Maynooth University
National University of Ireland Maynooth

Topics in Model Evaluation and Comparison

A dissertation submitted for the degree of

Doctor of Philosophy

By:

Darshana Jayakumari

Under the supervision of:

Dr. Rafael A. Moral

Hamilton Institute

Maynooth University

August 2024

To Amma and Abey.

Declaration

I hereby declare that I have produced this manuscript without the prohibited assistance of any third parties and without making use of aids other than those specified.

The thesis work was conducted from September 2020 to August 2024 under the supervision of Dr. Rafael A. Moral in the Hamilton Institute, Maynooth University.

Darshana Jayakumari.

Maynooth, Ireland,

August 2024.

Sponsor

This work was supported by a Science Foundation Ireland grant number 18/CRT/6049.

Collaborations

Dr. Rafael A. Moral: As my supervisor, Dr. Moral (Maynooth University) supervised and collaborated on the work of all chapters.

Prof. John Hinde: Prof. John Hinde (University of Galway) collaborated on all the chapters by reading and providing valuable observations and guidance.

Prof. Jochen Einbeck: Prof. Jochen Einbeck (Durham University) collaborated on on all the chapters by reading and providing valuable observations and guidance.

Dr. Idemauro Lara: Dr. Lara collaborated on Chapter 5 by providing the data and advice at the model formulation and by reading and providing suggestions on the written chapter.

Dr. Julien Mainguy: Dr. Mainguy collaborated on the Chapter 4 by providing the initial idea and by reading and providing suggestions on the written chapter.

Publications

The chapters contained in this thesis have been either published, submitted to peer-reviewed journals or manuscript is under preparation. Chapter 3 has been published in *Modelling Insect Populations in Agricultural Landscapes*. Chapter 4 has been submitted to Journal of Statistical Computation and Simulation and is currently under peer review. Chapter 5 is under preparation and will be submitted to the Brazilian Journal of Probability and Statistics.

Peer-reviewed journal article:

- Jayakumari, Darshana, et al. "Tools for Assessing Goodness of Fit of GLMs: Case Studies in Entomology." Modelling Insect Populations in Agricultural Landscapes. Springer, Cham, 2023. 211-235.

Submitted articles (under review):

- Darshana Jayakumari, Jochen Einbeck, John Hinde, Julien Mainguy, Rafael de Andrade Moral. "A goodness-of-fit diagnostic for count data derived from half-normal plots with a simulated envelope." arXiv:2405.05121v1

Contents

Abstract

Statistical modelling of data in real-world scenarios often require models that can accommodate variability and dependence in a plethora of different ways. Within the generalized linear modelling framework, various extended models are available to address these commonly found problems. In this context, goodness-of-fit assessment is pivotal for ensuring reliable inferential results. This includes, but is not limited to, graphical tools, which can be helpful when deciding whether a data sample can be a plausible realisation of a fitted model. In this thesis, we focussed on the development of diagnostic measures and the comparison of different modelling approaches, when applied to diverse types of data.

Initially, we present an overview of diagnostic analyses stemming from generalized linear models, whilst illustrating different tools using two case studies from experiments in ecology and agriculture. Then, we extend the graphical model selection method known as half-normal plots with a simulated envelope. The simulated envelope is such that, under a well-fitted model, the majority of points should fall within its bounds. Nonetheless, closely related models tend to produce very similar graphs. We propose a new distance-based framework that acts as an added quantitative summary to the half-normal plot with a simulated envelope. This new measure can effectively determine the most appropriate model when closely related models are included. An extensive simulation study was carried out taking into

account many different scenarios. The results showed that the distance framework exhibits robust performance in finding the true model and is comparable to BIC; in some instances, it even displays superior efficacy.

Finally, we present a comparative analysis of different modelling frameworks applied to interval-censored longitudinal data, which is bounded in the interval $(0, 1)$. We considered three approaches, where the first and second involved mixed and marginal models using a transformation of the interval-censored response, and the third incorporated the interval-censored nature in the likelihood. We found that the accounting for the interval-censored nature of the data improved model goodness-of-fit. However, the conclusions drawn from all three approaches were qualitatively similar.

Acknowledgements

I would like to thank my Amma for her immense support and love through out my life and my husband Abey Jose who supported me with all the love and patience.

I would also like to thank my supervisor Rafael de Andrade Moral who has motivated me and helped me in every step in this PhD and has been a friend to me more than a supervisor.

I would also like to thank my collaborators John Hinde and Jochen Einbeck for their support and collaboration through out my PhD. This would not be complete without mentioning Idemauro with whom I had the opportunity to collaborate during my third chapter, and treated me more like family than a visiting student.

Next, I would like to thank Maria Rodriguez who went out of her ways to support me and motivated me to apply for this PhD and helped me take decisions when I could not and mentored me to be a strong person.

I would also like to thank my hiking group “Over the hills" who welcomed me with open arms and helped me to pursue the best thing I enjoy in my life. Every Saturday has been pure joy and life lessons that helped to overcome stressful times in my PhD. I would like to particularly mention Ray for patiently answering all my questions even when we were struggling for breath; Willie and Sean who always motivated and gave me the best possibilities and advice in difficult times.

I would also like to thanks my friends Daire, Ahmed, Aswathy, Shauna, Chang, Akash, Fergal, Nathan, Pramit, Tzirath, Amit, Jonny, Sneha, Sreeraj and all other cohort members and colleagues in Hamilton Institute for always being there for me and helped me to integrate into a foreign country and supported me in stressful times.

I would also like to thank Prof David Malone and Prof Ken Duffy and other directors of the CRT program who supported me to get more exposure and to learn and grow in such a diverse and enriching academic environment during my PhD. I would also like to thank the admin staff in Hamilton Institute Rosemary, Kate and admin staff from the CRT Joanna, Janet, Pat and Peg for making my life easier by supporting me with all the admin work.

I would like to thank the faculty at Maynooth University for providing me with the resources to pursue this graduate research in the Hamilton Institute.

I would also like to thanks all my friends and family who helped me and supported me during my PhD life.

Finally I would like to thank the Science Foundation Ireland, who financed my research and without whom the undertaking of this project would not have been possible.

List of Figures

List of Tables

CHAPTER 1

Introduction

In this chapter we discuss the motivation behind the work presented in this thesis and outline the content in the chapters included in the thesis.

Motivation

The word ecology comes from two Greek words, “oikos" meaning household and “logos" meaning study (Odum et al., 1971). Thus ecology is the study of household, which includes investigating all the organisms, habitats and the processes involved in sustaining life. The development of using science to understand organisms and relationships between them and its surrounding have helped mankind to understand many natural processes. This also involved using statistics to aid in comprehending patterns in the behaviour of ecosystems. Data can be collected from ecological experiments in natural or controlled environments. There are different types of data involved in modelling ecological process. The common examples include: counts, continuous, proportions, categorical, interval data, among others. The statistical processes involved in modelling depends on the nature of

the data, the design of the experiment or observational study, and the research questions. In the first part of this thesis we focus on modelling count data, while later we turn our focus to longitudinal data.

Count data consists of discrete data points that reflect the occurences of an event in a specified period of time or unit of space (Coxe et al., 2009). The basic modelling strategy for count data involves using Poisson regression and extensions. The Poisson model is an equidispersion model, i.e. it assumes that the variance is equal to the mean. But when analysing real count data, especially in ecological studies, it is very rare that the data is equidispersed (Harrison, 2014). Typically, (1) the variability can be greater than predicted by the model which is known as overdispersion; or (2) the variability can be lower than predicted by the model which is known as underdispersion. Cases of underdispersion are not frequently encountered in ecological studies measuring animal abundance, which typically exhibit overdispersion. An example of a scenario where underdispersed data can be observed is when analysing species richness.

The dispersion can be modelled in different ways within the generalized linear modelling framework which is an extension to classical linear models (Nelder and Wedderburn, 1972). The generalized linear model stems from the exponential family of distributions that gives the user flexibility to model the mean-variance relationship with additional parameters to account for unexplained variance. Commonly used extensions of the Poisson distribution within the generalized linear modelling framework are the Negative binomial distribution and the Quasi-Poisson, which itself is not a true probability model, since it only specifies first and second moments. Quasi-likelihood based models are also sensitive to differences in sample size and highly skewed data. It should be noted that some models like the Negative binomial can only account for overdispersion, while the Quasi-Poisson can be

used for modelling both overdispersion and underdispersion. A further scenario that can occur in count data is the occurrence of an excess number of zero counts that could lead to heavily skewed data, termed zero-inflation. In instances where zero-inflation is incorrectly treated as overdispersion this could potentially lead to incorrect estimation of model parameters, standard errors and incorrect specification of the distribution of the test statistics (Perumean-Chaney et al., 2013). This could be handled by using appropriate models that can handle an excess number of zero counts, such as the zero-inflated Poison and zero-inflated Negative Binomial distributions. The other scenarios of skewness in the data could be dealt with by applying appropriate transformations and specifying more flexible variance functions.

Model fitting is followed by checking whether the model we have fitted is adequate, considering the assumptions and biological nature of the problem. The most common methods for assessing model goodness-of-fit involve using residual analysis and graphical methods. The first paper presented in this thesis (Chapter discusses tools and diagnostics for assessing goodness-of-fit in GLMs. The second paper (Chapter 4) introduces a novel metric to assess goodness-of-fit of GLMs applied to count data. We showcase the methodology using examples arising from studies in 2.6 and 2.7.

The third paper (Chapter 5) focusses on exploring marginal and mixed modelling frameworks (Verbeke et al., 1997) applied to longitudinal interval censored data, and how these methodologies compare with interval-censored regression. Longitudinal data is characterised by repeated measurements from a single subject. This type of data is inherently correlated and can be modelled using marginal models with different working correlation structures and by using random effects within a mixed modelling framework. In the particular example dataset explored, the range

of response variable is bounded between 0 and 1, therefore we use interval-censored beta regression to model the data.

Thesis Outline

The remaining chapters of this thesis are organised as follows:

Chapter 2 discusses the five ecological case studies used in this thesis. This chapter aims to provide a brief introduction to the examples explored in the subsequent chapters. This will also include exploratory analysis. We intend to discuss the original published source and provide information for interested parties who may wish to fully reproduce the analysis in this thesis using these datasets.

Chapter 3 is a review chapter discussing the tools for assessing goodness-of-fit of GLMs illustrated with two case studies in entomology. This chapter aims to give a brief introduction to the exponential family and GLMs, and further delve into the goodness-of-fit methods. We intends to give the reader a deep understanding o residual analysis and a class of residual plots known as half-normal plots with a simulated envelope. This is followed by discussing influence measures used in statistical inference. We consider that this chapter will serve as a background chapter for the subsequent chapters of this thesis. The first example in this chapter examines the effect of an entomopathogenic nematode called *Heterorhabditis bacteriophora* against a tick commonly found in cattle named *Rhipicephalus* under laboratory and field conditions (Filgueiras et al., 2023a). The second example discusses the usage of fungi as bioagents against plant parasites. These fungi are used as a sustainable method against the chemical pesticides that poses potential hazardous effects to the environment and human health (Silva et al., 2022a). The chapter also discusses the motivation for a distance-based metric as an alternative or a complementary step to half-normal plots.

Chapter 4 introduces a novel distance-based framework based on half-normal plots with a simulated envelope, and presents an extensive simulation study aimed at evaluating its performance. This chapter aims to provide a good understanding on the usability of the distance metric. The simulation framework setup assesses the effectiveness of the distance metric as a model selection method in the cases of mild and strong overdispersion, as well as cases of mild and strong zero inflation. The modelling framework involves the Poisson and extensions of the Poisson namely Negative Binomial distributions with quadratic and linear variance functions, the Quasi-Poisson, and the Zero Inflated Poisson and Negative Binomial models. This chapter also includes two example datasets to explore the application of the suggested metric to real life cases. The first example investigates the effect of soil dry mass as an environmental factor on the distribution of hunting spiders (Smeenk-Enserink and Van der Aart, 1974). The second example considers the age frequency data of walleye fish from gillnet surveys in Canada (Mainguy and Moral, 2021).

Chapter 5 evaluates the application of mixed and marginal models on interval censored data and how that compares to interval censored regression in the context of the beta distribution. It aims to give an introduction to the marginal and mixed modelling approaches with a special focus on how correlated data are dealt with in both cases. This chapter also aims to give the reader a brief review on interval censored regression. This chapter is motivated by an example provided in (Garzón-Barrero et al., 2016). The experiment focusses on evaluating the efficacy of innovative sugarcane bagasse particle boards compared to traditional Medium density particle boards. We considered different formulations for fixed effects and random effects for mixed modelling and interval censored regression, and used different correlation structures for the data in the marginal modelling approach to analyse the data.

Chapter 5 briefly discusses the results of the studies in each of the three chapters. This section also explains the implications and the limitations of the work in this thesis and further delves into the future directions and proposes new lines of investigation.

All proposed methods in this thesis were implemented using R (R Core Team, 2022) software and are accessible at the author's Github[1] via two public repositories. The repository https://github.com/DARSHANAJAYA/Goodness-of-fit-Distance-metric and related to Chapters 3, 4, and repository https://github.com/DARSHANAJAYA/Fungi-study- relates to chapter 5. These repositories include all the scripts to reproduce the analyses and the plots provided in the chapters.

[1] https://github.com/DARSHANAJAYA

CHAPTER 2

Case Studies

In this chapter, we discuss the various case studies utilised as part of this thesis, provide details as to their notable features, and provide links to where they are located for practitioners interested in reproducing our work.

2.1 Introduction

In this thesis we examine five datasets; three of which involve count responses and two involve continuous responses. We introduce a count dataset and a continuous dataset to demonstrate modelling with GLMs and a graphical model selection method known as half-normal plots in Chapter 3. In Chapter 4, we illustrate a novel goodness-of-fit methodology using two count datasets. Finally, in Chapter 5 we explore the use of different modelling strategies when analysing a longitudinal dataset with an ordinal response variable, which can be treated as continuous and interval-censored. The datasets cconsidered in this thesis are comparitively small and has no apparent skewness or zero inflation to be addressed in the model specification.

2.2 Biological control of ticks

Rhipicephalus (Boophilus) microplus (Canestrini, 1888) (Acari: Ixo- didae), otherwise known as cattle ticks, impose significant impact on beef and dairy cattle husbandry by causing severe economic losses (Rodriguez-Vivas et al., 2018). The species can be controlled via pesticides, however there are a number of reasons to look for alternatives to chemical control. One of the reasons is that ticks may develop high resistance to the chemicals as a consequence of repeated usage. Also, the presence of chemicals in dairy products has also elevated the aversion to them (Klafke et al., 2017). There is, therefore, a need for further research on biocontrol agents that are safer for the environment. The application of entamopathogenic nematodes (EPNs) has been identified to be an effective means of pest management. The symbiotic relationship between the EPNs and the ticks allows for sustained pest elimination. The contact of EPNs and ticks induces rapid septicemia and leads to their mortality. At different developmental stages, there is a varied susceptibility of the ticks to the nematode infection and its been shown that engorged females are more susceptible.

Filgueiras et al. (2023a) explored the susceptibility of ticks to the EPNs at different engorgement levels and at different body weights and tick sizes. The impact of different tick populations from different geographical locations was also evaluated. The experimental setup consisted of a field trial with two groups comprising of a treatment group and a control group. The treatment group was treated with *Heterorhabditis bacteriophora* and the control group had no treatment. Each group had sex replicates (plots) of *Megathyrsus maximus* grass, which provides an ideal environment for the survival of *R. microplus*. The treatment was done by using infected dead specimens of *Tenebrio molitor* larvae and prior to one week of the start of the experiments the treatment groups were seperated and buried in the soil at random points in each of the plots. The total number of ticks in each plot

Mean		Variance	
Control	NEPs	Control	NEPs
91.63	71.92	13003.55	46984.25

Table 2.1: Mean and Variance of Control group and NEPs for the ticks data

was counted weekly for a period of four weeks. The dataset includes 48 rows with 5 covariates. The five covariates are treatment, the number of blocks, number of weeks, Number of ticks and an additional column that includes the cumulative number of ticks.

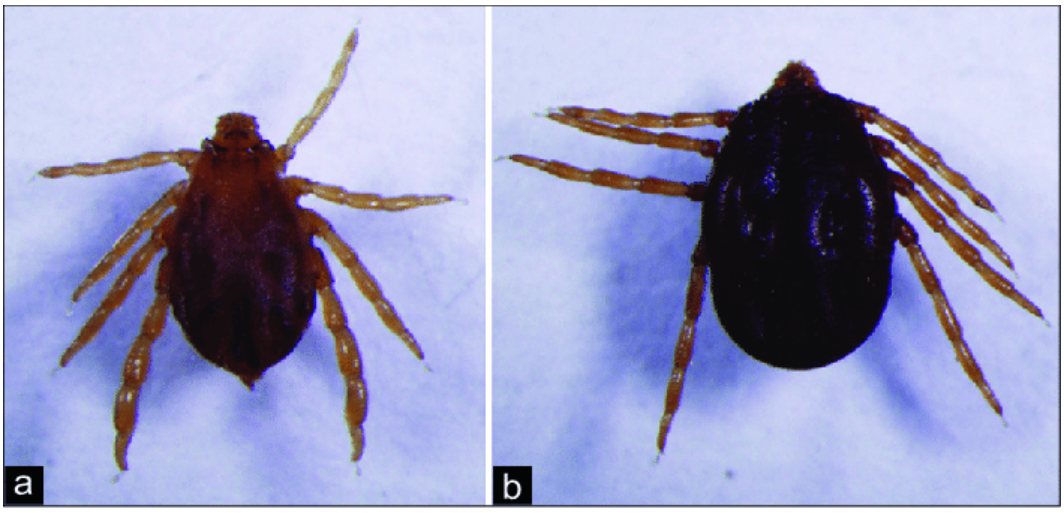


Figure 2.1: Images of (a) *Rhipicephalus (Boophilus) microplus male and (b) Rhipicephalus (Boophilus) microplus* female (Brites-Neto et al., 2015).

2.3 Sustainable management of parasitic nematodes using bioagents – the 'plant height' data

The microbial treatment of agricultural pests is on high demand, as it helps to reduce resistance to chemical pesticides and aids in the sustainable management of pest populations by improving the organic crop development (Eilenberg et al.,

2001). The usage of natural enemies against the parasitic nematodes is beneficial to human health by minimizing the utilization of chemical pesticides. These low impact bio agents have narrow spectrum and higher selectivity for the hosts, when compared to a broader spectrum selectivity exhibited by the chemical pesticides. In this case study we considered filamentous fungi of the order *Hypochreales*. Among these fungi, two soil-born cosmopolitan fungi, namely *Purpureocillium lilacinum* and *Pochonia chlamydosporia*, are known for their pest controlling capabilities (De Souza et al., 2015). These two fungi are abundantly found in the rhizosphere of the vegetation and secretes hydrolytic enzymes against the parasitic nematodes. They also have the added advantage of supporting plant growth by establishing endophytic associations. So usage of fungal nematicides not only helps in getting rid of parasitic nematodes but aid in the overall development of the plant growth.

There are two different methods of mass production of the fungal propagules: by solid or submerged fermentations. Solid fermentation is a low cost method that uses less water with plenty of oxygen supply for the propagation of the fungal structures for the dissemination of the infection. Although this is a sustainable method in producing aerial conidiophores[2], it takes around 10 or 15 days to achieve high sporulation with a higher risk of contamination due to the uncontrollable nature of the environmental conditions (pH, water activity, aeration, nutrition levels). Submerged fermentation method is a low cost method that has a considerably lower risk of contamination and provides higher yield. This method also has an advantage of shorter duration for propagation of submerged conidia which has different structure and characteristics compared to aerial conidia. A notable fungal structure for the pest control named microsclerotium is of selective consideration as an alternative to aerial conidium. These structures has added stability and resilience

[2]A specialized hyphal branch of some fungi that produces conidia.

under field conditions and is also tolerant towards UV-B and heat when compared to the latter.

Silva et al. (2022b) explored the properties of dried fungal propagules after submerged fermentation. The experiment was conducted in a randomized block design with ten replicates for each treatment. A total of six treatments were considered, which included five fungal treatments and one control. This experiment was repeated three times at different time points under greenhouse conditions. The treatment consisted of surface sterilisation using 70% ethanol followed by sodium hypochlorite and further back to ethanol and distilled water. Following this, Arabic gum was added as a coating to seeds of the common bean cv. 'IAC Milenío'. The treatment was done using two isolates of *P. chlamyhdosporia* and three isolates of *P. lilacinum*. The treated seeds were then placed on a filter paper to dry out the excess water. In the final phase the seeds were potted with a combination of soil and sand and kept in greenhouse conditions. These plants were watered daily with tap water whenever required. After 45 days, the plants were removed from the pots and the roots were cleaned from soil. Subsequently, the dry mass measurements were carried out for the aerial part and the root part of the plants. The dataset includes 180 rows and 4 columns. The 4 columns includes the treatment considered, experiment, plant considered and the measured dry weight.

Figure 2.4 shows the distribution of dry weight measured in grams for two isolates of *P. chlamydosporia* and three isolates of *P. lilacinum*, as well as the control group.

Spider data

Smeenk-Enserink and Van der Aart (1974) examined the spatial distribution of *Alepecosa accentuata*, a hunting spider species, found in the dune area 'Meijendel'

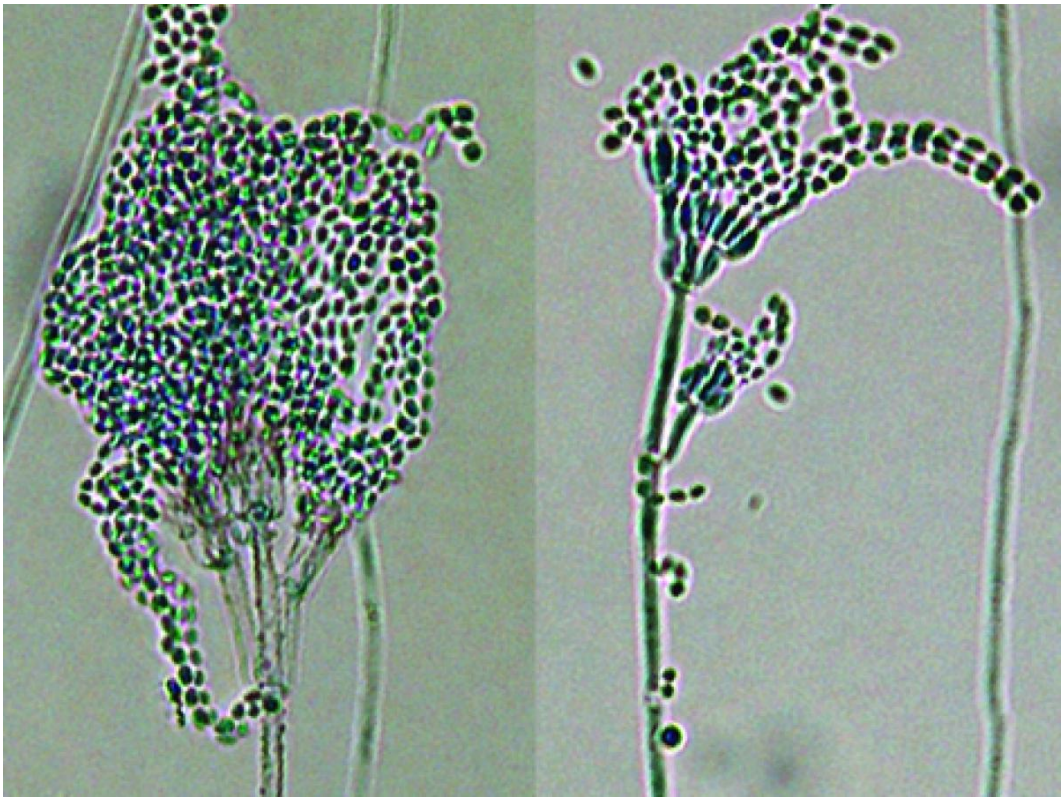


Figure 2.2: *Purpureocillium lilacinum* conidiophores, phialides and conidia. (Source: https://www.adelaide.edu.au/mycology/fungal-descriptions-and-antifungal-susceptibility/hyphomycetes-conidial-moulds/purpureocillium)

situated between the The Hague and Wassenaar in the Netherlands, and compared it to environmental characteristics. These are considered to be non-specialist predators, which means that they feed on multiple different species of prey. The motivation for conducting this study included the need for unravelling the functional differences in the predatory nature with the environmental characteristics, such as soil dry mass. The study was conducted using a pitfall catch method that gives insights to the density and movement of the spiders. One hundred pitfall traps were set up in four square-grid arrangements of 25 pitfalls each in the Bierlap dune valley. The smallest distance between two pitfalls was 10m. Among these, 28 pitfalls were selected and the water content in the soil was estimated

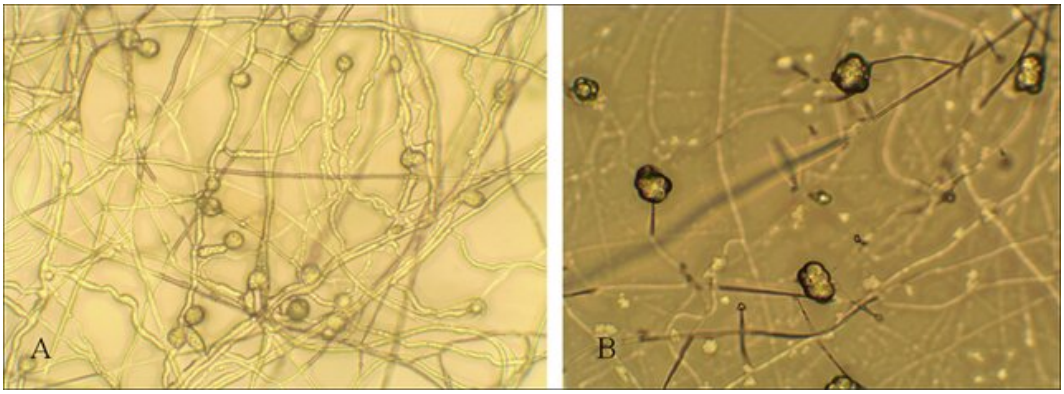


Figure 2.3: Images of the fungus *Pochonia chlamydosporia* under light microscope. (A) conidia and mycelium at 10× magnification, (B) chlamydospores at 40× magnification (Oliveira et al., 2022).

gravimetrically. This experiment comes with challenges as a number of hunting spider species can be found in the same area and the spatial separation of the spiders is not very evident.

Figure 2.6 displays a scatter plot of soil dry mass and the number of spiders belonging to the species *Alepecosa accentuata* found in the dune area of Meijendel, collected in 28 pitfall traps.

