**University of São Paulo
"Luiz de Queiroz" College of Agriculture**

**Statistical modelling and computational tools applied to the natural
sciences**

Maíra Blumer Fatoretto

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

**Piracicaba
2020**

Maíra Blumer Fatoretto
Bachelor in Statistics

Statistical modelling and computational tools applied to the natural sciences

Advisor:

Prof^ª Dr^ª: **CLARICE GARCIA BORGES DEMÉ-
TRIO**

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2020

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

Fatoretto, Máira Blumer

Statistical modelling and computational tools applied to the natural sciences / Máira Blumer Fatoretto. -- Piracicaba, 2020.
70 p.

Tese (Doutorado) -- USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Modelos mistos 2. Reamostragem bootstrap 3. Modelo Bivariado
4. Dados de contagens 5. Modelagem conjunta 6. Software R I. Título.

ACKNOWLEDGMENTS

I would like to thank my parents and my sister for supporting me. My husband Pedro for always being by my side, supporting me and encouraging me even in the hard times, and mainly understanding all the dedication and hours of work that I was absent. My grandmother Silvia who passed away this year, but always has been by my side praying for my success.

I would like to thank my great friend and unofficial co-supervisor Prof. Dr. Rafael de Andrade Moral from Maynooth University, for all his teaching, patience, and support throughout my doctorate. Without it, this work would not be possible. He showed me every day how to be an excellent researcher, advisor and to love statistics every day. I can only thank him for all his help, for showing me the best paths in this scientific area, and for always being present even that he could not be officially registered as a co-supervisor. To my adviser, Prof. Dr. Clarice Garcia Borges Demetrio who agreed to guide me and helped me throughout this period, for her patience and immense knowledge. To Prof. Dr. Caroline Brophy from Trinity College Dublin, from receiving me on my Ireland period, together with Prof. Dr. Rafael for, even without knowing me, treated me very well, with great patience, affection, and attention, contributing a lot to all this work.

Finally, I would like to thank the entire Graduate Program in Statistics and Agricultural Experimentation (PPGEEA) of the Department of Exact Sciences at ESALQ USP. All the teachers and professionals who work in the department and who helped me during this period. To my friends from ESALQ/USP and Maynooth University, Eduardo, Gustavo, Patricia, who helped me a lot and contributed to all my professional growth and in special to Bruna that received me in her house during my internship period. In addition to all the excellent professors and researchers that I met at national and international congresses and conferences.

This work was supported by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), Brazil.

"Without data you are just another person with an opinion."

W. Edwards Deming

"The beautiful thing about learning is nobody can take it away from you."

B. B King

SUMÁRIO

RESUMO	6
ABSTRACT	7
1 INTRODUCTION	9
References	10
2 OVERDISPERSED FUNGUS GERMINATION DATA: STATISTICAL ANALYSIS USING R	13
2.1 Introduction	13
2.2 Case Study	13
2.3 Statistical Methods	15
2.3.1 Introduction to Generalized Linear Models (GLMs)	15
2.3.1.1 Binomial model	16
2.3.1.2 Quasi-binomial model	17
2.3.1.3 Beta-binomial model	17
2.3.1.4 Logistic-normal model	17
2.4 Analysis the of case-study	18
2.4.1 Grouping	26
2.5 Discussion	28
References	29
3 A SIMULATION STUDY USING BOOTSTRAP PROCEDURES APPLIED TO GENERALIZED LINEAR MIXED MODELS	31
3.1 Introduction	31
3.2 Motivation	32
3.3 Methodology	33
3.3.1 Background of Generalized Linear Mixed Models	33
3.3.2 Bootstrap methods	34
3.4 Simulation studies	36
3.5 Results	38
3.5.1 Simulation results	38
3.5.2 Practical results	42
3.6 Discussion	44
References	46
4 A BIVARIATE FRAMEWORK TO JOINTLY MODEL COUNT AND CONTINUOS RESPONSES	49
4.1 Introduction	49
4.2 Case study	50
4.3 Joint Modelling of a Count and a Continuous Response	52
4.4 Estimation and Inference	52
4.4.1 Univariate modelling	52
4.4.2 Bivariate modelling	53
4.5 Case study analysis	55
4.6 Simulation studies	55
4.6.1 Simulation results	58
4.7 Discussion	61
References	61
5 FINAL CONSIDERATIONS	67
APPENDIX	69

RESUMO

Modelagem estatística e ferramentas computacionais aplicadas a ciências naturais

Dados da área da ciências naturais podem envolver diferentes características tornando-se interessantes de serem analisados e possibilitando muitas contribuições do ponto de vista estatístico. Uma de suas grandes áreas é a entomologia, ciência que estuda os insetos. Nessa área, podemos encontrar dados de proporções, dados de contagens, dados contínuos, respostas univariadas ou multivariadas, além de medidas repetidas. Nesse sentido, o objetivo deste trabalho é desenvolver modelos e métodos computacionais que auxiliem na interpretação desse tipo de dados, trazendo ganhos para a área estatística e tornando-a mais acessível a pesquisadores de todas as áreas. Nesse trabalho, nós utilizamos modelos lineares generalizados mistos para analisar dados de proporção superdispersos, em experimento de dose-resposta, com o objetivo de comparar a resistência de fungos a altas temperaturas. Feita essa comparação e verificada a qualidade do ajuste, utilizando gráficos meio-normal com envelope de simulação, propusemos um tutorial de como comparar diferentes modelos de superdispersão para dados de proporção, utilizando o software R. Outro objetivo foi comparar a similaridade dos fungos (diferentes tratamentos). Para isso, foram propostos métodos de reamostragem bootstrap não paramétricos e comparados por meio de simulação, verificando a performance de diferentes tipos de intervalos de confiança. Além disso, foi desenvolvida e proposta uma extensão dos modelos normais bivariados, com o objetivo de acomodar dados de contagem. Essa extensão foi desenvolvida pela adição de um novo parâmetro na matriz de variâncias e covariâncias para melhor acomodar dados de contagem, subdispersos ou superdispersos, modelando assim a média e a dispersão dos dados observados.

Palavras-chave: Análise bivariada, Bootstrap não paramétrico, Correlações, Modelos lineares generalizados mistos, Superdispersão.

ABSTRACT

Statistical modelling and computational tools applied to the natural sciences

Natural sciences data can involve different characteristics, making it very interesting to be analyzed and may bringing many contributions from a statistical point of view. One of its major areas is entomology, a science that studies insects and where we can find proportions, count or continuous data, univariate or multivariate responses, in addition to repeated measures. In this sense, to develop models and computational methods that support these observations and to bring gains to the statistical area, at the same time as making this analyzes more accessible is the aim of this work. Generalized linear mixed models were used to analyze proportion and overdispersed data in dose-response studies and, the goodness-of-fit was assessed using half-normal plots with simulation envelopes. To compare treatments of these data different non-parametric bootstrap resampling methods were proposed and, a simulation study was carried out to verify the performance of different types of confidence intervals. Besides, an extension of the bivariate normal model was developed to accommodate count data. This extension was developed by adding a new parameter in the variance-covariance matrix to better accommodate over- or underdispersion for count data modeling the mean and dispersion of the observed data.

Keywords: Bivariate analysis, Correlations, Generalized linear mixed models, Non-Parametric Bootstrap, Overdispersion.

1 INTRODUCTION

Natural sciences data are present in several fields of study, being one of these areas the Entomology, that englobe the study of biology, taxonomy, or insect behavior. Insects affect many aspects of life hence all kinds of natural and modified ecosystems support communities of insects. Entomology also includes how the insects interact with the environment, the effects of the insect in crops, and how insect activities might be controlled (GULLAN and CRANSTON, 2014).

Studies carried out in entomology may present different characteristics, including discrete or continuous data, experiments with one or more variable responses, longitudinal data, repeated measurements, even so, hierarchical structures, making this area very interesting to present and propose different statistical models aiming to capture all characteristics present in the data.

All data analyses should consider the experimental design, possible correlations between observation units, and the characteristics of the response variables (continuous or discrete). When analyzing discrete data, the theory of generalized linear models, developed by Nelder and Wedderburn (1972), provides a range of options, using distributions that belong to the exponential family. The exponential family provides an elegant and comprehensive mathematical structure as it encompasses normal, Bernoulli/binomial, Poisson models, among others (MOLENBERGHS *et al.*, 2010). Common characteristics in entomological experiments are count or proportion data in the called dose-response experiments. To analyze discrete data, the Poisson and the binomial model could be considered, respectively, but to use these models it is necessary to be careful with overdispersion, a phenomenon where we observe greater data variability than expected by the standard models (HINDE and DEMÉTRIO, 1998). Fitting a model that does not take account for overdispersion yields underestimated standard errors of the parameter estimates, leading to incorrect interpretation of the significance of the treatment effects (MORAL *et al.*, 2017). These problems could be solve by modeling besides the mean, the dispersion of the data, using as an example the double generalized linear models (SMYTH, 1989; MCCULLAGH and NELDER, 1989; SMYTH and VERBYLA, 1999) or including a random effect to model the extra-variation present in the data, using Generalized Linear Mixed Models. Proposed by Breslow and Clayton (1993), the generalized linear mixed model (GLMM) combines GLM and random effects that allow for the accommodation of existing dependence structures between observations (as in longitudinal data) and overdispersion. Although flexible, for the construction of GLMMs, several aspects need to be taken into account, such as the correct specification of the model distribution, the definition of the model components, and how to include random effects and covariance structures.

The theory for discrete data models considering mainly GLM and GLMM has been studied and developed for many years. Nevertheless, these are still emerging for the multivariate case, mostly in the case where different distributions for the response variables are observed, characteristic often found in many types of natural sciences experiments. Then, there is a growing demand for new theories, data analysis models, and computational tools that could help researchers to interpret and get all information on these complex data structures. Starting from this lack, the objectives of this work are to study and develop models to analyze data from experiments and to solve problems related to natural sciences, involving discrete data (proportion and count) with complex characteristics, such as overdispersion, correlated structures, and multivariate responses.

In chapters 2 and 3 we present different statistical studies involving *Isaria Fumosorosea* that are commonly found in soil and infects several species of arthropods. This fungus has different types of isolates that may be collected from distinct places or insects. Fungi-based biopesticides have the capacity to infect a large number of pests and to remain in the environment. The fungus spreads conidia or spores, that germinate on the insect. The germinated spores will penetrate the outer skin of the insect and start to grow inside of it. As the fungus continues to grow inside the insect, the insect will

die. However, ultraviolet radiation can inhibit the conidia germination, influencing fungicidal efficacy. Rojas (2015) conducted the experiment at the Department of Entomology and Acarology of ESALQ to study the UV-B radiation tolerance of 14 different isolates of *Isaria Foomosorosea* to select isolates with high UV-B radiation tolerance for developing a new mycopesticide. The experiment was conducted as a randomized complete block design with four blocks occurring over four days (one on each day). Within each block, one conidia suspension of 14 isolates was prepared and was used for inoculation in five Petri dishes corresponding to each exposure time (0, 2, 4, 6, and 8h). The 70 dishes in each block were placed in a wooden chamber with four fluorescent lamps and remained there until their corresponding exposure time. The response was the proportion of germinated conidia, and the higher this proportion was, the higher the isolate tolerance is.

Proportion data from dose-response experiments are often overdispersed and several models could incorporate this extra-variation. In Chapter 2, we present a review of GLMM and discuss how to fit and select the best model to describe these overdispersed fungi data, detailing step-by-step using the R software (R CORE TEAM, 2018) (FATORETTO *et al.*, 2018). This work stands as a tutorial for researchers of applied areas, supporting how to analyze overdispersed proportional data using generalized linear mixed models, always focusing on how to identify and to take account of specific issues of the experimental design.

After selected the best model, in Chapter 3, the similarity between different types of isolates (treatments) will be verified using bootstrap confidence intervals approaches aiming to present to the researcher, options of more resistant isolates. Through simulations, we compare different types of bootstrap resampling methods, checking which ones perform better for data with this complex structure. From that, we also compare different types of bootstrap intervals, checking which one presents the best coverage rate. Here, we present a new way of comparing treatments for fitted models where it is difficult to obtain good performance with analytical confidence intervals, due to the complexity of the data structure and the modeling.

In Chapter 4, we work with a different dataset, also from entomology. Here, we proposed a bivariate normal model using as motivation a dataset from an experiment carried out at the Laboratory of Insect Biology, Department of Entomology and Acarology, ESALQ - USP. The experiment aimed to examine the development and fertility of *Podisus nigrispinus* (Hemiptera: Pentatomidae), an important predator, found in agricultural and forest systems, in several countries of Central and South America, under two different diets using the prey (*Anticarsia gemmatalis* and the prey *Diatraea saccharalis*). The female insects were submitted to these two different diets for a while, and the development was measured by the weight after 18 days. Fertility was measured by the number of eggs produced during the study. Considering a count and a continuous variable, we fitted a bivariate model using normal approximation, modeling the mean and dispersion by adding a new parameter in the variance-covariance matrix. The new parameter allowed us to make this normal approximation closer to the nature of the count data, and also allowed us to accommodate overdispersion, underdispersion and, heterogeneity of variances. In Chapter 5, we make our final considerations.

References

- AKAIKE, H., 1974a A new look at the statistical model identification. IEEE transactions on automatic control **19** (6): 716–723.
- BRESLOW, N. E. and D. G. CLAYTON, 1993a Approximate inference in generalized linear mixed models. Journal of the American Statistical Association **88** (421): 9–25.

- FATORETTO, M. B., R. D. A. MORAL, C. G. B. DEMÉTRIO, C. S. DE PÁDUA, V. MENARIN, V. M. A. ROJAS, C. P. D’ALESSANDRO, and I. DELALIBERA JR, 2018 Overdispersed fungus germination data: statistical analysis using r. *Biocontrol Science and Technology* **28**: 1034–1053.
- GULLAN, P. J. and P. S. CRANSTON, 2014 *The insects: an outline of entomology*. John Wiley & Sons.
- HINDE, J. and C. G. DEMÉTRIO, 1998 Overdispersion: models and estimation. *Computational Statistics & Data Analysis* **27** (2): 151–170.
- LAIRD, J. H., NAN M & WARE, 1982 Random-effects models for longitudinal data. *Biometrics* **38**: 963–974.
- LEE, J. A., YOUNGJO & NELDER, 1996 Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**: 619–678.
- MCCULLAGH, P. and J. A. NELDER, 1989 Generalized linear models, volume 37 of. *Monographs on statistics and applied probability* .
- MOLENBERGHS, G. and G. VERBEKE, 2006 *Models for discrete longitudinal data*. Springer Science & Business Media.
- MOLENBERGHS, G., G. VERBEKE, C. G. DEMÉTRIO, A. M. VIEIRA, *et al.*, 2010 A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* **25** (3): 325–347.
- MORAL, R. A., J. HINDE, and C. G. DEMÉTRIO, 2017 Half-normal plots and overdispersed models in r: The hnp package. *Journal of Statistical Software* **81** (10): 1–23.
- NELDER, J. A. and W. WEDDERBURN, 1972 Generalized linear models. *Journal of the Royal Statistical Society Series A* **135**: 370–384.
- PINHEIRO, D., JOSE & BATES, 2006 *Mixed-effects models in S and S-PLUS*. Springer, New York, 528 p.
- R CORE TEAM, 2018 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- SMYTH, G. K., 1989 Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society: Series B (Methodological)* **51**: 47–60.
- SMYTH, G. K. and A. P. VERBYLA, 1999 Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics: The official journal of the International Environmetrics Society* **10**: 695–709.

2 OVERDISPERSED FUNGUS GERMINATION DATA: STATISTICAL ANALYSIS USING R

Abstract

Proportion data from dose-response experiments are often overdispersed, characterized by a larger variance than assumed by the standard binomial model. Here, we present different models proposed in the literature that incorporate overdispersion. We also discuss how to select the best model to describe the data and present, using R software, specific code used to fit and interpret binomial, quasi-binomial, beta-binomial, and binomial-normal models, as well as to assess goodness-of-fit. We illustrate applications of these generalized linear models and generalized linear mixed models with a case study from a biological control experiment, where different isolates of *Isaria fumosorosea* (*Hypocreales: Cordycipitaceae*) were used to assess which ones presented higher resistance to UV-B radiation. We show how to test for differences between isolates and also how to statistically group isolates presenting a similar behavior.

Keywords: entomopathogenic fungi; generalized linear models; mixed models; proportion data; random effects.

2.1 Introduction

The class of generalized linear models (GLM) was introduced by Nelder and Wedderburn (1972) as a general framework for handling a range of statistical models for normal and non-normal data. The GLM framework allows us to analyze proportion data.

An extension of these models are the generalized linear mixed models (GLMM), which include random effects in the linear predictor, and can be used to adequately accommodate extra variability. When the variation is greater than expected by the standard Poisson and binomial models, natural starting points for the analysis of count and proportion data, respectively, this phenomenon is referred to as overdispersion (HINDE and DEMÉTRIO, 1998). In the presence of overdispersion, an alternative is to include random effects in the linear predictor of the model to explain the extra-variability.

Germination data are usually overdispersed (HINDE and DEMÉTRIO, 1998). This may be due to individual variability of the experimental units, or due to some correlation between these units. For example, experimental units consisting of a Petri dish or a test tube that are set up using the same solutions have a higher correlation than units with different solutions. The main problem of not considering overdispersion in the model when it is present is the under estimation of standard errors and, consequently, erroneous interpretation and conclusions drawn.

One of the interests of fungus germination data is to assess the Ultraviolet radiation (UV-B) tolerance of reproductive structures (conidia) of entomopathogenic fungi. The conidia germination is valued to select the most tolerant fungal isolate that could be used for the development of biopesticides. However, some isolates may present similar resistance and, in this case, it is necessary to cluster them according to the statistical results. In this paper, we discuss how to select models to properly analyze overdispersed proportion data, using R. We then illustrate applications of generalized linear models and generalized linear mixed models, using fungal germination data. Finally, once the final model is selected, we present methods for grouping isolates and testing for differences between them.

