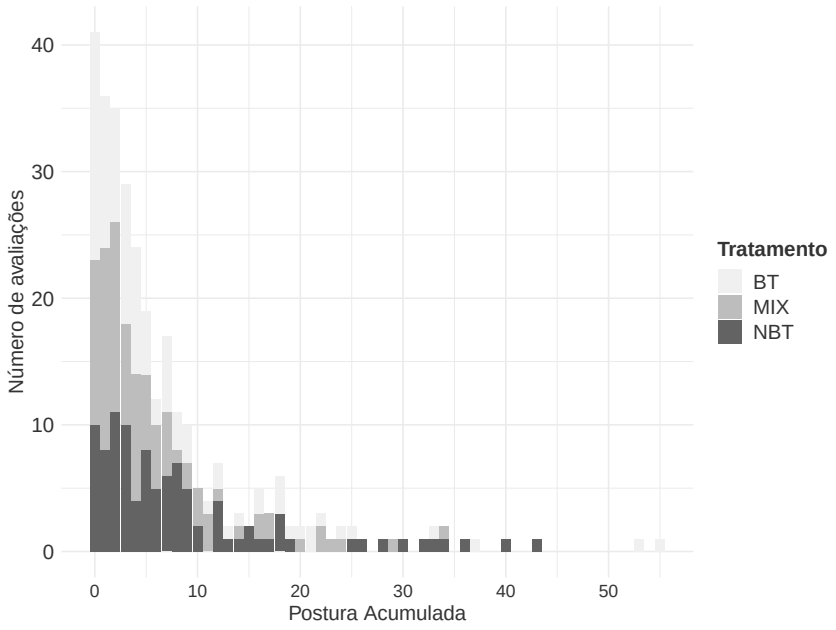
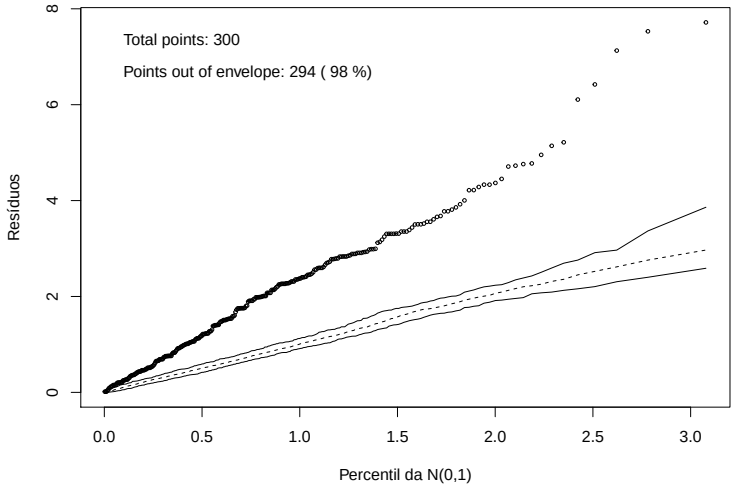
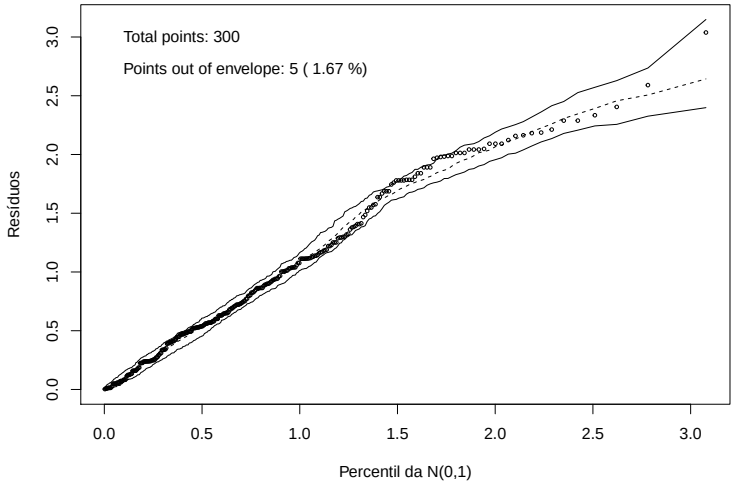


Figura 2.3: Distribuição do número de ovos postos por *Diatraea saccharalis* em cada tratamento do experimento.

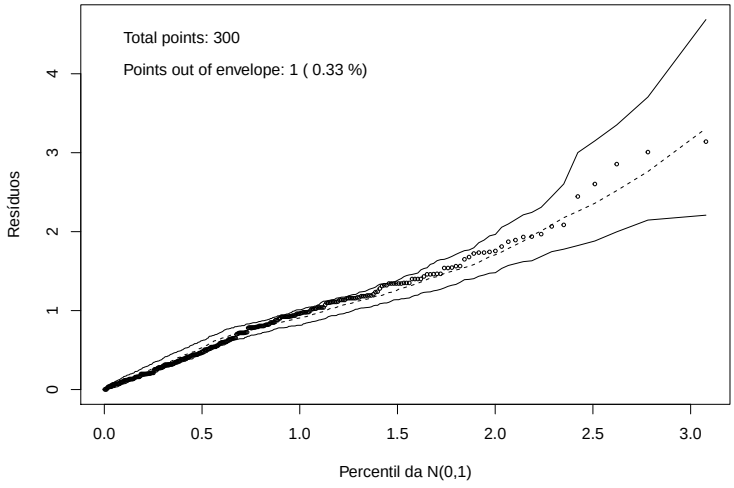
Em seguida, foram ajustados três modelos contendo as quatro variáveis, conforme descrito na seção 2.5, sendo esses o modelo de Poisson, binomial negativo e quase-Poisson. Para avaliar os ajustes dos modelos, foi realizada uma comparação pelo AIC. O modelo ajustado utilizando a distribuição binomial negativa possui o menor valor de AIC, sendo este de 1716, enquanto o modelo com a distribuição de Poisson possui um AIC de 2525. O modelo quase-Poisson não possui AIC, visto que tal distribuição não possui uma verossimilhança. Outra maneira de avaliar a qualidade do ajuste é utilizando o gráfico meio normal de probabilidade com envelope simulado, apresentados na Figura 2.4.



(a) *Half-normal plot* para o modelo Poisson



(b) *Half-normal plot* para o modelo binomial negativo



(c) *Half-normal plot* para o modelo quase-Poisson

Figura 2.4: *Half-normal plot* para os modelos ajustados aos dados de oviposição de *Diatraea saccharalis*.

Novamente, para o modelo Poisson, o gráfico referente a Figura 2.4 (a) possui quase todos os resíduos observados fora do envelope de simulação, mostrando que a distribuição não possui um ajuste satisfatório. Os outros dois modelos se apresentam bem ajustados neste caso, com apenas 5 e 1 pontos fora do envelope para as distribuições binomial negativa e quase-poisson, respectivamente.

Utilizando o teste de razão de verossimilhanças, constatou-se que a variedade de planta e a posição na gaiola não são variáveis significativas para explicar o comportamento da postura de ovos acumulada em $72h$. De tal forma que as covariáveis que explicam a postura dos ovos são as variáveis Vaso e Tratamento.

As estimativas para os parâmetros associados a estas variáveis explicativas, para as estruturas binomial negativa e quase-Poisson são apresentadas na Tabela 2.3.

Tabela 2.3: Estimação dos parâmetros para os modelos selecionados no experimento de oviposição de *Diatraea saccharalis*

Parâmetro	Binomial negativa			quase-Poisson		
	Estimativa	E. Padrão	p-valor	Estimativa	E. Padrão	p-valor
β_0	3,167	0,15	<0,0001	3,42	0,13	<0,0001
β_1	-0,399	0,04	<0,0001	-0,512	0,04	<0,0001
β_3	-0,412	0,14	0,003	-0,414	0,14	0,002
β_4	-0,250	0,14	0,072	-0,175	0,13	0,166
ϕ	1,262			6,149		

Com tais modelos ajustados, pode-se realizar a predição para cada um dos casos do experimento por meio dos intervalos de confiança para as médias, representados na Figura 2.5.

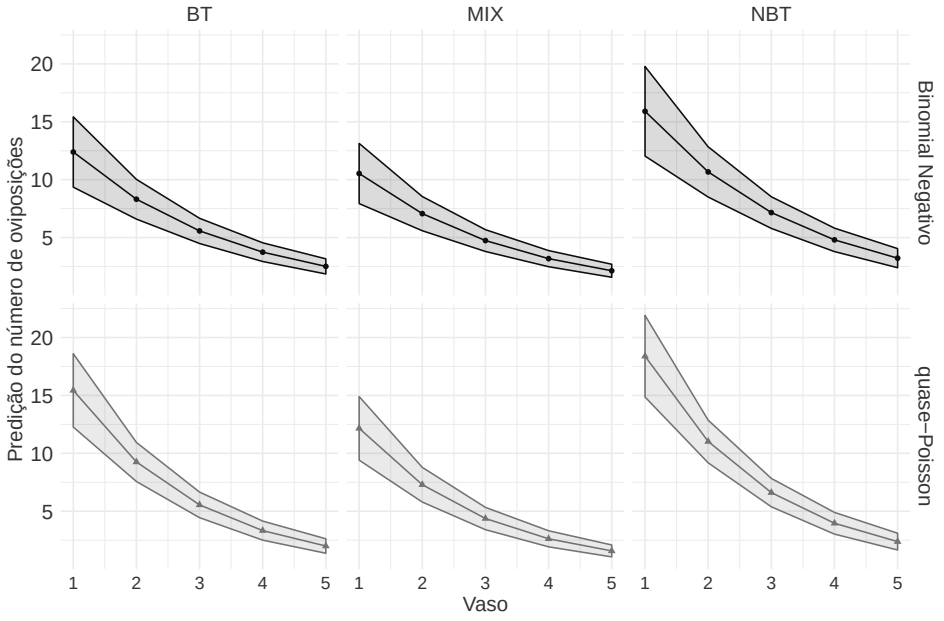


Figura 2.5: Intervalo de confiança (95%) para a predição do número de ovos postos nos modelos ajustados.