2.4 Walleye data

To estimate instantaneous mortality of walleye (*Sander vitreus*) fish specimens in Québec, Canada, fish were collected as a part of a large standardized gillnet provincial monitoring program by the Service de la Faune Aquatique in 2012 (Mainguy and Moral, 2021). The site of data collection was the Baskatong Reservoir near the city of Mont Laurier. The gillnets considered comprised of eight panels with increasing mesh sizes to facilitate capturing fish with a wide range of lengths. This helps to reduce the catchability bias related to fish size and gives better estimates of their age. The ages were evaluated based on the examination of year-increment

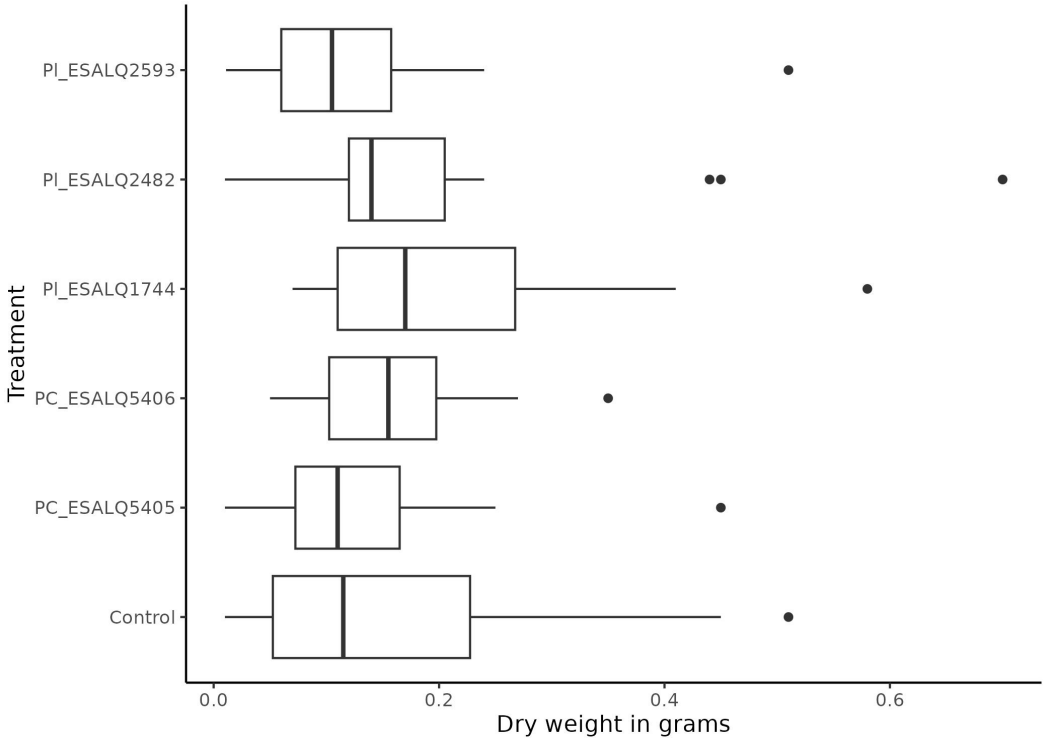


Figure 2.4: Box plots of the dry weight variable measured in the 'plant height' dataset for ten replicates of six different treatments: three isolates of *P. lilacinum*, two isolates of *P. chlamydosporia* and a control.

annuli of the otoliths[3] using a microscope. The age frequency data (Figure 2.7) is also referred to as catch curve data and is typically overdispersed (Nelson, 2019). The dataset consists of 99 observations with 3 columns (age of the fish, count of the fishes, year considered).

2.5 Mold growth data

This dataset was obtained from a comparative study carried out by Garzón-Barrero et al. (2016) to evaluate the effectiveness of a novel sugarcane-based bagasse particleboard when compared to the traditional medium density particle boards. The

[3]Annual growth increment of fish otoliths or earstone located behind the brain of the bony fishes.

Figure 2.5: Image of spider species *Alepecosa accentuata*.(Wikipedia contributors, 2021)

medium density particle boards are generally made from *Pinus* and *Eucalyptus* trees. The increased demand of particle boards in the construction necessitates an alternative for particle board manufacturing. One of the alternative options for traditional particle boards is the sugarcane bagasse particle boards. The durability of the particle boards are evaluated by quantifying the percentage of mold growth in a controlled environment. The usage of an polyurethane resin adhesive as the binding agent in the sugarcane based particle boards compared to synthetic resins like phenolformaldehyde resins gives an added advantage of biodegradability. For the experiment the sugarcane particle boards were manufactured under laboratory conditions using a castor oil based polyurethane bicomponent resin as an adhesive. The medium density particle boards were produced on an industrial scale using

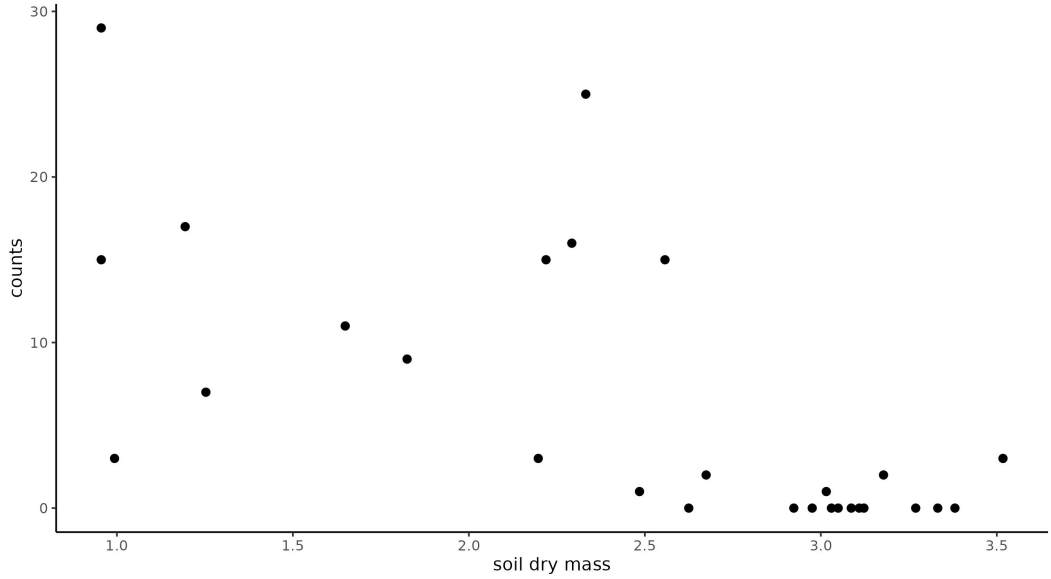


Figure 2.6: Number of hunting spiders of the species *Alepecosa accentuata* collected in 28 pitfall traps versus soil dry mass.

Eucalyptus particles with urea formaldehyde resin. The percentage of mold growth was measured in an ordinal scale with 11 levels, ranging from from 0 to 10, with 0 representing a rate of mold growth between $90\% - 100\%$, 1 corresponding to $80\% - 89\%$, 2 corresponding to $70\% - 79\%$, and so on, with 10 corresponding to 0% mold growth. This data can be seen either as ordinal, or interval-censored and bounded between 0% and 100%. The experiment was set up in a 2×2 factorial design with two wooden materials(BCP, MDP) and two settings for the coating (With coating, No coating) with six replicates, and the data was measured weekly for a period of four weeks. Figure 2.9 displays the rate of mold growth over time for all treatments in the experiment.

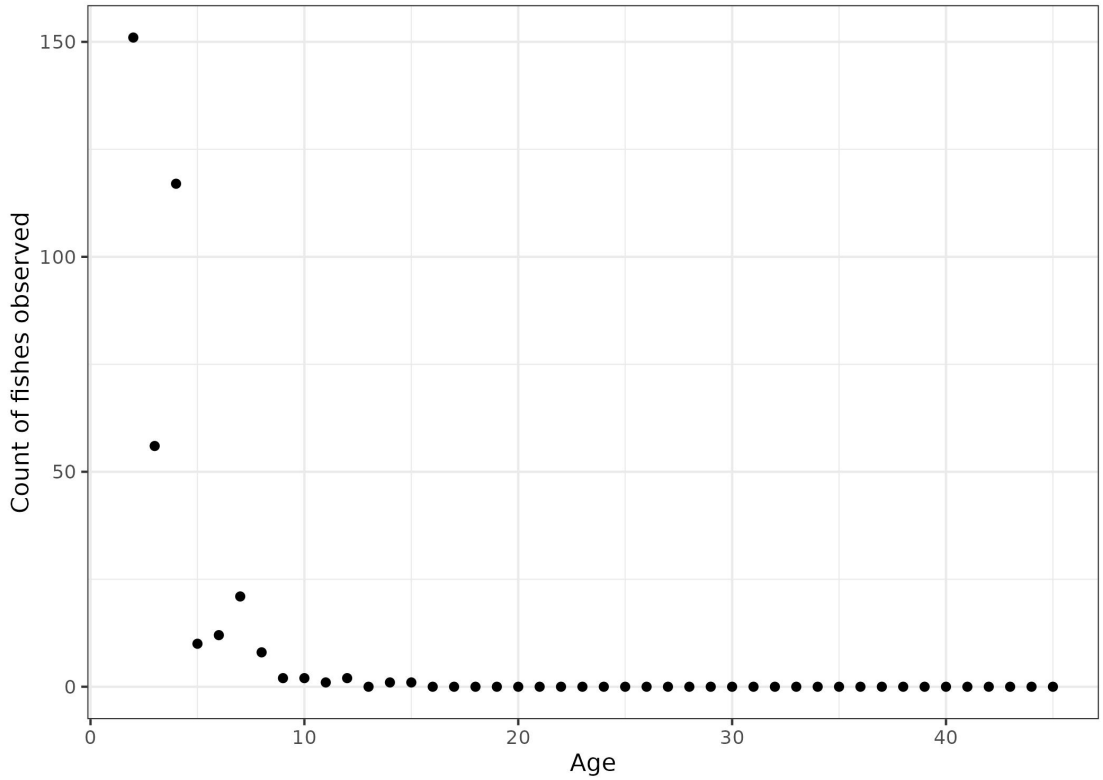


Figure 2.7: Age distribution of walleye fish captured in 2012 by the Service de la Faune Aquatique, in Québec, Canada.

Figure 2.8: Experiment setup to evaluate the mold growth in a controlled environment (Barrero, 2015).

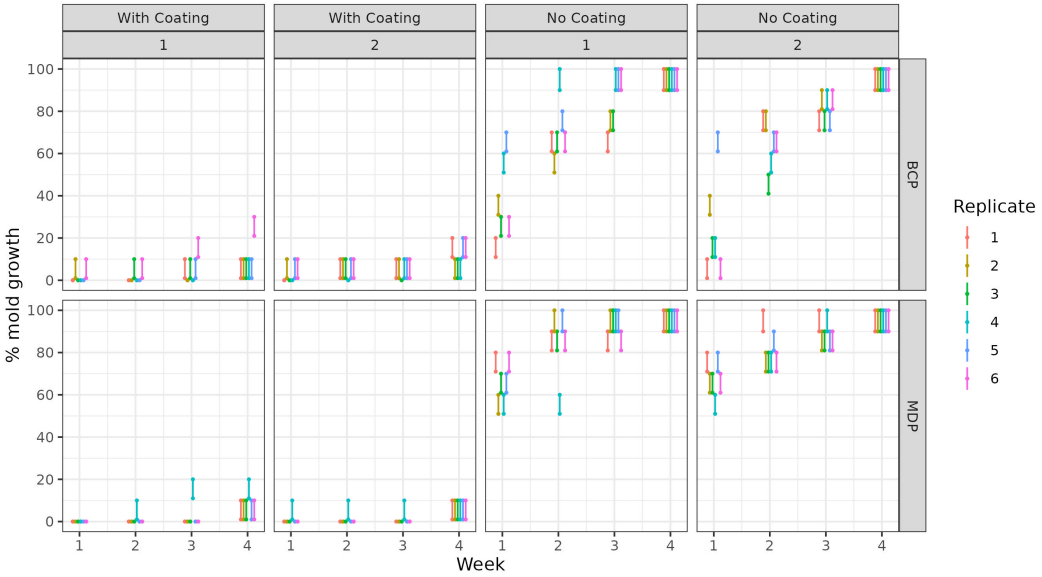


Figure 2.9: Rate of mold growth for two different materials (MDP, BCP)over a period of 4 weeks for conditions of Face and coating for the boards.

.

CHAPTER 3

Tools for Assessing Goodness-of-fit of GLMs: Case Studies in Entomology

In this chapter, we discuss the analysis of data that typically arise from entomological studies using generalized linear models. We focus on techniques that can be used to assess model goodness-of-fit, which is an important step in statistical modelling to ensure the reliability of the inferences made. Specifically, we demonstrate the utility of half-normal plots with a simulated envelope as a complementary tool for assessing model assumptions. We illustrate the concepts with two examples, one involving count responses and another involving continuous responses.

This chapter was published in Jayakumari, D., Hinde, J., Einbeck, J., Moral, R.A. (2024) Tools for assessing goodness-of-fit of GLMs: Case studies in entomology. In Moral, R.A., Godoy, W.A.C. (2024) Modelling insect populations in agricultural

landscapes, Springer.

3.1 Introduction

Given a scientific hypothesis, an experiment or observational study can be carried out to collect data that may confirm or provide evidence against the said hypothesis. Statistical models represent an attempt to explain patterns of variation found in a response variable through the use of specific distributional assumptions and predictor variables. These patterns change depending on the nature of the response. A useful model should be able to capture most of the relevant variation in the data and distinguish between a true signal and noise, while at the same time maintaining parsimony.

Consider the following multiple linear regression model:

$$\begin{aligned} \mathbf{Y} &\sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I}_n\sigma^2), \\ \boldsymbol{\mu} &= \mathbf{X}\boldsymbol{\beta}, \end{aligned} \tag{3.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ is the vector of responses of dimension n, $\boldsymbol{\mu}$ is the vector of means, $\mathbf{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients of dimension p, $\mathbf{I}_n$ is the $n \times n$ identity matrix and σ^2 is the variance parameter. This model makes three main assumptions: (i) the response variable is assumed to be normally distributed; (ii) the means are allowed to differ across the Y_is, according to the specified linear model, however their variances are assumed to be the same (σ^2 for all Y_i); and (iii) the responses are assumed to be independent (which under the assumption of a normal distribution is implied from the diagonal covariance matrix with $\mathrm{Cov}(Y_i, Y_j) = 0$, for $i \neq j$).

When fitting model (3.1) to real data, it is important to assess whether the aforementioned assumptions are met before treating any inferential results as reliable. Assumption (i) can be checked in many different ways. We may assess it through hypothesis testing using, for example, the Shapiro-Wilk test of normality based on a standardised version of the model residuals Shapiro and Wilk (1965). We may also carry out graphical assessments, that include quantile-quantile plots, which will be discussed in detail in Section 3.4. Assumption (ii) can also be checked through formal hypothesis tests, such as the Bartlett test for variance homogeneity Snedecor and Cochran (1989), and graphically, by looking at a plot of residuals versus fitted values (see Figure 3.3 for examples). Frequently, assumption (iii) will be deemed to be true or not based on the design of the experiment or observational study, without resorting to formal hypothesis testing. Tests are available, but typically they are onl;y carried out when analysing longitudinal or time-series data, which may be assumed to be correlated. For these cases, calculating the empirical auto-correlation and partial auto-correlation functions is especially helpful.

Nevertheless, all assumptions discussed above can be summarised by a single distributional assumption, which is denoted by (3.1). If the distributional assumption in (3.1) is true, then the observed data must be a plausible realisation of the estimated model $\mathrm{N}(\hat{\boldsymbol{\mu}}, \mathbf{I}_n\hat{\sigma}^2)$. In other words, theoretically we should be able to generate the observed values of the response variable we obtained in the experiment or observational study by simulating from our fitted model. Goodness-of-fit assessment methods that rely on simulation (such as half-normal plots with a simulation envelope) are based on this principle. They involve simulating multiple times from a fitted model and comparing the results obtained with the observed response or a function of the response. This is especially useful in the context of more general models (e.g. based on a wider family of probability distributions), for which assumptions (i) and (ii) do not hold (and consequently carrying out tests for

normality and homogeneity of variances would be pointless in this case). Note that here we confine our attention to settings where the assumption (iii) of independent responses is plausible by virtue of the form of data collection; more general data structures may require model extensions, such as the mixed modelling framework for multiple components of variation.

In entomological studies, it is often the case that non-normal continuous data and discrete data (counts and proportions) are collected. For these types of data, model (3.1) would not be suitable for analysis, and different distributional assumptions would be required. For instance, to analyse count data, the Poisson model is one of many alternatives; to analyse proportion data, the binomial model could be used; and gamma and inverse Gaussian models are able to flexibly accommodate right-skewed data with positive support. These are all examples of generalized linear models, for which many different extensions are also available.

There is a plethora of modelling options available for the analysis of entomological data. In many cases, more than one distribution can suitably accommodate the variability in the data. In Section 3.2, we provide a general definition and overview of generalized linear models. Later, in Sections 3.3, **??** and 3.4, we present an overview of goodness-of-fit assessment tools and techniques. We conclude by illustrating their use with real datasets in Section 3.5. All analyses and figures in this chapter are generated using R R Core Team (2022).

3.2 Generalized Linear Models

The modelling of non-normal data could involve, as alternatives, distributions belonging to the exponential family (EF) of distributions. The EF includes discrete (e.g. Poisson, negative binomial, and binomial), as well as continuous (e.g. gamma and inverse Gaussian) distributions, representing flexible alternatives to the normal

distribution when modelling discrete, or continuous and strictly positive, or skewed data in entomology.

The probability density function (pdf) of a random variable Y whose distribution belongs to the EF of distributions can be written, in the canonical form, as

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}, \tag{3.2}$$

where $b(\cdot)$ and $c(\cdot)$ are functions of the dispersion parameter ϕ, the canonical parameter θ and the data and dispersion, respectively.

The generalized linear model (GLM) is an extension to the classical linear model where the response variable is assumed to follow a distribution which belongs to the EF. By developing an inferential framework that encompassed distributions belonging to the EF, Nelder and Wedderburn (1972) allowed for fitting GLMs using a unified estimation process. It should also be noted that the exponential family of distributions is often inadequate to represent observed response variables.

The GLM consists of three components:

1. Random component: this is the assumed distribution for the response variable, which belongs to the EF. This component is termed 'random' because it is a probability distribution that is used to model the variability in the data.

2. Systematic component: takes the form of a linear predictor, which consists of a linear combination of the predictor variables and unknown parameters. It may be written as

   $$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

   where $\mathbf{X}_{n\times p}$ is the design matrix and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.

3. Link function: this *links* the random component to the systematic component through a monotonic and differentiable function $g(\cdot)$. Typically we aim to link the parameter corresponding to the mean $\boldsymbol{\mu}$ of the distribution to the linear predictor through the transformation $\boldsymbol{\eta} = g(\boldsymbol{\mu})$. As link functions are invertible, we have that the mean $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ and so is determined by the linear predictor. The link function that transforms the mean $\boldsymbol{\mu}$ to the natural parameter $\boldsymbol{\theta}$ is known as the canonical link. Note that using the canonical link corresponds to specifying a linear predictor for the canonical parameter, which may, or may not, be desirable depending on the substantive questions of interest. Table 3.1 presents six commonly used EF distributions with their canonical link functions, the dispersion (or scale) parameter, and the form of the variance function (discussed in subsequent sections).

Distribution	Representation	$g(\mu)$	$V(\mu)$	ϕ
Normal (Gaussian)	$\mathrm{N}(\mu, \sigma^2)$	μ	1	σ^2
Gamma	$\mathrm{Gamma}(\mu, \alpha)$	μ^{-1}	μ^2	α^{-1}
Inverse Gaussian	$\mathrm{IG}(\mu, \sigma^2)$	μ^{-2}	μ^3	σ^2
Poisson	$\mathrm{Pois}(\mu)$	$\log(\mu)$	μ	1
Negative binomial	$\mathrm{NB}(\mu, k)$	$\log\left(\frac{\mu}{\mu+k}\right)$	$\mu\left(\frac{\mu}{k}+1\right)$	k^{-1}
Binomial	$\mathrm{Binom}(m, \pi)$	$\log\left(\frac{\mu}{m-\mu}\right)$	$\frac{\mu}{m}(m-\mu)$	1

Table 3.1: Representation, canonical link functions ($g(\mu)$), variance functions ($V(\mu)$), and dispersion parameter (ϕ) for six commonly used distributions within the generalized linear modelling framework. Here and throughout this chapter, log denotes the natural logarithm (i.e. log base e).

The estimation of GLMs is typically done using the maximum likelihood (ML) method. For the normal model, ML estimates are equivalent to the ones obtained via ordinary least squares (which aims to minimise the sum of squared differences between observed and fitted values). Under the assumed distribution, the ML

estimates are the ones that maximise the likelihood function, that is the likelihood of the observed data being generated by that distribution. Let $f(y_i; \boldsymbol{\beta}, \phi)$ be the probability density or mass function distribution function for observation i, where $\boldsymbol{\beta}$ is the vector of regression parameters to be estimated and ϕ is the dispersion parameter of the assumed distribution. The likelihood function for a single observation is defined as $L(\boldsymbol{\beta}, \phi; y_i) = f(y_i; \boldsymbol{\beta}, \phi)$. Assuming observations are independent, the overall likelihood for the full sample is given by the joint probability density or mass function

$$L(\boldsymbol{\beta}, \phi; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\beta}, \phi).$$

To avoid numerical problems when working with likelihood functions, it is commonplace to work with the log-likelihood $l(\boldsymbol{\beta}, \phi; \mathbf{y}) = \log L(\boldsymbol{\beta}, \phi; \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\beta}, \phi)$ instead. Since logarithms are monotonic functions, the parameter values that maximise $l(\cdot)$ will also maximise $L(\cdot)$. Therefore, the ML estimates will be the parameter values that maximise $l(\cdot)$, i.e.

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l(\boldsymbol{\theta}; \mathbf{y}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi)^\top$.

The maximised log-likelihood value is used within different goodness-of-fit criteria, and in some cases it can itself be used as a goodness-of-fit measure. When comparing model fits, a higher value of the log-likelihood would indicate a better reproduction of the observed data. However, a saturated model, for instance, would reproduce every single observation; while the log-likelihood would be larger than for a less complex model, saturated models overfit the data and are not flexible enough to generate predictions for other sets of predictor values. Therefore, selecting a model is a task that involves balancing flexibility and explainability.

The residual deviance measure is defined as the difference between the log-likelihood of the saturated model and log-likelihood of the fitted model (also called 'current' model), scaled by a factor of two, i.e., it conveys how far the model is from fully reproducing the observed data. (Strictly speaking the deviance here is the *scaled deviance* where the basic deviance is ϕ times this and for the EF gives a fitting criteria that does not involve ϕ and plays the same role as the residual sum of squares for the normal model. More recent usages tend to blur this distinction and, of course, for models where $\phi = 1$ it is not an issue, but in other models the user needs to take care as to what version if being reported.) In GLM theory, the deviance is an important measure, because by subtracting the residual (scaled) deviances between two nested models, we obtain a statistic called 'likelihood-ratio' (since the difference between (scaled) deviances is equivalent to a ratio between model likelihoods in the natural scale). It can be proven that likelihood ratios, under the null hypothesis that the simplest model is most adequate to explain the data, for a fixed or known value of the scale parameter ϕ, asymptotically follow a χ^2 distribution with the number of degrees of freedom equal to the difference between the number of estimated parameters between the models being compared.

Now we briefly present the most commonly used EF models when analysing entomological data, for which we provide examples of their probability density or mass functions in Figure 3.1.

3.2.1 The Normal model

The normal distribution is the most commonly used distribution in statistics and is widely used to model real life eventsWeisstein (2012). It is a continuous and symmetric probability distribution, with a probability density function (pdf) given by:

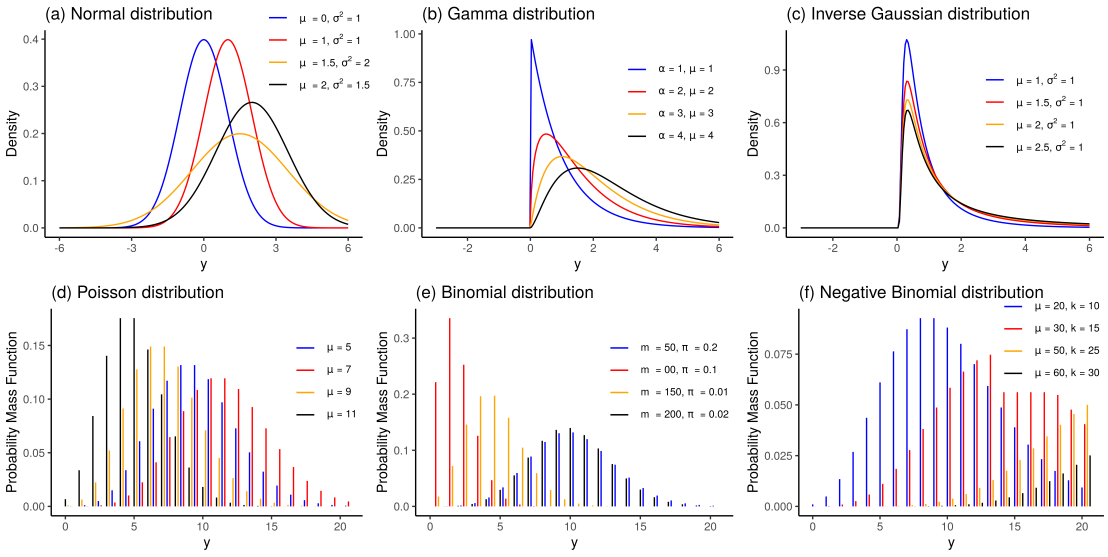


Figure 3.1: Probability density/mass functions for six distributions belonging to the exponential family. The top row shows the density of the continuous distributions for different sets of parameter values. The bottom row shows the probability mass function of discrete distributions.

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \quad -\infty < y < \infty.$$

It is also known as the 'Gaussian' distribution as it was introduced by Carl Friedrich Gauss in 1809. The distribution has two parameters (μ, σ) and is denoted by $\mathrm{N}(\mu, \sigma^2)$ where $\mu \in (-\infty, \infty)$ is the mean and $\sigma^2 > 0$ is the variance. The density function of the normal distribution resembles a bell shape and is known as the 'bell curve' and is centred around the mean μ. When $\mu = 0$ and $\sigma^2 = 1$ the distribution is known as the 'standard normal' distribution. It can be shown

that approximately 68.2% of the area under the curve is contained in the interval $(\mu \pm \sigma)$; approximately 95.4% of the area under the curve is contained in the interval $(\mu \pm 2\sigma)$; while approximately 99.7% of the area under the curve is contained in the interval $(\mu \pm 3\sigma)$, see Figure 3.2. The effect of different means and variances can be seen in Figure 3.1(a).

The Central Limit Theorem, a key theorem in statistics, states that for a sufficiently large sample of independent and identically distributed variables, the sampling distribution of the mean is approximated by a normal distribution, no matter the shape of the population distribution. This makes it possible for other distributions to be approximated by a normal distribution, which makes it easier to solve complex problems by using the properties of the normal distribution.

Referring to Equation 3.2 and rewriting the pdf of the normal distribution we obtain:

$$f(y) = \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \quad -\infty < y < \infty;$$

which gives the canonical parameter $\theta = \mu$, the dispersion parameter $\phi = \sigma^2$, $b(\theta) = \theta^2/2$, and $c(y, \phi) = -\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$.

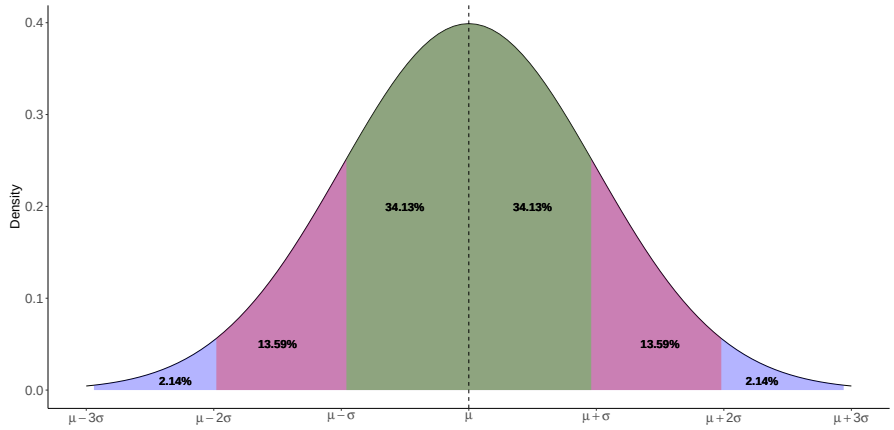


Figure 3.2: The normal distribution pdf curve. We have that 68.2% of the area under the curve falls between $\mu \pm \sigma$, around 95.4% of the area falls between $\mu \pm 2\sigma$, and approximately 99.7% of the area under the curve falls between $\mu \pm 3\sigma$.