2.2 Case Study

An experiment in a randomized block complete design was conducted in a laboratory to study the UV-B radiation tolerance on fourteen isolates of *Isaria fumosorosea* (Wize). The entomopathogenic

fungus *Isaria fumosorosea* (Wize) Brown & Smith (Ascomycota: Hypocreales: Cordycipitaceae) is commonly found in soil and infecting several species of arthropods (ZIMMERMANN, 2008). Mycoinsecticides based on *I. fumosorosea* were developed and commercialized in the world being used to control insects of the order Hemiptera (whitefly, aphids and scales), Thysanoptera (thrips), Acari (spider mites) and Coleoptera (ZIMMERMANN, 2008). However, in Brazil there is only one bioproduct based on this pathogen to *Diaphorina citri* control. The objective of this experiment was to select one strain of *I. fumosorosea* with high UV-B radiation tolerance for developing of a new mycopesticide based on *I. fumosorosea*. The *I. fumosorosea* strains were obtained from soil and insects of different Brazilian biomes and preserved in the entomopathogen collection of the Laboratory of Pathology and Microbial Control of Insects at the “Luiz de Queiroz” College of Agriculture (ESALQ), University of São Paulo (USP), located in Piracicaba, SP, Brazil. The conidia suspension of each isolate was inoculated in one Petri dish containing culture medium and all Petri dishes were placed to a wooden chamber with four fluorescent lamps emitting UV-B light. Irradiation experiments were conducted in a chamber with four UV-B 313EL lamps (Q-lab Cleveland, OH, USA). The lamps were aged prior to the start the experiments, resulting in a stable level of irradiation of 1312.4 mW/m² and 5.20 KJ/h. Therefore, the exposure times were 0, 2, 4, 6 and 8 hours, which correspond to total UV-B irradiation of 0, 4.76, 9.52, 14.28 and 19.04 KJ/m², respectively. The Petri dishes were covered with a 0.13 mm-thick cellulose diacetate film (Málaga Ltda), which had a cutoff point at 290 nm. This permitted the passage of most UV-B and UV-A (290–360 nm), that was selected based on the spectral characteristics of fungal responses reviewed by Paul et al. (1997). After exposure to UV-B light, the Petri dishes were maintained in a growth chamber set to $26 \pm 2^\circ\text{C}$ under a 12 h photophase for 48 hours to allow the recovery and germination of conidia. The control treatment consisted of Petri dishes containing the conidia suspension of each isolate without exposure to UV-B radiation. Germination of this treatment was quantified after 24 hours of incubation in a growth chamber at 26°C and 12 hours of photophase. The direct count of germinated and non-germinated conidia was performed under a light microscope following the methodology of Oliveira et al. (2015). After that, the dishes were incubated for 48 hours to allow for fungi recovery and conidia germination. The direct counting of the number of germinated and non-germinated conidia was carried out using an optical microscope (400x), and conidia were considered as germinated if their germ tube was twice as long as the spore. All treatments were repeated four times in the same conditions. In each of the replicates (blocks), was prepared one conidia suspension of each isolate and used for inoculation in five Petri dishes corresponding to each exposure time (0, 2, 4, 6 and 8h), adding up to 14 isolates $\times$ 5 exposure time (hours) $\times$ 4 blocks (replicates) = 280 observations. In this case, we may expect correlation between the isolates per block. There is also an inherent variability of the application of fungal suspensions in the center of the Petri dishes due to the use of hand-held instruments (micropipette), or conduction by different people.

The exploratory plots of the data (Figure 2.1) show the proportions of germinated conidia for each fungus, over time of exposure. At the beginning of the experiment, the isolate is less exposed to UV-B radiation, thus, has a higher germination rate. However, the longer the exposure time, greater is the damage to the isolate, which causes the germination reduction and, in some cases, the death of the conidia. There is large variability between blocks within isolates (Figure 2.1(a)) and also between isolates (Figure 2.1(b)). Also, there are different starting points for each curve at time 0 (Figure 2.1), i.e., the isolates have different initial proportions of conidia germination.

Figure 2.1: Plots of (a) observed proportion of germinated conidia vs. exposure time per isolate and (b) averages of observed proportion of germinated conidia vs exposure time for each isolate.

2.3 Statistical Methods

Here, the outcome of interest is the proportion of germinated conidia, hence the response variable is the ratio between the number of germinated conidia, represented by the random variables $Y_i, i = 1, \dots, n$, and the total number of conidia in the i -th Petri dish, $m_i, i = 1, \dots, n$.

If we assume each conidia has the same germination probability, then we can assume the binomial distribution for our variables of interest Y_i . The binomial model is a reasonable starting point to analyze these data, and extensions are available, such as the ones discussed in latter sections. This is a specific example of a generalized linear model (NELDER and WEDDERBURN, 1972).

2.3.1 Introduction to Generalized Linear Models (GLMs)

The class of GLMs consists of three components: (i) the random component of the model, i.e., a univariate distribution belonging to the exponential family, with probability density function (pdf)

$$f(y_i; \theta_i; \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\} \quad i = 1, \dots, n, \quad (2.1)$$

where $b(\cdot)$ and $c(\cdot)$ are known functions, θ_i is called the natural parameter and ϕ the dispersion parameter (for the normal distribution, for example, we have $\theta = \mu$ and $\phi = \sigma^2$, the mean and variance, respectively); (ii) a linear predictor related to the explanatory variables

$$\eta_i = \beta^T \mathbf{x}_i,$$

where β is a vector of p unknown parameters and $\mathbf{x}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{in}]'$ is the i -th column of the $n \times p$ design matrix; and (iii) a link function $g(\mu_i) = \eta_i$, relating the systematic to the random component. We have that the mean is $E[Y_i] = \mu_i$ and the variance is $\text{Var}[Y_i] = \phi b''(\theta_i) = \phi V(\mu_i)$, where $V(\mu_i)$ is called variance function.

Nelder and Wedderburn (1972) proposed the analysis of deviance, generalizing the ideas in standard ANOVA, to assess the significance of effects in the linear predictor. In this framework, a measure that compares a fitted model to the saturated model (i.e. a model that has one parameter per observation), and, for known ϕ , can be used as a measurement of goodness-of-fit for the fitted model, is the residual deviance. For the binomial model, the residual deviance can be written as

$$D_B = 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right],$$

where $\hat{\mu}_i$, $i = 1, 2, \dots, n$, are the fitted values for the current model. Asymptotically (i.e. for a large sample size), D_B has an approximate χ^2 distribution with $n-p$ degrees of freedom (df).

To compare nested models, which are models that contain common terms and one has at least one additional term, by writing D_p for the residual deviance of the full model and D_q for the reduced model, the statistic $D_q - D_p \sim \chi_{q-p}^2$ can be used to test the hypothesis that true coefficient values of omitted terms are zero. This test corresponds to a likelihood-ratio test, and if $D_q - D_p > \chi_{q-p; (1-\alpha)}^2$, the upper $100 \times \alpha$ percentile of the χ_{q-p}^2 distribution, we reject the null hypothesis that the additional parameter is zero at a significance level of α , which means that the parameters tested are important to describe the data and should remain in the model (DEMÉTRIO *et al.*, 2014).

To detect possible model failure we can use diagnostic plots. In general, we compare observed and fitted values using some type of residuals, usually the deviance residuals. Atkinson (1985) proposed the addition of a simulated envelope to half-normal plots to check the goodness-of-fit of a given model to a set of observations. The half-normal plots show the residuals (or any other model diagnostic) should behave if the observed data were a plausible realization of the fitted model, making it possible to detect overdispersion in the data. If many points are outside the envelope, then the fitted model is not suitable for analyzing the data. A fit is usually considered satisfactory when the number of points outside of the simulation envelope falls below 5%. In R, these plots are implemented as the `hnp` package (MORAL *et al.*, 2017).

2.3.1.1 Binomial model

Let be a random variable representing counts of successes out of a sample of size n , $i=1, \dots, n$, then a reasonable assumption would be that $Y_i \sim \text{Bin}(m_i, \pi_i)$, with probability function (pf)

$$P[Y_i = y_i] = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}, \quad y_i = 0, 1, \dots, m_i.$$

We have that $E[Y_i] = \mu_i = m_i \pi_i$ and

$$\text{Var}[Y_i] = m_i \pi_i (1 - \pi_i) \quad (2.2)$$

The GLM framework allows us to model the expected proportions in terms of explanatory variables $\mathbf{x}_1, \dots, \mathbf{x}_n$ (HINDE and DEMÉTRIO, 1998). The usual (canonical) link function for the binomial GLM is the logit

$$g(\mu_i) = \ln \left(\frac{\pi_i}{1 - \pi_i} \right) \quad (2.3)$$

This link function maps the probabilities π_i to the real line, and yields the characteristic sigmoid curve used to describe the behavior of the response variable in dose-response experiments.

Due to π_i is bounded between 0 and 1, $\text{Var}[Y_i] = (1 - \pi_i) E[Y_i] < E[Y_i]$ and hence, when the variability of the data is larger than the mean, this model is not able to accommodate overdispersion. Therefore, extensions of the binomial model can be used to analyze overdispersed data, such as the ones described in the following sections.

2.3.1.2 Quasi-binomial model

The simplest way to deal with overdispersion is based on a quasi-likelihood approach, which requires the specification of the first two moments of the distribution, i.e, the mean and the variance. For the binomial GLM, equation (2.2) is replaced by

$$\text{Var}[Y_i] = \phi m_i \pi_i (1 - \pi_i), \quad (2.4)$$

where ϕ is called the dispersion parameter. Differently from the binomial model, in which ϕ is known ($\phi = 1$), here it is considered as an unknown parameter and a way of estimating ϕ is through

$$\tilde{\phi} = \frac{X_B^2}{n - p}$$

where $X_B^2 = \sum_{i=1}^n \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}$ is the generalized Pearson statistic for the binomial model, a measure of goodness-of-fit, such as the residual deviance.

2.3.1.3 Beta-binomial model

One of the assumptions of the binomial model is that the probability of success is the same for every Bernoulli trial ¹. An alternative is to allow the probability of success to vary according to some distribution. Due to this probability is continuous and can assume values in the interval (0,1), the beta distribution is a suitable option. Let $Y_i|P_i \sim \text{Binomial}(m_i, P_i)$, with $P_i \sim \text{Beta}(a_i, b_i)$. Then marginally, has a beta-binomial distribution with mean and variance given by

$$\text{E}[Y_i] = m_i \pi_i \quad (2.5)$$

and

$$\text{Var}[Y_i] = m_i \pi_i (1 - \pi_i) [1 + \phi (m_i - 1)], \quad (2.6)$$

where $\pi_i = \frac{a_i}{a_i + b_i}$ and $\phi = \frac{1}{a_i + b_i + 1}$ with $(a_i + b_i)$ considered to be constant. This model is often used when the binomial denominator m_i have variation (Demétrio et al., 2014).

Hence, if $\phi = 0$ the variance function in equation (2.6) will be of the same form as in the binomial model (2.2). If the m_i are equal, the variance function in equation (2.6) will be of the same form as in the quasi-binomial model (2.4).

2.3.1.4 Logistic-normal model

Another way to incorporate overdispersion is to assume that there is random variation in the linear predictor η_i . This can be done by including a random effect in the linear predictor, i.e., $\eta_i = x_i' \beta + Z_i$, with $Z_i \sim N(0, \sigma^2)$. Using a logit link (2.3), we obtain the logistic-normal model. The expectation of Y_i is the same as in equation (2.5), and the variance may be approximated by

$$\text{Var}[Y_i] \approx m_i \pi_i (1 - \pi_i) [1 + \sigma^2 (m_i - 1) \pi_i (1 - \pi_i)].$$

This is the simplest case of a GLMM, because it includes fixed and random effects in the linear predictor. In some applications these random effects can be interpreted, e.g. when there is random variation in the conidia concentration of the solutions, this may increase data variability, and observational-level random effects can be included in the linear predictor to explain this extra-variability.

¹If the random variables in a finite or infinite sequence $X_1, X_2, \dots, X_n$ are independent and identically distributed, and if each random variable X_i has the Bernoulli distribution with parameter p , then it is said that $X_1, X_2, \dots, X_n$ are Bernoulli trials with parameter p (DEGROOT and SCHERVISH, 2012)

Here, we use likelihood-ratio tests to assess the significance of the variance components. The null hypothesis for a variance component σ^2 is $H_0 : \sigma^2 = 0$ versus the alternative $H_1 : \sigma^2 > 0$. Due to this test is on the boundary of the parameter space (i.e., the variance cannot be negative), the resulting asymptotic distribution of the test statistic is a mixture of chi-square distributions. For more details, see Verbeke and Molenberghs (2003) .

2.4 Analysis the of case-study

Fitting the binomial model

We begin by fitting a standard binomial logit model (M1), using the maximal linear predictor for this case, i.e., with different logistic regression curves (one for each of the different isolates, with different intercept by block):

$$\eta_{jik} = \gamma_j + \beta_{0i} + \beta_{1i}t_k, \quad j = 1, \dots, 4, \quad i = 1, \dots, 14 \quad \text{and} \quad k = 1, \dots, 5, \quad (2.7)$$

where γ_j is the effect of j -th block, β_{0i} and β_{1i} are the intercept and slope related to the i -th isolate, respectively and t_k is the k -th exposure time. Here, the outcome variable is the proportion of germinated conidia and we have two covariates: the identification of the fungus isolate, represented as a categorical variable in the model, and the dose of exposure time represented in the model as a continuous variable. The R code to carry out the analysis is given in Box 2.1.

Box 2.1: R code to fit the standard binomial model to the germination data.

```
# read data
germination <- read.table('data.txt', header=TRUE, dec=',', sep='\t')
# examining first 6 rows of the dataset
head(germination)
  FUNG    DOSE  BLOCK TOTAL GERM
1 1296      0     1    126  126
2 1296      2     1    112  111
3 1296      4     1    110  108
4 1296      6     1     68   66
5 1296      8     1     74   24
6 1296      0     2     88   88
resp <- with(dados, cbind(GERM, TOTAL-GERM))
# fitting binomial model (M1) using block and fung as factor, dose as numeric and
# the linear predictor (7)
M1 <- glm(resp ~ BLOCK + FUNG*DOSE, family=binomial, data=germination)
# check adequacy of model using hnp
require(hnp)
```

Looking at the half-normal plot with simulated envelope for the deviance residuals (Figure 2.2(a)), there is evidence that the model does not fit the data satisfactorily. We are looking here for clear patterns of systematic deviations from the expected shape of the envelope, which will mean that the observed values are not a plausible realization of the fitted model and thus it should not be used to analyze the data. Here, the reason for this is that there is more variability than the binomial model accommodates, and hence the data is overdispersed.

As suggested by Demétrio et al. (2014), as a next step we may try to accommodate the extra variability by estimating the dispersion parameter with a quasi-binomial model.

Figure 2.2: Germination data – Half-normal plot with simulation envelopes for deviance residuals of: (a) binomial model; (b) quasi-binomial model; (c) Beta Binomial model; all of them with maximal linear predictor (Equation (2.7)).

Fitting the quasi-binomial model

As we can see from equation (2.4), the quasi-likelihood approach estimates the dispersion parameter to allow for greater variability. The quasi-binomial model (M2), with the same linear predictor (2.7), is easily fitted in R using the quasi-binomial family in the `glm` function, and the `hnp` function can still be used to assess goodness-of-fit (Box 2.2).

Box 2.2: R code to fit the quasi-binomial model to the germination data.

```
# fitting the quasi-binomial model (M2) using block and fung as factor, dose as
# numeric and the linear predictor (7)
M2 <- glm(resp ~ BLOCK + FUNG*DOSE, family=quasibinomial, data= germination)
# check adequacy of model using hnp
hnp(M2, paint.out=TRUE, print=TRUE)
```

Again, there is strong evidence of an inadequate model fit, with 80.71% of the observed residuals lying outside the simulated envelope (Figure 2.2(b)).

Fitting the beta-binomial model

The beta-binomial distribution is an alternative approach to account for overdispersion. We can easily fit the beta-binomial model (M3) with the same linear predictor (2.7) in R (Box 2.3) by making use of the `aods3` package (Lesnoff & Lancelot, 2013).

Box 2.3: R code to fit the beta binomial model to the germination data.

```
require(aods3)
# fitting the beta-binomial model (M3) using block and fung as factor, dose as
# numeric and the linear predictor (7)
M3 <- aodml(resp ~ BLOCK+FUNG*DOSE, family='bb', data= germination)
# check adequacy of model using hnp
hnp(M3, paint.out=TRUE, print=TRUE)
```

It appears that this model provides a better fit to the data when compared to the quasi-binomial model. However, there is still a considerable amount of points outside of the simulated envelope when

looking at the half-normal plot of the deviance residuals (Figure 2.2(c)). The next step is to try a model that incorporates random effects to capture this extra-variation.

Fitting the logistic-normal model

In each of the blocks, the same suspensions were used for the inoculation. Then we can add a random effect to account for the correlation between proportions taken on the same isolate and block. Using the logistic-normal model (M4), we can add to the linear predict (2.7) a normal random effect on the intercept

$$\eta_{jik} = \gamma_j + \beta_{0i} + \beta_{1i}t_k + b_{0ij}, \quad j = 1, \dots, 4, \quad i = 1, \dots, 14 \quad \text{and} \quad k = 1, \dots, 5 \quad (2.8)$$

where b_{0ij} is the random effect for the intercept of the i -th isolate and j -th block with variances σ_I^2 . We may easily fit different logistic-normal model (M5) in which a random effect to slope is also included in the linear predictor (2.8)

$$\eta_{jik} = \gamma_j + \beta_{0i} + \beta_{1i}t_k + b_{0ij} + b_{1ij} t_k \quad (2.9)$$

where b_{1ij} is the random effect for the slope of the i -th isolate and j -th block with variances σ_S^2 and $\begin{bmatrix} b_{0ij} \\ b_{1ij} \end{bmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right]$ with $\Sigma = \begin{bmatrix} \sigma_I^2 & \sigma_{IS} \\ \sigma_{IS} & \sigma_S^2 \end{bmatrix}$. These two normal random effects will account for part of the variability induced by the use of the different fungal suspensions.

Another practical explanation for the extra-variation could be the random variation of the conidia concentration. This can be taken into account in the model by adding in the linear predictor (2.9) an observation-level random effect (M6)

$$\eta_{jik} = \gamma_j + \beta_{0i} + \beta_{1i}t_k + b_{0ij} + b_{1ij} t_k + z_{jik}, \quad (2.10)$$

where $z_{jik} \sim N(0, \sigma_O^2)$ is the observation level random effect. For the models presented in equations

(2.9) and (2.10) we assume $\begin{bmatrix} b_{0ij} \\ b_{1ij} \end{bmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right]$ with $\Sigma = \begin{bmatrix} \sigma_I^2 & \sigma_{IS} \\ \sigma_{IS} & \sigma_S^2 \end{bmatrix}$.

We now consider fitting these models making use of the lme4 package (PINHEIRO and BATES, 2006) as showed in Box 2.4. When adding random effects, some convergence problems may arise due to numerical integration problems, so it is important to take into consideration different approaches. Here we are using the bobyqa method (POWELL, 2009) with maximum number of function evaluations allowed equal to 500.000, as it yielded better convergence in our case. The purpose of bobyqa is to minimize a function of many variables by a trust region method that forms quadratic models by interpolation. Another option is to use the Nelder-Mead method, which is the default in glmer.