Observa-se que, para ambos os modelos, os valores preditos e seus respectivos intervalos de confiança possuem valores semelhantes, sendo ambos os modelos bons preditores para o número de ovos postos em cada uma das posições de vaso.

O tratamento NBT, assim como visto nas medidas de dispersão, possui maiores números de ovos postos preditos em todos os vasos, seguido pelo tratamento BT e MIX, respectivamente.

2.6.2 Estudo de motivação II

A distribuição de parasitismo por espécie hospedeira pode indicar se existe alguma preferência das fêmeas parasitóides pelas espécies hospedeiras. Visando entender tal ponto, as medidas descritivas para o número de pupas parasitados por cada espécie são apresentadas na Tabela 2.4.

Tabela 2.4: Medidas descritivas para o número de pupas parasitados em relação as espécies parasitóides no experimento conduzido por Reigada em 2009.

Parasitóide	Mínimo	Média	Mediana	Máximo	Desvio Padrão	Total
Diapriidae	0	0,82	0	5	1,41	41
Nasonia vitripennis	0	1,66	1	5	1,88	83
Pachycrepoideus vindemiae	0	2,66	3	5	1,92	133
Spalangia endius	0	1,98	1	5	2,10	99

Conforme descreve a Tabela 2.4, parasitóides da espécie *Pachycrepoideus vindemiae* possuem maior taxa média de parasitismo, 34% maior que a espécie *Spalangia endius*, com média próxima a 2 pupas parasitadas. Para entender a mesma questão em relação aos hospedeiros, a Tabela 2.5 exibe as medidas descritivas para o número de pupas parasitados por hospedeiro.

Tabela 2.5: Medidas descritivas para o número de pupas parasitados em relação as espécies hospedeiras no experimento.

Hospedeiro	Mínimo	Média	Mediana	Máximo	Desvio Padrão	Total
C. albiceps	0	1,27	0,0	5	1,66	51
C. macellaria	0	1,30	0,0	5	2,03	52
C. megacephala	0	2,58	2,5	5	1,52	103
C. putoria	0	2,80	4,0	5	2,15	112
L. eximia	0	0,95	0,0	5	1,63	38

As espécies hospedeiras *C. putoria* e *C. megacephala* foram as mais escolhidas pelas espécies parasitóides, com médias de $2,80$ e $2,58$ pupas parasitados, respectivamente. A espécie *L. eximia* foi a que obteve menor quantidade de pupas parasitados, com média de apenas $0,95$ pupas. A preferência de cada espécie também pode ser observada na Figura 2.6.

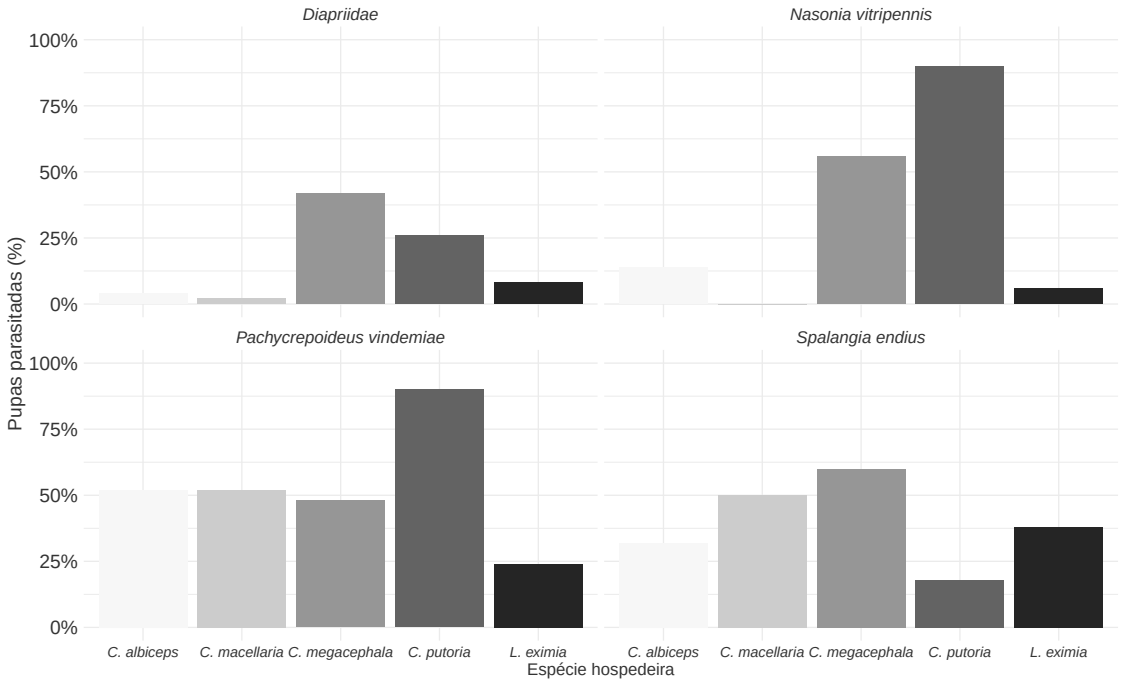


Figura 2.6: Taxas de parasitismo de cada espécie parasitóide em cada espécie hospedeira no experimento de Reigada em 2009.

Pachycrepoideus vindemiae se destaca, chegando a parasitar 90% das pupas de *C. putoria*, cenário que também pode ser visto para *Nasonia vitripenis* que, em contrapartida, não teve taxa de parasitismo alta nas outras espécies. A espécie Dipriidae foi a que obteve menor taxa de parasitismo para quase todas as espécies, sendo considerada a pior espécie parasitóide neste caso.

A seguir, ajustaram-se três modelos para entender o comportamento de parasitismo das espécies de vespas nas pupas de moscas varejeiras. As distribuições utilizadas nos ajustes são a Poisson, COM Poisson e Poisson-Tweedie.

Pelo AIC, o modelo Poisson-Tweedie possui o melhor ajuste, ou seja, menor valor de AIC (664), enquanto o modelo com a distribuição de Poisson teve o maior valor no critério (728). Sendo assim, seleciona-se o modelo Poisson-Tweedie e estimativas dos parâmetros para este modelo são apresentadas na Tabela 2.6.

Tabela 2.6: Estimativas dos parâmetros do modelo Poisson-Tweedie ajustado ao número de pupas parasitadas no experimento.

	Estimativa	E. Padrão	p-valor
β_0	-0,561	0,275	0,0427
β_1	0,698	0,261	0,0080
β_2	1,199	0,244	<0,0001
β_3	0,918	0,252	0,0003
β_4	0,015	0,271	0,9569
β_5	0,729	0,237	0,0024
β_6	0,802	0,234	0,0007
β_7	-0,283	0,291	0,3310
ϕ	1,808		

Com o modelo Poisson-Tweedie ajustado, é possível realizar as predições para todos os cenários do experimento, por meio dos respectivos intervalos de confiança de 95%. Estas projeções para cada parasitóide e hospedeiro podem ser vistas na Figura 2.7.

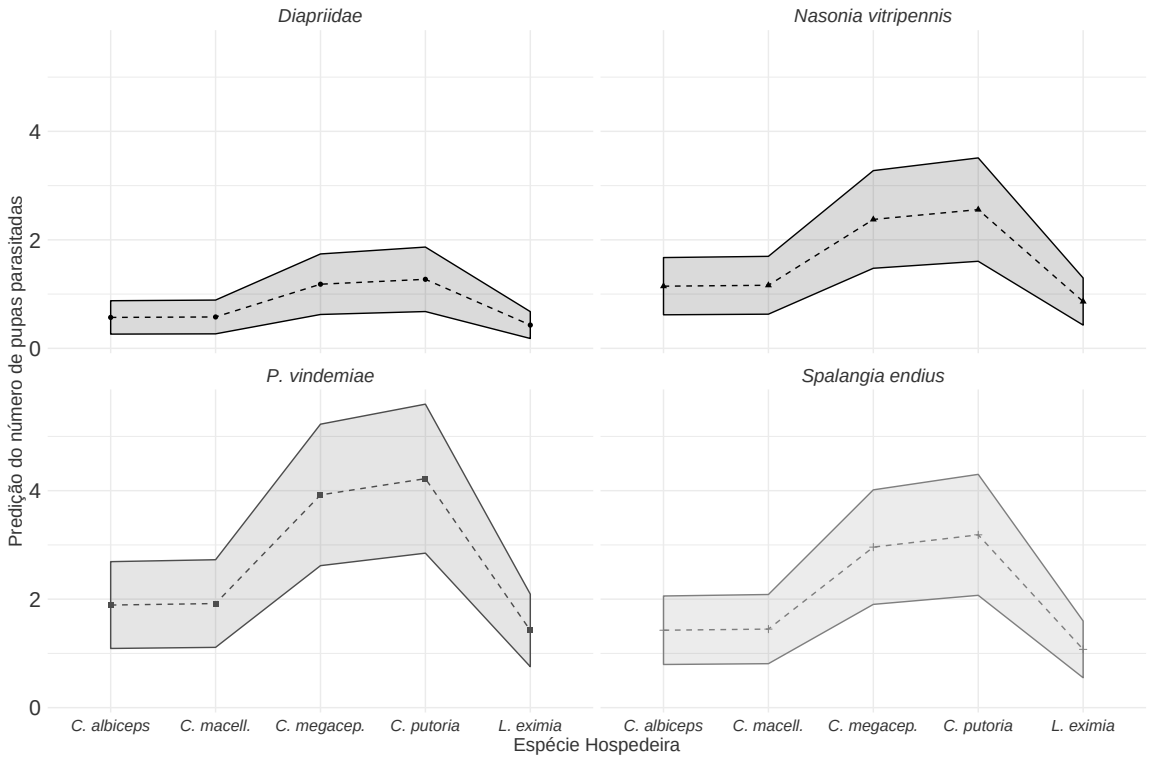


Figura 2.7: Intervalo de confiança (95%) para a predição do número de pupas parasitadas.