3.2.2 The Gamma model

The gamma distribution is generally seen in examples where the outcome is strictly positive and skewed. Real life examples of the gamma distribution include the modelling of rain fall Coe and Stern (1982), faults in the equipment maintenance Van Noortwijk (2009), and insect populations Matis et al. (1992). It can be parameterised in different ways. It is common to use a shape parameter $\alpha > 0$ and a scale parameter $\beta > 0$, such that if $Y \sim \text{Gamma}(\alpha, \beta)$ the pdf is given by

$$f(y; \alpha, \beta) = \frac{y^{\alpha-1} e^{-\beta y} \beta^{\alpha}}{\Gamma(\alpha)}, \quad y > 0, \tag{3.3}$$

where $\Gamma(\cdot)$ is the gamma function, defined as

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy.$$

The gamma distribution may also be considered a generalised form of other distributions. For instance, when $\alpha = 1$, it is reduced to the exponential distribution with parameter β, i.e. a $\text{Gamma}(1, \beta)$ distribution is equivalent to $\text{Exponential}(\beta)$. Moreover, a $\text{Gamma}(\nu/2, 1/2)$ distribution is equivalent to a χ^2_ν distribution.

For modelling purposes a more useful parameterisation is to use the mean μ. We have that the expected value and variance are

$$\mathrm{E}(Y) = \frac{\alpha}{\beta}, \ \mathrm{Var}(\mathbf{Y}) = \frac{\alpha}{\beta^2}.$$

Therefore, we may reparameterise the pdf (3.3) using $\mu = \alpha/\beta$, to obtain the following exponential family pdf in canonical form:

$$f(y) = \exp\left\{\left(\frac{-y}{\mu} - \log(\mu)\right)\alpha + \alpha\log\alpha + (\alpha-1)\log y - \log(\Gamma(\alpha))\right\}, \quad y > 0.$$

It is clear from above that the canonical parameter $\theta = -1/\mu$, and the dispersion parameter $\phi = 1/\alpha$. This gives $b(\theta) = \log(-1/\theta)$, and

$$c(y,\phi) = \frac{1}{\phi}\log\left(\frac{1}{\phi}\right) + \left(\frac{1}{\phi} - 1\right)\log y + \log\left(\Gamma\left(\frac{1}{\phi}\right)\right), \quad y > 0.$$

See Figure 3.1(b) for different shapes of the gamma distribution.

3.2.3 The Inverse Gaussian model

The inverse Gaussian distribution is a continuous distribution similar to the gamma distribution, but with a sharper peak and greater skewness Folks and Chhikara (1978). It has a single mode and a long tail in the density function which helps modelling data sets with extreme values. The name inverse Gaussian is related to the fact that its cumulant generating function is the inverse of the Gaussian distribution's. The pdf of an $\mathrm{IG}(\mu, \sigma^2)$ distribution is given by

$$f(y) = \sqrt{\frac{1}{2\pi\sigma^2 y^3}}\, e^{-\frac{(y-\mu)^2}{2\mu^2\sigma^2 y}}, \quad y > 0,$$

where $\mu \in (-\infty, \infty)$ is the mean and $\phi = \sigma^2 > 0$ is the dispersion parameter. We have that $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \mu^3\sigma^2$.

We may re-write the pdf of the inverse Gaussian distribution in the canonical exponential family form as

$$f(y) = \exp\left\{\left(-\frac{y}{2\mu^2} + \frac{1}{\mu}\right)\sigma^2 - \frac{1}{2}\left(\log(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 y}\right)\right\}, \quad y > 0,$$

where we identify the canonical parameter $\theta = -1/2\mu^2$, $b(\theta) = \sqrt{-2\theta}$, and

$$c(y, \phi) = \frac{-1}{2}\left(\log 2\pi y^3\phi + \frac{1}{y\phi}\right).$$

For different shapes of the inverse Gaussian distribution, see Figure 3.1(c).

3.2.4 The Poisson model

Unlike a Normal distribution which is a continuous distribution, the Poisson distribution is a discrete probability distribution that is used to model data in the form of counts for a specified interval of time (or space). Examples of data that can be modelled using a Poisson distribution include the number of eggs laid by insects over a specified period of time, or the number of insects counted per m^2.

If Y has a Poisson distribution, we may write $Y \sim \mathrm{P}(\mu)$, and write its probability mass function (pmf) is given by

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y \in \{0, 1, 2, \ldots\}$$

where $\mu > 0$ is the mean parameter. A property of the Poisson distribution is that the mean is equal to the variance, i.e. $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$, known as the 'equi-dispersion' property. A consequence of this is that the Poisson model is not able to appropriately model data for which the variance is greater than the mean (a phenomenon known as 'over-dispersion' Hinde and Demétrio (1998a)). Real entomological data often exhibits over-dispersion, and therefore extensions of the Poisson model would be more suitable for analysis, such as the quasi-Poisson, negative binomial and Poisson-normal models Demétrio et al. (2014).

Re-writing the pmf in the canonical exponential family form we obtain:

$$f(y) = \exp(y \log \mu - \mu - \log(y!)), \quad y \in \{0, 1, 2, \ldots\}.$$

We can identify the canonical parameter $\theta = \log \mu$, as well as the dispersion parameter $\phi = 1$. Moreover, $b(\theta) = \mu$, and $c(y, \phi) = -\log(y!)$. See Figure 3.1(d) for different shapes of the Poisson pmf.

3.2.5 The Negative Binomial model

The negative binomial distribution, also known as the 'Pascal' distribution, is the distribution of the number of failures before the first $k \in \{1, 2, \ldots\}$ successes in a sequence of independent Bernoulli trials[4] with probability of success $0 \leq \pi \leq 1$. This distribution can also be expressed as a sum of k independent geometric random variables, since the geometric distribution describes the number of failures before the first success in a sequence of independent Bernoulli trials. The pmf of the negative Binomial distribution is given by

$$f(y|\mu, k) = \binom{y+k-1}{y} \left(\frac{\mu}{\mu+k}\right)^n \left(\frac{k}{\mu+k}\right)^k, \quad y \in \{0, 1, 2, \ldots\}$$

where $\mathrm{E}(Y) = \mu$ and $\phi = k^{-1}$ is the dispersion parameter. In this parameterisation the negative binomial distribution is in the EF and has a quadratic variance function, given by $\mathrm{Var}(Y) = \mu + \mu^2\phi$; we will refer to this parameterisation of the negative binomial distribution as negbin-quad.

The negative binomial distribution can also be viewed as arising from a two-stage model with a Poisson distribution where the parameter is assumed to follow a gamma distribution, reflecting additional heterogeneity over the observed counts. By considering different parameterisations of the gamma distribution we obtain different parametric forms of the negative binomial. For fixed values of the mean μ they are the same distribution but when we consider allowing the mean to vary (as in regression models) they exhibit different mean-variance behaviour. In particular, there is one variant with a linear variance function $\mathrm{Var}(Y) = \mu + \mu\psi$,

[4] A Bernoulli trial is an experiment with only two possible outcomes: 'success' or 'failure'.

referred to here as negbin-lin; but note that in this form the resulting negative binomial model is not in the EF.

Depending on the data, either parameterisation can be the best choice to accommodate the extra variability. Note that these variance functions are inflated with respect to the Poisson distribution, and we have that $\mathrm{Var}(Y) > \mathrm{E}(Y)$, with a limiting case of equi-dispersion obtained when the additional parameter ϕ, or ψ, is 0 (note that these correspond to a gammma distribution with zero variance, i.e. a degenerate constant distribution, hence the reduction to the equi-dispersed Poisson distribution). Therefore, the negative binomial distribution can be considered a one-parameter extension of the Poisson distribution for modelling overdispersed counts. See Figure 3.1(e) for examples of pmf shapes for the negbin-quad distribution.

3.2.6 The Binomial model

The number of successes out of $m \in \{0, 1, 2, \ldots\}$ independent Bernoulli trials with same probability of success $0 \leq \pi \leq 1$ follows binomial distribution, denoted as $\mathrm{Binom}(m, \pi)$. The values assumed by a binomial random variable are discrete and bounded between 0 and m, i.e. they can be referred to as 'discrete proportions'. In entomology, there are many cases where this type of data arises, such as in experiments measuring the proportion of viable eggs, sex ratios, or dose-response experiments where the focus is on mortality (or survival) of insects. The pmf of the binomial distribution is given by

$$f(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y \in \{0, 1, \ldots, m\}.$$

We have that $\mathrm{E(Y)} = m\pi$ and $\mathrm{Var(Y)} = m\pi(1-\pi)$.

Re-writing the pmf of the binomial distribution in the canonical exponential family

form we obtain

$$f(y) = \exp\left\{ y \log\left(\frac{\pi}{1-\pi}\right) + \log\binom{m}{y} - m\log(1-\pi) \right\}, \quad y \in \{0, 1, \ldots, m\}.$$

We can identify the canonical parameter $\theta = \log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi)$ and the dispersion parameter $\phi = 1$. We also have $b(\theta) = -m\log\left(\frac{1}{1+e^{\theta}}\right)$, and $c(y, \phi) = \log\binom{m}{y}$.

The binomial model is naturally under-dispersed, since $\text{Var}(Y) = m\pi(1-\pi) = \text{E}(Y)(1-\pi) < \text{E}(Y)$. In many applications, we may find that this mean-variance relationship does not hold, and the variability in the data is greater than accommodated by the standard binomial model. In such cases, extensions of the binomial model can be used, such as the quasi-binomial, beta-binomial and logistic-normal models Fatoretto et al. (2018). See Figure 3.1(f) for different shapes of the binomial distribution pmf.

3.3 Residuals

Here we provide a brief introduction to different types of residuals used when assessing goodness-of-fit and performing model selection.

Residuals can be considered as an information metric that gives an idea about how well the specified model fits the data. They are based on the deviation between fitted/predicted values from observed values. Since they can be used to detect outliers and abnormalities in the data, they are considered an integral part of exploratory analysis.

When working with the classical linear model (Eq. 3.1), different types of residuals can be used to verify the assumptions of linearity, independence, homogeneity of variances and normality, through either visual displays or formal hypothesis testing. In this model we have the fundamental decomposition of an observation

y_i into its fitted value, $\hat{y}_i$, and the residual $r_i = y_i - \hat{y}_i$ with

$$y_i = \hat{y}_i + (y_i - \hat{y}_i) = \hat{y}_i + r_i$$

and moreover the vectors $\hat{\mathbf{y}}$ and $\mathbf{r}$ are orthogonal. Hence, a very useful and basic form of diagnostic display is to plot residuals ($\mathbf{r}$) versus fitted values ($\hat{\mathbf{y}}$). This plot can help to assess whether the assumptions of variance homogeneity and linearity are met. An ideal plot would show the residuals distributed randomly around zero, with no trend (Figure 3.3(a)). When the variance is not constant, one would see changes in variability throughout the plot, such as the example in Figures 3.3(b) and 3.3(d), where the variance changes proportionately with the fitted values. When the linearity assumption is not met, the residuals versus fitted values plot will show a trend or curve, rather than points distributed randomly around zero, such as the examples in Figures 3.3(c) and 3.3(d).

Naturally, when working with other generalized linear models that are not the normal model, we would not expect the assumption of constant variance to hold, since some of the variance functions are proportional to the mean (Table 3.1). However, the residuals and the fitted values still form the basic building blocks of quantities of interest and useful displays. We now introduce the most commonly used residual types. Note that the type of residuals used for the model selection process should depend on the models considered, and the nature of the response variable.

3.3.1 Raw Residuals

The *raw residuals* (r_i) are defined as the difference between the observed data and fitted/predicted values:

$$r_i = y_i - \hat{\mu}_i \; ; \quad i = 0, 1, 2, \ldots, n \tag{3.4}$$

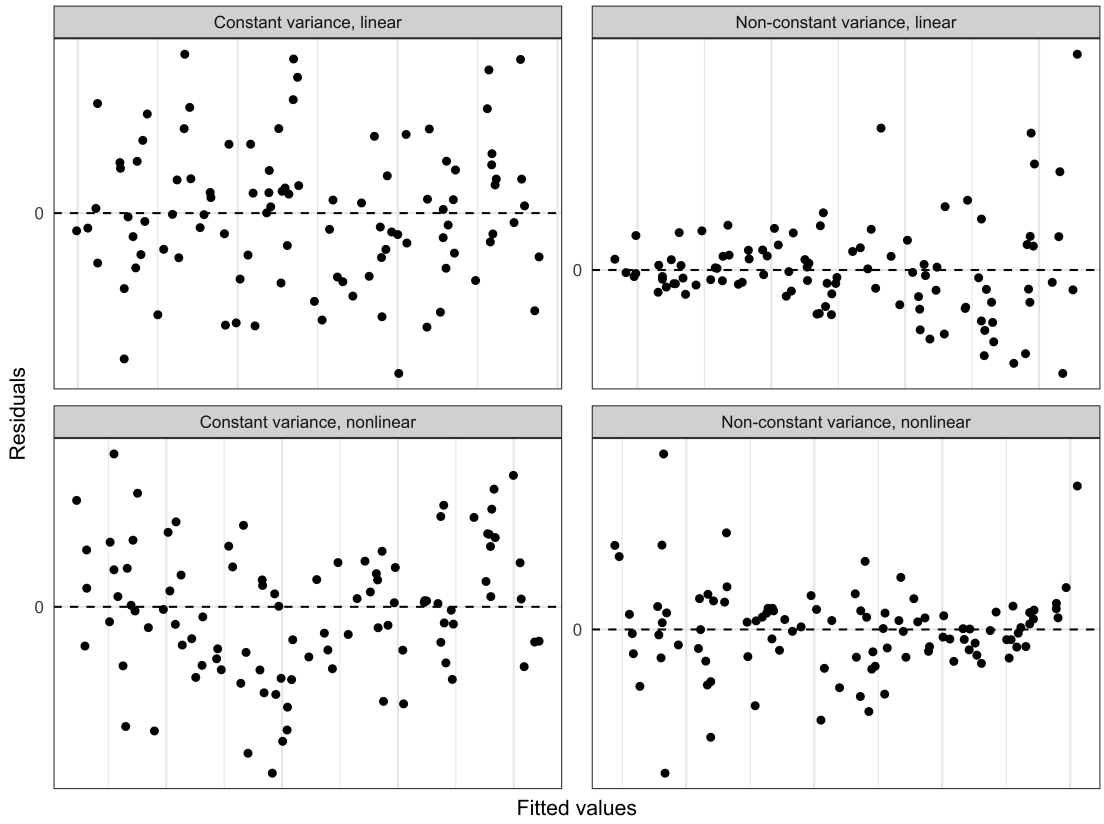


Figure 3.3: Examples of patterns expected when looking at 'residuals versus fitted values' plots for four different scenarios. (a) Constant variance and linearity assumptions are met; (b) variance is not constant, but linearity assumption is met; (c) plot shows a trend, therefore linearity assumption is not met, however the variance seems to be constant; (d) neither the constant variance nor linearity assumptions are met.

where y_i is the observed value and $\hat{\mu}_i$ is the fitted value. Large $|r_i|$ shows a higher discrepancy between the observed data and the predicted value, which may indicate that either observation y_i is an outlier under the distributional assumption or that the model does not have a good fit, if that is the case for many observations. Very small $|r_i|$ for many observations could indicate overfitting. However, this will depend on the scale of the response variable. Moreover, in classical linear regression the raw residuals should follow a normal distribution with zero mean and variance

$\sigma^2 > 0$. But in non-normal scenarios the raw residuals behave differently, and may be asymmetric, and have non-constant variance. This is why it is better and commonplace to use scaled versions of the residuals.

3.3.2 Pearson Residuals

The 'Pearson residuals' r_i^P are defined as,

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}, \tag{3.5}$$

which are the raw residuals scaled by the estimated standard deviation. This formulation addresses the problem of non-constant variance; therefore under well-fitted models r_i^P should display a constant variance behaviour when plotted against the fitted values.

The Pearson statistic X^2 is calculated by summing the squared Pearson residuals, i.e. $X^2 = \sum_{i=1}^{n}(r_i^P)^2$. It can be shown that $1/phi$ times this statistic , asymptotically, follows a χ^2 distribution with $n - p$ degrees of freedom Jørgensen (2013). This statistic can be used to test the goodness-of-fit of a GLM when the dispersion parameter ϕ is known. This is especially useful for the Poisson and binomial GLMs, for which $\phi = 1$, fixed. In the cases where ϕ is unknown Pearson residuals can be misleading and should not be used. Under a well-fitted Poisson or binomial GLM, we would expect X^2 to be similar to $n - p$. If $X^2 >> n - p$ this may be an indication that the variability in the data is larger than expected by the model, and therefore extended models that accommodate extra-variability would be more appropriate for analysis. But Pearson residuals are not ideal for spotting unusual outliers or extreme values and can be sometimes misleading for discrete outcomes and has to used with caution

3.3.3 Deviance Residuals

The *deviance residuals* r_i^D are associated with the concept of deviance D, which is a measure that considers the departure of the fitted model from the saturated model. The saturated model has as many as parameters as observations, and reproduces all observed values exactly, i.e. $\hat{\mu}_i = \hat{y}_i = y_i$. This model clearly overfits the data, however it is useful to quantify how far the fitted (or current) model is from reproducing the data exactly. We may write

$$D = 2(l^* - l),$$

where l^* and l are the maximised log-likelihoods of the saturated and current models, respectively. The deviance can be represented as the sum of the deviance measures from each data point, such that $D = \sum d_i^2$, where d_i is the i^{th} component of deviance. The deviance residual r_i^D is then given by

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}.$$

Under a well-fitting model, the distribution of the deviance residuals approximates a normal distribution, and they are a common choice for likelihood-based methods Pierce and Schafer (1986).

3.4 Half-Normal Plots with a Simulated Envelope

A quantile-quantile plot, *q-q plot*, is a graphical method used to compare the distributions of two samples by plotting the quantiles against each other. This is mainly employed in cases where we assume a distribution for a response variable and would like to check if that is a reasonable assumption. The quantiles of the theoretical distribution are plotted against the quantiles of the data and if the

distributional assumption is reasonable, the points would fall on, or close to, the identity line $y = x$ (see Figure 3.4(a)).

A normal q-q plot compares the observed data against a normal distribution. Figure 3.4 shows two q-q plots: in Figure 3.4(a) the theoretical model considered is the normal model and the data is also simulated from the normal distribution. The observed behaviour is similar to a $y = x$ plot which means the assumed model is reasonable. But from Figure 3.4(b) we conclude the assumed distribution is not a good approximation for the data.

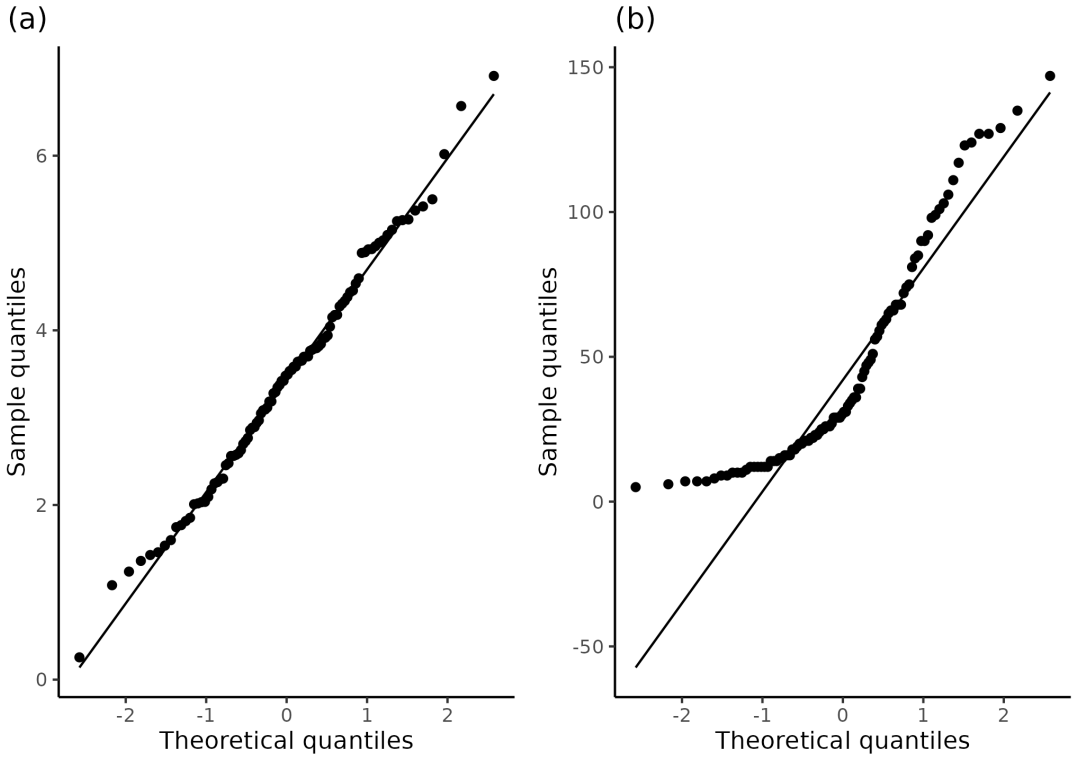


Figure 3.4: Quantile-quantile plots showing (a) agreement between an assumed distribution and the sample distribution, and (b) disagreement (i.e. the assumption is not a reasonable one).

Half-normal plots can be considered as an extension to the q-q plot where the ordered absolute value of a model diagnostic (e.g. residuals, leverage, Cook's dis-

tance, etc.) is plotted against the expected order statistics of the half normal distribution; these are particularly useful in smaller samples where the full normal plot can be rather sparse and natural sampling variation can be potentially misleading. It is based on the statistical principle that if the assumptions and the model fit are adequate, then theoretically we should be able to generate the observed values of the response variable. It is a graphical method now primarily used to identify outliers and assess distributional assumptions. The half-normal plot was orginally introduced by Daniel Daniel (1959) for the analysis of factorial experiments, especially those involving un-replicated designs. The paper also introduces 'guardrails' for giving better interpretation of the results. A major revision to this method was proposed by Zahn (1975). Since then, several alternative methods were introduced to reduce subjectivity.

If Y follows a Normal distribution, then $|Y|$ follows a half-normal distribution Zahn (1975), with pdf given by:

$$f(x) = \sqrt{\frac{2}{\pi\sigma^2}}\, e^{-\frac{y^2}{2\sigma^2}}, \quad y \geq 0$$

The expected ordered statistics of the half-normal distribution, hereby referred to as 'half-normal scores', are approximated by:

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

where i is the $i-$th order statistic, $1 \leq i \leq n$, and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function of the normal distribution de Andrade Moral et al. (2017).

The application of half-normal plots as a model selection method is implemented with an added simulated envelope as suggested by Atkinson (1985a) to aid interpretability. To construct the plot, first, model diagnostics are calculated from the

fitted model, the absolute value is taken and they are then sorted from minimum to maximum. Second, 99, or more, simulated realisations of the response variable are created from the fitted model, that is using the same model matrix and distributional assumptions and parameter values as given by the fitted model. The next step is to fit the same model to these simulated responses and re-calculate the same model diagnostics, take absolute values and sort them. The final step is to form the envelope by computing the percentiles of interest for each order statistic from the set of (99) simulated values together with the original real data one. Typically the chosen percentiles are 2.5% and 97.5%. For this case, up to 5% of the points (order statistic values) may fall outside of the envelope to indicate a well-fitted model. If much more than 5% of the points lie outside of the envelope, it means that the observed data is not a plausible realisation of the fitted model.

We present two sample half-normal plots with a simulated envelope for model residuals in Figure 3.5. In Figure 3.5(a) all the residual points fall inside the simulated envelope, which means that the model fits the data well, i.e. the data is a plausible realisation of the assumed probability distribution. In Figure 3.5(b), however, most residual points falls outside the envelope, which means that the model is not a good fit for the data.

The half-normal plot with a simulated envelope is simple to interpret, however if the estimation procedure for a model is time-consuming, it might be computationally expensive to produce.

3.5 Examples

3.5.1 Biological control of ticks

To study the efficacy of biologically controlling ticks using nematodes in grasslands, an experiment was set up in a randomised complete block design with six blocks in

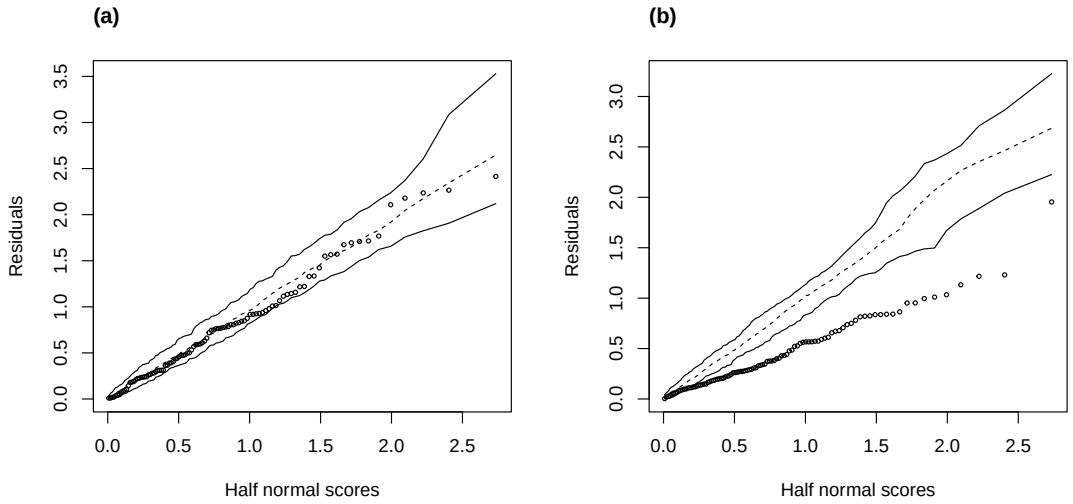


Figure 3.5: Half-normal plot with a simulated envelope for model residuals for (a) a case where the model fits the data well, and (b) a case where with poor goodness-of-fit.

a field of *Megathyrsus maximus* grass, in the state of Goiás, Brazil Filgueiras et al. (2023b). The field was divided into six groups of two plots each, totalling twelve plots. Within each group (block), one plot was treated with the entomopathogenic nematode *Heterorhabditis bacteriophara* by introducing infected *Tenebrio molitor* larvae one week before the experiment commenced, while the other plot received no treatment (control). A day before the experiment commenced, six females of the tick *Rhipicephalus microplus* were placed in each of the twelve plots. The total number of ticks in each plot was observed after 1, 2, 3, and 4 weeks (Figure 3.6)

It seems that the control plots had larger numbers of ticks when compared to the nematode-treated plots, apart from the first and second week in block 6, where over 1,000 ticks were recovered after one week of experiment. This type of behaviour occurs in field experiments involving arthropods, where population sizes and reproduction rates vary between plots. In this particular plot the ticks reproduced rapidly and their population exploded after one week. However, there is an

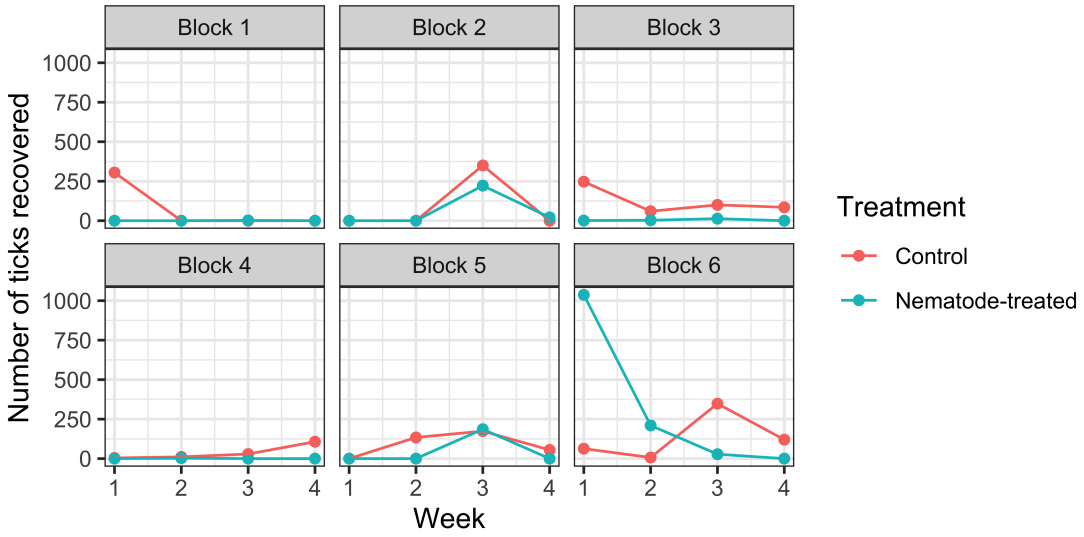


Figure 3.6: Number of ticks of the species *Rhipicephalus microplus* recovered in each of twelve plots in a field of *Megathyrsus maximus* grass. In six of these plots, the entomopathogenic nematode *Rhipicephalus microplus* was introduced a week prior to commencement of the experiment.

exponential decline after 1 week, which could reflect successful control of the tick via the introduction of the entomopathogenic nematode.