Box 2.4: R code to fit the logistic-normal models (M4) and (M5) with linear predictors shown in equations (2.8) and (2.9).

```
require(lme4)
# fitting logistic normal-model (M4) with the linear predictor (8)
M4 <- glmer(resp ~ BLOCK+FUNG*DOSE+(1|FUNG:BLOCK), family=binomial,
            data=germination, control=glmerControl(optimizer='bobyqa',
            optCtrl=list(maxfun=500000)))
# fitting logistic normal-model (M5) with the linear predictor (9)
M5 <- glmer(resp ~ BLOCK+FUNG*DOSE+(DOSE|FUNG:BLOCK), family=binomial,
            data=germination, control=glmerControl(optimizer='bobyqa',
            optCtrl=list(maxfun=500000)))
# check adequacy of models using hnp
hnp(M4, paint.out=TRUE, print=TRUE)
hnp(M5, paint.out=TRUE, print=TRUE)
```

The inclusion of these random effects alone were not enough to explain the extra variation (see Figure 2.3(a) and Figure 2.3(b)). A good alternative, with a practical explanation, to model this extra-variation is to add in the linear predictor (2.9) a random effect to each experimental unit (M6). This could explain variation induced by random variation of conidia concentration by dish when preparing the solutions (Box 2.5).

Box 2.5: R code to fit the logistic-normal model (M6) with the linear predictor shown in equation (2.10).

```
# fitting logistic normal-model (M6) with the linear predictor (10)
ind <- gl(nrow(germination),1)
M6 <- glmer(resp ~ BLOCK+FUNG*DOSE + (DOSE|FUNG:BLOCK) + (1|ind), family=binomial,
            data=germination, control=glmerControl(optimizer='bobyqa',
            optCtrl=list(maxfun=500000)))

# check adequacy of models using hnp
hpn(M6, paint.out=TRUE, print=TRUE)
```

The half-normal plot (Figure 2.3(c)) shows no evidence of an inadequate model with most observed residuals lying inside the simulated envelope. It is important to note that every time we produce a simulated envelope, it will slightly change and hence the number of observations lying within the envelope may differ. Now we can assess the significance of the variance components and compare fitted models with one another (Box 2.6). The suitability of the logistic-normal model is consistently verified by comparing the Akaike Information Criteria (AIC) (AKAIKE, 1974a) for the binomial (M1, AIC = 6437), beta-binomial (M3, AI = 1896), and logistic-normal models (M7, AIC = 1749.3). A smaller AIC gives evidence of a better-fitted model, and hence the logistic-normal model would be selected here.

Figure 2.3: Germination data – Half-normal plots with simulation envelopes of deviance residuals for (a) logistic-normal model with linear predictor (Equation (2.8)); (b) logistic-normal model with linear predictor (Equation (2.9)) and (c) logistic-normal model with linear predictor (Equation (2.10)).

Box 2.6: R code for comparing models (M4), (M5) and (M6).

```
# get deviance for models
-2*logLik(M4)
log Lik'. 3036.34
-2*logLik(M5)
log Lik'. 2110.31
Dif1 <- 3036.34-2110.31
-2*logLik(M6)
log Lik'. 1679.27
Dif2 <- 2110.31-1679.27
# p-value for comparing M4 and M5
p <- 0.5*pchisq(Dif1,df=1, lower.tail=F) +
0.5*pchisq(Dif1,df=2, lower.tail=F);p
[1] 4.212566e-202 # M5 (most complex) is selected because p < 0.05
# p-value for comparing M5 and M6
p <- 0.5*pchisq(Dif2,df=0, lower.tail=F) +
0.5*pchisq(Dif2,df=1, lower.tail=F);p
[1]1.315649e-96 # M6 (most complex) is selected because p < 0.05
```

The linear predictor (2.8) includes random intercepts only, assuming all measurements on the i -th isolate and j -th block are correlated. The linear predictor (2.9) includes random linear dose effects b_{0ij} correlated with the random intercepts b_{1ij} and the linear predictor (2.10) includes all effects in the linear predictor (2.9) besides the observation level random effect z_{ijk} . Here, by using likelihood-ratio tests, we can conclude that model (M6) is the most suitable to analyze the data (see Table 2.1).

Table 2.1: Summary of the results of likelihood ratio (LR) tests for the comparison of generalized linear mixed models (M4), (M5) and (M6).

Model	Distribution of random effects	Covariance matrix	Null Hypothesis	LR test ($D_q - D_P$)
M4:	$b_{0ij} \sim N(0, \Sigma)$	$\Sigma = (\sigma_I^2)$		
M5:	$\begin{bmatrix} b_{0ij} \\ b_{1ij} \end{bmatrix} \sim N_2 \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right]$	$\Sigma = \begin{pmatrix} \sigma_I^2 & \sigma_{IS} \\ \sigma_{IS} & \sigma_S^2 \end{pmatrix}$	$H_0 : \sigma_s^2 = \sigma_{IS}^2 = 0$	$3036.34 - 2110.3 = 926$ Reference distribution: $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ $p < 0.0001$
M6:	$\begin{bmatrix} b_{0ij} \\ b_{1ij} \\ z_{ijk} \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right]$	$\Sigma = \begin{pmatrix} \sigma_I^2 & \sigma_{IS} & 0 \\ \sigma_{IS} & \sigma_S^2 & 0 \\ 0 & 0 & \sigma_O^2 \end{pmatrix}$	$H_0 : \sigma_O^2 = 0$	$2110.3 - 1679.3 = 431$ Reference distribution: $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ $p < 0.0001$

After the selection of random effects is concluded, we may now test for parallelism of linear predictors. In this experiment, if we select the linear predictor with separate lines, this will mean that each isolates curve presents different intercepts and slopes (see Box 2.7). If we select the parallel lines linear predictor, the curves present different intercepts but the same slope for all isolates (linear predictor (2.11)). Finally, by selecting the coincident line linear predictor, all isolates present the same intercept and slope (linear predictor (2.12)).

$$n_{jik} = \gamma_j + \beta_{0i} + \beta_1 t_k + b_{0ij} + b_{1ij} t_k + z_{jik} \quad (2.11)$$

$$n_{jik} = \gamma_j + \beta_0 + \beta_1 t_k + b_{0ij} + b_{1ij} t_k + z_{jik} \quad (2.12)$$

Box 2.7: R code to fit linear predictors (11) and (12), and LR test for models (M6), (M7), (M8).

```
# fitting the logistic normal-model (M7) with fung, dose and without interaction
# effect - linear predictor (11) (Parallel lines)
M7 <- glmer(resp ~ BLOCK+FUNG+DOSE + (DOSE|FUNG:BLOCK) + (1|ind), family=binomial,
            data= germination, control=glmerControl(optimizer='bobyqa',
                                                    optCtrl=list(maxfun=50000)))

# fitting the logistic normal-model (M8) with dose, without fung and interaction
# effect - linear predictor (12) (Coincident lines)
M8 <- glmer(resp ~ BLOCK + DOSE + (DOSE|FUNG:BLOCK) + (1|ind), family=binomial,
            data= germination, control=glmerControl(optimizer='bobyqa',
                                                    optCtrl=list(maxfun=50000)))

# LR tests for nested models
anova(M6, M7, M8)
```

The parallel lines logistic-normal model (M7) was selected through likelihood-ratio tests (see Table 2.2). The fitted curves presented in Figure 2.4 are constructed using the linear predictor (2.11) and the mean of the estimated effects of the blocks, presented in Table 2.3 (see Box 2.8). The half-normal plot to model (M7) is presented in Figure 2.5.

Table 2.2: Likelihood-ratio tests for the logistic-normal models with separate, parallel and coincident linear predictors.

Test	χ^2	df	p-value
Separate vs parallel	12.46	13	0.4904
Parallel vs coincident	27.49	13	0.0090

Figure 2.4: Fitted proportions using (M7a) for (a) each isolate in the same plot and (b) each isolate in different plots with points representing the observed values.

Table 2.3: Parameter estimates and Standard Errors for the selected model (M7), without intercept (M7a), before and after grouping the treatments (M12a)

Model M7a - Before Grouping		Model M12a - After Grouping	
Effect	Estimate (Std. Error)	Effect	Estimate (Std. Error)
Block 1	8.2183(0.6456)	Block 1	8.2307(0.6646)
Block 2	7.3519(0.6428)	Block 2	7.3440(0.6606)
Block 3	8.3111(0.6771)	Block 3	8.2635(0.6906)
Block 4	8.6385(0.6510)	Block 4	8.6476(0.6697)
Fung 1741	-1.3479(0.7608)	Group 2	8.6473(0.6697)
Fung 1998	-1.2948(0.7455)	Group 3	-2.6047(0.6115)
Fung 2778	-1.4004(0.7457)	Exposure time	-0.9002(0.0542)
Fung 3300	-2.4699(0.7361)		
Fung 3302	-2.4699(0.7361)		
Fung 3307	-3.2694(0.7398)		
Fung A152-I	-2.5329(0.7397)		
Fung C23-I	-2.5329(0.7397)		
Fung 149-I	-2.3015(0.7302)		
Fung E71-I	-2.5032(0.7298)		
Fung IT1-I2	-1.1730(0.7507)		
Fung IT2-10	-1.3875(0.7767)		
Fung SB4-I7	-1.4418(0.7440)		
Exposure time	-0.9003(0.0541)		
Observation (σ_O^2)	0.8100	Observation (σ_O^2)	0.8024
Intercept Block:Fung (σ_I^2)	1.7906	Intercept Block:Fung (σ_I^2)	1.7272
Slope Block:Fung (σ_S^2)	0.1078	Slope Block:Fung (σ_S^2)	0.1092
Covariance (σ_{IS})	-0.3602	Covariance (σ_{IS})	-0.3560

Box 2.8: R code to create the function to calculate exposure time for germination of 75% and 95%.

```
# fitting logistic normal-model without intercept (-1) to obtain every block effects
M7a <- glmer(resp ~ -1+BLOCK+FUNG+DOSE+(DOSE|FUNG:BLOCK)+ (1|ind),family=binomial,
             data= germination, control=glmerControl(optimizer='bobyqa',
             optCtrl=list(maxfun=50000)))

# get block effects 1 to 4, fung effects 2 to 14 and Dose.
block<-(summary(M7a)$coef[1:4,1])
fungus<-(summary(M7a)$coef[5:17,1])
slope<-(summary(M7a)$coef[18,1])
# the next functions presented only works for this specific model!
# creating function to calculate the estimated exposure time from 1° Fung Isaria 1296
dg_1296 = function(percentil){
  dg_1296 = as.data.frame(round((log(percentil/(1-percentil)) -
  (1/4)*(block[1]+block[2]+block[3]+block[4]))/slope, 2))
  colnames(dg_1296) = paste('germination',percentil*100,'%')
  return (dg_1296)
}
cbind(dg_1296(0.75),dg_1296(0.9))
# creating function to calculate the estimated exposure time from other fungus
dg = function(percentil){
  dg = as.data.frame(round( (log(percentil/(1-percentil)) -
  (1/4)*(block[1]+block[2]+block[3]+block[4]) - fungus[1:13])/slope, 2)
  colnames(dg) = paste('germination',percentil*100,'%')
  return (dg)
}
cbind(dg(0.75),dg(0.9))
```

Figure 2.5: Germination data – Half-normal plot with simulation envelopes of deviance residuals for logistic-normal model with the parallel lines (M7) with linear predictor (Equation 2.11).

We may now use the selected model to determine the exposure time that resulted in some specified proportion of germination, say 75% and 90%. We can obtain these estimates using the code presented in Box 2.8.

Table 2.4: Exposure time (in hours) for 75% and 90% germination by fungus

	75% germination	90% germination
Fungus	Exposure time	Exposure time
1296	7.81	6.59
IT1-I2	6.51	5.29
1998	6.37	5.15
1741	6.31	5.09
2778	6.25	5.03
IT2-I10	6.27	5.05
SB4-I7	6.21	4.99
C23.I	5.69	4.47
E149.I	5.25	4.03
3300	5.07	3.85
E71.I	5.03	3.81
A152.I	5.00	3.78
3302	4.93	3.70
3307	4.18	2.97

The *I. fumosorosea* 1296 was the most tolerant isolate because presented a time to reduce viability of 7.81 and 6.59 hours compared to the most susceptible isolate, *I. fumosorosea* 3307 which presented 4.18 and 2.97 hours (see Table 2.4).

2.4.1 Grouping

Because of the similarity observed between isolates in Figure 2.4(a), some of them were grouped and a likelihood-ratio test was performed to test these differences. The groups were constructed according to the predicted curves in order to group the ones that had a similar behavior. For example the test of groups to be created consisted of 5 groups (Table 2.5).

Table 2.5: Grouping 1 - first grouping according to the proximities between the curves presented in Figure 2.4

Group 1	Group 2	Group 3	Group 4	Group 5
1296	1741	C23.I	3300	3307
	1998		A152.I	
	2778		E149.I	
	IT1-12		E71.I	
	IT2-10		3302	
	SB4-I7			

Box 2.9: R code to create and to test groupings.

```
# creating the Factor Grouping 1
germination$Grouping1 <- germination$FUNG
levels(germination$Grouping1)
[1] "1296" "1741" "1998" "2778" "3300" "3302"
[7] "3307" "A152.I" "C23.I" "E149.I" "E71.I"
[12] "IT1-12" "IT2-10" "SB4-I7"
levels(germination$Grouping1) <- c(1,2,2,2,4,4,5,4,3,4,4,2,2,2)

# model fitted with factor Grouping 1 instead of Fung
M8 <- glmer(resp ~ BLOCK+Grouping1+DOSE+(DOSE|FUNG:BLOCK)+(1|ind), family=binomial,
            data=germination, control=glmerControl(optimizer='bobyqa',
            optCtrl=list(maxfun=50000)))

# testing grouping (equality between isolates of the same group)
anova(M8,M7,test='Chisq')
Df  AIC    BIC   logLik deviance Chisq    Chi    Df    Pr(>Chisq)
M8: 13    1718.1 1765.4 -846.06  1692.1
M7: 22    1735.7 1815.7 -845.86  1691.7 0.396    9      1
# M8(simplest) is selected, because p>0.05, i.e. the isolates within
# the groups do not differ statistically
```

This test was not significant (see Box 2.9), hence there is no evidence against a simpler adequate model, and the isolates within groups are concluded to not differ statistically.

Once the five groups described above were constructed, more similarities were sought between these isolates by testing the clustering of these groups in the attempt to identify the most resistant group. Grouping 2 was created to test resistance similarities between *Isaria* 1296 and isolates of Group 2 (Table 2.5), but this hypothesis was rejected (see Box 2.10). Grouping 3 was created to test similarities between *Isaria* C23.I and Group 2 (Table 2.5). Grouping 4 was created to test similarities between Groups 2, 3 and 4 (Table 2.5), but this hypothesis was rejected (Box 2.10). Finally, Grouping 5 was created to test equality of Groups 4 and Group 5 (Table 2.5).

Box 2.10: R code to created and test different groupings of isolates.

```

# creating factor Grouping 2
germination$Grouping2 <- germination$FUNG
levels(germination$Grouping2) <- c(2,2,2,2,4,4,5,4,3,4,4,2,2,2)
# creating factor Grouping 3
germination$Grouping3<- germination$FUNG
levels(germination$Grouping3) <- c(1,2,2,2,4,4,5,4,2,4,4,2,2,2)
# creating factor Grouping 4
germination$Grouping4<- germination$FUNG
levels(germination$Grouping4) <- c(1,2,2,2,2,2,5,2,2,2,2,2,2,2)
# creating factor Grouping 5
germination$Grouping5<- germination$FUNG
levels(germination$Grouping5) <- c(1,2,2,2,4,4,4,4,2,4,4,2,2,2)
# model fitted with factor Grouping 2
M9 <- glmer(resp ~ BLOCK+ Grouping2 + DOSE + (DOSE|FUNG:BLOCK)
            + (1|ind), family=binomial, data=germination,
            control=glmerControl(optimizer='bobyqa',
                                optCtrl=list(maxfun=50000)))
# model fitted with factor Grouping 3
M10 <- glmer(resp ~ BLOCK + Grouping3 + DOSE + (DOSE|FUNG:BLOCK)
            + (1|ind), family=binomial, data=germination,
            control=glmerControl(optimizer='bobyqa',
                                optCtrl=list(maxfun=50000)))
# model fitted with factor Grouping 4
M11 = glmer(resp ~ BLOCK + Grouping4 + DOSE + (DOSE|FUNG:BLOCK)
            + (1|ind), family=binomial, data= germination,
            control=glmerControl(optimizer='bobyqa',
                                optCtrl=list(maxfun=50000)))
# model fitted with factor Grouping 5
M12 = glmer(resp ~ BLOCK + Grouping5 + DOSE + (DOSE|FUNG:BLOCK)
            + (1|ind),family=binomial, data= germination,
            control=glmerControl(optimizer='bobyqa',
                                optCtrl=list(maxfun=50000)))
# testing equality between isolates of Group 2 and Isolate 1926
anova(M9,M8,test='Chisq')
Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
M9: 12 1721.1 1764.7 -848.56  1697.1
M8: 13 1718.1 1765.4 -846.06  1692.1 5.0031      1    0.0253 *
# M8 (most complex) was selected, because p < 0.05, i.e. Isolate 1296
# differs from isolates in Group 2
# testing equality between isolates from Group 2 and Isolate C23.I
anova(M10,M8,test='Chisq')
Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
M10: 12 1717.2 1760.8 -846.61  1693.2
M8: 13 1718.1 1765.4 -846.06  1692.1 1.1056      1    0.293
# M10 (simplest) was selected, because p > 0.05, i.e. Isolate C23.I
# does not differs from isolates in Group 2
# testing equality between Group 2 plus Isolate C23.I and Group 4
anova(M11,M10,test='Chisq')
Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
M11: 11 1727.3 1767.2 -852.63  1705.3
M10: 12 1717.2 1760.8 -846.61  1693.2 12.031      1 0.0005231 ***
# M10 (most complex) was selected, because p < 0.05, i.e. the isolates
# in Group 2 plus Isolate C23.I differ from the isolates in Group 4

# testing equality between Group 4 and Isolate 3307
anova(M12,M10,test='Chisq')
Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
M12: 11 1717.3 1757.3 -847.64  1695.3

```



```

M10: 12 1717.2 1760.8 -846.61 1693.2 2.0582 1 0.1514
# M12 (simplest) was selected, because p > 0.05, i.e. the isolates
# in Group 4 does not differ from the Isolate 3307

# fitting M12 without intercept (-1) to obtain every block effects
M12a = glmer(resp ~ -1 + BLOCK + Grouping5 + DOSE + (DOSE|FUNG:BLOCK)
            + (1|ind), family=binomial, data= germination,
            control=glmerControl(optimizer='bobyqa',
                                optCtrl=list(maxfun=50000)))

```

According to tests shown above, we can group the isolates in three groups in which the fungi belonging to distinct groups are significantly different at a significance level of 5%.

Group 1: *I. fumosorosea* 1296

Group 2: *I. fumosorosea* 1741, 1998, 2778, IT1-12, IT2-I10, SB4-I7, C23.I

Group 3: *I. fumosorosea* 3300, 3302, A152.I, E149.I, E71.I, 3307

Fitted proportions (M12a) are shown in Figure 2.6 and the half-normal plot for the deviance residuals in Figure 2.7. This graphics indicates that this is also an adequate model fit, and hence this model can be used to make inference on the data. The parameter estimates are given in Table 2.3.

Figure 2.6: Fitted group proportions using a parallel lines logistic-normal for (a) each isolate in the same plot and (b) each isolate in different plots with points representing the observed values.

