Count Mixed-Effects Regression Models in Parasite Ecology

Simão Correia
Mestrado em Estatística Computacional e Análise de Dados
Departamento de Matemática da Faculdade de Ciências da Universidade do Porto
2023

Orientador
Ana Rita Pires Gaio, Professora Auxiliar, Faculdade de Ciências da Universidade do Porto

Coorientador
Luísa Virgínia de Sousa Magalhães, Investigadora Júnior, CESAM & Departamento de Biologia da Universidade de Aveiro


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Acknowledgements

Chega ao fim mais uma etapa significativa do meu percurso académico. Contudo, é imperativo reconhecer que este trabalho não teria sido possível sem o inestimável apoio de várias pessoas. Assim, gostaria de expressar a minha mais profunda gratidão a todos aqueles que estiveram ao meu lado ao longo deste processo.

Em primeiro lugar, gostaria de expressar a minha profunda gratidão à minha orientadora, a Professora Doutora Ana Rita Gaio, por ter aceite em orientar-me neste trabalho. Agradeço por toda a exigência e rigor demonstrada, bem como por todos os valiosos ensinamentos que enriqueceram consideravelmente o meu percurso tanto a nível profissional quanto pessoal. Agradeço também pelo entusiasmo contagiante e pela confiança que sempre demonstrou, mesmo ciente de que as minhas bases académicas são do ramo da Biologia.

À minha coorientadora Luísa Magalhães, pelo apoio e amizade de sempre. Por estar sempre disponível para me seguir (e ajudar) em mais uma maluqueira. Pelo sorriso e gargalhadas, mesmo nos momentos mais difíceis. E aqui vai mais uma... já são quantos? Obrigado!

A todos os meus amigos e colegas que me acompanharam e apoiaram ao longo do curso de Estatística Computacional e Análise de Dados e todos estes anos. Em especial, gostaria de destacar o meu grande amigo e companheiro nestas maluqueiras de doutoramento e mestrado em simultâneo, Guilherme Jeremias. Agradeço pela motivação e apoio que partilhamos para enfrentar este desafiante percurso no qual nos decidimos meter. Pela compreensão nos momentos de desespero devido à falta de tempo, e acima de tudo, por nunca ter faltado aquele convite para ir surfar!

Por fim, aos mais importantes. A toda a minha família pelo sempre inalcançável apoio que me dão para que todos os meus objetivos sejam possíveis. Em particular à minha mãe por todos os sacrifícios para me ver e fazer crescer, e por sempre estar disponível para me ouvir desabafar, e ao meu irmão sempre pronto para levar comigo a melgar a cabeça e para perder nuns jogos de matraquilhos.

Um muito obrigado a todos!

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto
Departamento de Matemática

MSc. Computational Statistics and Data Analysis

Count Mixed-Effects Regression Models in Parasite Ecology

by Simão CORREIA

Count data, such as species abundance, are frequently used to analyse ecological phenomena, because the response variable only takes nonnegative integer values. The Poisson distribution is the most common and widely used for modelling count data. Nonetheless, ecological data distribution is often skewed with many zeros, along with repeated assessments that promotes an inefficient or incorrect statistical inference, unless serious attention is given to the excess of zeros, correlation structure, and how to model them effectively. In this situation, other distributions, such as Negative Binomial or Generalised Poisson, must be used. Moreover, there is increasing interest in statistical approaches for dealing with excess zeros in ecological research with Zero-Inflated and Hurdle Models. This project fulfils two main objectives: the study of regression models for count data and their application to the ecology of parasites infecting the European cockle, *Cerastoderma edule*, from the Ria de Aveiro. The data was obtained from the COACH project "CoOperative ApproACH applied to conservation and management of cockles", a project developed by researchers from Centre of Environmental and Marine Studies and University of Aveiro. In particular, this project aimed to identify which environmental variables have a determining impact on the abundance of parasite infecting cockles. The Poisson, Negative Binomial, Generalised Poisson and Binomial distributions, both with linear and non-linear additive predictors, were applied. Despite the amount of zero counts observed for metacercariae, the Poisson model did not greatly violate the equidispersion assumption, showing the importance of conducting regression analysis step by step, rather than making decisions solely based on the appearance of the data. Nevertheless, the Negative Binomial model seemed to be the one that best fitted the data. Cockle's shell length and

water salinity and pH seemed to be the most relevant explanatory variables. Additionally, dissolved oxygen also showed to be an important variable. However, the accuracy of the models' predictions was not very satisfactory.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto
Departamento de Matemática

Mestrado Integrado em Estatística Computacional e Análise de Dados

Modelos de Regressão Mistos para Dados de Contagem no Estudo Ecológico da Parasitologia

por Simão CORREIA

Os dados de contagem, como a abundância de espécies, são frequentemente utilizados para analisar fenómenos ecológicos, uma vez que a variável de resposta apenas assume valores inteiros não negativos. A distribuição de Poisson é a mais comum e amplamente utilizada para modelar dados de contagem. No entanto, a distribuição de dados ecológicos é frequentemente enviesada com muitos zeros, juntamente com avaliações repetidas que promovem uma inferência estatística ineficiente ou incorreta, a menos que seja dada uma atenção séria ao excesso de zeros, à estrutura de correlação e à forma de os modelar eficazmente. Nesta situação, devem ser utilizadas outras distribuições, como a Binomial Negativa ou a Poisson Generalizada. Para além disso, existe um crescente interesse em abordagens estatísticas para lidar com o excesso de zeros no ramo da Ecologia com modelos zeros-inflacionados (Zero-Inflated) ou modelos de barreira (Hurdle). Este projeto cumpre dois objetivos principais: o estudo de modelos de regressão para dados de contagem e a sua aplicação à ecologia de parasitas que infetam o berbigão europeu, *Cerastoderma edule*, da Ria de Aveiro. Os dados foram obtidos no âmbito do projeto COACH "Uma abordagem cooperativa à conservação e gestão de berbigão", um projeto desenvolvido por investigadores do Centro de Estudos do Ambiente e do Mar e da Universidade de Aveiro. Em particular, este projeto teve como objetivo identificar quais as variáveis ambientais que têm um impacto determinante na abundância de parasitas que infetam o berbigão. Foram aplicadas as distribuições de Poisson, Binomial Negativa, Poisson Generalizada e Binomial, ambas com preditores aditivos lineares e não lineares. Apesar da quantidade de contagens nulas observadas para metacercariae, o modelo de Poisson não

violou severamente o pressuposto da equidispersão, mostrando a importância de efetuar a análise de regressão passo a passo, em vez de tomar decisões apenas com base na aparência dos dados. No entanto, o modelo Binomial Negativo pareceu ser o que melhor se ajustava aos dados. O comprimento da concha do berbigão e a salinidade e o pH da água pareceram ser as variáveis explicativas mais relevantes. Adicionalmente, o oxigénio dissolvido também se revelou uma variável importante. No entanto, a exatidão das previsões dos modelos não foi muito satisfatória.

Contents

List of Figures

List of Tables

Glossary

Abiotic factors	Non-living part of an ecosystem that shapes its environment.
Abundance	Total number of individuals of a species in a given area.
Asexual multiplication	Type of reproduction in which the offspring comes from a single parent organism, and not from the union of gametes as in sexual reproduction. Produced offspring is usually clone of the parent.
Benthic macrofauna	Organisms visible to the naked eye ($>$ 0.5 mm) that inhabit at the bottom of a body of water, buried at the sediment or attached to a fixed substrate.
Biomass	Total mass of an organism in a given area or volume.
Biotic factors	Living part of an ecosystem that shapes its environment.
DO	Dissolved Oxygen in the water column
Ecosystem engineers	Species that modify, maintain, and/ or create habitat to other species by directly or indirectly modulate the availability of resources.
Eh	Reduction-oxidation potential (redox potential).
Eukaryote	Any single-celled or multicellular organism whose cells contain a clearly defined nucleus. Animals and plants are examples of eukaryotes.
Infaunal species	Benthic species that live buried in the sediment.
Intertidal zone	Area above the water level at low tide and underwater at high tide.

Invertebrate	An animal that lacks a vertebral column.
Keystone species	Species that helps define an entire ecosystem due to the critical role in maintaining the structure of an ecological community.
Macroparasite	Parasite that are visible to the naked eye (> 0.5 mm).
MGS	Sediment median grain size.
Molluscs	Phylum of invertebrate species that include class of animals such as bivalves (e.g., cockles), gastropods (e.g., snails), or cephalopods (e.g., octopuses).
Parasite	Organism that benefits at the expense of another organism
Population Density	Number of individuals of a species in a population relative to a given area.
Prevalence	Fraction of individuals with a specific characteristic in a given population or area.
Prokaryote	Organisms whose cells lack a nucleus and other organelles, such as bacteria.
SL	Cockle's shell length.
TOM	Total organic matter content of the sediment.
Trematode	A class of parasitic flatworms.
Trematode Cercariae	Second free-living swimming larval stage of trematode parasites that emerge from the first intermediate host and infect the second intermediate host.
Trematode Metacercariae	Second parasitic stage of trematode that encyst inside the second intermediate host.
Trematode Miracidia	First free-living ciliated larval stage of trematode parasites that hatch from an egg and infect the first intermediate host.

Trematode Sporocyst A parasitic saclike larva of trematodes infecting the first intermediate host that produces cercariae by asexual multiplication.

Vertebrate An animal that possesses a vertebral column or a backbone.

Chapter 1

Introduction

1.1 Count data

Ecologists like to count. The importance of count data for ecological research is undeniable, allowing the analyses of several ecological descriptors, such as species abundance and biodiversity, population size, or the occurrence of specific events. However, counting can be a challenging process [1].

From an ecological perspective, biological systems can be very large, enduring the difficulty of count every individual or species. Additionally, certain species can be extremely mobile, unnoticeable [2], or exhibit complex behaviours, making it challenging to detect and count them accurately and precisely [3]. Therefore, counting every single individual without resorting to sampling populations of interest or methods that can potentially introduce biases is both impracticable and unreasonable [4]. Besides, due to the natural evolution of ecological systems over time, counting must take this natural variability into consideration.

Similarly, modelling count data in statistics is not any easier. Count data follows a discrete distribution and is constrained to non-negative integer values [5, 6]. This unique characteristic of count data poses challenges when applying conventional statistical methods, such as ordinary linear regression models, which assume a continuous and normally distributed response variable, conditional on the regressors. When ordinary linear regression models are applied to count data, several issues may arise. Firstly, count data violates the assumption of normality, as it is discrete and has a limited range [7]. Secondly, the assumption of constant variance is easily violated, as the variance is equal to

the mean and therefore also varies with the regressors. Finally, count data are not compatible with the direct modelling of the mean of the response, as counts cannot be negative. To address these challenges, a common approach in the analysis of count data is to start by fitting a Poisson (mixed-effects) model and then to evaluate its assumptions to check whether it is necessary to change the distribution of the response variable (possibly, and most commonly, to a Negative Binomial distribution) [8, 9]. In these models, the link function connects the linear predictor to the expected count, accommodating the constrains of count data [8, 9].

The most widely used regression model for count data is the Poisson regression [8]. However, its assumption of equidispersion, that the distribution's mean is equal to its variance, in contrast to other count-based regression models, is one of its drawbacks. If that assumption is not met, and under or overdispersion is observed, the Poisson regression model may provide inaccurate standard errors for the model coefficients [5, 10]. Overdispersion happens when the variance of the conditional response exceeds the mean. This is the case for many ecological data. To account for this excess variability, one may use the negative binomial distribution instead. The negative binomial distribution is an extension of the Poisson distribution that allows for overdispersion; the variance has a quadratic relationship with the mean, through the addition of an additional parameter ϕ [10]. Whenever ϕ converges to 1, the negative binomial distribution converges to the Poisson distribution.

In addition to the Poisson and Negative Binomial distributions already mentioned, other typical distributions for count data include the Generalised Poisson [11, 12] and the Conway-Maxwell-Poisson [13]. While a Negative Binomial regression can only model overdispersed phenomena, Generalised Poisson and Conway-Maxwell-Poisson introduce a new parameter that enables modelling of both underdispersion and overdispersion, making it applicable to a variety of count data scenarios. Chapter 2 presents further details on the distributions for count data. In this manner, the nature of the conditional response variable will determine the selection of the distribution function.

Further challenges of analysing count data, particularly those associated with ecological sampling, may include the skewness related to zero-inflation, and/or repeated or longitudinal assessments [8, 10, 14]. If substantial consideration is not given to the excess of zeros and/or to the correlation structure, and how to model them effectively, this promotes an inefficient or incorrect statistical inference. On the other hand, longitudinal or

repeated measurements data are handled by regression models with mixed effects or regression models estimated by the Generalised Estimating Equations (GEE) [15, 16]. This statistical method is tailored to accommodate the inherent complexities of data collected over multiple time points or repeated measured from the same subjects.

There has been a considerable growing interest in statistical tools that deal with excess zeros in ecology research, with zero-inflated (ZI) and hurdle (H) models being commonly employed to fit such data [14]. These models differ in the way they deal with zeros. In models of inflated zeros, null counts can be originated from a true absence of observations in the counting process, being designated as true zeros, or introduced due to process issues, designated as false zeros. From a practical point of view, true zeros correspond to individuals that, in fact, are not present when the sampling process is carried out, while false zeros refer to individuals that were not observed due to problems in the sampling process. In the case of hurdle models, the counting process is truncated at zero and, therefore, cannot produce null counts. Moreover, the model separates the counts into zeros (absences) *versus* non-zeros (presences).

The ecological study of host-parasite interactions is one illustration of this intricacy.

1.2 Motivation

Bivalves are a dominating component of the coastal benthic macrofauna, both in terms of abundance and biomass [17]. These organisms, which are regarded as keystone species and perform significant roles in the ecosystem support the marine environment's resilience [18]. The filter-feeding habit and bioturbation activity of marine bivalves enable them to perform several crucial ecological functions, including carbon storage and energy cycling [19]. Additionally, they serve as a connection between primary producers and higher trophic levels in the ecological food webs [20]. Bivalves are also considered ecosystem engineers by altering the environment and promoting life conditions for other infaunal species due to their burrowing activity. Additionally, bivalves provide the foundation of significant commercial activity, playing a crucial socioeconomic role [21, 22].

The infaunal suspension-feeder bivalve *Cerastoderma edule*, the European edible cockle, is a common and widely dispersed bivalve species along the northeast Atlantic coast from Norway, in northern Europe [23, 24], to Mauritania, in northern Africa [25]. This bivalve holds significant importance in Europe, and particularly in Portugal, due to its ecological, economic, and cultural value. Cockle harvesting is an essential economic activity in many

coastal regions of Europe [26, 27]. The species is commercially valuable and contributes to the livelihoods of local fishermen and seafood processors. The sale of cockles, whether for domestic consumption or export, helps generate income and employment opportunities for coastal communities. In 2015, Portuguese cockle harvesting represented 20% of total European captures, accounting for a 4.5 million euros revenue [28]. Cockles are, therefore, integral part of Portugal's and Europe's culinary tradition with an important cultural footprint [29].

Beyond its socioeconomic value, *C. edule* plays a crucial role in maintaining the health of coastal ecosystems [30]. As filter feeders, cockles help improve water quality by filtering and removing organic particles and pollutants from the water [31]. This helps maintain the health of coastal ecosystems, supporting other marine life and recreational activities [32]. To ensure its extended importance, sustainable management practices and conservation efforts are essential to protect the species and its habitat for future generations.

In many European regions, cockle populations have been suffering from periodic mass mortalities with increasing frequency and intensity in the last decades [33, 34]. The great variability of effectives in the populations has serious consequences for natural stocks [33]. Emergent diseases, overfishing, inefficient management and degradation of the environmental conditions have been pointed as the main drivers of cockles' production decline that leads to high economic and ecological impacts. This scenario has severe consequences for the social structure of coastal communities, and for the wider ecosystem services and societal benefits provided by cockles. In addition to anthropogenic influence, cockle population dynamics are naturally controlled by abiotic and biotic factors such as temperature or parasitism, respectively.

In fact, from prokaryotic to eukaryotic species, it is recognized that cockles are hosts to a wide variety of parasites and diseases [35, 36]. The infection of some of these parasites leads to sub-lethal impacts in the host. However, high prevalence and abundance outbreaks can occur and devastatingly impact the host wild populations, related fisheries, and aquaculture industries [37]. There are several studies demonstrating the effect of parasites on cockles individuals or population dynamics. Nevertheless, the factors that trigger parasites abundance are still unclear. Some studies predict that in a climate change scenario, marine diseases are likely to become more frequent and severe [38, 39]. Many studies have been conducted on environmental abiotic variables and their use to predict

parasite abundance and prevalence. Nonetheless, results are still unresolved with conclusions being regionally dependent and/or variable according to the host/parasite model used. For instance, while studying trematode parasites in their bivalve host, some authors described higher parasite prevalence in the case of increased salinities [40], but the reverse was reported as well [41]. Additionally, these studies occasionally include geographical and temporal data [42, 43]. Locations or years that are closest to one another are anticipated to have less parasite variation than years that are further apart. The same is anticipated for years or areas that exhibit comparable abiotic circumstances. However, the sampling effort or the fact that there are issues with the counting procedures that have previously been highlighted might make it difficult to understand the data.

In order to forecast outbreaks of parasite prevalence and abundance, it is of uttermost importance to investigate and analyse how environmental variables affect parasite prevalence and abundance. Nevertheless, as was already mentioned, the modelling of this kind of data displays several challenges. In that regard, employing the edible cockle *C. edule* and the parasite community infecting this bivalve as host-parasite model, the major objectives of this research are the study of regression models for counting data, namely for data with a large percentage of zeros, and their application to ecological parasitology data, specifically, identifying the main drivers of parasite infection.

1.3 Thesis Structure

This project was divided into several chapters, with a brief introduction to the topic to be addressed and a description of the chapter's content at the beginning of each chapter. **Chapter 1 – Introduction** presents an introductory contextualization regarding the regression models for count data and their challenges, the importance of bivalves and the objectives of this project. **Chapter 2 – Count Data** includes the theoretical basis for generalized linear models and zero-inflated models, which supports the main methodologies addressed in this thesis. The regression models used in modelling count data and the logistic regression are discussed in detail in this chapter. **Chapter 3 – R** displays the different libraries available on R software to deal with Generalised Linear (Mixed) and Generalised Additive (Mixed) Models and their functionalities. **Chapter 4 – Application to Ecological Parasitology Data** encompasses the entire data pre-processing process, descriptive analysis of the same and application of the portrayed methodology. **Chapter 5 –Model Formulation** will present the results obtained from the application of the methodologies

portrayed in the previous chapters **Chapter 6 – Final Remarks** will cover all the conclusions drawn from this project, along with suggestions and/or problems to be addressed in future works.

Chapter 2

Count Data

Count data, as the name suggests, comprise integer values that result from counting. They can be accounted for by a random variable often taking values starting at zero. However, theoretically, any positive integer could serve as the lower boundary.

Models are simplified representations of reality, and they can be deterministic or probabilistic. In statistical modelling, the latter is prevalent, incorporating probabilistic components [44]. In practical scenarios, researchers often aim to analyse how one or more explanatory variables, measured in units like individuals or objects, influence a response variable or outcome. This analysis is very often conducted using regression models [8].

The classical statistical model for regression is the linear model [45]. This model addresses the conditional mean of the response as a linear combination of explanatory variables and assumes a normal distribution for the errors. However, in certain cases, linearity and normality may not be realistic, and no transformations can make them valid. Modelling count data is such a situation. These models are a specific category of discrete regression models. To address situations not fitting the linear model framework, generalized linear models were introduced [9]. These, assume the response follows a distribution from the exponential family, and establish a potentially non-linear relationship between the response's average and a linear combination of explanatory variables, through an adequate link function. Estimation in these models relies on the maximum likelihood method, with iterative numerical techniques used to maximize the likelihood function.

In this chapter, we explore count data distributions, such as the Poisson, Negative Binomial, Generalised Poisson and Binomial, and generalised linear models.

2.1 Poisson distribution

The reference probabilistic model, and often the first choice for count data, is the Poisson model. The Poisson distribution is a probability distribution that describes the number of events that occur in a fixed interval of time or space when these events occur independently at a constant rate [8]. A random variable Y follows a Poisson distribution with parameter λ, $P(\lambda)$, if its probability function is given by

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \qquad y = 0, 1, 2, \ldots, \tag{2.1}$$

It can be easily shown that $E[Y] = \lambda$.

The primary property of the Poisson model is equidispersion, i.e., the variance of the counts is the same as their mean value. However, the property of equidespersion may be violated by the data, meaning that over- or underdispersion is often encountered in real events, requiring the use of different distributions that do not rely on this assumption.

Another key property of the Poisson distribution is its additivity, i.e, the sum of a finite number of independent Poisson-distributed random variables, is also Poisson-distributed. Mathematically, if you have random variables $Y_1, Y_2, \ldots, Y_n$, each of which follows a Poisson distribution with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, respectively, and these random variables are independent from each other, then the sum $Y = X_1 + X_2 + \ldots + X_n$ follows a Poisson distribution with a parameter equal to the sum of the individual parameters:

$$Y \sim P(\lambda_1 + \lambda_2 + \ldots + \lambda_n)$$

The Poisson distribution is characterised in several ways, but two in particular stand out, namely the Poisson as the Law of Rare Events and the Poisson Counting Process [8].

The Law of Rare Events applies when the counts occur in a large number n of independent Bernoulli trials with the success probability π of each trial being small. The Poisson probability distribution corresponds to the limiting case $n \to +\infty$ and $\pi \to 0$, with $n\pi = \lambda > 0$ constant.

In turn, the Poisson Counting Process characterizes complete randomness, and extends the Poisson distribution to describe events occurring over continuous time intervals, assuming events happen randomly and independently.

A counting process $N(t)$ can be defined as a count of events up to time t, where $N(t)$ is a nonnegative, integer value that must meet the property that $N(s) \leq N(t)$ if $s \leq t$,

and $N(t) - N(s)$ is the number of events in the interval (s, t). If λ is the constant rate of occurrence λ of the event of interest and $N(s, s+h)$ is the number of occurrences in the time interval $(s, s+h)$. Then, $N(s, s+h)$, for nonlimit h, can be shown to follow a Poisson distribution with mean λh:

$$P[N(s, s+h) = r] = \frac{e^{-\lambda h}(\lambda h)^r}{r!} \qquad r = 0, 1, 2, \ldots \tag{2.2}$$

Normalizing the length of the exposure time interval to be unity, $h = 1$, leads to the Poisson density $P(\lambda)$.

In Figure 2.1 it is possible to observe examples of various instances of the Poisson distribution corresponding to different mean, (μ), values.

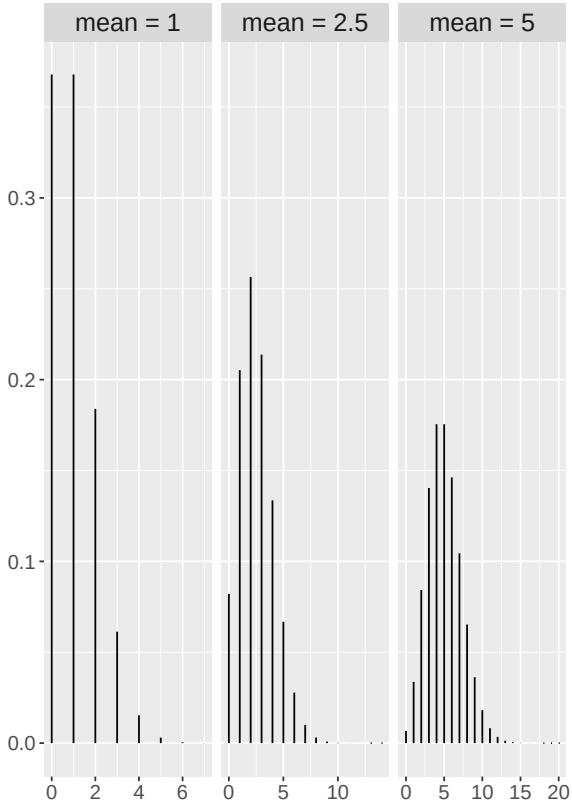


FIGURE 2.1: Poisson distribution for various mean (μ) values.

2.2 Negative Binomial

When dealing with count data that exhibits overdispersion, the most commonly used probability distribution is the Negative Binomial distribution. The Negative Binomial

distribution is a probability distribution that describes the number of Bernoulli trials required for a specified number of successes to occur. The Negative Binomial distribution has two parameters: the mean (μ) and the dispersion parameter (k). The mean represents the average count, and the dispersion parameter controls the level of overdispersion [10]. A random variable Y follows a Negative Binomial distribution with parameters k and p,

$$Y \sim NB(k, p),$$

if Y represents the number of failures previous to k successes, in a set of independent events with the same probability of success, p. The probability function of Y is given by

$$P(Y = y) = \binom{y + k - 1}{k - 1} \cdot (1 - p)^y \cdot p^k \tag{2.3}$$

There are several descriptions of the Negative Binomial, mostly depending on the variance definition, however in this work the focus will fall on two in particular: the linear Negative Binomial (NB1) and the traditional Negative Binomial regression model, the quadratic Negative Binomial (NB2).

The above definition is $NB1(k, p)$. It can be shown that

$$E(Y) = k(1 - p)/p \tag{2.4}$$

and,

$$\begin{aligned} Var(Y) &= \frac{1}{p} \cdot E(Y) \\ &= (1 + \alpha) \cdot E(Y) \\ &= E(Y) + \alpha E(Y) \end{aligned} \tag{2.5}$$

for $\alpha = \frac{1}{p} - 1$. In particular, the variance is linear on the mean.

The quadratic negative binomial distribution (NB2) is an example of a Poisson-gamma mixture distribution. More precisely, $Y \sim NB2(\mu, \alpha)$ if $Y \sim P(\mu V)$ with $V \sim \Gamma(\frac{1}{\alpha})$

In this case, the probability mass function for the $NB2(\mu, \alpha)$ distribution is given by:

$$\begin{aligned} P(Y = y|\mu, \alpha) &= \frac{\Gamma(y + \frac{1}{\alpha})}{\Gamma(y + 1) \cdot \Gamma(\frac{1}{\alpha})} \cdot \left(\frac{1}{1 + \alpha\mu}\right)^{\frac{1}{\alpha}} \cdot \left(\frac{\alpha\mu}{1 + \alpha\mu}\right)^y \\ &= \binom{y + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1 + \alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1 + \alpha\mu}\right)^y \qquad \alpha > 0, 1, 2, \dots \end{aligned} \tag{2.6}$$

The Poisson distribution is obtained as a limit distribution of the Negative Binomial distribution when $\alpha = 0$.

By transforming $k = \frac{1}{\alpha}$ and $p = \frac{1}{1+\alpha\mu}$ we obtain

$$P(Y = y|k, \mu) = \binom{y+k-1}{k-1} p^k (1-p)^y \tag{2.7}$$

Showing the equivalence between the NB2 and NB1 formulations. It can be shown that, if $Y \sim NB2(\mu, \alpha)$, then $E(Y) = \mu$ and $Var(Y) = \mu + \alpha\mu^2$ (or $Var(Y) = \mu + \frac{\mu^2}{k}$). In particular, the variance is quadratic on the mean.

The Figure 2.2 displays examples of Negative Binomial distributions for different values of mean, (μ) and k. As k increases in relation to μ^2, the term $\frac{\mu^2}{k}$ tends towards zero, resulting in the convergence of the Negative Binomial distributions towards a Poisson distribution (where $Var(Y) = \mu$).

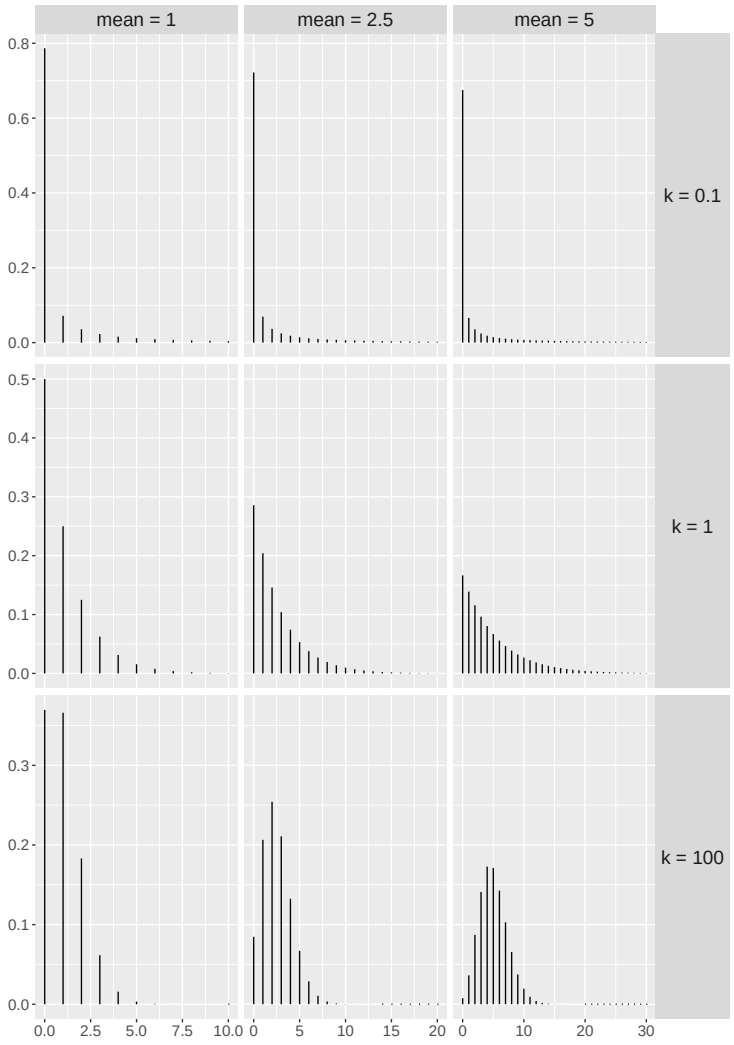


FIGURE 2.2: Negative Binomial probability distribution for various mean (μ) and k values.

Of note, we mention here that a Negative Binomial distribution can only model overdispersed phenomena, not underdispersion.

2.3 Generalised Poisson

Similar to the Negative Binomial distribution, there are other distributions that enable the modelling of overdispersion of count data. The Generalised Poisson (GP) distribution is one of those distributions. This distribution is a flexible extension of the Poisson distribution and, in addition to accommodating data with overdispersion, it also permits the modelling of underdispersed data [46].

Generalised Poisson is a two-parameter discrete distribution. One of the parameters measures the location and the other measures the dispersion. The distribution is unimodal, and it can be skewed to the right or the left. It approaches the normal distribution when the location parameter gets very large [11].

A discrete random variable Y is said to follow a Generalised Poisson distribution with parameters θ and λ

$$Y \sim \text{GP}(\theta, \lambda)$$

if its probability distribution is given by

$$P(Y = y) = \begin{cases} \theta(\theta + \lambda y)^{y-1} e^{-\theta - \lambda y} / y!, & y = 0, 1, \ldots \\ 0, & \text{for } y > m \text{ when } \lambda < 0. \end{cases} \tag{2.8}$$

and zero otherwise, where $\theta > 0$, $\max(-1, -\theta/4) \leq \lambda < 1$ and m is the largest positive integer for which $\theta + m\lambda > 0$ when λ is negative. The probability function reduces to the Poisson distribution when $\lambda = 0$. The parameters θ and λ are independent, but the lower limits on λ and $m \geq 4$ are imposed to ensure that there are at least five classes with nonzero probability when λ is negative [11].

Contrarily to the distributions described so far, the Generalised Poisson distribution belongs to the Lagrangian distributions $L(f, g, x)$, where $f(z) = e^{\theta(z-1)}$, $\theta > 0$, and $g(z) = e^{\lambda(z-1)}$, $0 < \lambda < 1$.

The mean of the Generalized Poisson distribution is given by

$$E[Y] = \frac{\theta}{1 - \lambda} \tag{2.9}$$

and the variance can be defined as

$$Var[Y] = \frac{\theta}{(1-\lambda)^3} \tag{2.10}$$

When $\lambda = 0$ the GP distribution reduces to the Poisson distribution. When $\lambda > 0$ it indicates that the data have higher variability than what would be expected in a Poisson distribution (overdispersion), while $\lambda < 0$ occurs when data is underdispersed, meaning that the data have lower variability than in a Poisson distribution.

Examples of the Generalised Poisson distribution are discernible in Figure 2.3. As previously stated, when values of λ are closer to 0, the Generalised Poisson distribution tends to a Poisson distribution. For λ values greater than 0, the data exhibits overdispersion and is more akin to a Negative Binomial distribution. Figure 2.3 was created in `R` using the `VGAM` package, which unfortunately does not support nonnegative values of λ, making it impossible to portray underdispersed data in the plot.

