Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

Use R!

Albert: Bayesian Computation with R
Bivand/Pebesma/Gómez-Rubio: Applied Spatial Data Analysis with R
Cook/Swayne: Interactive and Dynamic Graphics for Data Analysis:
With R and GGobi
Hahne/Huber/Gentleman/Falcon: Bioconductor Case Studies
Paradis: Analysis of Phylogenetics and Evolution with R
Pfaff: Analysis of Integrated and Cointegrated Time Series with R
Ritz/Streibig: Nonlinear Regression with R
Sarkar: Lattice: Multivariate Data Visualization with R
Spector: Data Manipulation with R

Christian Ritz • Jens Carl Streibig

Nonlinear Regression with R

Springer

Christian Ritz
Department of Basic Sciences
and Environment (Statistics)
Faculty of Life Sciences
University of Copenhagen
Thorvaldsensvej 40
DK-1871 Frederiksberg C
Denmark
ritz@life.ku.dk

Jens Carl Streibig
Department of Agriculture and Ecology
(Crop Science)
Faculty of Life Sciences
University of Copenhagen
Hoejbakkegaard Allé 13
DK-2630 Taastrup
Denmark
jcs@life.ku.dk

ISBN: 978-0-387-09615-5 e-ISBN: 978-0-387-09616-2
DOI: 10.1007/978-0-387-09616-2

Library of Congress Control Number: 2008938643

Printed on acid-free paper

springer.com

To Ydun Marie Ritz and in memory of Erik Ritz

Preface

This book is about nonlinear regression analysis with **R**, in particular, how to use the function `nls()` and related functions and methods.

Range of the book

Nonlinear regression may be a confined and narrow topic within statistics. However, the use of nonlinear regression is seen in many applied sciences, ranging from biology to engineering to medicine and pharmacology. Therefore, this book covers a wide range of areas in the examples used. Appendix A lists the disciplines from which data are used in this book.

What not to expect

This book is not a textbook on nonlinear regression. Basic concepts will be briefly introduced, but the reader in need of more explanations will have to consult comprehensive books on nonlinear regression such as Bates and Watts (1988) or Seber and Wild (1989). Instead, this book may be particularly well-suited as an accompanying text, explaining in detail how to carry out nonlinear regression with **R**. However, we also believe that the book is useful as a stand-alone, self-study text for the experimenter or researcher who already has some experience with **R** and at the same time is familiar with linear regression and related basic statistical concepts.

Prerequisites

Experience with **R** at a level corresponding to the first few chapters in Dalgaard (2002) should be sufficient: The user should be acquainted with the basic objects in **R** such as vectors, data frames, and lists, as well as basic plotting and statistics functions for making scatter plots, calculating descriptive statistics, and doing linear regression.

How to read the book

Chapter 2 is essential for getting started on using `nls()`. Section 3.2 and Chapter 4 are at times somewhat technical regarding the use of **R**, whereas Section 6.3 is technical on a statistical level. These parts of the book could be skipped on a first reading.

The **R** extension package `nlrwr` is support software for this book, and it is available at CRAN: http://cran.r-project.org/web/packages/nlrwr/index.html. All datasets and functions used are available upon loading `nlrwr`. All code snippets used in the book are also found in the `scripts` folder that comes with the package. This means that all **R** code snippets shown in this book can be run once the support package `nlrwr` has been installed and loaded. Appendix A provides a list of all datasets used, with a reference to the package where they are found, and Appendix C lists the main functions used in this book together with a package reference.

Acknowledgments

We would like to thank Claire della Vedova and Christian Pipper for proofreading parts of the book. The first author also wishes to thank the participants of the short course *Non-linear regression with **R***, held at the Faculty of Life Sciences, University of Copenhagen, in September 2007, for useful comments and suggestions. We are also grateful to Spencer Graves for his valuable comments on an almost final version of the book. All remaining breaches or errors rest solely on the authors.