2.5 Discussion

The aim of this paper was to propose a tutorial, using software R, on how to fit and to assess goodness-of-fit for a range of different models for overdispersion data. We described and showed, step by step, how to fit and interpret the simplest, binomial model, as well as the more complex logistic-normal model.

We use, as a motivation for this tutorial, a dataset from an experiment involving germination of conidia from different species of entomopathogenic fungi, focusing on how to identify and account for

Figure 2.7: Germination data – Half-normal plot with simulation envelopes of deviance residuals for fitted group proportions using a parallel lines logistic-normal model (M12).

specific issues arising from experimental design (such as correlated observations) and extra-variation (i.e. overdispersion).

For the case study presented here, we concluded that the logistic-normal model provided a good fit to the data, and included three random effects: a random intercept per block, a random slope per block and an observation-level random effect to capture overdispersion. Due to the inclusion of three random effects, some convergence problems occurred because of numerical integration problems, so it is important to take into consideration different numerical integration approaches (e.g., Laplace approximation, Adaptive Gauss-Hermite quadrature, Monte-Carlo methods) available to tackle this problem. As a practical outcome, the selected model allowed us to identify the strain of *I. fumosorosea* that presented the highest UV-B radiation tolerance. We were also able to group similar isolates using likelihood-ratio tests, which yields important information in the context of biological control.

The simpler binomial model could have led to mistaken conclusions, such as the selection of the separate linear predictors model or different groupings of isolates. The use of a GLMM solved this issue and allowed us to model the correlation between observations within each block. Hence, we advocate for the use of half-normal plots with simulation envelopes, and other goodness-of-fit assessment techniques, to ensure inference made using the fitted model is reliable.

Finally, we would like to stress that this paper does not present a recipe for analyzing this type of data, since each experiment is different and is carried out to test different types of hypotheses. It is of utmost importance that the modeling takes into account the experimental design and the nature of the data, and that the appropriate hypothesis tests are performed. Every dataset presents its own peculiarities which may guide the statistical analyses one way or another.

References

- AKAIKE, H., 1974a A new look at the statistical model identification. *IEEE transactions on automatic control* **19**: 716–723.
- BRESLOW, N. E. and D. G. CLAYTON, 1993a Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**: 9–25.

- DEGROOT, M. H. and M. J. SCHERVISH, 2012 *Probability and statistics*. Pearson Education.
- DEMÉTRIO, C. G., J. HINDE, and R. A. MORAL, 2014b Models for overdispersed data in entomology. In *Ecological modelling applied to entomology*, pp. 219–259, Springer.
- HINDE, J. and C. G. DEMÉTRIO, 1998 Overdispersion: models and estimation. *Computational Statistics & Data Analysis* **27**: 151–170.
- LAIRD, J. H., NAN M & WARE, 1982 Random-effects models for longitudinal data. *Biometrics* **38**: 963–974.
- LEE, J. A., YOUNGJO & NELDER, 1996 Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**: 619–678.
- MCCULLAGH, P. and J. A. NELDER, 1989 *Generalized linear models*. Chapman and Hall/CRC, London, second edition, 532 p.
- MOLENBERGHS, G., G. VERBEKE, C. G. DEMÉTRIO, A. M. VIEIRA, *et al.*, 2010 A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* **25**: 325–347.
- MOLENBERGHS, G., GEERT & VERBEKE, 2006 *Models for discrete longitudinal data*. Springer, New York, 687 p.
- MORAL, R. A., J. HINDE, and C. G. DEMÉTRIO, 2017 Half-normal plots and overdispersed models in r: The hnp package. *Journal of Statistical Software* **81**: 1–23.
- NELDER, J. A. and W. WEDDERBURN, 1972 Generalized linear models. *Journal of the Royal Statistical Society Series A* **135**: 370–384.
- OLIVEIRA, D. G. P., G. PAULI, G. M. MASCARIN, and I. DELALIBERA, 2015 A protocol for determination of conidial viability of the fungal entomopathogens *beauveria bassiana* and *metarhizium anisopliae* from commercial products. *Journal of microbiological methods* **119**: 44–52.
- PINHEIRO, J. and D. BATES, 2006 *Mixed-effects models in S and S-PLUS*. Springer, New York, 528 p.
- POWELL, M. J., 2009 The bobyqa algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge pp. 26–46.
- VERBEKE, G. and G. MOLENBERGHS, 2003 The use of score tests for inference on variance components. *Biometrics* **59**: 254–262.
- ZIMMERMANN, G., 2008 The entomopathogenic fungi *isaria farinosa* (formerly *paecilomyces farinosus*) and the *isaria fumosorosea* species complex (formerly *paecilomyces fumosoroseus*): biology, ecology and use in biological control. *Biocontrol science and technology* **18**: 865–901.

3 A SIMULATION STUDY USING BOOTSTRAP PROCEDURES APPLIED TO GENERALIZED LINEAR MIXED MODELS

Abstract

The bootstrap method is widely used when more complex models are needed to explain the studies' data characteristics. Based on resampling, or from the original distribution of the population or from the sample dataset, by bootstrap is possible to make inferences and obtain main information, often difficult to obtain analytically. In this work, we proposed a comparison between different types of resampling applying the nonparametric bootstrap method (hierarchical and non-hierarchical approach). As a motivation, we used a fungicidal proportion data, obtained in a dose-response experiment in the Department of Entomology at the University of São Paulo. The main objective was to present an alternative and visual method to find isolates of the fungus *Isaria Fumorosoria*, more resistant to high temperatures, and with the same behavior. We used simulations to replicate the original design and select the best resampling process. Further, we fitted generalized linear mixed models using maximum likelihood and restricted maximum likelihood estimation and constructed confidence intervals aiming to compare isolate types. It was also possible to conclude, for the scenario under study, that resampling following the original characteristics of the experiment, in this case in a hierarchical manner, presents better performance and intervals with better coverage rate. As a practical result, it was possible to present alternative isolates, with equal resistance, that can also be used by the researcher.

Keywords: Confidence interval; Entomology; Nonparametric bootstrap; Propotion data.

3.1 Introduction

Studies focussing on entomology, the area of biology that studies insects, are very useful for biological control (LEVINS *et al.*, 1980). Insects represent the most diverse biological classification group, capable of affecting the agroecosystem in which they inhabit. They could cause damage or even that contribute to the ecosystem, because some species which feed on pests, could reduce economic damage in the crops (VAN EMDEN and WILLIAMS, 1974; THIES *et al.*, 2011). Data from entomological studies can have many characteristics, and a natural experiment in this area is of the dose-response type. Dose-response experiments in this field usually involve insects that live in agroecosystems and cause economic damage to producers, called pests. The interest lies within assessing the effects of different reagents, aiming to determine a lethal dose to be applied to the crops, causing mortality of pests. Outcomes in dose-response experiments are often counts or proportions that can be analyzed using the Generalized Linear Model (GLM) theory (NELDER and WEDDERBURN, 1972). However, these data are usually overdispersed, due to the presence of additional variability, which may lead to the need to develop extended models with extra complexity. Overdispersion occurs for many reasons among them, the variability of the experimental material and the correlation between individual responses (HINDE and DEMÉTRIO, 1998; DEMÉTRIO *et al.*, 2014). However, in these models with more complex structures, it could be problematic to rely on asymptotic results to make inferences.

The generalized linear mixed models (GLMM), an extension of GLM, are commonly used to accommodate overdispersion caused by extra-variability or correlation. These models include fixed and random effects in the linear predictor, allowing us to model different types of data with more flexibility (BRESLOW and CLAYTON, 1993). However, in this case, the standard errors and confidence intervals can be difficult to obtain analytically, and the asymptotic approximation may not give appropriate results. To deal with these problems, bootstrap methods can be a good alternative. The bootstrap method,

initially proposed by Efron (1979), consists of resampling the observed data repeatedly, creating new datasets, aiming to find characteristics of interest in the study population. This resampling method can be useful for constructing confidence intervals (SHAO and TU, 1995; THAI *et al.*, 2013; CHAMBERS and CHANDRA, 2013; MARTÍNEZ and LOUZADA NETO, 2001). When applying bootstrapping techniques, it is advisable to incorporate the original data generating process, such as repeated measures or hierarchical structure (BRUIN, 1999; BRESLOW and CLAYTON, 1993). Two types of bootstraps could be considered, the parametric or the nonparametric. For the parametric bootstrap, it is necessary to specify assumptions or a particular model for the population under investigation, alternatively, in nonparametric bootstrap, we do a resampling procedure from the original datasets (EFRON, 1981; DAVISON *et al.*, 1997; KARLSSON, 2009; MANLY, 2006).

The motivation in this paper comes from a dose-response experiment, carried out in 2015 in the Department of Entomology of the Luiz de Queiroz College of Agriculture at the University of São Paulo (ESALQ), in Piracicaba, Brazil. The aim was to compare different isolates of the same fungus species, detecting those that are more resistant to high temperatures, that is, with a higher proportion of surviving cells. This is done to select the most resistant isolates to ensure elimination of pests. We fitted one curve for each isolate using a binomial GLMM, including four random effects, due to the presence of extra variation of the experimental units (FATORETTO *et al.*, 2018).

The main objective of this paper was to compare the curves and test for differences, detecting similarity between the isolates. Isolates are fungi of the same species but could be founded in different locations or infect different insects. Here, a possible way of doing so is to compare the curves' confidence intervals by checking whether they overlap. However, in these models with hierarchical structures, the construction of analytical confidence intervals is not trivial, therefore we propose to use bootstrap confidence intervals. Then, nonparametric bootstrap and confidence intervals were compared using simulation techniques. Here, we compared the obtained results considering hierarchical and nonhierarchical resampling structures of the original dataset. After the resampling, the models were fitted to the datasets using the maximum likelihood and the restricted maximum likelihood estimation methods (ML and REML respectively), aiming to find better confidence intervals with less computational effort. Besides, we constructed percentile and high-density confidence intervals, checking for better coverage rates. The extension of REML for GLMMs is not as straightforward as in the linear mixed model, but approximate REML was implemented (MILLAR, 2011).

In the following sections, details of the experiment are described in Section 3.2. Section 3.3 presents the theory of models and estimation methods used. Section 3.4 presents the simulation studies and Section 3.5 the results obtained. Finally, Section 3.6 presents the discussion of this work.

3.2 Motivation

Some insects live in agroecosystems and cause economic damage to producers. They are referred to as pests. Biological control, which involves the promotion of pest suppression using fungi, referred to as natural enemies. Hence, no chemical pesticides are used, assuring better crop quality and disease risk prevention for humans (KOGAN, 1998).

The fungus *Isaria fumosorosea* is commonly found in soil and infects several species of arthropods (invertebrates insects with jointed legs). There are variations of these fungi that affect different places or infect different types of insects. These variations are called Isolates. Fungi-based biopesticides have the capacity to infect a large number of pests and to remain in the environment. The fungus spreads conidia or spores that attach to the insect and then germinate. The germinated spores will penetrate the outer skin of the insect and start to grow inside of it. As the fungus continues to grow inside the insect, the insect will die. However, ultraviolet radiation influences and can inhibit the conidia germination and

hence the fungicidal efficacy.

An experiment, using a randomized complete block design, was conducted in a laboratory at ESALQ, to study the UV-B radiation tolerance of 14 different isolates of *Isaria fumosorosea* to select isolates with high UV-B radiation tolerance for developing a new mycopesticide. Within each of four blocks, one conidia suspension of each isolate was prepared, and was used for inoculation in five Petri dishes corresponding to each exposure time (0, 2, 4, 6 and 8h). The seventy dishes were placed in a wooden chamber with four fluorescent lamps and remained there until their corresponding exposure time. Here the response is the proportion of germinated conidia and the higher this proportion, the higher the isolate tolerance is, meaning that the isolate has more resistance.

FATORETTO *et al.* (2018) presented a logistic-normal model that provided a good fit to this data. The model included fixed effect for block, isolate and exposure time, and three random effects: a random intercept and slope per isolate and block, and a observation-level random effect. The model was specified as $Y_{ijk}|b_{0ij}, b_{1ij}, z_{ijk} \sim \text{binom}(m_{ijk}, \pi_{ijk})$, where Y_{ijk} is the total number of conidia in the i -th block, j -th isolate and k -th exposure time. The linear predictor was specified as:

$$\eta_{ijk} = \ln \left(\frac{\pi_{ijk}}{1 - \pi_{ijk}} \right) = \gamma_i + \beta_{0j} + \beta_1 t_k + b_{0ij} + b_{1ij} t_k + z_{ijk}, \quad (3.1)$$

where $i = 1, \dots, 4$ for block; $j = 1, \dots, 14$ for isolate and $k = 1, \dots, 5$ for exposure time; γ_i is the effect of i -th block, β_{0j} is the intercept referring to the j -th isolate, β_1 is the slope across time common for all isolates and t_k is the k -th exposure time, b_{0ij} and b_{1ij} are the random effects for the intercept of the i -th block

and j -th isolate, and z_{ijk} is the observational-level random effect with

$$\begin{bmatrix} b_{0ij} \\ b_{1ij} \\ z_{ijk} \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right]$$

with $\Sigma = \begin{pmatrix} \sigma_I^2 & \sigma_{IS} & 0 \\ \sigma_{IS} & \sigma_S^2 & 0 \\ 0 & 0 & \sigma_O^2 \end{pmatrix}$. The first two normal random effects account for the variability induced by the use of the different fungal suspensions.

They identified a particular isolate (labelled number 1296) that was the most tolerant. Further, constructed likelihood-ratio tests to verify the similar behavior between isolates to find isolates alternatives giving more options to the researcher. However, the model (eq. 3.1) presents multiple hierarchies, then can be more efficient, to compare the similarity between isolates, by approximate the empirical distribution of the statistics using a nonparametric bootstrap approach and constructing bootstrap confidence intervals, which is the propose of this paper.

3.3 Methodology

3.3.1 Background of Generalized Linear Mixed Models

The generalized linear mixed model (GLMM) is a flexible framework which allows for accommodating many characteristics present in discrete or continuous data, such as repeated measurement, overdispersion and other. It is a rather straightforward extension of the generalized linear model (GLM) and an elegant framework proposed by Nelder and Wedderburn in 1972, that provide a unified theory for application of observations that come from some distribution in the exponential family (MOLENBERGHS and VERBEKE, 2006). According to McCulloch (2003), the idea of the generalized linear mixed model is incorporate in addition to the usual fixed effects, one or more random components into the linear predictor of a generalized linear model. The GLMMs are introduced by Pierce and Sands (1975) considering

the random components with assumed normal distributions. These models are very adaptable, allowing a rich set of predictor variables, accommodating correlated observations and extra-variation. Later, Lee and Nelder (1996) showed the use of GLMMs to hierarchical datasets, in which the distribution of random effects are extended to conjugates of arbitrary distributions from the GLM family (LEE *et al.*, 2018).

The GLMMs have the following definition

$$\begin{aligned} Y_i | \mathbf{u} &\sim \text{indep. } f_{Y_i | \mathbf{u}}(y_i | \mathbf{u}) \\ \boldsymbol{\eta}_i &= \mathbf{g}(\boldsymbol{\mu}_i) = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_i \mathbf{u} \\ \mathbf{u} &\sim f_U(\mathbf{u}), \end{aligned}$$

where Y_i has a distribution belonging to the exponential family, $g(\cdot)$ is the link function, applied to the conditional mean of Y_i given $\mathbf{u}$ to obtain the conditional linear predictor $\boldsymbol{\eta}_i$, and the linear predictor has two components, the fixed effects portion ($\mathbf{x}'_i \boldsymbol{\beta}$) and the random effects portion ($\mathbf{z}'_i \mathbf{u}$) (McCULLOCH, 2003).

Estimation of GLMMs can be done using maximum likelihood (ML). The likelihood function may be written as

$$L = \int \prod_i f_{Y_i | \mathbf{u}} f_U(\mathbf{u}) d\mathbf{u}, \quad (3.2)$$

where the integration is over the q -dimensional distribution of $\mathbf{u}$ (BRESLOW and CLAYTON, 1993; MCCULLOCH and SEARLE, 2001). In the normal case, this integral has an analytic solution, however, in the other cases, there is no analytics solutions and need for numerical integration to resolve its, as Laplace approximation or Gauss-Hermite quadrature 3.2 (MOLENBERGHS and VERBEKE, 2005).

An alternative for estimating variance components is the restricted (or residual) maximum likelihood (REML). Defined previously by Patterson and Thompson (1971), the restricted likelihood function was described as the marginal likelihood of a set of residual contrasts, hence, alternatively, it may be computed as an integrated likelihood, following a Bayesian or profile likelihood function (BELLIO and BRAZZALE, 2011). The extension of REML for GLMMs is not straightforward as in the linear mixed model, although approximated REML can be implemented by integrated likelihood. Breslow and Clayton (1993) presented the use of REML to estimate parameters in GLMMs using the marginal quasi-likelihood method (GOLDSTEIN, 1986), but they concluded that for binary data this method could results in biased estimates. Noah and Lee (2007) published a comparison between other proposed methods and concluded that REML could be extended to non-normal models via hierarchical-likelihood (h-likelihood) as the best alternative. In this case, they considered the h-likelihood, instead of the classical likelihood, a special kind of extended likelihood, where the scale of the random parameters is specified to satisfy the canonical scale. As an alternative for many models that do not have a canonical scale, Lee *et. al.* (2006) showed that for orthogonal parameters, the integrated likelihood is approximately the adjusted profile likelihood. For more details, see Lee *et. al* (2006) and Noah and Lee (2007). For these reasons, we consider the use of REML estimation in this paper.

3.3.2 Bootstrap methods

The bootstrap method proposed by Efron (1979) is an alternative approach to asymptotic methods for estimating the standard error of parameters and to provide confidence intervals without making assumptions about the distribution of the data. Its main feature is to treat the data as if they were the population and draw samples (with replacement) from the data as if sampling from the population.

The bootstrap was originally developed for parametric inference, given independent and identically distributed data. However, nowadays, bootstrap procedures may be classified into two main categories, namely nonparametric and parametric bootstrap methods.

In the nonparametric bootstrap, the unknown original distribution may be replaced by the empirical distribution of the sample. The empirical distribution function (EDF) $\hat{F}$ gives equal weight probabilities n^{-1} at each sample value y_i

$$\hat{F} = \frac{\#\{y_i \leq y\}}{n}.$$

The values of $\hat{F}$ are fixed $(0, 1/n, 2/n, \dots, n/n)$, so $\hat{F}$ is equivalent to its points of increase, the ordered value $y_{(1)} \leq \dots \leq y_{(n)}$ of the data (DAVISON *et al.*, 1997).

The feature of the nonparametric bootstrap is to simulate data from the $\hat{F}$. On the other hand, in the parametric bootstrap, the underlying distribution function F is estimated from the data using a parametric model and bootstrap samples are generated by simulating from a distribution. The difference from the nonparametric bootstrap is that the samples are drawn from a parametric estimate of the population rather than the nonparametric estimate $\hat{F}$ (EFRON and TIBSHIRANI, 1994a). In this work we will use the nonparametric case and so will now describe in more detail.

