Chapter 1
Introduction

1.1 What Is in the Book?

Does your data have repeated measurements; is it nested (hierarchical)? Is it sampled at multiple locations or sampled repeatedly over time? Or is your response variable heterogeneous? Welcome to our world, the world of mixed effects modelling. The bad news is that it is a complicated world. Nonetheless, it is one that few ecologists can avoid, even though it is one of the most difficult fields in statistics. Many textbooks describe mixed effects modelling and extensions, but most are highly mathematical, and few focus on ecology.

We have met many scientists who have proudly showed us their copy of Pinheiro and Bates (2000) or Wood (2006), but admitted that these were really too technical for them to fully use. Of course, these two books are extremely good, but probably outside the reach of most non-mathematical readers.

The aim of this book is to provide a text on mixed effects modelling (and extensions) that can be read by anyone who needs to analyse their data without the (immediate) need to delve into the underlying mathematics. In particular, we focus on the following:

1. Generalised least squares (GLS) in Chapter 4. One of the main underlying assumptions in linear regression models (which include analysis of variance models) is homogeneity (constant variance). However, our experience has shown that most ecological data sets are heterogeneous. This is a problem that can be solved by using non-parametric tests, transformations, or analysing the raw data with GLS, which extends the linear regression by modelling the heterogeneity with covariates.
2. Mixed effects models and additive mixed effects models in Chapters 5, 6, and 7. We focus on regression and smoothing models for nested data (also called panel data or hierarchical data), repeated measurements, temporal correlated data, and spatial correlated data.
3. Generalised linear modelling (GLM) and generalised additive modelling (GAM) for count data, binary data, proportional data, and zero-inflated count data in Chapters 8–11.

A.F. Zuur et al., *Mixed Effects Models and Extensions in Ecology with R*,
Statistics for Biology and Health, DOI 10.1007/978-0-387-87458-6_1,

4. Generalised estimation equations (GEEs) in Chapter 12. GEE can be used to analyse repeated measurements and longitudinal repeated measurements (over time) data. These can be continuous, binary, proportional, or count data.
5. Generalised linear mixed models (GLMMs) and generalised additive mixed models (GAMMs) in Chapter 13. GLMMs and GAMMs are used to model nested data and temporal and spatial correlation structures in count data or binomial data. These models combine mixed effects modelling and GLM and GAM.

When writing any technical book, a common starting point is to decide on the existing expertise of your target reader. Do we assume no existing expertise or do we assume a certain level of statistical background?

We decided that the entrance level for this text would be good knowledge of linear regression. This means we have assumed a familiarity with the underlying assumptions of linear regression, the model selection process, hypothesis testing procedures (*t*-test, *F*-test, and nested models), backward and forward selection based on the Akaike information criterion (or related information criteria), and model validation (assessing the underlying assumptions based on graphical or numerical tools using the residuals). Appendix A gives a short review of these procedures, and we recommend that you first familiarise yourself with the material in this appendix before continuing with Chapter 2. If you feel uncomfortable with the information in the appendix, then we recommend that you have a look at the regression chapters in, for example, Montgomery and Peck (1992), Fox (2002), or Quinn and Keough (2002). In fact, any book on linear regression will do. Also, our own book, Zuur et al. (2007), can be used.

The next question is then to decide who the book is to be aimed at. Since 2000, the first two authors of this book have given statistical courses for environmental scientists, biologists, ecologists, and other scientists; they have seen about 5000 participants in this time. The material covered in these courses is based on modules described in Zuur et al. (2007). For example, a popular course is the following one:

- Day 1: Data exploration.
- Day 2: Linear regression.
- Day 3: GLM.
- Day 4: GAM.
- Day 5: Catching up.

This is a 40-hour course and has been incorporated into MSc and PhD courses in several countries in Europe as well as being given as in-house and open courses at many universities and research institutes, mainly at biology departments. The problem with this course is that although you can teach people how to do linear regression, GLM, or GAM, the reality is that nearly all ecological data sets contain elements like nested data, temporal correlation, spatial correlation, data with lots of zeros, and heterogeneity. Hence, most ecologists for most of the time will need to apply techniques like mixed effects modelling, GLMM, GAMM, and models that can cope with lots of zeros (zero-inflated GLM). And it is for the user of this type of data that this book is primarily aimed at.