Since the response variable consists of counts, the normal distribution is not suitable for analysis. The Poisson model is a reasonable starting point, since it is suitable to analyse count data. We observe, however, that the variance is much greater than the mean for all plots over time (Table 3.2). This indicates extra-variability, or over-dispersion, and therefore extensions to the Poisson model could be more appropriate to analyse this dataset.

Firstly, we ignore the time dependence between observations made on the same plot at different occasions. We fit the normal, Poisson and negative binomial models to the ticks data, using the same linear predictor, which included the effects of block, treatment, week, and an interaction between treatment and week. The normal model assumes variance homogeneity, and a quick glance at the residuals versus

Treatment	Week	Mean	Variance
Control	1	103.2	18831.77
	2	35.3	2847.07
	3	166.8	23544.17
	4	61.2	2730.17
Nematode-treated	1	173.0	179159.20
	2	36.0	7268.40
	3	75.2	10187.37
	4	3.5	73.50

Table 3.2: Mean and variance of the number of recovered *Rhipicephalus microplus* ticks in plots treated or not with the entomopathogenic nematode *Heterorhabditis bacteriophara*, over four weeks of observation.

fitted values plot (Figure 3.7) reveals that this assumption is not met. Moreover, since the lower bound of the response variable is zero, there is an obvious lower bound for the residuals as well. The lack-of-fit of the normal model is confirmed by the half-normal plot with a simulated envelope for the raw residuals (Figure 3.8(a)). The Poisson model fit is also not adequate according to the half-normal plot in Figure 3.8(b), and this is due to the extra-variability in the data. The negative binomial model, however, seems to fit the data well ((Figure 3.8(c)).

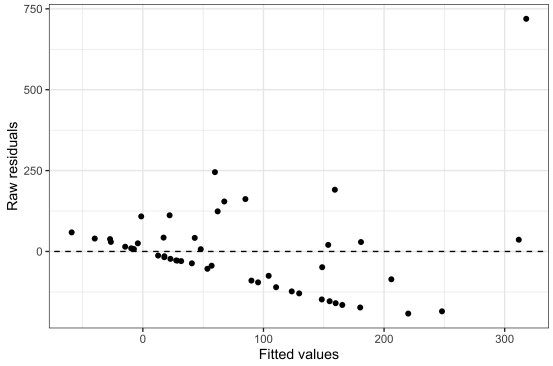


Figure 3.7: Residuals versus fitted values for the normal model fitted to the ticks data.

The importance in assessing goodness-of-fit before drawing inferential conclusions from a statistical model is enhanced when we look at the results presented in Ta-

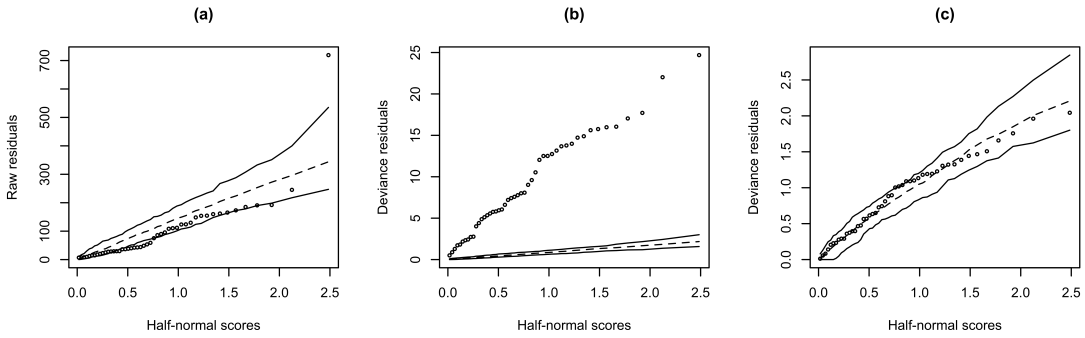


Figure 3.8: Half-normal plots with a simulated envelope for the (a) normal model using raw residuals, (b) Poisson and (c) negative binomial models using deviance residuals, fitted to the ticks data.

ble 3.3. According to the normal model, there are no significant effects of time (weeks) or treatment (the entomopathogenic nematode) on the number of ticks retrieved from the field, and therefore would lead researchers to conclude that the nematode is inefficient in controlling the tick population size. This lack of significance is due to the large variability in the data overall, which results in an overestimation of the standard errors. The Poisson model, on the other hand, detects a significant interaction between time and treatment. This is due to the assumption of equi-dispersion, which results in the underestimation of the overall variability of the data. Finally, the negative binomial model, which suitably incorporates the over-dispersion in the data, yields inferential results confirming that the nematode is indeed efficient in controlling the pest and shows the interaction to be unnecessary, indeed there is no evidence of any time effect.

3.5.2 Sustainable management of parasitic nematodes using bioagents – the 'plant height' data

The use of microbial agents as pesticides has been shown to be a more sustainable approach than chemical pesticides on agricultural pests. Silva et al. (2022b) carried

Model	Source	d.f.	Test statistic	$p-$value
Normal	Week	3, 35	1.29	0.29
	Treatment	1, 35	0.16	0.69
	Week × Treatment	3, 35	0.52	0.67
Poisson	Week	3	1444.10	< 0.01
	Treatment	1	57.14	< 0.01
	Week × Treatment	3	638.18	< 0.01
Negative binomial	Week	3	3.76	0.29
	Treatment	1	4.72	0.03
	Week × Treatment	3	4.98	0.17

Table 3.3: Test statistics and associated $p-$values for the effects in the linear predictor of the models fitted to the ticks data. F test statistics are used for the normal model and likelihood-ratio χ^2 statistics are used for the Poisson and negative binomial models.

out an experiment where they assessed the effectiveness of using a filamentous fungi of the order Hypocreales, namely two strains of *Pochonia chlamydosporia* and three strains of *Purpureocillium lilacinum* as potential bioagents against plant parasitic nematodes. Seeds of the common bean cultivar “IAC Milênio" were treated with suspensions prepared using each fungal strain, as well as a negative control that used only Arabic gum. The seeds were planted and ten potted plants were used for each treatment as observational units. After 45 days, the height of the plants was measured in cm. This experiment was repeated three times, totalling 30 plants receiving each treatment.

The box plots in Figure 3.9 show that the plant height is very homogeneous across treatments, especially for experiments 1 and 2. However, for experiment 3 it appears that plants treated with *P. lilacinum* strain ESALQ2593 were slightly taller, suggesting a significant interaction between experiments and treatments.

We fitted normal, gamma and inverse Gaussian models to the plant height data, all including the effects of experiment, treatment, and the two-way interaction between experiment and treatment in the linear predictor and response variable

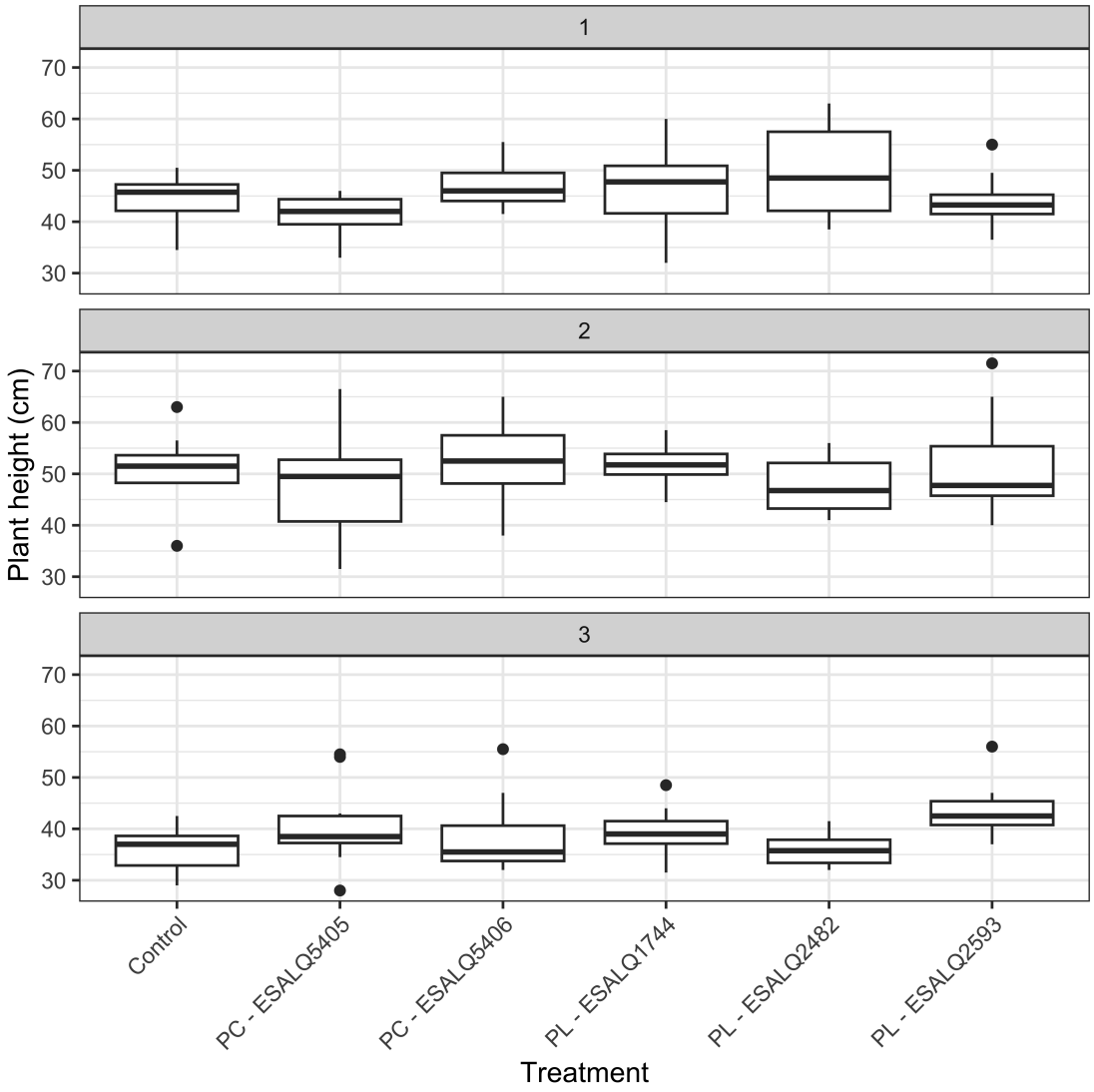


Figure 3.9: Box plots of the height for the ten plants within each experiment (each panel numbered 1, 2 and 3) according to each treatment, which included a negative control, two strains of *Pochonia chlamydosporia* (PC - ESALQ5405 and PC - ESALQ5406), and three strains of *Purpureocillium lilacinum* (PL - ESALQ1744, PL - ESALQ2482 and PL - ESALQ2593).

being the height of plant. We used the canonical link functions, i.e. identity for the normal model, inverse for the gamma model, and $g(\mu) = 1/\mu^2$ for the inverse Gaussian model. From Table 3.4, we observe that while the gamma and inverse

Gaussian models agree with respect to the significance of the two-way interaction between experiment and treatment, the normal model yields a p-value larger than 0.05 for this effect (although close to the 5% significance threshold). The half-normal plots with a simulated envelope indicate, however, that the normal model is an inadequate representation of the data (Figure 3.10), and therefore inferential results from this model should not be taken into account. On the other hand, the gamma and inverse Gaussian models seem to fit the data well.

Model	Source	d.f.	Test statistic	$p-$value
Normal	Experiment	2, 162	42.82	< 0.01
	Treatment	5, 162	1.13	0.35
	Experiment $\times$ Treatment	10, 162	1.81	0.06
Gamma	Experiment	2, 162	45.41	< 0.01
	Treatment	5, 162	1.16	0.33
	Experiment $\times$ Treatment	10, 162	2.09	0.03
Inverse Gaussian	Experiment	2, 162	46.04	< 0.01
	Treatment	5, 162	1.14	0.34
	Experiment $\times$ Treatment	10, 162	2.25	0.02

Table 3.4: F test statistics and associated $p-$values for the effects in the linear predictor of the models fitted to the plant height data.

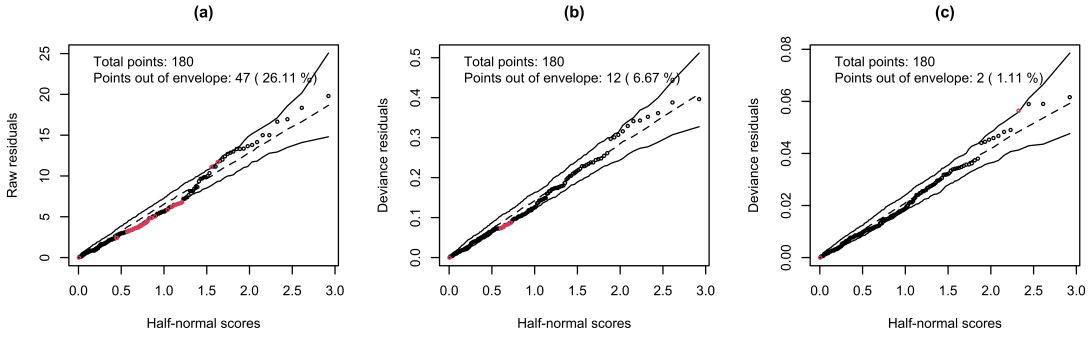


Figure 3.10: Half-normal plots with a simulated envelope for the (a) normal, (b) gamma, and (c) inverse Gaussian models fitted to the plant height data.

3.6 Discussion

It is very uncommon to see linear relationships and in this chapter, we aimed to present the generalized linear modelling framework, an extension of the classical linear linear with a specific focus on the use of diagnostic analyses to assess model goodness-of-fit, specifically through the use of the half-normal plot with a simulated envelope. We demonstrated that the normal model is not the most suitable option for analysis of discrete data, which is commonly found in entomological studies. In terms of software, although we used R throughout the chapter, there are other implementations of the models and techniques presented here through SPSS, Python, SAS, among others. The focus here has been on a single response variable; multivariate extensions to jointly model responses of interest are more complicated, and subject of active research.

CHAPTER 4

A goodness-of-fit diagnostic for count data derived from half-normal plots

Traditional methods of model diagnostics may include a plethora of graphical techniques based on residual analysis, as well as formal tests (e.g. Shapiro-Wilk test for normality and Bartlett test for homogeneity of variance). Goodness-of-fit diagnostics have also been extended to other quantitative metrics, such as information criteria based on a model's likelihood. In this chapter we derive a new distance metric based on the half-normal plot with a simulation envelope, a graphical model evaluation method, and investigate its properties through simulation studies. One advantage of the proposed metric is that it allows for the comparison between models that do and do not have a full likelihood. This newly introduced distance metric quantitatively encompasses the model evaluation principles and removes the subjective bias when closely related models are involved. We validate the technique by means of an extensive simulation study carried out using count data, and illustrate with a case study.

This chapter is currently under review at Journal of Statistical Computation and Simulation, and can be found as a pre-print at Darshana Jayakumari, Jochen Einbeck, John Hinde, Julien Mainguy, Rafael de Andrade Moral. "A goodness-of-fit diagnostic for count data derived from half-normal plots with a simulated envelope." arXiv:2405.05121v1

4.1 Introduction

Analyses of counts is ubiquitous to many research disciplines, being applied to a broad range of applied areas; for instance, the number of patients admitted to a hospital in a single day (Du et al., 2012), the progeny of insects in a biological control experiment in entomology (Borges, 2013), or the number of different animal species in a particular area in ecology (Cunningham and Lindenmayer, 2005). A first assumption when modelling count data typically involves the Poisson distribution (Hilbe, 2014). This single-parameter distribution only accounts for equidispersion and this restrictive property may not properly accommodate a count variable's mean-variance relationship in practical situations (Richards, 2008). It is common practice, therefore, to use extensions of the Poisson distribution that allow for more flexible mean-variance modelling, accommodating over- or under-dispersion Brooks et al. (2019), as well as excess zero counts (Hilbe and Greene, 2007).

One of the goals of fitting a model to data is to carry out inference. However, for inference about a fitted model to be reliable, goodness-of-fit tests and diagnostic analyses should be carried out to ensure that the model represents an adequate fit to the data (Ding et al., 2018). For example, residual plots can be constructed by plotting a function of the residuals, typically a scaled version of the ordinary residuals, against the predictors or fitted values (Tsai et al., 1998). Once a model is deemed to be adequate, the inferential process typically involves hypothesis

testing to compare nested models (through, e.g., likelihood-ratio tests) and assess the effects of covariates, however non-nested model selection may also be carried out by comparing measures of out-of-sample predictive performance, which include information criteria (Anderson and Burnham, 2004).

Likelihood-based methods are omnipresent in statistical analyses involving parametric models. However, quasi-likelihood estimation of marginal models also represents a useful approach when the inferential objective lies in modelling the mean of the data under an assumed variance-covariance structure. Examples of this approach include quasi-Poisson and quasi-binomial models where the underlying base variance function is scaled by a dispersion parameter $\phi > 0$, thus accommodating over- or under-dispersion (Ver Hoef and Boveng, 2007; Consul, 1990). This is also applicable to generalized estimating equations (GEE), which allow for the accommodation of complex dependencies in the data through prior specification of the variance-covariance structure (Zeger et al., 1988), and multivariate covariance generalized linear models (McGLM), which represent an extension to the GEE approach and allow the joint modelling of multiple responses (Bonat and Jørgensen, 2016).

The marginal modelling frameworks pose challenges in terms of nested and non-nested model comparisons due to the impossibility of calculating a full likelihood measure to be used either in direct hypothesis tests, or to compute likelihood-based information criteria. This makes comparing similar, but separate, models a difficult task, and in this context graphical methods represent a complementary approach when attempting to assess model fit. One graphical technique that can be used to assess empirically whether an observed data sample is a plausible realisation of a fitted model is the half-normal plot with a simulation envelope Hinde and Demétrio (1998b). The envelope indicates the expected variability under the assumed model

and allows an assessment of the appropriateness of the model for the data at hand, indicated by the observed diagnostic half-normal plot mostly lying withing the envelope. It can be used to compare, for instance, the fit of Poisson and quasi-Poisson models, even though a full likelihood can be computed for the former model but not the latter (Moral et al., 2017).

In this chapter we address these model comparison issues by introducing a goodness-of-fit metric based on distances calculated from half-normal plots with a simulated envelope. We review the graphical technique, introduce our proposed distance metric, and describe simulation studies to explore the approach in Section 4.2. We present the results from the simulation studies in Section 4.3, and show the utility of the proposed approach using two case studies in Section 4.4. Finally, we provide a discussion and conclusions in Sections 4.5 and 4.6, respectively.

4.2 Methods

4.2.1 Half-normal plots with a simulated envelope

A QQ-plot is a graphical method used in ascertaining the distribution of a sample by plotting the ordered sample quantiles versus a particular distribution's theoretical quantiles (Lodder and Hieftje, 1988). A normal QQ-plot is useful to identify departures from normality. A half-normal plot is similar to the normal QQ-plot, but it plots the ordered absolute values of the sample against the expected order statistics of the half-normal distribution instead of the normal; this approach is especially useful with smaller datasets. For a sample of size n, the expected half-normal order statistics can be approximated by

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right), \quad i=1,\ldots,n,$$

where Φ^{-1} is the quantile function of the standard normal distribution.

Atkinson (1985b) proposed the use of half-normal plots with any model-based diagnostic statistics, as well as the addition of a simulated envelope to highlight departures from what would be expected under the fitted model. The envelope is such that, if the observed data is a plausible realisation of a fitted model, for most diagnostic measures the sample values would fall within the bounds of the simulated envelope. It does not constitute a hypothesis testing procedure, however it is useful as an empirical method to detect outliers, poor goodness-of-fit, and over- or under-dispersion when analysing counts or discrete proportions (Hinde and Demétrio, 1998b), depending on the particular diagnostic statistic used. In the cases of counts and proportions, for over dispersion the points falls systematically above envelope and for underdispersion the points falls systematically below the envelope.

The algorithm for constructing a $100(1-\alpha)\%$ simulated envelope is as follows:

(i) Fit a model to the data and compute the ordered absolute values of a chosen diagnostic statistic (e.g., Pearson residuals, Cook's distances, or leverage values).

(ii) Simulate different samples from the fitted model, using the same model matrix and the distribution and parameter estimates from the fit in step *(i)*. Atkinson (1985b) suggests performing 19 simulations, but smoother envelopes are obtained with more simulated samples. The `hnp` package for R uses 99 simulations as the default (Moral et al., 2017).

(iii) Re-fit the model to each simulated sample and compute, for each fit, the ordered absolute values of the same diagnostic statistic used in step *(i)*.

(iv) Compute the $100\alpha/2$ and $100(1-\alpha/2)$ percentiles over the set of simulated model diagnostics for each order statistic to form the lower and upper bounds

of the envelope, respectively. Typically, we take $\alpha = 0.05$.

4.2.2 A distance measure derived from a half-normal plot with simulated envelope

For illustration, Figure 4.1 displays three half-normal plots with a simulated envelope from different models fitted to data simulated from a negative binomial model with a quadratic variance function (NB-quad), i.e. from an overdispersed count model. All simulation studies and case studies discussed here use Pearson residuals for the production of the half-normal plots, however any type of diagnostic can be used in practice. The points painted in red are the ones that are falling outside the envelope. The first panel (a) shows the results from the true model, and 95% all sample residuals are expected to fall within the simulated envelope. Fitting a Poisson model (which assumes that the variance is equal to the mean), the extra variability (over-dispersion) in the data cannot be accommodated, which is evident from panel (b), with most of the sample residuals falling outside of the envelope. Panel (c) is the half-normal plot for a quasi-Poisson model (variance $\propto$ mean), and while this model can account for some of the overdispersion, it is clear that the linear variance function cannot fully account for the quadratic variance function of the underlying model and a considerable number of points falls outside the envelope.

Graphical goodness-of-fit assessment and model selection can be challenging in this context when closely related models are considered. Moreover, likelihood-based procedures and associated quantities are not available for marginal models such as the quasi-Poisson, although pseudo-likelihoods are sometimes used, see Bonat and Jørgensen (2016). However, it is possible to derive an objective metric d for model selection based on the sample residual values and the simulated envelope. A natural way of measuring how far the observed residuals are from the expected

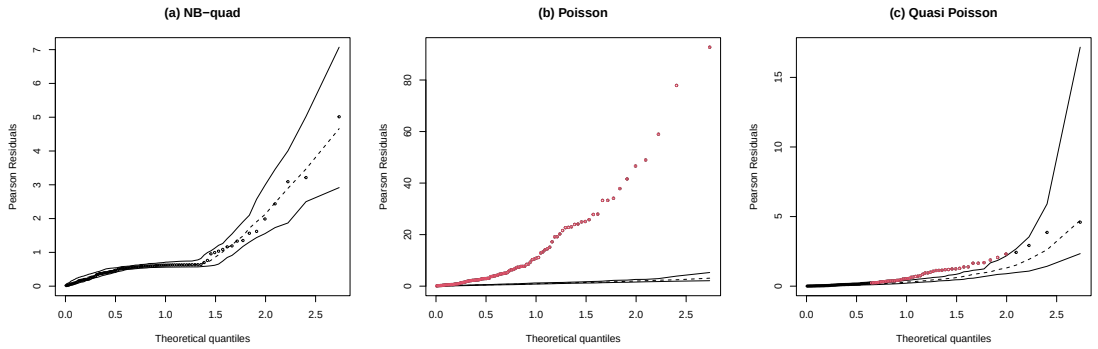


Figure 4.1: Half-normal normal plots with a simulated envelope for three different models fitted to data simulated from a negative binomial model with a quadratic variance function (NB-quad).

median behaviour is to use a quantity d given by

$$d = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} |r_i - m_i|^p, \tag{4.1}$$

where r_i is the i^{th} ordered absolute residual, m_i is the median of the simulated envelope corresponding to the i^{th} residual, and p denotes the power used. Here, we restrict to $p \in \{1, 2\}$, with $p = 1$ corresponding to the absolute difference between the residual point and the median (corresponding to the L_1 norm), and $p = 2$ to the squared Euclidean distance.

4.2.3 Simulation studies

To study the properties of the proposed metric (4.1), we performed simulation studies based on different underlying count distributions, including both overdispersion and zero-inflation. As we were interested in testing many distributions and the whole exercise was computationally intensive, as each simulation the calculation of d requires further simulations for the formation of the simulated envelope, to speed up computation time we used a single covariate model, i.e.

$$\begin{aligned} Y_i &\sim \mathcal{D}(\mu_i, \phi, \nu) \\ g(\mu_i) &= \beta_0 + \beta_1 x_i \end{aligned}$$

where $i = 1, \ldots, n$ indexes the sample, $g(\cdot)$ is a link function (e.g. the log link when $\mathcal{D}$ is Poisson) for the mean μ_i, ϕ is a dispersion parameter, and ν is a zero-inflation parameter for the parent distribution $\mathcal{D}$. We generated the values of the covariate x_i using a standard normal distribution. We used three different sample sizes $n \in \{20, 50, 100\}$ and 6 different parent distributions $\mathcal{D} \in \{\text{Poisson}, \text{Quasi-Poisson}, \text{NB-lin}, \text{NB-quad}, \text{ZIP}, \text{ZINB}\}$, where NB-lin represents a negative binomial model with a linear variance function $V(\mu_i) = \mu_i + \phi\mu_i$, NB-quad is the negative binomial with a quadratic variance function $V(\mu_i) = \mu_i + \zeta\mu_i^2$; where $\zeta = \dfrac{1}{\phi}$, ZIP is the zero-inflated Poisson model, and ZINB is the zero-inflated negative binomial model (Ver Hoef and Boveng, 2007; Yau et al., 2003; Lambert, 1992). For the Poisson and ZIP models the dispersion is fixed, $\phi = 1$; for the other models we used $\phi \in \{0.5, 7\}$ to introduce scenarios of weak and strong overdispersion, respectively. For the non zero-inflated models, $\nu = 0$, i.e. no zero-inflation; for the ZIP and ZINB models we used $\nu \in \{0.2, 0.6\}$ to introduce weak (20%) and strong (60%) zero-inflation. Briefly, the zero inflation model is a mixture model that consists of a count data model and a point mass function at zero. The zero counts can arise from both the count data model and the point mass function. The count data models can be Poisson or a Negative binomial model with quadratic variance function, referred to as ZIP and ZINB, respectively.