Por meio da análise da Figura 2.7, ratifica-se que as espécies hospedeiras *C. albiceps*, *C. macellaria* e *L. eximia* possuem menores valores preditos e as outras duas espécies hospedeiras (*C. megacephala* e *C. putoria*) possuem maiores valores nas predições. Além disso, a espécie parasitóide *Diapriidae* é a espécie com menores valores preditos, independente do hospedeiro, e a espécie *Pachycrepoideus vindemiae* é a espécie parasitóide com maior destaque, possuindo maiores projeções de pupas parasitadas.

2.7 Discussão

Nas duas situações práticas consideradas, os modelos que utilizam distribuições Poisson não foram tão satisfatórios, isso se deve ao fato de tal distribuição exigir algumas pressuposições, como a equidispersão dos dados. As demais distribuições consideradas neste trabalho, como binomial negativa, quase-Poisson ou Poisson-Tweedie são mais flexíveis para os ajustes dos modelos e resultam em qualidades melhores que a distribuição Poisson.

Para o primeiro estudo de motivação, as distribuições binomial negativa e quase-Poisson sugerem resultados semelhantes e satisfatórios, mostrando que é possível realizar uma predição da contagem de ovos utilizando apenas o Tratamento e o Vaso (que também pode ser representado pela distância percorrida dentro da gaiola a partir do ponto de soltura). O Tratamento significativo representa a preferência das mariposas por alguma variedade e o Vaso (ou distância percorrida) como um fator significativo que pode estar relacionado com a baixa distância percorrida por estes insetos, como observado nos estudos de Francischini et al. (2019) e Caixeta (2010), no caso de mariposas de gênero masculino.

No segundo estudo de motivação, a distribuição Poisson-Tweedie foi a que obteve o melhor ajuste, resultando em um menor AIC. Tal distribuição possui um ajuste satisfatório devido à estrutura da variável resposta no conjunto, pois a mesma é composta por excessos de zeros e tal distribuição consegue se ajustar bem com este tipo de problemática.

Referências

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19:716–723.

Batista, D. T. (2020). *Modelos para dados de contagem não equidispersos com aplicação à ecologia e em estudos longitudinais*. PhD thesis, Escola Superior de Agricultura Luiz de Queiroz.

Bliss, C. (1985). The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22:134–167.

Caixeta, D. F. (2010). *Dispersão de Machos de Diatraea saccharalis (Fabricius) (Lepidoptera: Crambidae) em Cana-de-açúcar*. PhD thesis, Faculdade de Ciências Agrárias e Veterinárias – Universidade Estadual de São Paulo.

Cesnik, R. (2007). Melhoramento da cana-de-açúcar: marco sucro-alcooleiro no Brasil. *ComCiência - SBPC/Labjor Brasil*, pages 1–4.

Chichera, R. A., Pereira, F. F., Kassab, O., Barbora, R. H., Pastori, P. L., e Rossoni, C. (2012). Capacidade de busca e reprodução de Trichospilus diatraeae E Palmistichus elaeisis (Hymenoptera: Eulophidae) em pupas de Diatraea saccharalis (Lepidoptera: Crambidae). *Interciencia*, 37:852–856.

Conway, R. e Maxwell, W. (1962). A queuing model with state dependent service rates. *Journal of Industrial Engineering*, 12:132–136.

Dunn, P. K. (2017). *Tweedie: Evaluation of Tweedie Exponential Family Models*. R package version 2.3.0.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, (A):309–368.

Francischini, F. J. B., Cordeiro, E. M. G., Campos, J. B., Alves-Pereira, A., Viana, J. P. G., e Wu, X. (2019). Diatraea saccharalis history of colonization in the Americas. The case for human-mediated dispersal. *PLoS ONE*, 14.

Gallo, D., Nakano, O., Silveira Neto, S., Carvalho, R. P. L., Baptista, G. C. d., Berti Filho, E., Parra, J. R. P., Zucchi, R. A., Alves, S. B., e Vendramim, J. D. (1988). *Manual de entomologia agrícola*. Agronômica Ceres, São Paulo, 2 edition.

Girón Pérez, K. (2013). *Suscetibilidade de Diatraea saccharalis a Cry1Ab e comportamento larval em cana-de-açúcar*. PhD thesis, Universidade Federal de Viçosa.

Jørgensen, B. (1997). The Theory of Dispersion Models. *Monographs on Statistics and Applied Probability. London: Chapman & Hall*.

Moral, R. A., Hinde, J., e Demétrio, C. G. B. (2017). Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software*, 81:23p.

Nelder, J. A. e Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistics Society*, 135:370–384.

Pearson, K. e Erdmann, H. O. M. F. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 186:343–414.

Poisson, S. D. (1837). Recherches sur la probabilité des jugements en matière criminelle et en matière civile.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramalho, J. J. D. S. (1996). *Modelos de Regressão para Dados de Contagem*. PhD thesis, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa.

Reigada, C. (2009). *Dinâmica tritrófica experimental em populações de moscas varejeiras*. PhD thesis, Universidade Estadual Paulista.

Rocha, E. B., Leandro, R. A., Demétrio, C. G. B., Amaral, S. W. G., e Ribeiro, P. J. (2014). Aplicação dos modelos lineares generalizados na análise do número de estômatos em coentro (Coriandrum sativum L.): estimação bayesiana utilizando INLA. *Revista da Estatística UFOP*, 3:212.

Sakuno, C. I. R. (2021). Efeitos da movimentação de *Diatraea saccharalis* (fabricius) (lepidoptera: Crambidae) em sistemas compostos por cana-de-açúcar geneticamente modificada e refúgio. Master's thesis, Tese de Mestrado em Entomologia. Escola Superior de Agricultura "Luiz de Queiroz".

Sellers, K. e Lotze, T. (2011). *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*. R package version 0.3.4.

Sellers, K. F., Borle, S., e Shmueli, G. (2011). The com-poisson model for count data: a survey of methods and applications. *Applied Stochastic Models in Business and Industry*, 28:104–116.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., e Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society*, 54:127–142.

Teles, J. (1995). Modelos lineares generalizados - uma aplicação à medicina. Master's thesis, Tese de Mestrado em Estatística e Investigação Operacional. Faculdade de Ciência da Universidade de Lisboa.

Terra, W. R., Costa, R. H., e Ferreira, C. (2006). Plasma membranes from insect midgut cells. *Academia Brasileira de Ciências*, 78:255–269.

Venables, W. N. e Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447.

3 MODELOS PARA DADOS DE CONTAGEM CATEGORIZADOS: UM ESTUDO DO COMPORTAMENTO DE ANIMAIS E INSETOS

Resumo

Um tipo de variável resposta muito presente em experimentos zoológicos e entomológicos são as categorizadas nominais pois, normalmente, é de interesse do pesquisador entender o comportamento ou preferência por alguma categoria específica de algum animal e/ou inseto. Para estudar a abordagem de modelos para essas variáveis respostas, apresentam-se dois estudos de motivação, sendo o primeiro referente à preferência de sitios de oviposição de duas espécies de vespas parasitóides e um segundo estudo sobre o comportamento de suínos em ambientes com ou sem enriquecimento. O objetivo do presente trabalho é revisar modelos para dados categorizados com três ou mais categorias. Em ambos os casos, os ajustes de modelos foram comparados utilizando o teste da razão de verossimilhanças (TRV) e a qualidade destes ajustes foi avaliada pelo critério de informação de Akaike (AIC). Com base nos estudos de motivação, observou-se que o modelo multinomial pode ser adequado para alguns casos, como o primeiro experimento de motivação, porém, em outros cenários, os modelos alternativos Dirichlet-multinomial e multinomial negativo apresentam ajustes mais satisfatórios, como no segundo experimento, o qual o modelo com a mistura Dirichlet-multinomial teve o melhor ajuste.

Palavras-chave: Modelos de mistura, Dirichlet-Multinomial, Parasitismo de *Platygastridae*, Enriquecimento ambiental.

3.1 Introdução

Na ciência é frequente deparar com estudos cuja variável resposta é categorizada pois tais tipos de dados são importantes em análises comportamentais e patológicos, como em psicologia com processos cognitivos (Riefer e Batchelder, 1988), na medicina com a avaliação de pacientes com esquizofrenia (Keefe et al., 1999) e nas ciências agrárias com fruticultura e zootecnia (Salvador, 2019).

Especificamente nas ciências agrárias, entender características, comportamentos, preferências, fisiologia ou outros fatores de animais e insetos são o interesse de estudos zoológicos e entomológicos, respectivamente. Frequentemente, quando se objetiva entender o comportamento desses indivíduos, são utilizadas variáveis respostas categorizadas, no qual os indivíduos são avaliados e classificados em categorias mutuamente exclusivas, estas, podendo ser dicotômicas (com apenas duas categorias de classificação) ou politômicas (com três ou mais categorias de classificação).

A análise de dados categorizados, também conhecida como análise de dados discretos (pois normalmente a variável resposta é associada à uma distribuição discreta de probabilidade (Giolo, 2017)), requer técnicas apropriadas que também dependem de sua natureza. Variáveis categorizadas, além de serem classificadas como dicotômicas ou politômicas, podem ser classificadas como ordinais, se seguem uma ordem natural, ou nominal, caso não possuam tal ordenação (Agresti, 2002).