A usual nonparametric bootstrap algorithm is as follows:

1. Generate a bootstrap sample of size n by sampling with replacement from the entire data, that is, the joint vector of design variables and corresponding responses;
2. Compute the estimate of interest $(\hat{\theta}^*)$;
3. Repeat steps (1) and (2) B times;
4. Compute the average of the parameter estimates

$$\hat{\theta}_B = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^*), \quad (3.3)$$

where $\hat{\theta}_b^*$ is the estimated parameter for the b th bootstrap sample, $(b = 1, \dots, B)$;

5. Calculate the bootstrap standard error

$$\widehat{se}_B = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b^* - \hat{\theta}_B)^2};$$

The limit of $\widehat{se}_B$ as B goes to infinity is the ideal bootstrap estimator ($se_{\hat{F}}(\hat{\theta}^*)$), hence, the empirical standard deviation approaches the population standard deviation as the number of replications grows large (EFRON and TIBSHIRANI, 1994a):

$$\lim_{B \rightarrow \infty} \widehat{se}_B = se_{\hat{F}}(\hat{\theta}^*).$$

The ideal bootstrap estimator and its approximation ($\widehat{se}_B$) are called *nonparametric bootstrap* estimates because they are based on $\hat{F}$.

6. Calculate a confidence interval (different methods will be described below)

There are several methods that can be used to construct bootstrap confidence intervals, such as bootstrap- t , bootstrap percentile, bias-corrected percentile bootstrap (BCPB) and others (EFRON and TIBSHIRANI, 1994a; DAVISON *et al.*, 1997).

A $100(1 - \alpha)\%$ bootstrap percentile confidence interval (CI) can be constructed, for example, by computing the $100\frac{\alpha}{2}\%$ and $100(1 - \frac{\alpha}{2})\%$ percentiles of the bootstrap distribution forming the bounds

$$(\hat{\theta}_{(\frac{\alpha}{2})}^*, \hat{\theta}_{(1-\frac{\alpha}{2})}^*), \quad (3.4)$$

where, usually $\alpha = 0.05$ (THAI *et al.*, 2013).

Another option is the Highest Density Interval (HDI) useful for asymmetric distributions. The HDI is the interval which contains the required mass such that all points within the interval have a higher probability density than points outside the interval. This is often applied to a Bayesian posterior distribution, but can be applied to any distribution, including priors (BARRY, 2011).

For single-level models, it is possible to implement a simple procedure for random sampling with replacement. But, in some cases, with multilevel or hierarchical data, such as the motivation study in this paper (Section 3.2), it would be better to respect the original experimental design when resampling from the data, and to use bootstrap techniques that reproduce their hierarchical dependence structure.

In this paper, we chose to use the nonparametric bootstrap to present a more generalized and easy to understand approach to the applied areas, once it relieves from having to make parametric assumptions about the form of the underlying population (EFRON and TIBSHIRANI, 1994b).

Following the bootstrap algorithm described above, we study the nonparametric bootstrap considering two resampling methods: without respecting the hierarchical structure of the experiment and respecting the hierarchical structure. After, the new datasets resampled were fitted using the same model fitted to the original dataset (eq. 3.1). Two confidence intervals were calculated for each parameter: the percentile and the HDI. Finally, we compared the intervals performance of the fitted binomial GLMM (eq. 3.1) models by two estimation methods: ML and REML. The parameters of the fitted models were estimated by the ML and REML methods using the glmmTMB package (MAGNUSSON *et al.*, 2017) in software R (R CORE TEAM, 2018). The use of glmmTMB package compared to lme4 (BATES *et al.*, 2007), widely used R packages for fitting GLMMs, provides an advantage in computational speed, underlying implementation using Template Model Builder (TMB) (FOURNIER, 1993). This method enables quick implementation, maximizing the Laplace approximation of the marginal likelihood where the random effects are automatically integrated out (KRISTENSEN *et al.*, 2015).

3.4 Simulation studies

In this Section, we present simulation studies, designed to assess the performance of nonparametric bootstrap confidence intervals under three resampling methods, focussing on the case-study experimental design.

The data were simulated from a randomized complete block design, with four blocks ($I = 4$) and two factors (isolate and exposure time) with fourteen ($J = 14$) and five ($K = 5$) levels, respectively, totaling $n = 280$ observations, inspired by the original design described in Section 3.2. True parameter values were fixed as shown in Table 3.2 for all simulation scenarios. These values are based on the estimated values from the original dataset presented by Faretto *et al.* (2018).

The study started with $B = 1000$ and $B = 2000$ bootstrap samples for each resampling method. After obtaining the bootstrapped samples, the logistic-normal models were fitted to each sample using two methods of estimation, maximum likelihood (ML), and restricted maximum likelihood (REML).

We compared three resampling algorithms: (1) Resampling from plots, without considering the hierarchy between block and plot; (2) Resampling in the same way as the data generating mechanism, first, resampling from blocks, and then within each block sampled; and (3), resampling in the same way as method (2), but also, creating a new index for block levels (Table 3.1). In method (3), each bootstrap sample has all block levels, even if some block was not sampled.

Table 3.1: Step-by-step of the three resampling algorithms compared. In all methods, following the case-study: I is the number of blocks; J is the number of isolates and K is the exposure time

Method 1 - Nonparametric bootstrap sampling from plots
i) resample all IK plots with equal probability; ii) fit the logistic-normal model (eq. 3.1); and iii) compute parameter estimates $\gamma_i^{(b)}$, $\beta_{0j}^{(b)}$, $\beta_1^{(b)}$, $\sigma_B^{(b)}$, $\sigma_O^{(b)}$, $\sigma_I^{(b)}$, $\sigma_S^{(b)}$ and $\sigma_{IS}^{(b)}$.
Method 2 - Hierarchical nonparametric bootstrap sampling from blocks
i) resample I blocks with equal probability; ii) within each block, resample JK plots with equal probability; iii) fit the logistic-normal model (eq. 3.1); and iv) compute parameter estimates $\gamma_i^{(b)}$, $\beta_{0j}^{(b)}$, $\beta_1^{(b)}$, $\sigma_B^{(b)}$, $\sigma_O^{(b)}$, $\sigma_I^{(b)}$, $\sigma_S^{(b)}$ and $\sigma_{IS}^{(b)}$.
Method 3 - Hierarchical nonparametric bootstrap sampling from blocks and renaming blocks
i) resample I block with equal probability; ii) within each block, resample JK plots with equal probability; iii) create a new block factor with I levels and JK plots within each level; iv) fit the logistic-normal model (eq. 3.1) using the new block variable; and v) compute parameter estimates $\gamma_{new_i}^{(b)}$, $\beta_{0j}^{(b)}$, $\beta_1^{(b)}$, $\sigma_B^{(b)}$, $\sigma_O^{(b)}$, $\sigma_I^{(b)}$, $\sigma_S^{(b)}$ and $\sigma_{IS}^{(b)}$.

For each bootstrap sampling method (1, 2 and 3), $R = 100$ simulations were performed. For each r -th simulated dataset the bootstrap estimated parameters ($\hat{\theta}_B^r$), the percentile confidence interval (CI) and the highest density interval (HDI) were computed for each parameter (eq. 3.4). Then the coverage rate of these intervals was calculated. Their performance was compared using a nominal coverage rate of 95%. The coverage rate was defined as the percentage of the $R = 100$ simulated datasets in which the intervals contain the true parameter values (Table 3.2). For each scenario, we considered the coverage rate of 95% as poor ($< 80\%$), low (from 80% to 90%) and good (from 90% to 100%) as suggested by Thai et. al. (2013). Besides, the estimation methods (ML and REML) and the interval types (CI and HDI) were compared using methods 2 and 3, $B = 1000$ samples, and $R = 100$ simulations.

Table 3.2: True parameter values used in the simulation studies

Parameter	Value
γ_1 - Block 1	8.0
γ_2 - Block 2	7.0
γ_3 - Block 3	8.0
γ_4 - Block 4	8.0
β_{02} - Isolate 2	-1.5
β_{03} - Isolate 3	-1.5
β_{04} - Isolate 4	-1.5
β_{05} - Isolate 5	-2.5
β_{06} - Isolate 6	-2.5
β_{07} - Isolate 7	-3.0
β_{08} - Isolate 8	-2.5
β_{09} - Isolate 9	-2.0
β_{010} - Isolate 10	-2.0
β_{011} - Isolate 11	-2.5
β_{012} - Isolate 12	-1.2
β_{013} - Isolate 13	-1.4
β_{014} - Isolate 14	-1.5
β_1 - Exposure time	-0.9
σ_O^2 - Observation	0.81
σ_I^2 - Intercept Block:Isolate	2.00
σ_S^2 - Slope Block:Isolate	0.10
σ_{IS} - Covariance	-0.40

3.5 Results

3.5.1 Simulation results

Considering a simulated dataset using the parameters' values described in Table 3.2 we sampled $B = 2000$ bootstrap samples for each method presented in section 3.4. Then, using the ML estimator, we fitted the GLMMs for each sample. The results showed that all methods yielded slightly asymmetric distributions for the parameter estimates (Figure 3.1 and 3.2), hence the HDI would be more adequate than the percentile CIs. Method 1 presented more bias for all parameters when compared to the others, what was to be expected because this resampling is non-hierarchical, however however less variation. The behavior of methods 2 and 3 is very similar for the β 's parameters that present the isolate effects. We only see differences between these methods for the γ 's parameters (block effects) and for the variance components that also depend on the blocks. Method 3 presented a slight bias for the parameter γ_2 , this occurred, probably, because the true value of this parameter is smaller than the other block effects (Table 3.2), once the estimated block effect converges to the mean. Thus, this information was lost when the levels of the blocks sampled were renamed. Besides, considering these two methods for some parameters, we can also observe a greater bias even in methods 2 and 3, e.g. $\beta_4, \beta_7, \beta_{14}$ and β_1 (Figures 3.1). Other characteristics possible to observe for some parameters are leptokurtic distributions (heavy tails) have positive excess kurtosis. For the variance components, we can clearly see the asymmetry. Also, method 1 presented a slightly greater bias than the others again for almost all evaluated components (Figures 3.2).

Figure 3.1: Density curves of estimated parameters considering $B = 2000$ bootstrap samples using method 1, method 2 and method 3 - The red line shows the true values.

Figure 3.2: Density curves of estimated variance components considering $B = 2000$ bootstrap samples using method 1, method 2 and method 3 - The red line shows the true values.

The coverage rate using a nominal rate of 95%, considering the three proposed methods and $R = 100$ simulations, were obtained aiming to compare the performance of the percentile confidence intervals (CI) (Table 3.3). We observed a poor coverage rate ($< 80\%$) for some parameters in method 1, moreover, this method had a larger computational effort, taking longer to compute than the other two. Method 2 and method 3 did not show important differences between coverage rates.

Table 3.3: Coverage rate of the 95% of CI of parameters by ML obtained by bootstrap methods 1, 2 and 3, using $B = 1000$ bootstrap samples and $R = 100$ simulations. Poor coverage rate ($< 80\%$) is typeset in red color.

Method	γ_1	γ_2	γ_3	γ_4	β_{02}	β_{03}	β_{04}	β_{05}	β_{06}	β_{07}	β_{08}	β_{09}	β_{010}	β_{011}	β_{012}	β_{013}	β_{014}	β_1	σ_I	σ_S	σ_{IS}	σ_O
1	0.76	0.76	0.73	0.75	0.79	0.82	0.78	0.75	0.75	0.78	0.83	0.78	0.83	0.82	0.77	0.68	0.8	0.72	0.93	0.87	0.93	1
2	0.89	0.82	0.84	0.87	0.92	0.9	0.9	0.94	0.96	0.88	0.91	0.89	0.91	0.87	0.92	0.88	0.91	0.8	0.97	0.92	0.94	0.98
3	0.98	0.82	0.98	0.98	0.87	0.88	0.87	0.9	0.92	0.91	0.94	0.87	0.87	0.91	0.9	0.88	0.89	0.87	1	0.97	1	0.95

The ML estimation method presented a good coverage rate (close to 95% for all parameters), and there were no major differences in the coverage rate intervals for fixed and random components considering methods 2 and 3 (Figure 3.3a,c). The REML estimation method presented a better coverage rate for method 3 than for method 2, for fixed effects and variance components (Figure 3.3b and 3.3d). Both interval types presented similar behavior for method 3. The biggest difference was observed with

REML and method 2, in which the HDI presented the worst coverage rate (Figure 3.3b). When it comes to the processing time, REML is twice as fast as ML.

Figure 3.3: Coverage rate of the 95% of CI and HDI of parameters by ML and REML obtained by bootstrap method 2 and 3, using $B = 1000$ bootstrap samples and $R = 100$ simulations. (a) and (c) ML estimation method. (b) and (d) REML estimation method

Taking a look at the width and lower limits of the percentile confidence intervals (Figure 3.4), for parameters estimated using REML, the interval, presented a slightly better precision than ML (Figure 3.4a). REML presented less variability between lower limits (Figure 3.4b), but apparently, there was not a large difference between estimation methods. There is no extensive coverage rate difference between simulation with $B = 1000$ and $B = 2000$, considering the REML method and CIs (Table 3.4).

Figure 3.4: Boxplots for comparison between REML and ML estimation methods. (a) Width of 100 percentile confidence interval for each parameter using ML and REML estimation methods and $B=1000$. (b) Lower limits of 100 percentile confidence interval for each parameter using estimated by ML and REML estimation methods and $B=1000$.

Table 3.4: Coverage rate of the 95% of CI parameters by REML obtained by bootstrap method 3, using $B = 1000$ and $B = 2000$ bootstrap samples and $R = 100$ simulations. Coverage rate closer 95% is typeset in bold font.

B	γ_1	γ_2	γ_3	γ_4	$\beta_{0,2}$	$\beta_{0,3}$	$\beta_{0,4}$	$\beta_{0,5}$	$\beta_{0,6}$	$\beta_{0,7}$	$\beta_{0,8}$	$\beta_{0,9}$	$\beta_{0,10}$	$\beta_{0,11}$	$\beta_{0,12}$	$\beta_{0,13}$	$\beta_{0,14}$	β_1	σ_I	σ_S	σ_{IS}	σ_O
1000	0.95	0.88	0.95	0.95	0.94	0.88	0.88	0.90	0.92	0.93	0.86	0.93	0.87	0.92	0.88	0.95	0.96	0.93	0.97	0.93	0.99	0.95
2000	0.94	0.88	0.93	0.94	0.89	0.93	0.93	0.85	0.94	0.92	0.94	0.90	0.95	0.90	0.95	0.93	0.93	0.91	0.97	0.94	0.99	0.93

3.5.2 Practical results

As an alternative to what was presented by Faretto et. al. (2018), here we compare the resistance to UV-radiation of different isolates by comparing confidence intervals obtained using bootstrap method 3, presented in the previous section. Firstly we fitted the model (eq. 3.1) using the original data and obtained the predicted curves (Figure 3.5). The curves presented a plausible realization to the data-set (Figure 3.5b) and Isolate 1 presented the most resistance to UVB-radiation (Figure 3.5a).

Then, we plotted the predicted curves by the REML method using the data resampled by the bootstrap method 3 and their respective percentile confidence intervals of 95% (Figure 3.6). This plot showed the same behavior between the fitted curves for the same isolate, showing the good performance of the resampled bootstrap method. In addition, the blue curves (percentage confidence intervals) apparently encompasses 95% of the black curves for most of the isolates, close to the nominal level of 95% and also confirming the good resampling performance. Then, we may use these intervals to check the similarity between isolates (Figure 3.7). Isolate 1 showed greater resistance to high temperatures, so the other isolates were compared to it. We compared the intervals obtained from Isolate 1 with the others and we could observe differences between 1 and 7, 1 and 8, and, 1 and 11 only, as there is no intersection between the estimated intervals.

Figure 3.5: Predicted curves using average of parameter estimates considering 1000 bootstrap samples. (a) Predicted curves over exposure time for each isolated. (b) Predicted curves and observed data over exposure time for each isolate.

Further we compare this results testing the similarity between isolates using the ML estimation and likelihood-ratio tests (Table 3.5). Some differences could be observed. The tests presented significant differences (5% significance level) in the resistance of Isolate 1 comparing with 5, 6, 9, and 10. This difference is not observed in the confidence interval method, since the curves of these isolates show interactions (Figure 3.7). In the next section we will discuss a little about the possible causes of these differences found.

Table 3.5: Summary of the results of likelihood ratio (LR) tests for the comparison of isolate 1 with others.

Test	Chisq	1	Pr(>Chisq)
Isolate 1 vs Isolate 2	3.0618	1	0.0802
Isolate 1 vs Isolate 3	2.9925	1	0.0837
Isolate 1 vs Isolate 4	3.4765	1	0.0622
Isolate 1 vs Isolate 5	10.4670	1	0.0012
Isolate 1 vs Isolate 6	11.7750	1	0.0006
Isolate 1 vs Isolate 7	17.1870	1	<0.0001
Isolate 1 vs Isolate 8	10.9680	1	0.0009
Isolate 1 vs Isolate 9	6.3577	1	0.0117
Isolate 1 vs Isolate 10	9.4747	1	0.0021
Isolate 1 vs Isolate 11	11.0460	1	0.0009
Isolate 1 vs Isolate 12	2.4230	1	0.1196
Isolate 1 vs Isolate 13	3.0796	1	0.0793
Isolate 1 vs Isolate 14	3.6937	1	0.0546

Figure 3.6: One thousand predicted curves (black line) for each isolate with their respective percentile confidence interval (blue line)

3.6 Discussion

Alternative bootstrap methods have been proposed for linear mixed models, however few studies have been proposed for generalized linear mixed models. In this work, simulation studies were carried out to investigate the behavior of different nonparametric bootstrap methods for generalized linear mixed models. Due to parametric simulation involves simulating new sets of data from the fitted parametric model, then has the usual disadvantage that datasets generated from a poorly fitting model may not have the statistical properties of the original data. This applies particularly when data are overdispersed unless the overdispersion has been successfully modeled (DAVISON *et al.*, 1997). However, the parametric approach could also be used. By using different resampling methods it was possible to assess the performance of different types of confidence intervals. The hierarchical sampling methods presented a better coverage rate, which was expected, since they represent a plausible realization of the experiment.

As a practical objective of this work, the interest was to compare resistance curves of different

Figure 3.7: Percentile confidence intervals for Isolate 1 comparing with the other isolates.

isolates. The resampling proved to be satisfactory, and the models fitted from these data showed plausible estimatives in the bootstrap method. In addition, it is possible to compare Isolates with the Isolate 1 (the most resistant) visually, it becomes simple for researchers in applied areas.

One of the interesting features observed in this work was the difference in the results when comparing isolates using the bootstrap interval and the likelihood ratio test. For some comparisons, the tests showed significant difference, while the confidence intervals did not, that is, the test was less conservative, given that it is asymptotic and depends on a larger number of samples. Besides, the likelihood ratio did a pairwise comparison (the significance level is by comparison, then the joint level will be greater), while in the bootstrap is by comparisonwise (the level of significance is jointly).