We performed $1,000$ simulations for each scenario. For each simulation we generated a response variable $\mathbf{y}$ from the parent distribution, and fitted all 6 models under consideration. For each model fit, we produced a half-normal plot with a simulated envelope based on Pearson residuals:

$$r_i^P = \frac{y_i - \hat{y}_i}{\sqrt{V(\hat{\mu}_i)}}$$

and calculated the distance metric (4.1), with $r_i \equiv r_i^P$, using $p \in \{1, 2\}$.

All computations were performed in R (R Core Team, 2024). We used the `glm` func-

tion to fit the Poisson and quasi-Poisson models, and the `glm.nb` function from package MASS (Venables and Ripley, 2002) to fit the NB-quad model, `gamlss` function of package gamlss (Rigby and Stasinopoulos, 2005) to fit the NB-lin model, and `zeroinfl` function of package pscl (Jackman, 2020) to fit the ZIP and ZINB models. Half-normal plots were generated using the package hnp (Moral et al., 2017) with envelopes based on 99 simulations.

All code used to produce the simulations is available at `https://github.com/DARSHANAJAYA/Goodness-of-fit-Distance-metric.git`.

4.3 Results

Given the simulation study design, the results are divided into twelve experiments based on the different parent model considered. Results are presented in figures consisting of three plots chosen to illustrate the relevance of the proposed distance goodness-of-fit metric. Panel (a) shows a boxplot of the proposed distance metric (4.1) on the log scale for the different fitted models. Figure 4.2 to Figure 4.4 include four fitted models (Poisson, quasi-Poisson, NB-lin, and NB-quad), while Figures 4.5 and 4.A.4 have six fitted models (Poisson, quasi-Poisson, NB-lin, NB-quad, ZIP, and ZINB). Panel (b) shows a bar plot illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run, i.e. points towards a particular model. Panel (c) shows a barplot representing the number of times a particular fitted model has the smallest BIC value. Since the BIC cannot be calculated for models estimated via quasi-likelihood methods, the quasi-Poisson model is omitted from this panel. All the plots are faceted across two variables: sample size and the value of p. The best performing model is expected to give the minimal value for the computed distance metric in the case of panel (a) and highest frequency level for panels (b) and (c). It should be noted that this methodology could be performed using a training-test

data split to assess performance with unseen data; this is the subject of future work. For illustration purposes we have chosen to only report five experiments in this section; for the results from the other experiments please refer to Appendix S1.

Simulation experiment 1

This simulation experiment assessed the effectiveness of the distance metric when the parent model is an equidispersion one (Poisson). The inherent randomness of the simulation causes the dataset to be not exactly equidispersed, but either slightly overdispersed or underdispersed. While Figure 4.2(a) and Figure 4.2(b) show that the quasi-Poisson model performs better than the competitor models with the lowest distance value, although the BIC favours the (true) Poisson model (Figure 4.2(c)) in experiment 1.

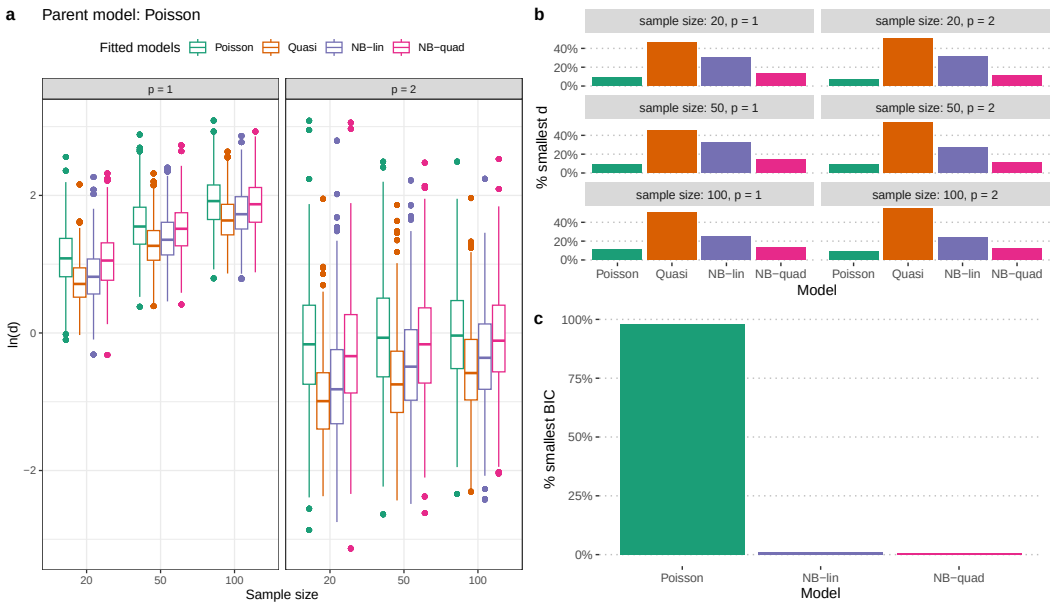


Figure 4.2: Figures generated when the parent model is the Poisson.

Simulation experiment 2

This experiment portrays the results when the parent model under consideration is one that can accommodate overdispersion, a negative binomial model with a linear variance function (NB-lin). The dispersion parameter chosen was 0.5 and is deemed a case of mild overdispersion. The results show that the quasi-Poisson and NB-lin perform the best, as is expected since both of these models have a linear variance function. Figure 4.3(a) and (b) show that using the absolute differences (equivalent to the L_1 norm) improves the performance of the distance metric in identifying the true parent model. In Figure 4.3(c) BIC favours the Poisson model as the best-fitting one.

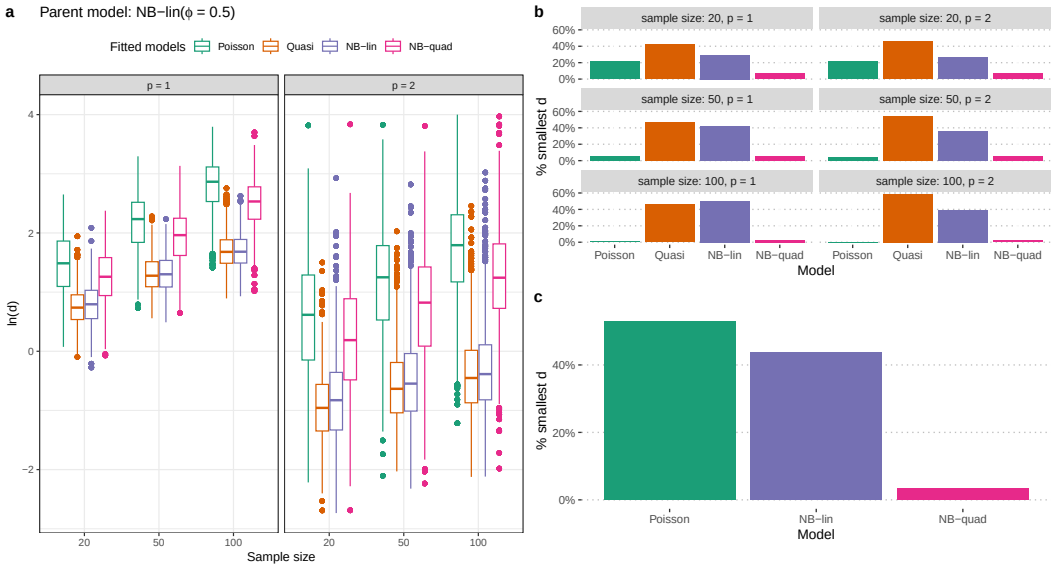


Figure 4.3: Figures generated when the parent model is the NB-lin with a dispersion parameter value of 0.5.

Simulation experiment 3

This experiment illustrates the results when the parent model is a GLM with a quadratic variance function (NB-quad), a model that can accommodate overdispersion. In this specific case, the dispersion parameter value is chosen to be 7 and

is considered an instance of strong overdispersion. Figure 4.4(a) and (b) show that the proposed goodness-of-fit diagnostic picks the best model as the parent model at sample sizes of 50 and 100, which is in agreement with the BIC (Figure 4.4(c)).

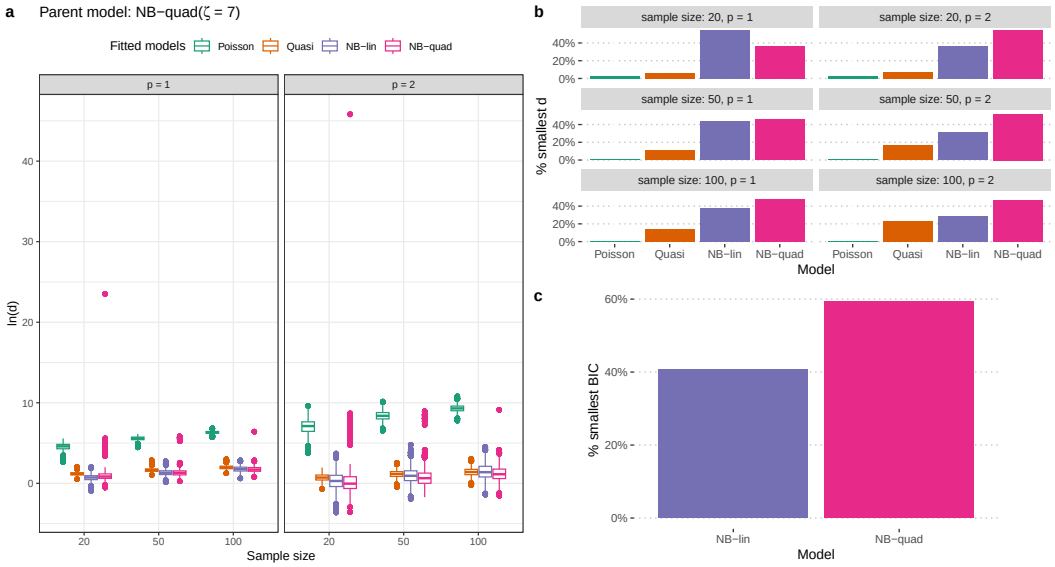


Figure 4.4: Figures generated when the parent model is the NB-quad with a dispersion parameter value of 7.

Simulation experiment 4

Simulation experiment 4 involves a ZIP model as the parent model with a zero-inflation parameter value of 0.2, a low level of zero-inflation. The results in Figure 4.5(a) and (b) show that the distance metric selects both the ZIP and ZINB models as the best performing models. This likely occurs due to the randomness in the simulated datasets where an inherent overdispersion can occur by chance in the simulation. In Figure 4.5(c), BIC favours the parent model, rejecting the additional complexity of the ZINB model.

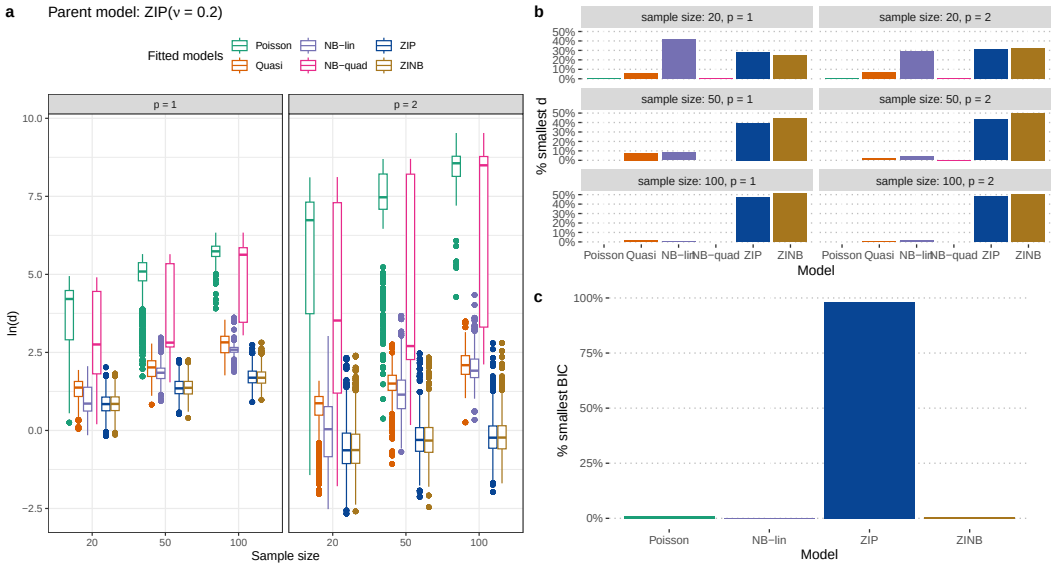


Figure 4.5: Figures generated when the parent model is the ZIP with a zero inflation factor value of 0.2.

Simulation experiment 5

This experiment uses a ZINB model as the parent model with a dispersion parameter value of 0.5, and zero inflation parameter value of 0.6, which provides a scenario where the parent model is zero-inflated negative binomial with mild overdispersion and high zero-inflation. Figures 4.6(a) and (b) show that the distance metric shows better performance in selecting the true model when the squared Euclidean distance is considered. In Figure 4.6(c), BIC favours the parent model as the true model in all instances.

4.4 Case Studies

We now present the analysis of two case studies using the distance metric proposed in equation (4.1) to aid goodness-of-fit assessment and model selection. All R code and data are available at https://github.com/DARSHANAJAYA/Goodness-of-fit-Distance-metric.git.

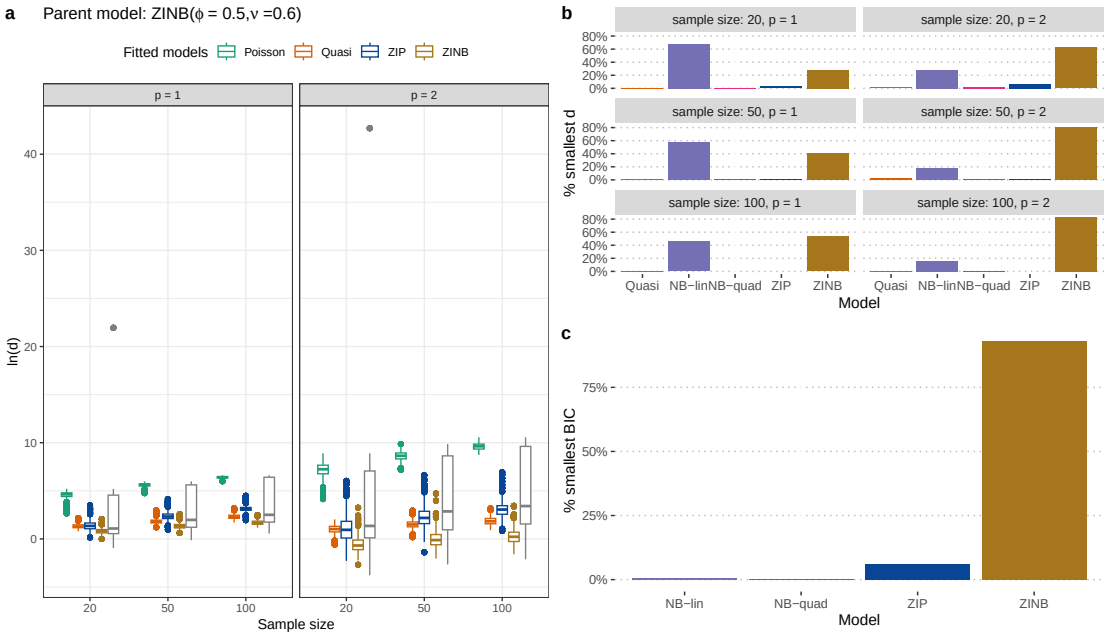


Figure 4.6: Figures generated when the parent model is ZINB with a with a zero inflation factor value of 0.6 and a dispersion parameter value of 0.5.

4.4.1 Spider data

This dataset is from the paper by (Smeenk-Enserink and Van der Aart, 1974) and is also included in the R package mvabund (Wang et al., 2012). We considered the count of the hunting spiders from the species *Alopecosa accentuata* that were caught in 100 pitfall traps (1-m radius) over a period of 60 weeks, taking the soil dry mass as a covariate. The dataset consists of two columns and 28 rows. The response variable considered is the count of the spiders and the covariate is the corresponsing soil dry mass, and was $\log(x+1)$ transformed prior to model fitting. We fitted the Poisson, quasi-Poisson, NB-lin, NB-quad, ZIP, and ZINB models. The half-normal plots corresponding to the fitted models were constructed 100 times, of which one is presented for each model in Figure 4.7. The process of producing the half-normal plot 100 times is not necessary when employing this technique, however we do so here to understand the degree of uncertainty of the distance metric in this case

study. The median of the distance metric was estimated from the 100 iterations and was used to assess the fit of the models considered (Table 4.1). The distance metric favours NB-Lin for the squared Euclidean distance and ZINB for the L_1 norm and this is explained as the estimated zero inflation parameter of the ZINB model is 2.75×10^{-5}, which can be considered meaningless and so close to no zero-inflation. The principle of parsimony would favour NB-lin, and thus the proposed distance metric shows that NB-lin captures the patterns in the data better than the other models considered, which is in line with the BIC values. The estimates of the soil mass from the summary of the NB-lin indicated that there is a higher number of hunting spiders associated with a lower level of soil dry mass.

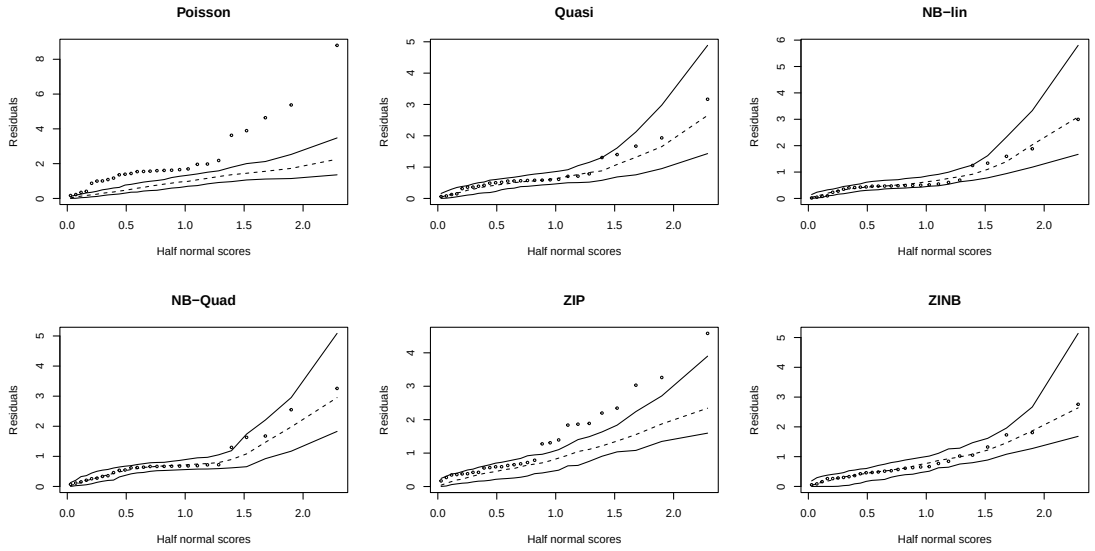


Figure 4.7: Spider data: half-normal plots with a simulated envelope for the Poisson, quasi-Poisson, NB-lin, NB-quad, ZIP and ZINB models.

4.4.2 Walleye data

In this case study we used a subset of the catch curve data analysed by (Mainguy and Moral, 2021). Catch-curve analyses relies on age frequencies to estimate the instantaneous mortality of fish. We looked at the walleye (*Sander vitreus*) gillnet

	p = 1			p = 2			BIC
	Median	**IQR**	**SD**	**Median**	**IQR**	**SD**	
Poisson	34.53	0.42	0.34	91.06	2.01	1.52	246.18
Quasi Poisson	2.66	0.29	0.22	0.71	0.19	0.16	-
NB - lin	2.20	0.17	0.12	0.33	.04	0.05	141.83
NB - quad	2.95	0.17	0.15	1.13	0.14	0.11	148.82
ZIP	13.16	0.34	0.28	14.21	0.64	0.43	194.09
ZINB	1.29	0.16	0.11	0.14	0.04	0.03	141.83

Table 4.1: Median, interquartile range(IQR) and standard deviation (SD) of the distance metric (eq. 4.1) calculated with $p = 1$ and $p = 2$, obtained from 100 half-normal plots generated for six different models fitted to the spider data, and associated BIC values.

survey data from the year 2012 from the Baskatong reservoir, Québec, Canada. The response variable considered is the count of fish which is modelled according to age fitted as a predictor variable, such that the rate at which counts decrease with age can be used to estimate mortality. Here we fitted the same six models as for the analysis of the spider data and produced 100 half-normal plots with a simulated envelope for each model fit (one is displayed for each model in Figure 4.8). The median distance metric is shown in Table 4.2. Using the squared Euclidean distance, the distance metric favours the NB-quad model, whereas for the L_1 norm it favours the ZINB model, with the NB-quad ranked closely. This result is not unexpected since if the NB-quad provides a good fit, then the ZINB also should even in the case zero-inflation is not present. For the walleye data, the zero-inflation parameter is estimated as 1.42×10^{-6}, which reflects very low or no zero-inflation. The BIC favours the NB quad model.The inclusion of the extra zero-inflation parameter was penalised by the BIC since it did not sufficiently explain additional variability in the data, and as such was the ZINB-quad was not selected over a simpler model. This is in line with the findings from (Mainguy and Moral, 2021), indicating the extra variability in the walleye data was best accommodated by the NB-quad model. The proposed distance metric has an added advantage of

	p = 1			p = 2			BIC
	Median	**IQR**	**SD**	**Median**	**IQR**	**SD**	
Poisson	16.32	0.37	0.28	53.90	2.25	1.65	164.76
Quasi Poisson	7.12	0.36	0.25	12.02	0.98	0.77	-
NB - lin	1.21	0.25	0.19	0.13	0.08	0.07	104.99
NB - quad	1.098	0.17	0.13	0.13	0.06	0.05	93.78
ZIP	18.04	0.35	0.27	60.36	1.43	1.09	171.53
ZINB	2.57	0.31	0.23	0.35	0.09	0.07	101.35

Table 4.2: Median, interquartile range(IQR) and standard deviation (SD) of the distance metric (eq. 4.1) calculated with $p = 1$ and $p = 2$, obtained from 100 half-normal plots generated for six different models fitted to the walleye data, and associated BIC values.

avoiding subjective bias that is present in the graphical model selection method apparent from this case study.

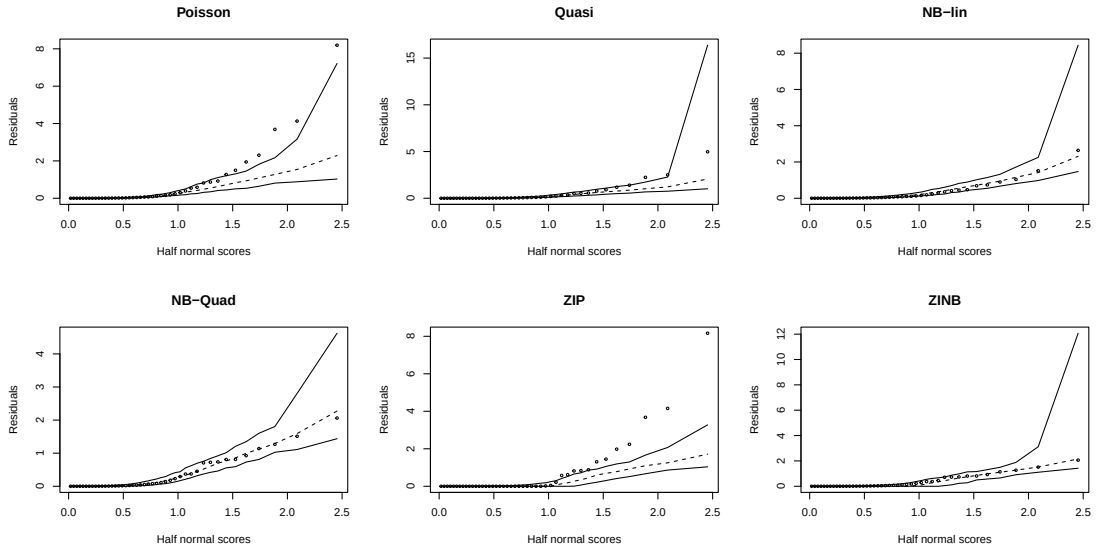


Figure 4.8: Walleye data: half-normal plots with a simulated envelope for the Poisson, quasi-Poisson, NB-lin, NB-quad, ZIP and ZINB models.

4.5 Discussion

This chapter focussed on defining a quantitative summary to the qualitative graphical model selection and goodness-of-fit assessment method known as a half-normal

plot with a simulated envelope. A simple and effective distance metric was introduced that could capture how far the observed data deviates from the expected behaviour according to the fitted model. We considered two forms of distances through the power coefficient p and found that they were useful in differentiating the fit of count data models with a linear variance function (quasi-Poisson or NB-lin) and those with a quadratic variance function (NB-quad, ZINB). We carried out further simulation studies to understand the effectiveness of adding a measure of envelope width in the distance (4.1), as well as an extra penalty term when a residual falls outside the envelope. However, these seemed to have no real impact on the final conclusions, and therefore were not used in our formulation.It should also be noted that large datasets may impact the results when simulated envelopes are used in the context of GLMs. See the appendix for the alternative formulation and results from the additional simulation studies.

When overdispersed counts are analysed, which corresponds to a commonly-encountered situation in ecology (Richards, 2008), determining whether such extra variation should be modelled as a linear or quadratic function of the mean is not trivial (Ver Hoef and Boveng, 2007). When only applying a correction factor to the standard error through a quasi-Poisson approach (Knape, 2016), or quasi-binomial one when modelling overdispersed discrete proportions instead (Bolnick et al., 2014), this may sometimes be sufficient to account for the detected overdispersion and then use the quasi-AIC to identify the best-fitting model. However, using a scaling parameter to directly model how data are dispersed, such as the one used with the NB-quad and NB-lin, may offer a better approach than relying on the former (Hilbe, 2014). With the proposed distanced-based method described in this chapter, identifying which of overdispersed model extensions provides a better fit to the data can now be assessed on a similar basis from not only an adequacy (i.e., goodness-of-fit) perspective, but also to help with model selection to identify the

model that should be retained for inferential purposes. As such, the proposed metric which however only assesses the fit without accounting for model complexity, can then be complemented with other commonly-used ones, such an information criteria like the BIC that was used in this study as a complement to further fine-tune the model selection process. It should also be noted that large datasets may impact the results when simulated envelopes are used in the context of GLMs.

Future work arising from this study would be to explore the impact of a mis-specified link function and missing covariates on the methodology and the behaviour of the proposed distance metric. Another area of research that could be be further investigated are the response patterns when mixed models are included in the study. Worm plots are diagnostic plots similar to half normal plots but also helpful to identify the skewness in the data. This makes worm plots a good choice for the implementation of distance metrics in the future.

4.6 Conclusion

The proposed distance metric framework provides a competitive goodness-of-fit diagnostic to check the adequacy of count data models. This was validated by a comprehensive simulation study that showed that our proposed metric is comparable, or in some cases superior, to BIC in identifying the true model that generated the data. It represents, therefore, a complementary approach in goodness-of-fit assessment that adds an objective measure of fit when comparing half-normal plots with a simulated envelope from competing models, especially when results are similar and is particularly useful when likelihood-based methods are not available.