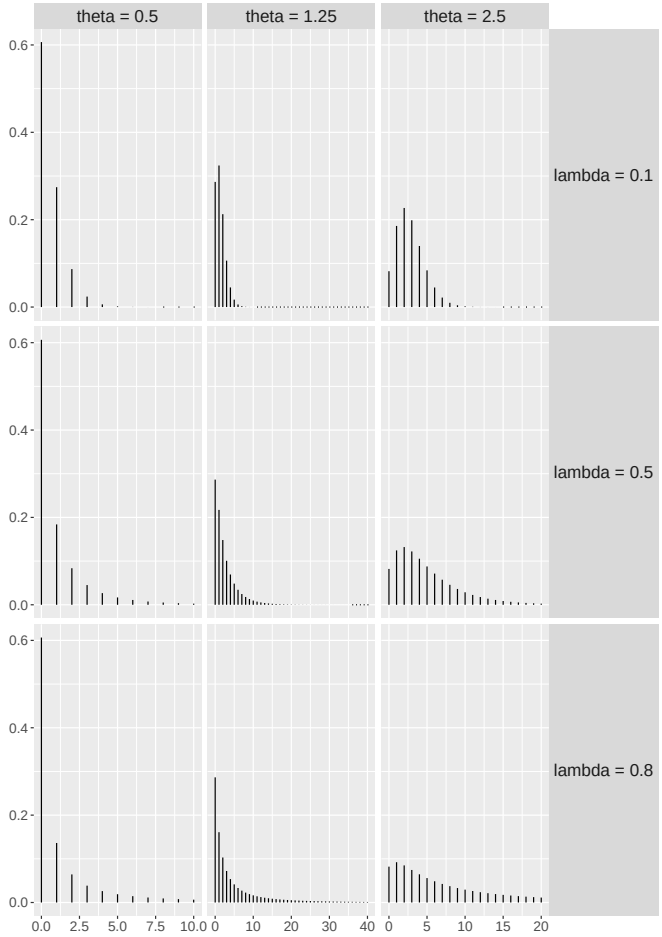


FIGURE 2.3: Generalised Poisson distribution examples for different λ and θ values.

2.4 Binomial

For situations that only two possible outcomes (success or failure) are possible, the Binomial distribution is a valuable and the go-to probability distribution [47, 48]. It is a fundamental distribution in statistics and probability theory that models the number of successes in a fixed number of independent experiments. More precisely, that a random variable Y follows the Binomial distribution $B(n, \pi)$, with parameters n and π, if Y counts the number of successes in n Bernoulli trials whenever the probability of success is π.

The probability function of Y is given by:

$$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y} \tag{2.11}$$

In this case,

$$E(Y) = n\pi \tag{2.12}$$

and,

$$Var(Y) = n\pi(1 - \pi) \tag{2.13}$$

As in Poisson distribution, the Binomial distribution does not have a dispersion parameter.

In Figure 2.4 it is possible to observe different Binomial distributions for different probabilities of success, π, and different number of successes in Bernoulli trials, n. For a small π (¡ 0.5) the distribution is asymmetrical with a positive skewness, whereas, for large probabilities of success (π ¿ 0.5) the Binomial distribution is positively skewed. When π is equal to 0.5, the distribution is symmetrical.

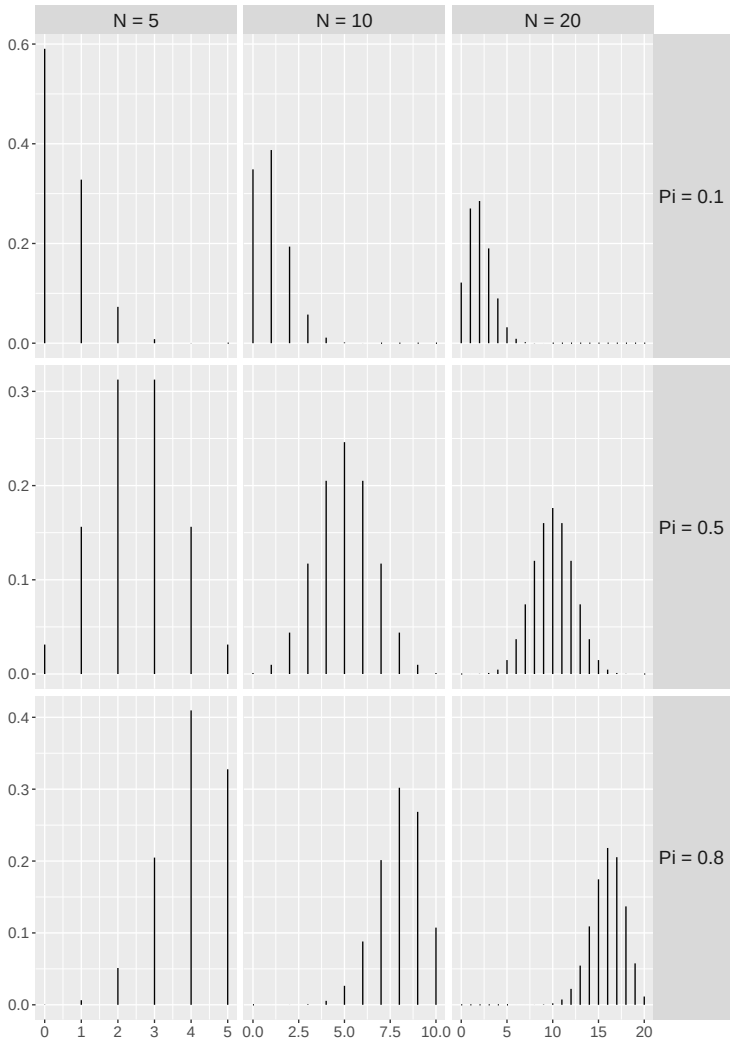


FIGURE 2.4: Binomial distributions for different number of successes, n, and probabilities of success (μ).

2.5 Generalised Linear Models

Generalised Linear Models (GLMs) [49] are a class of regression models that extend the ordinary linear model to a broader range of relationships between a response variable and its predictors. Opposed to linear regression models, which directly model the conditional mean of the response through a linear combination of the predictors, GLMs use the linear predictor to model a transformation of the conditional mean of the response [9, 50]. The distribution of the conditional response is also allowed to vary in a family of distributions, which include the normal distribution as a special case. GLMs consist of three components: the random component, the linear predictor (or systematic component), and the link function.

2.5.1 Random Component

The random component specified the probability distribution of the response variable, Y_i, conditioned by a set of predictors $X_1, \ldots, X_p$, i.e., $Y_i|X_1, \ldots, X_p$, stating that it has to be a member of the exponential family of distributions.

2.5.1.1 Exponential Family

A random variable Y follows a distribution belonging to the Exponential Dispersion Model family (EDM) [51] if it has a probability (density) function that can be modelled into the form of

$$P(Y = y|\theta, \phi) = \exp\left(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

where θ is the canonical parameter, ϕ is a dispersion parameter, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are functions that vary depending on the exponential family. Specifically, $a(\cdot)$ is a function that depends only of the dispersion parameter ϕ, $b(\cdot)$ is a function that depends on the vector of location parameters θ and $c(\cdot)$ depends on the random variable, Y, and the dispersion parameter, ϕ.

For instance, the Poisson distribution belongs to the EDM family since it can be written as

$$\begin{aligned} P(Y = y|\theta) &= \tfrac{\theta^y}{y!}\exp(-\theta) \\ &= \tfrac{1}{y!}\exp(y\log(\theta) - \theta) \end{aligned}$$

In this case, $a(\phi) = 1$, $b(\theta) = \exp^{\theta}$, and $c(y, \phi) = -log_e(y!)$.

The Negative Binomial and Generalised Poisson distributions do not belong to the EDM family. However, if the dispersion parameter producing the overdispersion (or underdispersion) is treated as a known, fixed constant, it can be used as a member of the EDM family [9, 52].

2.5.2 Linear Predictor (Systematic Component)

The linear predictor (or systematic component) represents the linear structure η produced by the predictors (exploratory variables):

$$\eta = \beta_0 + \sum_{j=1}^{p} X_j\beta_j$$

where, $\beta = (\beta_0, \beta_1, \ldots, \beta_j)$ is the unknown vector of regression parameters.

2.5.3 Link Function

The link function is the bridge between the linear predictor and the conditional expected value of the response variable. It transforms the unbounded and linear scale of the linear predictor, η, into a suitable range for the response variable. The link function is denoted by $g(\cdot)$ and we have

$$g(E(Y|X_1, X_2, \ldots, X_p)) = \eta$$

The choice of the link function ensures that the predicted values from the linear predictor are on the appropriate scale for the chosen distributions. Thus, different GLMs use different link functions depending on the probability distribution applied. For example

$$\begin{aligned} \textbf{Gaussian distribution}\text{identity link}: \quad g(\mu) &= \mu \\ \textbf{Poisson distribution}\text{log link}: \quad g(\mu) &= \log(\mu) \\ \textbf{Binomial distribution}\text{logit link}: \quad g(\mu) &= \log(\tfrac{\mu}{1-\mu}) \end{aligned}$$

where, for brevity, we use $\mu = E(Y|X_1, X_2, \ldots, X_p)$.

Chapter 3

R Functionalities

R [53] is a versatile and powerful open-source programming language that it is primarily used for statistical computing and data analysis. Due to a strong focus on extensibility, enabling the development and incorporation of a vast variety of packages and extensions, R software is constantly evolving. As such, this flexibility has led to R becoming an increasingly popular tool for the analysis of different data-related tasks.

In this project, we will be dealing with longitudinal count data in a regression framework, which will require the inclusion of random effects for the modelling of the time-dependencies across different observations from the same experimental unit. In R, the fitting of such mixed-effects regression models is not an easy task. The inexistence of closed-form solution led to the development of several numerical approximations, which, in turn, were included in different R-packages. In this chapter we summarize the most well-known such packages and the different functionalities of each will be compared and contrasted. Additionally, a more thorough exploration of the recently developed R package DHARMa will be carried out due to its value in interpreting and validating the models that have been fitted to the data.

3.1 R packages

Several R packages are available for fitting Generalised Linear Mixed Models (GLMMs) and Generalised Additive Mixed Models (GAMMs). These packages offer a distinct range of capabilities and features, enabling researchers to tailor their modelling approaches to the specific requirements of their data. The R packages `nlme` [54, 55], `lme4` [56], `GLMMadaptive` [57], `glmmML` [58], `glmmADMB` [59], `glmmTMB` [60], `mgcv` [61], `gamm4` [62], `VGAM` [63], `gamlss` [64],

`MCMCglmm` [65], `brms` [66], `R-INLA` [67], `R2jags` [68], and `pscl` [69] are particularly noteworthy for their contribution to count data analysis in a regression framework.

Starting with `nlme` and `lme4`, these two versatile `R` packages are the most widely used for fitting GLMMs. Both packages employ the Restricted Maximum Likelihood (REML) criterion with Laplace approximation as their default method for GLMM fitting, while also allowing for the specification of Maximum Likelihood (ML) estimates. The main differences observed between them pertain to the available link functions. For instance, `nlme` cannot be employed to fit outcomes with non-Gaussian distributions. In contrast, `lme4` can be used to fit mixed-effects regressions with a (conditional) response following a Binomial, Poisson or Negative Binomial distribution. Additionally, `nlme` offers the capability to specify the variance-covariance matrix for the random effects, a feature not supported by `lme4`. Nevertheless, both packages lack the ability to handle zero-inflated and hurdle models.

For data that requires the use of zero-inflated models, more suitable options include packages like `GLMMadaptive`, `glmmML`, `glmmADMB`, or `glmmTMB`. `GLMMadaptive` is a package concerned with fitting GLMMs using an adaptive Gauss-Hermite quadrature approximation, facilitating flexible modelling of non-Gaussian response distributions. It supports a broad range of variance and covariance structures for random effects, making it appropriate for a wide range of data types. `glmmML`, on the other hand, estimates GLMMs using maximum likelihood and is particularly proficient at fitting models with binomial and Poisson distributions. While it is user-friendly and efficient, it may not accommodate as many distributional assumptions as other packages. `glmmADMB` is an `R` package that estimates GLMMs using the ADMB (Automatic Differentiation Model Builder) software. It is ideal for models with complex random effect structures and non-linear predictor terms. For likelihood estimation, ADMB employs a Laplace approximation, which can yield accurate results but may be computationally intensive for big datasets. However, it is noting that `glmmADMB` is not suitable for models where the degree of zero-inflation varies across observational units, making it most appropriate for scenarios where all observational units share an equal probability of producing structural zeros. `glmmTMB` offers an alternative to the `glmmADMB` package, using a Template Model Builder (TMB) that is more flexible and efficient in computing GLMMs, thereby reducing computational demands. It approximates the probability with a Laplace approximation, which can be highly accurate for diverse response distributions, and uses the ML estimator. Nevertheless, `glmmTMB`

lacks an alternative option for REML or Gauss-Hermite quadrature estimations to integrate over random effects, which could lead to suboptimal performance when limited information is available for each random effect level.

In a Bayesian framework, the most suitable `R` packages for fitting GLMMs include `pscl`, `MCMCglmm`, `brms`, `R-INLA`, or `R2jags`. The widely used `pscl` package excels in fitting zero-inflated and hurdle Generalised Linear Models by employing Maximum Likelihood estimation to incorporate predictor variables in the zero-inflation component. However, one limitation of the `pscl` package is its inability to account for correlation within sampling units arising from repeated samples (mixed effects). Neglecting random effects in the modelling process can lead to overly optimistic statistical inferences, making them less conservative. `MCMCglmm` and `brms` packages are capable of fitting zero-inflated GLMMs with predictors of zero-inflation, albeit they employ different criterion methods for the Markov chain Monte Carlo (MCMC) sampling. `MCMCglmm` utilises the Metropolis-Hastings updates or the slice sampling method as the deviance information criterion, while `brms` is built on the Stan programming language, implementing Hamiltonian Monte Carlo and the No-U-Turn Sampler (NUTS). One of the primary issues with Metropolis-Hastings algorithms is their relatively slow convergence for high-dimensional models, compared to the faster convergence of Hamiltonian Monte Carlo algorithms for the same type of data. `R-INLA` shares a similar limitation with `glmmADMB` in terms of variable degrees of zero-inflation across observational units. The `R2jags` package implements Bayesian analysis of GLMMs in JAGS, including monitoring convergence of MCMC models using Rubin and Gelman Rhat statistics and implementing parallel processing of MCMC models for multiple chains.

Finally, for GAMMs, the most commonly used packages include `mgcv`, `gamlss`, `VGAM`, and `gamm4`. The `gamlss` package offers flexibility by fitting Generalised Additive Models with predictors on all parameters of a distribution, covering a wide range of zero-inflated and hurdle distributions. On the other hand, `mgcv` can fit zero-inflated GAMMs, but only when using a Poisson distribution for the predictors of zero-inflation.

The `R` packages utilising these methods, as previously described, are presented in Table 3.1, with a description and comparison of the main functions that each package provide.

R packages and main functions available in R for GLMM and GAMM

TABLE 3.1: R packages and main function available for Generalised Linear and Generalised Additive Mixed Models (GLMMs & GAMMs) computing in R.

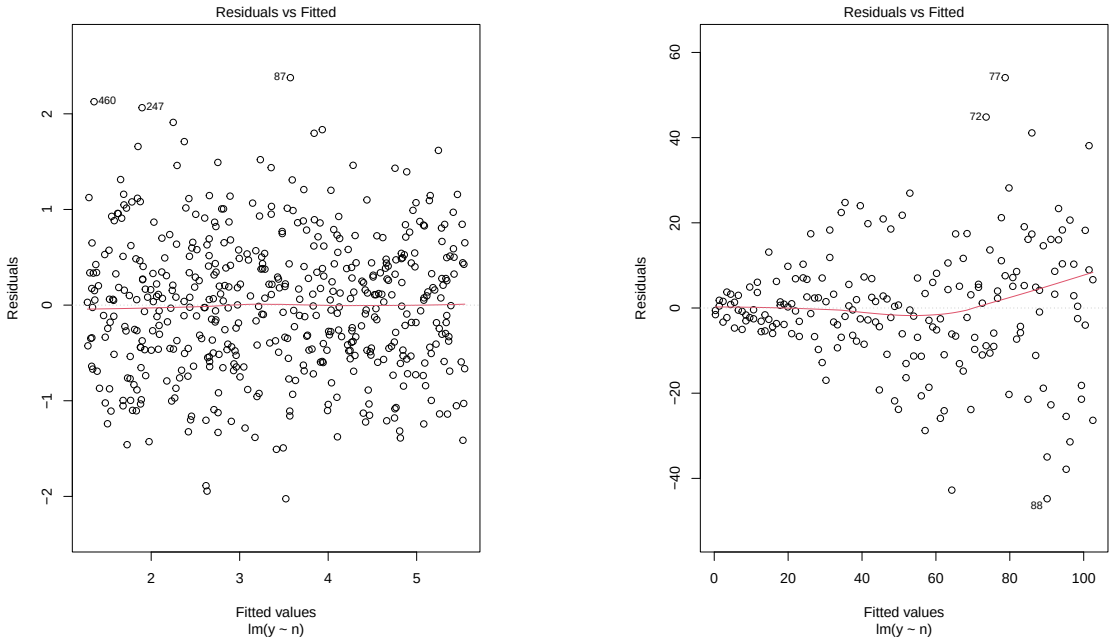
Packages	Functions	Description	Approach		Ref
			Frequentist	Bayesian	
nlme	lme()	Fits a linear mixed-effects model, allowing for nested random effects. The within-group errors are allowed to be correlated and/or have unequal variances.	✓		[54, 55]
	nlme()	Fits a non-linear mixed-effects model, allowing for nested random effects. The within-group errors are allowed to be correlated and/or have unequal variances.			
	gls()	Fits a linear model using generalised least squares.			
	gnls()	Fits a non-linear model using generalised least squares.			
	summary.'function'()	Summarises the fitted object information for the function of interest. For example, summary.lme provides information about the coefficients of the fitted model.			
	corGaus()	Constructs a Gaussian spatial correlation structure that needs to be initialised.			
	corLin()	Constructs a Linear spatial correlation structure that needs to be initialised.			
lme4	lmer()	Fits Linear Mixed Models	✓		[56]
	nlmer()	Fits Non-linear Mixed Models			
	glmer()	Fits a Generalised Linear Mixed Model			
	glmer.nb()	Fits a Generalised Linear Mixed Model with Negative Binomial distribution			
	fixef()	Extracts fixed effects estimates			
	getME()	Extract components from a fitted mixed effects model			
GLMMadaptive	mixed_model()	Model-fitting function with four required arguments (fixed: formula for fixed effects; random: formula for random effects; family; and data)	✓		[57]
	effectPlotData()	Predictions with confidence interval for constructing effects plots			
glmmADMB	glmmadmb()	Fits Generalised Linear Mixed Models and extensions.	✓		[59]
	coefplot2()	coefplot2() belongs to the coefplot2 package. However, this function can read glmmADMB objects to plot coefficients.			
	coef()	Extract fixed effect coefficients.			
	ranef()	Extract random effect coefficients.			
glmmTMB	glmmTMB()	Fits a generalised linear mixed model (GLMM) using Template Model Builder (TMB).	✓		[60]
	fixef()	Extract fixed-effects estimates			
	ranef.glmmTMB()	Extract random-effects estimates			
	confit.glmmTMB()	Calculate confidence intervals			
	getME.glmmTMB()	Extract Generalise Components from a Fitted Mixed Effects Model			
	dtruncated_poisson	Probability function for k-truncated Poisson distribution.			
	dtruncated_nbinom2	Probability function for k-truncated Negative Binomial distribution.			
pscl	hurdle()	Fits an hurdle regression models for count data via maximum likelihood		✓	[69]
	hurdletest()	Wald test of the null hypothesis that no zero hurdle is required in hurdle regression models for count data.			
	zeroinfl()	Fits a zero-inflated regression models for count data via maximum likelihood.			
	predprob()	Compute predicted probabilities from fitted models			
MCMCglmm	MCMCglmm()	Markov chain Monte Carlo Sampler for Multivariate Generalised Linear Mixed Models with special emphasis on correlated random effects arising from pedigrees and phylogenies		✓	[65]
	summary.MCMCglmm()	Summarises GLMM fits from MCMCglmm			
	predict.MCMCglmm()	Predicted values for GLMMs fitted with MCMCglmm			
	residuals.MCMCglmm()	Return the residuals for a GLMMs fitted with MCMCglmm			
	simulate.MCMCglmm()	Simulated response vectors for GLMMs fitted with MCMCglmm			
brms	brm()	Fits a Bayesian generalized (non-)linear multivariate multilevel models using Stan for full Bayesian inference.		✓	[66]
	arma()	Set up an autoregressive moving average (ARMA) term of order (p, q) in brms. it exists purely to help set up a model with ARMA terms.			
	conditional_effects()	Display conditional effects of one or more numeric and/or categorical predictors including two-way interaction effects.			
R-INLA	inla()	Provides a method of fitting GLM and GLMMs models through a bayesian approach		✓	[67]
	summary.fixed	Display the fitted model summary for the fixed effects			
	summary.random	Display the fitted model summary for the random effects			
mgcv	gam()	Fits a Generalised Additive Model to data.	✓		[61]
	gamm()	Fits a Generalised Additive Mixed Model to data.			
	bam()	Fits a Generalised Additive Model to a very large data set.			
	vis.gam()	Produces plot views of the gam model predictions.			
gamlss	gamlss()	Returns an object of class "gamlss", which is a generalized additive model for location scale and shape	✓		[64]
	gamlss.MX()	Fits a K fold non parametric mixture of gamlss family distributions			
	gamlss.MX()	Fits a K fold non parametric mixture of gamlss family distributions			
	gamlssNP()	This function fits a finite (or normal) mixture distribution where the kernel distribution can belong to any gamlss family			
VGAM	vglm()	Fits a generalized linear model (RR-VGLM)	✓		[63]
	vgam()	Fits a vector Generalised Additive Model			
	rrvglm()	Fits a reduced-rank vector generalized linear model (RR-VGLM)			
	residualsvglm	Displays the residuals for a vector generalized linear model (VGLM) object.			
gamm4	gamm4()	Fits the specified generalized additive mixed model (GAMM) to data	✓		[62]
	object$gam	Summarises the fitted gamm results.			
	object$mer	Summarises the mixed model part.			

During this project, data modelling was performed using functions available mainly from the glmmTMB and gamm4 packages. These packages were chosen to fit the data to generalised linear mixed models and generalised additive mixed models, respectively, due to the number of distributions available, allowing for a simple comparison between the models.

3.2 DHARMa

`DHARMa` (Diagnostics for HierArchichal Regression Models) [70] is a recent R-package, introduced in 2016, designed to assess the goodness-of-fit, and diagnose potential problems in hierarchical (or multilevel) regression models.

Residuals analysis is a crucial step in regression modelling to assess model assumptions, identify potential issues, improve model performance, and validate the model's reliability. It provides valuable insights about the quality and appropriateness of a regression model for the data at hand. For the linear model, for instance, in the display of the residuals vs. fitted-values, the points should fluctuate arbitrarily around the horizontal 0-line without exhibiting any pattern, as that structure is coherent with homocesdasticity. For instance, the fitted vs. residual plot of two separate linear regressions is shown in Figure 3.1. In the first plot, as would be expected in a homoscedastic linear model, the residuals and fitted values are uncorrelated and the points are randomly dispersed around the horizontal line at $y = 0$. Even though the conditional mean of the residuals is still close to 0, the second plot displays heteroskedasticity (heterogenous variance of errors), as the spread of the residuals grows along the x-axis.



Homoscedastic linear model. Heteroskedastic linear model.

FIGURE 3.1: Example of Residuals vs. fitted plots for linear model.

The interpretation of the residual plots, however, is one of the challenges when using generalised linear (mixed) models. Due to the discrete nonnegative nature of count data, for example, the residual plots from Poisson regressions are much more intricate than those from linear models. Quite often, the interpretation of the residuals plot of GLMMs is almost impossible and thus, unreliable.

A good example is provided by the developer of the `DHARMa` package, comparing the interpretations from residual plots corresponding to two Poisson Mixed Model, one lacking a quadratic effect and one that fits the data perfectly (respectively top and bottom plots in Figure 3.2). Looking closely, it is possible to identify a slight overdispersion on the range of the top plots' residuals. However, there is no way to distinguish between the need to accommodate for an overdispersion correction, or to add a quadratic effect.

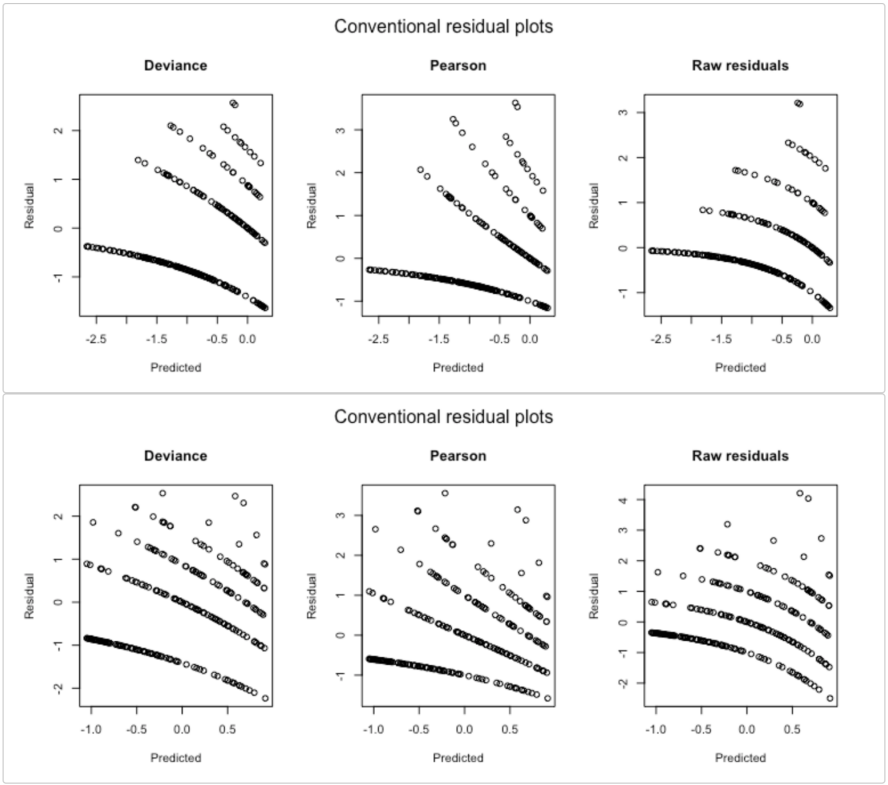


FIGURE 3.2: Comparison of the residuals between a poor fitting Poisson Mixed Model (top) and a good fitting Poisson Mixed Model (bot).

The DHARMa package was developed with the intention to solve the readability issues with residuals plots for generalised linear (mixed) models, as well as producing diagnostic tools for verifying model assumptions and validation. The package uses a simulation-based approach to create a readily interpretable scaled (quantile) residuals for

fitted (generalised) linear mixed models, which can be interpreted as intuitively as residuals from a linear regression. Equivalently, the package computes the order of the empirical quantile of the observation within the simulated data. This will be defined as the residual for that precise observation. It will clearly be within the range 0 to 1 (Figure 3.3).

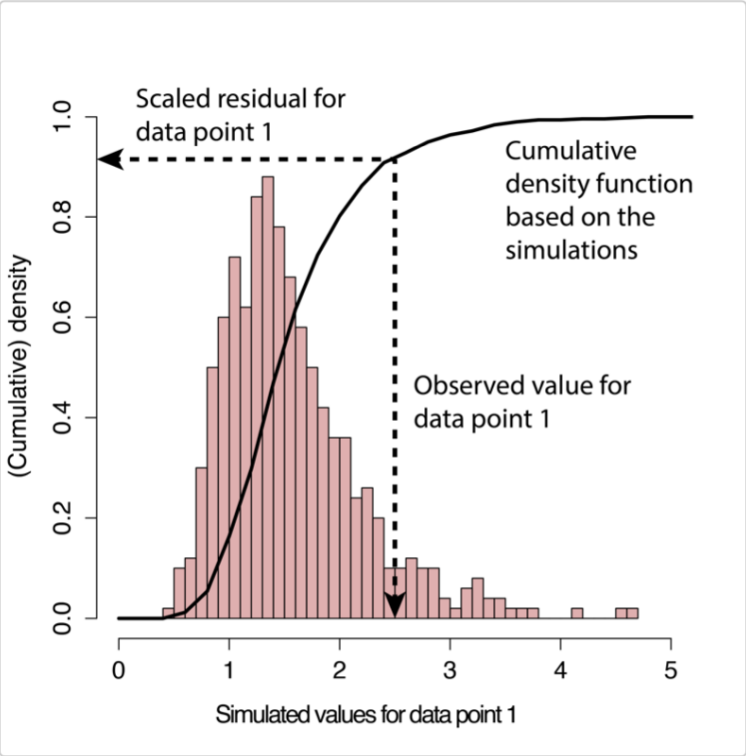


FIGURE 3.3: Visual representation of the residuals standardization steps. Adapted from the `DHARMa` vignette.

In order to evaluate the accuracy of a fitted statistical model, a specific process is employed. First, a new simulated response data is generated from the fitted model for each observation. Then, for each individual observation, the empirical cumulative density function is computed, using the simulated data. Notably, a residual of 0 for a given response observation indicates that all simulated values surpass the observed value, whereas a residual of 0.5 suggests that half of the simulated values are larger than the observed value.

This simulation has the particular benefit of ensuring that these residuals consistently follow a uniform distribution, irrespective of the type of distribution used for the fitted model (whether it is Poisson, Negative Binomial, Binomial, or even if it involves random effects (mixed model)), provided that the model is correctly specified. This is grounded in the principle that if the observed data originates from the same data-generating process as the simulated data, all cumulative distribution values should have an equal probability

occurring. Consequently, the distribution of residuals should remain flat, regardless of the underlying model structure.

Additionally, `DHARMa` also provides several diagnostic functions and tools that can be calculated directly on the fitted model object. A number of these functions that are particularly relevant for model validation are listed and described below:

- `simulateResiduals()`: This function generates simulated residuals from fitted models to assess the goodness of fit. These simulated residuals can be used to compare them with the residuals from the actual fitted model.

- plot: `DHARMa` provides several plotting functions to visualise the goodness of fit, including `plot()`, `plotResiduals()`, `plotQQunif()`, `plotSimulatedResiduals()`, `plotConventionalResiduals()`, and `plot.DHARMa()`. These plots help to evaluate the assumptions of your model.

- `testDispersion()`: This function tests the dispersion of the residuals, determining if the variance structure of the model is appropriate. As default, this function applies the non-parametric test developed in `AER` package. This test compares the variance of the simulated residuals to the observed residuals. Alternatively, `DHARMa` can implement a Perason's chi-squared test (χ^2) if specified so.

- `testZeroInflation()`: This function compares the observed number of zeros with the anticipated number of zeros from simulations to see whether the fitted generalised linear model can cope with the quantity of zeros on the dataset using the Kolmogorov-Smirnov test.

- `testOutliers()`: It conducts an outlier test to identify influential observations in the data. It offers two options (binomial and bootstrap) to test for the outliers. The binomial test assumes that the model is accurate and do not reject the null hypothesis when the probability of an observation being greater than all simulations is $1/(nSim + 1)$, following a Binomial distribution. This test, however, is more suitable for continuous distributions. The bootstrap method implements an alternative method for integer-valued distributions.

- `testQuantiles()`: Using a quantile test (a non-parametric t-test), the function fits quantile regression on the residuals and compares the quantile location to the expected location (of the uniform distribution). This function returns a p-value, for the quantile in the plot, adjusted for multiple comparisons using Bejamini and Hochberg.
- `testTemporalAutocorrelation()`: If the data is a time series or has temporal structure, this function runs a Durbin-Watson test on the uniformly scaled residuals, to checks for temporal autocorrelation, and plots the residuals against time.
- `createData()`: This function generates example data for simulating residuals.
- `getResiduals()`: Retrieves the residuals from the model for further analysis.
- `getSimulations()`: Retrieves the generated simulations residuals from a model in a standardized way
- `getFixedEffects()`: Extract and returns the fixed effects of a supported model.

Although `DHARMa` is intended to provide support for a wide variety of `R` packages that are compatible with Generalised Linear Mixed Models (GLMMs), namely `MASS`, `lme4`, `mgcv`, `gamm4`, `glmmTMB`, `spaMM`, `GLMMadaptive`, `phyr`, and `brms`, it is important to note that not all models within these packages are fully compatible with `DHARMa`. In response to that, `DHARMa` offers the `checkModel()` function, which serves the purpose of verifying whether the fitted model is indeed supported.

However, in situations where the applied models, derived from the data, and/or packages are not compatible with `DHARMa`, it is still feasible to conduct a thorough residual analysis using `DHARMa`. To accomplish this, the user must first generate a new set of simulated response data and subsequently create a DHARMa object by utilizing the `createDHARMa` function. This approach ensures that even when the `R` package being applied is not support, `DHARMa` can still facilitate the essential analysis of residuals.

Finally, it is essential to grasp and interpret the plots generated by `DHARMa`. Figure 3.4 illustrates the discernible patterns associated with models suffering from overdispersion issues. As depicted, the quantile-quantile plot for the uniform distribution reveals a departure from uniformity (low expected values fall below the reference line, whereas high expected values are above it) and an excess of residual values at the extremes (around 0 and 1) in the residual vs. expected plot.