This book is also aimed at readers who want to gain the required knowledge by working through examples by downloading the code and data and try it for themselves before applying the same methods on their own data.

Two of the authors of this book are statisticians and speaking from their experience, having a book like this that first explains complicated statistical methods in a non-mathematical context and demonstrates them in case studies before digging into the underlying mathematics can still be extremely useful, even for the statistician!

The final question was what to write? We have already partially answered this question in the paragraphs above: statistical techniques that can cope with complicated data structures like nested data, temporal and spatial correlation, and repeated measurements for all types of data (continuous, binary, proportional, counts, and counts with lots of zeros).

1.1.1 To Include or Not to Include GLM and GAM

One of our dilemmas when writing this book was whether we should require the reader to be familiar with GLM and GAM before reading this book. We decided against this and have included GLM and GAM chapters in this book for the following reasons.

1. During the pre-publication review process, it became clear that many instructors would use this book to explain the full range of methods beyond linear regression. It, therefore, made sense to include GLM and GAM, allowing students to buy a single book containing all the methods beyond linear regression.
2. Most statistical textbooks written 5 or 10 years ago tend to discuss only logistic regression (for absence–presence and proportional data) and Poisson regression (for count data). In reality, Poisson regression hardly ever works for ecological count data due to its underlying assumption that the variance equals the mean of the data. For most ecological data sets, the variance is larger than the mean; this phenomenon is called overdispersion. Negative binomial GLMs and GAMs have become increasingly popular to deal with overdispersion. However, we still cover Poisson GLM as a pre-requisite to explain the negative binomial (NB) GLM.
3. Many ecological data sets also contain large number of zeros, and during the last 5 years, a new set of models have become popular in ecology to deal with this. These include zero-inflated Poisson GLMs and GAMs and zero-inflated negative binomial GLMs and GAMs. Zero inflated means that we have a data set with lots of zeros, more than we expect based on the Poisson or negative binomial distribution. The excessive number of zeros may (or may not!) cause overdispersion. Using these zero-inflated models means that we can often solve two problems at once: overdispersion and the excessive number of zeros. But again, before we can explain these zero-inflated models, we have to ensure that the reader is fully familiar with Poisson and logistic GLMs.

This explains why we have included text on the Poisson GLM, negative binomial GLM, and zero-inflated Poisson and the increasingly useful negative binomial GLMs and GAMs.

A few applications of zero-inflated Poisson GLMMs and zero-inflated negative binomial GLMMs/GAMMs have been published recently. However, there is hardly any fully tested software around that can be used to fit these zero-inflated GLMMs and GAMMs. So, although we decided to include the zero-inflated GLMs and GAMs in this book, we leave zero-inflated GLMMs and GAMMs for a future text.

1.1.2 Case Studies

A common criticism of statistical textbooks is that they contain examples using 'ideal' data. In this book, you will not find ozone data or Fisher's iris data to illustrate how well certain statistical methods work. In contrast, we have only used data sets from consultancy projects and PhD research projects, where for many our first reaction was "How are we ever going to analyse these data?"

As well as the chapters on applied theory, this book also contains ten case study chapters with each case study showing a detailed data exploration, data analysis, discussion and a 'what to write in a paper' section. In the data exploration and data analysis section, we describe our thinking process, and in the 'what to write in a paper' section, we emphasise the key points for a paper.

It should be noted that our analysis approach for these data may not be the only one; as it is often the case, multiple statistical analyses can be applied to the same data set.

The data used in the case studies, and in the main text, are all available from the book's website at www.highstat.com. The computer code is also available for downloading. If you want to use any of the data from this book for publications, please contact the owner of the data for permission. Contact details are given at the beginning of the book.

1.1.3 Flowchart of the Content

The flowchart in Fig. 1.1 gives a schematic overview of this book. In Part I, we start discussing the limitations of the linear regression model and show how these limitations can be solved with additive modelling, including random effects (resulting in mixed effects models), and temporal and spatial correlation. In Part II, we discuss GLM, GAM, and zero-inflated models. In Part III, we combine Parts I and II and discuss GEE, GLMM, and GAMM. Finally, in Part IV, we present ten case studies, each of them showing a detailed example using real data.