Acknowledgments

This publication has emanated from research supported by a grant from Science Foundation Ireland, under Grant number 18/CRT/6049.

Appendix

4.A Simulation results

This appendix displays the figures related to the simulation scenarios not presented in the main body of text of the chapter.

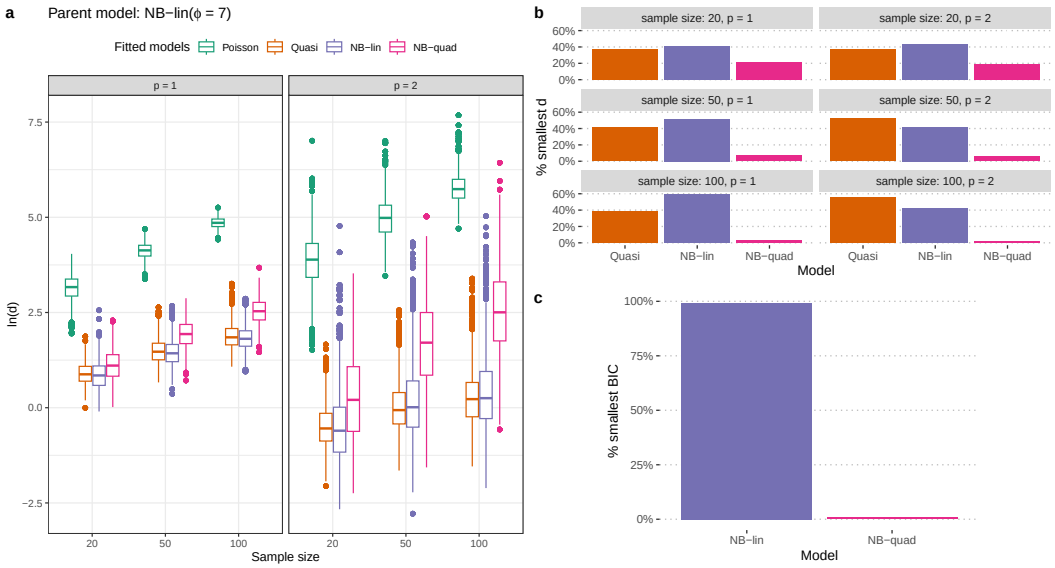


Figure 4.A.1: Figures generated when parent model is NB-lin with a with an dispersion parameter value of 7. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted model. Panel (b) shows the bar plot illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation.

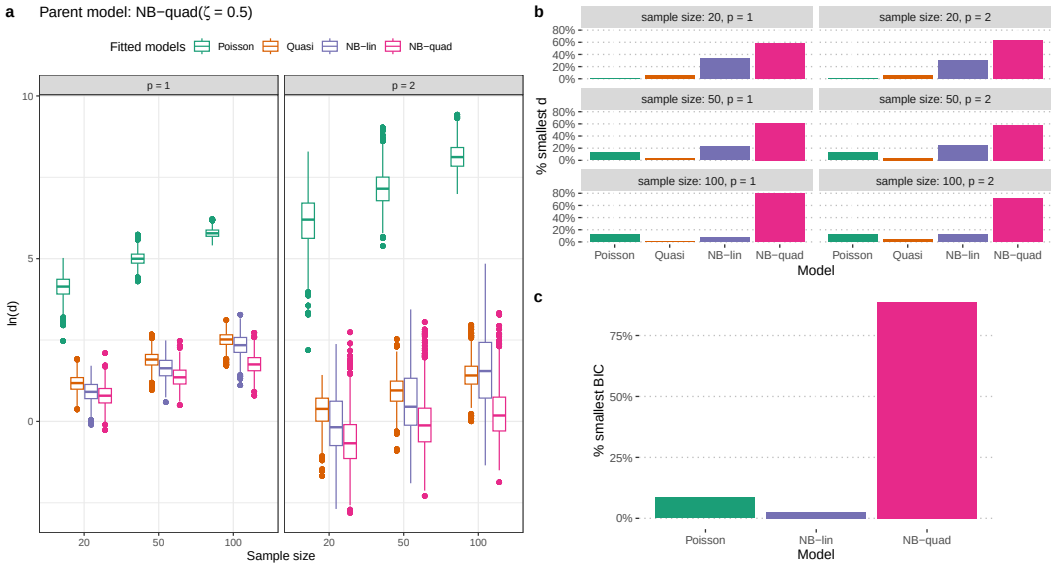


Figure 4.A.2: Figures generated when the parent model is NB-quad with a dispersion parameter value of 0.5. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

The under dispersed data was generated by forcing the Poisson distribution to halve the values produced by rpois function from the stats package to induce the deisred mean-variance relationship. Among the models considered quasi- Poisson model realises the underdispersed data and the distance metric also shows the same.

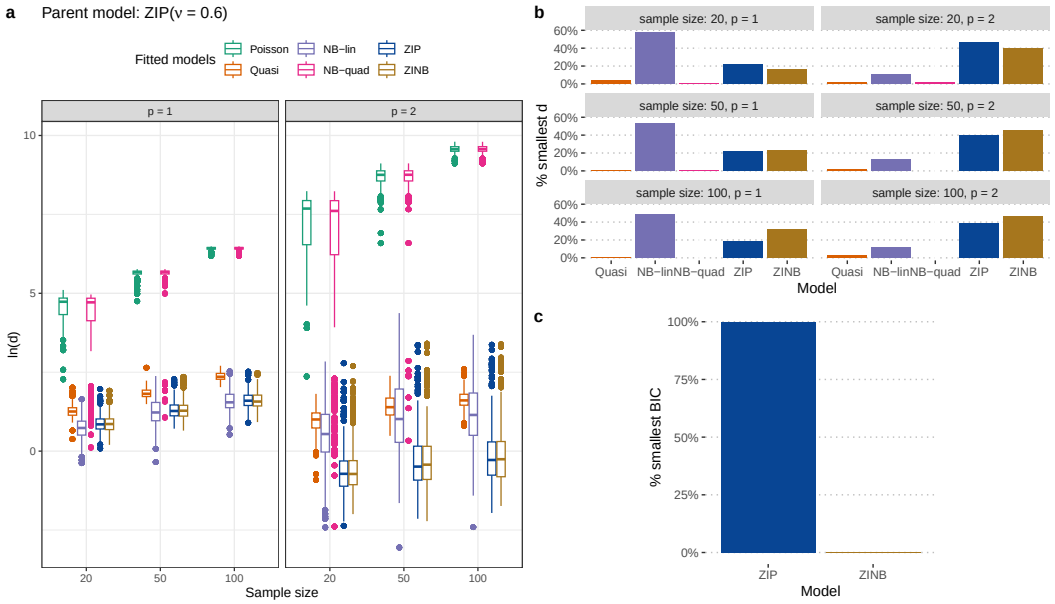


Figure 4.A.3: Figures generated when parent model is ZIP with a with zero inflation factor value of 0.6. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

4.B Reduced simulation setup using response residuals

The appendix illustrates the results from a simulation setup consisted of using ordinary residuals. We did a reduced simulation with 500 simulation to evaluate the performance of using response residuals in the proposed distance metric.

4.C Simulation results using AIC

This appendix illustrates the plot results when another information criterion like AIC is used. It has been shown the results are comparable to BIC and also the proposed distance metric.

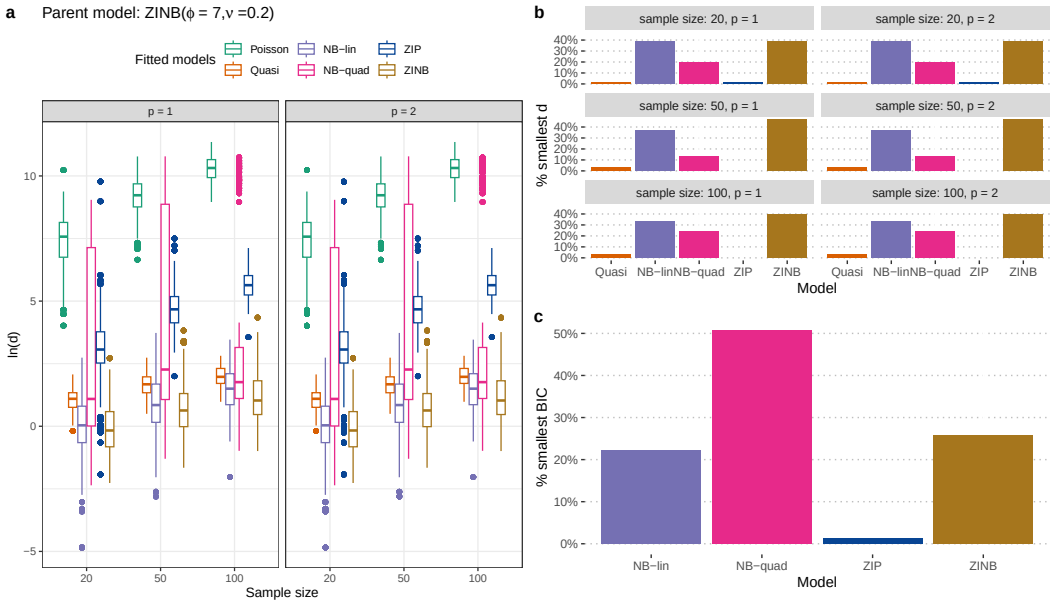


Figure 4.A.4: Figures generated when parent model is ZINB with a with zero inflation factor value of 0.2 and a dispersion parameter value of 7.Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

4.D Additional simulations

This appendix discusses an extension to the simulation study with two additional actors integrated into the distance metric proposed in 4.2.3.

We constructed a statistic based on distances from the residual points to parts of the envelope $\mathcal{E}_i = \{x \in \mathcal{R} | x \in (l_i, u_i)\}$, namely the envelope median m_i, upper (u_i) and lower (l_i) bounds. The statistic is given by:

$$d = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \frac{(|r_i - m_i|)^p g(b_i)^{I(r_i \in \mathcal{E}_i)}}{f(w_i)}$$

where r_i is the $i-$th ordered residual and $g(b_i)$ is a function of the distance of the

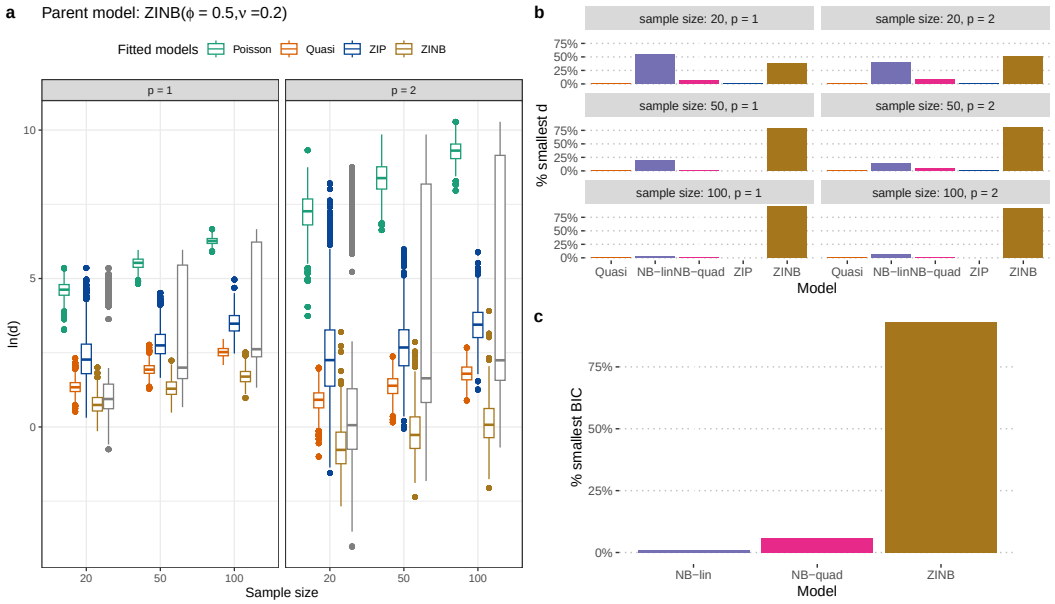


Figure 4.A.5: Figures generated when parent model is ZINB with a with zero inflation factor value of 0.2 and a dispersion parameter value of 0.5. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

residual point to the boundary of the envelope:

$$b_i = \begin{cases} r_i - u_i, \text{ if } r_i > u_i \\ l_i - r_i, \text{ if } r_i < l_i \end{cases}$$

The indicator function $I(r_i \in \mathcal{E}_i)$ is equal to 1 if the residual point is contained in the envelope and equal to 0 otherwise, therefore the penalty function g only influences the metric if the point is outside of the envelope. The variable p can take two different values depending on the value of p. When the value of $p = 1$, the L_1 norm is considered and for $p = 2$, the squared Euclidean distance is considered. The function $f(w)$ is the function for envelope width and acts as a scaling factor and has been tested for three different variations:

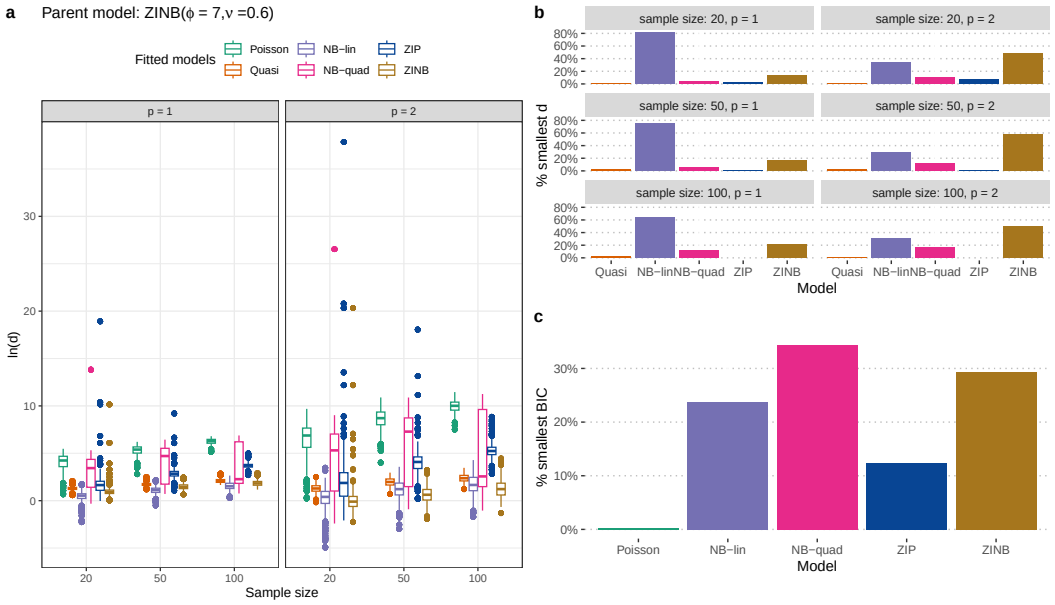


Figure 4.A.6: Figures generated when parent model is ZINB with a with zero inflation factor value of 0.6 and a dispersion parameter value of 7. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

- no scaling: $f(w) = 1$
- inverse linear scaling: $f(w) = w$
- squared inverse scaling: $g(w) = w^2$

And we tested five different variations of $g(b)$:

- constant/no penalty: $g(b) = 1$,
- unlimited linear increase: $g(b) = \alpha + \gamma b$,
- saturated increase (ratio): $g(b) = \dfrac{\alpha + \gamma_1 b}{1 + \gamma_2 b}$,

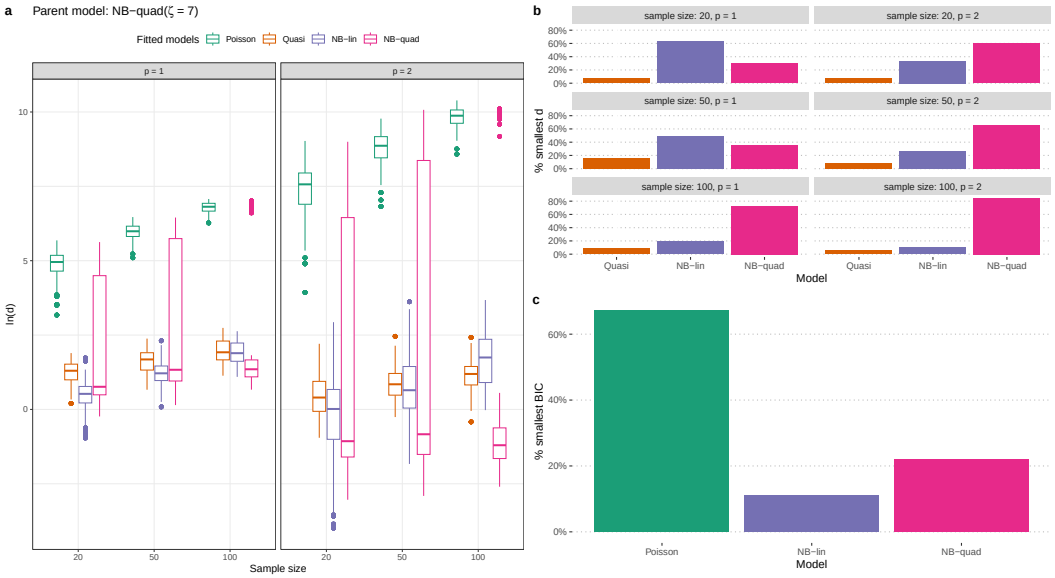


Figure 4.B.1: Figures generated when parent model is Negative Binomial with quadratic variance when the response residuals are considered. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

- saturated increase (logistic): $g(b) = \dfrac{\alpha + \gamma}{1 + \exp -\delta + (b - \eta)}$ and
- saturated increase (hyperbolic tangent): $g(b) = \alpha + \gamma \tanh \delta b$.

The hyper-parameters α, β, γ, γ_1, γ_2, η and δ are assumed to be known and fixed. The penalties are introduced to differentiate, for instance, residual points that are close to either u_i or l_i, but inside the envelope (and therefore expected under the fitted model), from points barely outside of the envelope, which should be more penalised, since that would not be expected under the fitted model most of the time.

We carried out a simulation study with 1,000 simulated samples from each of three

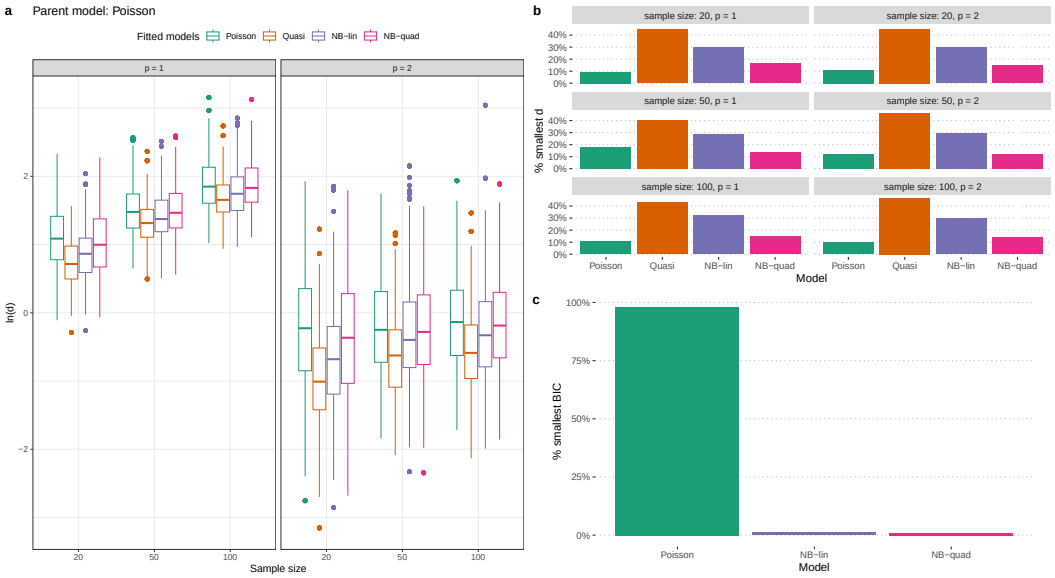


Figure 4.B.2: Figures generated when parent model is Poisson when the response residuals are considered. Panel (a) shows a boxplot of the base distance considered in the log scale for the fitted models. Panel (b) shows the bar plots illustrating the number of times a particular fitted model has the distance metric computed to be the minimum in a single simulation run. Panel (c) shows the barplot demonstrating the number of times a particular fitted model has the BIC value computed to be the minimum in a single simulation run.

sample sizes (20, 50, and 100) and three parent models (Poisson, negative binomial with a quadratic variance function with strong and mild overdispersion, negative binomial with a linear variance function with strong and mild overdispersion). We fitted three models to each simulated sample (Poisson and negative binomial with quadratic and linear variance functions), produced a half-normal plot with a simulated envelope for the Pearson residuals and computed d_i.

4.D.1 Results

The results are shown as three barplots with each barplot corresponding to each parent model. Each row in the bar plot shows the each of the $g(b)'s$ considered and each column denotes the sample sizes considered $(20, 50, 100)$. Each bar in

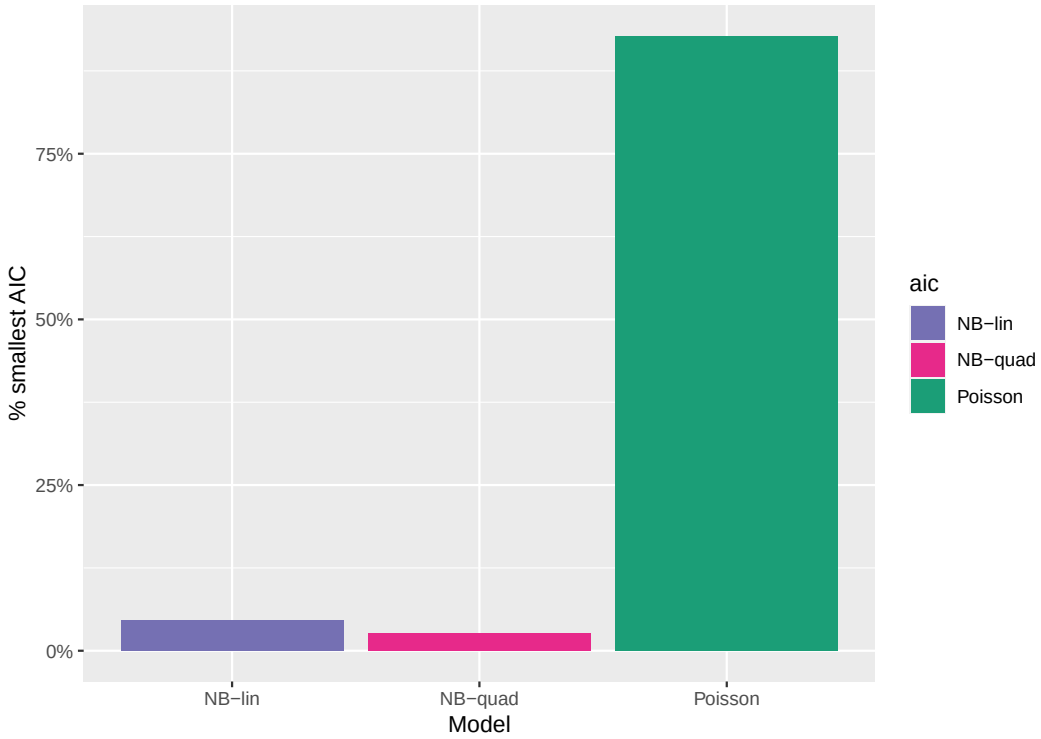


Figure 4.C.1: Figure shows the barplot when the parent model is Poisson, demonstrating the number of times a particular fitted model has the AIC value computed to be the minimum in a single simulation run.

the individual block corresponds to the combination of type of distance ($p = 1$: L_1 norm, $p = 2$: squared Euclidean distance) and $f(w)'s$ considered and the y axis shows the log transformed sum of the distance values for each fitted model. The fitted models are given in the legend; the distance metric values were calculated for NB-lin, NB-quad and Poisson as parent models, respectively. It is clearly evident from all barplots (Figure 4.D.1, Figure 4.D.2 and Figure 4.D.3) there is no difference between the penalty functions ($g(b)$) in terms of performance. There is negligible effect of scaling factor on the efficiency of the distance metric as it does not provide an added ability in selecting the best performing model. The reasons for the null performance of added factors is presumed to be because we are recreating perfect scenarios where parent models are fitted to data generated

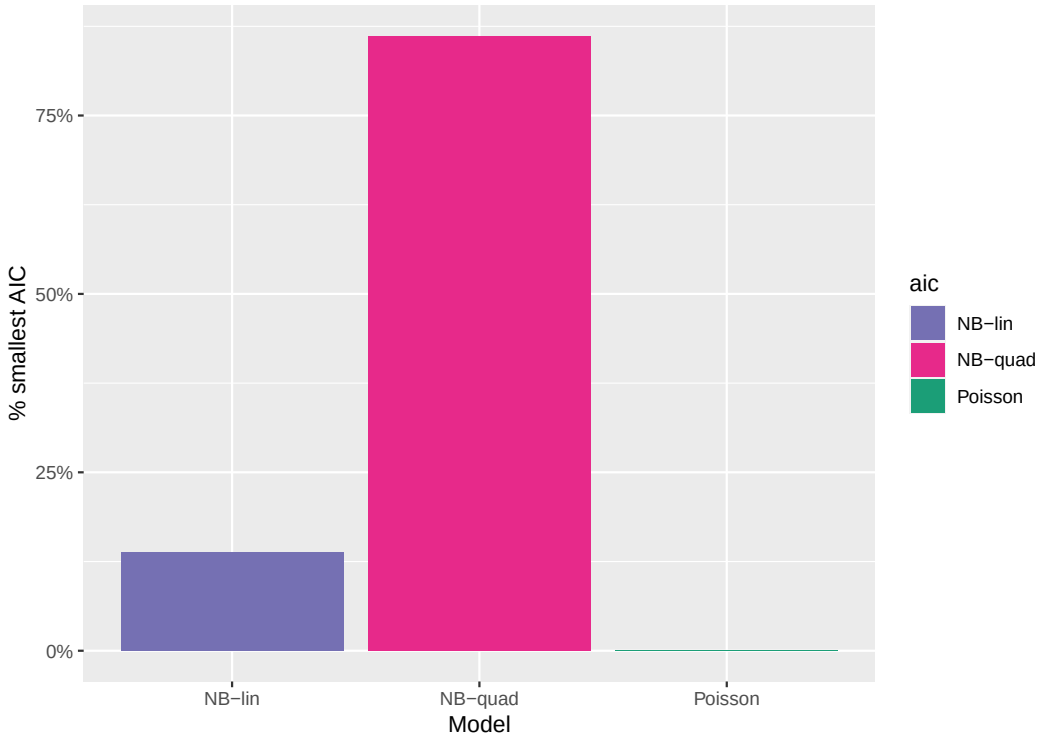


Figure 4.C.2: Figure shows the barplot when the parent model is Negative Binomial with a quadratic variance function with a dispersion value of 5, demonstrating the number of times a particular fitted model has the AIC value computed to be the minimum in a single simulation run.

by themselves or by closely related models. The poor performance of $g(b)$ might also be attributed to the small hyper-parameter values and since all the scenarios are perfect there are only few residuals falling outside the envelope.