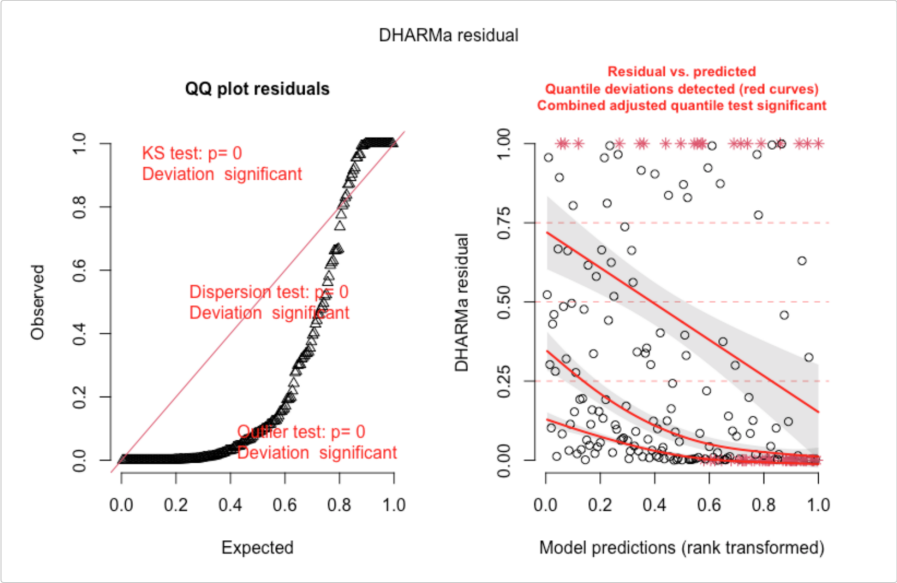


FIGURE 3.4: DHARMa plots for overdispersed data. Plots adapted from DHARMa package vignette

Conversely, Figure 3.5 presents the anticipated patterns observed in underdispersed models. The underdispersion problem shows up as a deviation from uniformity (opposite pattern as overdispersion) in the quantile-quantile plot and an excessive concentration of residual values around 0.5 in the residual vs. expected plot.

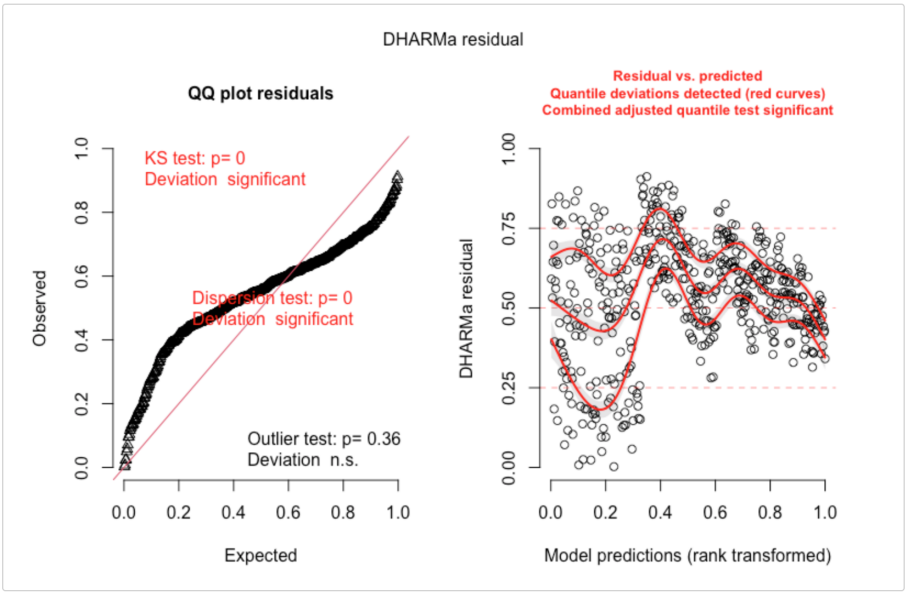


FIGURE 3.5: DHARMa plots for underdispersed data. Plots adapted from DHARMa package vignette.

Nonetheless, it should be highlighted that the absence of a discernible residual pattern does not conclusively establish that the model is correctly specified. Instead, it should be

regarded as a working hypothesis, recognizing that DHARMa may not identify all structural problems within the model. On the other hand, the presence of a significant residual pattern does not necessarily render the model unfit for use. Furthermore, it is important to assess the magnitude of the residual pattern when the uniform distribution is statistically rejected. In this context, significance reflects the signal-to-noise ratio rather than the strength of the pattern itself. For models with substantial sample sizes, it is not uncommon for residual diagnostics to demonstrate significance, even in the absence of severe issues. Therefore, it is imperative for users to exercise their judgment in determining whether deviations from the uniform distribution hold relevance for their analysis. Ultimately, DHARMa's purpose is to highlight disparities between observed and expected data, leaving the responsibility of discerning whether these disparities pose issues for the analysis in the hands of the user.

Chapter 4

Application to Ecological Parasitology Data

The data under study in this project was obtained as part of the COACH project – *Cooperative approach applied to conservation and management of cockles* (http://coach.web.ua.pt). The COACH project was funded by *Fundação Oceano Azul* and *Oceanário de Lisboa* through the *FUNDO para a Conservação dos Oceanos* and developed by researchers from CESAM – Centre for Environmental and Marine studies – at the University of Aveiro.

4.1 COACH project

The edible cockle *Cerastoderma edule* (Bivalvia: Cardiidae) is an indigenous, infaunal, suspension-feeder bivalve living in semi-sheltered marine systems along the north-eastern coast of the Atlantic Ocean [23, 25]. Due to their critical function in ecosystems, cockles are an important ecological species [18, 20]. Moreover, cockles are the primary source of income for many fishermen, particularly in Portugal [22]. In Aveiro, a Portuguese fishing village bathed by the Ria de Aveiro coastal lagoon, cockles represent one of the most important marine resources. The capture of this bivalve in the Ria de Aveiro can exceed 1000 tons per year. However, the global changes and the high harvesting pressure are compromising the conservation of cockles' population of Ria de Aveiro, contributing to the observed declining of fishing stocks in recent years. In this sense, the COACH project aims to promote the conservation of this important natural resource along with the ecosystem services it provides while assuring its sustainable exploitation and the economic and social development of the local community. This was achieved by gathering

multifactorial information on the biology, habitat, and fishing of cockles in the Ria de Aveiro, namely studying the effects and identifying the main abiotic factors influencing the abundance and prevalence of parasites in cockles. For this purpose, 18 cockle's beds in the Ria de Aveiro were selected, covering the entire distributional range of this species in this coastal system (Figure 4.1).

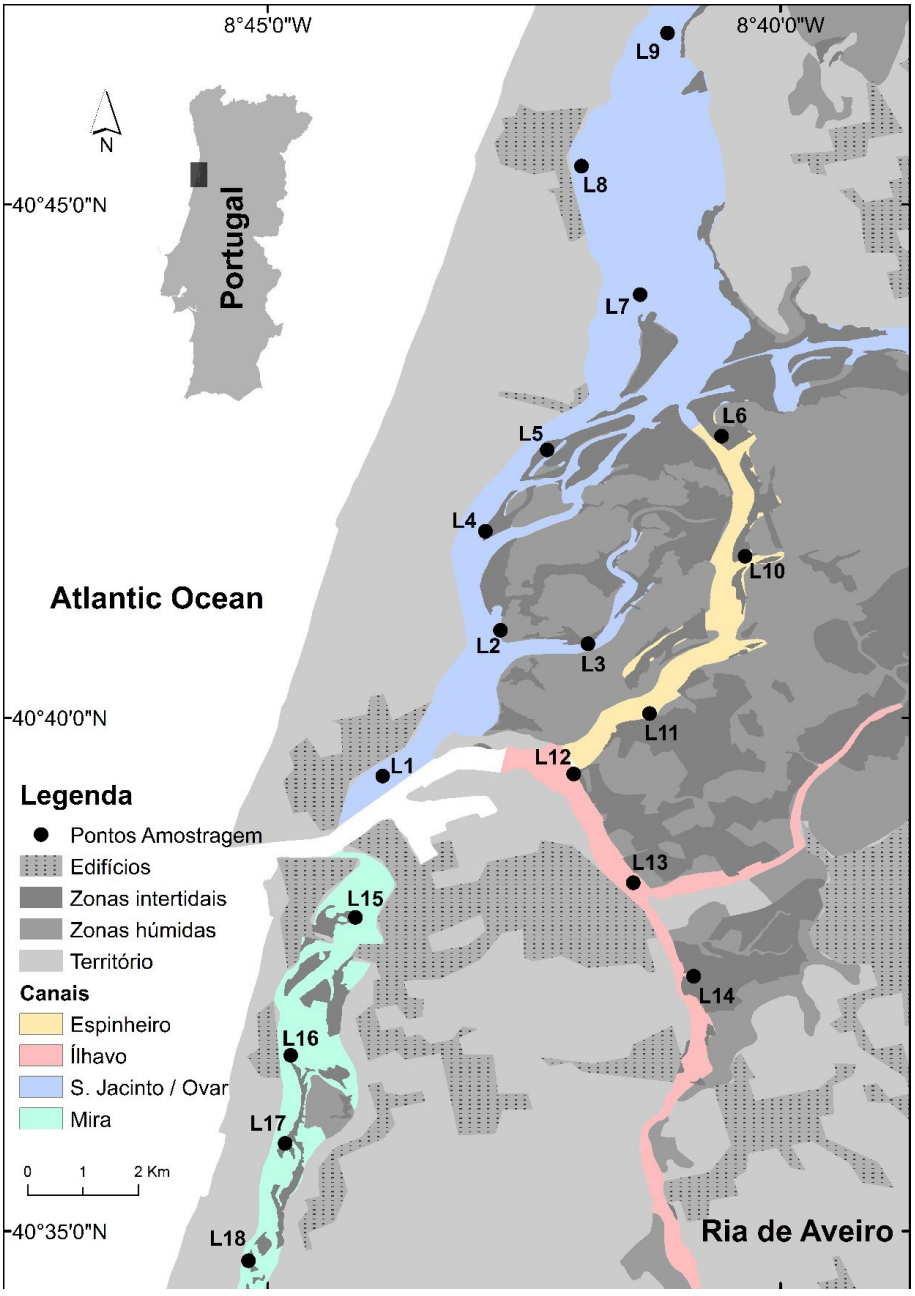


FIGURE 4.1: Study area. Geographical location of the 18 sampling sites along the Ria de Aveiro coastal lagoon (Portugal). Figure created using ArcGIS software.

The Ria de Aveiro (40° 38'N, 8° 44'W) is a large coastal lagoon located on the North-west coast of mainland Portugal. This highly intricate coastal system comprises an area

of approximately 83 km^2, 45 km length and a width of 8.5 km [71]. It is divided into four main channels (S. Jacinto / Ovar, Espinheiro, Ílhavo and Mira) that radiate from the ocean mouth into several branches and ecosystems (mudflats, saltmarshes, freshwater marshes, and alluvial forests). The Ria de Aveiro is connected to numerous rivers (Antuã, Boco, Cáster, and Vouga), resulting in a strong horizontal gradient of salinity and water temperature across the lagoon, experiencing semi-diurnal tidal cycles with amplitudes ranging from 0.6 m to 3.2 m [72, 73]. Thus, Ria de Aveiro is one of the most significant biodiversity hotspots of south-western Europe, being a Special Protected Area, protected by the EU Birds Directive (79/409/CEE), and is a component of the Natura 2000 network (EU Habitats Directive).

4.2 Sampling and Data collection

Sampling occurred monthly between June 2020 and May 2021 at 18 locations in the Ria de Aveiro (Figure 4.1). At each location in each month, cockles were harvested at low tide, from the intertidal zone, by collecting the top 5 cm layer of sediment of six quadrats of 0.25 m^2 and sieving through a mesh sieve with 1 mm openings. The density *(d)* of cockles per square meter was then calculated.

$$d = \frac{\text{total number of cockles}}{6 \text{ quadrats} \times 0.25\, m^2} \tag{4.1}$$

Each cockle's shell length (SL) was measured with a calliper to the lowest mm. Whenever possible, fifteen cockles representing the SL of each sampling site were dissected. Dissected cockles were squeezed between two glass slides and observed under a stereomicroscope. Following the available identification keys [35, 74, 75], all macroparasites were identified up to the species level. Parasite species identified in each of the observed cockles were counted to assess parasite abundance (number of parasites per cockle) and prevalence (percentage of infected cockles) [76]. For some parasitic species, namely trematodes parasites using cockles as first intermediate host (sporocysts), only prevalence was registered.

At the same time of cockle sampling, abiotic data was collected to characterize the habitat. Two sediment samples from the sediment surface were taken from each sampling site to perform grain-size analysis (MGS) and to determine total organic matter (TOM) content. Sediment grain-size analysis (MGS) was conducted by wet sieving the silt and

clay fraction (fin particles, diameter < 0.063 mm) and dry sieving the remaining sediment fractions (sand and gravel). The median grain size was defined in ϕ. This value was obtained through the mean value (P50) of the cumulative frequency of each fraction ($\phi = -\log_2(P50)$) [77]. Regarding TOM analysis, sediment was dried at 60 °C for 48 hours before being pulverized to powder with a mortar and pestle. The difference between the dried samples (about 1 g) and the combusted samples was used to calculate TOM content [78]. Water temperature (°C), pH, salinity, dissolved oxygen (%, DO) and redox potential (mV, Eh) were measured in the nearest water column using a handheld multiparameter probe.

4.3 Dataset

The dataset obtained contained 18 cockle's beds, 12 sampling months, 13 species of macroparasites (3 species of trematode sporocysts (*Bucephalus minimus*, *Monorchis parvus* and *Gymnophallus choledochus*), 7 species of trematode metacercariae ((*Himasthla elongata*, *H. interrupta*, *H. quissetensis*, *Himasthla* sp., *Renicola roscovitus*, *Diphterostomum brusinae*, *Gymnophallus minutus*), two copepods ((*Mytilicola orientalis* and *Hermannella rostrata*) and one rhabdocoela (*Paravortex cardii*)), cockle's shell length and condition index, and 8 abiotic variables (salinity, dissolved oxygen, reduction-oxidation potential, pH and temperature of the water and organic matter content and median grain size of the sediment) measured for each of the cockle's beds.

The dataset contained 2849 rows, each representing a cockle analysed in a particular site and month, and 391 missing values. The missing values were caused by low cockle density at the time of sampling, which prevented from collecting 15 cockles for analysis, or due to sampling impossibilities, specifically at site 7 in the months of October and December of 2020, and February and April of 2021 (Table 4.1).

For the analysis, parasites were divided into four groups based on the stages of their life cycle (for trematode parasites) and the types of behaviours they exhibited. In this manner, *Gymnophallus minutus* was separately grouped from remaining trematode parasites at metacercariae stage. This separation is related to the fact that *Gymnophallus minutus* may display some aggregation behaviour patterns in highly contrasted environments, as long as first intermediate host (*Scrobicularia plana*) is present in the ecosystem [79]. Species of trematode sporocysts represented another group, and the remaining parasites were

TABLE 4.1: Number of analysed cockles per month (columns) and sampling site (rows)

	1	2	3	4	5	6	7	8	9	10	11	12
1	15	15	15	15	15	14	15	15	15	6	15	15
2	15	15	15	15	15	15	15	15	15	15	15	15
3	15	15	15	15	15	15	15	15	15	15	15	15
4	15	15	15	15	15	15	15	15	15	15	15	15
5	15	15	15	15	15	15	15	15	15	15	15	15
6	15	15	15	15	15	15	15	15	15	15	15	15
7	15	15	15	15	0	15	0	15	0	15	0	15
8	15	15	15	15	15	15	15	15	15	15	15	15
9	15	1	4	1	1	1	0	0	0	0	0	15
10	15	15	15	15	15	15	15	15	15	15	15	15
11	15	15	15	15	15	15	15	15	15	15	15	15
12	15	15	15	15	15	15	15	15	15	15	15	15
13	15	15	15	14	15	15	15	15	15	15	15	15
14	15	15	15	15	15	15	15	15	15	15	15	15
15	15	15	15	15	15	15	15	15	15	15	15	15
16	15	15	15	15	15	15	15	15	15	15	15	15
17	15	15	15	15	15	15	15	15	3	2	15	15
18	10	0	1	5	0	3	1	0	0	0	1	6

grouped as "Other". Thus, the final database consisted of 4 dependent variables and 9 explanatory variables (Table 4.2).

TABLE 4.2: Desciption of dataset variables

Variable	**Description**	**Type of variable**
Month	Sampling Month	Explanatory
Site	Cockle bed sampling site	Explanatory
Metacercariae	Number of trematode parasites individuals at metacercaria life stage per cockle	Dependent
Gymnophallus	Number of emph{Gymnophallus minutus} individuals per cockle	Dependent
Sporocysts	Number of trematodes sporocyst species infecting cockles	Dependent
Other	Number of copepod and rhabdocoela species individuals per cockle	Dependent
SL	Shell length of the cockle (mm)	Explanatory
Sal	Salinity of the water at the sampling time	Explanatory
DO	Dissolved oxygen in the water column at the time of sample $(mg.l^{-1})$	Explanatory
Eh	Water column redox potential at the time of sample	Explanatory
pH	Water column pH at the time of sample	Explanatory
Temp	Water column temperature at the time of sample $(°C)$	Explanatory
TOM	Sediment organic matter content (ϕ)	Explanatory

In this project report, analyses will only be performed for the dependent variable *Metacercariae* since this study is strictly academic and aims to learn the intricacies of the modelling of count data with excess of zeros. All the following analyses were carried out using R, version 4.3.1.

4.4 Descriptive Analysis

As previously described, this project aimed to analyse the effect of abiotic variables on the abundance of parasites, in this case trematode parasites infecting cockles as second intermediate host (metacercariae stage). Trematodes are common macroparasites in coastal systems [80]. Their life cycle typically involves complex alternations between parasitic and free-living stages, and multiple hosts. Free-swimming larvae, miracidia, hatches from the egg and actively search for the first intermediate host, often molluscs. Inside the first intermediate host, miracidia transform into sporocyst (or redia) and undergo asexual multiplication to form cercariae, free-living stage which are released into the water column. Cercariae will look for the second intermediate host, that can vary from invertebrates (e.g., molluscs) to vertebrates (e.g., fishes), and penetrates their tissue, transforming into metacercariae. When the final host, a vertebrate, predates the second intermediate host, metacercariae will mature into adult stage. In the final stage, adult trematodes sexually lay eggs that are excreted through the definitive host's faeces, restarting the cycle [81].

Understanding the behaviour of the data is important before fitting a model. Therefore, the number of outliers, the relationship between variables and dependency, and the quantity of zeros were all investigated as part of the data exploratory process. Figure 4.2 shows the Cleveland dot plot for all the relevant variables.

We begin the descriptive analysis using **Metacercariae**, the dependent variable. The counts observed for the number of metacercariae infecting a cockle varied between 0 and 12. In total, 2472 non-infected cockles (with zero metacercariae) were observed, which corresponded to approximately 86.7 % of the observations. One could think of starting immediately by fitting a zero-inflated or hurdle model. However, we shall start from the usual Poisson model, investigate the validity of its assumptions, and improve the model whenever there are conditions that fail. This procedure of starting from the Poisson model has been described in the literature [82]. Of the remaining cockles, most of the observations showed a single metacercariae (238 cockles, 86.7 %), 80 (8.4 %) were infected with 2 metacercariae and 31 (2.8 %) cockles displayed 3 metacercariae. Observations above 3 metacercariae per cockle represented each less than 1 % of total observations. None of these larger counts looks too large or too extreme for the usual values found in the literature, however observations exceeding 3 metacercariae in this study are uncommon and are thus potential outliers if issues arise. The bar plot with the number of counts and respective percentages per number of metacercariae are represented in Figure 4.3.

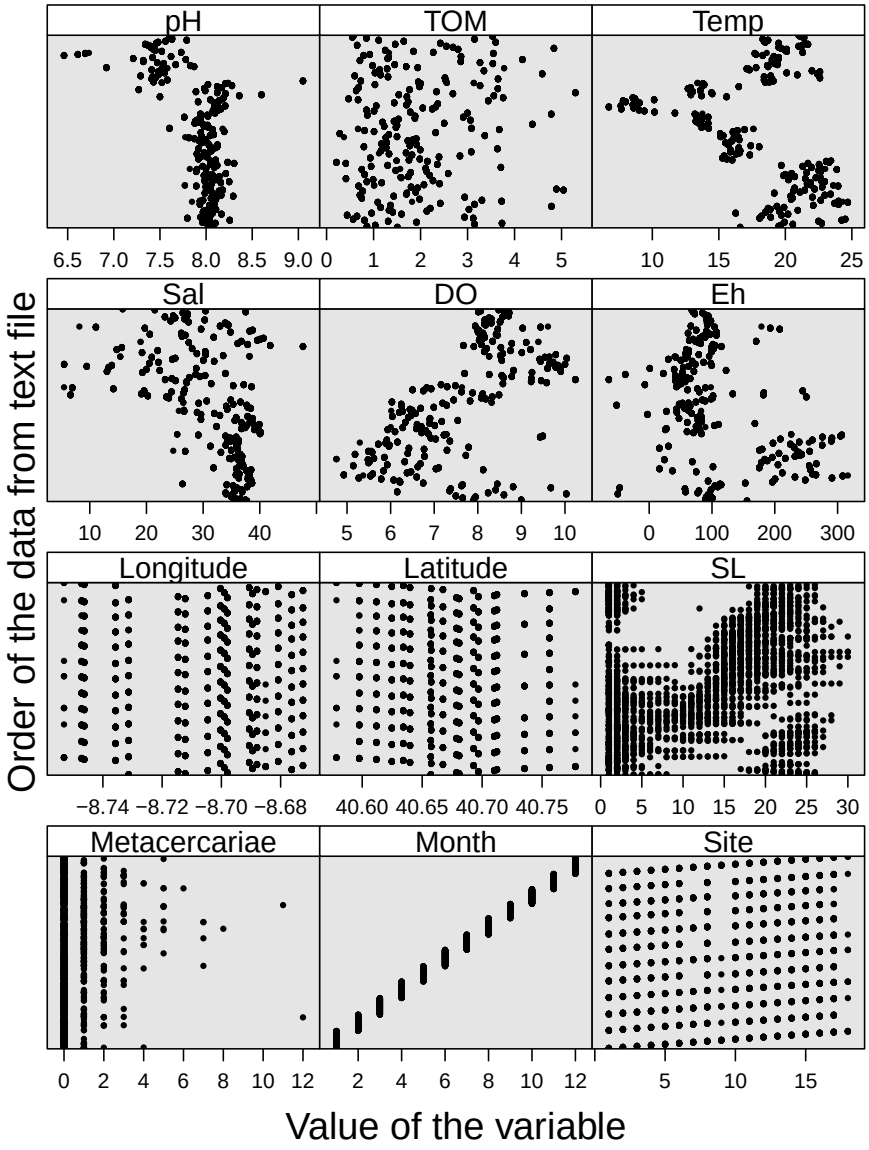


FIGURE 4.2: Cleveland dotplot for the variables used variables during the analysis. The horizontal axes represented the value of the variable, while the vertical axes shows the order of the observation in the dataset. Figure created with the help of the `ggplot2` package from `R` software.

Spatially, with a total of 55 infected cockles over the course of a year, sampling site 6 was the one that contributed the most for the counts (higher prevalence). However, the largest abundance was found at sampling site 15, with 123 metacercariae detected in total. Additionally, site 15 also had the highly infected cockle with 12 metacercariae discovered infecting a single cockle. The lowest prevalence and abundance of metacercariae were found at sampling sites 9, 17, and 18, where no metacercariae were found infecting cockles (Figure 4.4).

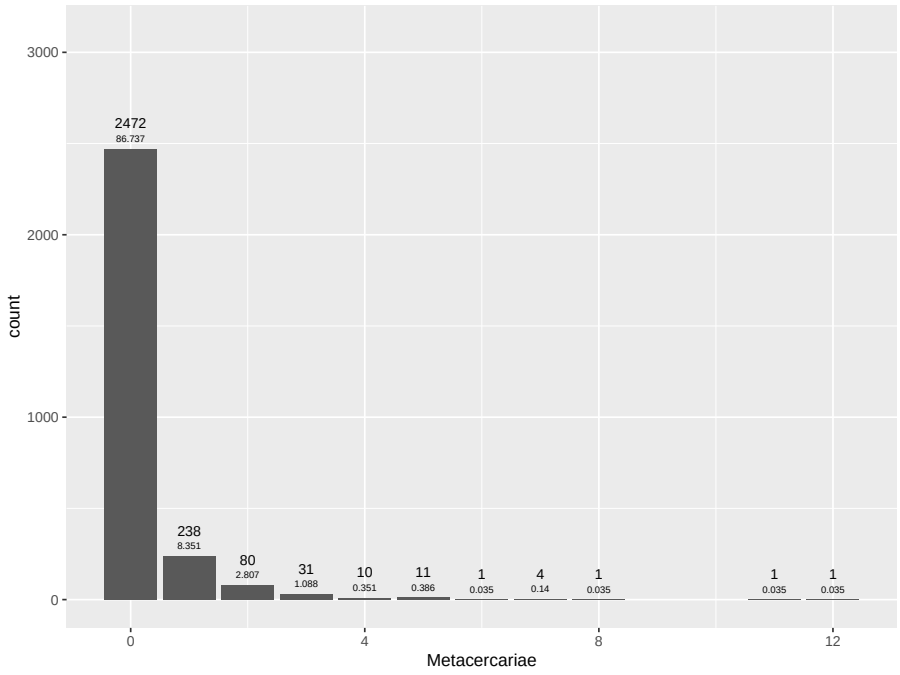


FIGURE 4.3: Bar plot for the number of observed metacercariae counts with total number of counts and respective percentage. Figure created with the help of the `ggplot2` package from `R` software.

The month with the highest prevalence and abundance of metacercariae was month 9 (February 2021) with 60 infected cockles counting for a total of 113 metacercariae. Nonetheless, the single highest infection was observed in month 2 (July 2020). Months 3 (August 2020) and 4 (September 2020) were the months with the lowest prevalence of parasites with only 13 and 15 infected cockles, respectively. Figure 4.5 displays the number of metacercariae for each month.

In contrast, and predicting a possible future need for a dichotomization of the counting variable, Figure 4.6 displays a barplot showing the absence (0) and prevalence (1) of cockles with metacercariae infections in each of the months.

The combination of metacercariae counts per sampling site and month is shown in Figure 4.7. In this image, the x-axis denotes the sampling month, while the y-axis the quantity of metacercariae per cockle, and each plot represents a sampling site. It should be highlighted that certain cockles under analysis had repeated counts (namely zeros), hence we do not always observe 15 points (counts) per month.

Figure 4.8 depicts the time evolution of the mean of parasites across time (total number of parasites counted from a site and month divided by the number of cockles analysed), separated by site. The range of metacercariae varied from 0 to 3 $metacercariae.cockle^{-1}$. The greatest average number of metacercariae was displayed in site 14 at month 8. An

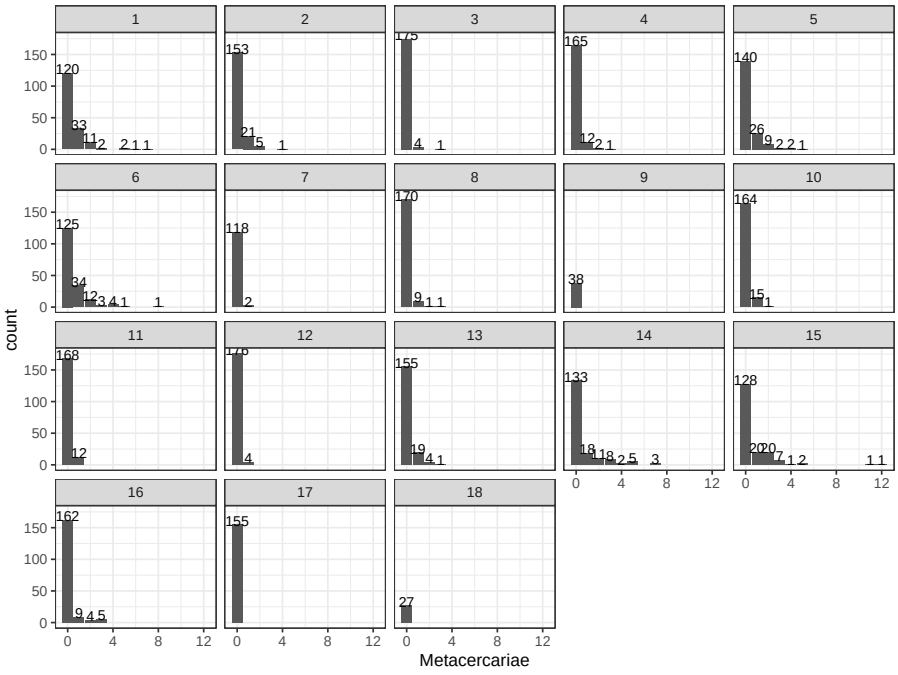


FIGURE 4.4: Bar plot for the number of observed metacercariae counts per site. Figure created with the help of the `ggplot2` package from `R` software.

average trend of parasite distribution across all locations is shown in this figure, as a superimposed smooth line, and it shows a peak at month 8.

Still related with the metacercariae abundance per month and sampling site, Figure 4.9 displays a 3D visualisation of an adjusted surface with the sample sites in the x-axes, the month in the y-axes, and the counts on the z-axes. As it is possible to observe, the surface lift is very low, reaching maximum values of 0.8. It should be noted however, for all these exploratory graphics, the method of adjusting a line or surface to the counts makes the incorrect assumption that the dots are independent of one another.

We now move to the description of the associations between the explanatory variables and the response.

Regarding the abiotic and biotic variables, none of the 7 variables (pH, organic matter content, temperature, salinity, dissolved oxygen, redox potential, or shell length) appear to have any evident outlier. These variables are dependent on the season and time when the data collection was conducted, hence it is possible to observe some variation. For instance, it is noticeable that temperature has an interval where the readings are lower, which must be correlated to the winter months where temperatures are naturally lower. A similar case is presented for salinity. In this case, this variable is not only dependent on the season of the year, but also on tidal stage and their currents (flood or ebb current),

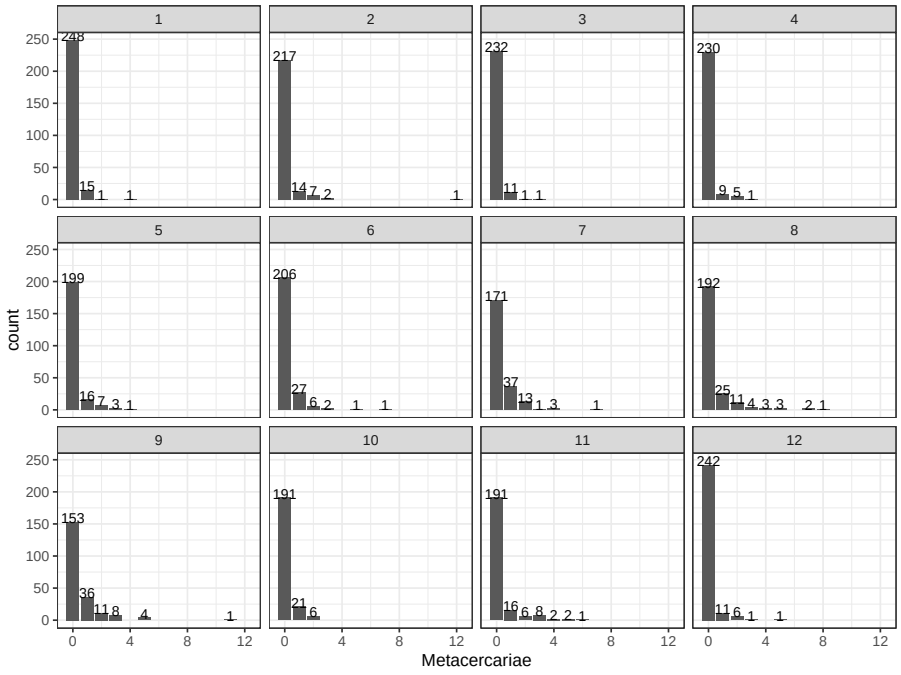


FIGURE 4.5: Bar chart for number of observed metacercariae counts per month. Figure created with the help of the `ggplot2` package from `R` software.

and the site localization on the coastal system. In this instance, the sample strategy was to cover the Ria de Aveiro's whole cockle distribution range, from downstream to upstream. Thus, it is anticipated that places upstream (close to the river) will be more influenced by fresh water and thus see a drop in salinity. Only the pH variable deviated from initial expectations (values were expected to be around pH 8.0). However, this may be a result of a modification to the used reading equipment that took place during the most recent sampling months. Therefore, it was decided to not exclude any of the values for the analysis.