Quando a variável resposta é politômica os modelos usualmente empregados são o modelo dos logitos generalizados ou o modelo de chances proporcionais, para os casos nominal e ordinal, respectivamente. Tratando-se de uma variável resposta politômica, a principal distribuição de probabilidade assumida para a variável resposta é a distribuição multinomial (Agresti, 2002), uma extensão da distribuição binomial, amplamente utilizada em casos dicotômicos (sucesso e fracasso).

Existem outras classes de distribuições que podem ser assumidas para a variável resposta, como a multinomial negativa (Bates e Neyman, 1952) e Dirichlet-multinomial (Mosimann, 1962), sendo estas mais úteis para problemas relacionados à superdispersão. Em Salvador (2019) faz-se uma discussão completa dessas distribuições aplicadas a casos de superdispersão em estudos agrários.

Este trabalho tem como objetivo estudar e comparar modelos para variáveis categorizadas politômicas nominais com aplicações em entomologia e zootecnia com base em dois experimentos motivacionais. O primeiro é um estudo entomológico que visa compreender o parasitismo de vespas em três categorias de ovos e o segundo tem como objetivo compreender o comportamento de suínos durante um período de tempo em duas condições ambientais. Os ajustes de modelos foram comparados utilizando o critério de informação de Akaike (AIC) (Akaike, 1974) e, adicionalmente, o gráfico meio normal de probabilidade com envelope simulado (*half normal plot*) (Moral et al., 2017).

Este capítulo está dividido em seções, organizadas da seguinte maneira: na seção 3.2 apresenta-se uma revisão de modelos para dados categorizados, descrevendo o modelo dos logitos generalizados, o modelo de regressão multinomial negativo e o modelo de regressão Dirichlet-multinomial; nas seções 3.3 e 3.4 apresentam-se os dois estudos de motivação, sendo o primeiro na área de entomologia e o segundo em zootecnia; na seção 3.5 encontra-se a metodologia utilizada nas análises, com os preditores lineares para cada experimento; a seção 3.6 conta com os resultados dos ajustes dos modelos para os dois estudos de motivação e, por fim, na seção 3.7 tem-se a discussão dos resultados apresentados na seção anterior.

3.2 Revisão de modelos para dados categorizados

Dados categorizados podem ser classificados em dicotômicos se possuem dois níveis ou politômicos quando possuem três ou mais níveis. Quando se trata de uma variável politômica, a distribuição multinomial é usualmente assumida para a estrutura aleatória dos modelos empregados (Giolo, 2017).

O presente trabalho refere-se aos dados politômicos nominais, que não possuem ordem natural, e para os quais podem-se assumir distribuições apropriadas, como a distribuição multinomial, multinomial negativa e Dirichlet-multinomial.

3.2.1 Modelo dos logitos generalizados

A distribuição multinomial proposta por volta de 1700 pode ser caracterizada como uma generalização multidimensional da distribuição binomial. Esta segunda foi proposta por Jacob Bernoulli em 1713 em seu estudo *Ars Conjectandi* (A arte de conjecturar - tradução nossa) e é uma generalização da distribuição de Bernoulli para n eventos, os quais podem ter dois resultados (sucesso, representado por 1, ou fracasso, representado por 0), com uma probabilidade de sucesso p. Tal distribuição, assim como a multinomial e outras distribuições, foram sintetizadas por Forbes et al. (2010).

Considera-se $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_j)$ um vetor de variáveis aleatórias resultante da realização de n ensaios multinomiais independentes em que π_j é a probabilidade associada à j-ésima categoria. A distribuição multinomial verifica a probabilidade de alguma combinação de números de sucessos para cada um dos n ensaios e sua distribuição de probabilidade é dada por:

$$P[Y_1 = n_1, ..., Y_j = n_j] = \frac{n!}{n_1!...n_j!}\pi_1^{n_1}\pi_2^{n_2}...\pi_j^{n_j} \tag{3.1}$$

em que n_j é o número de ocorrências da categoria j nas n realizações do experimento multinomial.

Essa distribuição é denotada por $\boldsymbol{Y} \sim multi(n, \boldsymbol{\pi})$. Para o vetor de variáveis aleatórias Y_j, os valores esperados para a j-ésima variável componente é:

$$E[Y_j] = n\pi_j \tag{3.2}$$

E as variâncias e covariâncias são definidas por:

$$Var[Y_j] = n\pi_j(1 - \pi_j) \tag{3.3}$$

$$Cov[Y_j, Y_k] = -n\pi_j\pi_k \tag{3.4}$$

em que $j \neq k$ e $j, k = 1, ..., J$

O modelo dos logitos generalizados ou modelo de regressão multinomial tem como objetivo ajustar modelos para dados com 3 ou mais variáveis respostas categorizadas. Para tal, uma categoria é escolhida como referência e é comparada com as demais, sendo utilizado, normalmente, o conhecimento do pesquisador para definir qual a categoria deve ser escolhida.

Considere um experimento com J categorias ($j = 1, 2, ..., J$). Sendo $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_p})$ um vetor de variáveis resposta e $\boldsymbol{\beta_j} = (\beta_{j1}, \beta_{j2}, ..., \beta_{jp})$ como o vetor de parâmetros a serem estimados e $\pi_j(x) = P(Y = j|\boldsymbol{x})$, utilizando a categoria J como referência, Agresti (2002) define este modelo por:

$$\eta = ln\left(\frac{\pi_p}{\pi_J}\right) = \beta_{j0} + \beta_{j1}x_1 + ... + \beta_{jp}x_p = \boldsymbol{x^T\beta_P}, \tag{3.5}$$

em que $j = 1, ..., J-1$ e os parâmetros de regressão ($\boldsymbol{\beta}$) são diferentes para cada categoria de resposta.

Ainda, segundo Agresti (2002), para realizar a estimativa dos parâmetros de regressão é utilizado o método da máxima verossimilhança via Newton-Raphson, visto que a função de verossimilhança para este modelo não possui forma analítica fechada.

3.2.2 Modelo de regressão multinomial negativo

A distribuição multinomial negativa pode ser descrita como uma generalização da distribuição binomial negativa para o caso dicotômico (Le Gall, 2006) e foi apresentada originariamente por Bates e Neyman (1952), como binomial negativa multivariada.

Neste contexto, considere $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_j)$ um vetor aleatório com distribuição multinomial negativa. A distribuição multinomial negativa tem sua função de probabilidade definida por:

$$P[Y_1, ..., Y_j] = \Gamma\left(\kappa + \sum_{j=1}^{J} y_j\right) \frac{\pi_0^v}{\Gamma(\kappa)} \prod_{j=1}^{J} \frac{\pi_j^{y_j}}{y_j!} \tag{3.6}$$

em que: κ representa o número de ocorrências; π_j é a probabilidade de sucesso para o j-ésimo ensaio; π_0 é a probabilidade pertencente a classe de referência; $\kappa > 0$.

Os valores esperados, as variâncias e as covariâncias são definidos por:

$$E[Y_j] = \kappa \frac{\pi_j}{\pi_0} \tag{3.7}$$

$$Var[Y_j] = \kappa \frac{\pi_j(\pi_0 + \pi_j)}{\pi_0^2} \tag{3.8}$$

$$Cov[Y_j, Y_k] = -\kappa \frac{\pi_j \pi_k}{\pi_0^2} \tag{3.9}$$

em que $j, k = 1, ..., J$ e $j \neq k$.

O modelo de regressão utilizando essa distribuição é definido por:

$$\eta = ln(\beta_{j0} + \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}), \tag{3.10}$$

3.2.3 Modelo de regressão Dirichlet-multinomial

A distribuição Dirichlet-multinomial (Mosimann, 1962) é uma mistura hierárquica, tal como proposta por Pearson e Erdmann (1894), utilizando a função de duas variáveis aleatórias, no caso: a multinomial e a Dirichlet. A mistura hierárquica das distribuições de Dirichlet e multinomial é normalmente usada em dados que possuem superdispersão e denominada Dirichlet-multinomial. Salvador (2019) utilizou essa distribuição para descrever problemas de superdispersão em dados politômicos.

A distribuição de Dirichlet é uma generalização da distribuição beta, proposta com o objetivo de analisar dados de proporção que sejam independentes (Connor e Mosimann, 1969). Considerando $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_J)$ e $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_J)$ vetores aleatórios, tem-se que em um primeiro estágio $\boldsymbol{Y}|\boldsymbol{\pi} \sim multi(n, \boldsymbol{\pi})$ e em um segundo estágio, tem-se que $\boldsymbol{\pi} \sim Dirichlet(\alpha)$. Logo $\boldsymbol{Y}$ segue uma distribuição Dirichlet-multinomial com função de densidade de probabilidade definida por:

$$f(\boldsymbol{y}|\boldsymbol{\alpha}) = \frac{n!}{n_1! n_2! ... n_J!} \frac{(\sum_{j=1}^{J} \alpha_j)}{(n + \sum_{j=1}^{J} \alpha_j)} \prod_{j=1}^{J} \frac{\Gamma(n + \alpha_j)}{\Gamma(\alpha_j)}, \tag{3.11}$$

com o vetor de parâmetros $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_J)$ sendo estritamente positivo.

Os valores esperados, as variâncias e as covariâncias para a distribuição Dirichlet-multinomial são dados por:

$$E[Y_j] = n\mu_j \tag{3.12}$$

$$Var[Y_j] = n\mu_j(1-\mu_j)\left[1 + \frac{(n-1)}{1+\phi}\right] \tag{3.13}$$

$$COV[Y_j, Y_k] = -n(n+\phi)(1+\phi)\mu_j\mu_k \tag{3.14}$$

em que: $\mu_j = \frac{\alpha_j}{\sum_{j=1}^{J}\alpha_j}$, $j \neq k$ e $j, k = 1, 2, ..., J$

Considerando $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_p})$ um vetor de variáveis explanatórias e $\boldsymbol{\beta_j} = (\beta_{j1}, \beta_{j2}, ..., \beta_{jp})$ como o vetor de parâmetros a serem estimados, o modelo de regressão Dirichlet-multinomial utiliza a função de ligação log-linear e é definido por:

$$\eta = ln(\beta_{j0} + \boldsymbol{x^T\beta}) \tag{3.15}$$

Os parâmetros desconhecidos de um modelo de mistura podem ser estimados utilizando métodos como: máxima verossimilhança, método dos momentos, métodos bayesianos, entre outros. Para o caso de variáveis aleatórias contínuas, a função densidade de probabilidade para os modelos de mistura é descrita por meio de probabilidades condicionais (Salvador, 2019).