In this work, we can observe that to perform the bootstrap resampling it is necessary to consider the data hierarchy. Resampling hierarchically the data becomes more faithful to the experiments, showing better results. For complete block designs, when the block effects do not present significant differences, it is interesting to rename these sampled blocks as in the case of this simulation. This renaming contributes to the resampled data continuous balanced, and to the performance of the models fitted.

The coverage did not show significant gains comparing 1000 or 2000 bootstrap samples. Of course, larger the sample should be better the obtained results. However, we succeeded to show here that a resampling with $B = 1000$ also presented plausible results.

In addition, the simulations showed a significant gain in terms of processing time when the

restricted maximum likelihood was used instead of maximum likelihood, without loss of accuracy.

We also compared other types of confidence intervals, correcting the estimation bias, but as they do not present gains in relation to the intervals presented here, they were not shown.

Other approaches can be explored, such as the parametric and semi-parametric bootstrap, which uses residual distribution for resampling. As the next steps for this work, we will compare these other approaches, checking if there are any gains compared to the approach proposed in this paper.

References

- BARRY, J., 2011 Doing bayesian data analysis: A tutorial with r and bugs. *Europe's Journal of Psychology* **7**: 778.
- BATES, D., D. SARKAR, M. D. BATES, and L. MATRIX, 2007 The lme4 package. R package version **2**: 74.
- BELLIO, R. and A. R. BRAZZALE, 2011 Restricted likelihood inference for generalized linear mixed models. *Statistics and Computing* **21**: 173–183.
- BRESLOW, N. E. and D. G. CLAYTON, 1993 Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**: 9–25.
- BRUIN, J., 1999 Mixed effects logistic regression | r data analysis examples.
- CHAMBERS, R. and H. CHANDRA, 2013 A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics* **22**: 452–470.
- DAVISON, A. C., D. V. HINKLEY, *et al.*, 1997 *Bootstrap methods and their application*, volume 1. Cambridge university press.
- DEMÉTRIO, C. G., J. HINDE, and R. A. MORAL, 2014 Models for overdispersed data in entomology. In *Ecological modelling applied to entomology*, pp. 219–259, Springer.
- EFRON, B., 1979 Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 1–26, Ann. Stat .
- EFRON, B., 1981 Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**: 589–599.
- EFRON, B. and R. J. TIBSHIRANI, 1994a *An introduction to the bootstrap*. CRC press.
- EFRON, B. and R. J. TIBSHIRANI, 1994b *An introduction to the bootstrap*. CRC press.
- FATORETTO, M. B., R. D. A. MORAL, C. G. B. DEMÉTRIO, C. S. DE PÁDUA, V. MENARIN, V. M. A. ROJAS, C. P. D'ALESSANDRO, and I. DELALIBERA, 2018 Overdispersed fungus germination data: statistical analysis using r. *Biocontrol Science and Technology* pp. 1–20.
- FOURNIER, D., 1993 Ad model builder .
- GOLDSTEIN, H., 1986 Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73**: 43–56.
- HINDE, J., C. G. DEMÉTRIO, *et al.*, 1998 Overdispersion: models and estimation. *Computational Statistics and Data Analysis* **27**: 151–170.

- KARLSSON, A., 2009 Bootstrap methods for bias correction and confidence interval estimation for non-linear quantile regression of longitudinal data. *Journal of Statistical Computation and Simulation* **79**: 1205–1218.
- KOGAN, M., 1998 Integrated pest management: historical perspectives and contemporary developments. *Annual review of entomology* **43**: 243–270.
- KRISTENSEN, K., A. NIELSEN, C. W. BERG, H. SKAUG, and B. BELL, 2015 Tmb: automatic differentiation and laplace approximation. arXiv preprint arXiv:1509.00660 .
- LEE, Y. and J. A. NELDER, 1996 Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**: 619–656.
- LEE, Y., J. A. NELDER, and Y. PAWITAN, 2018 *Generalized linear models with random effects: unified analysis via H-likelihood*, volume 153. CRC Press.
- LEVINS, R., M. WILSON, *et al.*, 1980 Ecological theory and pest management. *Annual Review of Entomology* **25**: 287–308.
- MAGNUSSON, A., H. SKAUG, A. NIELSEN, C. BERG, K. KRISTENSEN, M. MAECHLER, K. VAN BENTHAM, B. BOLKER, and M. BROOKS, 2017 glmmTMB: generalized linear mixed models using template model builder. r package version 0.1. 0.
- MANLY, B. F., 2006 *Randomization, bootstrap and Monte Carlo methods in biology*. Chapman and Hall/CRC.
- MARTÍNEZ, E. Z. and F. LOUZADA NETO, 2001 Estimação intervalar via bootstrap. *Revista de Matemática e Estatística* **19**: 217–251.
- MCCULLOCH, C. E., 2003 Generalized linear mixed models. In *NSF-CBMS regional conference series in probability and statistics*, pp. i–84, JSTOR.
- MCCULLOCH, C. E. and S. R. SEARLE, 2001 Generalized, linear, and mixed models (wiley series in probability and statistics) .
- MILLAR, R. B., 2011 *Maximum likelihood estimation and inference: with examples in R, SAS and ADMB*, volume 111. John Wiley & Sons.
- MOLENBERGHS, G. and G. VERBEKE, 2005 Models for discrete longitudinal data. springer-verlag. New York .
- MOLENBERGHS, G. and G. VERBEKE, 2006 *Models for discrete longitudinal data*. Springer Science & Business Media.
- NELDER, J. A. and R. W. WEDDERBURN, 1972 Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135**: 370–384.
- NOH, M. and Y. LEE, 2007 Reml estimation for binary data in glmms. *Journal of Multivariate Analysis* **98**: 896–915.
- PATTERSON, H. D. and R. THOMPSON, 1971 Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**: 545–554.
- PIERCE, D. and B. SANDS, 1975 Extra-bernoulli variation in regression of binary data. Oregon State University, Department of Statistics Technical Report **46**.

- R CORE TEAM, 2018 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- SHAO, J. and D. TU, 1995 The jackknife and bootstrap. 1995.
- THAI, H.-T., F. MENTRÉ, N. H. HOLFORD, C. VEYRAT-FOLLET, and E. COMETS, 2013 A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical statistics* **12**: 129–140.
- THIES, C., S. HAENKE, C. SCHERBER, J. BENGTSSON, R. BOMMARCO, L. W. CLEMENT, P. CERYNGIER, C. DENNIS, M. EMMERSON, V. GAGIC, *et al.*, 2011 The relationship between agricultural intensification and biological control: experimental tests across europe. *Ecological Applications* **21**: 2187–2196.
- VAN EMDEN, H. and G. WILLIAMS, 1974 Insect stability and diversity in agro-ecosystems. *Annual Review of Entomology* **19**: 455–475.

4 A BIVARIATE FRAMEWORK TO JOINTLY MODEL COUNT AND CONTINUOUS RESPONSES

Abstract

Multivariate data are present in many studies in the natural sciences, including entomology (study of insects). Generally, the interest is in more than one response variable and in how factors and covariates affect them, simultaneously. However, when the interest is in one count variable and one continuous variable, we need to use different approaches to model these variables jointly. In this work, motivated by a data set from entomology, in which a continuous and a count response were observed simultaneously, a mean-dispersion bivariate model was developed. The proposed model is based on the bivariate normal distribution, and adds a new parameter to the variance-covariance matrix to better accommodate the main characteristics of count data, such as overdispersion or underdispersion. This modelling framework showed good results and also allowed for greater flexibility in the case of heterogeneity of variances, and allows for the correlation between the two responses to vary according to an experimental treatment.

Keywords: Count data; Entomology; Joint modelling; Multivariate data; Poisson model.

4.1 Introduction

In many observational or experimental studies, multiple response variables are of interest, and they may be correlated. In statistical modelling techniques, multiple response variables can be analyzed in isolation using univariate methods, where the dependence between them are not considered in the modelling. However, the correlation between the response variables can provide valuable insight, and their incorporation in the modelling framework can lead to more reliable tests and improved inference, especially when this correction is of high magnitude. Multivariate regression techniques facilitate the inclusion of correlated responses and are useful for exploring data patterns that may exist in more than one dimension (RAYKOV and MARCOULIDES, 2008; EVERITT and HOTHORN, 2011).

For the majority of univariate analysis, many inferences are based on the normal distribution, and once many multivariate techniques are extensions of univariate analysis, the majority of multivariate procedures have the multivariate normal distribution as their underpinning. For example, in bivariate analysis, the relationship between two continuous variables with unbounded symmetric distribution could be described by a joint probability distribution using the bivariate normal density (JOBSON, 2012). However, other distributions according to characteristics of the response variables can also be considered. The analysis of non-normal univariate or multivariate data involves, mostly, the generalized linear models (GLM) formulated by Nelder and Wedderburn (1972), that represent an elegant and encompassing mathematical framework to model response variables whose distribution belongs to the exponential family (such as normal, binomial, Poisson, gamma and inverse Gaussian). A feature of the exponential family of distributions is the *mean-variance* relationship, i.e., the fact that the variance is a function of the mean.

Johnson (1997) presented a range of discrete multivariate models that can jointly analyze count and binary data. Nevertheless, when it comes to joint modelling of one discrete and one continuous variable, suitable techniques and current software implementations are still emerging (BONAT and JØRGENSEN, 2016). Bonat et. al. (2018) proposed a flexible framework to model response variables of different nature jointly (e.g. unbounded and bounded continuous variables and discrete variables). They created the multivariate covariance generalized linear modelling framework (McGLM), implemented as the `mcglm`

package for R software (R CORE TEAM, 2020). In their approach, marginal models are fitted through quasi-likelihood functions based on first and second-moment assumptions (BONAT *et al.*, 2017; BONAT, 2018). The framework includes flexibility in the specification of the covariance structure and provides reliable tests for fixed effects. However, one drawback of this framework is that it does not provide an associated probability distribution. Another possibility for modelling discrete and continuous multivariate response variables jointly are copula models (NIKOLOULOPOULOS and KARLIS, 2009; KRUPSKII and JOE, 2013). The copula model allows us to incorporate the correlation between the variables and also allows flexibility when choosing the correlation structure. However, the use of different copulas generates completely different results. In this paper, we present another option for modelling count and continuous data jointly using an approximation based on the central limit theorem and a one-parameter extension of the variance-covariance structure.

When analyzing count data, the Poisson model is a natural first choice, however, the Poisson distribution can be approximated asymptotically by the normal distribution (MOOD, 1973). A multivariate extension based on the normal distribution is advantageous in the sense that this distribution presents known and closed formulas. A drawback for the Poisson model is the assumption of equality of mean and variance, which is not met very often in practice. This extra overdispersion in count data may be caused by a deficiency of relevant covariates or heterogeneity of samples, or repeated measures, and it is necessary to model this extra-variability to obtain reliable inference about the parameters (HINDE and DEMÉTRIO, 1998; VER HOEF and BOVENG, 2007). These causes of variation can also impact continuous data. However, the normal distribution, often used to fit to this data, assumes homoscedastic variance, which could not reflect the real characteristics of the data. Aitkin (1987) proposed the joint modelling of mean and dispersion in the normal regression analysis using a Fisher scoring algorithm for the simultaneous maximum likelihood estimation. This proposal brought a powerful approach to deal with the heterogeneity of variance, which was modeled rather than being transformed away. McCullagh and Nelder (1989) presented the joint modelling of mean and dispersion in the generalized linear models, using the extended quasi-likelihood function (NELDER and PREGIBON, 1987).

In this paper, we present a joint modelling of bivariate data, which also could be extended to multivariate data of more than two responses, to accommodate two counts, or two continuous response variables, or a combination of one count and one continuous variable using the bivariate normal approximation. To obtain a versatile framework to deal with count data, we included an extra parameter in the variance-covariance matrix to capture different characteristics present in this type of data, making it very useful also for continuous data. The normal bivariate model proposed allows us to model jointly mean and dispersion dealing with over- and underdispersion, and heteroscedastic variance. Furthermore, it is possible to fit the model considering different correlation structures that can change depending on a covariate of interest. To exemplify the method developed, a case study applied to the area of entomology is presented.

This paper is organized as follows. Section 4.2 presents the case-study that is a motivation for this paper. Section 4.3 presents the bivariate framework to jointly model count and continuous responses. Section 4.4 presents estimation and inference for the novel regression model based on the likelihood paradigm. The case study analysis and results are presented in Section 4.5. The properties of the maximum likelihood and profile likelihood estimators are assessed in Section 4.6 through simulation studies. Finally, we present discussion in Section 4.7.

4.2 Case study

The data comes from an experiment examining *Podisus nigrispinus* (Hemiptera: Pentatomidae) a predatory stinkbug, found in agricultural and forest systems in several countries of Central and South

America, when fed two different prey (*Anticarsia gemmatalis* and *Diatraea saccharalis*, two agricultural pests). This stinkbug has an important role as a control agent biological for different cultures, for example, in the forest area, the predator can be useful in the biological control of defoliating caterpillars in *Eucalyptus* plantations.

Given this importance, experiments aimed at assessing the dynamics of prey consumption by *P. nigrispinus* have been developed. In these studies, the response variables are the weight and the fertility of the female predator. The weight is one way to measure the development of an insect over its life cycle, and the fertility measures how well the insect can establish in an ecosystem. This information is essential for the choice of biological control strategies in pest management programs (PARRA, 2002).

The experiment was carried out at the Laboratory of Forest Ecology and Entomology, Department of Entomology and Acarology, ESALQ - USP, Piracicaba, Brazil. The insects were placed in individualized incubator chambers, where they were fed and monitored daily. The two diets or treatments consisted of offering three caterpillars of either *Anticarsia gemmatalis* or *Diatraea saccharalis* daily. The experiment started with 50 stinkbug insects for each diet, however, as there was a high mortality rate, the final measurements were made only with 9 stinkbug replicates of the *Anticarsia gemmatalis* treatment and 18 of the *Diatraea saccharalis* treatment. After 18 days of receiving the diet, the weight of females was measured. Next, the females were allocated with a male that also received the same diet. The formed couples were kept until the death of the females. All eggs laid by the females were placed in Petri dishes and counted. The correlation between the two outcomes (female weight and number of eggs) was examined by a scatterplot (Figure 4.1). This indicates that stinkbugs being fed the *Anticarsia gemmatalis* caterpillar diet performed more poorly overall than those fed the *Diatraea saccharalis* caterpillar diet. A strong correlation between female weight and number of eggs was suggested for the *Diatraea saccharalis* diet; this indicates that a modelling approach that allows the strength of the correlation between bivariate responses to change depending on an experimental treatment would be suitable here.

Figure 4.1: Scatterplot of *Podisus nigrispinus* female weight after 18 days and total number of eggs laid throughout the insect's life cycle by treatment (a diet consisting of *Anticarsia gemmatalis* or *Diatraea saccharalis*).

4.3 Joint Modelling of a Count and a Continuous Response

Bivariate models may be applied to analyze bivariate data jointly. However, when the outcomes have different distributions, the useful bivariate normal may not capture all information present in these data. In this paper, we propose a joint modelling approach, based on the multivariate normal distribution, that is suitable for jointly analyzing count and continuous bivariate data.

The marginal distribution structure is the same for either a count or a continuous variable. Both assume a normal distribution, with mean $\mu_i; i = 1, 2$ and standard deviation depending on the mean and a dispersion parameter ($\phi_i \mu_i; i = 1, 2$). It is also possible to capture the heterogeneity of variances, overdispersion, or underdispersion through the modelling of the ϕ parameters. We model the mean μ through a monotonic and differentiable link function $g(\mu) = \eta$, such that the linear predictor is $\eta = \mathbf{X}\beta$.

Let $\mathbf{Y}_1$ and $\mathbf{Y}_2$ be vectors of count or continuous random variables. We then assume that marginally $\mathbf{Y}_1 \sim N(\mu_1, \Sigma_1)$ and $\mathbf{Y}_2 \sim N(\mu_2, \Sigma_2)$, with the joint distribution of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ taken to be multivariate normal, i.e.,

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12} & \Sigma_2 \end{bmatrix} \right), \quad (4.1)$$

where Σ_1 denotes the covariance matrix for the first response; Σ_2 denotes the covariance matrix for the second response and Σ_{12} denotes the covariance matrix between the two responses, where Σ_{12} depends on a correlation parameter ρ , Σ_1 and Σ_2 .

We write $\Sigma_i = \text{diag}(\mathbf{X}_\phi \phi) \text{diag}(g^{-1}(\mathbf{X}_\beta \beta)) = \mathbf{D}_\phi \mathbf{D}_\mu$, where the function $\text{diag}(\mathbf{x})$ is defined as a square matrix with the elements of the vector $\mathbf{x}$ in its main diagonal; $\mathbf{X}_\phi$ and $\mathbf{X}_\beta$ are the design matrices for dispersion and regression parameters respectively; and ϕ and β unknown parameter vectors. We may write for μ_{ij} and ϕ_{ij}

$$\Sigma_i = \begin{pmatrix} \phi_{i1}\mu_{i1} & 0 & \dots & 0 \\ 0 & \phi_{i2}\mu_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \phi_{in}\mu_{in} \end{pmatrix},$$

where i represents the response number ($i = 1, 2$); j is the observation ($j = 1, \dots, n$) and the variance-covariance matrix (Σ_i) will involve the mean and dispersion parameters.

4.4 Estimation and Inference

4.4.1 Univariate modelling

Taking one of the response outcomes $\mathbf{Y}_i$, from here we will omit the index i to improve readability. We assume that asymptotically $\mathbf{Y} \sim N(\mu, \Sigma = \mathbf{D}_\phi \mathbf{D}_\mu)$, and write the log-likelihood as

$$\log \mathcal{L}(\mu, \Sigma | \mathbf{y}) = l(\mu, \Sigma | \mathbf{y}) = -\frac{1}{2} \{ \ln |\Sigma| + (\mathbf{y} - \mu)' \Sigma^{-1} (\mathbf{y} - \mu) \} + \text{const}. \quad (4.2)$$

The estimating equations can be obtained using the chain rule. Differentiating (4.2) w.r.t. Σ yields

$$\frac{\partial l}{\partial \Sigma} = -\frac{1}{2}(\Sigma^{-1})^T + \frac{1}{2}(\Sigma^{-1})^T(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'(\Sigma^{-1})^T. \quad (4.3)$$

Now further differentiating eq. (4.3) w.r.t. $\mathbf{D}_\phi$ and w.r.t. ϕ yields (see Appendix A for calculations):

$$-\frac{1}{2}\text{diag}(\mathbf{X}_\phi\phi)^{-1} \circ \mathbb{I} + \frac{1}{2}\text{diag}(\mathbf{X}_\phi\phi)^{-1}\mathbf{D}_\mu^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'\text{diag}(\mathbf{X}_\phi\phi)^{-1} \circ \mathbb{I}. \quad (4.4)$$

Finally, solving eq. (4.4) w.r.t. ϕ yields (see Appendix A for calculations):

$$\hat{\phi} = (\mathbf{X}'_\phi\mathbf{X}_\phi)^{-1}\mathbf{X}'_\phi\{(\mathbf{D}_\mu^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \circ \mathbb{I})\mathbb{J}\}, \quad (4.5)$$

where $\mathbb{I}$ is an $n \times n$ identity matrix and $\mathbb{J}$ is an $n \times 1$ vector of 1s, and $\circ$ the Hadamard product. In this case, for positive values, the dispersion parameter (ϕ) is always greater than zero. If $\phi < 1$ there is underdispersion and $\phi > 1$ overdispersion.