Figure 4.10 shows a pairwise scatterplot and Spearman correlation matrix for all the relevant variables of the study. The variables names are shown on the top and left side of the Figure, and their density plots are represented along the diagonal of the matrix, with the scatterplots for each variable-variable relationship shown below the diagonal, and linear correlation coefficients corresponding to those scatterplots above the diagonal (the density plots should not be taken into account for categorical variables (e.g., Site)).

The top row and the left-most column show the relationship between the dependent variable, **Metacercariae**, and the covariates. Some weak patterns are observed, none with particular relevance to draw attention to. All the other panels were used for detecting collinearity. There were some moderate correlations observed between variables, around

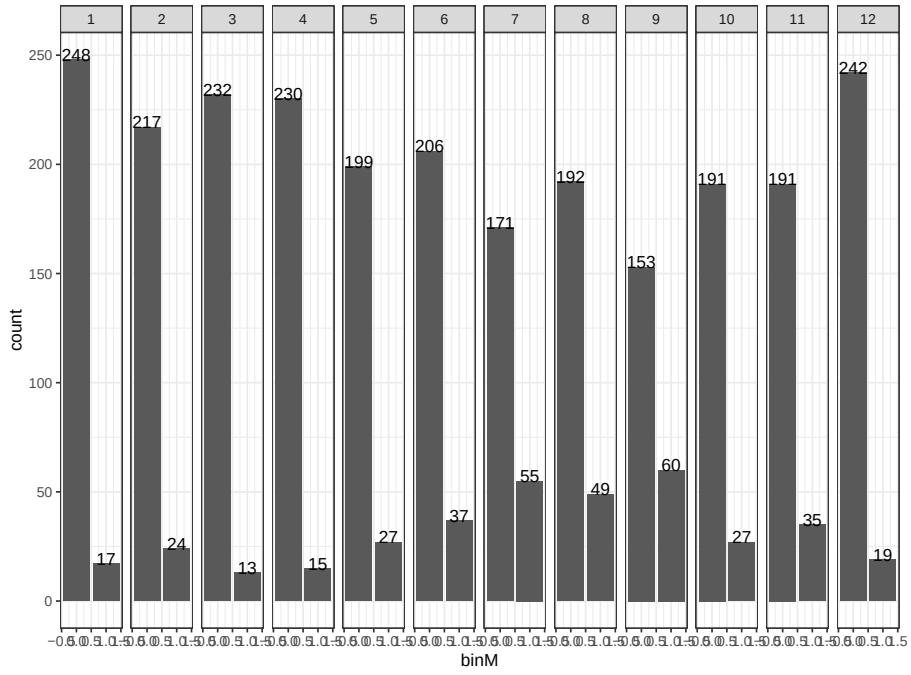


FIGURE 4.6: Bar plot for the presence and absence of metacercariae counts per site. Figure created with the help of the `ggplot2` package from `R` software.

0.5. The correlations observed with Longitude and Latitude should not be taken into consideration as these variables will not be used in the model (we only have 18 sites, which is not enough to carry out spatial models, along with the longitudinal structure). Most of the moderate correlations involved the variable month. This is expected because, as already mentioned, these abiotic variables are season dependent (for example, higher temperatures in summer compared to winter). Salinity and Dissolved Oxygen in the water column also showed to be a moderately collinear with a significant negative correlation ($\rho = -0.561$). It is well known that the solubility of oxygen is dependent on salinity, as the amount of oxygen dissolved in water decreases as salinity levels rise. Thus, having covariates that derived from the same source can set off an alarm. Nevertheless, since correlations were only moderate, there were no covariates removed from the dataset prior to modelling.

Regarding the relationship between the dependent variable and the covariates, the abiotic variables observations used for this study were punctual collections that occurred at the moment of sampling (see section 4.2; page 33). As previously mentioned, trematodes have a complex life cycle. Metacercariae, more specifically, are the outcome of cercariae infection, which are trematode free-living larval stages that emerge from the first intermediate host under optimum circumstances, namely at ideal temperatures. Cercariae

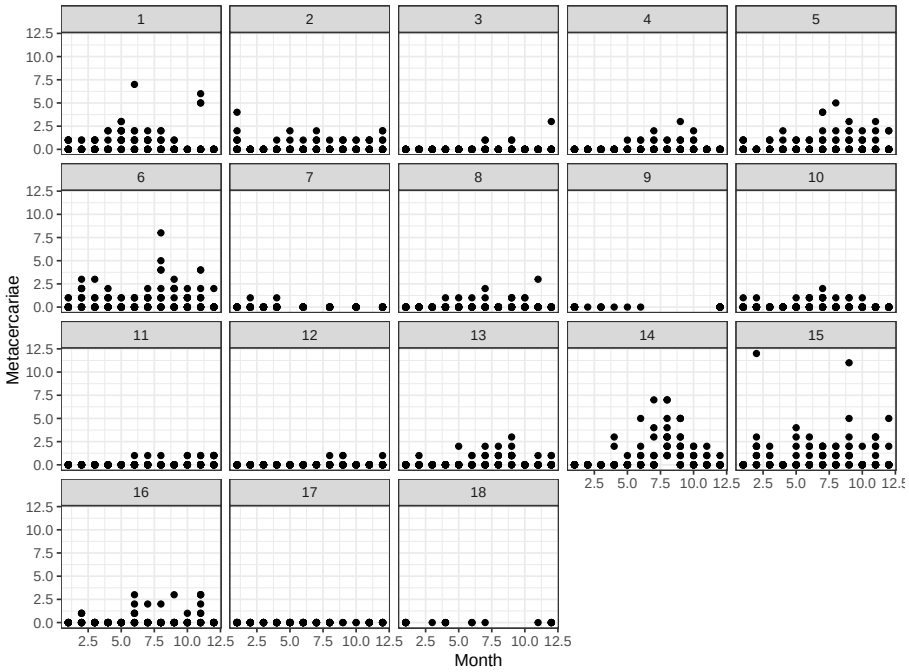


FIGURE 4.7: Dotplot with metacercariae counts per month and sampling site. Figure created with the help of the `ggplot2` package from `R` software.

exhibit a lifespan of approximately 48-hour during which they must find and infect the second intermediate host, here cockles, in order to survive. Consequently, given that data collection includes 1-month intervals, the metacercariae counts infecting cockles seen in a given month may be a result of the abiotic circumstances observed in the previous month.

Table 4.3 presents the correlation between the response variable, Metacercariae, and the abiotic variables with 0 (no lag) or 1 (a month lag) lags. The highest Spearman correlation absolute values for each of these variables are shown in bold. Only 1 lag was used for this analysis, as infections would be detectable in the following sampling.

TABLE 4.3: Correlation between the environmental variables, with and without a one-month lag, and the dependent variable Metacercariae. The highest absolute Spearman correlation coefficient, in absolute value, are shown in bold.

Lag	0	1
Salinity	-0.067	**-0.440**
DO	**0.087**	0.040
Eh	-0.095	**-0.111**
pH	**0.030**	-0.003
TOM	**0.097**	0.055
Temp	-0.167	**-0.171**

The lags with highest absolute value varied from variable to variable. That way, the analysis was maintained without the use of lags since this would result in the loss of

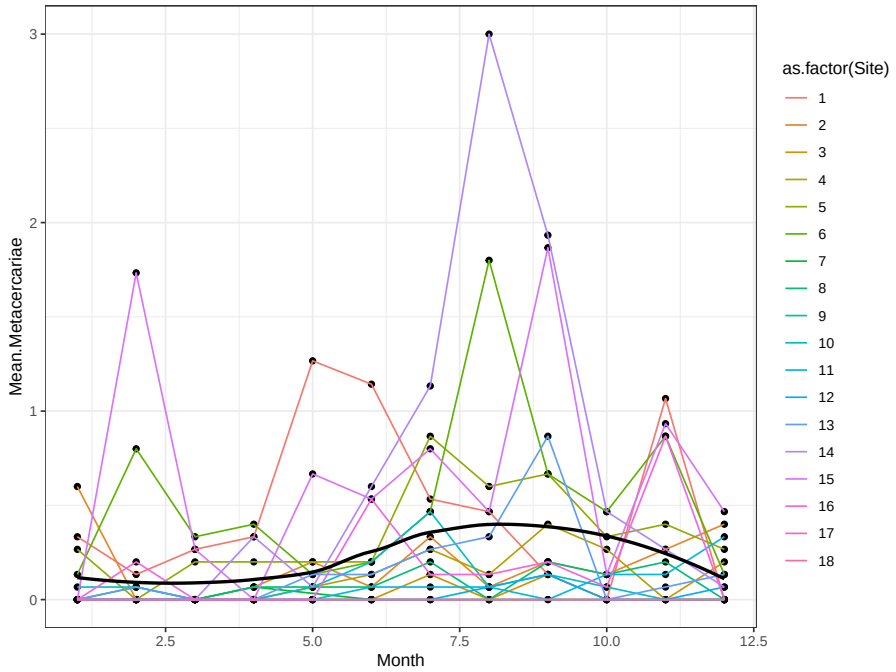


FIGURE 4.8: Evolution of the mean number of metacercariae over the 12 months of sampling per site. Figure created with the help of the `ggplot2` package from `R` software.

information. Moreover, it would not be sensible from a biological standpoint to set distinct lags for the environmental variables because they affect the parasites (and the whole community) concurrently.

Figure 4.11 illustrates how the variables under study and the dependent variable are related, similar to the previously boxplot analysis. These plots display the counts on the y-axis in relation to the variables' values on the x-axis, with a smooth to demonstrate how the relationship. There seems to be an approximately linear association between the counts and the variables dissolved oxygen, redox potential, temperature, and organic matter content. However, the variables Month, Site, Salinity and Shell Length, did not exhibit the same behaviour. This way, it will be important to exercise caution when deciding whether to model this data without the use of GAMs.

The associations between the variable month and some explanatory variables, namely salinity, dissolved oxygen, and pH, were examined using boxplots (Figure 4.12).

In the first six months of the sample, the median of salinity ranged between 35 to 40, decreasing between 20 to 25 in months 7 to 9 (December 2020 to February 2021), and then increased once again in the last three months of sampling. Salinity in coastal systems rises during the warmer months (summer) as a result of a decrease in freshwater inputs and

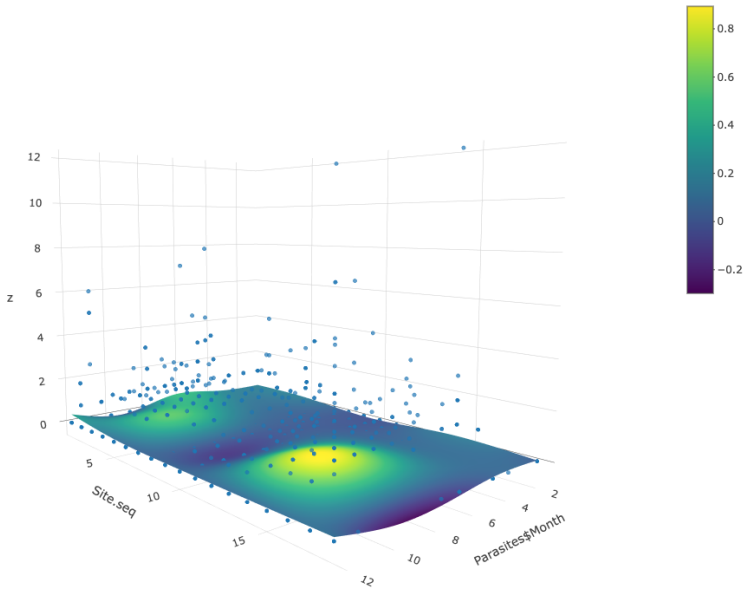


FIGURE 4.9: 3D visualisation of a surface fitted to metacercariae counts by month and site. Figure created with the help of the `plotly` package from `R` software.

an increase in evaporation caused by the higher temperatures. This is also noticed by the relationship between temperature and month on Figure 4.12 **b)**.

In the initial sampling months, the temperature median was higher and gradually dropped until it reached the lowest records in month eight. The subsequent months saw a further increase in the temperatures. Salinity, as previously stated, also has an impact on the chemistry of coastal systems, mostly by lowering the oxygen solubility and, consequently, the concentrations in the water. Thus, Figure 4.12 **c)** shows the opposite behaviour between dissolved oxygen and month compared to the previously mentioned relation between salinity and month, with an increase in dissolved oxygen observed in the last sample months and a peak in month 9.

The redox potential exhibited stable median values throughout the year of sampling, except for months 2 to 4 where the values were considerably higher (Figure 4.12 **d)**). However, in all months, the values remained positive. The redox potential is an important water parameter as it indicated the anaerobic condition of the system. Higher redox potential values suggest aerobic conditions, while lower redox values indicate anaerobic and reduced conditions. Since redox potential and oxygen availability are connected, it was anticipated that oxygen and redox potential would follow a similar trend across months. In this study, the correlation was not found as other variables, such as organic matter

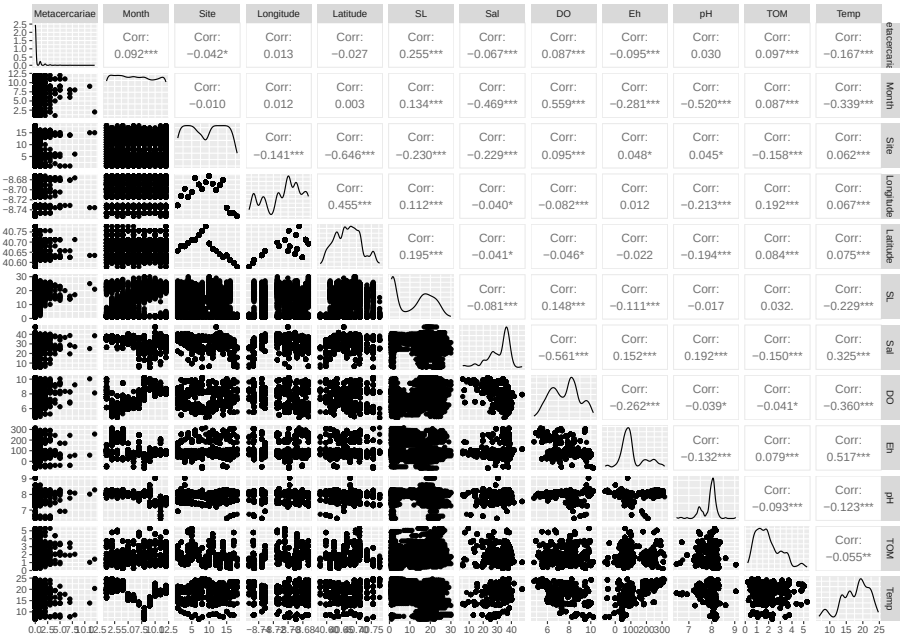


FIGURE 4.10: Density plot (diagonal), scatterplot (below the diagonal) and pairwise Spearman correlation matrix (above the diagonal) for all variables of interest to the study. Asterisks represent statistically significant correlations. Figure created with the help of the `GGally` package from `R` software.

content, can also influence the redox potential status.

The pH (Figure 4.12 **e)**) and organic matter content (Figure 4.12 **f)**), of all variables, showed the least fluctuating behaviour over the month, practically displaying no variation throughout the sampling months. The pH of coastal water systems usually hovers around 8.0 / 8.1, and the lower values found in the last three months analyzed might be due to a change in reading equipment.

On the other hand, Total Organic Matter (TOM) showed similar median values throughout all months with slight variations.

Figure 4.12 **g)**, depicts shell length across the studied months. Overall, the shell length increase up to a maximum in month 10 and then suddenly decreases. Cockles display an annual reproduction cycle, occurring in the warmer months, followed by a larval stage that last up to three months until new a recruitment settles. Thus, this development and increase in the shell length over the course of sampling might be an indicator of the development of a cohort and the appearance of a new one.

Finally, we investigated whether the infection status (infected versus non-infected) of cockles corresponded to different median values of the explanatory variables (Figure 4.13). For most of the variables, no relevant differences between infected or non-infected cockles were identified.

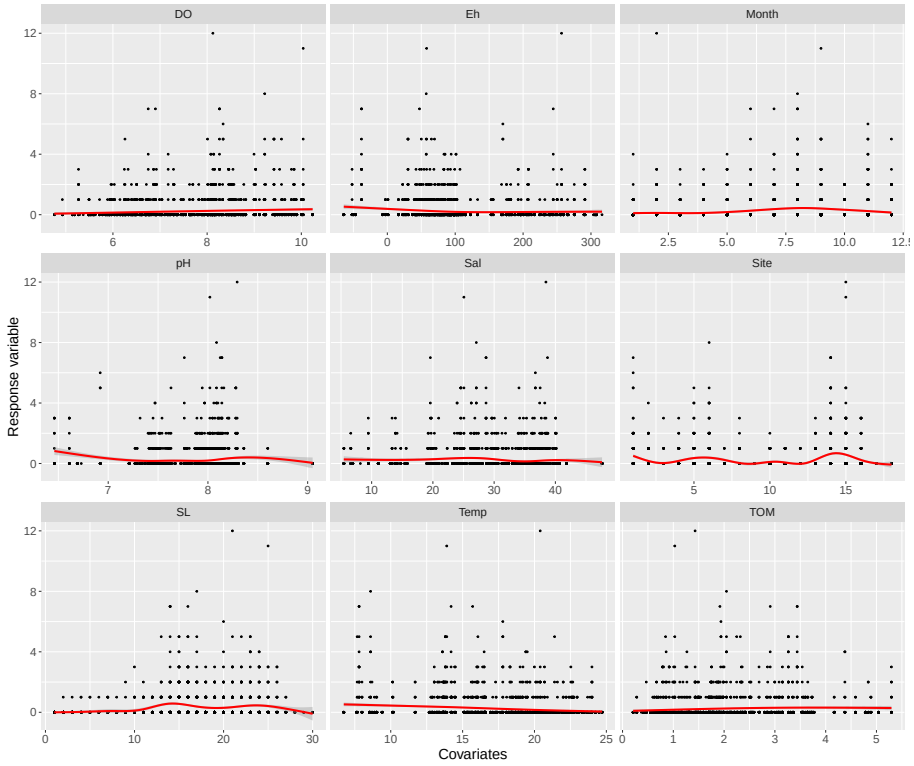


FIGURE 4.11: Number of metacercariae counts plotted against each of the variables to analyse relationship between the dependent variables (Metacercariae) and the explanatory variables. Smoother was added to aid visual interpretation of the relationship. Figure created with the help of the `ggplot2` package from `R` software.

For salinity, the medians of infected cockles were greater in several months than it was for non-infected ones, such as months 1, 3, 5, 6, and 9. In the opposite direction, months 4 and 11 showed lower medians for infected cockles, while the remaining months showed very similar values between infected and non-infected cockles (Figure 4.13 **a)**).

The Dissolved Oxygen boxplots for each infection state are shown in turn in Figure 4.13 **b)**. Again, the highest values differ from month to month, with the highest values occurring in months 2, 3, 5 and 12 for infected cockles, in comparison to non-infected cockles. It should also be noted that, with the exception of months 4 and 8, where both the median and third quartile show values higher levels of dissolved oxygen for non-infected cockles, the third quartile is almost always higher for infected cockles, even in situations where the median is similar between infected and non-infected cockles.

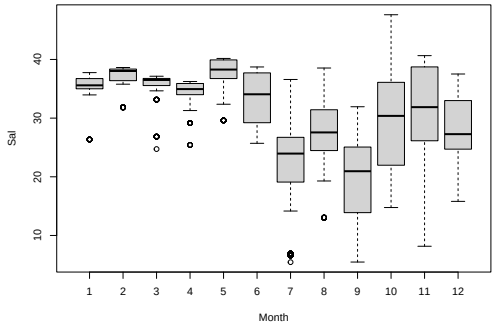
The same pattern can be seen for the redox potential, where the third quartile is frequently greater for infected cockles even if many months exhibit identical medians regardless of infection status (Figure 4.13 **c)**).

Since the pH of coastal systems should not vary greatly, the pH values for every given

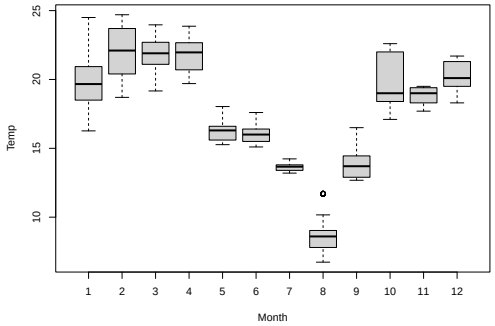
month or infection state turned out to be relatively comparable (Figure 4.13 **d)**). The same thing occurred with temperature, as can be shown in plot **e)** of Figure 4.13, where, with very few exceptions, the medians and boxplots across states of infection in each month are relatively comparable.

Figure 4.13 **f)** shows the boxplots for the organic matter content (TOM) and, once more, the behaviour was consistent with that observed for the other variables.

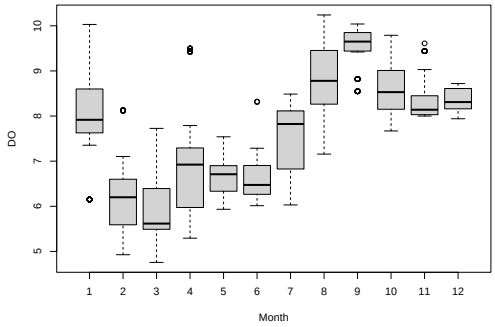
Finally, the shell length boxplot (textbfg)) was the only one to exhibit unusual behaviour, with the median and quartile values for infected cockles being significantly larger in almost every month for infected cockles. However, in months that shell length as generally higher (months 7 to 10), the median was more evenly distributed between infected and non-infected cockles.



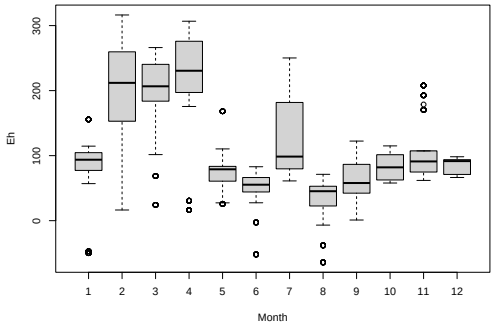
a) Salinity versus Month



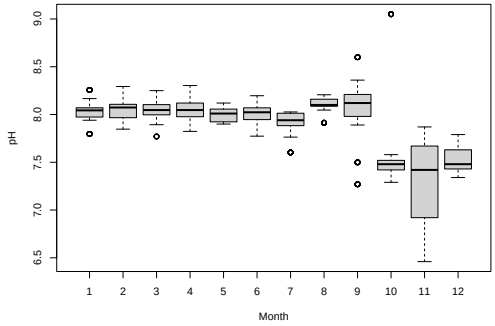
b) Temperature versus Month



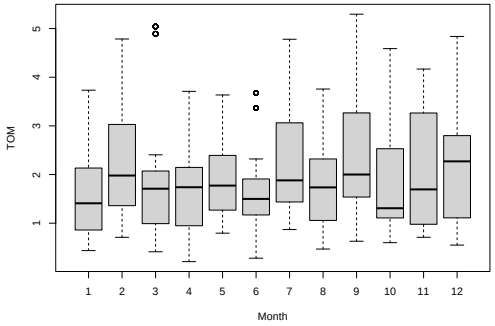
c) Dissolved Oxygen versus Month



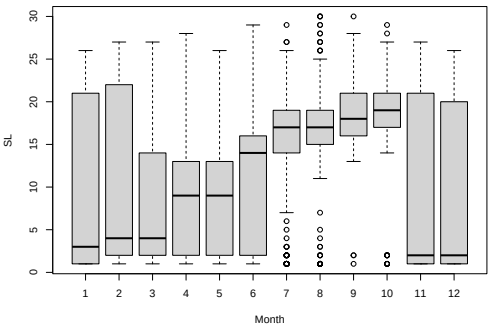
d) Redox Potential versus Month



e) pH versus Month

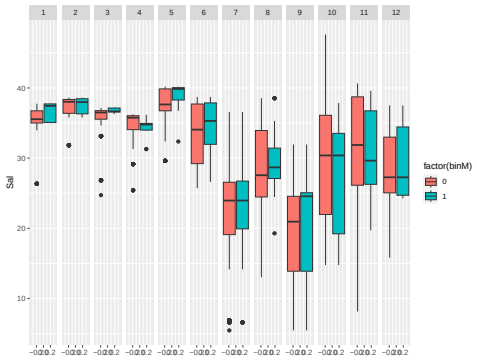


f) Organic Matter content versus Month

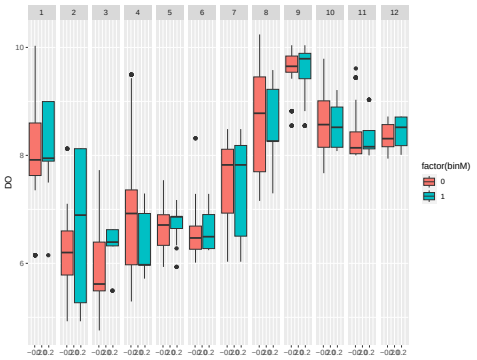


g) Cockle's Shell Length versus Month

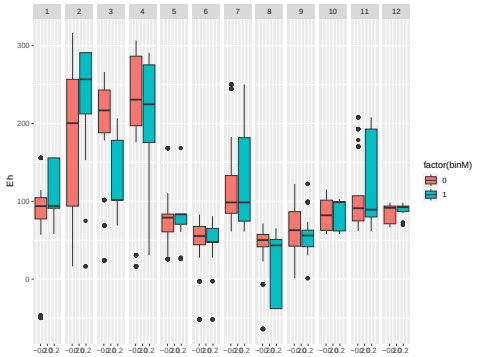
FIGURE 4.12: Boxplots of the explanatory variables versus Month. Figure created with the help of the `ggplot2` package from `R` software.



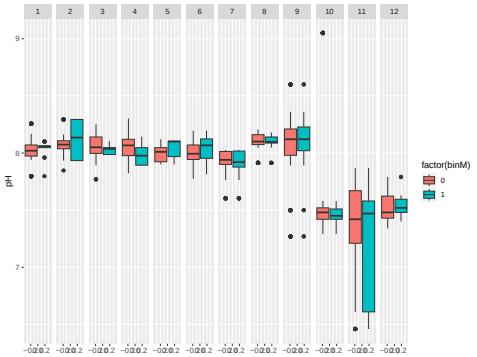
a) Salinity versus Month



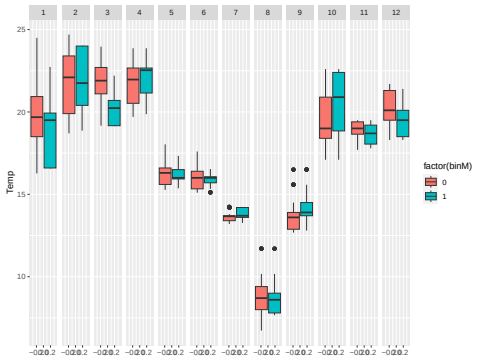
b) Dissolved Oxygen versus Month



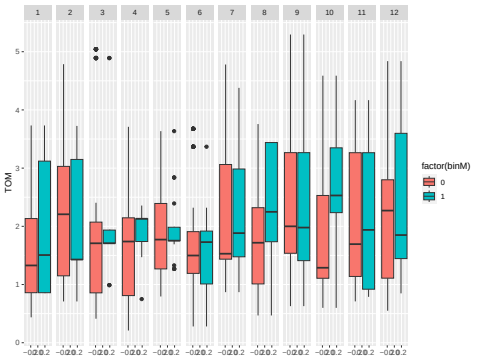
c) Redox potential versus Month



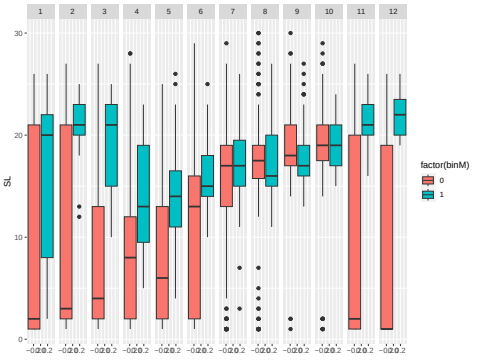
d) pH versus Month



e) Temperature versus Month



f) Organic Matter content versus Month



g) Cockle's Shell Length versus Month

FIGURE 4.13: Boxplots of the explanatory variables versus Month for non-infected (in blue) and infected (in red) cockles. Figure created with the help of the `ggplot2` package from `R` software.

Chapter 5

Model Formulation

This chapter will present the modelling of the data under study, through the application of Generalised Linear Mixed Models (GLMMs) and/ or Generalised Additive Mixed Models (GAMMs). To avoid numerical estimation problems and due to the different scales observed for the covariates, prior to starting, a standardisation procedure was performed to all continuous independent variables, in this case Shell Length (SL), Salinity (Sal), Dissolved Oxygen (DO), Redox Potential (Eh), pH, Organic Matter Content (TOM) and Temperature (Temp). Standardization was made by subtracting the sample mean and dividing by the sample standard deviation. For any give random sample $x = (x_1, \ldots, x_n)$, the standardization of x considers $(x - mean(x))/sd(x)$.

It should be recalled that the goal of this project was to determine how environmental factors affect the abundance of metacercariae parasites infecting cockles in the Ria de Aveiro, Portugal. Given the purpose of the study and given that we are dealing with count data, an initial model encompassing every variable and following the Poisson distribution, served as the foundation for all subsequent models.

- $Metacercariae_{ijt}$ represents the number of metacercariae in the j^{th} cockle from the sampling site i at month t;
- $Metacercariae_{ijt} \sim P(\mu_{ijt})$

$$\begin{aligned} \log(\mu_{ijt}) &= \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 DO_{jt} + \beta_3 Eh_{jt} + \beta_4 pH_{jt} + \\ &\quad + \beta_5 TOM_{jt} + \beta_6 Temp_{jt} + \beta_7 SL_{jt} + \beta_8 Month_{jt} \end{aligned}$$

where b_{0i} represents a random intercept for the Sampling Site (which was sampled longitudinally) and $b_{0i} \sim N(0, \sigma^2_{Site}) i = 1, \ldots, 18$.

In the subsequent models, we successively eliminated the least significant variable in the above model until all of the variables in the model were statistically significant and then interactions terms were considered.

5.1 Poisson model

The model chosen for the Poisson distribution was as follows:

$$\log(\mu_{ijt}) = \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 DO_{jt} + \beta_3 pH_{jt} + \beta_4 SL_{jt} \tag{5.1}$$

with an AIC of 2765.8. In Table 5.1 it is possible to observe the output results for the model. For a typical cockle, the abundance of metacercariae appeared to be positively correlated with cockles' shell length, dissolved oxygen and pH of the water and negatively correlated with salinity.

TABLE 5.1: Output of the Poisson mixed model.

Random effects			
	Variance	Std. Deviation	
Intercept (Site)	1.695	1.302	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-2.7603	0.3368	<0.001
SL	1.0543	0.0616	<0.001
Sal	-0.1771	0.0534	<0.001
DO	0.1103	0.0521	0.034
pH	0.0999	0.0378	0.008
AIC: 2765.8			

5.1.1 Model Validation

The first thing to do after fitting a Poisson Mixed Model is to verify whether the equidispersion assumption is valid, for instance by estimating the dispersion parameter. This should *surround* 1, in order to validate the assumption.

There are two formulas for the calculation of an estimate of the dispersion parameter, the Mean Deviance estimator and the Pearson estimator [51, 83].

The Mean Deviance estimator ($\hat{\phi}_D$) can be obtained as

$$\hat{\phi}_D = \frac{D(y, \hat{\mu})}{N - p'}$$

where $D(y, \hat{\mu})$ is the deviance of the model, N is the total number of observations and p' is the number of parameters estimated by the model.

Considering the particular case of the Poisson model, the expression can be given as

$$\hat{\phi}_D = 2\sum_i \frac{y_i \log(y_i/\hat{\mu}_i) - (y_i/\hat{\mu}_i)}{N - p'}$$

which has no known distribution. It can be shown that this Mean Deviance estimator is asymptotically unbiased and consistent.

For the definition of the Pearson estimator ($\hat{\phi}_P$), we first consider the Pearson χ^2 statistic

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

If a correction is made to the number of degrees of freedom, then the estimator is obtained as

$$\hat{\phi_P} = \frac{X^2}{N - p'}$$

where, once again, N is the total number of observations and p' is the number of parameters estimated by the model.

Considering the particular case of the Poisson model, with equal weights among all observations, the previous estimate is

$$\hat{\phi_p} = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(N - p')}$$

Knowing that the Pearson residuals for a Poisson model can be defined as $residuals_P = \frac{y_i - \hat{\mu_i}^2}{\sqrt{\hat{\mu}_i}}$, then we obtain that

$$\hat{\phi_p} = \sum(residuals_P^2)/(N - p') \tag{5.2}$$

The distribution of the Pearson estimator is not known, however the Pearson's statistics (χ^2) follows asymptotically a distribution $X_2(N - p')$. Likewise the Mean Deviance estimator, the Pearson estimator is asymptotically unbiased and consistent.