3.3 Estudo de motivação I

O primeiro estudo de motivação provém de um experimento com duas espécies parasitóides de ovos do percevejo *Euschistus heros*: *Trissolcus basalis* (TB) e *Telenomus podisi* (TP). O parasitismo é amplamente estudado para implementação de programas de controle biológico e, ambas as espécies abordadas neste experimento são utilizadas como agentes de controle biológico de percevejos pragas da soja, principalmente *Euschistus heros* e *Nezara viridula*.

Esse experimento foi conduzido no Laboratório de Ecologia de Interações do Departamento de Ecologia e Biologia Evolutiva da Universidade Federal de São Carlos (UFSCar) no ano de 2020. Tal experimento teve como objetivo identificar qual é o grupo de ovos que são primeiramente visitados pelas espécies quando o forrageio por ovos se dá na presença de ovos de diferentes qualidades do hospedeiros, durante o processo de parasitismo. Para avaliar essa escolha, foram dispostos em uma placa de Petri (15×2 cm) três diferentes grupos de ovos de *E. heros*, de forma aleatória, sendo estes: (1) ovo sadio - sem parasitismo prévio, (2) ovo já parasitado pela espécie TB, (3) ovo já parasitado pela espécie TP, ilustrado pela Figura 3.1.

Figura 3.1: Croqui do experimento do parasitismo de *Trissolcus basalis* e *Telenomus podisi* sobre ovos de *E. heros* conduzido no Laboratório de Ecologia de Interações do Departamento de Ecologia e Biologia Evolutiva da Universidade Federal de São Carlos (UFSCar).

Fêmeas previamente acasaladas foram introduzidas em placas de Petri contendo três grupos de ovos de *E. heros* e o primeiro parasitismo foi registrado, em um período máximo de 35 minutos. Após esse tempo, a fêmea foi retirada da arena experimental e os ovos separados e armazenados para confirmação da ocorrência de parasitismo. O experimento contou com 12 repetições para cada espécie.

A variável resposta desse experimento é primeira categoria de escolha (ovo sadio, ovo já parasitado pela espécie TB ou ovo já parasitado pela espécie TP) e, como variável explanatória, tem-se as espécies de parasitóides.

3.4 Estudo de motivação II

O segundo estudo de motivação é uma adaptação do experimento realizado por Castro (2016) com suínos machos reprodutores, no período de março a julho de 2014 em uma granja comercial, objetivando entender se o comportamento dos animais é influenciado pelas condições ambientais, entendendo que diferentes condições ambientais podem evitar comportamentos anormais e agressivos nos animais, impedindo que sejam descartados na seleção para reprodução.

O experimento foi realizado em um delineamento inteiramente casualizado com duas repetições e contou com 4 linhagens de suínos machos: duas linhas puras: 1.010 (originária da raça Landrace) e 1.020 (originária da raça Large White) e duas cruzadas: 65 (originário das raças Pietrain, Duroc, Landrace e Large White) e 415 (originário da fêmea de linhagem 65 com macho da raça Pietrian) em 2 condições

ambientais: com enriquecimento (CE) e sem enriquecimento (SE). A condução do experimento ocorreu em 8 baias com 16 machos cada, totalizando 128 machos, ilustrado na Figura 3.2.

Figura 3.2: Instalação de crescimento dos suínos divididos em baias no experimento conduzido por Castro (2016).

Durante 27 dias, os animais foram acompanhados e seu comportamento quanto à posição na qual se encontravam dentro da baia (sentados, em pé ou deitados) foi registrado em períodos de 15 minutos, entre o período da manhã e início da tarde, totalizando 12 registros diários. Para o presente trabalho foi realizada uma adaptação do experimento original, contando com apenas 6 dias.

3.5 Métodos

Com o objetivo de observar as distribuições e comportamentos das variáveis independentes em relação à variável resposta ao longo do tempo, são apresentadas as medidas de dispersão, observando a proporção de ocorrência de cada categoria, além da observação gráfica das variáveis de interesse. Posteriormente, foram ajustados modelos com distribuições multinomial, multinomial negativa e/ou Dirichlet-multinomial.

Para o primeiro experimento definido na seção 3.3, tem-se a equação 3.16 definida como o preditor linear para o modelo irrestrito:

$$\eta = \beta_{j0} + \beta_{jk} Tratamento_k, \tag{3.16}$$

em que: β_{j0} representa uma constante geral para a j-ésima categoria e β_{jk} representa o parâmetro associado ao efeito do k-ésimo tratamento em relação a j-ésima categoria.

Após o ajuste do modelo restrito e irrestrito utilizando a distribuição multinomial, realizou-se a comparação da qualidade dos ajustes pelo teste de hipóteses denominado teste de razão de verossimilhança (TRV), para avaliar qual modelo está melhor ajustado. Para tal, as hipóteses testadas são: $\boldsymbol{H_0}$: $\eta = \beta_{j0}$ e $\boldsymbol{H_1}$: $\eta = \beta_{j0} + \beta_{jk} Tratamento_k$.

Ainda, considerando $\widehat{\theta_0}$ como o estimador de máxima verossimilhança para o modelo restrito com p parâmetros e $\hat{\theta}$ o estimador de máxima verossimilhança para o modelo irrestrito com q parâmetros. O TRV segue uma distribuição Qui-quadrado com $q - p$ graus de liberdade e tem sua estatística definida por:

$$TRV = -2ln[\ell(\widehat{\theta_0}) - \ell(\hat{\theta})] \tag{3.17}$$

em que ℓ representa o logarítmo da função de máxima verossimilhança e $p < q$.

E para o segundo experimento descrito na seção 3.4, o preditor linear para o modelo irrestrito é da forma:

$$\eta = \beta_{j0} + \beta_{jk} Tratamento_k + \beta_{jl} Linhagem_l + \beta_{jm} Hora_m, \tag{3.18}$$

em que: β_{j0} representa uma constante geral para a j-ésima categoria e β_{jk} representa o parâmetro associado ao efeito do k-ésimo tratamento em relação à j-ésima categoria; β_{jl} representa o parâmetro associado ao efeito da l-ésima linhagem em relação à j-ésima categoria; β_{jm} representa o parâmetro associado ao efeito da m-ésima hora em relação à j-ésima categoria.

Para o segundo estudo de motivação, os ajustes são avaliados por meio do AIC (Akaike, 1974) e em ambos experimentos os parâmetros são estimados pelo método da máxima verossimilhança (EMV), primeiramente apresentado por Fisher (1922). Sejam $\boldsymbol{\theta}$ os parâmetros de interesse e $X_1, ..., X_n$ n variáveis aleatórias com função de densidade dada por $f(X_1, ..., X_n; \boldsymbol{\theta})$. O EMV de $\boldsymbol{\theta}$ é dado por:

$$\hat{\boldsymbol{\theta}} =_{argmax} L(\boldsymbol{\theta}; X_1, ..., X_n), \tag{3.19}$$

em que $L(\boldsymbol{\theta}; X_1, ..., X_n) = f(X_1, ..., X_n; \boldsymbol{\theta})$ representa a função de verossimilhança, com $\boldsymbol{\theta} \in \Theta$, com Θ sendo o espaço paramétrico.

Adicionalmente, para ambos os estudos, são analisados os resíduos para cada um dos modelos ajustados, utilizando os gráfico meio normal de probabilidade com envelope simulado (Moral et al., 2017) com o objetivo de verificar os ajustes dos modelos. Após a seleção do modelo melhor ajustado, são feitas

predições para as probabilidades de ocorrência de cada categoria de resposta e estas comparadas com as proporções observadas.

Todas as análises são realizadas utilizando o software estatístico R Core Team (2019), aplicando os pacotes base, nnet (Venables e Ripley, 2002), MGLM (Zhang et al., 2017) e hnp (Moral et al., 2017), com um nível de significância de $\alpha = 0,05$.

3.6 Resultados

3.6.1 Estudo de motivação I

Primeiramente, com o objetivo de entender a distribuição dos dados pelas categorias de resposta, realizou-se uma análise exploratória cujas proporções e totais podem ser observados na Tabela 3.1.

Tabela 3.1: Medidas descritivas para os tipos de ovos parasitáveis no experimento de parasitismo de vespas conduzido na UFSCar em 2020.

Ovos	Proporção	Total
T. basalis	25%	6
T. podisi	17%	4
Sadio	58%	14

Pode-se observar que a categoria de ovo Sadio (sem um parasitismo prévio) foi, no geral, a mais escolhida pelos parasitas para depositar seus ovos, com 58% das escolhas, o que é interessante pois demonstra a capacidade dos parasitóides em discriminar ovos sadios. A categoria de ovo com menor número de escolhas foi a previamente parasitada pela espécie TP. Também pode-se observar os percentuais de ovos segundo as escolhas dos parasitóides, como apresentado na Figura 3.3.

