Alain F. Zuur and Elena N. Ieno

The World of Zero-Inflated Models. Volume 2: Using GLMM

ISBN: 978-1-7399636-1-3
First published August 2024
First print

www.highstat.com

I dedicate this book to two little stars who shine even brighter with time.
- Alain F. Zuur -

I dedicate this book to you, dad! I am very grateful for everything you have done for me. It´s my turn to take care of you in these difficult times.
- Elena N. Ieno -

Contents

Preface

Acknowledgements

We greatly appreciate the efforts of those who wrote R (R Core Team, 2023) and its many packages. This volume would not have been possible without the efforts of the `glmmTMB` (Brooks et al., 2017), `ggplot2` (Wickham et al., 2024a), `mgcv` (Wood, 2017), `plyr` (Wickham, 2023), `dplyr` (Wickham et al., 2023), `cowplot` (Wilke, 2024), `performance` (Lüdecke et al., 2024), `MASS` (Ripley, 2023), `DHARMa` (Hartig, 2022), `ggmap` (Kahle et al., 2023), `sp` (Pebesma and Bivand, 2024), `gstat` (Pebesma and Graeler, 2023), `pscl` (Jackman et al., 2024), `tidyr` (Wickham et al., 2024b), `ggeffects` (Lüdecke, 2024), and `ordbetareg` (Kubinec, 2023) programmers. We hope that they will keep up the excellent work.

We thank ir. Thierry Onkelinx for reviewing Chapters 11 and 12.

We are greatly indebted to all the scientists who either supplied data directly for this book or made their data available through their publications. We used data from Bodensteiner et al. (2019), Twining et al. (2020), Pearse et al. (2018), Rosti et al. (2022), Warner et al. (2016), Lesmerises et al. (2017), and Härkönen et al. (2014). We have no criticisms on any of the analyses carried out in these publications. If we ended up with a different analysis, the primary reason was a pedagogical choice.

While writing this book, we used ChatGPT (ChatGPT-4o) for grammar and style corrections. It's a great tool for non-native English speakers and writers. The text was entirely written by the authors. ChatGPT was also utilised to assist in writing some of the R code.

Data sets and R code used in this book

All datasets featured in this book are available for download at https://highstat.com/. Simply click on the Books menu and select the relevant

book. You can also download all the R code from the same website. To open the ZIP files containing the R code, use the password: <omitted in online pdf>

Cover art

The cover drawing is by Jon Thompson (www.yellowbirdgallery.org). Mr Thompson was born in 1939 to Irish parents and has lived most of his life in Scotland. In the 1980s, he was drawn to the Orkney Islands. He is continually inspired by the landscape and bird life of Orkney. He has been creating bird art for 30 years in a variety of media, including drawing, painting, sculpture, and jewellery, never attempting to reproduce nature, but to draw parallels with it. A close-up view of a bird feather is all the inspiration he needs.

Alain F. Zuur
Newburgh, Scotland

Elena N. Ieno
Allicante, Spain

10

Introduction

This is Volume 2 of our book series 'The World of Zero-Inflated Models'. In Volume 1, we used datasets for which ordinary generalised linear models (GLM) and zero-inflated models were sufficient. In this volume, we increase the complexity of the datasets and models by allowing for a dependency structure. We do this via random effects in generalised linear mixed effects models (GLMMs).

Writing this volume took longer than anticipated. Volume 1 was published in 2021, and then the aftermath of the pandemic kicked in. We had to convert all our courses to online and on-demand formats, which took some time. In 2023, we had to convert everything back to onsite and hybrid courses. We used the material from this book in various online courses, which turned out to be an excellent peer-reviewing process. Then there was (and still is) the delightful chaos caused by the first author's two young children. They require a generous amount of pleasant attention. Finally, there is our own writing ambition. Just when we thought the book was complete, we discovered that the `glmmTMB` package had introduced the ordered beta distribution for analysing proportional data with zeros and ones. Naturally, we felt compelled to include this exciting development.

We also stumbled upon a paper by van der Veen et al. (2023), extending the work of Niku et al. (2019b), discussing generalised linear latent variables (GLLVM). If you sample multiple species at the same site, it is convenient to convert this into a univariate diversity index and apply a GLM or a GLMM (with or without zero-inflation components). However, a GLLVM is a multivariate GLMM that allows for the analysis of the individual species in a multivariate framework. A GLLVM can be fitted with the `gllvm` function from the `gllvm` package (Niku et al., 2023), which also allows for zero-inflated Poisson, zero-inflated negative binomial, Tweedie, and ordered beta distributions. Hence, we immediately wrote another 150 pages on GLLVMs and applied them to zero-inflated data. However, this made the book excessively long—too large to fit through a letterbox (even a generously sized one).

Therefore, we decided to split the material. This volume covers univariate GLMMs, and we will simultaneously release Volume 3, focusing on GLLVM. This was not our original plan, as we intended to include zero-inflated GAMMs

in Volume 3. These will now appear in Volume 4. Who knew the world of zero-inflated models could expand faster than our to-do list?

So, what is in this book?

Chapter 11 contains an extensive explanation of linear mixed-effects models. Originally, we used a dataset of bears and ants, but after discovering that the covariates only explained 2% of the variation, we decided to completely rewrite this chapter with a different dataset on painted turtles. At that point, we had forgotten that the chapter on zero-inflated binomial GLMMs also uses painted turtle data. So, we hope you like turtles.

In Chapter 12, we first introduce Poisson GLMM using a squirrel dataset and discuss marginal and conditional predicted values. The chapter also covers zero-inflated Poisson and generalised Poisson GLMMs.