In this project, we used the Pearson estimator ($\hat{\phi}_P$), as it has been recently shown to be unbiased even for small sample sizes.

If $\hat{\phi}_P$ is *close to* 1, then the assumption of equidispersion might hold. If it is substantially larger than 1, then we have overdispersion (the variance is larger than the mean). And if it is much smaller than 1, then we have underdispersion (the variance is smaller than the mean).

For model 5.1, we obtained $\hat{\phi}_P = 1.10$, which is not very far away from 1. However, to understand if the model was able to cope with this dispersion, we conducted a simulation study.

We used the *simulate()* function in R, to simulate a large number of truly Poisson responses (we considered 1000) from the model. For each simulated response, we ran the model and computed the dispersion statistic for each model. Finally, we considered a histogram of those values. Then, we superimposed the dispersion statistic for the observed data in the histogram; if our value is close to the center of the histogram, then we can conclude that the observed data complies with a Poisson regression, at least with respect to the mean-variance relationship.

In Figure 5.1, it is possible to observe the histogram with the results of the simulated study and, in a red dot, the dispersion statistic calculated for the Poisson model 5.1. These results do not make sense, however, since the estimates obtained from the simulations should all be around 1, as the simulations come from a Poisson regression. What happens is that the *simulate()* function is doing two levels of simulations: one at the random effect level and another one at the conditional regression model. The effect is to simulate data from a log-normal response, with greater variability that a Poisson variable – see Rui Miranda's master thesis at the Faculty of Sciences of the University of Porto for further details.

The outcomes of a different simulation study for the dispersion statistics are shown in Figure 5.2. Instead of utilising the *simulate()* function in this instance, a matrix containing 1000 Poisson distributions was formed with λ equal to the Poisson distribution of the first selected model. The dispersion statistic for each distribution was then computed, and a histogram was made. The dispersion statistic values in this histogram ranged from 0.6 to approximately 1.7 with the median at 0.995. Since the dispersion statistic of the obtained model (depicted as a red dot) was about in the middle of the histogram, it appears that there were no dispersion concerns.

Another possibility in studying dispersion statistics is the use of the `DHARMa` package in R and its *testDispersion* function. This function in `DHARMa` simulates datasets, calculates the variance of all simulated data and compares the variance of the observed residuals against the variance of the simulated residuals via their ratio. More information on this is given in section 3.2 of chapter 3.

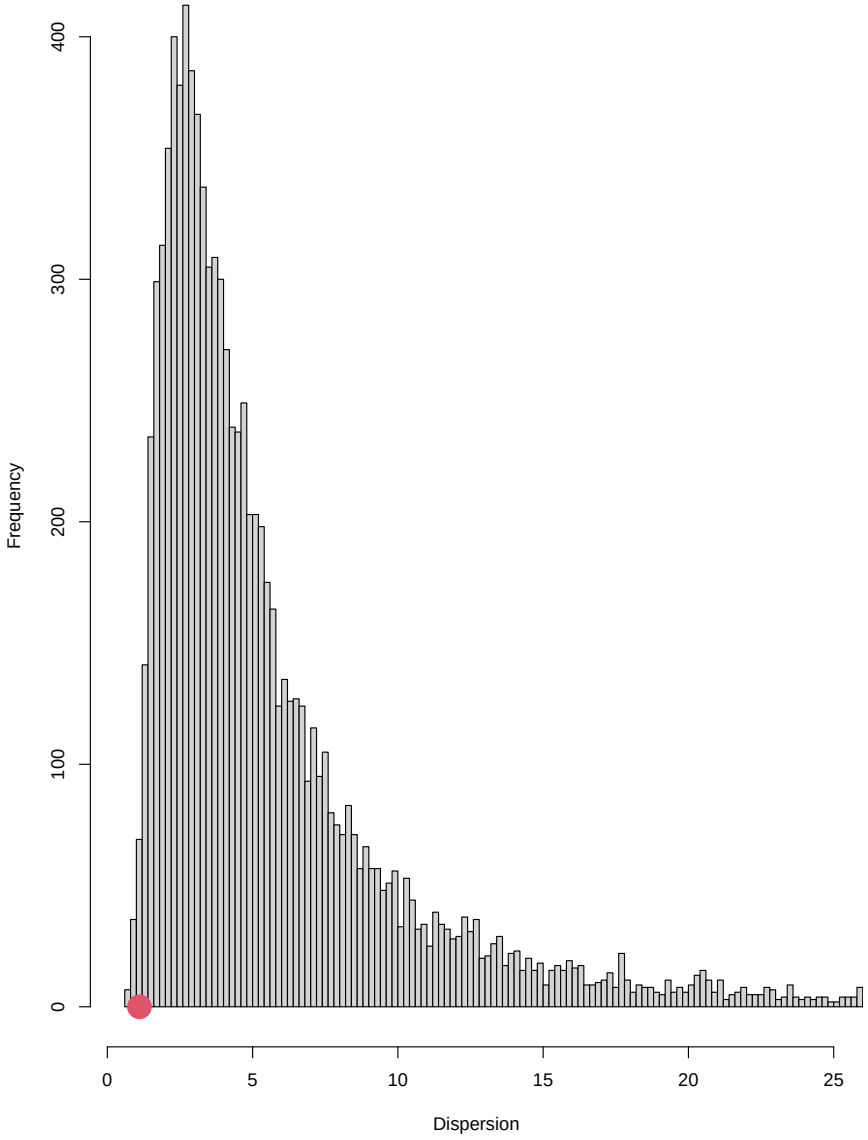


FIGURE 5.1: Histogram of the dispersion statistic frequency for the simulated datasets, using the *simulate()* function, with the dispersion statistic obtained by the model superimposed as a red dot.

Figure 5.3 shows the *testDispersion* plot and its results. The obtained p-value of 0.736 shows no evidences to reject the null hypothesis, in which the variance of the observed residuals are equal to the variance of the simulated residuals, so there are no over- or underdispersion problems in this model.

The dispersion statistics will now only be analysed using `DHARMa`, in the validation of the remaining models.

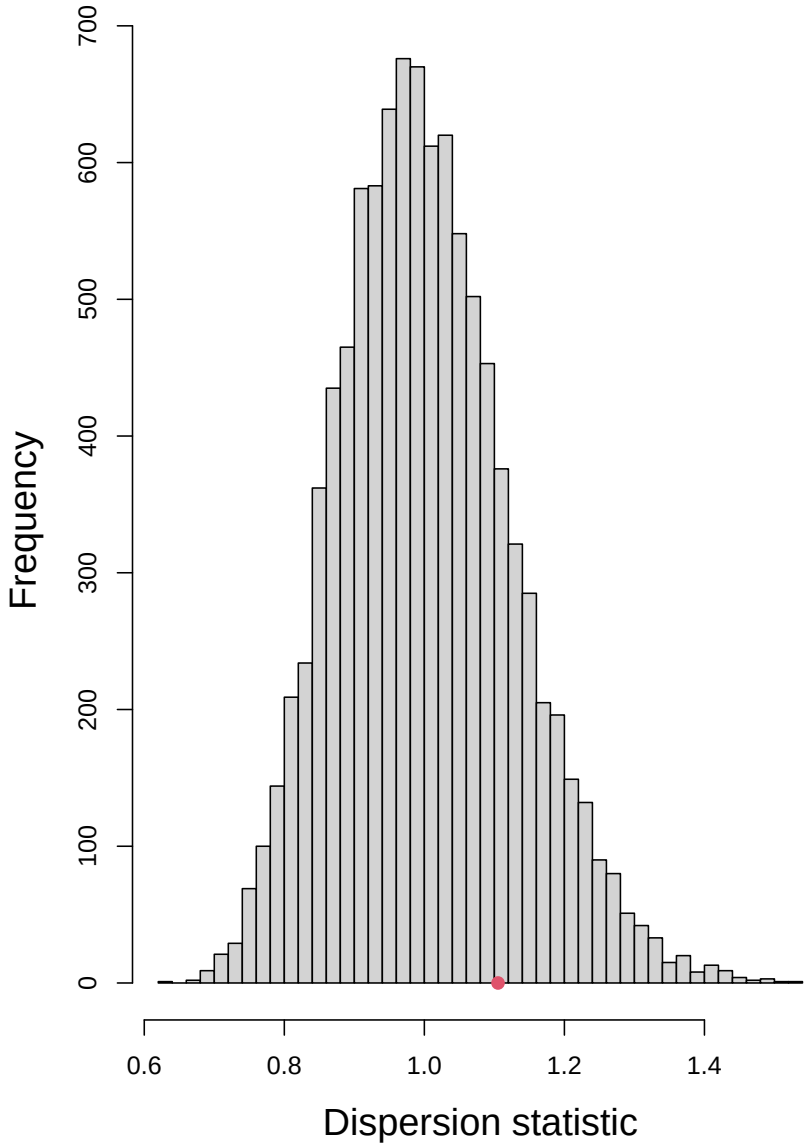


FIGURE 5.2: Histogram of the dispersion statistic of the second simulation study, with the dispersion statistic obtained by the model superimposed as a red dot.

Following the examination of the equidispersion condition, we then evaluated the model's capacity to handle zeros (in this case 86.7 %, in the response). Similarly to the study of the dispersion statistic, this analysis can be done through a simulation study or using the `DHARMa` package and its *testZeroInflation()* function.

For the simulation, as before, a large number of truly Poisson responses with mean equal to the fitted values of the initially obtained model is simulated, and a Poisson model is fitted. But, instead of computing the dispersion statistic for each model, the number of expected zeros is calculated. Then, a histogram with the number of expected zeros is

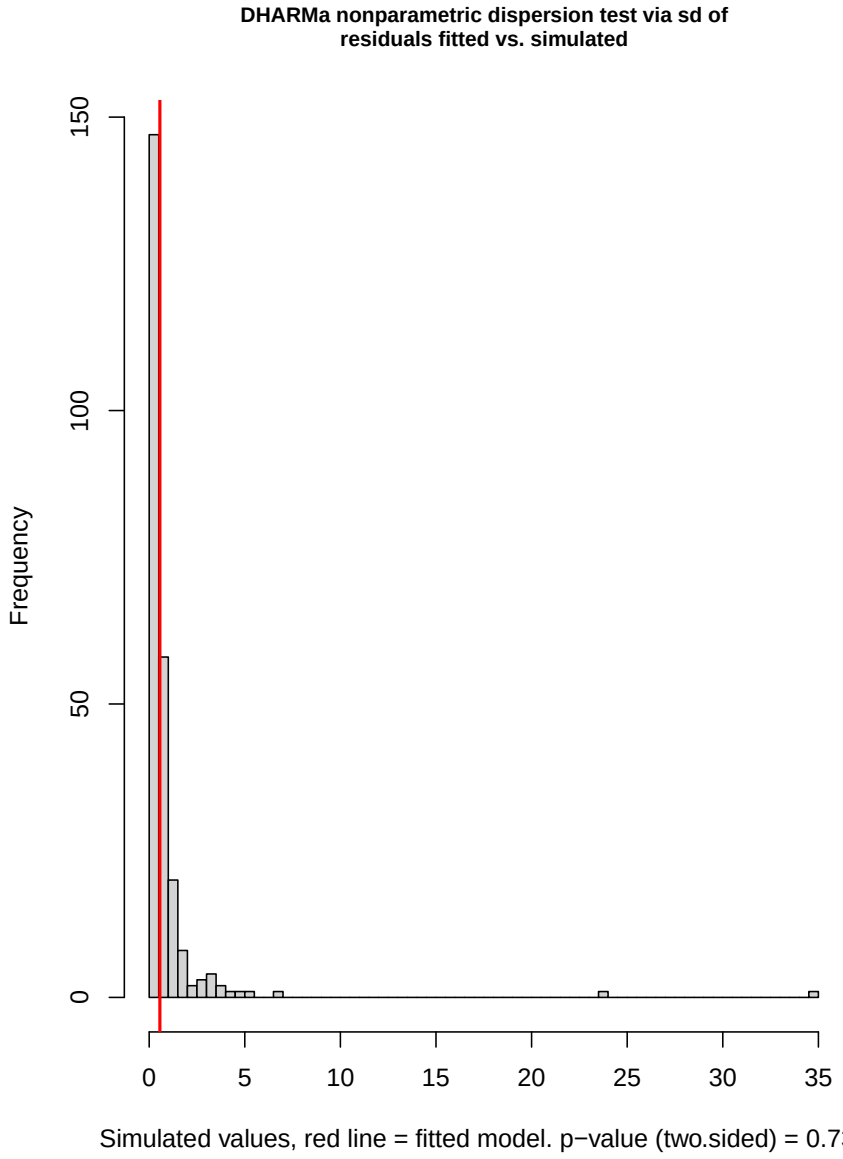


FIGURE 5.3: `DHARMa` plot for the *testDispersion()* of the Poisson Mixed model.

created and the number of zeros of our initial dataset is superimposed as a red dot. If the number of zeros in our dataset is relatively in the middle of the distribution of zeros found in the histogram, it is likely that the model can cope this observed amount of zeros. This analysis is relevant in assessing whether the Poisson model should be changed to a zero-inflated or hurdle model.

Figure 5.4 displays the histogram for the percentage of zeros of 1000 simulated datasets. The number of zeros that the simulated datasets presented varied from approximately 67 % to 96.5 % of zeros, with the median being at 85 %. Therefore, our model seems to be able to cope with the percentage of zeros of our dataset.

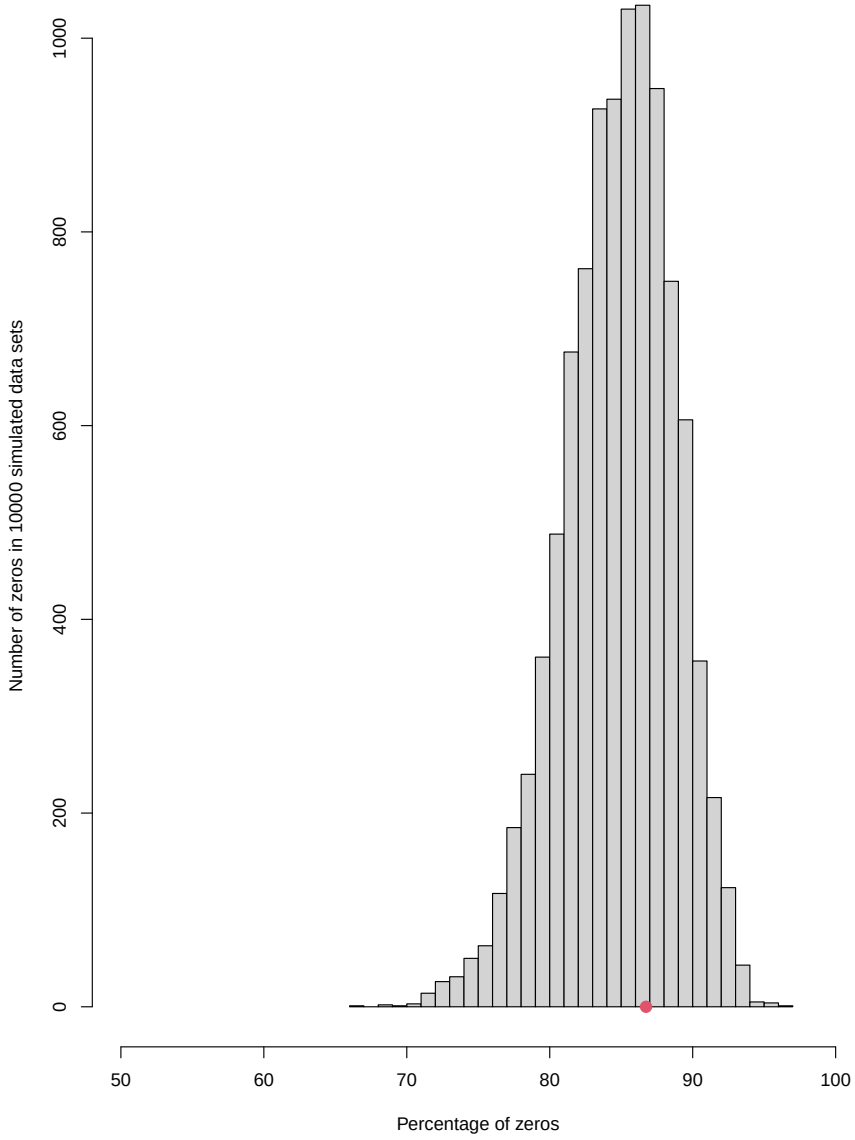


FIGURE 5.4: Histogram of the percentage of observed zeros frequency for the simulated datasets, using the *simulate()* function, with the percentage of zeros obtained by the model superimposed as a red dot.

In order to see if the same finding still holds, the `DHARMa` package was also used to study the capacity of our model to cope with zeros. The findings and corresponding statistical test are shown in Figure 5.5, demonstrating once more that our model can accommodate the number of zeros observed.

As for the dispersion statistic, for the remaining models and consequent model validation, only the `DHARMa` package will be used since the returned results seem to those

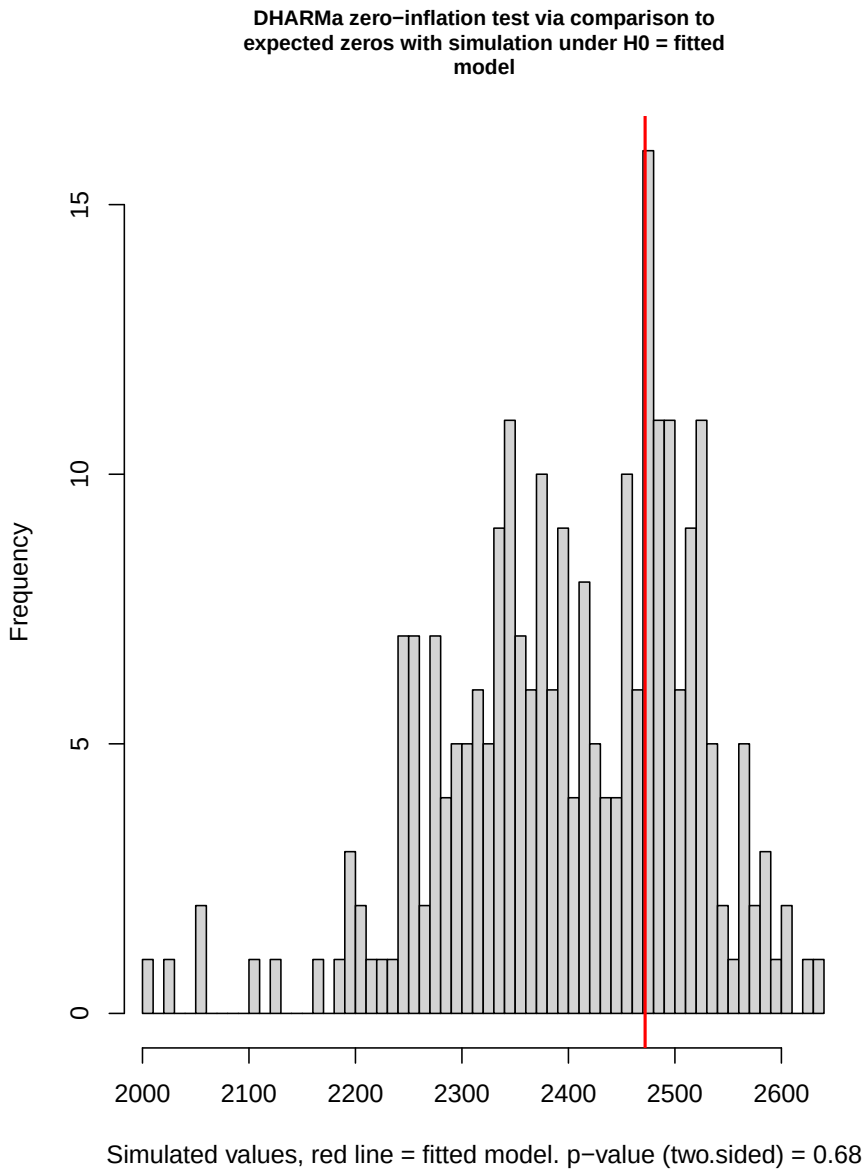


FIGURE 5.5: `DHARMa` plot for the *testZeroInflation()* function for the Poisson Mixed model.

obtained by the different methods.

In the `DHARMa` residual vs. predictor plot of Figure 5.6, it is expected that the points will bounce randomly around the horizontal line $y = 0.5$. The superimposed dashed red line obtained by quantile regression, raises the possibility of non-linear relationship between predictors and log-response, which would be modelled by an Additive Poisson Model. In fact, this scenario was already taken into account in the exploratory data analysis. A similar trend can be observed in the plot displayed in Figure 5.7, which shows a slight dependence of the Pearson residuals on the month.

The plot in Figure 5.8 is difficult to interpret but resembles similar plots for Poisson

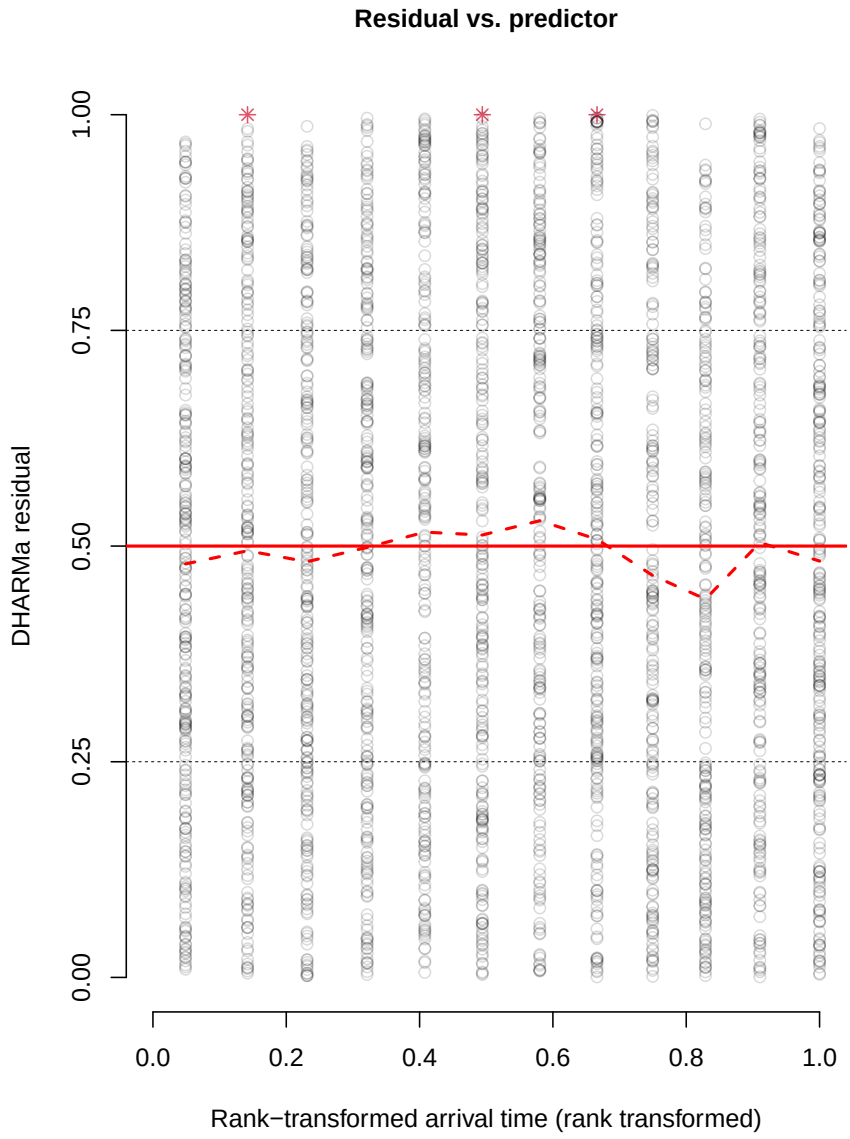


FIGURE 5.6: `DHARMa` residual vs. predictor plot for the Poisson Mixed model. Line in red means that statistically significant problems were detected.

regressions in the literature. It is possible to observe clear bands of points due to the number of zeros and the discrete nature of the data. Also, there is a group of outliers, that should be approached carefully.

In the 3D scatter plot in Figure 5.9, the month is depicted on the x-axis, while the metacercariae counts are shown on the y-axis. The contour of the area in red represents the fitted values of the Poisson Mixed Model. Similar to what is observed in Figure 4.9 of the section 4.4 of the chapter 4, the fitted values are close to zero due to the vast number of observations equal to zero recorded throughout the months. The Poisson probability

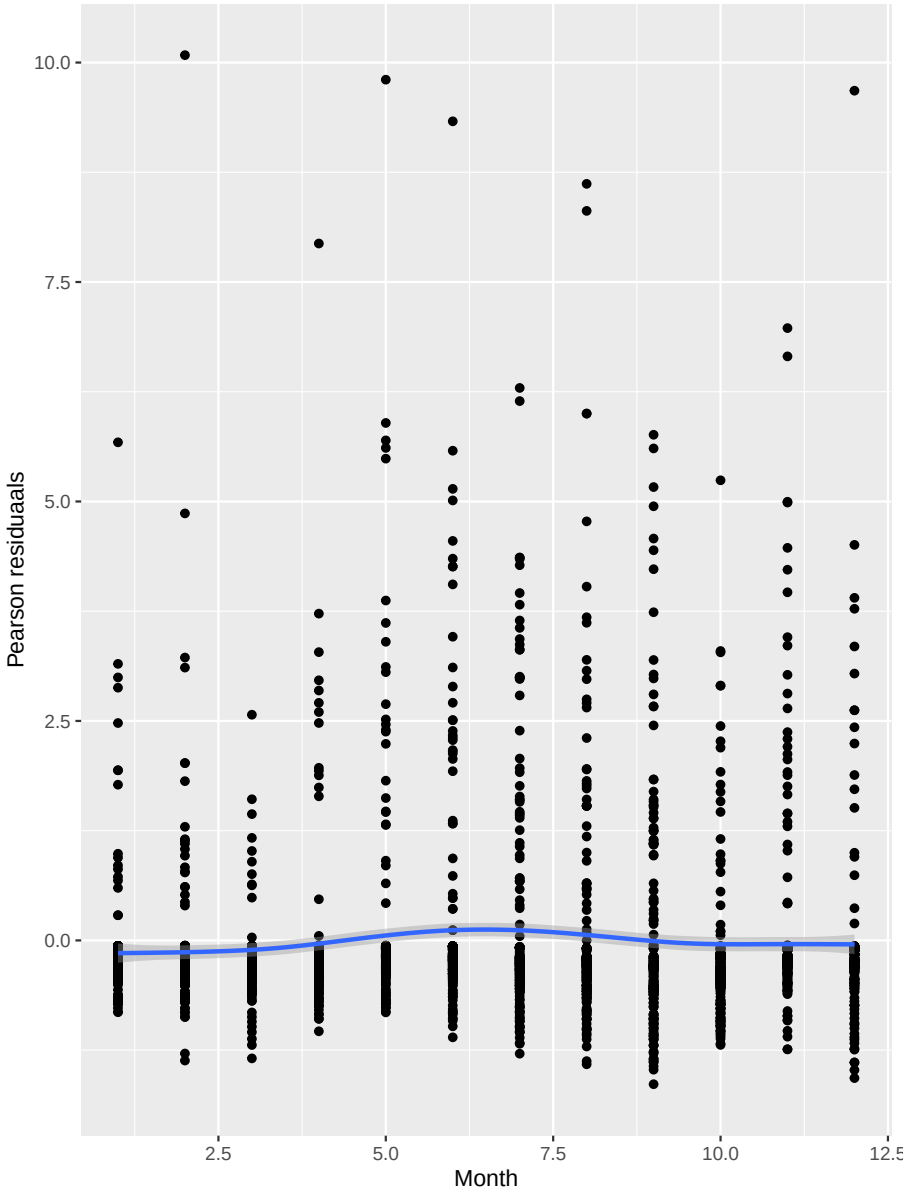


FIGURE 5.7: Plot of Pearson residuals per sampled month with smoother added for visual interpretation of the relationship.

function for each of the months is overlayed in purple. Despite some observations reaching 12 metacercariae, the model essentially predicts counts between 0 and 2 metacercariae, with cases that exceed 2 metacercariae being rare (the exception is month 9, which can model up to 4 metacercariae). It should be emphasized that random effects are not taken into consideration to make this graph, and they do interfere with the conditional Poisson probability functions.

Although this first model 5.1 has no severe issues, its prediction ability is not very satisfactory (Table 5.2). Either one would discard counts less than or equal to 3 or 4, in

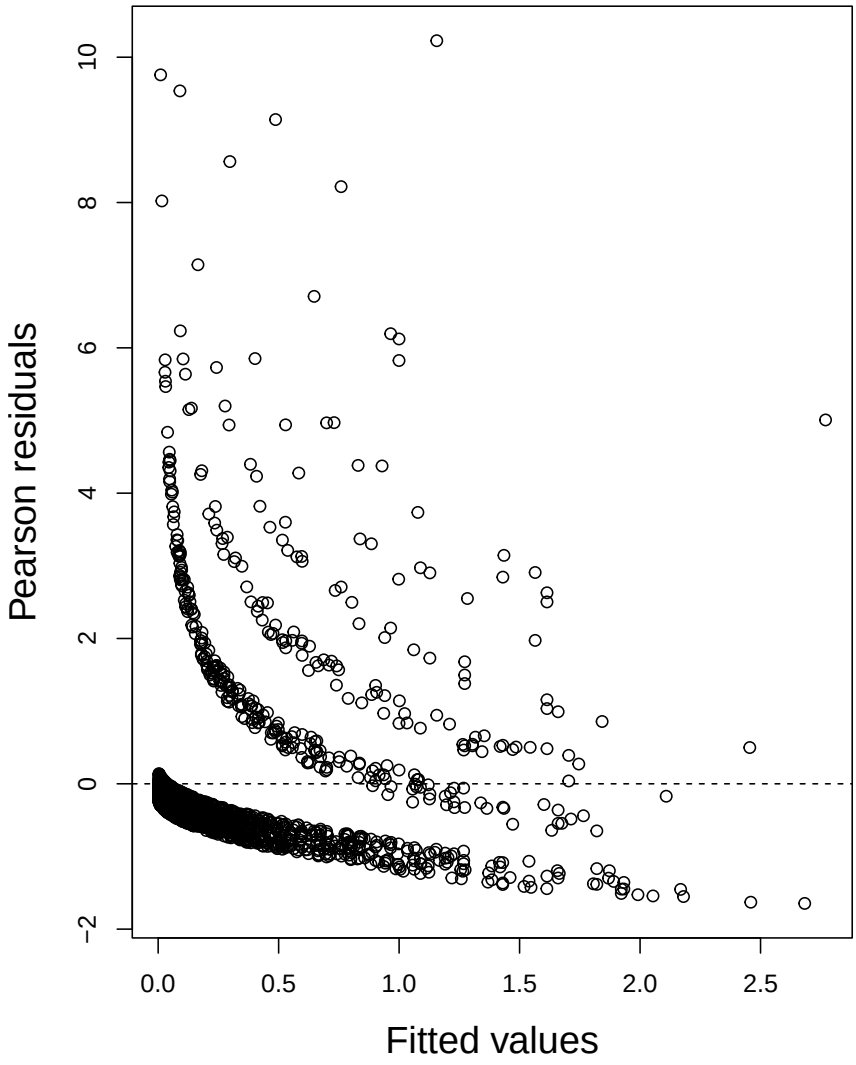


FIGURE 5.8: Residuals vs. fitted plot for the Poisson Mixed model.

the response, or else one would have to accept that the right predictors are only going up to 2. We then decided to approach the problem with different models, both in the hope of improving the model's predictions and for didactic reasons, as the study of ecological modelling is of great interest to the author. Hence, alternative distributions and zero-inflated models were also considered and contrasted with the Poisson model.