This volume has benefitted vastly from the many comments and suggestions from the anonymous reviewers of earlier versions. We are also thankful to John Kimmel for his encouragement and guidance throughout this book project.

Finally, we would like to thank the **R** Core Development Team for making all this happen by developing a great open source project. The book has been written using LaTeX 2ε and Sweave, yet another powerful invention in the wake of the **R** project.

Copenhagen
July 2008

Christian Ritz
Jens C. Streibig

Contents

1

Introduction

Throughout this book, we consider a univariate response, say y, that we want to relate to a (possibly multivariate) predictor variable x through a function f. The function f is not completely known, but it is known up to a set of p unknown parameters $\beta = (\beta_1, \ldots, \beta_p)$. We will use various Greek and Latin letters to denote parameters, often using the symbols typically used in the particular models. The relationship between the predictor and the response can be formulated as follows:

$$y = f(x, \beta) \tag{1.1}$$

This book is about the situation where the function f is nonlinear in one or more of the p parameters $\beta_1, \ldots, \beta_p$. In practice, the parameters have to be estimated from the data. Consider a dataset consisting of n pairs $(x_1, y_1), \ldots, (x_n, y_n)$. (The number of parameters occurring in f should be less than the number of observations; that is, $p < n$.) The relationship in Equation (1.1) is for the ideal situation where both the predictor values $x_1, \ldots, x_n$ and the response values $y_1, \ldots, y_n$ are observed without error. In reality, there will be measurement errors that will distort the picture such that none of the pairs $(x_1, y_1), \ldots, (x_n, y_n)$ will fit Equation (1.1) exactly. Therefore, we will assume that the value x_i is predicting the value y_i according to Equation (1.1) apart from some measurement error. In other words, it is more realistic to entertain the idea that the relationship in Equation (1.1) is correct only *on average*. We can formalise this notion by introducing the conditional mean response $E(y_i|x_i)$ (conditional on the predictor value x_i) and recast Equation (1.1) as follows:

$$E(y_i|x_i) = f(x_i, \beta) \tag{1.2}$$

Equation (1.2) reads as follows: Given the predictor value x_i, we will expect the response to be centered around the value $f(x_i, \beta)$. Therefore, we will refer to f as the mean function.

In the formulation above, it is implicitly assumed that the data analyst has some prior knowledge about which kind of function f should be used (at least roughly). Thus nonlinear regression methods are suited for analysing data for which there is an empirically or theoretically established functional relationship between response and predictor.

Each measurement will be distorted by some error related to the measurement process. The observation y_i will differ from the expected mean $E(y_i|x_i)$ by some amount, which we will denote ε_i. The perturbations may be due to minute changes in the measurement process. Therefore, the complete specification of the model of the relationship between the response and the predictor is given by the nonlinear regression model:

$$y_i = E(y_i|x_i) + \varepsilon_i = f(x_i, \beta) + \varepsilon_i \tag{1.3}$$

We will think of the term ε_i as the error term for observation i; that is, the distortion in the response y_i away from the expected value $f(x_i, \beta)$ caused by various unknown sources of variation. The error ε_i will vary from measurement to measurement. Typically, the errors are assumed to be normally distributed with mean 0 and some unknown standard deviation σ that is estimated from the data. We will go into more detail about the assumptions underlying model (1.3) in Chapter 5. We will in this book specify models by means of the mean function involved but implicitly having the complete specification as given in Equation (1.3) in mind. The variables x and y often will be replaced with the actual names being used in a given context and for a given dataset.

Now we will present three examples. The first example is motivating the use of a two-parameter nonlinear regression model. In the second example, a nonlinear regression model involving two predictor variables is introduced. The third example presents a grouped data structure in a nonlinear regression framework.