There are various ways to use this book. You can start reading the case studies, find one that matches your data, and apply the same steps on your own data. Then look up the corresponding theory. The alternative is to read the theory first, perhaps

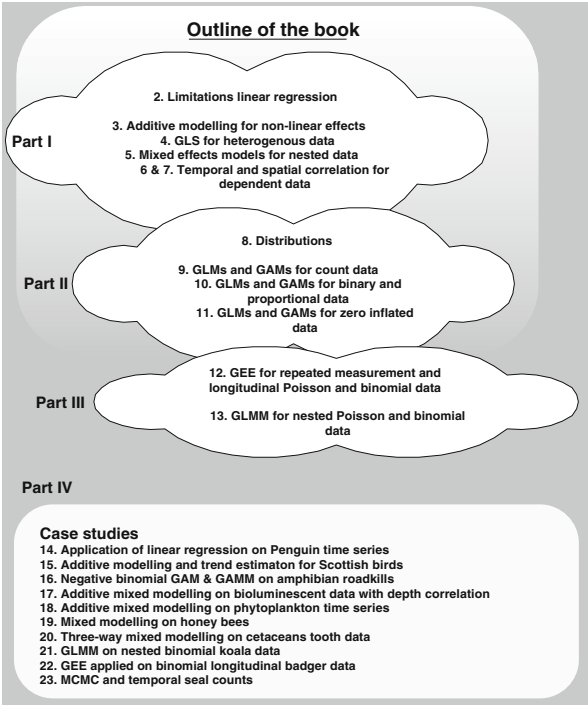

Fig. 1.1 Outline of this book. In Part I, the limitations of linear regression are discussed, and various solutions are discussed (additive modelling for non-linear patterns, GLS for heterogeneity, mixed effects modelling for nested data, and correlation structures to deal with dependence). In the second part, GLM and GAM are introduced, and in the third part, these methods are extended towards GLMM and GAMM. In the last part, case studies are presented

concentrate on the numerous examples, and find a matching case study. Yet, a third option is to read the book from A to Z (which we obviously advise our readers).

Some sections are marked with an asterisk. These are more technical sections, or expand on ideas in the main text. They can be skipped on the first reading.

1.2 Software

There are many software packages available for mixed effects modelling, for example MLWIN, SPLUS, SAS, Stata, GENSTAT, and R. All have excellent facilities for mixed effects modelling and generalised linear mixed modelling; see West et al. (2006) for a comparison. As to GAM and GAMM, we can only recommend SPLUS or R. Stata seems to be particularly suited for negative binomial models, but has limited GAM facilities (at the time of writing).

Our choice is R (www.r-project.org), because it is good and it is free. There is no point teaching students a complicated computer language in a 2500 USD package if a future employer is unwilling to buy the same package. Because R is free, this is not an issue (unless the employer demands the use of a specific package).

If you are an instructor and use this book for teaching, we advise you start your class with an introductory course in R before starting with this book. We have tried teaching R and statistics at the same time, but have found this is rather challenging for the student.

The pre-requisite R knowledge required for this book is fairly basic and is covered in Appendix A; important commands are `boxplot`, `dotchart`, `pairs`, `lm`, `plot`, `summary`, and `anova`. Some basic R skills in data manipulating and plotting will also be useful, especially if the data contain missing values.

Instructors can contact us for an R survival guide that we wrote for our own courses. It contains all essential R code for pre-required knowledge for this book.

1.3 How to Use This Book If You Are an Instructor

We wrote this book with teaching in mind. When we teach, we tend to have groups consisting of 10–25 people (environmental scientists, biologists, etc.), mostly consisting of PhD students, post-docs, consultants, senior scientists, and the occasional brave MSc students. As people can only fully appreciate the text in this book if they have good knowledge of linear regression and basic R knowledge, our courses contain the following:

- Day 1: Revision of linear regression and R (half a day).
- Day 1 and 2: GLS.
- Day 3: Mixed effects modelling and additive mixed modelling.
- Day 4: Adding temporal and spatial correlation to linear regression, mixed effects models, and additive (mixed) models.
- Days 5 and 6: GEE, GLMM, and GAMM.

Each day is 8 hours of teaching and exercises. The case studies and detailed examples in the sections can be used as exercises. The schedule above is challenging, and depending on the pre-knowledge and number of questions, 48 hours may not be enough.