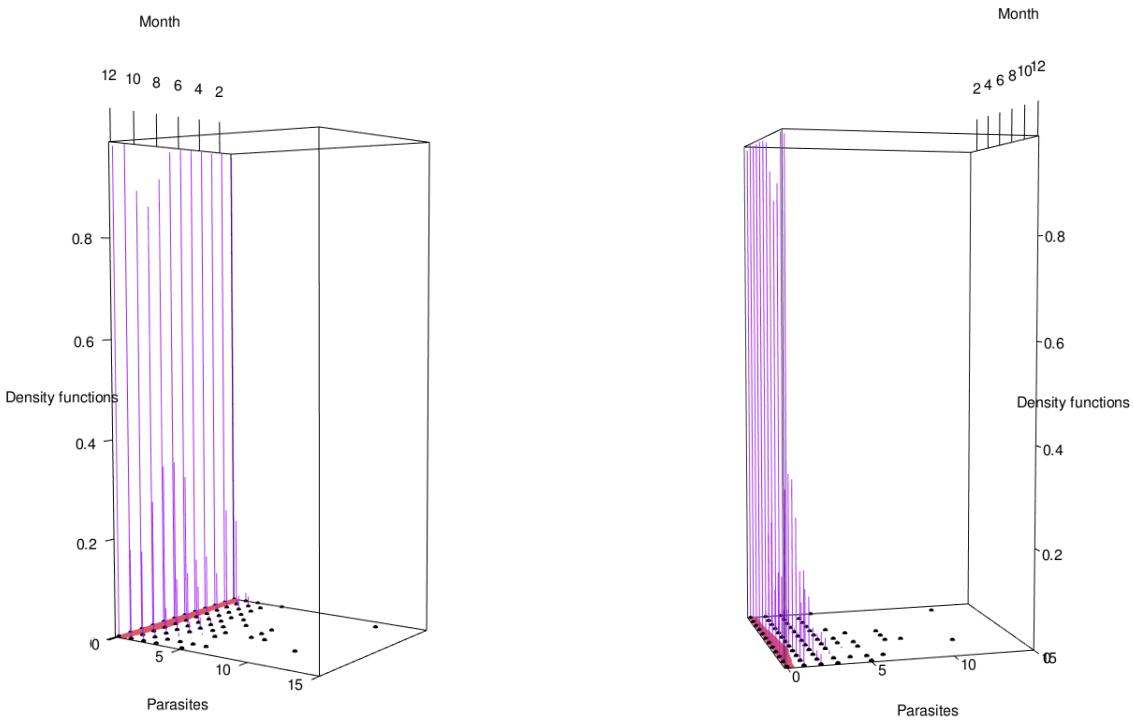


FIGURE 5.9: 3D scatter plot of the fitted Poisson Mixed model. The red line represents the fitted values, the Poisson probability function curves for each month are represented in purple, and the black dots represent the observed values.

TABLE 5.2: Confusion matrix of the obtained model with the Predicted values in the columns and the Observed values in the rows.

		Predicted			
		0	1	2	3
Observed	0	2249	198	24	1
	1	153	77	8	0
	2	29	45	6	0
	3	9	17	5	0
	4	1	8	1	0
	5	1	7	3	0
	6	0	1	0	0
	7	1	3	0	0
	8	0	1	0	0
	11	0	0	0	1
	12	0	1	0	0

5.2 Negative Binomial model

The Negative Binomial distribution was the subject of the second model to be used. After model selection, done essentially by the backwards elimination procedure, the final model

was

$$\log(\mu_{ijt}) = \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 pH_{jt} + \beta_3 SL_{jt} \tag{5.3}$$

The output results of the fitted model are represented in Table 5.3. The AIC of the model was 2547.1, which was lower than the Poisson Mixed Model (Table 5.1), with a Negative Binomial parameter k of 0.626. For a typical cockle, the abundance of metacercariae was shown to be positively correlated with cockle's shell length and pH and negatively correlated with salinity.

TABLE 5.3: Output of the Negative Binomial mixed model.

Random effects			
	Variance	Std. Deviation	
Intercept (Site)	1.744	1.321	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-2.8463	0.3450	<0.001
SL	1.2196	0.0864	<0.001
Sal	-0.3094	0.0631	<0.001
pH	0.0842	0.0499	0.092
AIC: 2547.1			
Dispersion parameter k**:** 0.626			

5.2.1 Model Validation

The capacity of the obtained model 5.3 to handle the quantity of zeros was also tested through the *testZeroInflation* function of `DHARMa`. The obtained p-value of 0.76 shows that our model can account for the observed number of zeros. The histogram and test results are displayed in Figure 5.10.

As for the previous Poisson mixed model, in the residual vs. predictor plot of Figure 5.11, the superimposed dashed and solid red lines are similar, therefore, the adjustment of a linear predictor seems reasonable. However, there is a small deviation of the dashed red line, that might increase our suspicion of a non-linear pattern, which would consequently require the use of an additive model. A similar pattern of non-linearity was described in the descriptive analysis regarding some of the variables.

In the residuals vs. fitted values plot, represented in Figure 5.12, it is possible to observe some outlier values that should be approached carefully.

The 3D scatter plot represented in Figure 5.13 is similar to that of Figure 5.9 (the month is depicted on the x-axis, while the metacercariae counts are shown on the y-axis), but for

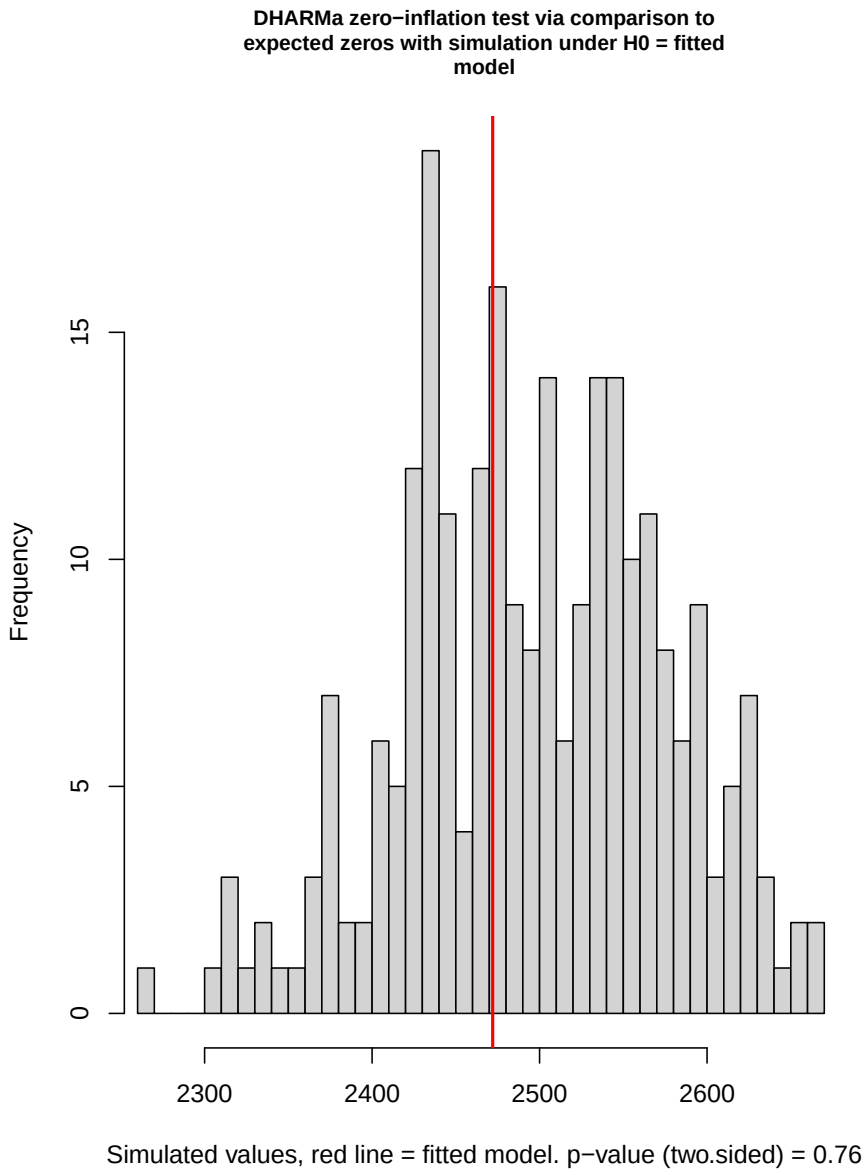


FIGURE 5.10: `DHARMa` plot for the *testZeroInflation()* function for the Negative Binomial Mixed model.

the Negative Binomial Mixed Model. The contour of the area in red represents the fitted values and overlayed in purple are the probability function values for each of the months. We see that the model is able to predicted higher counts than the Poisson model, namely on months 8 to 10, reaching up to 8 *metacercariae.cockle*$^{-1}$. This is not surprising, as the model can handle a large variance. Nonetheless, it is important to emphasize, once again, that random effects were not taken into consideration for this graph, and the predictions obtained in the confusion matrix (described next) are way lower.

The Negative Binomial mixed model actually presented a lower AIC than the Poisson

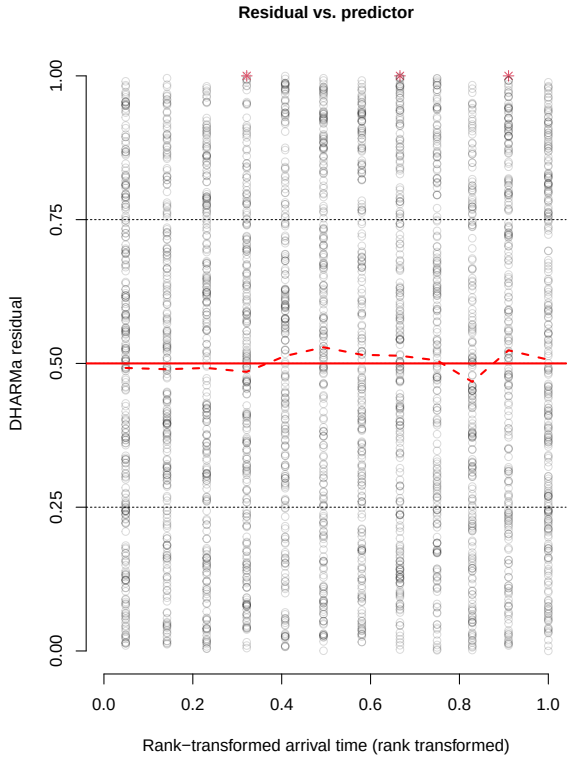


FIGURE 5.11: `DHARMa` residual vs. predictor plot for the Negative Binomial Mixed model. Line in red means that statistically significant problems were detected.

mixed model (5.1) and managed to predict slightly higher counts (up to 4 – compare NB confusion matrix (Table 5.4) with P confusion matrix (Table 5.2)). The latter, however, was not particularly relevant, since the model only predicted (wrongly) a single count of 4 and had similar correct counts for counts of 1 (33.2 %) and 2 (8.8 %) .

TABLE 5.4: Confusion matrix of the obtained Negative Binomial mixed model with the Predicted values in the columns and the Observed values in the rows.

		Predicted				
		0	1	2	3	4
Observed	0	2235	201	31	4	1
	1	148	79	11	0	0
	2	28	44	7	1	0
	3	9	19	3	0	0
	4	2	7	1	0	0
	5	1	9	1	0	0
	6	0	1	0	0	0
	7	0	4	0	0	0
	8	0	1	0	0	0
	11	0	0	0	1	0
	12	0	1	0	0	0

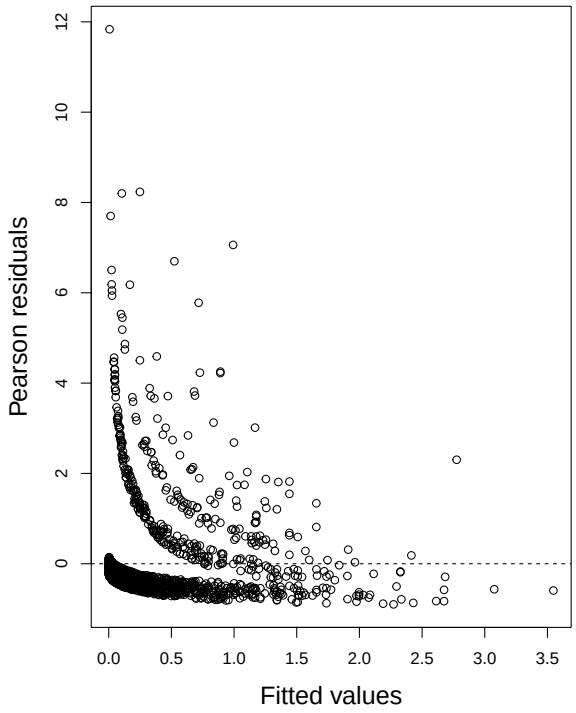


FIGURE 5.12: Residuals vs. fitted plot for the Negative Binomial Mixed model.

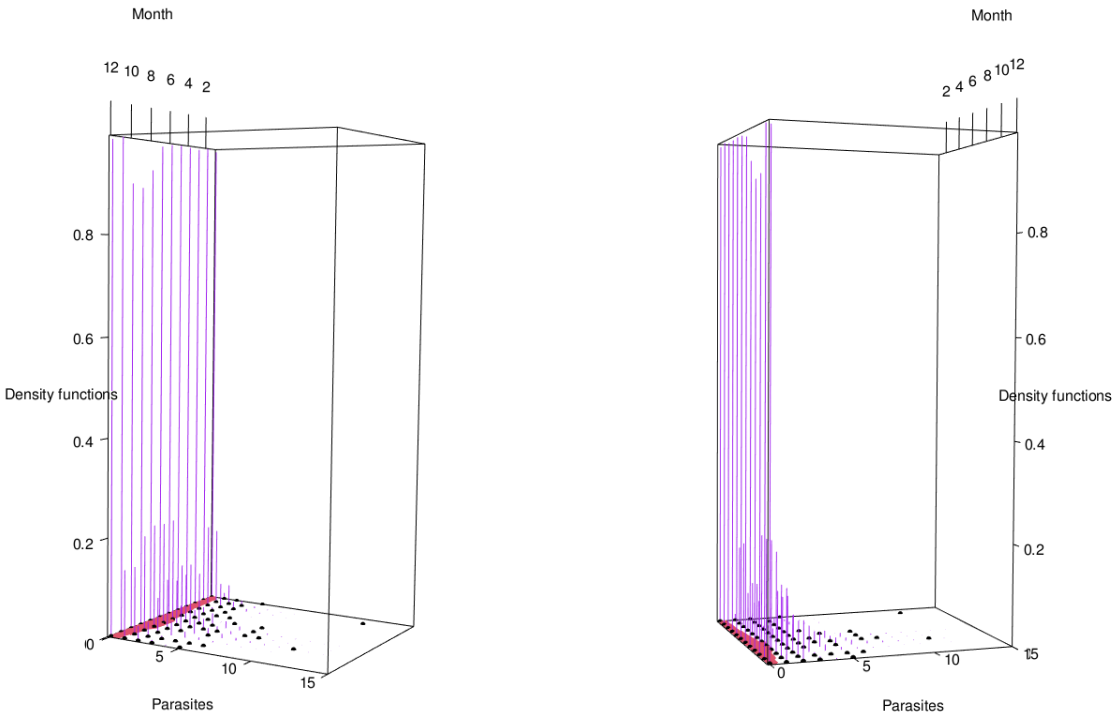


FIGURE 5.13: 3D scatter plot of the fitted Negative Binomial Mixed model. The red line represents the fitted values, the Negative Binomial probability function curves for each month are represented in purple, and the black dots represent the observed values.

5.3 Generalised Poisson model

The third applied model was the Generalised Poisson Mixed Model. As previously discussed in chapter 2, the Generalised Poisson distribution allows modelling for over- and underdispersed data. Additionally, this distribution has a more flexible parameter (ϕ) than the Negative Binomial, which provides a better fit for overly dispersed data, for example [84].

After model selection, the obtained Generalised Poisson Mixed Model was

$$\log(\mu_{ijt}) = \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 pH_{jt} + \beta_3 Eh_{jt} + \beta_4 SL_{jt} \tag{5.4}$$

The model showed a AIC of 2581.7, and a dispersion parameter (ϕ) equal to 1.87. The dispersion parameter shapes the Generalised Poisson variance by scaling it proportionally to the mean ($Var(Metacercariae_{ijt}) = \phi \times \mu_{ijt}$). The output of the Generalised Poisson Mixed Model is represented in Table 5.5. The abundance of metacercariae showed to be positively correlated with cockle's shell length and water pH, whereas the water salinity and redox potential (Eh) were negatively correlated. For a typical cockle, a one-standard-deviation increase in the cockle's shell length is associated with an average increase in the expected number of parasites infecting cockles by 2.7 ($RR = \exp(0.9941) = 2.7$) For a typical cockle, the abundance of metacercariae showed to be positively correlated with cockle's shell length and water pH, whereas the water salinity and redox potential (Eh) were negatively correlated.

TABLE 5.5: Output of the Generalised Poisson mixed model.

Random effects			
	Variance	Std. Deviation	
Intercept (Site)	1.357	1.165	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-2.5786	0.3086	<0.001
SL	0.9941	0.0719	<0.001
Sal	-0.2251	0.0518	<0.001
pH	0.0985	0.0537	0.066
Eh	-0.1021	0.0584	0.080
AIC: 2581.7			
Dispersion parameter ϕ: 1.87			

5.3.1 Model Validation

We first checked how the model handled with the percentage of zeros. In Figure 5.14 it is represented the histogram and the statistical result of the *testZeroInflation* function. The p-value obtained was of 0.83, not showing evidence to reject the null hypothesis.

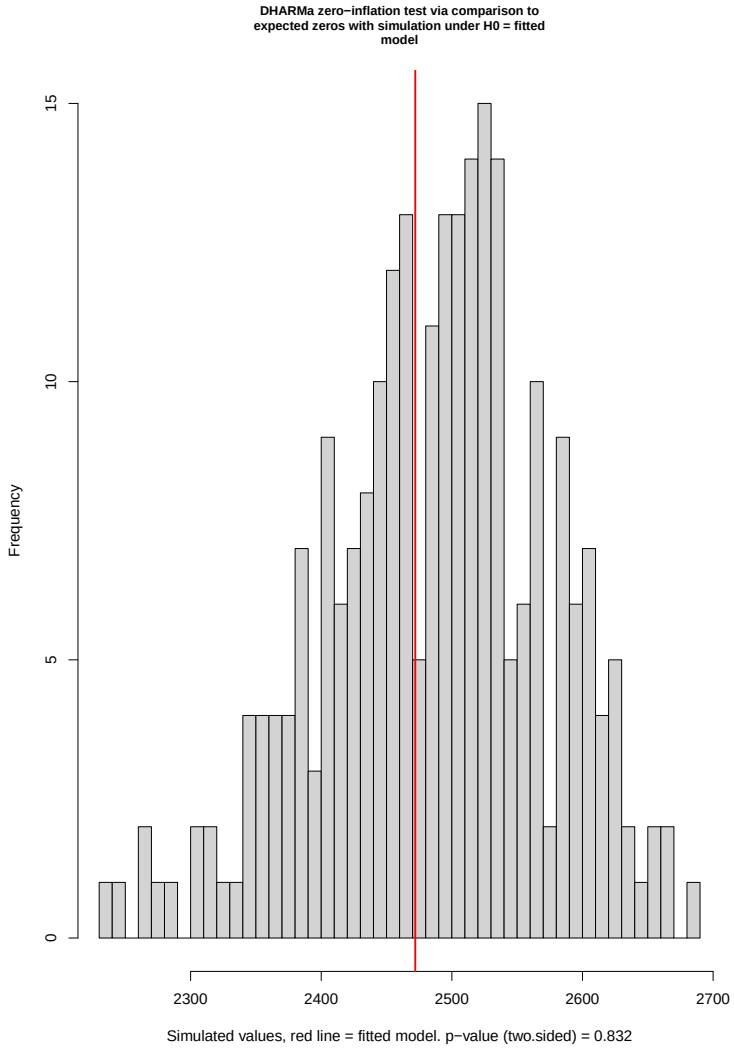


FIGURE 5.14: `DHARMa` plot for the *testZeroInflation()* function for the Generalised Poisson Mixed model.

In Figure 5.15 it is represented a QQ plot to check whether the scaled quantile residuals are uniformly distributed. The obtained p-value for the Kolmogorov-Smirnov test ($p - value = 0.363$) does not give evidence to reject the null hypothesis that the scaled quantile residuals follow a uniform distribution.

In this model, the deviation between the dashed and solid red lines in the residuals vs. predictor plot *Figure 5.16 is exacerbated in comparison to the patterns seen in the previous Poisson and Negative Binomial model. This increases suspicion that our data

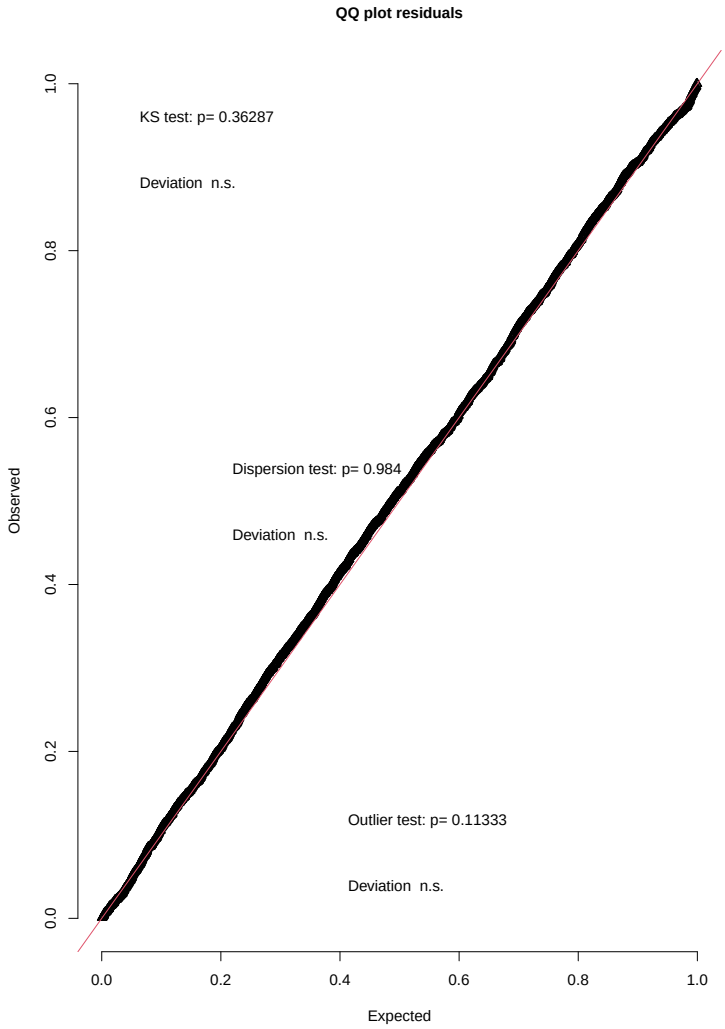


FIGURE 5.15: texttttDHARMa residuals quantile-quantile plot for the Generalised Poisson Mixed model.

may require an additive model and presents non-linear behaviour. Although in the previous models presented, a linear model seems reasonable, in the Generalised Poisson model the deviation appears significant enough that it should not be ignored. Furthermore, the `DHARMa` package assesses the deviation from uniformity and presents results deemed suspect in red, rather than in the usual black.

In the non-transformed fitted vs. residuals values plot, represented in Figure 5.17, it is possible to observe some outliers values that should be approached carefully.

The 3D scatter plot for the Generalised Poisson Mixed Model and the estimated probability functions for each month are represented in Figure 5.18. As in the case of the Negative Binomial regression, the Generalised Poisson model predicted higher counts compared to the Poisson mixed model. In this case, counts reached up to 11 $metacercariae.cockle^{-1}$,

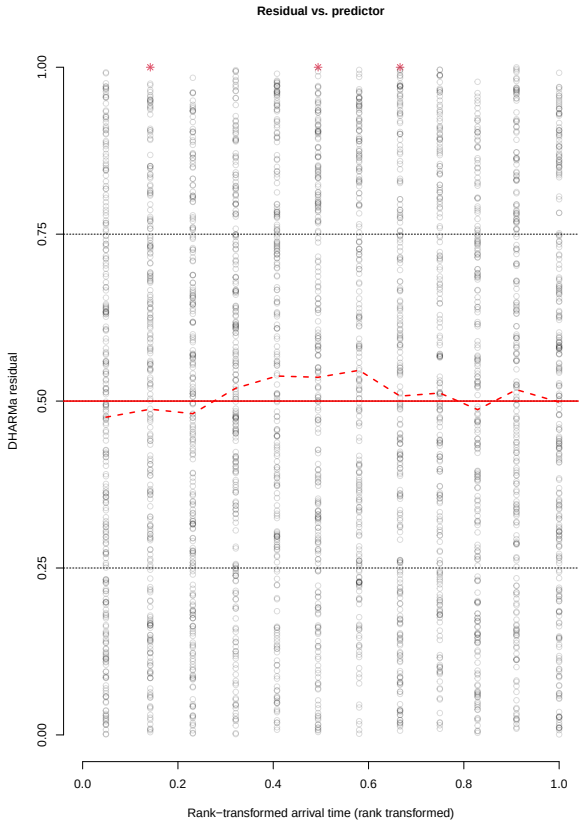


FIGURE 5.16: `DHARMa` residual vs. predictor plot for the Generalised Poisson Mixed model. Line in red means that statistically significant problems were detected.

however, most of the predictions did not go further than 3.

The Generalised Poisson mixed model had a similar AIC compared to the Negative Binomial model and lower than the Poisson model and did not exhibited any serious issues despite the possibility of once more having to fit generalised additive mixed models instead of generalised linear mixed models. The Generalised Poisson Mixed Model displayed better predictions for counts of 1 (34.5 % of correct predictions), however, displayed worse predictions than the earlier models for the remaining counts, with the highest predicted count being 2 and only correctly predicted 6.3 % of the times (Table 5.6).

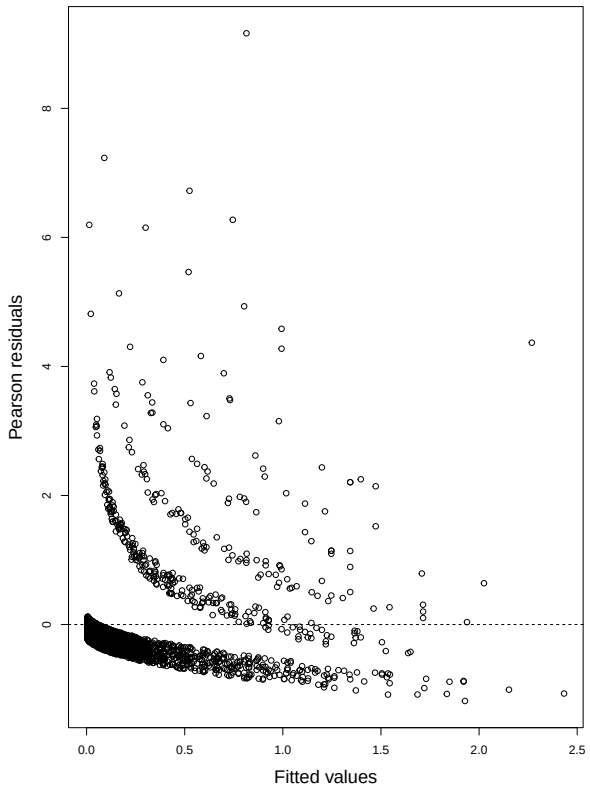


FIGURE 5.17: Residuals vs. fitted plot for the Generalised Poisson Mixed model.

TABLE 5.6: Confusion matrix of the obtained Generalised Poisson mixed model with the Predicted values in the columns and the Observed values in the rows.

		Predicted		
		0	1	2
Observed	0	2250	205	17
	1	152	82	4
	2	28	47	5
	3	10	19	2
	4	1	9	0
	5	1	10	0
	6	0	1	0
	7	0	4	0
	8	0	1	0
	11	0	0	1
	12	0	1	0

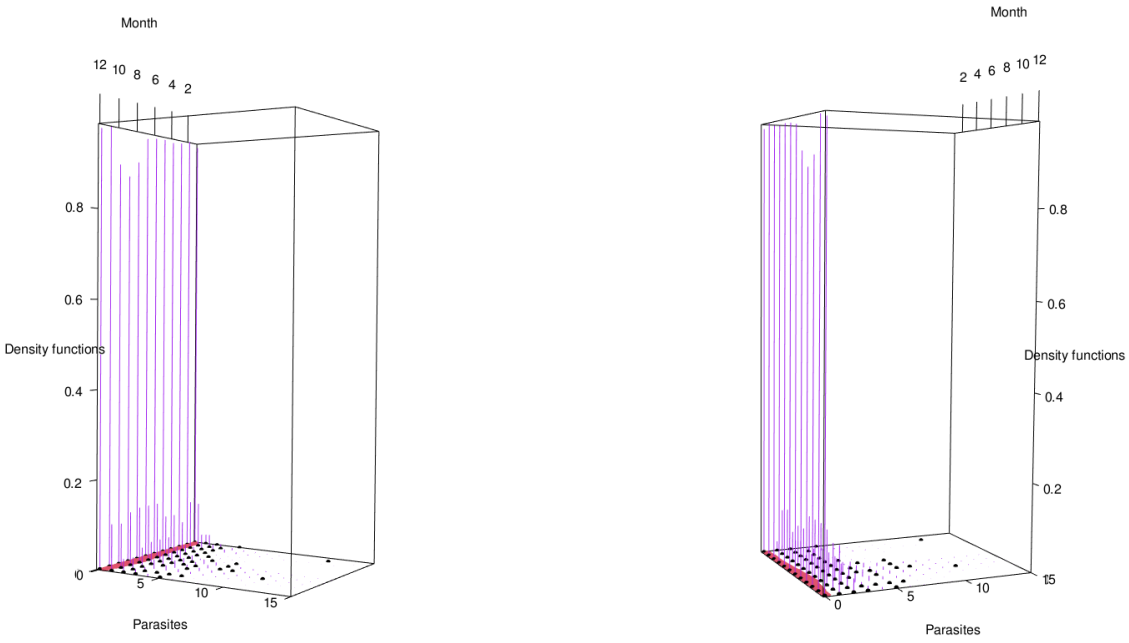


FIGURE 5.18: 3D scatter plot of the fitted Generalised Poisson Mixed model. The red line represents the fitted values, the Generalised Poisson probability function curves for each month are represented in purple, and the black dots represent the observed values.

5.4 Zero-Inflated Poisson model

The next two models to be applied are models that account for the inflation of zeros. As we have mentioned in chapter 4,given the high percentage of zeros observed in this study (about 86.7 %), one could instantly think of fitting a zero-inflated or hurdle model. That would be true in the majority of situations. However, in the last three models described (Poisson 5.1; Negative Binomial 5.3, and Generalised Poisson 5.4), no major concerns were raised regarding the number of zeros. Anyway, these models were applied to the dataset because of their didactic significance.

The Zero-Inflated Poisson mixed model was the first model to be applied. The zero-inflated Poisson model employs two components, each corresponding to different generating processes [8, 82, 85, 86]. A binary (*logit*) distribution governs the first process, while the Poisson distribution governs the second, in a way that if a random variable $Y \sim ZIP(\mu, \pi)$

$$P(Y = y|\mu, \pi) = \begin{cases} \pi + (1+\pi)e^{-\mu}, & \text{if } y = 0 \\ (1-\pi)\frac{\mu^y e^{-\mu}}{y!}, & \text{if } y \geq 1 \end{cases}$$

and,

$$\begin{aligned} E(Y) &= (1-\pi) \times \mu \\ Var(Y) &= (1-\pi) \times (\mu + \pi \times \mu^2) \end{aligned}$$

The adopted formula for this model was

$$\begin{aligned} \log(\mu_{ijt}) &= \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 DO_{jt} + \beta_3 SL_{jt} \\ logit(\pi_{ijt}) &= \gamma_0 + \gamma_1 SL_{jt} \end{aligned} \tag{5.5}$$

with an AIC of 2585.6. As it can be seen in the output results table depicted in Table 5.7, the variable *SL* is not statistically significant on the conditional model. However, this model was selected due to display the best AIC of all the tested zero-inflated Poisson mixed models. The probability of a cockle to be infected by metacercariae showed to be negatively correlated with cockle's shell length (Binomial (zero-inflation) model). On the other hand, for the conditional model, the abundance of metacercariae in the typical cockle showed to be positively correlated with cockle's shell length (despite not statistically significant) and dissolved oxygen in the water, and negatively correlated with salinity.

TABLE 5.7: Output of the Zero-Inflated Poisson mixed model.