1.1 A stock-recruitment model

In fisheries biology, there are several theoretical models describing the relationship between the size of the spawning stock (the spawning biomass) and the resulting number of fish (the recruitment). The data frame `M.merluccius` in the package `nlrwr` contains three variables: `spawn.biomass` (the stock), `num.fish` (the recruitment), and `year` (which we will not use). Figure 1.1 shows the plot of recruitment versus stock. We may be able to discern an increase that flattens out as spawning biomass increases, but we also notice that there is considerable scatter/variation in the data.

One of these models is the Beverton-Holt model:

$$f\big(S, (\alpha, k)\big) = \frac{\alpha S}{1 + S/k} \tag{1.4}$$

```
> plot(num.fish ~ spawn.biomass,
+     data = M.merluccius, xlab = "Spawning biomass (1000 tonnes)",
+     ylab = "Recruitment (million fish)")
```

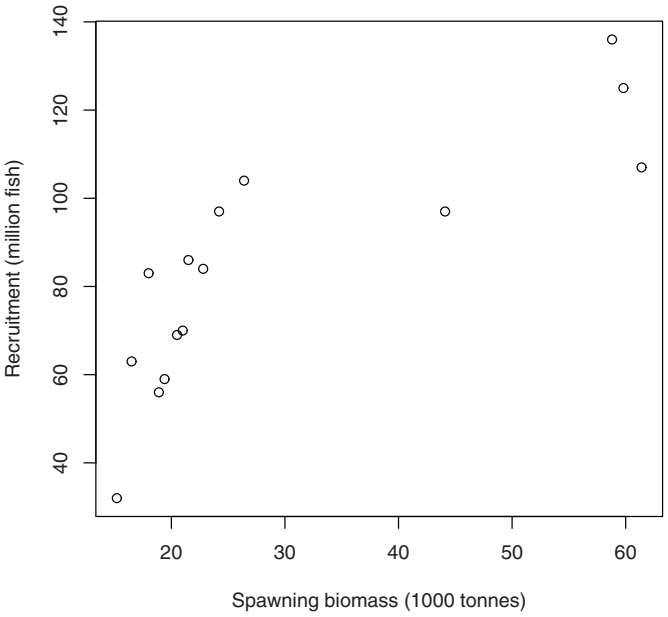

Fig. 1.1. Recruited number of fish plotted against the size of the spawning stock.

The parameter α is the slope at 0, whereas k is the stock resulting in a recruitment halfway between 0 and the upper limit, which is equal to $\alpha \cdot k$. In practice, there is a lot of variability in this kind of data (Cadima, 2003, p. 47), so it is not reasonable to assume that the relationship in Equation (1.4) is deterministic (Carroll and Ruppert, 1988, p. 139), and therefore a nonlinear regression model with the mean function in Equation (1.4) comes into play. We will return to this dataset in Section 7.6.

1.2 Competition between plant biotypes

We want to assess the relative competitive abilities of two biotypes of *Lolium rigidum*. One has developed resistance to glyphosate, and the other is a sensitive wild type (Pedersen et al., 2007). The experimental layout in the greenhouse was an incomplete factorial design. The density of the resistant and

sensitive biotypes was based upon the actual density counted after germination (see Fig. 1.2). This sometimes differed from the intended density. The model we use is a hyperbolic model (Jensen, 1993). It describes the competitive ability of the sensitive biotype in relation to the resistant biotype. More specifically, the relationship between biomass per plant of the sensitive biotype (the response) and the densities is described by the function

$$f\big(x, z, (a, b, c)\big) = \frac{a}{1 + b(x + cz)} \tag{1.5}$$

where x and z are the density per unit area of the sensitive and resistant biotype, respectively. The interpretation of the parameters is:

- a is the theoretical biomass of a plant at zero density.
- b is a measure of the intraspecific competition between plants of the sensitive biotype.
- c is the substitution rate, and it is effectively the exchange rate between the biotypes: If c equals 1, then the two biotypes are equally competitive, and if c is greater than 1, the sensitive biotype is more competitive than the resistant one and vice versa if c is smaller than 1.