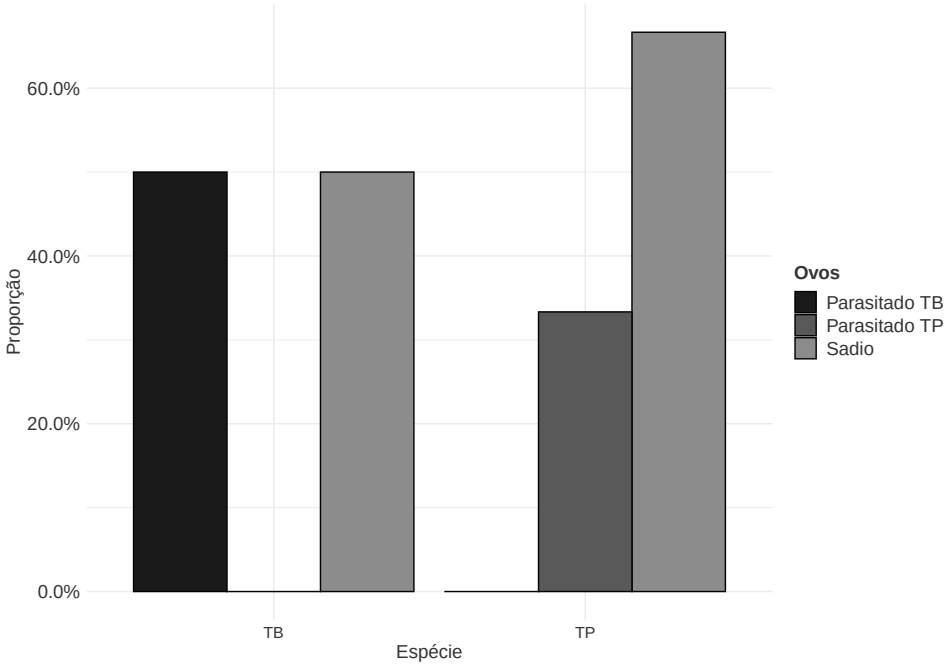


Figura 3.3: Proporção de escolha de categorias de ovos pelas vespas *Trissolcus basalis* e *Telenomus podisi* no experimento de parasitismo conduzido da UFSCar em 2020.

Com base na Figura 3.3, pode-se observar que as taxas de superparasitismo, ou seja, ovos parasitados pela mesma espécie, são menores para *Telenomus podisi*, o que demonstra uma capacidade de discriminação maior dessa espécie e, logo, uma competição intra-específica menor. Tais fatores são importantes para a continuidade da população de parasitóides no campo visando a necessidade de controle biológico e, também, a redução de custos com tais controles nas lavouras. A espécie TB, observando sua primeira escolha, tem taxas de superparasitismos de 50%, o que pode acarretar em altos custos para manutenção da população no campo e competição em estágio larval, podendo ser letal a espécie (Cingolani et al., 2013).

A seguir, foram ajustados dois modelos multinomiais para a escolha dos parasitóides que foram comparados pelo do teste da razão de verossimilhanças (TRV), Os valores estimados dos parâmetros dos modelos estão apresentados na Tabela 3.2.

Tabela 3.2: Modelos ajustados e teste da razão de verossimilhanças para o efeito de tratamento relacionado a categoria de escolha de sítio de oviposição.

Modelo	Preditor Linear	Nº parâmetros	ℓ	TRV	p-valor
1	$\eta = \beta_{j0}$	2	-23,03		
2	$\eta = \beta_{j0} + \beta_{jk} tratamento_k$	4	-15,95	14,15	0,0008

Com base no TRV, pode-se concluir que o efeito de tratamento é significativo e interfere na classificação da escolha dos parasitóides. Definindo a categoria de ovos sadios como referência, tem-se que o modelo dos logitos generalizados referente à categoria de ovos previamente parasitados pela espécie TP utiliza os parâmetros β_{10} e β_{11} e que o modelo referente à categoria de ovos previamente parasitados pela espécie TB utiliza os parâmetros β_{20} e β_{21}, com as estimativas dos parâmetros apresentadas na Tabela 3.3.

Tabela 3.3: Estimativas dos parâmetros e erros padrões para o modelo irrestrito ajustado referente a categoria de escolha de sítio de oviposição.

Parâmetro	Estimativa	Erro Padrão
β_{10} (Ovo TP)	-9,67	51,63
β_{11} (Ovo TP)	8,98	51,62
β_{20} (Ovo TB)	0,0001	0,57
β_{21} (Ovo TB)	-10,45	65,80

Para avaliar a qualidade do ajuste por meio dos resíduos do modelo, construiu-se um gráfico meio normal de probabilidade com envelope simulado, apresentado na Figura 3.4.

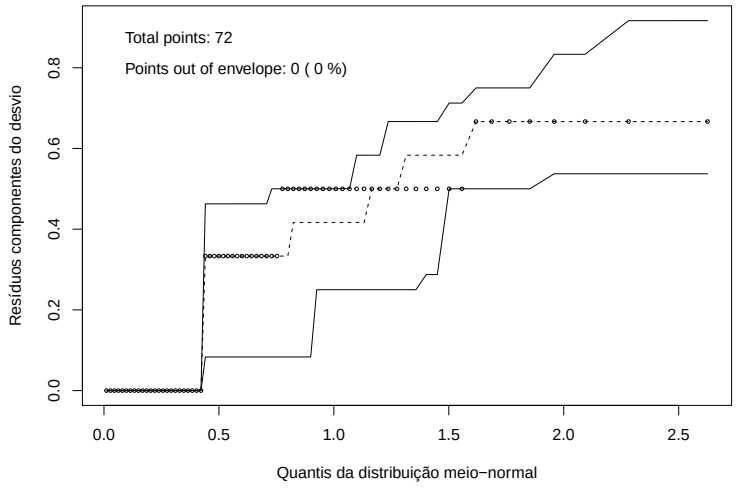


Figura 3.4: Half-normal plotpara os modelos ajustados aos dados de parasitismo de vespas no experimento conduzido na UFSCar.

Todas as observações no gráfico de diagnóstico encontram-se dentro do envelope de simulação, o que indica que este modelo teve um ajuste razoável aos dados, apesar de seu aspecto atípico. Agora,

com o modelo ajustado, pode-se verificar as probabilidades preditas pelo modelo, apresentadas na Tabela 3.4.

Tabela 3.4: Proporções observada e predita utilizando o modelo dos logitos generalizados ajustado para as categorias de ovos escolhidos por *Trissolcus basalis* e *Telenomus podisi*.

		Observado	Predito
TP	Ovos TP	33%	33%
	Ovos TB	0%	0%
	Ovos sadios	67%	67%
TB	Ovos TP	0%	0%
	Ovos TB	50%	50%
	Ovos sadios	50%	50%

Como observado, as probabilidades observadas e preditas pelo modelo são coincidentes, ou seja, iguais, isso pode ser devido ao fato do experimento envolver apenas dois tratamentos, três categorias de escolha e uma baixa variabilidade nas escolhas. Para ambos os tratamentos, os parasitóides buscam ovos somente parasitados pela mesma espécie ou ovos sadios para realizar o parasitismo. Em ambos os casos, as probabilidades dos parasitas escolherem ovos parasitados pela outra espécie é zero.

3.6.2 Estudo de motivação II

Objetivando compreender a distribuição dos dados, fez-se uma análise descritiva inicial para ambos os tratamentos, conforme Figura 3.5.

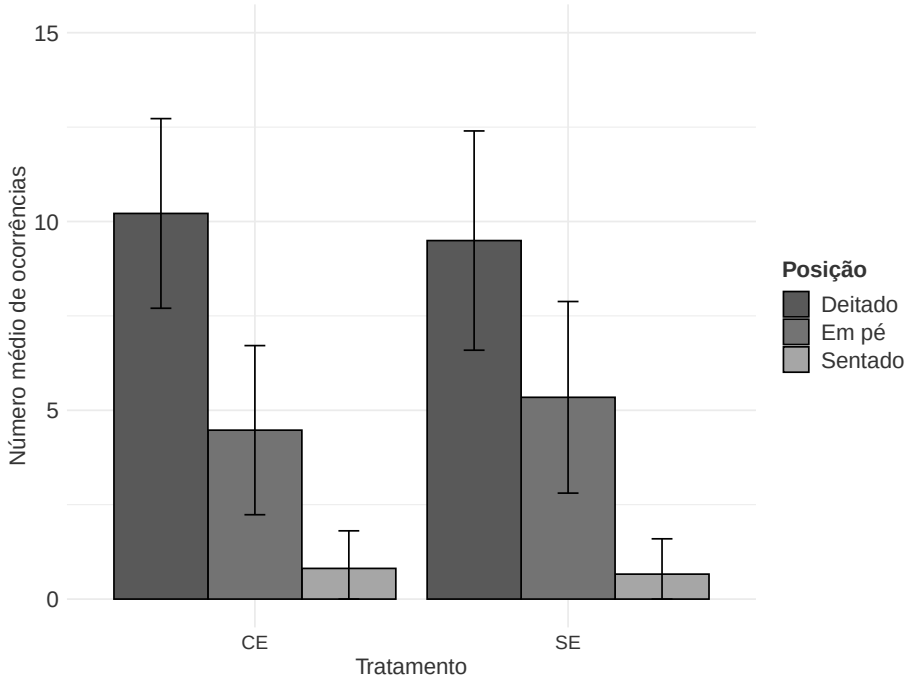


Figura 3.5: Distribuição do posicionamento dos animais no experimento desenvolvido por Castro (2016) em relação aos tratamentos.

Pode-se observar que a maioria dos animais tem preferência por ficar deitado, seguido dos animais em pé e depois sentados, independente da condição ambiental. A Tabela 3.5 apresenta os percentuais segundo as posições e linhagens dos suínos.

Tabela 3.5: Proporção de animais em cada posição dentro dos tratamento e linhagens analisados.