We have taught our courses in more than 20 different countries and noticed that there is a huge difference in mathematical and statistical knowledge of students. We have had groups of 60 MSc students where 20 had never seen any statistics at all, 20 were familiar with basic statistics, and 20 had done regression and GLM during their undergraduate courses and were keen to move on to GLMMs and GAMMs! This applies not only to MSc courses but also to postgraduate courses or courses at research institutes. Hence, teaching statistics is a challenge.

Before starting with the mixed effects modelling material, you need to ensure that all students are familiar with concepts like interaction, comparing full and nested models, model validation, sketching fitted values, and dealing with nominal variables.

1.4 What We Did Not Do and Why

During the writing of this book and when it was finished, we received comments from a large group of people, including the referees. This resulted in an enormous amount of ideas and suggestions on how to improve the text, and most of these

suggestions were included in the final version, but a few were not. As some of these topics are important for all readers, we decided to briefly discuss them.

Originally, our plan was to provide all the data in nicely prepared ASCII files and use the `read.table` command to import the data into R. However, data preparation is also part of the analyses, and we therefore decided to provide the data in the same format as was given to us. This means we put the reader through the same data preparation process that they would need to go through with their own data. With the `read.table` command, one has to store the data somewhere physically in a directory, e.g. on the C or D drive, and access it from there. However, not everyone may be able to store data on a C drive due to security settings or has a D drive. To avoid any confusion, we created a package (don't call it a library!) that contains all data sets used in this book. This means that any data set used in this book can be accessed with a single command (once the package has been installed). Our package is available from the book website at www.highstat.com. There, you can also find all the R code and data files in ASCII format, should you wish to use the `read.table` command.

It has also been suggested that we include appendices on matrix algebra and giving an introduction to R. We think that this would duplicate material from other books as many statistical textbooks already contain appendices on matrix algebra. As for R, we suggest you get a copy of Dalgaard (2002) and spend some time familiarising yourself with it. Appendix A shows what you need to know to get started, but R warrants spending additional time developing your expertise. We realise this means that you need to buy yet more books, but information on matrix algebra and R programming can also be obtained free from the Internet.

We have also deliberately decided not to add more mathematics into the text. If, after completing the book, you have a desire to dig further into the mathematical details, we recommend Pinheiro and Bates (2000) or Wood (2006).

1.5 How to Cite R and Associated Packages

This is an important issue. Without the effort of the people who programmed R and the packages that we have used, this book would not exist. The same holds for you; you have access to a free package that is extremely powerful. In recognition, it is appropriate therefore to cite R or any associated package that you use. Once in R, type

```
> citation()
```

and press enter. Do not type the > symbol. It gives the following text.

```
To cite R in publications use:
R Development Core Team (2008). R: A language and environment
for statistical computing. R Foundation for Statistical
Computing, Vienna, Austria. ISBN 3-900051-07-0,
URL http://www.R-project.org.
```

```
...
We have invested a lot of time and effort in creating R,
please cite it when using it for data analysis. See also
'citation("pkgname")' for citing R packages.
```

The last lines suggest that for citing the `mgcv` or `nlme` packages (which we will use a lot), you should type

```
> citation("nlme")
> citation("mgcv")
```

It gives full details on how to cite these packages. In this book, we use a large number of packages. Citing them each time would drastically increase the number of pages; so for the sake of succinctness, we mention and cite them all below. In alphabetic order, the packages used in the book and their citations are as follows: `AED` (Zuur et al., 2009), `BRugs` (Thomas et al., 2006), `coda` (Plummer et al., 2007), `Design` (Harrell, 2007), `gam` (Hastie, 2006), `geepack` (Yan, 2002; Yan and Fine 2004), `geoR` (Ribeiro and Diggle, 2001), `glmmML` (Broström, 2008), `gstat` (Pebesma, 2004), `lattice` (Sarkar, 2008), `lme4` (Bates and Sarkar, 2006), `lmtest` (Zeileis and Hothorn, 2002), `MASS` (Venables and Ripley, 2002), `mgcv` (Wood, 2004; 2006), `ncf` (Bjornstad, 2008), `nlme` (Pinheiro et al., 2008), `pscl` (Jackman, 2007), `scatterplot3d` (Ligges and Mächler, 2003), `stats` (R Development Core Team, 2008), and `VGAM` (Yee, 2007). The reference for R itself is R Development Core Team (2008). Note that some references may differ depending on the version of R used. While writing this book, we used versions 2.4.0–2.7.0 inclusive, and therefore, some references are to packages from 2006, while others are from 2008.