The data are in the data frame `RScompetition` in the package `drc`. The plot of the data is shown in Fig. 1.2. For the label on the x axis, we use mathematical annotation (see `?plotmath` for the details). We use the construct `as.numeric(as.factor(RScompetition$z))` to convert the integer values of the variable z in the range 0 to 64 to integer values from 1 to 10, which are suitable as values of the argument `pch`, which is controlling the plot symbol (Dalgaard, 2002, pp. 8, 173). We will return to this dataset in Subsection 4.4.1.

1.3 Grouped dose-response data

Christensen et al. (2003) describe an experiment that was designed in order to compare the potency of the two herbicide treatments in white mustard plants (*Sinapis alba*).

The data are from a dose-response experiment, which means that a biological stimulus is recorded for a range of doses of some toxic substance. The aim of such experiments is typically to assess the toxicity of the substance applied. The data consist of measurements for the herbicide glyphosate at six different doses and for the herbicide bentazone at seven different doses; for all doses, there are four replicates. The dose is in the unit g/ha and the response is dry matter measured in g/pot. Furthermore, both herbicides have a control of zero dose with eight replicates. The data are available as the data frame `S.alba` in the package `drc`. The data frame consists of a total of 68 observations and three variables: the response `DryMatter`, the predictor `Dose`, and the factor `Herbicide`, with two levels (`Bentazone` and `Glyphosate`) identifying the two treatments.

```
> plot(biomass ~ x, data = RScompetition,
+     log = "", xlab = Density ~
+         (plants/m^2), ylab = Biomass ~
+         of ~ sensitive ~ biotype ~
+         (g/plant), pch = as.numeric(as.factor(RScompetition$z)))
```

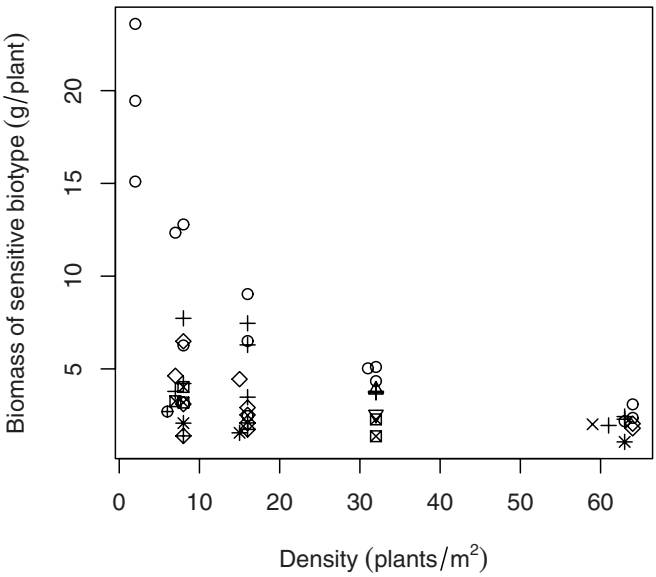

Fig. 1.2. Plot of biomass of sensitive biotype in response to its density. The biomass of the sensitive biotype growing together with the resistant biotype has different symbols.

To get a first impression of the data, we make a plot. As we have two herbicide treatments, we choose a conditional scatter plot using the function `xyplot()` in the package `lattice`. The resulting plot is shown in Fig. 1.3 and displays a marked s-shaped or sigmoid curve for each herbicide treatment. One commonly used empirically based model for such dose-response data is the log-logistic model (Streibig et al., 1993) with mean function

$$f\big(\text{dose}, (b, c, d, e)\big) = c + \frac{d - c}{1 + \exp[b\{\log(\text{dose}) - \log(e)\}]} \tag{1.6}$$

The parameters c and d refer to the lower and upper horizontal asymptotes of the s-shaped curve. The parameter e in model (1.6) is the inflection point of the curve (that is, the point around which it is symmetric on a logarithmic