	Ambientes					
Linhagem	**Sem enriquecimento**			**Com enriquecimento**		
	Deitado	**Sentado**	**Em pé**	**Deitado**	**Sentado**	**Em pé**
65	61,2%	2,3%	36,5%	67,1%	3,6%	29,3%
415	67,5%	7,5%	25,0%	66,6%	6,4%	27,0%
1010	60,7%	4,3%	35,0%	68,6%	5,2%	26,2%
1020	54,9%	2,7%	42,4%	61,3%	5,6%	33,1%

Observa-se em termos exploratórios, que há indícios de não diferenças entre as linhagens em ambos os tratamentos, nos quais a maioria dos animais preferem a posição deitada e poucos ficam na posição sentada, destacando-se a linhagem pura 1020 com o menor proporção de animais deitados e maior proporção de animais em pé, em ambos os tratamentos.

Posteriormente, foram ajustados cinco modelos dos logitos generalizados com distribuição multinomial para a variável resposta, visando selecionar o melhor modelo. Os resultados para a sequência de modelos, considerando AIC e TRV são apresentados na Tabela 3.6.

Tabela 3.6: Teste de razão de verossimilhanças e AIC para modelos dos logitos generalizados com distribuição multinomial.

Modelo	Preditor Linear	Nº parâmetros	AIC	TRV	p-valor
1	$\eta = \beta_{j0}$	2	9490,693		
2	$\eta = \beta_{j0} + \beta_{jk} trat_k$	4	9471,876	68,635	<0,0001
3	$\eta = \beta_{j0} + \beta_{jl} linhag_l$	8	9434,057	45,810	<0,0001
4	$\eta = \beta_{j0} + \beta_{jk} trat_k + \beta_{jl} linhag_l$	10	9414,065	69,802	<0,0001
5	$\eta = \beta_{j0} + \beta_{jk} trat_k + \beta_{jl} linhag_l + \beta_{jm} hora_m$	12	9409,962	8,103	0,0174

Por ambos os critérios (AIC e TRV), o modelo com melhor ajuste é o modelo 5, com 12 parâmetros. Para verificar a qualidade do ajuste do modelo selecionado, construiu-se o gráfico meio normal de probabilidade com envelope simulado, apresentado na Figura 3.6.

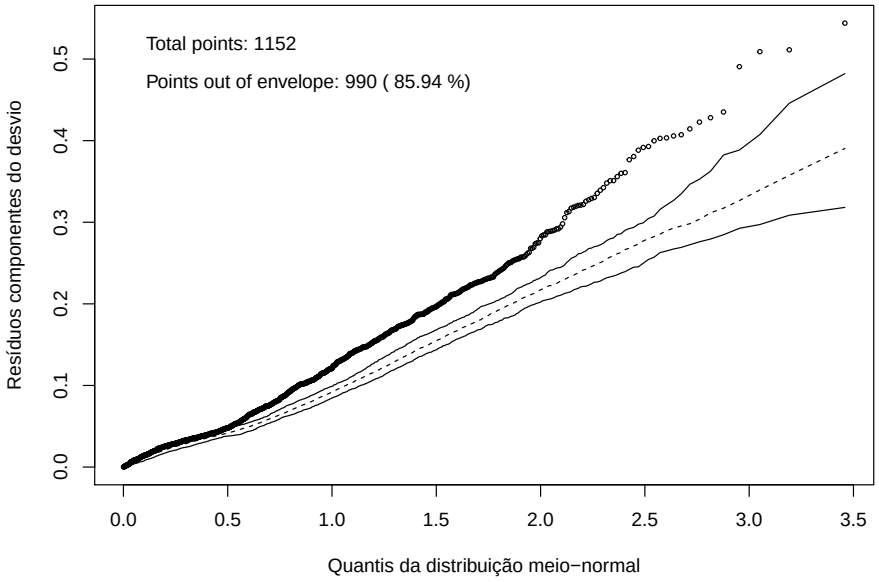


Figura 3.6: Gráfico meio normal de probabilidade com envelope simulado para verificar a qualidade de ajuste do modelo multinomial no estudo sobre enriquecimento ambiental no comportamento de suínos.

Com 85% dos pontos fora do envelope de simulação, é possível afirmar que o modelo selecionado não possui um bom ajuste, apesar de ser selecionado como o melhor modelo em dois critérios. Uma abordagem mais adequada para ajustar um modelo para este conjunto de dados é trabalhar com modelos hierárquicos.

Sendo assim, foram ajustados outros dois modelos com distribuições cujas estruturas envolvem misturas, sendo estas a distribuição multinomial Negativa e Dirichlet-multinomial. O modelo com melhor ajuste é o Dirichlet-multinomial, com $AIC = 2.623,130$, aproximadamente metade do valor do critério para o modelo de regressão multinomial negativo $(4.538,238)$. Ainda, para afirmar que o modelo possui um bom ajuste, é necessário analisar os resíduos para verificar a qualidade do ajuste do modelo com o gráfico meio normal de probabilidade, apresentado na Figura 3.7.

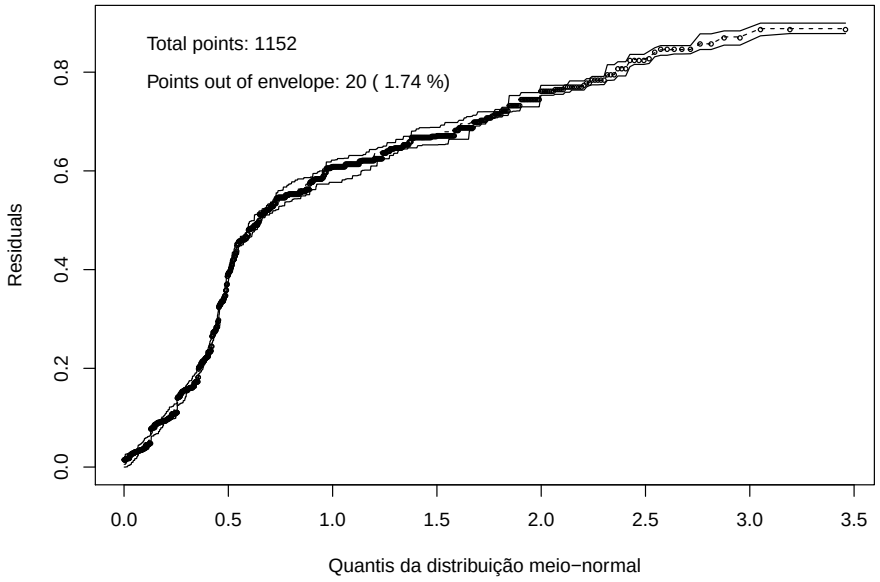


Figura 3.7: *Half-normal plot* para verificar a qualidade de ajuste do modelo Dirichlet-multinomial no estudo sobre comportamento de suínos.

Verifica-se que, com o uso de uma distibuição mais adequada, tem-se quase a totalidade dos resíduos contidos dentro do envelope de simulação, o que indica que o modelo Dirichlet-multinomial possui um bom ajuste para este problema. Pode-se realizar a predição das probabilidades de ocorrência de cada cenário utilizando o modelo ajustado. Tais predições podem ser observadas na Figura 3.8.

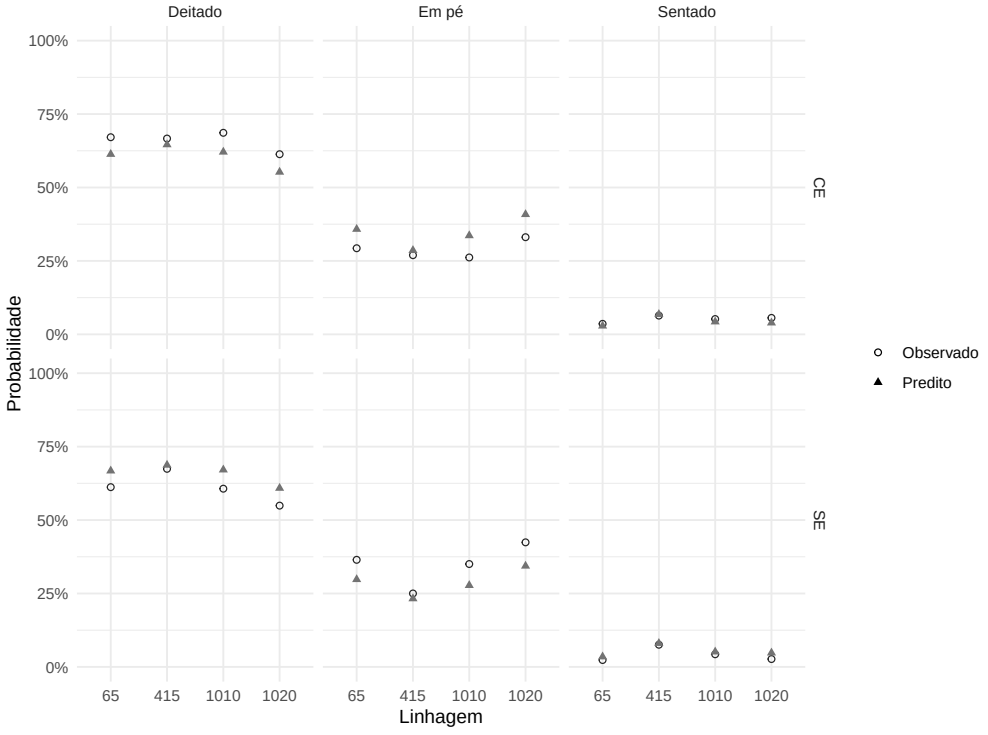


Figura 3.8: Probabilidades observadas e preditas para as linhagens e tratamentos utilizando o modelo Dirichlet-multinomial no estudo de enriquecimento ambiental relacionado ao comportamento de suínos.

Verifica-se que as probabilidades preditas para cada cenário são próximas às probabilidades observadas, o que pode ser considerado mais uma confirmação de que o modelo proposto está bem ajustado.

3.7 Discussão

Em ambos os experimentos foi possível utilizar distribuições apropriadas para dados categorizados. Para o primeiro experimento envolvendo ovos parasitáveis o modelo dos logitos generalizados assumindo distribuição multinomial mostrou-se satisfatório. Pelo EMV foi rejeitada a hipótese nula, havendo indícios de que o tratamento é significativo e, nesse caso, representa a preferência de cada espécie parasitóide na escolha dos novos sítios de oviposição.