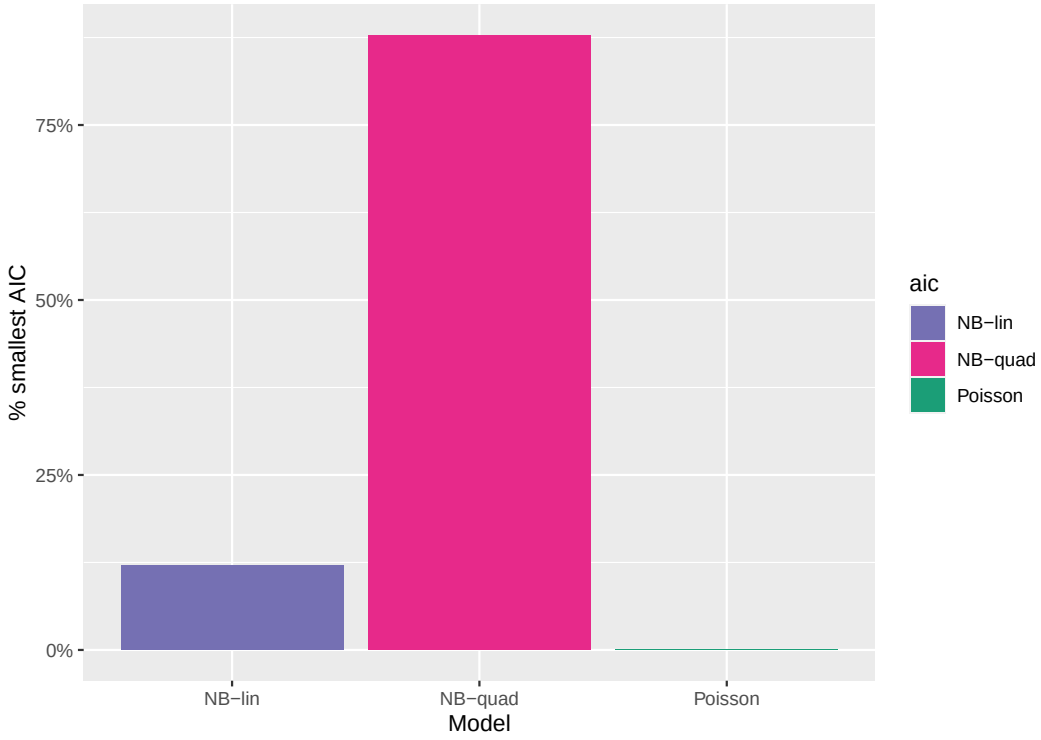


Figure 4.C.3: Figure shows the barplot when the parent model is Negative Binomial with a quadratic variance function with a dispersion value of 2, demonstrating the number of times a particular fitted model has the AIC value computed to be the minimum in a single simulation run.

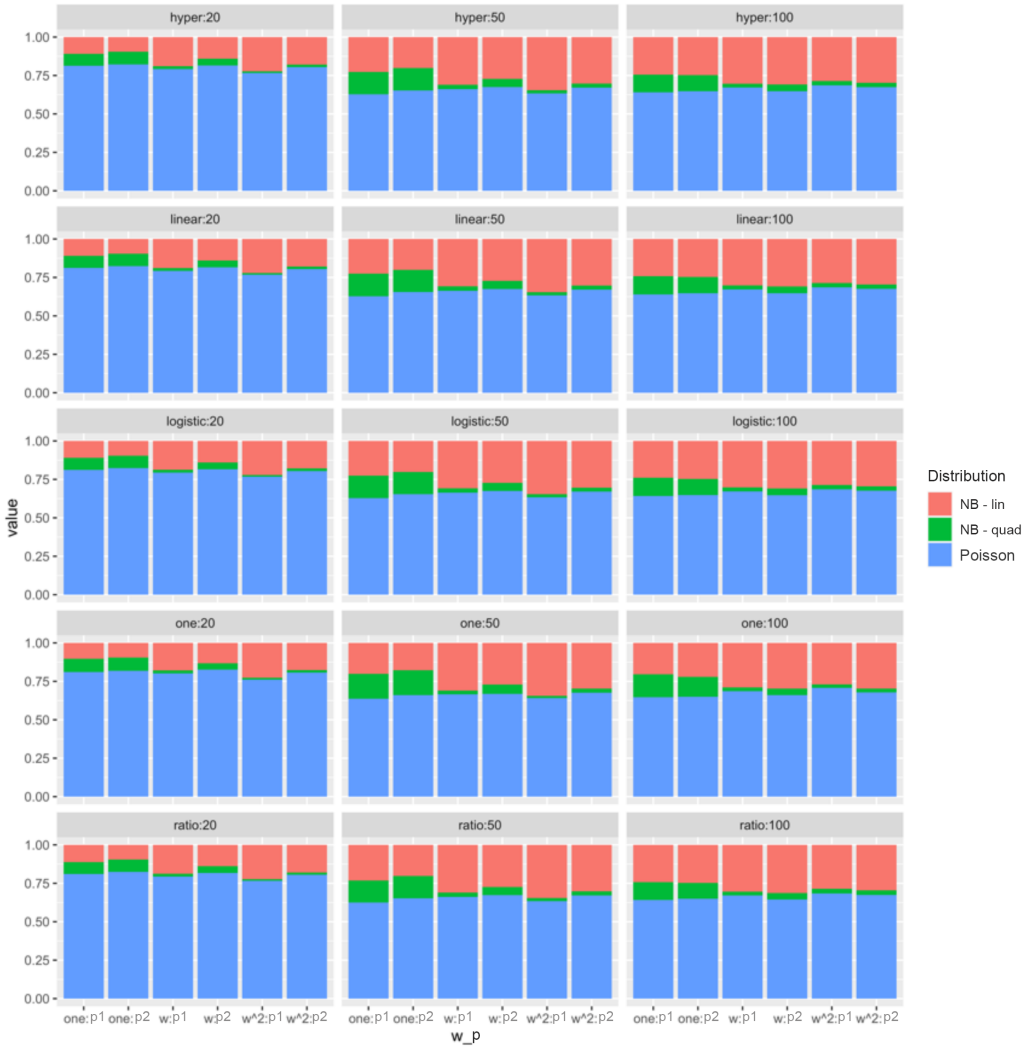


Figure 4.D.1: Figure generated when the parent model is the Poisson, every row corresponds to function b that corresponds to the distance of the residual from the simulated envelope, b considered(hyperbolic tangent,constant linear increase, logistic function, constant/no penalty and ratio. Every bar in the plot consist of combination of function w that corresponds to the distance between the lower and upper envelope considered, w(one, w, w^2))

Figure 4.D.2: Figure generated when the parent model is the NB-quad with a dispersion of 2, every row corresponds to function b that corresponds to the distance of the residual from the simulated envelope, b considered(hyperbolic tangent,constant linear increase, logistic function, constant/no penalty and ratio. Every bar in the plot consist of combination of function w that corresponds to the distance between the lower and upper envelope considered, w(one, w, w^2)

Figure 4.D.3: Figure generated when the parent model is the NB-lin with a dispersion of 5, every row corresponds to function b that corresponds to the distance of the residual from the simulated envelope, b considered(hyperbolic tangent,constant linear increase, logistic function, constant/no penalty and ratio. Every bar in the plot consist of combination of function w that corresponds to the distance between the lower and upper envelope considered, w(one, w, w^2)

CHAPTER 5

Mixed and marginal models applied to interval-censored bounded data

In this chapter, we discuss the analysis of a longitudinal dataset using mixed and marginal modelling approaches. The study involves the evaluation of the percentage of mold growth on two types of wooden boards in an ordinal scale. This is a comparative study between the traditional medium-density particle board and the innovative sugarcane bagasse-based boards. We employ mixed and marginal modelling approaches, using a transformation of the ordinal responses, as well as an interval-censored approach.

5.1 Introduction

In statistics, longitudinal data is characterised by measurements taken repeatedly at different time points on the same experimental or observational unit (Verbeke et al., 1997). Analysing this type of data gives the flexibility to understand the evolution of processes over time, but comes with the challenge of appropriately accommodating the possible correlation between the data collected within the same experimental or observational unit. To do so, additional parameters need to be included to account for the different correlation structures that help explaining the overall variability in the data. The two types of modelling strategies adopted to understand these data are the marginal modelling approach and the mixed modelling approach (Fitzmaurice et al., 2008). Marginal modelling incorporates correlation structures to account for the variability, whereas the mean response is related to the covariates (Liu, 2015). Mixed modelling involves considering fixed effects that explain the relationship between the response variable and the predictor variables, and random effects that can accommodate the variability within a group level. For this reason marginal modelling is also known as population-level analysis as it gives population averaged results and mixed modelling is known as subject-level analysis (Lee and Nelder, 2004).

In this chapter we applied subject-specific modelling to an interval-censored variable. Interval-censored responses are observed when the exact response value is unknown but an interval encompassing them is known. It occurs commonly in survival analysis, where the response is the time until the occurrence of an event of interest (Radke, 2003). In survival analysis the modelling of interval-censored data includes using the proportional hazards model (Finkelstein, 1986), in which the baseline distribution and regression parameters are fitted simultaneously by maximum likelihood estimation. Another approach involves the application of proportional odds method using the sieve maximum likelihood estimator (Huang

and Rossini, 1997). The three most common approaches used for inference are (1) maximum likelihood (Huang, 1996) (2) generalized estimating equations (Lin et al., 1998); and the (3) imputation (Pan, 2000). Approach (1) involves specifying differences of cumulative distribution functions in the likelihood to incorporate the interval-censored data. It is well known in the literature that while this approach is helpful, its pitfalls are reduced precision in the estimates (Gentleman and Geyer, 1994). Approach (2) comes with the limitation to assess the asymptotic validity and the properties of the estimators, as well as their efficiency (Zhang and Sun, 2010). Approach (3) involves the use of multiple imputation methods to impute the interval-censored data by sampling from the current estimate of the conditional error (Wei and Tanner, 1991), or by imputing using a (semi-)parametric model such as the Cox proportional hazards model (Pan, 2000).

The main objectives of this chapter are to give the reader a comparative analysis of the results obtained from two different modelling techniques applied to longitudinal data. We explored two of most common approaches for modelling longitudinal data, namely marginal and mixed modelling frameworks, while at the same time working with imputed data versus incorporating the interval-censored data directly in the likelihood. The remainder of the chapter is organised as follows: Section 5.2 provides an overview of mixed and marginal modelling, interval censored data and beta regression models. In Section 5.3 we describe the motivational study and in Section 5.3.1 we outline the modelling strategies. In Section 5.4 we present the results from the modelling. Finally, in Section 5.5 we provide a discussion of results and future considerations.

5.2 Overview of methods

5.2.1 Mixed and marginal model specifications

In this subsection we introduce the specification of mixed and marginal modelling frameworks from a longitudinal data analysis perspective.

Let Y_{it} be the response variable, and $\mathbf{x}_{it}$ be the $p \times 1$ vector of covariates at time t for subject i where $i = 1, \ldots, K$ and $t = 1, \ldots, n_i$. For a marginal modelling approach the marginal expectation is given by:

$$\mathrm{E}(Y_{it}) = \mu_{it}.$$

We may assume a linear predictor for the mean

$$h(\mu_{it}) = \mathbf{x}_{it}^{\top}\boldsymbol{\beta},$$

where $h(\cdot)$ is a monotonic and differentiable link function and $\boldsymbol{\beta}$ is the vector of parameters of order $p \times 1$. We also assume that

$$\mathrm{Var}(Y_{it}) = \phi V(\mu_{it})$$

where $V(\cdot)$ is the variance function.

For a mixed model, there is an additional term to account for the subject-specific variations. The conditional expectation is given by

$$\mathrm{E}(Y_{it}|b_i) = \mu_{it},$$

where the b_i are assumed to be independent random effects, which are distributed according to a specified probability distribution. We also have that

$$h(\mu_{it}) = \mathbf{x}_{it}^{\top}\boldsymbol{\beta} + \mathbf{z}_{it}^{\top}\mathbf{b}$$

and

$$\mathrm{Var}(Y_{it}|b_i) = \phi V(\mu_{it})$$

where $\mathbf{z}_{it}$ is the design matrix for q random effects of order $n \times q$ and $\mathbf{b}$ is the associated parameter vector of order $q \times 1$.

One major difference between marginal and mixed models is that mixed models are based on the assumption of a conditional probability distribution, i.e. $Y_{it}|b_i \sim \mathcal{D}$, whereas marginal models make only first- and second-moment assumptions for Y_{it}, which does not include specifying a full probability distribution.

A key application of marginal modelling or population averaged models is in the field of epidemiology where we compare two groups with different treatment effects. Mixed modelling is implemented when individual-level heterogeneity is examined (Hu et al., 1998).

5.2.2 Interval-censored data

Interval censoring occurs when the exact observed value is not known, but an interval containing it § is known. It is a common phenomenon in survival analysis, when typically we analyse the time T until the occurrence of an event of interest. If the event happens between two known interval bounds, L and U, such that $L \leq T \leq U$, this data is termed as interval censored. The most common examples include clinical trials involving AIDS patients where the occurrences of the viral and bacterial infections (Betensky and Finkelstein, 1999) with correlated interval censored endpoints, or by using the day to day maintenance data from machines (Han et al., 2024).

Right- and left-censored data could be considered as special instances of interval censored data, where $U = \infty$ and $L = -\infty$, respectively (Gómez et al., 2009). An exactly observed data point happens when $L = U$. In practice, right censored is seen more frequently than left censored and interval censored data. Interval censored data is more prevalent in clinical studies and longitudinal studies. It is

also possible to find all types of censoring in a single dataset (Zheng and Zelen, 2009). There are different methodologies in the literature to deal with censored data. These methodologies include weighting, imputation, maximum likelihood, and Bayesian methods (Lotspeich et al., 2024).

5.2.3 Beta regression

The beta distribution $\mathcal{B}$ can be used when the response variable values lie in the $(0, 1)$ interval. It is a very flexible distribution, with its probability density function accommodating right-skewed, left-skewed, uniform and unimodal shapes (Cribari-Neto and Zeileis, 2010). Its probability density function is given by:

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1$$

where $p, q > 0$ are shape parameters, and $\Gamma(\cdot)$ is the gamma function.

An alternative parameterisation, given by Ferrari and Cribari-Neto (2004), uses the transformations $\mu = \frac{p}{p+q}$ and $\phi = p + q$, yielding

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

where $0 < \mu < 1$ is the mean and $\phi > 0$ the dispersion parameter.

Let $Y_1, \ldots, Y_n$ be a random sample, with

$$Y_i \sim \mathcal{B}(\mu_i, \phi_i),\ i = 1, \ldots, n.$$

The beta regression model is given by

$$g(\mu_i) = \boldsymbol{\eta}_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^\top$ is a $k \times 1$ vector of unknown regression parameters ($k < n$), $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^\top$ is the vector of covariates and $g(\cdot) : (0, 1) \mapsto \mathbb{R}$ is a link function. The most commonly used link functions are the logit, probit, complementary

log-log, and Cauchy. Typically $\phi_i = \phi$, however it can also be modelled with covariates within a distributional regression approach (Stasinopoulos et al., 2017).

The variance of Y as a function of the mean and dispersion parameters is given by

$$\text{Var}(Y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi_i}.$$

5.3 Motivational study: the mold growth data

As an application related to agricultural sciences, we present a dataset obtained by Garzón-Barrero et al. (2016). In this study, the researchers aimed to evaluate the performance of an alternative sugarcane bagasse particle board (BCP) using castor oil polyurethane resin, in comparison with commercial medium density wood particle board (MDP), under natural and accelerated test conditions. A better performance was considered when less mold infestation was found.

The experiment was carried out at the University of São Paulo, Animal Science and Food Engineering Department, Pirassununga, state of São Paulo, Brazil. Samples of both materials were produced in laboratories following specific technical standards recommended by Brazilian legislation. The experiment is a 2×2 factorial experiment with 2 types of wood and 2 types of coating with six replicates. The two types of wooden material considered are: (1) BCP (sugarcane bagasse particle board) and (2) MDP(Medium density particle board). The two variations for the coating are: (1) With coating (2) No coating. In total, 48 panels (24 of each material) were used in the study, with the following dimensions: $270mm \times 50mm \times 12mm$ (length, width, thickness). For each material the two different sides (called “faces") are considered (1 and 2). Here, we are considering only the data from natural conditions, that is, the particle boards were evaluated after 12 months of exposure to natural weather in Pirassununga, and the response variable observed was an ordinal scale reflecting the percentage of fungal infestation,

measured weekly at four time points, totalling 192 observations. In the original study, 11 ordinal categories were considered according to the perceived percentage of mold growth: 0 (91% to 100%), 1 (81% to 90%), 2 (71% to 80%), 3 (61% to 70%), 4 (51% to 60%), 5 (41% to 50%), 6 (31% to 40%), 7 (21% to 30%), 8 (11% to 20%), 9 (1% to 10%) and 10 (0%). The dataset consists of 192 observations and 6 columns. The covariates considered are type of the material considered, if there is a coating on the wooden board or not, number of replications, day of the experiment considered, face of the wooden board and an ordinal variable that indicates the percentage of mold growth.

More details about this study can be found in Garzón-Barrero et al. (2016).

5.3.1 Modelling strategies

We used three modelling strategies. Two of them involved transforming the ordinal response variable (equivalent to simple imputation of the interval-censored response) and fitting mixed and marginal models, and one treated it as an interval-censored response within a mixed modelling approach.

The transformation of the ordinal response involved two simple steps:

1. Impute categories with the midpoint of the interval they represent (e.g. category 0 was transformed to 0.95, category 1 was transformed to 0.85, and so on);

2. For category 10, which represents exactly 0%, the value used was 0.001, to avoid numerical problems when estimating the models, since the beta distribution does not allow for perfect zeros (and ones, but these are not present in our transformed data).

5.3.1.1 Approach 1: Mixed modelling of the transformed response

The first approach involved using a beta mixed modelling framework considering the transformed response variable. In the maximal model, we included a random intercept per experimental unit, since the multiple observations over time are correlated. We also included the effects of material type (BCP and MDP), coating (yes and no), day of measurement (as a linear effect), all three two-way interactions between these factors, and the three-way interaction between material type, coating and day of measurement, all as fixed in the linear predictor. We used a logit link, and assumed the random intercepts were normally distributed with mean zero. This can be written as

$$\begin{aligned} Y^*_{ijkt}|b_i &\sim \mathcal{B}(\mu_{ijkt}, \phi) \\ \log\left(\frac{\mu_{ijkt}}{1-\mu_{ijkt}}\right) &= \beta_0 + \beta_1 \text{day}_t + \beta_2 \text{coating}^{\text{no}}_j + \beta_3 \text{material}^{\text{MDP}}_k + \\ &\quad \beta_4 \text{day}_t \text{coating}^{\text{no}}_j + \beta_5 \text{day}_t \text{material}^{\text{MDP}}_k + \beta_6 \text{coating}^{\text{no}}_j \text{material}^{\text{MDP}}_k + \\ &\quad \beta_7 \text{day}_t \text{material}^{\text{MDP}}_k \text{coating}^{\text{no}}_j \text{material}^{\text{MDP}}_k \\ b_i &\sim \text{N}(0, \sigma^2_b) \end{aligned} \tag{5.1}$$

where σ^2_b is the variance associated with the random intercepts. Here, $\mu_{ijkt} = \text{E}(Y^*_{ijkt}|b_i)$ represents the conditional mean. The random intercepts b_i are indexed by the combinations of side and replicate within a treatment, totalling $2\times6\times2\times2 = 48$ groups.

This model was implemented using the `gamlss` package (Stasinopoulos et al., 2018) for fitting generalised linear mixed models in R software (R Core Team, 2024). The significance of the fixed effects was assessed via likelihood-ratio tests between nested models.

5.3.1.2 Approach 2: Marginal modelling of the transformed response

As stated before, the marginal model only makes first- and second-order moment assumptions. The assumed mean structure was the same as in Equation 5.1 above, with $\mu_{ijkt} = \mathrm{E}(Y^*_{ijkt})$ representing the marginal mean, instead of the conditional one.

The marginal modelling considers the correlation between the observations in different pre-specified ways. Some of the correlation structures available are, for instance, the independence structure, which assumes covariance between two observations is equal to zero; the autoregressive structure, which considers increasing correlation between the adjacent timepoints and decreasing correlation as the distance between the timepoints increases. The correlation structure that involves the most parameters to be estimated is the unstructured. This particular structure assumes unique correlation between each pair of observations. However, the estimates obtained through the generalized estimation equations (GEE) frameowork are valid irrespective of the correlation structure used (Agresti, 2012).

Under the principle of parsimony, we used the *exchangeable* correlation structure, which assumes that the correlations between all pairs of observations within the same cluster (combinations of side and replicates) are the same. The exchangeable correlation structure is also known as the “compound symmetry" structure (West et al., 2022). For our case study, the exchangeable correlation structure for one group (combination between face and replicate within a treatment) is given by

$$\mathrm{Cov}(Y_{ijkt}, Y_{i'jkt}) = \begin{cases} 1, & i = i' \\ \rho, & i \neq i' \end{cases}$$

where ρ is the correlation between observations within the same group/cluster.

This model was implemented using `glmgee` function in the `glmtoolbox` (Vanegas

et al., 2023) available in R.

5.3.1.3 Approach 3: Mixed modelling of the interval-censored response

The third approach involved the beta mixed modelling framework applied to the interval censored data. Now the conditional distributional assumption is made for the original interval-censored response variable Y_{ijkt}, rather than for the transformed variable Y^*_{ijkt}. This model can be estimated via maximum likelihood, which involves marginalising over the random intercepts. Let Y_{ijkt} be the observed responses, while Y^u_{ijkt} and Y^l_{ijkt} the upper and lower limits of the interval-censored responses. Ignoring random effects, the likelihood for the beta regression with a mix between uncensored and interval-censored responses is

$$L(\mu_{ijkt}, \phi; \mathbf{y}, \mathbf{y}^u, \mathbf{y}^l) = \prod_{i,j,k,t} f(y_{ijkt}; \mu_{ijkt}, \phi)^{\delta}_{ijkt}[F(y^u_{ijkt}; \mu_{ijkt}, \phi) - F(y^l_{ijkt}; \mu_{ijkt}, \phi)]^{1-\delta_{ijkt}},$$

where δ is the censoring indicator, equal to 0 if an uncensored response is observed, and to 1 if an interval-censored response is observed; $f(\cdot)$ is the density function of the beta distribution, and

$$F(y; \mu, \phi) = I_y(\mu, \phi) = \frac{B_y(\mu\phi, \phi(1-\mu))}{B(\mu\phi, \phi(1-\mu))}$$

is the regularised incomplete beta function, with

$$B_y(\mu\phi, \phi(1-\mu)) = \int_0^y t^{(\mu\phi-1)}(1-t)^{(\phi(1-\mu)-1)}dt$$

the incomplete beta function and

$$B(\mu\phi, \phi(1-\mu)) = \frac{\Gamma(\mu\phi)\Gamma(\phi(1-\mu))}{\Gamma(\phi)}$$

the complete beta function.

To estimate the mixed model, the random effects need to first be marginalised. Therefore, the likelihood becomes:

$$L(\mu_{ijkt}, \phi; \mathbf{y}, \mathbf{y}^u, \mathbf{y}^l) = \prod_i \int_{-\infty}^{\infty} \prod_{j,k,t} f(y_{ijkt}; \mu_{ijkt}, \phi)^{\delta}_{ijkt} \times \\ \times [F(y^u_{ijkt}; \mu_{ijkt}, \phi) - F(y^l_{ijkt}; \mu_{ijkt}, \phi)]^{1-\delta_{ijkt}} g(b_i; \sigma_b^2) db_i,$$

where $g(\cdot)$ is the probability density function of the normal distribution. Because this integral has no analytic solution, numerical methods need to be used to approximate it, such as penalised quasi-likelihood and the Laplace approximation (which approximate the integrand), or Gauss-Hermite quadrature (which approximates the integral). Here, we use the penalised quasi-likelihood approach as implemented in `gamlss` (Stasinopoulos et al., 2017).

The model was implemented using the interval-censored beta distribution created using the `gamlss.cens` package (Stasinopoulos et al., 2018) for R software. The significance of the fixed effects was assessed via likelihood-ratio tests between nested models. For all analyses the zero values were substituted by 0.0001 and the one values at the end of the intervals were substituted by 0.9999.

5.4 Results

In this section, we present and discuss the results for the motivational study, on the resistance of two types of materials (one standard and other proposed) for the manufacture of custom furniture, considering the three proposed methods. In this context, the estimates of fixed parameters are presented in the Table 5.1, considering the three modelling frameworks (mixed model with transformed and interval-censored response, marginal model with transformed response), used for data analysis

Parameter	Mixed model for transformed response	Marginal model for transformed response	Mixed model for interval-censored response
β_0	-4.82^* (0.379)	-4.56^* (0.325)	-4.19^* (0.334)
β_1	0.09^* (0.018)	0.08^* (0.012)	0.09^* (0.016)
β_2	2.91^* (0.422)	2.61^* (0.464)	2.81^* (0.394)
β_3	-0.65 (0.549)	-2.24 (1.56)	-0.45 (0.482)
β_4	0.08^* (0.022)	0.09^* (0.017)	0.09^* (0.021)
β_5	-0.001 (0.028)	0.06 (0.052)	-0.01 (0.024)
β_6	2.66^* (0.624)	4.09^* (1.69)	2.37^* (0.565)
β_7	-0.07^* (0.033)	-0.13^* (0.056)	-0.06 (0.030)
σ_b^2	0.128	–	0.152
ϕ	0.25	0.05	0.251
ρ	–	0.029	–

Table 5.1: Parameter estimates (standard errors) from the three modelling strategies used (mixed modelling for transformed and interval-censored responses, and marginal modelling for the transformed response). The $*$ indicates significance at a 5% level based on the Wald t-test for fixed effects.

Based on the models fitted to the data, the estimates of the fixed effects for the three structures considered, as presented in the Table 5.1, are very close, with some exceptions as can be observed for example for $\hat{\beta}_3$), whose point estimate is relatively larger in the marginal model. An agreement on the significance of the effects can also be observed using the Wald Test ($*$) at a 5% level. In general, the proximity of point estimates for fixed effects is expected when fitting marginal and mixed models, but the same cannot be confirmed for precision statistics, such as standard errors.

To understand the sensitivity of using transformation we used two more values for transformation and got the estimates for mixed and marginal models. Table 5.2 shows the estimates with the extreme values in data considered are $(0.0001, 0.9999), (0.001, 0.999), (0.01, 0.99)$ and Table 5.3 shows the same for the marginal model. It is clear that first and second column have similar values for marginal modelling framework than mixed modelling framework. But using $(0.01, 0.99)$ has very different estimate values compared to modelling frameworks using $(0.0001, 0.9999)$ and $(0.001, 0.999)$

In practical terms, for the case study, it was found that the three-way interaction between day, coating and material was significant for the models that considered the transformed response, whereas it was not for the model considering the interval-censoring. Aside from that, all three approaches identified the day × coating and material × coating two-way interactions as significant at a 5% level.