Conditional model:			
Random effects			
	Variance	Std. Deviation	
Intercept (Site)	1.527	1.236	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-1.4867	0.3507	<0.001
SL	0.2802	0.1743	0.108
Sal	-0.1317	0.0696	0.059
DO	0.1550	0.0608	0.011
Zero-inflation model:			
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	0.7025	0.1374	<0.001
SL	-1.4512	0.2872	<0.001
AIC: 2585.6			

5.4.1 Model Validation

As for the Poisson mixed model validation (subsection 5.1.1), the first step that we need to check is if the model is able to cope with dispersion and the amount of zeros. For that, the *testDispersion()* and *testZeroInflation()* functions of the `DHARMa` package were applied (Figure 5.19)

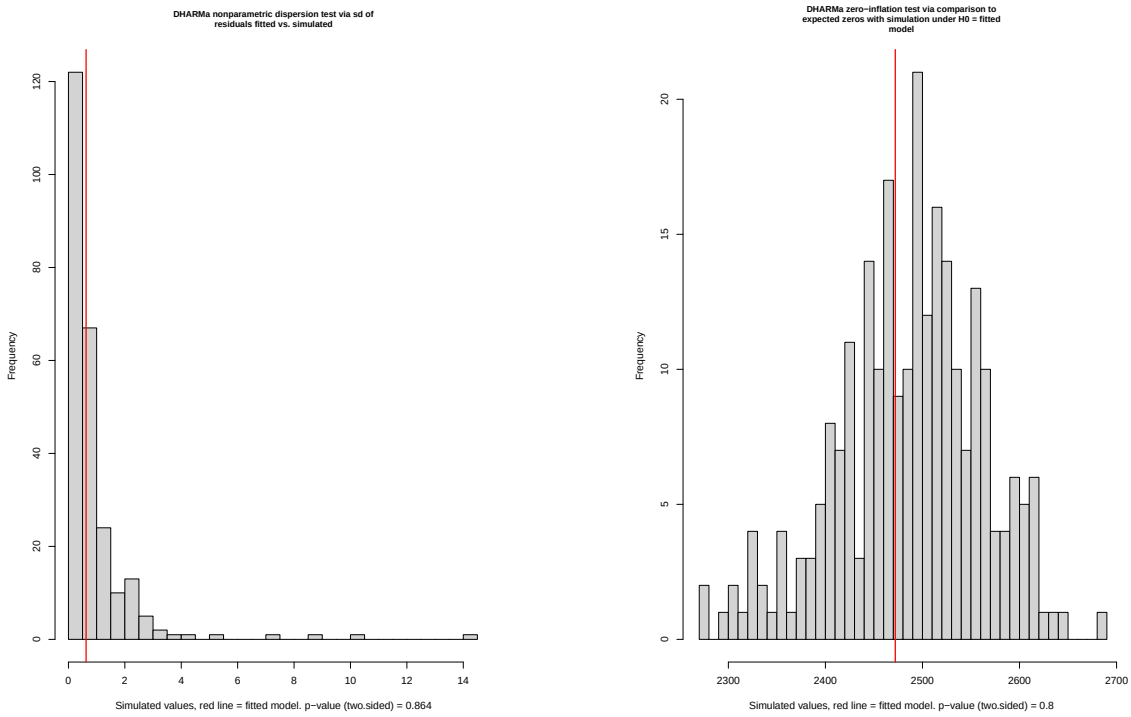
Since both p-values ($p - value = 0.864$ and $p - value = 0.8$ for dispersion and zero-inflation, respectively) did not provide evidence to reject the null hypothesis, the model appeared to be able to handle the dispersion and the number of zeros.

The overlaid dashed red line from quantile regression in Figure 5.20 residual vs. predictor plot raises the idea of a non-linear connection between the predictors and the model, and displays several outliers (red stars in the plot).

The overlaid dashed red line from quantile regression in Figure 5.20 residual vs. predictor plot shows a resemblance to the solid red line (similar to the Poisson and to the Negative Binomial mixed models), suggesting a linear predictor might adjust well to our data. However, once again, the deviation observed should be approached carefully for the need of using additive models.

The same pattern of outliers is visible in the non-transformed fitted vs. residuals plot Figure 4, represented in Figure 5.21.

Once again, we display the 3D scatter plot (Figure 5.22), but this time, the probability functions from the four analysed models (Poisson (in blue), Negative Binomial (in green),



a) `DHARMa` plot for *testDispersion()* function

b) `DHARMa` plot for *testZeroInflation()* function

FIGURE 5.19: `DHARMa` plots for Zero-Inflated Poisson mixed model.

Generalised Poisson (in red), and Zero-Inflated Poisson (in black)) have been merged. The Negative Binomial and the Generalised Poisson mixed models seem to be the ones that can predict higher counts, while the Poisson and the Zero-Inflated Poisson mixed models seem to only cope to a maximum of 3 counts.

With predicted counts of up to 2 $metacercariae.cockle^{-1}$, the prediction performance of this model shown to be not very satisfactory (Table 5.8. Additionally, this model AIC proved to be poorer (**AIC(ZIP) =** 2585) but similar compared to earlier models (**AIC(NB) =** 2547 or **AIC(GP) =** 2581). Nonetheless, in terms of AIC, the Zero-Inflated Poisson mixed model seemed to improve concerning to the Poisson mixed model (**AIC(P) =**) 2765).

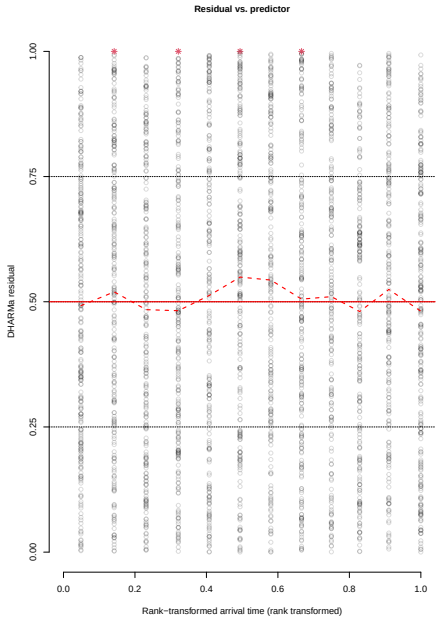


FIGURE 5.20: `DHARMa` residual vs. predictor plot for the Zero-Inflated Poisson Mixed model. Line in red means that statistically significant problems were detected.

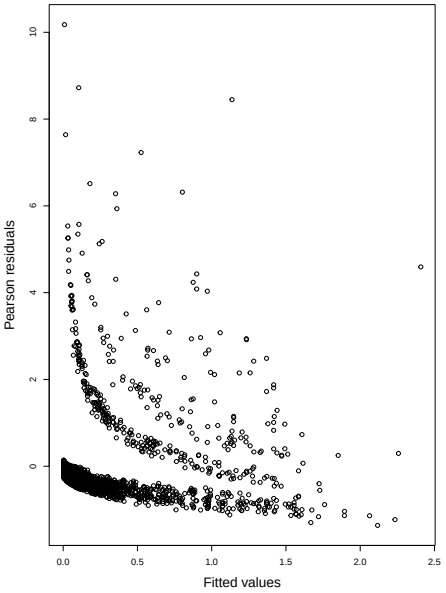


FIGURE 5.21: Residuals vs. fitted plot for the Zero-Inflated Poisson Mixed model.

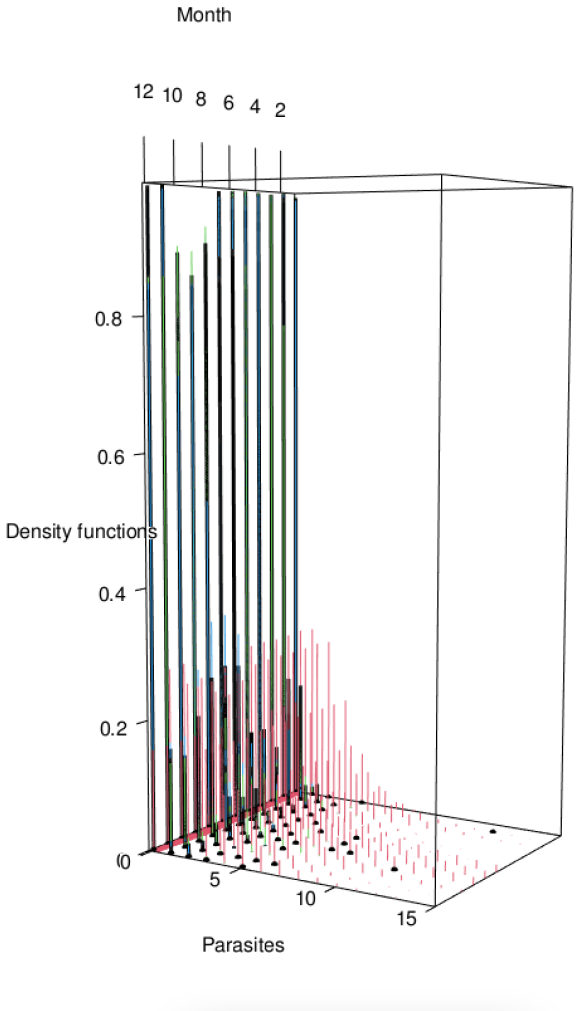


FIGURE 5.22: 3D scatter plot of the fitted models. The red line represents the fitted values, the black dots represent the observed values, and the different probability functions of the fitted models for each month are represented in black (Zero-Inflated Poisson mode), blue (Poisson model), green (Negative Binomial model), and red (Generalised Poisson).

TABLE 5.8: Confusion matrix of the obtained Zero-Inflated Poisson mixed model with the Predicted values in the columns and the Observed values in the rows.

		Predicted		
		0	1	2
Observed	0	2238	209	25
	1	144	91	3
	2	25	52	3
	3	8	21	2
	4	1	9	1
	5	1	10	0
	6	0	1	0
	7	0	4	0
	8	0	1	0
	11	0	0	1
	12	0	1	0

5.5 Hurdle Poisson model

The second model applied to deal with zero inflation was the Hurdle Poisson model. Equivalent to zero-inflated models, the Hurdle Poisson model also consists of two parts [8, 87]. One of the part processes zero counts, while the other component processes positive counts. A binary (*logit*) method is applied for the zero counts. In contrast to zero-inflated models, truncated Poisson distribution is used for processing the positive counts.

When truncated at 0, for $Y = y$, the zero-truncated Poisson distribution has probability function

$$\begin{aligned} P(Y = y|\mu) &= \frac{\mu^y \times e^{-mu}}{y!} \times \frac{1}{1-P(Y=0|\mu} \\ &= \frac{\mu^y \times e^{-\mu}}{(1-e^{-\mu}) \times y!} \end{aligned}$$

and is denoted as $Y \sim ZTP(\mu)$.

Thus, for a random variable $Y \sim HP(\mu, \pi)$, the probability function is

$$\begin{aligned} P(Y = y|\mu, \pi) &= \begin{cases} (1-\pi), & \text{if } y = 0 \\ \pi \times f_{truncatedPoisson}, & \text{if } y \geq 1 \end{cases} \\ &= \begin{cases} (1-\pi), & \text{if } y = 0 \\ \pi \times \frac{\mu^y \times e^{-\mu}}{(1-e^{-\mu}) \times y!} & \text{if } y \geq 1 \end{cases} \end{aligned}$$

and the Expected values (E) and variance (Var) are given as

$$\begin{aligned} E(Y) &= \pi \frac{\mu}{1-e^{-mu}} \\ Var(Y) &= \frac{\pi}{1-e^{-\mu}} \times (\mu + \mu^2) - \left(\frac{\pi \times \mu}{1-e^{-\mu}}\right)^2 \end{aligned}$$

The obtained model was

$$\begin{aligned} \log(\mu_{ijt}) &= \beta_0 + b_{0i} + \beta_1 DO_{jt} + \beta_2 SL_{jt} \\ logit(\pi_{ijt}) &= \gamma_0 + a_{0i} + \gamma_1 SL_{jt} + \gamma_2 Sal_{jt} + \gamma_3 pH_{jt} \end{aligned} \tag{5.6}$$

In Table 5.9, it is possible to see the output results for the Hurdle Poisson mixed model. The probability of a cockle to be infected by metacercariae is positively correlated with cockle's shell length and water pH, while salinity showed to influence this probability negatively. On the other hand, the typical cockle was positively correlated with cockle's shell length and dissolved oxygen.

The computed AIC for this model was 2608.9, which was lower than the Poisson mixed model but higher than the Generalised Poisson mixed model and the Zero-Inflated

TABLE 5.9: Output of the Hurdle Poisson mixed model.

Conditional model:			
Random effects			
	Variance	Std. Deviation	
Intercept (Site)	0.7144	0.8452	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-0.98402	0.3020	0.001
SL	0.5261	0.1369	<0.001
DO	0.2960	0.0661	<0.001
Conditional model AIC: 874.6			
Zero-inflation model:			
Random effects			
	Variance	Std. Deviation	
Intercept (Site)	1.6290	1.2760	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-2.8855	0.3353	<0.001
SL	1.1755	0.0890	<0.001
Sal	-0.2911	0.0687	<0.001
pH	0.1339	0.0635	0.035
Binomial model AIC: 1734.3			
AIC: 2608.9			

Poisson mixed model. For that reason (i.e., model did not improve in comparison to earlier models), it was decided not to continue with this model, hence model validation was not performed.

5.6 Binomial model

The preceding models' simulations and confusion matrices, in particular, shown that most of the models were not capable to model counts higher than 4 *metacercariae.cockle*$^{-1}$. In the particular case of the Poisson mixed model, the mean (μ) obtained for our model ranged from $\mu = 0.003$ and $\mu = 2.77$ (Figure 5.23).

For a Poisson distributions with mean 0.003 ($P(0.003)$), 99.7 % are distributed on 0, while the remaining 0.3 % are 1's. On the other hand, for a $P(2.8)$, only 23 % are counts of 0 or 1, while 77 % are counts higher than 1. Nonetheless, the fitted values median was around $\mu = 0.3$. For a $P(0.3)$, around 74 % of the values are 0s, 22 % are 1s, and roughly 4 % are values higher than 1. Not that 4 % (or higher for other fitted values) is a negligible percentage. However, this distinction is not entirely relevant to the effect of the parasite on the individual from a biological standpoint. Therefore, we decided to transform the

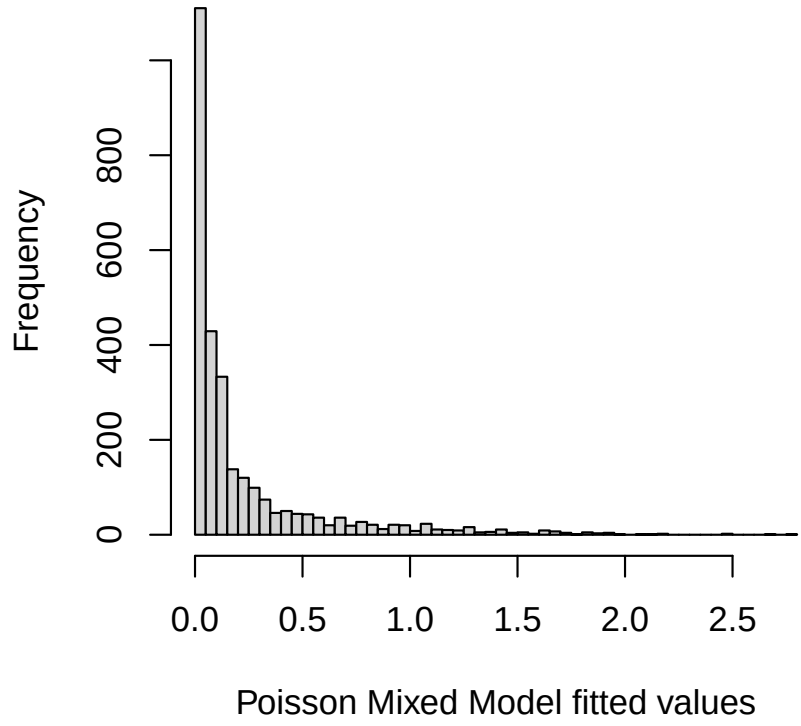


FIGURE 5.23: Histogram of the fitted values for the Poisson Mixed model 5.1.

variable into a dichotomous variable (infected cockles - 1 and non-infected cockles - 0) and apply a binomial model.

The obtained Binomial model was

$$logit(\mu_{ijt}) = \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 pH_{jt} + \beta_3 SL_{jt} \tag{5.7}$$

The calculated AIC for this model was of 1734.3. The probability of a cockle being infected showed to be positively correlated with the pH of the water and with cockle's shell length. On the other hand, salinity appears to negatively influence the probability of cockles to be infected. The model results output are represented in Table 5.10.

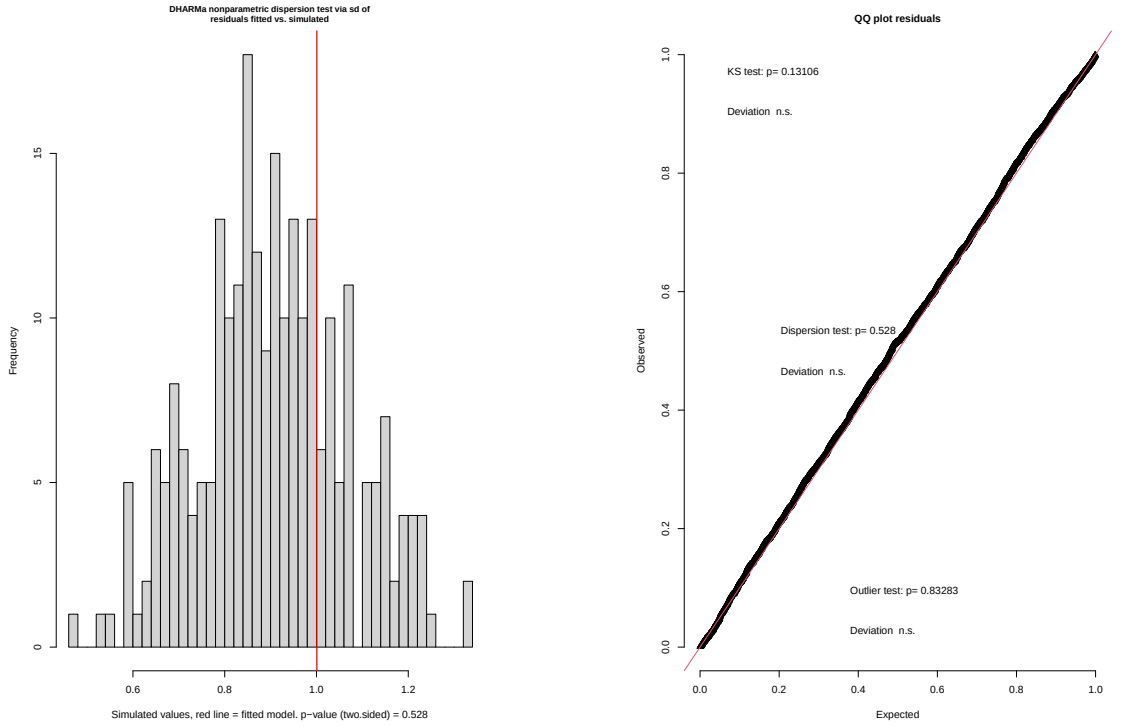
5.6.1 Model Validation

The plot in Figure 5.24 demonstrates that the Binomial model do not raises dispersion issues ($p - value = 0.528$; there is no evidence to reject the null hypothesis). In the same

TABLE 5.10: Output of the Binomial mixed model.

Random effects			
	Variance	Std. Deviation	
Intercept (Site)	1.6290	1.2760	
Fixed effects			
	Coef.	Std. Error	p-value
Intercept	-2.8855	0.3353	<0.001
SL	1.1755	0.0890	<0.001
Sal	-0.2911	0.0687	<0.001
pH	0.1339	0.0635	0.035
AIC: 1734.3			

Figure 5.24, it is also visible the quantile-quantile plot with the results for the Kolmogorov-Smirnov ($p - value = 0.131$) and dispersion tests, used to ensure that the transformed residuals follow a uniform distribution.



a) `DHARMa` plot for *testZeroInflation()* function.

b) `DHARMa` residuals quantile-quantile plot function.

FIGURE 5.24: `DHARMa` plots for the Binomial Mixed model

The Figure 5.25 (Pearson residuals vs. fitted probabilities)seem to show the existence of some outliers.

Nonetheless, the Binomial model again demonstrates the primary issue that has been raised in all models, a non-linear relationship with the *logit* predictor, in this case. This

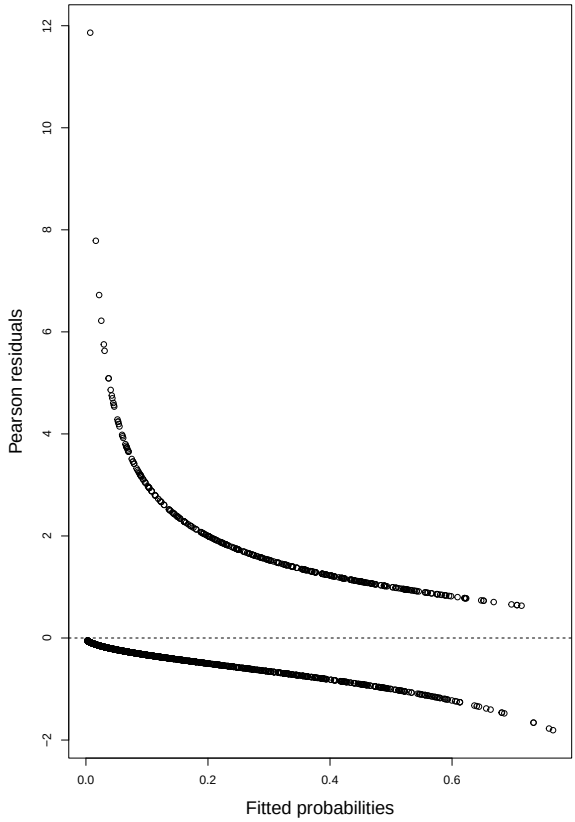


FIGURE 5.25: Residuals vs. fitted plot for the Binomial Mixed model.

may indicate need to fit splines into the model using additive generalised models. This plot of the non-scaled residuals vs. predicted values is shown in Figure 5.26

Finally, the model's performance was assessed using the ROC curve (Figure 5.27). The AUC for the model was 0.8142, indicating that the model accurately predicts values 81.4 % of the time. Given this high-performance value obtained for the model, this should give us a high level of confidence on how the model is predicting infection. However, since the majority of the values in our dataset are zeros, this AUC value is rather deceptive, as the model is practically only modelling zeros (only 15.6 % of 1's were correctly modelled). This information can be seen in the confusion matrix in Table 5.11.

As the 1's are being poorly modelled, one possibility would be to use logistic regression for imbalanced data, which is not the target of this project's study.

In general, compared to discrete outcomes, binary outcomes analysis is easier, and the results are more straightforward to understand. However, information about the abundance frequency is lost when data are categorised into binomial outcomes, and, additionally, reduces also the ability to identify small peaks of abundance that can be important.

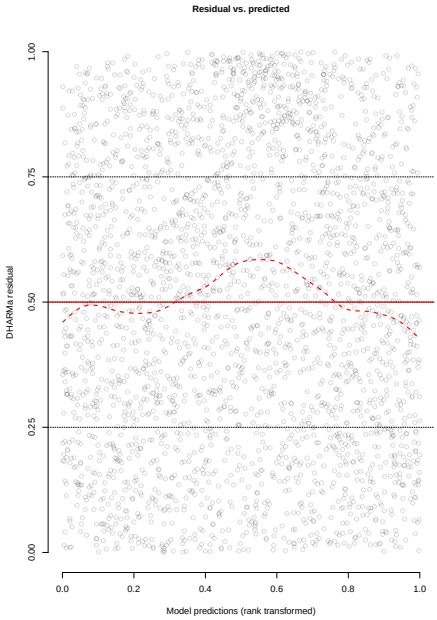


FIGURE 5.26: `DHARMa` residual vs. predictor plot for the Binomial Mixed model. Line in red means that statistically significant problems were detected.

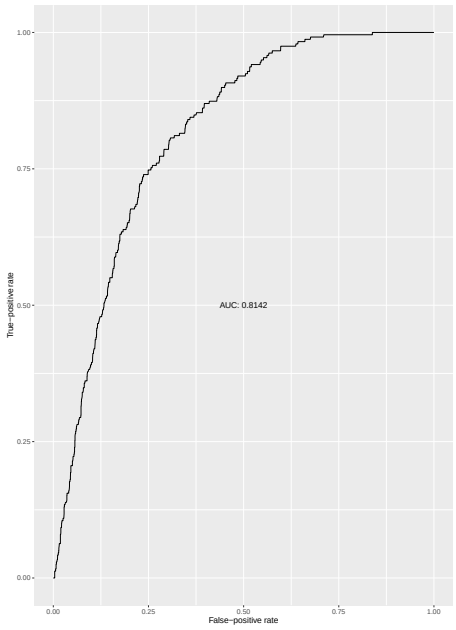


FIGURE 5.27: ROC curve for prediction of Binomial Mixed model false-positive rate against its true-positive rate.

TABLE 5.11: Confusion matrix of the obtained Binomial mixed model with the Predicted values in the columns and the Observed values in the rows.

		Predicted	
		0	1
Observed	0	2409	63
	1	319	59

5.7 Generalised Additive Mixed Models

In the following section, we will apply Generalised Additive Mixed Models (GAMMs) on our dataset. The main concern that all the Generalised Linear Mixed Models (GLMMs) applied to our dataset revealed was the possibility of a non-linear relationship between the predictors and the response variable. Generalised Additive (Mixed) Models are a model family that allows the extension of GLMMs by enabling the addition of a smoothing function (or spline) to the explanatory variables [88, 89].

As a result, if a random variable Y with parameters μ, mean, and ϕ, dispersion parameter, that follows one of the distributions allowed by GLMMs, GAMMs display a general form of

$$g(Y = y|\mu, \phi) = \beta_0 + b_0 + \beta_1 Y_1 + \ldots + \beta_q X_q + f(X_{q+1}) + \ldots + f(X_p)$$

where, $X_1, \ldots, X_p$ are the explanatory variables, and $f(\cdot)$ is the smoothing function.

Therefore, the non-linear relationship between an explanatory variable and the response variable is represented by splines. A spline uses basis functions (a set of complex polynomial functions) that are used to approximate complex curves. More specifically, a spline is a sum of basis functions (pieces) weighted by coefficients and connected by jointing points (knots), where each component is selected to have the least mean square curvature possible. A simple basis can be represented by a polynomial of order 4. $f(\cdot)$ is said to be a polynomial of 4^{th} order if f contains

$$f(x) = \beta_1 + x\beta_2 + x^2\beta_3 + x^3\beta_4 + x^4\beta_5 + \epsilon$$

where $\beta_1(x) = 1$, $\beta_2(x) = x$, $\beta_3(x) = x^2$, $\beta_4(x) = x^3$, and $\beta_5(x) = x^4$.

There are several types of splines, including B-splines, cubic splines, and thin-plate splines. One of the most commonly used is the cubic spline.

Cubic splines are piecewise polynomial functions that are used to smooth and continuous approximation and interpolation of data points. These splines are defined by a set of cubic polynomials, each of which is applied to a different interval between data points. Suppose that for a set of points $(x_0, y_0), \ldots, (x_n, y_n)$. Within each interval $[x_i, x_{i+1}]$, the cubic spline can be represented as $S(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$ where x_i and x_{i+1} are the boundary points of the defining spline interval, and a_i, b_i, c_i, and d_i are the coefficients specific to the interval $[x_i, x_{i+1}]$.

Finally, to prevent overfitting, a penalty term is frequently introduced to the likelihood function in GAMMs. This penalty prevents the spline from fitting the noise in the data, resulting in smoother curves. A smoothing parameter or penalty parameter controls the degree of penalization.

5.7.1 Poisson model

Once again, we started the analysis by fitting the data to a Poisson distribution. The adopted model for this distribution was

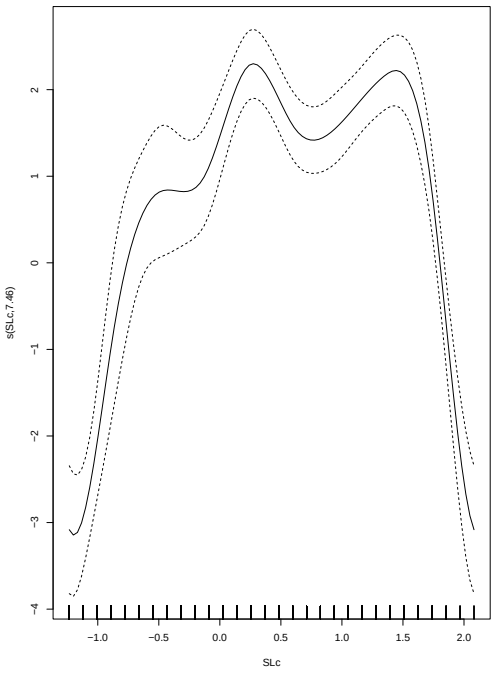
$$\log \mu_{ijt} = \beta_0 + b_{0i} + \beta_1 Sal_{jt} + \beta_2 DO_{jt} + f(SL_{jt}) + f(pH_{jt}) + f(TOM_{jt}) + f(Month_{jt}) \tag{5.8}$$

with and AIC of 2901.4, and an explained deviance of 32.7 %. The output results for the model are represented in Table 5.12. For a typical cockle, metacercaria abundance per cockle was positively correlated with salinity and dissolved oxygen in the water column.

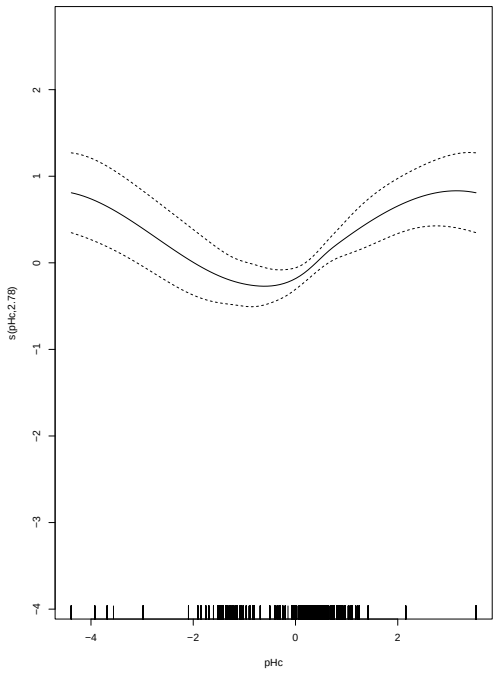
TABLE 5.12: Output of the Poisson additive mixed model.

Smooth terms			
	EDF	p-value	<0.001
SL	7.455	<0.001	
pH	2.784	<0.001	
TOM	6.424	<0.001	
Month	7.210	<0.001	
Site (re)	0.959	<0.001	
Parametric terms			
	Coef.	Std. Error	p-value
Intercept	-3.3755	0.1974	<0.001
SL	0.1807	0.6519	0.006
DO	0.4465	0.1047	<0.001
AIC: 2901.4		**Deviance explained:** 32.7 %	

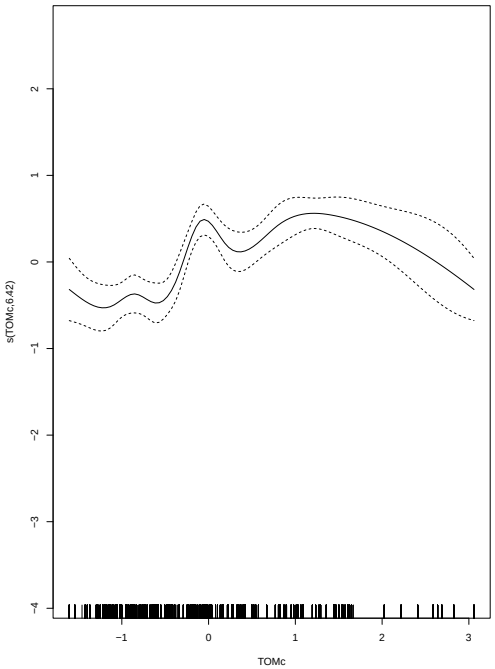
Note that for smooths the coefficients are not printed, since each smooth has several coefficients. Instead, it has represented the effective degrees of freedom (EDF). The EDF represents the complexity of the smooth, with a EDF of 1 representing a straight line, 2 a quadratic curve, until higher EDF that represent more complex smooths (wigglier curves). The EDF values ranged from 0.959 to 7.455. The splines behaviour for each variable are visible in Figure 5.28.



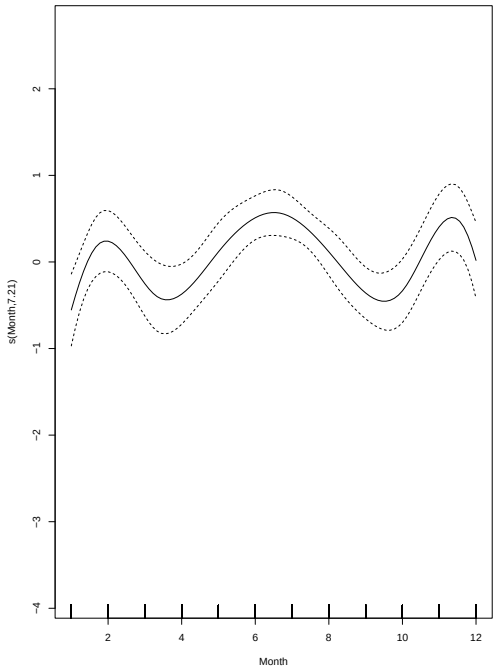
a) Cockle's Shell length



b) pH



c) Organic Matter content



d) Month

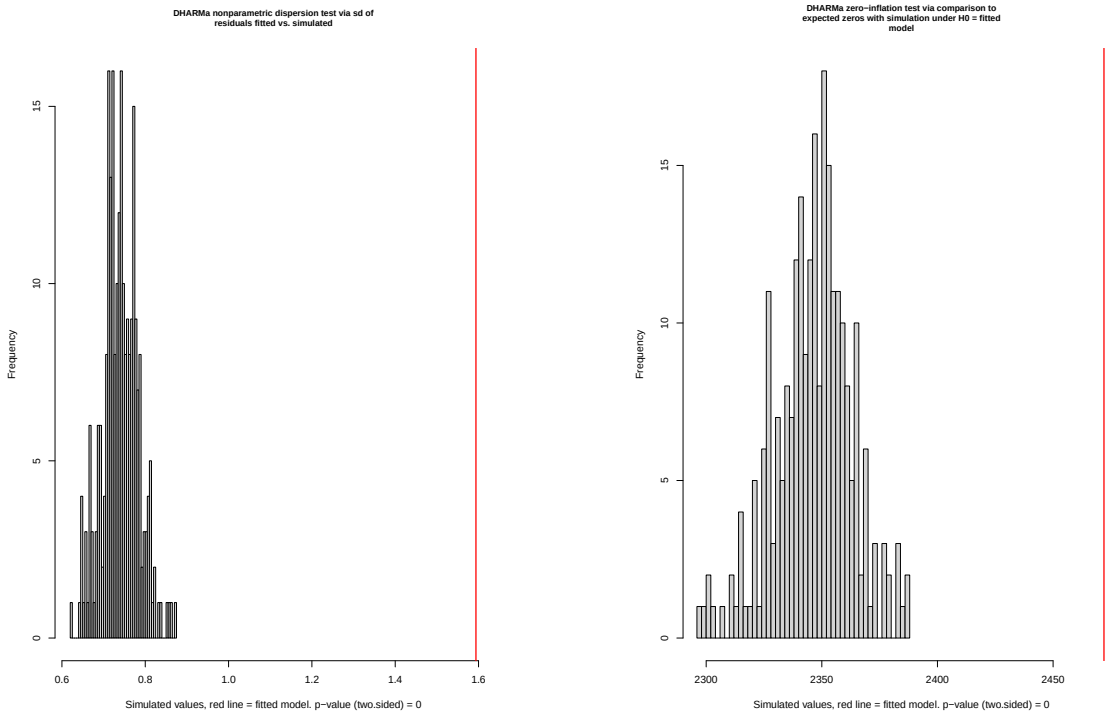
FIGURE 5.28: Line plot for the variables that a spline was applied to observe the trend.