It is not as straightforward to obtain the maximum likelihood estimator for $\boldsymbol{\beta}$, because $\boldsymbol{\mu}$ is also a function of Σ in the likelihood. Therefore we firstly approximate the estimate of the regression coefficients using the known estimator for the regression parameters in the linear models and then update $\hat{\boldsymbol{\beta}}$ using $\hat{\boldsymbol{\mu}}$ estimated in the preceding stage. The maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, conditional on $\boldsymbol{\mu}, \phi, \rho$, with $\Sigma(\boldsymbol{\mu}, \phi, \rho)$ is given by (LAIRD and WARE, 1982; MOLENBERGHS and VERBEKE, 2000)

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\mu}, \phi, \rho) = (\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Sigma^{-1}\mathbf{Y}.$$

Because we are including different link functions, not limited to identity, we included a matrix of weights $\mathbf{W}$ in the estimation of $\boldsymbol{\beta}$, following the estimation procedure of the GLM framework. The estimation for the vector $\boldsymbol{\beta}$ is given considering

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}), \quad \mathbf{W} = \frac{1}{\mathbf{V}(\boldsymbol{\mu})[g'(\boldsymbol{\mu})]^2} \quad \text{and} \quad \mathbf{z} = \boldsymbol{\eta} + \mathbf{G}(\mathbf{y} - \boldsymbol{\mu}), \quad (4.6)$$

where $\mathbf{W}$ is a diagonal matrix of weights, $\mathbf{V}(\boldsymbol{\mu}) = \mathbf{D}_\phi\mathbf{D}_\mu$ and $\mathbf{G} = \text{diag}\{g'(\mu_1), \dots, g'(\mu_n)\}$ (NELDER and WEDDERBURN, 1972; MCCULLAGH and NELDER, 1989).

Using $\boldsymbol{\eta}^{(1)} = g(\mathbf{y})$ for the first iteration, then

$$\hat{\boldsymbol{\beta}}^{(1)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\eta}^{(1)} \quad \text{and} \quad \hat{\boldsymbol{\beta}}^{(n)} = (\mathbf{X}^T\mathbf{W}^{(n-1)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(n-1)}\mathbf{z}^{(n-1)}. \quad (4.7)$$

This algorithm yields the same results as numerically maximising the likelihood function using, e.g. the BFGS algorithm, however it is much faster in terms of computational burden. The good behavior of these estimators can also be verified in the simulations studies in Section 4.6.

4.4.2 Bivariate modelling

To model both outcomes jointly it is necessary to estimate the matrix Σ_{12} (eq. 4.1). Martinez and Benedito (2013) introduced a method to construct the multivariate covariance matrix using Kronecker products. Bonat and Jørgensen (2016) applied the same method to develop the technique called Multivariate Covariance Generalized Linear Models. Let $\mathbf{Y}_{nR \times 1} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_R\}$, with n = number of observations and R = number of outcomes; Σ_i the $n \times n$ covariance matrix within the outcome i for $i = 1, \dots, R$. Let Σ_b be the $R \times R$ correlation matrix between outcomes, with the diagonal elements equal to 1 and off-diagonal elements equal to ρ . Then

$$\mathbf{E}(\mathbf{Y}) = \{g^{-1}(\mathbf{X}_1\boldsymbol{\beta}_1), \dots, g^{-1}(\mathbf{X}_R\boldsymbol{\beta}_R)\}$$

$$\text{Var}(\mathbf{Y}) = \mathbf{\Sigma} = \text{Bdiag}(\tilde{\mathbf{\Sigma}}_1, \dots, \tilde{\mathbf{\Sigma}}_R)(\mathbf{\Sigma}_b \otimes \mathbf{I})\text{Bdiag}(\tilde{\mathbf{\Sigma}}_1^T, \dots, \tilde{\mathbf{\Sigma}}_R^T)$$

where the matrix $\tilde{\mathbf{\Sigma}}_i$ denotes the lower triangular matrix of the Cholesky decomposition of $\mathbf{\Sigma}_i$. The operator Bdiag denotes a block diagonal matrix and $\mathbf{I}$ denotes an $n \times n$ identity matrix. Then, the covariance matrix, could be defined as

$$\mathbf{\Sigma}_{12} = \rho \tilde{\mathbf{\Sigma}}_i \tilde{\mathbf{\Sigma}}_{i'}^T$$

with $i \neq i'$.

Here we propose an extension to allow more than one correlation coefficient, for example, we could have one for each treatment in our case study data (Section 4.2). This requires increasing the flexibility of the $\mathbf{\Sigma}_b$ matrix to include more than one correlation coefficient. Let $\mathbf{X}_\rho$ denote a $N \times t$ design matrix, with t the to number of treatment levels, and ρ a vector for the correlation coefficients with $t \times 1$ dimension, then

$$\mathbf{\Sigma}_\rho = \text{diag}(\mathbf{X}_\rho \rho)$$

and

$$\mathbf{\Sigma} = \text{Bdiag}(\tilde{\mathbf{\Sigma}}_1, \tilde{\mathbf{\Sigma}}_2) \begin{bmatrix} \mathbf{I} & \mathbf{\Sigma}_\rho \\ \mathbf{\Sigma}_\rho & \mathbf{I} \end{bmatrix} \text{Bdiag}(\tilde{\mathbf{\Sigma}}_1^T, \tilde{\mathbf{\Sigma}}_2^T).$$

Hence, the proposed model is considering $R = 2$ outcomes. Then, for $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2]$, the full log likelihood is

$$l(\boldsymbol{\mu}, \mathbf{\Sigma}(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{\rho}) | \mathbf{y}) = -\frac{1}{2} \{ \ln |\mathbf{\Sigma}| + (\mathbf{y} - \boldsymbol{\mu})' \mathbf{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \} + \text{const.} \quad (4.8)$$

Here the covariance matrix is more flexible, allowing the modelling of data sets with different correlation profiles. A likelihood ratio test can be used to test if correlations coefficients differ, for example by levels of a treatment in an experiment.

Jørgensen and Knudsen (2004) suggested a method to jointly estimate all parameters of the $\mathbf{\Sigma}$ matrix using the Newton scoring algorithm, based on quasi-likelihood and the Pearson estimating functions, applying second-moment assumptions. As an alternative approach, here, the matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are estimated first separately, using the maximum likelihood method, presented in section (4.4.1). These estimates are then used as initial values to obtain the estimates $\hat{\boldsymbol{\rho}}$ by profiling the log-likelihood function w.r.t. $\boldsymbol{\rho}$ and maximizing (4.8) numerically. At each step, the estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$ are updated, using the estimation algorithm described in Section 4.4.1. Since the derivatives of l w.r.t. $\boldsymbol{\rho}$ cannot be obtained in closed form, the L-BFGS-B algorithm (ZHU *et al.*, 1997) was implemented in the `mle2()` function for R software (R CORE TEAM, 2020). Standard errors for the regression parameters are obtained based on the observed information matrix $\mathbf{I}(\boldsymbol{\theta})$, where $\mathbf{I}(\boldsymbol{\theta}) = -\mathbf{H}(\boldsymbol{\theta})$ (hessian matrix) is computed numerically, using the `hessian()` function of the package `numDeriv` (GILBERT *et al.*, 2006). Since $\hat{\boldsymbol{\phi}}$ depends on $\hat{\boldsymbol{\beta}}$, the standard errors for the dispersion parameters were obtained using a two-stage algorithm, considering $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\rho}}$ obtained from the previous stages.

We applied these methods to the case study data described in Section 4.2. Diagnostic analyses and goodness-of-fit assessment for the models fitted to the case study data set were carried out by producing bivariate residual plots with simulation polygons (MORAL *et al.*, 2020). Likelihood ratio tests were used to compare treatments and to assess the significance of the parameters.

4.5 Case study analysis

We fitted the bivariate model to the data set described in Section 4.2, including the effects of diet (treatment) in the linear predictors for the mean, dispersion and correlation. Table 4.1 shows the estimated parameters and standard errors. The females fed with *D. saccharalis* emerged with greater weight (β_{22}) and had greater fertility, producing more eggs (β_{12}). The estimates for ϕ_{11} and ϕ_{12} indicate overdispersion for the number of eggs, however since their standard errors are large we tested for their equality. The estimates for ϕ_{21} and ϕ_{22} are also similar.

Table 4.1: Parameter estimates (EST.) and standard errors (SE) for the normal bivariate model. β_{11} and β_{12} represent the mean number of eggs produced by stinkbugs on for diets *Anticarsia* and *Diatraea* respectively; β_{21} and β_{22} represent the mean weight for stinkbugs on diets *Anticarsia* and *Diatraea* respectively. ϕ_{11} and ϕ_{12} represent the dispersion of the number of eggs for diets *Anticarsia* and *Diatraea* respectively; ϕ_{21} and ϕ_{22} represent the dispersion of weight for diets *Anticarsia* and *Diatraea* respectively. ρ_1 represents the correlation between number of eggs and weight for *Anticarsia* diet and the parameter ρ_2 represents the correlation between number of eggs and weight for *Diatraea*.

	EST.	SE
β_{11}	4.0561	0.2907
β_{12}	5.3782	0.1609
β_{21}	62.5000	4.0684
β_{22}	93.9474	3.3824
ϕ_{11}	39.0292	19.5049
ϕ_{12}	106.6240	32.1814
ϕ_{21}	2.1186	1.0588
ϕ_{22}	2.3138	0.6983
ρ_1	-0.0389	0.3528
ρ_2	0.5192	0.1487
LogLik	-221.0125	

The bivariate residual plot with simulation polygons with 99 simulations, showed 3 out of 27 points out of their respective polygons, which indicates a reasonably good fit (Figure 4.2).

The selection of the model was performed using likelihood ratio tests. We verify the treatment effect on the mean and the dispersion by each outcome and jointly (Table 4.2 and 4.3). The effect of treatment on the correlation was also tested and presented no significance, but here due to the differences in correlations observed per treatment (Figure 4.1), the biological assumption, and the small sample size, we chose to maintain a different correlation effect per treatment. The test gave evidence for equal dispersion coefficients, but different regression coefficients for each diet. The regression parameters showed a difference between treatments for both outcomes, hence, the diet *D. saccharalis* contributed to a better development of the predator.

4.6 Simulation studies

Simulation studies were performed to assess the properties of the maximum likelihood estimators and the flexibility of the bivariate framework. The scenarios proposed here compare the performance of the proposed model with different sample size; under, over and equidispersion for the count outcome; heterogeneity of variances for the continuous outcome and negative, positive, low and high correlation between outcomes.

Figure 4.2: Bivariate plot with simulation polygons assuming a bivariate normal distribution for bivariate samples of size 99 simulated, with estimated marginal densities for residuals 1 and residuals 2. The an lighter shade corresponds to the estimated density of all simulated diagnostics, and the darker one corresponds to the estimated density of the observed diagnostics. The points outside their simulated polygons are displayed in red.

Table 4.2: Model fit measurements and comparisons between the complete and reduced models to test the treatment effect on the mean and on the dispersion for each outcome.

Model	np	l	$2(\text{diff } l)$	$P(> \chi^2)$
Full model	10	-221.0125		
$\beta_{11} = \beta_{12}$	9	-227.4322	12.8394	0.0003
$\beta_{21} = \beta_{22}$	9	-236.2779	16.4631	0.0000
$\phi_{10} = \phi_{11}$	9	-222.2381	2.4511	0.2936
$\phi_{20} = \phi_{21}$	9	-221.0235	0.02209	0.9890

¹np, number of parameters; l , log-likelihood; diff l , difference in log-likelihoods.

Table 4.3: Model fit measurements and comparisons between the complete and reduced models to test the treatment effect on the mean and on the dispersion for both outcomes jointly.

Model	np	l	$2(\text{diff } l)$	$P(> \chi^2)$
Full model	10	-221.0125		
$\beta_{11} = \beta_{12}$ and $\beta_{21} = \beta_{22}$	8	-232.1660	22.3070	0.0000
$\phi_{10} = \phi_{11}$ and $\phi_{20} = \phi_{21}$	8	-222.2347	2.4445	0.2946

²np, number of parameters; l , log-likelihood; diff l , difference in log-likelihoods.

We conducted the simulation study based on the experimental structure described in the case study (Section 4.2). Four simulation scenarios were designed considering one treatment factor with two

levels and the numbers of experimental units equal to half of the sample size. The count outcome was generated from a normal distribution with a logarithmic link function, fixing the regression coefficients at the values, $\beta_{11} = 2$ and $\beta_{12} = 3$ (the first index represents the response variable and the second index the treatment level). The dispersion parameters were fixed at the values $\phi_{11} = 0.5$ and $\phi_{12} = 2$. For the second continuous outcome, we considered $\beta_{21} = 10$, $\beta_{22} = 300$, $\phi_{21} = 1$, $\phi_{22} = 10$ and an identity link function. We produced simulations based on four different values for a single correlation parameter ($\rho = -0.8, 0.2, 0.5$ and 0.8). These scenarios present negative, weak, moderate, and strong positive correlation. To check the consistency of the estimators, four sample sizes were considered: 50, 100, 300, and 600. We generated 1000 data sets for each simulation scenario. The simulation process is summarised as follows

1. Define $\beta_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$; $\beta_2 = \begin{bmatrix} 10 \\ 300 \end{bmatrix}$; $\phi_1 = \begin{bmatrix} 0.5 \\ 2 \end{bmatrix}$; $\phi_2 = \begin{bmatrix} 1 \\ 10 \end{bmatrix}$;
2. Set up $\mathbf{X}_{\beta_1}$ and $\mathbf{X}_{\beta_2}$ as $n \times 2$ design matrices;
3. Set up $\mathbf{X}_{\phi_1}$ and $\mathbf{X}_{\phi_2}$ as $n \times 2$ dispersion design matrices;
4. Calculate $\mu_1 = g^{-1}(\mathbf{X}_{\beta_1}\beta_1) = \exp(\mathbf{X}_{\beta_1}\beta_1)$ and $\mu_2 = g^{-1}(\mathbf{X}_{\beta_2}\beta_2) = \mathbf{X}_{\beta_2}\beta_2$;
5. Obtain $\Sigma_1 = \text{diag}(\mathbf{X}_{\phi_1}\phi_1)\text{diag}(\mu_1)$ and $\Sigma_2 = \text{diag}(\mathbf{X}_{\phi_2}\phi_2)\text{diag}(\mu_2)$;
6. Set up $\mathbf{X}_\rho$ as a $n \times 1$ unit matrix;
7. Obtain $\Sigma_\rho = \text{diag}(\mathbf{X}_\rho\rho)$, where ρ is a scalar ($-0.8, 0.2, 0.5$ or 0.8);
8. Calculate $\Sigma = \text{Bdiag}(\Sigma_1, \Sigma_2) \begin{bmatrix} \mathbf{I} & \Sigma_\rho \\ \Sigma_\rho & \mathbf{I} \end{bmatrix} \text{Bdiag}(\Sigma_1^T, \Sigma_2^T)$;
9. Simulate from $\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma \right)$;
10. Fit the proposed model to the simulated data sets.

This process was followed for each simulation scenario. Further, we also simulated a scenario with two correlations parameters, one for each treatment. The study considered 1000 simulations, sample sizes equal to 50, 100 and 300, and four scenarios combining correlation coefficients ($\rho_1 = 0.2$ and $\rho_2 = 0.5$; $\rho_1 = 0.2$ and $\rho_2 = 0.8$; $\rho_1 = -0.8$ and $\rho_2 = 0.2$; $\rho_1 = -0.2$ and $\rho_2 = 0.5$).

The complete code used for the simulations and model fitting can be found in the supplementary materials.

4.6.1 Simulation results

To visualize estimation bias for each simulation scenario Figure 4.3 was constructed. It presents the average bias, defined as the average estimate of 1000 simulations minus the true value and their respective confidence interval calculated as the average bias plus and minus 1.96 times the standard error.

Figure 4.3: Estimation average bias of each simulation scenario (symbols) with their respective confidence interval of 95% obtained by the average standard error considering 1000 simulations.

The results of the simulation studies showed that for all correlation levels, the average bias and the standard errors tend to 0 as the sample size increases, meaning that all parameters have unbiased and consistent estimators. For small samples, the dispersion parameters are underestimated. This is also observed for the regression parameters when the correlation is very high ($\rho = 0.8$). For the correlation parameters, the bias is low even for small samples, but it is smaller when the sample size increases. The average standard error is lower for strong correlations, positive or negative ($\rho = -0.8$ and $\rho = 0.8$); this is clear for the correlation coefficient, but it is also possible to observe a slightly smaller average standard error for the dispersion coefficients. Figure 4.4 shows the empirical coverage rate of the asymptotic confidence intervals. The results showed that for the regression parameters the empirical coverage rates are close to the nominal level of 95% for all sample sizes and all simulation scenarios. For the dispersion and correlation parameters, the empirical coverage rates are lower than the nominal level, however, for $\rho = 0.2$ and $\rho = 0.5$, they become closer to the nominal level for large samples. The worst scenario is with a small sample and strong correlation, in this case, even for large samples the coverage rate is still lower than the nominal level. An alternative here is to consider other methods to construct the confidence intervals such as the bootstrap approach.

Figure 4.4: Coverage rate based on normal approximation confidence intervals with a nominal rate of 95% for different sample sizes (50, 100, 300, and 600) and correlation coefficients (-0.8, 0.2, 0.5 and 0.8).

To verify the coverage rate considering bootstrap confidence intervals, a parametric bootstrap simulation was performed, considering 1000 bootstrap samples and 100 simulations, respecting the same previous scenarios with a sample size equal to 50 and correlation coefficients equal to -0.8 and 0.8 . These simulations showed some asymmetry and bias, mainly for the dispersion and correlation coefficients, therefore bias-corrected percentile confidence intervals (BCP) were considered. Efron (1982) presented this method, in which the correction relies upon transforming the distribution to one that satisfies the symmetry condition (DICICCIO and ROMANO, 1988; BUCKLAND, 1984). Using this method, the empirical

coverage rates increased, becoming very close to the 95% nominal coverage (Table 4.4).

Table 4.4: Coverage rate based on bias-corrected percentile confidence intervals with 1000 bootstrap samples and 100 simulations, $n = 50$ and a nominal rate of 95%.