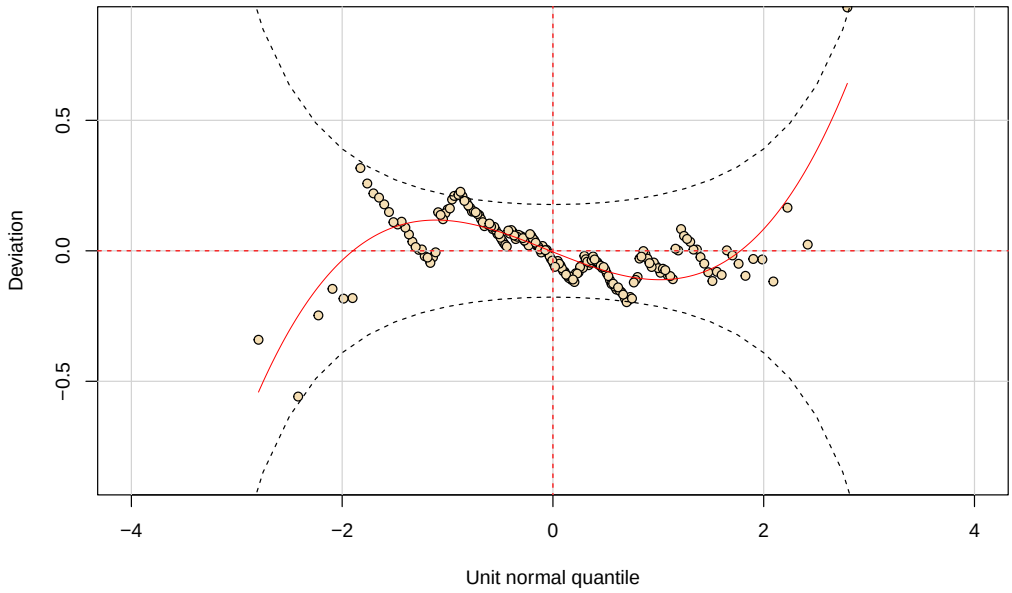
Also, it is important to consider that, for the mixed modelling framework for the transformed response, the mold growth increases over time, and when there is no coating available for the medium density particle boards. This implies when there is no coating available, the sugarcane bagasse based wooden board performs better than the traditional medium density particle boards.

Parameter	Mixed model for $value = 0.0001$	Mixed model for $value = 0.001$	Mixed model for $value = 0.01$
β_0	-4.82^* (0.379)	-4.4^* (0.353)	-3.71^* (0.302)
β_1	0.09^* (0.018)	0.07^* (0.017)	0.05^* (0.014)
β_2	2.91^* (0.422)	2.45^* (0.394)	1.71^* (0.343)
β_3	-0.65 (0.549)	-0.61 (0.517)	-0.37 (0.444)
β_4	0.08^* (0.022)	0.09^* (0.02)	0.13^* (0.018)
β_5	-0.001 (0.028)	0.002 (0.026)	-0.001 (0.022)
β_6	2.66^* (0.624)	2.64^* (0.586)	2.43^* (0.51)
β_7	-0.07^* (0.033)	-0.08^* (0.003)	-0.08 (0.027)

Table 5.2: Parameter estimates (standard errors) from Mixed modelling techniques using three different values for the extreme values of the transformed responses. The $*$ indicates significance at a 5% level based on the Wald t-test for fixed effects.

Parameter	Marginal model for $value = 0.0001$	Marginal model for $value = 0.001$	Marginal model for $value = 0.01$
β_0	-4.56^* (0.325)	-4.52^* (0.31)	-4.23^* (0.02)
β_1	0.08^* (0.012)	0.08^* (0.01)	0.067^* (0.01)
β_2	2.61^* (0.464)	2.57^* (0.95)	2.28^* (0.38)
β_3	-2.24 (1.56)	-2.04 (1.36)	-0.89 (0.52
β_4	0.09^* (0.017)	0.09^* (0.02)	0.107^* (0.015)
β_5	-0.06 (0.052)	0.06 (0.05)	-0.01 (0.02)
β_6	4.09 (1.69)	3.89^* (1.51)	2.75^* (0.69)
β_7	-0.13^* (0.056)	-0.12^* (0.05)	-0.08 (0.02)

Table 5.3: Parameter estimates (standard errors) from Marginal modelling techniques using three different values for the extreme values of the transformed responses. The $*$ indicates significance at a 5% level based on the Wald t-test for fixed effects.



,

or are the

Figure 5.1: Worm plot of the normalised quantile residuals from the mixed model applied to the transformed response (approach 1).

We considered a detrended Q-Q plot known as wormplots (Buuren and Fredriks, 2001) to evaluate model fit, since these are readily available within the `gamlss` framework. A worm plot is a quantile-quantile plot that compares the theoretical distribution of the residuals (in this case it is a normal distribution, in case the model is well fitted to the data) with the empirical distribution of the observed residuals. The data points form a worm-like string and for a well fitted model the worm string should align with the horizontal line in the centre. The plot (Figure 5.1) shows a reasonable fit for the model considered but there are some deviations that indicate unaccounted variability.

From the marginal modelling framework, we also conclude that the sugarcane bagasse particle boards perform better than the medium density particle boards

when there is no coating available, which is consistent with the results from the mixed modelling framework for both approaches.

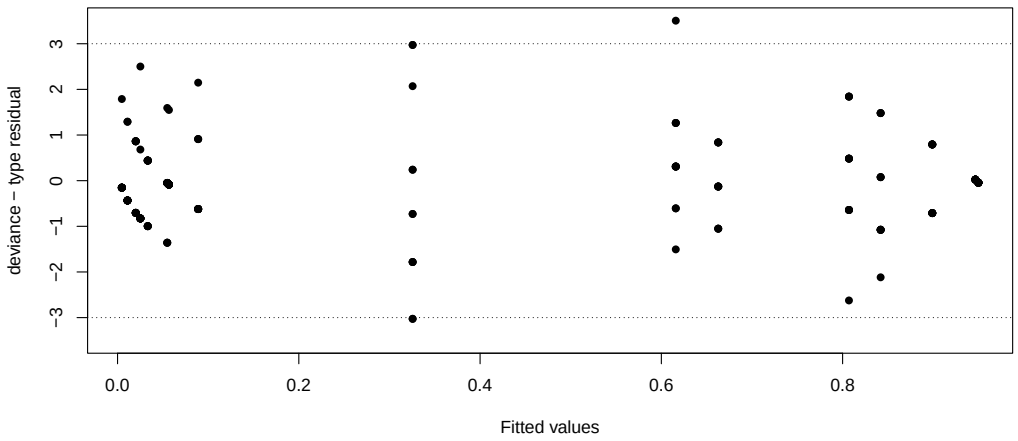


Figure 5.2: Residual analysis of the marginal fit for the dataset considering an exchangeable correlation structure. The plot shows a constant variance for most of the deviance residuals except for larger fitted values, which suggests a reasonably well-fitted model.

Finally, a diagnostic analysis for the mixed model using interval-censored responses is presented in Figure 5.3. The plot that shows a reasonable fit for the model as the residual points are deviating away along the extremes but still close to the horizontal axis along the centre. In comparison to the worm plot from from the mixd model that uses transformed responses (Figure 5.1), the model considering the interval-censored nature of the data presents an improved fit.

The fixed effect estimates for approach 3 shows that for every extra day there is a significant increase in mold growth. The lack of coating also significantly contributed to the increase in mold growth for the medium density particle board, which is in alignment with the other two approaches.

In our application, the percentage of mold growth was observed in control con-

ditions and this restricts the number of covariates used for modelling. In this restricted scenario, the results for both modelling show that compared to traditional Medium density boards, sugarcane bagasse boards have lesser mold growth when there is no coating on the wooden board.

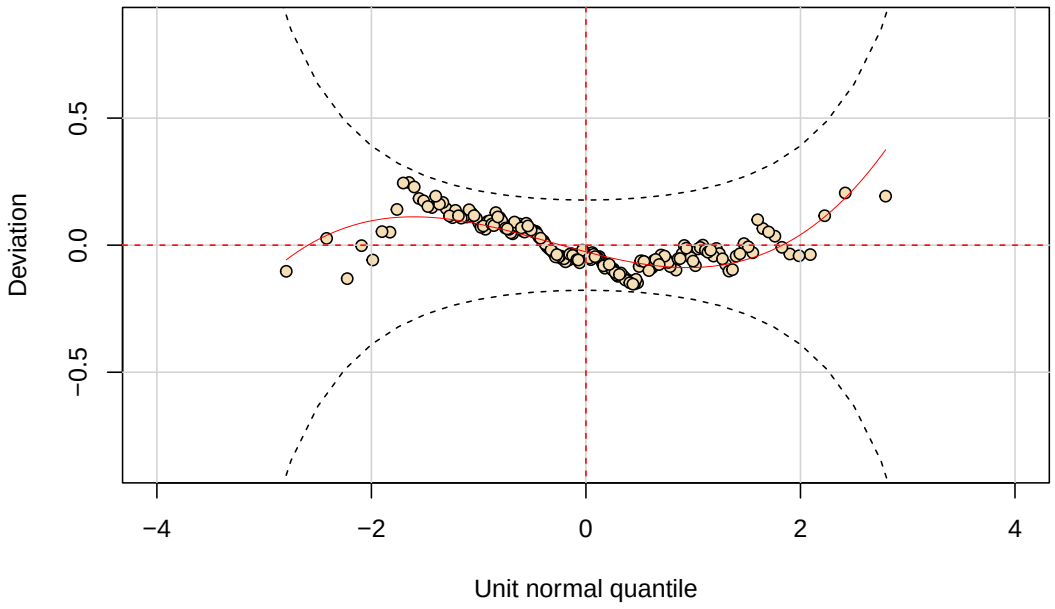


Figure 5.3: Worm plot for th mixed model using interval-censored data (approach 3).

5.5 Discussion

In this chapter we presented an alternative modelling strategy to log proportions model for modelling interval-censored bounded responses, motivated by a case study in agricultural engineering. The mixed model considering the interval-censored data has an added advantage of incorporating the hierarchical nature of the data while preserving the nature of the response variable. We emphasize that the choice of modelling strategy should be strongly related to the objectives

of the study and the nature of the observed data. Thus, if the problem in question is prediction, the marginal model can be a good alternative, since it can provide consistent point estimates regardless of the correlation structure. However, if the data dispersion is high and if the objective, in addition to prediction, is to select a better structure to accommodate such variability, marginal models may not be a suitable alternative. The main contribution of this study includes using interval censored beta regression that does not involve any transformation, which has been shown to reduce the bias in the estimates and performed better when assessing model fit using the diagnostic plots.

The challenges involved in this study included the difficulty in considering the complex relationship between the covariates in the dataset. Future work includes considering other correlation structures for the marginal models, using the Beta inflated distribution and a Bayesian approach for modelling the interval-censored data to allow for imputing the interval-censored data based on prior information.

CHAPTER 6

Final Remarks

In this chapter, we review and summarise the work presented in this thesis, examine any obstacles or difficulties encountered, and provide recommendations for future work.

In this thesis we proposed a goodness-of-fit diagnostic framework applied to count responses from a generalized linear modelling framework perspective, as well as modelling strategies applied to longitudinal data characterised by an interval-censored bounded response. We also presented real life ecological and agricultural case studies to discuss the proposed frameworks. Here we intend to give a brief overview to the work presented in the previous chapters and discuss future directions.

In the first part of this thesis we proposed a distance-based quantitative metric as an extension to half normal plots with a simulated envelope to provide an objective model selection method. Unlike traditional information criteria such as AIC and BIC, this model selection method is not constrained to the comparison of full likelihood-based models. We showed that the proposed distance metric was

efficient recovering the true model in simulation studies. We also demonstrated the effectiveness of using the suggested metric in the context of mild and strong overdispersion, and low and high zero inflation. The results shown were in agreement with BIC, which is a commonly used information criterion. We noted that the distance metric exhibited superior performance to BIC when the parent model considered was the zero-inflated negative binomial for a strong overdispersion and low zero inflation scenario.

Additionally, we showed that the distance metric is a reasonable selection alternative for smaller sample sizes, when compared to BIC. This is evident for the case when the parent model considered was the NB-lin with mild overdispersion. The only instance where the distance metric failed to recognise the parent distribution was when the parent model was the standard Poisson model. However, this behaviour was not unexpected, because the Poisson model is the limiting case of the model extensions used for comparison. The suggested distance metric was tested for three sample sizes and two distance norms were considered. Among the models considered the Quasi-Poisson was the only model that could possibly accommodate underdispersion and the metric fared effectively in recognising the best performing model in such instances. The main contribution of the distance metric is to be able to provide a quantitative comparison between the performance of quasi likelihood models and full likelihood models. If quasi models are not considered this still serves as a complementary extension to half normal plots and model selection and adequacy assessment procedures.

An extension to the simulation study was done to consider other factors that could potentially contribute to the distance metric's performance. The two factors considered were the envelope width and a penalty based on the distance of the residual from the envelope boundary when it falls outside the envelope. Three different

functional configurations for the envelope width and five different configurations for the distance of the residual from the envelope boundary were explored, however there was effectively no difference in performance when these features were added to the distance metric. This behaviour was elucidated by our examination of closely related models, whereby the probability of residuals extending beyond the envelope is exceedingly low or negligible. The different functional configurations of the distance of the residual from the envelope when the residual falls outside of the envelope also exhibited no influence on the distance. The envelope width and its functional configurations had an influence on the distance metric. However the basic building block of the distance metric (based on the difference between the residual and the median of the envelope) made a more pronounced impact on overall performance.

Future directions to this work include considering mis-specified link functions mis-specified models and highly skewed data. These models are expected to generate more residuals outside the envelope and bring more definition to the addition of the envelope width and the distance of the residual from the envelope to the distance metric. The inclusion of mixed models would also be beneficial to see how random effects may improve model fit and accommodate overdispersion. However, this comes with the added complexity of determining different types of residuals to be used. Another future direction would be to consider one-parameter distributions to potentially avoid the bias of the number of parameters in the models considered. Currently, we are in the process of adding the distance-based metric to the `hnp` package (de Andrade Moral et al., 2017) as a summary output that aids goodness-of-fit assessment using half-normal plots with a simulated envelope.

The final part of this thesis presented a comparative study of two modelling frameworks applied to longitudinal data. We used two variations for a mixed modelling

framework that considered random effects to account for the variability within a group. The first variation included a transformation of interval-censored responses from an ordinal scale, and the second variation employed no transformation and directly accommodated the interval-censored nature of the response in the likelihood. We also used a marginal modelling approach based on generalized estimating equations (GEE). Interestingly, our results were consistent across all modelling techniques used. However, the mixed modelling approach that accommodated the interval-censored response provided a better fit.

The model fit for all three methods indicates unaccounted variability and a complex relationship between the covariates. However, each modelling approach accommodated this variability in a different way. One challenge encountered was how to compare goodness-of-fit of the three different approaches. For the mixed modelling, one could use the distance metric proposed in the earlier part of this thesis. However, simulating new responses from the GEE model fit is non-trivial and an object of ongoing research.

All proposed methods in this thesis were implemented using `R` (R Core Team, 2022) software and are accessible at the author's Github[5] via two public repositories. The repository `https://github.com/DARSHANAJAYA/Goodness-of-fit-Distance-metric` and related to Chapters 3, 4, and repository `https://github.com/DARSHANAJAYA/Fungi-study-` relates to chapter 5. These repositories include all the scripts to reproduce the analyses and the plots provided in the chapters.

[5] `https://github.com/DARSHANAJAYA`

Bibliography

Agresti, A. (2012). *Categorical data analysis*, volume 792. John Wiley & Sons. 94

Anderson, D. and Burnham, K. (2004). Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63(2020):10. 53

Atkinson, A. C. (1985a). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Oxford University Press. 41

Atkinson, A. C. (1985b). *Plots, transformations, and regression: an introduction to graphical methods of diagnostic regression analysis*. Oxford University Press. 54, 55

Barrero, N. M. G. (2015). *Study of the durability of sugarcane bagasse particle boards and castor oil resin for application in civil construction*. Phd thesis, University of Sao Paolo, Piracicaba, Sao Paolo. Available at https://doi.org/10.11606/T.74.2016.tde-16032016-161005. xiii, 18

Betensky, R. A. and Finkelstein, D. M. (1999). A non-parametric maximum likelihood estimator for bivariate interval censored data. *Statistics in Medicine*, 18(22):3089–3100. 89

Bolnick, D. I., Snowberg, L. K., Caporaso, J. G., Lauber, C., Knight, R., and Stutz, W. E. (2014). Major h istocompatibility c omplex class ii b polymorphism influences gut microbiota composition and diversity. *Molecular ecology*, 23(19):4831–4845. 68

Bonat, W. H. and Jørgensen, B. (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 65(5):649–675. 53, 56

Borges, C. G. (2013). Annona mucosa jacq.(annonaceae): a promising source of bioactive compounds against sitophilus zeamais mots.(coleoptera: Curculionidae). *Journal of Stored Products Research*, 55:6–14. 52

Brites-Neto, J., Duarte, K. M. R., and Martins, T. F. (2015). Tick-borne infections in human and animal population worldwide. *Veterinary world*, 8(3):301. xiii, 9

Brooks, M. E., Kristensen, K., Darrigo, M. R., Rubim, P., Uriarte, M., Bruna, E., and Bolker, B. M. (2019). Statistical modeling of patterns in annual reproductive rates. *Ecology*, 100(7):e02706. 52

Buuren, S. v. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, 20(8):1259–1277. 99

Coe, R. and Stern, R. (1982). Fitting models to daily rainfall data. *Journal of Applied Meteorology and Climatology*, 21(7):1024–1031. 30

Consul, P. (1990). On some properties and applications of quasi-binomial distribution. *Communications in Statistics-Theory and Methods*, 19(2):477–504. 53

Coxe, S., West, S. G., and Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of personality assessment*, 91(2):121–136. 2

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in r. *Journal of statistical software*, 34:1–24. 90

Cunningham, R. B. and Lindenmayer, D. B. (2005). Modeling count data of rare species: some statistical issues. *Ecology*, 86(5):1135–1142. 52

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, 1(4):311–341. 41

de Andrade Moral, R., Hinde, J., and Garcia Borges Demétrio, C. (2017). Half-normal plots and overdispersed models in r: the hnp package. *Journal of Statistical Software*, 81(10). 41, 105

De Souza, V., Inomoto, M., Pascholati, S., Roma-Almeida, R., Melo, T., and Rezende, D. (2015). Fitonematoides: Controle biológico e indução de resistência. *Revisão Anual de Patologia de Plantas, 1st ed.; Dalio, RJD, Ed*, pages 242–292. 10

Demétrio, C. G., Hinde, J., and Moral, R. A. (2014). Models for overdispersed data in entomology. *Ecological modelling applied to entomology*, pages 219–259. 32

Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34. 52

Du, J., Park, Y.-T., Theera-Ampornpunt, N., McCullough, J. S., and Speedie, S. M. (2012). The use of count data models in biomedical informatics evaluation

research. *Journal of the American Medical Informatics Association*, 19(1):39–44. 52

Eilenberg, J., Hajek, A., and Lomer, C. (2001). Suggestions for unifying the terminology in biological control. *BioControl*, 46:387–400. 9

Fatoretto, M. B., Moral, R. d. A., Demétrio, C. G. B., de Pádua, C. S., Menarin, V., Rojas, V. M. A., D'Alessandro, C. P., and Delalibera Jr, I. (2018). Overdispersed fungus germination data: statistical analysis using r. *Biocontrol Science and Technology*, 28(11):1034–1053. 35

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815. 90

Filgueiras, M. D. G., Matos, R. S., Barreto, L. P., Mascarin, G. M., Rizzo, P. V., Freitas, F. M. C., de Azevedo Prata, M. C., Monteiro, C., and Fernandes, É. K. K. (2023a). From the laboratory to the field: efficacy of entomopathogenic nematodes to control the cattle tick. *Pest Management Science*, 79(1):216–225. 4, 8

Filgueiras, M. D. G., Matos, R. S., Barreto, L. P., Mascarin, G. M., Rizzo, P. V., Freitas, F. M. C., de Azevedo Prata, M. C., Monteiro, C., and Fernandes, É. K. K. (2023b). From the laboratory to the field: efficacy of entomopathogenic nematodes to control the cattle tick. *Pest Management Science*, 79(1):216–225. 43

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, pages 845–854. 86

Fitzmaurice, G., Davidian, M., Verbeke, G., and Molenberghs, G. (2008). Advances in longitudinal data analysis: an historical perspective. In *Longitudinal data analysis*, pages 17–42. Chapman and Hall/CRC. 86

Folks, J. L. and Chhikara, R. S. (1978). The inverse gaussian distribution and its statistical application—a review. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 40(3):263–275. 31

Garzón-Barrero, N. M., Shirakawa, M. A., Brazolin, S., de Lara, I. A. R., Savastano Jr, H., et al. (2016). Evaluation of mold growth on sugarcane bagasse particleboards in natural exposure and in accelerated test. *International Biodeterioration & Biodegradation*, 115:266–276. 5, 14, 91, 92

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81(3):618–623. 87

Gómez, G., Calle, M. L., Oller, R., and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in r. *Statistical Modelling*, 9(4):259–297. 89

Han, D., Brownlow, J. D., Thompson, J., and Brooks, R. G. (2024). Bayesian estimation of the mean time between failures of subsystems with different causes using interval-censored system maintenance data. *Quality and Reliability Engineering International*. 89

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2:e616. 2

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press. 52, 68

Hilbe, J. M. and Greene, W. H. (2007). 7 count response regression models. *Handbook of statistics*, 27:210–252. 52

Hinde, J. and Demétrio, C. G. (1998a). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170. 32

Hinde, J. and Demétrio, C. G. (1998b). Overdispersion: models and estimation. *Computational statistics & data analysis*, 27(2):151–170. 53, 55

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., and Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American journal of epidemiology*, 147(7):694–703. 89

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568. 87

Huang, J. and Rossini, A. (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, 92(439):960–967. 86

Jackman, S. (2020). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. United States Studies Centre, University of Sydney, Sydney, New South Wales, Australia. R package version 1.5.5.1. 59

Jørgensen, B. (2013). Generalized linear models. In *Encyclopedia of environmetrics*, pages 1152–1159. Wiley. 38

Klafke, G., Webster, A., Agnol, B. D., Pradel, E., Silva, J., de La Canal, L. H., Becker, M., Osório, M. F., Mansson, M., Barreto, R., et al. (2017). Multiple resistance to acaricides in field populations of rhipicephalus microplus from rio grande do sul state, southern brazil. *Ticks and tick-borne diseases*, 8(1):73–80. 8

Knape, J. (2016). Decomposing trends in swedish bird populations using generalized additive mixed models. *Journal of Applied Ecology*, 53(6):1852–1861. 68

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14. 58

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical science*. 86

Lin, D., Oakes, D., and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, 85(2):289–298. 87

Liu, X. (2015). *Methods and applications of longitudinal data analysis*. Elsevier. 86

Lodder, R. A. and Hieftje, G. M. (1988). Quantile analysis: a method for characterizing data distributions. *Applied spectroscopy*, 42(8):1512–1520. 54

Lotspeich, S. C., Ashner, M. C., Vazquez, J. E., Richardson, B. D., Grosser, K. F., Bodek, B. E., and Garcia, T. P. (2024). Making sense of censored covariates: Statistical methods for studies of huntington's disease. *Annual Review of Statistics and Its Application*, 11. 90

Mainguy, J. and Moral, R. A. (2021). An improved method for the estimation and comparison of mortality rates in fish from catch-curve data. *North American Journal of Fisheries Management*, 41(5):1436–1453. 5, 13, 65, 66

Matis, J. H., Rubink, W., and Makela, M. (1992). Use of the gamma distribution for predicting arrival times of invading insect populations. *Environmental entomology*, 21(3):436–440. 30

Moral, R. A., Hinde, J., and Demétrio, C. G. B. (2017). Half-normal plots and overdispersed models in R: The hnp package. *Journal of Statistical Software*, 81(10):1–23. 54, 55, 59

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384. 2, 24

Nelson, G. A. (2019). Bias in common catch-curve methods applied to age frequency data from fish surveys. *ICES Journal of Marine Science*, 76(7):2090–2101. 14

Odum, E. P., Barrett, G. W., et al. (1971). *Fundamentals of ecology*, volume 3. Saunders Philadelphia. 1

Oliveira, I. C., Vieira, Í. S., Freitas, S. G., Campos, A. K., Paz-Silva, A., Monteiro, C. F., de Gives, P. M., and de Araújo, J. V. (2022). Evaluation of nematophagous fungal mycelial growth and interactions with bovine gastrointestinal parasitic nematodes. *Ger. J. Vet. Res*, 2:39–45. xiii, 13

Pan, W. (2000). A multiple imputation approach to cox regression with interval-censored data. *Biometrics*, 56(1):199–203. 87

Perumean-Chaney, S. E., Morgan, C., McDowall, D., and Aban, I. (2013). Zero-inflated and overdispersed: what's one to do? *Journal of Statistical Computation and Simulation*, 83(9):1671–1683. 3

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986. 39

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 6, 23, 106

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 58, 93

Radke, B. R. (2003). A demonstration of interval-censored survival analysis. *Preventive veterinary medicine*, 59(4):241–256. 86

Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45(1):218–227. 52, 68

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554. 59

Rodriguez-Vivas, R. I., Jonsson, N. N., and Bhushan, C. (2018). Strategies for the control of rhipicephalus microplus ticks in a world of conventional acaricide and macrocyclic lactone resistance. *Parasitology research*, 117:3–29. 8

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611. 22

Silva, D. M., de Souza, V. H. M., Moral, R. d. A., Delalibera Júnior, I., and Mascarin, G. M. (2022a). Production of purpureocillium lilacinum and pochonia chlamydosporia by submerged liquid fermentation and bioactivity against tetranychus urticae and heterodera glycines through seed inoculation. *Journal of Fungi*, 8(5):511. 4

Silva, D. M., de Souza, V. H. M., Moral, R. d. A., Delalibera Júnior, I., and Mascarin, G. M. (2022b). Production of purpureocillium lilacinum and pochonia chlamydosporia by submerged liquid fermentation and bioactivity against tetranychus urticae and heterodera glycines through seed inoculation. *Journal of Fungi*, 8(5):511. 11, 46

Smeenk-Enserink, N. and Van der Aart, P. (1974). Correlations between distributions of hunting spiders (lycosidae, ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, 25(1):1–45. 5, 11, 64

Snedecor, G. W. and Cochran, W. G. (1989). Statistical methods, eight edition. *Iowa state University press, Ames, Iowa*, 1191(2). 22

Stasinopoulos, M., Rigby, B., Mortan, N., and Stasinopoulos, M. M. (2018). Package 'gamlss. cens'. 93, 96

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R*. CRC Press. 91, 96

Tsai, C.-L., Cai, Z., and Wu, X. (1998). The examination of residual plots. *Statistica Sinica*, pages 445–465. 52

Van Noortwijk, J. M. (2009). A survey of the application of gamma processes in maintenance. *Reliability Engineering & System Safety*, 94(1):2–21. 30

Vanegas, L. H., Rondon, L. M., and Paula, G. A. (2023). Generalized estimating equations using the new r package glmtoolbox. *R J.*, 15(2):105–133. 94

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0. 59

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772. 53, 58, 68

Verbeke, G., Molenberghs, G., and Verbeke, G. (1997). *Linear mixed models for longitudinal data*. Springer. 3, 86

Wang, Y., Naumann, U., Wright, S. T., and Warton, D. I. (2012). mvabund–an r package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3(3):471–474. 64

Wei, G. C. and Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, pages 1297–1309. 87

Weisstein, E. W. (2012). Normal distribution. from mathworld-a wolfram web resource. *from MathWorld-a Wolfram Web Resource. http://mathworld. wolfram. com/NormalDistribution. html*. 27

West, B. T., Welch, K. B., and Galecki, A. T. (2022). *Linear mixed models: a practical guide using statistical software*. Chapman and Hall/CRC. 94

Wikipedia contributors (2021). Alopecosa accentuata — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Alopecosa_accentuata&oldid=1028987743. [Online; accessed 26-July-2024]. xiii, 15

Yau, K. K., Wang, K., and Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(4):437–452. 58

Zahn, D. A. (1975). An empirical study of the half-normal plot. *Technometrics*, 17(2):201–211. 41

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060. 53

Zhang, Z. and Sun, J. (2010). Interval censoring. *Statistical methods in medical research*, 19(1):53–70. 87

Zheng, L. and Zelen, M. (2009). Urn sampling, interval censoring and proportional hazard models: tests and relationships. *Statistical Modelling*, 9(4):361–379. 90