5.7.1.1 Model Validation

We begin the validation of the model by checking its dispersion and capacity to cope with the quantity of zeros. These analysis were made using the *testDispersion* and *testZeroInflation* functions from the `DHARMa` package. The obtained plots with the respective test p-values are visible in Figure 5.29.

The obtained p-value for the dispersion statistic (p-value < 0.001), showed evidence to reject the null hypothesis, that the variance of the observed residuals are equal to the variance of the simulated residuals. Therefore, this model shows problems of overdispersion.

Similarly, the p-value < 0.001 obtained for the zero inflation also shows that the model is not able to cope with the amount of observed zeros.



a) `DHARMa` plot for *testDispersion()* function

b) `DHARMa` plot for *testZeroInflation()* function.

FIGURE 5.29: `DHARMa` plots for Poisson Additive Mixed model.

We were able to determine that this was not an appropriate model to apply to the data in this manner.

5.7.2 Negative Binomial model

Although we observed zero inflation difficulties in the prior model 5.8, these may be connected to the previously discovered overdispersion problems. As a result, we will start by applying a Negative Binomial additive model. If the issue persists, then we will use a zero-inflated model.

The obtained model for the Negative Binomial distribution was

$$\log \mu_{ijt} = \beta_0 + b_{0i} + + f(SL_{jt}) + f(DO_{jt}) + f(pH_{jt}) + f(Month_{jt}) + f(TOM_{jt}) + f(Month_{jt}) \tag{5.9}$$

The output results for the model are represented in Table 5.13. The AIC of the model was 2650.5 (Table 5.13), which was a significant improvement in comparison to the Poisson additive mixed model. The Negative Binomial parameter k was of 0.422, and the model had 36 % of deviance explained. No parametric terms showed significance for the model, while cockle's shell length, the organic matter content of the sediment, water dissolved oxygen and pH and month were significant smooth terms. There EDM ranged from 2.22 to 7.27 (Table 5.13), and the spline trends are represented in Figure 5.30.

TABLE 5.13: Output of the Negative Binomial additive mixed model.

Smooth terms			
	EDF	p-value	<0.001
SL	7.267	<0.001	
DO	2.403	0.008	
pH	2.6570	<0.001	
TOM	2.909	<0.001	
Month	2.223	<0.020	
Site (re)	0.842	<0.011	
Parametric terms			
	Coef.	Std. Error	p-value
Intercept	-3.3755	0.1974	<0.001
AIC: 2901.4		**Deviance explained:** 32.7 %	
Dispersion parameter k**:** 0.422			

5.7.2.1 Model Validation

The model's ability to handle a large number of zeros was further evaluated using `DHARMa`'s *ZeroInflation()* function. The p-value of 1 found indicates that the model can explain for the observed number of zeros. Figure 5.31 depicts the histogram and test results.

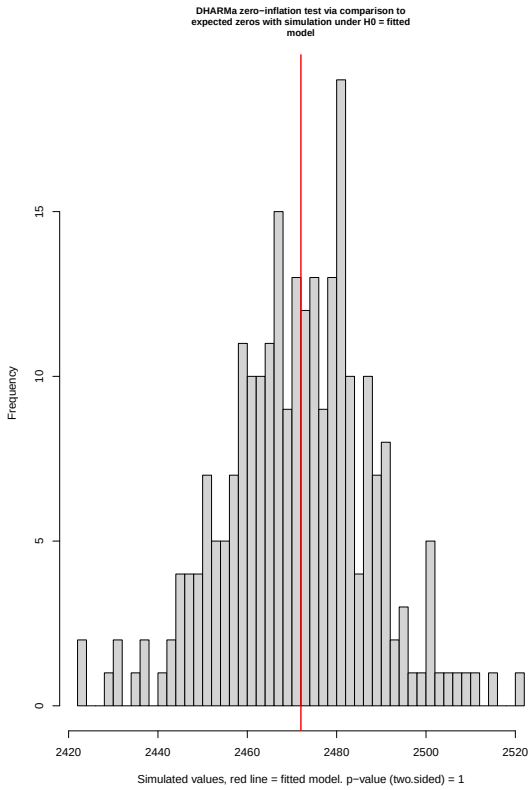
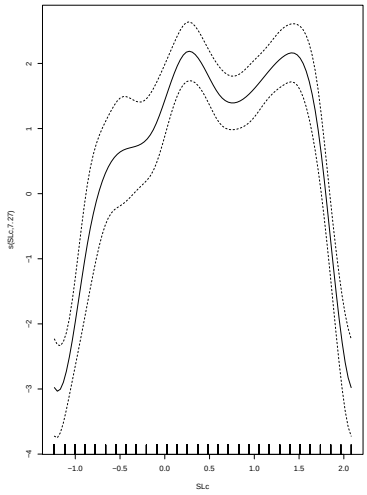


FIGURE 5.31: `DHARMa` plot for the *testZeroInflation()* function for the Negative Binomial Additive Mixed model.

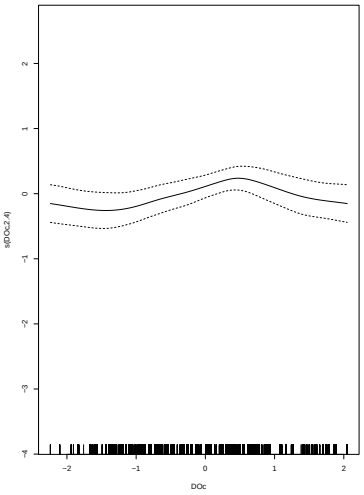
Despite that fact that the model did not raise any problems related with the amount of zeros or dispersion, the model's prediction remains rather inadequate (91.1 % of zeros were successfully modelled, while only 22.8 % of 1 counts and 1.3 % of 2 counts were adequately modelled; counts greater than 2 were not precisely modelled), as shown by the confusion matrix in Table 5.14.

TABLE 5.14: Confusion matrix of the obtained Negative Binomial mixed model with the Predicted values in the columns and the Observed values in the rows.

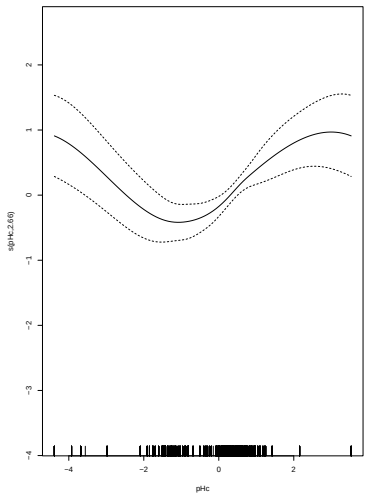
		Predicted Values		
		0	1	2
Observed	0	2253	219	0
	1	184	53	1
	2	49	30	1
	3	10	21	0
	4	4	6	0
	5	2	9	0
	6	0	1	0
	7	1	1	2
	8	1	0	0
	11	1	0	0
	12	1	0	0



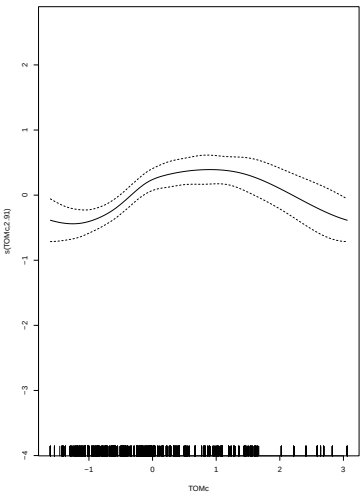
a) Cockle's Shell length



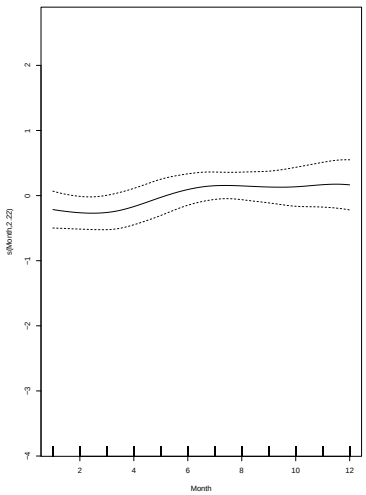
b) Dissolved Oxygen



c) pH



d) Organic Matter content



e) Month

FIGURE 5.30: Line plot for the variables that a spline was applied to observe the trend.

Chapter 6

Final Remarks and Future Perspectives

The purpose of this work was meticulously centred on the examination of regression models for count data, particularly data exhibiting an excessive number of zeros, and its application to real scenarios data, with a specific focus on the ecological study of parasites infecting the European cockle, *Cerastoderma edule*. More specifically, this project encompassed a theoretical component aimed at scrutinising and comparing the various `R` packages available for count data regression analysis (comprising both Generalised Linear Mixed Models and Generalised Additive Mixed Models). Additionally, it involved also a more practical component, which revolved around data collected during the *COACH project*. In addition to the application of these models, this project also aimed to identify which environmental variables have a determining impact on the abundance of parasite infecting cockles across different cockle beds in the Ria de Aveiro, Portugal.

To accomplish this, cockles were collected monthly from 18 sampling sites in the Ria de Aveiro over the course of a year. For each analysed cockle, the species of parasites infecting each individual and their abundance were recorded. Simultaneously, the variables collected to explain infection were cockle's shell length, organic matter content of the cockles' bed sediment, and the temperature, pH, dissolved oxygen, salinity and redox potential of the water at the time of sampling.

A substantial number of cockles were found to not be infected by any metacercariae. Consequently, it was expected that models capable of handling zero inflation, such as zero-inflated or hurdle models, would be necessary to apply. Nevertheless, the initial approach involved the application of a Poisson mixed model. The obtained model identified

salinity, dissolved oxygen, pH, and cockle's shell length as significant explanatory variables. When tested for the Poisson regression assumptions, specifically the assumption of equidispersion, the model did not any significant violations. This outcome underscores the importance of conducting regression analysis step by step, rather than making decisions solely based on the appearance of the data.

Although the Poisson mixed model did not present major issues overall, this project, given its strong didactic component focused on learning to apply several different models, proceeded to apply five other Generalised Linear Mixed Models, including the Negative Binomial, Generalised Poisson, Zero-Inflated Poisson, Hurdle Poisson, and Binomial.

In general, the models displayed similar results. All models identified cockle shell length, salinity, and pH as significant effects, with the Poisson model and the two models for zero-inflation (Zero-Inflated Poisson and Hurdle Poisson) also selecting dissolved oxygen as a significant explanatory variable. Furthermore, none of the models exhibited substantial violations except for a slight non-linear relationship between the explanatory variables and the response mean. Nonetheless, the Negative Binomial regression model displayed the best goodness of fit (lowest AIC). This highlights the need to test the remaining models (especially if not totally satisfied with the obtained model) to determine if they can somehow enhance the fit, even if no problems were detected during previous model fitting. In any case, the models generated were unable to predict counts close to those obtained (maximum 4 - Negative Binomial model - against 12 observed). These results should therefore be analysed with caution.

Regarding the Binomial model, it exhibited commendable performance, primarily due to the high number of zeros encountered during cockle analyses, as the model predominantly modelled zeros. Nevertheless, it would be intriguing to apply specific logistic regression models for imbalanced classes to observe how they would respond.

Finally, in an effort to address the issue of the non-linear relationship between explanatory variables and the response mean, Generalised Additive mixed models were also applied. The employed Poisson model exhibited violations of the equidispersion assumption and of its ability to handle zero inflation. When the Negative Binomial regression was applied, these issues appeared to be resolved, but the model's predictions remained relatively poor, offering no improvement over the previous models.

In conclusion, one of the primary limitations of this study pertained to the sampling process, specifically the representativeness of the collection in terms of cockle size at the

site. This led to the analysis of relatively small (young) cockles that had not yet been exposed to parasites. As future work, it would be intriguing to further explore this dataset. This could involve the application of untested distributions to potentially improve data fit, as well as the examination of other groups of parasites available in the database. Notably, a keen interest lies in the analysis of the dependent variable *Gymnophallus*, which, apart from its high number of zeros, exhibits parasite counts reaching into the hundreds per cockle. Applying a Tweedie distribution could be a potential solution in this case.

Bibliography

[1] C. S. Elphick, "How you count counts: the importance of methods research in applied ecology," pp. 1313–1320, 2008. [Cited on page 1.]

[2] M. C. MacSwiney G, F. M. Clarke, and P. A. Racey, "What you see is not what you get: the role of ultrasonic detectors in increasing inventory completeness in neotropical bat assemblages," *Journal of applied Ecology*, vol. 45, no. 5, pp. 1364–1371, 2008. [Cited on page 1.]

[3] M. W. Alldredge, K. Pacifici, T. R. Simons, and K. H. Pollock, "A novel field evaluation of the effectiveness of distance and independent observer sampling to estimate aural avian detection probabilities," *Journal of Applied Ecology*, vol. 45, no. 5, pp. 1349–1356, 2008. [Cited on page 1.]

[4] J. Greenwood, R. Robinson, and W. Sutherland, "Ecological census techniques: a handbook," *Press Syndicate of the University of Cambridge*, 1996. [Cited on page 1.]

[5] J. M. Hilbe, *Modeling Count Data*. Cambridge University Press, 2014. [Cited on pages 1 and 2.]

[6] A. Zeileis, C. Kleiber, and S. Jackman, "Regression models for count data in r," *Journal of statistical software*, vol. 27, no. 8, pp. 1–25, 2008. [Cited on page 1.]

[7] W. G. Cochran, "Some consequences when the assumptions for the analysis of variance are not satisfied," *Biometrics*, vol. 3, no. 1, pp. 22–38, 1947. [Online]. Available: http://www.jstor.org/stable/3001535 [Cited on page 1.]

[8] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, 2nd ed., ser. Econometric Society Monographs. Cambridge University Press, 2013. [Cited on pages 2, 7, 8, 74, and 79.]

[9] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed., ser. CRC Monographs on Statistics and Applied Probability. Chapman and Hall, 1989. [Cited on pages 2, 7, 15, and 16.]

[10] J. M. Hilbe, *Negative binomial regression*. Cambridge University Press, 2011. [Cited on pages 2 and 10.]

[11] P. C. Consul and F. Famoye, *Lagrangian probability distributions*. Springer, 2006. [Cited on pages 2 and 12.]

[12] P. C. Consul, *Generalized Poisson distributions: properties and applications*. M. Dekker, 1989. [Cited on page 2.]

[13] A. Huang, "Mean-parametrized conway–maxwell–poisson regression models for dispersed counts," *Statistical Modelling*, vol. 17, no. 6, pp. 359–380, 2017. [Cited on page 2.]

[14] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, G. M. Smith *et al.*, *Mixed effects models and extensions in ecology with R*. Springer, 2009, vol. 574. [Cited on pages 2 and 3.]

[15] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger, *Analysis of longitudinal data*. Oxford university press, 2002. [Cited on page 3.]

[16] K.-Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986. [Cited on page 3.]

[17] R. Sousa, J. L. Gutiérrez, and D. C. Aldridge, "Non-indigenous invasive bivalves as ecosystem engineers," *Biological invasions*, vol. 11, pp. 2367–2385, 2009. [Cited on page 3.]

[18] E. Morgan, R. M. O'Riordan, and S. C. Culloty, "Climate change impacts on potential recruitment in an ecosystem engineer," *Ecology and Evolution*, vol. 3, no. 3, pp. 581–594, 2013. [Cited on pages 3 and 31.]

[19] C. J. Philippart, J. J. Beukema, G. C. Cadée, R. Dekker, P. W. Goedhart, J. M. van Iperen, M. F. Leopold, and P. M. Herman, "Impacts of nutrient reduction on coastal communities," *Ecosystems*, vol. 10, pp. 96–119, 2007. [Cited on page 3.]

[20] C. Rakotomalala, K. Grangeré, M. Ubertini, M. Forêt, and F. Orvain, "Modelling the effect of cerastoderma edule bioturbation on microphytobenthos resuspension towards the planktonic food web of estuarine ecosystem," *Ecological Modelling*, vol. 316, pp. 155–167, 2015. [Cited on pages 3 and 31.]

[21] J. Wijsman, K. Troost, J. Fang, and A. Roncarati, "Global production of marine bivalves. trends and challenges," *Goods and services of marine bivalves*, pp. 7–26, 2019. [Cited on page 3.]

[22] J. Oliveira, F. Castilho, Â. Cunha, and M. J. Pereira, "Bivalve harvesting and production in portugal: An overview," *Journal of Shellfish Research*, vol. 32, no. 3, pp. 911–924, 2013. [Cited on pages 3 and 31.]

[23] N. Tebble, "British bivalve seashells: a handbook for identification," *(London, UK: Trustees of the British Museum (Natural History))*, 1966. [Cited on pages 3 and 31.]

[24] L. Dabouineau and A. Ponsero, "Synthesis on biology of common european cockle cerastoderma edule." 2009. [Cited on page 3.]

[25] P. J. Honkoop, E. M. Berghuis, S. Holthuijsen, M. S. Lavaleye, and T. Piersma, "Molluscan assemblages of seagrass-covered and bare intertidal flats on the banc d'arguin, mauritania, in relation to characteristics of sediment and organic matter," *Journal of Sea Research*, vol. 60, no. 4, pp. 255–263, 2008. [Cited on pages 3 and 31.]

[26] A. Ponsero, L. Dabouineau, and J. Allain, "Modelling of common european cockle cerastoderma edule fishing grounds aimed at sustainable management of traditional harvesting," *Fisheries Science*, vol. 75, no. 4, pp. 839–850, 2009. [Cited on page 4.]

[27] J. Beukema and R. Dekker, "Annual cockle cerastoderma edule production in the wadden sea usually fails to sustain both wintering birds and a commercial fishery," *Marine Ecology Progress Series*, vol. 309, pp. 189–204, 2006. [Cited on page 4.]

[28] F. FishStatJ, "Fisheries and aquaculture software," *FishStatJ-Software for Fishery Statistical Time Series. Rome: FAO Fisheries and Aquaculture Department*, 2015. [Cited on page 4.]

[29] D. N. Carss, A. C. Brito, P. Chainho, A. Ciutat, X. de Montaudouin, R. M. F. Otero, M. I. Filgueira, A. Garbutt, M. A. Goedknegt, S. A. Lynch *et al.*, "Ecosystem services

provided by a non-cultured shellfish species: The common cockle cerastoderma edule," *Marine Environmental Research*, vol. 158, p. 104931, 2020. [Cited on page 4.]

[30] J. D. Taylor, E. A. Glover, E. M. Harper, J. A. Crame, C. Ikebe, and S. T. Williams, "Left in the cold? evolutionary origin of laternula elliptica, a keystone bivalve species of antarctic benthos," *Biological Journal of the Linnean Society*, vol. 123, no. 2, pp. 360–376, 2018. [Cited on page 4.]

[31] I. M. McLeod, P. S. zu Ermgassen, C. L. Gillies, B. Hancock, and A. Humphries, "Can bivalve habitat restoration improve degraded estuaries?" in *Coasts and Estuaries*. Elsevier, 2019, pp. 427–442. [Cited on page 4.]

[32] A. van der Schatte Olivier, L. Jones, L. L. Vay, M. Christie, J. Wilson, and S. K. Malham, "A global review of the ecosystem services provided by bivalve aquaculture," *Reviews in Aquaculture*, vol. 12, no. 1, pp. 3–25, 2020. [Cited on page 4.]

[33] D. Burdon, R. Callaway, M. Elliott, T. Smith, and A. Wither, "Mass mortalities in bivalve populations: A review of the edible cockle cerastoderma edule (l.)," *Estuarine, Coastal and Shelf Science*, vol. 150, pp. 271–280, 2014. [Cited on page 4.]

[34] D. W. Thieltges, "Parasite induced summer mortality in the cockle cerastoderma edule by the trematode gymnophallus choledochus," *Hydrobiologia*, vol. 559, pp. 455–461, 2006. [Cited on page 4.]

[35] X. de Montaudouin, I. Arzul, A. Cao, M. J. Carballal, B. Chollet, S. Correia, J. Cuesta, S. Culloty, G. Daffe, S. Darriba *et al.*, *Catalogue of parasites and diseases of the common cockle Cerastoderma edule*. UA Editora-Universidade de Aveiro, 2021. [Cited on pages 4 and 33.]

[36] M. Longshaw and S. K. Malham, "A review of the infectious agents, parasites, pathogens and commensals of european cockles (cerastoderma edule and c. glaucum)," *Journal of the Marine Biological Association of the United Kingdom*, vol. 93, no. 1, pp. 227–247, 2013. [Cited on page 4.]

[37] R. D. Adlard, T. L. Miller, and N. J. Smit, "The butterfly effect: parasite diversity, environment, and emerging disease in aquatic wildlife," *Trends in Parasitology*, vol. 31, no. 4, pp. 160–166, 2015. [Cited on page 4.]

[38] C. A. Burge, A. Shore-Maggio, and N. D. Rivlin, "Ecology of emerging infectious diseases of invertebrates," *Ecology of Invertebrate Diseases*, pp. 587–625, 2017. [Cited on page 4.]

[39] C. D. Harvell, K. Kim, J. Burkholder, R. R. Colwell, P. R. Epstein, D. J. Grimes, E. E. Hofmann, E. Lipp, A. Osterhaus, R. M. Overstreet *et al.*, "Emerging marine diseases–climate links and anthropogenic factors," *Science*, vol. 285, no. 5433, pp. 1505–1510, 1999. [Cited on page 4.]

[40] F. Lei and R. Poulin, "Effects of salinity on multiplication and transmission of an intertidal trematode parasite," *Marine Biology*, vol. 158, pp. 995–1003, 2011. [Cited on page 5.]

[41] J. Koprivnikar and R. Poulin, "Effects of temperature, salinity, and water level on the emergence of marine cercariae," *Parasitology Research*, vol. 105, pp. 957–965, 2009. [Cited on page 5.]

[42] S. Correia, L. Magalhães, R. Freitas, H. Bazairi, M. Gam, and X. de Montaudouin, "Large scale patterns of trematode parasite communities infecting cerastoderma edule along the atlantic coast from portugal to morocco," *Estuarine, Coastal and Shelf Science*, vol. 233, p. 106546, 2020. [Cited on page 5.]

[43] L. Magalhães, S. Correia, X. de Montaudouin, and R. Freitas, "Spatio-temporal variation of trematode parasites community in cerastoderma edule cockles from ria de aveiro (portugal)," *Environmental Research*, vol. 164, pp. 114–123, 2018. [Cited on page 5.]

[44] J. K. Lindsey, "Generalized linear modelling," *Applying generalized linear models*, pp. 1–26, 1997. [Cited on page 7.]

[45] R. H. Myers and R. H. Myers, *Classical and modern regression with applications*. Duxbury press Belmont, CA, 1990, vol. 2. [Cited on page 7.]

[46] P. C. Consul and G. C. Jain, "A generalization of the poisson distribution," *Technometrics*, vol. 15, no. 4, pp. 791–799, 1973. [Cited on page 12.]

[47] P. Roback and J. Legler, "Beyond multiple linear regression," *Applied Generalized Linear Models and Multilevel Models in R*, p. 436, 2021. [Cited on page 14.]

[48] J. J. Faraway, *Extending the linear model with R: generalized linear, mixed effects and non-parametric regression models*. CRC press, 2016. [Cited on page 14.]

[49] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 135, no. 3, pp. 370–384, 1972. [Cited on page 15.]

[50] J. Fox, *Applied regression analysis and generalized linear models*. Sage Publications, 2015. [Cited on page 15.]

[51] P. K. Dunn and G. K. Smyth, *Generalized linear models with examples in R*. Springer, 2018, vol. 53. [Cited on pages 16 and 52.]

[52] P. Consul and F. Famoye, "Generalized poisson regression model," *Communications in Statistics-Theory and Methods*, vol. 21, no. 1, pp. 89–109, 1992. [Cited on page 16.]

[53] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/ [Cited on page 19.]

[54] J. Pinheiro, D. Bates, and R Core Team, *nlme: Linear and Nonlinear Mixed Effects Models*, 2023, r package version 3.1-163. [Online]. Available: https://CRAN.R-project.org/package=nlme [Cited on pages 19 and 22.]

[55] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. New York: Springer, 2000. [Cited on pages 19 and 22.]

[56] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015. [Cited on pages 19 and 22.]

[57] D. Rizopoulos, *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature*, 2023, https://drizopoulos.github.io/GLMMadaptive/, https://github.com/drizopoulos/GLMMadaptive. [Cited on pages 19 and 22.]

[58] G. Broström, *glmmML: Generalized Linear Models with Clustering*, 2023, r package version 1.1.5. [Online]. Available: https://CRAN.R-project.org/package=glmmML [Cited on page 19.]

[59] D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, and J. Sibert, "Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models," *Optimization Methods and Software*, vol. 27, no. 2, pp. 233–249, 2012. [Cited on pages 19 and 22.]

[60] M. E. Brooks, K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *The R Journal*, vol. 9, no. 2, pp. 378–400, 2017. [Cited on pages 19 and 22.]

[61] S. N. Wood, "Stable and efficient multiple smoothing parameter estimation for generalized additive models," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 673–686, 2004. [Cited on pages 19 and 22.]

[62] S. Wood and F. Scheipl, *gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4'*, 2020, r package version 0.2-6. [Online]. Available: https://CRAN.R-project.org/package=gamm4 [Cited on pages 19 and 22.]

[63] T. W. Yee, *Vector generalized linear and additive models: with an implementation in R*. Springer, 2015, vol. 10. [Cited on pages 19 and 22.]

[64] R. A. Rigby and D. M. Stasinopoulos, "Generalized additive models for location, scale and shape,(with discussion)," *Applied Statistics*, vol. 54, pp. 507–554, 2005. [Cited on pages 19 and 22.]

[65] J. D. Hadfield, "Mcmc methods for multi-response generalized linear mixed models: The MCMCglmm R package," *Journal of Statistical Software*, vol. 33, no. 2, pp. 1–22, 2010. [Online]. Available: https://www.jstatsoft.org/v33/i02/ [Cited on pages 20 and 22.]

[66] P.-C. Bürkner, "brms: An r package for bayesian multilevel models using stan," *Journal of statistical software*, vol. 80, pp. 1–28, 2017. [Cited on pages 20 and 22.]

[67] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 71, no. 2, pp. 319–392, 2009. [Cited on pages 20 and 22.]

[68] S. Yu-Sung and Y. Masanao, *R2jags: Using R to Run 'JAGS'*, 2021, r package version 0.7-1. [Online]. Available: https://CRAN.R-project.org/package=R2jags [Cited on page 20.]

[69] S. Jackman, "pscl: Classes and methods for r. developed in the political science computational laboratory, stanford university. department of political science, stanford university, stanford, ca. r package version 1.03. 5," *http://www. pscl. stanford. edu/*, 2010. [Cited on pages 20 and 22.]

[70] F. Hartig, *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*, 2022, r package version 0.4.6. [Online]. Available: https://CRAN.R-project.org/package=DHARMa [Cited on page 23.]

[71] J. M. Dias, J. Lopes, and I. Dekeyser, "Tidal propagation in ria de aveiro lagoon, portugal," *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, vol. 25, no. 4, pp. 369–374, 2000. [Cited on page 33.]

[72] A. Lillebø, O. Ameixa, L. Sousa, A. Sousa, J. Soares, M. Dolbeth, and F. Alves, "The physio-geographical background and ecology of ria de aveiro," *coastal lagoons in europe*, p. 21, 2015. [Cited on page 33.]

[73] N. Vaz, J. M. Dias, P. Leitao, and I. Martins, "Horizontal patterns of water temperature and salinity in an estuarine tidal channel: Ria de aveiro," *Ocean Dynamics*, vol. 55, pp. 416–429, 2005. [Cited on page 33.]

[74] X. De Montaudouin, D. W. Thieltges, M. Gam, M. Krakau, S. Pina, H. Bazairi, L. Dabouineau, F. Russell-Pinto, and K. T. Jensen, "Digenean trematode species in the cockle cerastoderma edule: identification key and distribution along the north-eastern atlantic shoreline," *Journal of the Marine Biological Association of the United Kingdom*, vol. 89, no. 3, pp. 543–556, 2009. [Cited on page 33.]

[75] G. Lauckner, "Diseases of mollusca: bivalvia," *Diseases of marine animals*, vol. 2, 1983. [Cited on page 33.]

[76] A. O. Bush, K. D. Lafferty, J. M. Lotz, and A. W. Shostak, "Parasitology meets ecology on its own terms: Margolis et al. revisited," *The Journal of parasitology*, pp. 575–583, 1997. [Cited on page 33.]

[77] V. Quintino, A. M. Rodrigues, and F. Gentil, "Assessment of macrozoobenthic communities in the lagoon of óbidos, western coast of portugal," *Scientia Marina (Barcelona)*, no. 2-3, 1989. [Cited on page 34.]

[78] E. Kristensen and F. Ø. Andersen, "Determination of organic carbon in marine sediments: a comparison of two chn-analyzer methods," *Journal of Experimental Marine Biology and Ecology*, vol. 109, no. 1, pp. 15–23, 1987. [Cited on page 34.]

[79] J. Fermer, S. Culloty, T. Kelly, and R. O'Riordan, "Temporal variation of meiogymnophallus minutus infections in the first and second intermediate host," *Journal of Helminthology*, vol. 84, no. 4, pp. 362–368, 2010. [Cited on page 34.]

[80] A. Dobson, K. D. Lafferty, A. M. Kuris, R. F. Hechinger, and W. Jetz, "Homage to linnaeus: how many parasites? how many hosts?" *Proceedings of the National Academy of Sciences*, vol. 105, no. supplement_1, pp. 11 482–11 489, 2008. [Cited on page 36.]

[81] P. Bartoli and D. I. Gibson, "Synopsis of the life cycles of digenea (platyhelminthes) from lagoons of the northern coast of the western mediterranean," *Journal of Natural History*, vol. 41, no. 25-28, pp. 1553–1570, 2007. [Cited on page 36.]

[82] A. F. Zuur and E. N. Ieno, *Beginner's guide to zero-inflated models with R*. Highland Statistics Limited United Kingdom, 2016. [Cited on pages 36 and 74.]

[83] Z. Yang, J. W. Hardin, and C. L. Addy, "A score test for overdispersion in poisson regression based on the generalized poisson-2 model," *Journal of statistical planning and inference*, vol. 139, no. 4, pp. 1514–1521, 2009. [Cited on page 52.]

[84] B. Yadav, L. Jeyaseelan, V. Jeyaseelan, J. Durairaj, S. George, K. Selvaraj, and S. I. Bangdiwala, "Can generalized poisson model replace any other count data models? an evaluation," *Clinical Epidemiology and Global Health*, vol. 11, p. 100774, 2021. [Cited on page 68.]

[85] K. Lam, H. Xue, and Y. Bun Cheung, "Semiparametric analysis of zero-inflated count data," *Biometrics*, vol. 62, no. 4, pp. 996–1003, 2006. [Cited on page 74.]

[86] D. Lambert, "Zero-inflated poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992. [Cited on page 74.]

[87] A. McDowell, "From the help desk: hurdle models," *The Stata Journal*, vol. 3, no. 2, pp. 178–184, 2003. [Cited on page 79.]

[88] S. N. Wood, *Generalized additive models: an introduction with R*. CRC press, 2017. [Cited on page 85.]

[89] W. N. Venables and B. D. Ripley, *Modern applied statistics with S-PLUS*. Springer Science & Business Media, 2013. [Cited on page 85.]