1.6 Our R Programming Style

One of the good things about R is also, perversely, a problem; everything can be done in at least five different ways. To many, of course, this is a strength of R, but for beginners it can be confusing. We have tried to adopt a style closely matching the style used by Pinheiro and Bates (2000), Venables and Ripley (2002), and Dalgaard (2002). However, sometimes these authors simplify their code to reduce its length, minimise typing, and speed up calculation. For example, Dalgaard (2002) uses the following code to print the output of a linear regression model:

```
> summary(lm(y ~ x1 + x2))
```

An experienced R user will see immediately that this combines two commands; the `lm` is used for linear regression, and its output is put directly into the `summary` command, which prints the estimated parameters, standard errors, etc. Writing optimised code, such as this, is good practice and in general something to be

encouraged. However, in our experience, while teaching statistics to R beginners, it is better to explicitly write code as easily followed steps, and we would write the above examples as

```
M1 <- lm(y ~ x1 + x2)
summary(M1)
```

We call this a – b – c programming; first a, then b, and finally c. This may not produce the most elegant or most efficient code, but its simplicity makes it easier to follow when learning R.

1.7 Getting Data into R

The most difficult thing in learning a new stats package is to import your data and start working with it. As an example of importing data in R, we use data from Cronin (2007), which is also used in Chapter 23. The following R code reads the data. We assume the data are available as a text (tab-delimited) file 'Seals.txt' on the C drive in the directory 'Bookdata'. The following code reads the data into R:

```
> Seals <- read.table(file = "C:\\Bookdata\\Seals.txt",
                      header = TRUE)
```

The > symbol is used to mimic the R commander. You should not type it into R! R commands are case sensitive; so make sure you type in commands exactly as illustrated. The `header = TRUE` option tells R that the first row contains headers (the alternative is FALSE). The data are stored in a data frame called `Seals`, which is a sort of data matrix. Information in a data frame can be accessed in various ways.

If you just type in `Abun` (the column with abundances), R gives an error message saying that it does not know what `Abun` is. There are various options to access the variables inside the object `Seals`. You can use commands like

```
> hist(Seals$Abun)
```

to make a histogram of the abundance. The `$` sign is used to access variables inside the object `Seals`. It is also possible to work along the lines of

```
> A <- Seals$Abund
> hist(A)
```

First, we define a new variable `A` and then work with this. The advantage is that you don't have to use the `Seals$` all the time. Option three is to access the data via columns of the object `Seals`:

```
> A <- Seals[,1]
> hist(A)
```

A fourth option is to provide the Seals object as an argument to the function that you use, e.g.

```
> lm(Abun ~ factor(Site), data = Seals)
```

The `data` option specifies that R has to use the data in the object Seals for the linear regression. Yet, a fifth option is to use the `attach(Seals)` command. This command tells R to look also inside the object `Seals`; hence, R will have access to anything that you put in there. Its advantage is that with one command, you avoid typing in lots of data preparation commands. In writing a book, it saves space. In classroom teaching, it can be an advantage too because students don't have to type all the $ commands.

However, at this point, the R experts tend to stand up and say that it is all wrong; they will tell you not to use the `attach` command. The reason is that you can attach multiple objects, and misery may happen if multiple objects contain the same variable names. This may cause an error message (if you are lucky). The other problem is that you may (accidentally) attach the same object twice. If you then make changes to a variable (e.g. a transformation), R may use the other (unchanged) copy during the analysis without telling you! Our advise is not to use the `attach` command, and if you decide to use it, be very careful!

1.7.1 Data in a Package

In this book, we use at least 30 different data sets. Instead of copying and pasting the `read.table` command for each example and case study, we stored all data in a package called `AED` (which stands for Analysing Ecological Data). It is available from the book website at www.highstat.com. As a result, all you have to do is to download it, install it (Start R, click on Packages, and select 'Install package from local zip file'), and then type

```
> library(AED)
> data(Seals)
```

Instead of the `Seals` argument in the function `data`, you can use any of the other data sets used in this book. To save space, we tend to put both commands on one line:

```
> library(AED); data(Seals)
```

You must type the ";" symbol. You can even use a fancy solution, namely

```
> data(Seals, package = "AED")
```