A zero-inflated Poisson GLMM is applied to a humphead fisheries dataset in Chapter 13. In Chapter 14, we discuss how to handle nested and crossed random effects, as well as auto-correlation, using a dataset on zero-inflated tree hyrax count data.

A detailed explanation of zero-inflated binomial GLMMs is provided in Chapter 15. We use a dataset on painted turtles and also touch upon beta-binomial models.

In contrast, Chapter 16 utilises beta GLMMs, zero-inflated beta GLMMs, zero-altered beta GLMMs, and ordered beta GLMMs for the analysis of zero-inflated caribou data.

Finally, Chapter 17 presents an application of the Tweedie GLMM to zero-inflated biomass data.

The Preface of this book outlines how to access the R code and data sets used in this volume.

11

Introduction to linear mixed-effects models

In this chapter, we will introduce the basic principles of linear mixed-effects models using a dataset on the thermal sensitivity of early life traits in the painted turtle (*Chrysemys picta*). The dataset includes observations from multiple hatchlings from the same clutches, suggesting a likely dependence among observations within each clutch. Ignoring this dependency during statistical analysis is referred to as pseudoreplication, which is discussed in Section 11.4.

In Section 11.5, we explain the underlying principles of a solution for pseudoreplication, namely the linear mixed-effects model. In Section 11.9, we discuss which variables are typically used as random intercepts, and which are not. The difference between linear regression models and linear mixed-effects models is discussed in Section 11.10. Some mathematical background is presented in Section 11.11, although this section can be skipped upon first reading. In Section 11.12, we discuss the differences between conditional and marginal predictions.

11.1 Painted turtles

Species with wide geographic ranges, like the painted turtle (*Chrysemys picta*), often face different climate conditions that require them to adapt. Bodensteiner et al. (2019) studied how early life traits of painted turtles vary with temperature across seven locations in the US (Idaho, Minnesota, Oregon, Illinois, Nebraska, Kansas, and New Mexico). They collected eggs from 9 to 16 clutches per location, totalling 79 clutches, and transported them to Iowa State University for a common garden incubation experiment.

Each egg was labelled, weighed, and randomly assigned to one of eight incubation temperature treatments: 26, 27, 27.5, 28, 28.5, 29, 29.5, and 30 degrees Celsius, matching the temperatures of the original locations. Once a hatchling emerged, it was weighed, and its midline carapace length, carapace width, and midline plastron length were measured.

The study focused on five response variables: incubation duration (days),

hatchling mass (grams), carapace length (mm), carapace width (mm), and plastron length (mm). Bodensteiner et al. (2019) modelled each response variable as a function of initial egg mass, temperature, location, and an interaction between temperature and location. Due to multiple observations from the same clutch, mixed-effects models with `Clutch` as a random intercept were used. Note that detailed 'Location' information is not provided, only the US states are mentioned. We will use that information instead.

The study found that most traits varied by location, suggesting local adaptation and/or maternal effects. This highlights the complex environmental impacts on traits, helping predict responses to climate change.

Let us summarise this experiment in statistical terms. In total, we have 743 hatchlings. For each hatchling, multiple response variables were measured. We have three covariates and one interaction term of interest. The first statistical complication is that these 743 hatchlings come from 79 clutches.

For simplicity, let us focus on one response variable for the moment, carapace width. The carapace is the hard, protective upper shell of a turtle, which covers its back. The carapace width can indicate the turtle's overall health and growth.

The carapace width of the 1-16 hatchlings from the same clutch may be more similar than the carapace width of hatchlings from different clutches. Hatchlings from the same clutch share the same mother and are genetically linked. If the carapace width of one hatchling is low, it is likely that the carapace width of the other hatchlings is also low (and vice versa). This suggests that the 743 measurements of carapace width are not independent of each other. We need to include the covariate 'clutch' in some form in the model to account for this dependency.

Please note that these 79 clutches were chosen randomly from a much larger pool of potential clutches. Our primary interest does not lie in modelling the relationship between carapace width, initial egg mass, temperature, and location for *these* specific clutches alone. Rather, our goal is to extrapolate this model to apply more broadly to clutches in general.

11.2 Import the turtle data

We first import the data from the `PaintedTurtles.csv` csv file using the `read.csv` function.

```
PT <- read.csv(file = "./Data/PaintedTurtles.csv",
               header = TRUE,
               na.strings = "NA",
               stringsAsFactors = TRUE,
               dec = ".")
```

Next, we load all the packages that we will need for this chapter. Just as in Volume 1, we will primarily use the `glmmTMB` package (Brooks et al., 2017). Alternative packages for linear mixed-effects models are `nlme` (Pinheiro et al., 2024) and `lme4` (Bates et al., 2024), among others.

```
library(lattice)
library(ggplot2)
library(mgcv)
library(plyr)
library(cowplot)
library(glmmTMB)
library(performance)
library(ggeffects)
library(MASS)
source("./Data/HighstatLibV15.R")
```

To ensure that variable names are short, and that the variable names in the R code match those in the formulas, we rename `CarapaceWidth` to `CWt`, `CarapaceLength` to `CLt`, and instead of `PlastronLength` we will use `PLt`.

```
PT2 <- plyr::rename(PT,
                    replace = c("CarapaceWidth"  = "CWt",
                                "CarapaceLength" = "CLt",
                                "PlastronLength" = "PLt"))
```

The variable `Clutch` defines the 72 clutches from which the 1–16 eggs were taken. It is coded as an integer with values 1 2 2 2 2 … 72 72 72 72. The variable `State` represents the state where the eggs were collected.