Nem sempre em experimentos com variáveis politômicas a distribuição multinomial para a variável resposta é adequada e fornece um bom ajuste, como no caso do segundo experimento de comportamento de suínos. Neste, a distribuição não se mostrou satisfatória, exigindo uma nova distribuição para a variável resposta. Aqui, a utilização de uma mistura hierárquica para a variável resposta apresentou um melhor ajuste, com o menor valor de AIC comparado aos outros dois modelos. O modelo de regressão Dirichlet-multinomial foi o mais adequado para descrever as três categorias propostas, contendo quase a totalidade dos resíduos dentro do envelope de simulação e apresentando as probabilidades preditas em valores próximos às observadas.

Referências

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience, New York, second edition.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19:716–723.

Bates, G. e Neyman, J. (1952). Contributions to the Theory of Accident Proneness I. An Optimistic Model of the Correlation Between Light and Severe Accidents. *University of California publications in statistics*, 30:215–253.

Castro, A. C. D. (2016). *Comportamento e desempenho sexual de suínos reprodutores criados em ambientes enriquecidos*. PhD thesis, Universidade de São Paulo / Escola Superior de Agricultura "Luiz de Queiroz".

Cingolani, M., Greco, N., e Liljesthröm, G. (2013). Multiparasitism of piezodorus guildinii eggs by telenomus podisi and trissolcus urichi. *BioControl*, 58:37–44.

Connor, R. J. e Mosimann, J. E. (1969). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Public Health*, 64:194–206.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, (A):309–368.

Forbes, C., Evans, M., Hastings, N., e Peacock, B. (2010). *Statistical Distributions*. Wiley.

Giolo, S. (2017). *Introdução à Análise de Dados Categóricos com Aplicações*. Blucher, Brasil, 1 edition.

Keefe, R., Arnold, M., Bayen, U., e Harvey, P. (1999). Source monitoring deficits in patients with schizophrenia; a multinomial modelling analysis. *Psychol Med*, 29:903–914.

Le Gall, F. (2006). The modes of a negative multinomial distribution. *Statistics and Probability Letters*, 76:619–624.

Moral, R. A., Hinde, J., e Demétrio, C. G. B. (2017). Half-Normal Plots and Overdispersed Models in R: The hnp Package. *Journal of Statistical Software*, 81:23p.

Mosimann, J. E. (1962). On the Compound Multinomial Distribution , the Multivariate β- Distribution , and Correlations Among Proportions. *Biometrika Trust, Oxford University Press*, 49:65–82.

Pearson, K. e Erdmann, H. O. M. F. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, 186:343–414.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riefer, D. M. e Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95:318–339.

Salvador, M. (2019). O problema da superdispersão em dados categorizados politômicos nominais em estudos agrários. Master's thesis, Tese de Mestrado em Estatística e Experimentação Agronômica. Escola Superior de Agricultura "Luiz de Queiroz".

Venables, W. N. e Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Zhang, Y., Zhou, H., Zhou, J., e Sun, W. (2017). Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13.

4 CONSIDERAÇÕES FINAIS

Este trabalho apresenta uma introdução à análise de dados discretos, abordando modelos com o uso de diferentes distribuições para uma melhor descrição da relação funcional. Esses tipos de dados são muito comuns na ciencias agrárias e essa classe de modelos traz inúmeras opções para análises estatísticas sem necessidade de transformações, possibilitando compreender comportamentos que permitem trabalhar com controle de pragas em mudanças genéticas, controles biológicos e compreender técnicas de reprodução animal.

A análise de tais dados ainda é alvo de muitos estudos e constantemente são propostas novas distribuições para tal, como a distribuição de Poisson ponderada para dados de contagem subdispersos apresentada por Louzayadio et al. (2021).

Entende-se que as distribuições tradicionais (como a Poisson para dados de contagem ou a multinomial para dados categorizados) são as mais utilizadas, porém existem casos em que as pressuposições não são satisfeitas ou os ajustes de modelos com tais distribuições não apresentam resultados satisfatórios, necessitando distribuições alternativas. Essas, como as distribuições COM-Poisson, Poisson-Tweedie e Dirichlet-multinomial, apresentam flexibilidades em relação às tradicionais, permitindo ajustar modelos com padrões de não equidispersão (subdispersão, superdispersão), heterocedasticidade e entre outros, tais como apresentados nos estudos dos capítulos 2 e 3.

Reitera-se que os estudos foram uma introdução aos modelos para dados de contagem e categorizados. Apesar dos ajustes satisfatórios para os quatro estudos apresentados utilizando as distribuições binomial negativa, quase-Poisson, multinomial e Dirichlet-multinomial, existe uma variedade de outras distribuições que podem ser aplicadas a outros conjuntos de dados semelhantes, dentre as quais se destacam: a distribuição Dirichlet-multinomial generalizada (Bouguila, 2008), multinomial-Poisson (Terza e Wilson, 1990), Weibull (Weibull, 1951) e as distribuições bayesianas.

Como uma extensão das análises apresentadas, pretende-se realizar estudos mais detalhados abordando distribuições distintas, além de entender outras misturas que podem ser utilizadas para a obtenção de um melhor ajuste para os estudos propostos. Também pretende-se aprofundar o estudo para dados politômicos sub e superdispersos utilizando misturas próprias para cada um dos casos.

Referências

Bouguila, N. (2008). Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474.

Louzayadio, C. G., Malouata, R. O., e Koukouatikissa, M. D. (2021). A weighted poisson distribution for underdispersed count data. *International Journal of Statistics and Probability*, 10(4):157–165.

Terza, J. e Wilson, P. (1990). Analyzing frequencies of several types of events: A mixed multinomial-poisson approach. *The Review of Economics and Statistics*, 72(1):108–115.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of Applied Mechanics*, 18:293–296.

ANEXOS

Anexo A

Linhas de comando R para o primeiro capítulo:

```
# Instalando os pacotes que serão utilizados:

if (!require('MASS')) install.packages('MASS'); library('MASS')
if (!require('hnp')) install.packages('hnp'); library('hnp')
if (!require('lme4')) install.packages('lme4'); library('lme4')
if (!require('COMPoissonReg')) install.packages('COMPoissonReg'); library('COMPoissonReg')
if (!require('tweedie')) install.packages('tweedie'); library('tweedie')

# Ajustando os modelos e verificando tais ajustes:

# 1. Ajuste do modelo com a distribuição Poisson:

m1 <- glm(Y ~ X,
          data = df,
          family = 'poisson')
summary(m1)

# 2. Ajuste do modelo com a distribuição Binomial Negativa:

m2 <- glm.nb(Y ~ X,
             data = df)
summary(m2)

# 3. Ajuste do modelo com a distribuição quase Poisson:

m3 <- glm(Y ~ X,
          data = df,
          family = 'quasipoisson')
summary(m3)

# 4. Ajuste do modelo com a distribuição COM Poisson:

m4 <- glm.cmp(Y ~ X,
          data = df
summary(m4)

# 5. Ajuste do modelo com a distribuição Poisson-Tweedie:

m5 <- glm(Y ~ X,
          data = df,
          family = tweedie(link.power = 0, var.power = 1.1))
```

```
summary(m5)

# Criando um data-frame para verificar as métricas de AIC e Likelihood:

Modelo = c('m1',
           'm2',
           'm3',
           'm4',
           'm5')

AIC = c(AIC(m1),
        AIC(m2),
        AIC(m3),
        AIC(m4),
        AICtweedie(m5))

VEROSSIMILHANCA = c(logLik(m1),
                    logLik(m2),
                    logLik(m3),
                    logLik(m4),
                    logLiktweedie(m5))

data.frame(Modelo, AIC, VEROSSIMILHANCA)

# Verificando a qualidade dos ajustes (Half-normal plots):

hnp(modelo_escolhido,
    print.on = T,
    xlab ='Percentil da N(0,1)',
    ylab =' Resduos',
    main ='Gráfico Normal de Probabilidades')

# Predição para os modelos escolhidos:

casos <-
(predict = predict(modelo_escolhido,
                   interval =' prediction',
                   newdata = casos,
                   type =' response'))
```

Anexo B

Linhas de comando R para o segundo capítulo:

```
# Instalando os pacotes que serão utilizados:

if (!require('nnet')) install.packages('nnet'); library('nnet')
if (!require('MGLM')) install.packages('MGLM'); library('MGLM')
if (!require('hnp')) install.packages('hnp'); library('hnp')

# Ajustando os modelos e verificando tais ajustes:

# 1. Ajuste do modelo com a distribuição multinomial:

m1 <- multinom(cbind(classe1, classe2, classe3) ~ X1,
               data = df)
summary(m1)

m2 <- multinom(cbind(classe1, classe2, classe3) ~ X2,
               data = df)
summary(m2)

# Teste de razão de verossimilhanças (TRV):
anova(m1,m2)

# 2. Ajuste do modelo com a distribuição multinomial negativa:

m3 <- MGLMreg(cbind(classe1, classe2, classe3) ~ X,
              dist = "NegMN",
              data = df)
summary(m3)

# 3. Ajuste do modelo com a distribuição Dirichlet-multinomial:

m4 <- MGLMreg(cbind(classe1, classe2, classe3) ~ X,
              dist = "DM",
              data = df)
summary(m4)

# Verificando o critério de AIC:
Modelo = c('m1',
           'm2',
           'm3',
           'm4)

AIC = c(AIC(m1),
        AIC(m2),
```

$AIC(m3),$
$AIC(m4))$

$data.frame(Modelo, AIC)$

Predições para o modelo Dirichlet-multinomial:
$casos <- matrix()$
$design\_mat <- t(m4@data\$X)$
$predict <- predict(m4, t(design\_mat))$