Parameter	Coverage rate	
	$\rho = 0.8$	$\rho = -0.8$
β_{11}	0.95	0.96
β_{12}	0.95	0.93
β_{21}	0.98	0.96
β_{22}	0.98	0.92
ϕ_{11}	0.94	0.95
ϕ_{12}	0.92	0.88
ϕ_{21}	0.93	0.95
ϕ_{22}	0.91	0.89
ρ	0.95	0.96

Table 4.5 shows the results of the simulation study considering two correlations coefficients, in this example, one for each treatment. The observed average bias was low, and the larger the sample size, the lower the average bias and associated standard errors. We observed that for stronger correlations, the standard error was lower, implying empirical confidence intervals with smaller coverage than the nominal rate. Further, it is clear that the scenarios that presented better coverage, mainly for larger samples, were scenarios 1 and 4.

Table 4.5: Estimator mean bias, MSE and coverage rate for the proposed model under 4 scenarios, which varied true parameter value. Results are based on 1000 simulations.

Scenario 1:										
Par.	True	Bias			MSE			Coverage		
		n=50	n=100	n=300	n=50	n=100	n=300	n=50	n=100	n=300
β_{11}	2	0.000	-0.004	0.000	0.050	0.036	0.021	0.94	0.94	0.94
β_{12}	3	-0.002	-0.001	0.000	0.061	0.044	0.026	0.94	0.93	0.95
β_{21}	10	-0.012	-0.018	-0.017	0.611	0.441	0.257	0.93	0.94	0.95
β_{22}	300	0.154	-0.059	-0.024	10.584	7.669	4.446	0.93	0.95	0.96
ϕ_{11}	0.5	-0.020	-0.013	-0.004	0.133	0.096	0.057	0.88	0.90	0.92
ϕ_{12}	2	-0.093	-0.054	-0.021	0.501	0.362	0.213	0.87	0.90	0.91
ϕ_{21}	1	-0.044	-0.015	-0.003	0.265	0.194	0.114	0.87	0.90	0.93
ϕ_{22}	10	-0.469	-0.080	0.083	2.501	1.844	1.069	0.87	0.90	0.91
ρ_1	0.2	-0.005	-0.004	-0.001	0.179	0.130	0.076	0.88	0.92	0.93
ρ_2	0.5	-0.013	-0.001	0.000	0.133	0.094	0.055	0.88	0.89	0.91
Scenario 2:										
Par.	True	Bias			MSE			Coverage		
		n=50	n=100	n=300	n=50	n=100	n=300	n=50	n=100	n=300
β_{11}	2	-0.002	-0.001	0.001	0.051	0.036	0.021	0.91	0.95	0.94
β_{12}	3	-0.002	0.001	0.001	0.062	0.044	0.026	0.92	0.94	0.94
β_{21}	10	0.021	0.006	-0.004	0.614	0.440	0.256	0.94	0.94	0.94
β_{22}	300	-0.093	0.207	0.078	10.682	7.599	4.435	0.92	0.94	0.95
ϕ_{11}	0.5	-0.015	-0.014	-0.005	0.134	0.096	0.057	0.89	0.89	0.93
ϕ_{12}	2	-0.059	-0.048	-0.031	0.450	0.321	0.187	0.86	0.87	0.87
ϕ_{21}	1	-0.036	-0.021	-0.012	0.267	0.193	0.113	0.89	0.91	0.93
ϕ_{22}	10	-0.283	-0.281	-0.130	2.252	1.597	0.939	0.86	0.86	0.88
ρ_1	0.2	-0.011	0.003	0.000	0.180	0.129	0.076	0.90	0.91	0.95
ρ_2	0.8	-0.006	-0.002	-0.002	0.058	0.040	0.023	0.83	0.86	0.87
Scenario 3:										
Par.	True	Bias			MSE			Coverage		
		n=50	n=100	n=300	n=50	n=100	n=300	n=50	n=100	n=300
β_{11}	2	-0.001	0.000	0.000	0.051	0.036	0.021	0.94	0.93	0.93
β_{12}	3	0.001	0.001	0.001	0.061	0.044	0.026	0.92	0.94	0.95
β_{21}	10	-0.013	-0.025	-0.007	0.615	0.442	0.257	0.92	0.93	0.95
β_{22}	300	-0.149	-0.321	0.245	10.699	7.599	4.444	0.94	0.93	0.94
ϕ_{11}	0.5	-0.014	-0.014	-0.001	0.112	0.080	0.047	0.83	0.85	0.89
ϕ_{12}	2	-0.109	-0.050	-0.029	0.524	0.384	0.225	0.86	0.91	0.93
ϕ_{21}	1	-0.030	-0.010	-0.002	0.225	0.163	0.095	0.84	0.86	0.89
ϕ_{22}	10	-0.239	-0.268	-0.100	2.701	1.918	1.130	0.88	0.91	0.94
ρ_1	-0.8	0.005	0.007	0.002	0.057	0.041	0.023	0.85	0.84	0.85
ρ_2	0.2	0.004	-0.003	-0.004	0.178	0.130	0.076	0.88	0.93	0.94
Scenario 4:										
Par.	True	Bias			MSE			Coverage		
		n=50	n=100	n=300	n=50	n=100	n=300	n=50	n=100	n=300
β_{11}	2	-0.003	0.001	-0.001	0.051	0.036	0.021	0.94	0.94	0.94
β_{12}	3	-0.001	0.002	0.001	0.061	0.044	0.025	0.94	0.94	0.95
β_{21}	10	-0.041	-0.013	-0.002	0.612	0.441	0.256	0.93	0.96	0.94
β_{22}	300	-0.137	-0.179	0.361	10.583	7.597	4.450	0.94	0.95	0.95
ϕ_{11}	0.5	-0.017	-0.010	-0.004	0.134	0.097	0.057	0.88	0.91	0.92
ϕ_{12}	2	-0.085	-0.062	-0.035	0.503	0.361	0.212	0.88	0.90	0.91
ϕ_{21}	1	-0.035	-0.014	-0.009	0.267	0.194	0.113	0.88	0.92	0.91
ϕ_{22}	10	-0.448	-0.264	-0.071	2.506	1.815	1.072	0.86	0.88	0.91
ρ_1	-0.2	0.004	0.000	0.002	0.179	0.130	0.076	0.88	0.94	0.93
ρ_2	0.5	-0.015	-0.008	-0.005	0.134	0.095	0.055	0.88	0.90	0.90

4.7 Discussion

In this work, we proposed a bivariate framework to jointly model count and continuous responses. Our approach represents an attractive and flexible framework for studying two variables jointly and is widely applicable. This alternative considered a normal approach, modelling the variance-covariance matrix so that the characteristics of the data, such as over- or underdispersion for count data, and heterogeneity of variances for continuous data can be accommodated. Since this method is based on the asymptotic approximation of a Poisson distribution to the normal distribution, it performed better with larger samples. The advantage of this methodology is that it is possible to incorporate correlation, while jointly modeling the response variables and assuming an underlying probability distribution. One drawback of our approach is that issues may arise for small sample sizes, including biased dispersion estimates and a true coverage rate of the confidence intervals for the parameters smaller than the nominal rate. So it is necessary to carefully analyze small samples, doing descriptive analyzes before fitting the model, and checking whether the obtained results have a practical interpretation.

We observed in the simulation studies that even for large samples, when $\rho = -0.8$ and $\rho = 0.8$, the coverage rate of the correlation parameters is close to 90% instead of 95%. This is possibly because the true correlation value is close to the boundary of the parametric space. When $|\rho| \approx 1$, the range limits are automatically limited by 1 or -1 , then the variation of the parameter space is smaller. This became more evident by observing Figure 4.3, which showed that the standard errors for the correlation coefficients are lower for stronger correlations. This is not reflected in the regression coefficients, due to the fact that $\hat{\beta}$ varies very little for different values of ϕ and ρ . However, the coverage rate for the dispersion parameters also depends on the correlation parameter value, and the stronger the correlation, the lower the coverage rate. These cited problems are expected when it comes to estimates at the limit of the parametric space, especially with small samples. Despite these limitations, we concluded that the model generally behaves as expected, with unbiased, consistent, and accurate estimates when sample size are large enough.

Several extensions to the proposed bivariate model are possible (and currently in progress). These extensions include modeling correlation between observations taken on the same experimental unit, e.g., longitudinal data or repeated measurements. In this case, the variance-covariance matrix needs to be extended, allowing to incorporate the dependent measures taken on the same subject. This extension can be carried out by adding to the current $\Sigma = \mathbf{D}_\phi \mathbf{D}_\mu$ the term $\mathbf{ZGZ}'$, where $\mathbf{Z}$ is the matrix of known covariates indicating the correlated observations or how the response evolves over time for each experimental unit subject and $\mathbf{G}$ a general covariance matrix. However, it is also necessary to include a link function for the dispersion parameter, so that it does not have a negative maximum likelihood estimate, and this is subject of ongoing work.

References

- BONAT, W., J. OLIVERO, M. GRANDE-VEGA, M. FARFÁN, and J. FA, 2017 Modelling the covariance structure in marginal multivariate count models: hunting in bioko island. *Journal of Agricultural, Biological and Environmental Statistics* **22**: 446–464.
- BONAT, W. H., 2018 Multiple response variables regression models in r: The mcglm package. *Journal of Statistical Software* **84**.
- BONAT, W. H. and B. JØRGENSEN, 2016 Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**: 649–675.

- BRESLOW, N. E. and D. G. CLAYTON, 1993 Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**: 9–25.
- BUCKLAND, S. T., 1984 Monte carlo confidence intervals. *Biometrics* pp. 811–817.
- DICICCIO, T. J. and J. P. ROMANO, 1988 A review of bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Methodological)* **50**: 338–354.
- EVERITT, B. and T. HOTHORN, 2011 *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.
- FATORETTO, M. B., R. D. A. MORAL, C. G. B. DEMÉTRIO, C. S. DE PÁDUA, V. MENARIN, V. M. A. ROJAS, C. P. D’ALESSANDRO, and I. DELALIBERA, 2018 Overdispersed fungus germination data: statistical analysis using r. *Biocontrol Science and Technology* pp. 1–20.
- GILBERT, P., M. P. GILBERT, and R. VARADHAN, 2006 The numderiv package.
- HINDE, J. and C. G. DEMÉTRIO, 1998 Overdispersion: models and estimation. *Computational Statistics and Data Analysis* **27**: 151–170.
- JOBSON, J. D., 2012 *Applied multivariate data analysis: regression and experimental design*. Springer Science & Business Media.
- JØRGENSEN, B. and S. J. KNUDSEN, 2004 Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics* **31**: 93–114.
- KRUPSKII, P. and H. JOE, 2013 Factor copula models for multivariate data. *Journal of Multivariate Analysis* **120**: 85–101.
- LAIRD, N. M. and J. H. WARE, 1982 Random-effects models for longitudinal data. *Biometrics* pp. 963–974.
- MARTINEZ-BENEITO, M. A., 2013 A general modelling framework for multivariate disease mapping. *Biometrika* **100**: 539–553.
- MCCULLAGH, P. and J. NELDER, 1989 *Generalized linear models*. Chapman and Hall.
- MOLENBERGHS, G. and G. VERBEKE, 2000 *Linear mixed models for longitudinal data*. Springer.
- MOOD, A. M., 1973 *Introduction to the Theory of Statistics.*, volume 3. McGraw-hill.
- MORAL, R. A., J. HINDE, and C. G. DEMÉTRIO, 2020 Bivariate residual plots with simulation polygons. *Journal of Computational and Graphical Statistics* **29**: 203–214.
- NELDER, J. A. and D. PREGIBON, 1987 An extended quasi-likelihood function. *Biometrika* **74**: 221–232.
- NELDER, J. A. and R. W. WEDDERBURN, 1972 Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135**: 370–384.
- NIKOLOULOPOULOS, A. K. and D. KARLIS, 2009 Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation* **39**: 172–187.
- PARRA, J. R. P., 2002 *Controle biológico no Brasil: parasitóides e predadores*. Editora Manole Ltda.
- PINHEIRO, J. and D. BATES, 2006 *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

- RAYKOV, T. and G. A. MARCOULIDES, 2008 *An introduction to applied multivariate analysis*. Routledge.
- TEAM, R. C., 2020 R: A language and environment for statistical computing. r foundation for statistical computing.
- VER HOEF, J. M. and P. L. BOVENG, 2007 Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data. *Ecology* **88**: 2766–2772.
- ZHU, C., R. H. BYRD, P. LU, and J. NOCEDAL, 1997 Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)* **23**: 550–560.

5 FINAL CONSIDERATIONS

In this thesis, we developed and extended statistical models and computational tools to help analysis of natural science data, in which the data sets here, are focused on entomological studies.

The first paper presented in Chapter 2 was published in *Biocontrol Science and Technology*, an applied journal. The paper works as a tutorial to help researches to analyze overdispersed data, a common characteristic in this area. This work being developed during the Masters's degree and was finished throughout the doctorate. Further, we proposed an extension of this work presented in Chapter 3, using the nonparametric bootstrap and different confidence intervals. This work is still in development, however, it will be submitted to the *Computational Statistics & Data Analysis* journal, also aiming to present an application of statistical tools in this applied area.

The new bivariate framework presented in Chapter 4 is the major contribution of this thesis and the production of future studies. In this work we presented a bivariate analysis for a count and continuous data, modeling the mean and dispersion, allowing a range of applications to analyze bivariate data. The paper will be submitted for publication in the *Journal of the Internacional Biometrics Society*. For future works we will develop a package in the R software, thus enabling the employability of the entire methodology developed. The package will be developed after the extension of the methodology developed for repeated measurement data.

Improvements in papers developed here are still the subject of ongoing research, and we appreciate any suggestions and comments. All data sets and programs used will be made available in the supplementary materials of the submitted journals. And, we would be happy to share code upon requests made to the author¹.

¹ mairafaretto@gmail.com

APPENDIX

Appendix A: Estimation and Inference for Chapter 4, Section 4.3

Univariate modelling

Taking one of the response outcomes $\mathbf{Y}_i$, from here we will omit the index i to improve readability. Asymptotically $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{D}_\phi \mathbf{D}_\mu)$, we can write the log-likelihood as

$$\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}) = l(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}) = -\frac{1}{2} \{\ln |\boldsymbol{\Sigma}| + (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\} + \text{const.} \quad (5.1)$$

Differentiating with respect to $\boldsymbol{\Sigma}$ is straightforward:

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}} = \frac{\partial l}{\partial \boldsymbol{\Sigma}} \left\{ -\frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\} \quad (5.2)$$

as $\frac{\partial}{\partial A} \ln |A| = A^{-T}$ and $\frac{\partial \mathbf{a}^T \mathbf{X}^{-1} \mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-T} \mathbf{a} \mathbf{b}^T \mathbf{X}^{-T}$. Then,

$$\frac{\partial l}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2} \boldsymbol{\Sigma}^{-T} + \frac{1}{2} \boldsymbol{\Sigma}^{-T} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-T} \quad (5.3)$$

and

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\Sigma}} \frac{\partial \boldsymbol{\Sigma}}{\partial \mathbf{D}_\phi} &= \left[-\frac{1}{2} \boldsymbol{\Sigma}^{-T} + \frac{1}{2} \boldsymbol{\Sigma}^{-T} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-T} \right] \circ \mathbf{D}_\mu \\ &= \left[-\frac{1}{2} (\mathbf{D}_\phi \mathbf{D}_\mu)^{-T} + \frac{1}{2} (\mathbf{D}_\phi \mathbf{D}_\mu)^{-T} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' (\mathbf{D}_\phi \mathbf{D}_\mu)^{-T} \right] \circ \mathbf{D}_\mu \\ &= -\frac{1}{2} (\mathbf{D}_\phi \mathbf{D}_\mu)^{-T} \circ \mathbf{D}_\mu + \frac{1}{2} (\mathbf{D}_\phi \mathbf{D}_\mu)^{-T} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' (\mathbf{D}_\phi \mathbf{D}_\mu)^{-T} \circ \mathbf{D}_\mu \\ &= -\frac{1}{2} \mathbf{D}_\phi^{-T} \mathbf{D}_\mu^{-T} \circ \mathbf{D}_\mu + \frac{1}{2} \mathbf{D}_\phi^{-T} \mathbf{D}_\mu^{-T} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' \mathbf{D}_\phi^{-T} \mathbf{D}_\mu^{-T} \circ \mathbf{D}_\mu \\ &= -\frac{1}{2} \mathbf{D}_\phi^{-T} + \frac{1}{2} \mathbf{D}_\phi^{-T} \mathbf{D}_\mu^{-T} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' \mathbf{D}_\phi^{-T}. \end{aligned}$$

If $\mathbf{A}$ is diagonal, then $\mathbf{A}^T = \mathbf{A}$ and

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\Sigma}} \frac{\partial \boldsymbol{\Sigma}}{\partial \mathbf{D}_\phi} \frac{\partial \mathbf{D}_\phi}{\partial \phi} &= \left[-\frac{1}{2} \mathbf{D}_\phi^{-1} + \frac{1}{2} \mathbf{D}_\phi^{-1} \mathbf{D}_\mu^{-1} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' \mathbf{D}_\phi^{-1} \right] \circ \mathbb{I} \\ &= -\frac{1}{2} \text{diag}(\mathbf{X}_\phi \phi)^{-1} \circ \mathbb{I} + \frac{1}{2} \text{diag}(\mathbf{X}_\phi \phi)^{-1} \mathbf{D}_\mu^{-1} (\mathbf{y} - \boldsymbol{\mu}) (\mathbf{y} - \boldsymbol{\mu})' \text{diag}(\mathbf{X}_\phi \phi)^{-1} \circ \mathbb{I}. \end{aligned} \quad (5.4)$$

Equalizing (5.4) to zero

$$-\frac{1}{2}\text{diag}(\mathbf{X}_\phi \hat{\phi})^{-1} \circ \mathbb{I} + \frac{1}{2}\text{diag}(\mathbf{X}_\phi \hat{\phi})^{-1} \mathbf{D}_\mu^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \text{diag}(\mathbf{X}_\phi \hat{\phi})^{-1} \circ \mathbb{I} = 0$$

$$\therefore \text{diag}(\mathbf{X}_\phi \hat{\phi})^{-1} \circ \mathbb{I} = \text{diag}(\mathbf{X}_\phi \hat{\phi})^{-1} \mathbf{D}_\mu^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \text{diag}(\mathbf{X}_\phi \hat{\phi})^{-1} \circ \mathbb{I}$$

$$\therefore \text{diag}(\mathbf{X}_\phi \hat{\phi}) = \mathbf{D}_\mu^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \circ \mathbb{I}.$$

Multiplying by $\mathbb{J}_{n \times 1}$ in both sides and factoring out $\hat{\phi}$ yields

$$\hat{\phi} = (\mathbf{X}'_\phi \mathbf{X}_\phi)^{-1} \mathbf{X}'_\phi \{(\mathbf{D}_\mu^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})' \circ \mathbb{I}) \mathbb{J}\} \quad (5.5)$$

where $\mathbb{I}$ is an $n \times n$ identity matrix and $\mathbb{J}$ is an $n \times n$ vector of 1s, and $\circ$ the Hadamard product.