Book proposal: *Integrating ecological models and data in* R

Ben Bolker

December 29, 2004

Modern statistical techniques and computational power let ecologists use models that correspond to specific biological questions, rather than (as is more traditional) squeezing their data to fit the assumptions of classical statistical models such as analysis of variance or linear regression. This book's goal is to teach the philosophy, and much of the nitty-gritty, of how to actually construct and use such models. It uses the R language (an open-source, freely available dialect of the S language and relative of the commercial S-PLUS system) as a platform to introduce the reader to all of the techniques required to build statistical models, fit them to data, and evaluate what the results mean in ecological terms. Along the way it also addresses basic philosophical issues in modern statistics (e.g. hypothesis testing and estimation, Bayesian vs. frequentist approaches, model selection, parametric vs. nonparametric models) in a concrete context.

The book starts by reminding students of some of the basic tools of statistical modeling (simple probability theory, probability distributions, useful mathematical functional forms); offers some reminders and opinions about the philosophy and purpose of statistical models; and then goes all the way through the actual practical mechanics of constructing, implementing, and estimating the parameters of statistical models for real data. Although the book emphasizes custom-building models rather than using pre-existing tools (e.g. generalized linear models or nonlinear regression), it does (1) teach many of the basic principles underlying these tools (likelihood, parsimony, etc.) and (2) put these tools into a more general perspective. Part of the point of custom-building models is to empower students and give them deeper knowledge — showing them the basic framework on which all advanced statistical techniques are built. More important, custom-building models gives much-needed flexibility in modeling. It only takes a small change in assumptions to put a problem beyond the scope of pre-existing tools. For example, in the framework of a standard logistic regression model (a binary response whose probability is driven by a continuous covariate), imposing a minimum probability of occurrence is impossible. If you know how to build your own models, then you can make the model fit the data rather than the other way around.

The book emphasizes a common-sense and eclectic approach to model-building:

graphing data in many different ways (there is a chapter on exploratory data and graphics in R that covers some basic ways of looking at a data set), getting a feel for what the parameters of a model mean in both mathematical terms (e.g. half-maximum, slope, y-intercept, etc.) and ecological terms (conversion efficiency, fraction of individuals resistant to predation), and always remembering to tie together the statistical/mathematical and ecological conclusions. It discusses a wide range of models, focusing on the tradeoffs between phenomenological models (which simply describe the mathematical relationships and statistical distributions apparent in the data) and mechanistic models (which attempt to derive these shapes and distributions from underlying ecological processes).

Audience The book assumes a basic knowledge of calculus and an introductory statistics course covering topics such as hypothesis testing, analysis of variance, and linear regression; the course that it's based on has catered to upper-level graduate students who have some fairly concrete ecological questions in mind and typically have already collected some of their own data that they can use as the foundations of a class project. The book also assumes basic computer literacy — e.g. use of Microsoft Excel — but no prior programming or use of advanced statistics packages such as SAS. First-year grad students who are either highly motivated or somewhat more mathematically/computationally sophisticated than the average ecology grad student can probably also profit from the book, as can postdocs or academics looking to learn these techniques for the first time or brush up on them.

Context, and what the book is not

- The closest analogue to this book is Hilborn and Mangel's *The Ecological Detective* (ED) [5]. This book is not really ED volume 2, but it's not too far off. ED is thoughtful and reader-friendly, and exposes the reader to a very wide range of concepts (many of the same concepts covered by this book), but really provides too little technical detail for the average reader to come out at the end being able to construct their own models. Hilborn and Mangel made the decision (sensible for what they wanted to accomplish) not to base the book on a particular computing platform, and thus do not show any actual code examples. (The downside of using a particular computing platform is that it will eventually become obsolete, making the book less useful. On the other hand, R is currently under extremely active development and use by researchers in many fields. It's also free, making it highly accessible to students and others with limited resources (academics at small institutions and in developing countries).)

- The other end of the spectrum is M. Crawley's *Statistical Computing: An Introduction to Data Analysis using S-PLUS* [2] (an updated edition of his excellent earlier *GLIM for Ecologists* [1] — an example of the problems of platform-specific books), which teaches the nuts and bolts of ecological data analysis in S-PLUS/R, along with a great deal of useful statistical philosophy and advice. The difference from this book is that Crawley focuses on standard, rather than custom-built, statistical approaches.

- The book does *not* cover dynamic modeling in ecology, theoretical ecology, or biological modeling; it focuses on statistical models, although it does use models that are more flexible and more closely linked to biological mechanisms than standard statistical models.
  - General modeling books: Wilson's *Simulating Ecological and Evolutionary Systems in C* [8] is an introduction to ecological simulation models; Haefner's *Modeling Biological Systems* [3] is more general (and platform-agnostic). There are plenty of other books on (dynamic) modeling of biological systems, some platform-based (C, STELLA, etc.), some agnostic.
  - Theoretical ecology books: Roughgarden's *Primer of Ecological Theory* [7] is MATLAB-based; other books (Nisbet and Gurney, Gurney and Nisbet, Gotelli, Case) are not.

- R books: `http://www.r-project.org/doc/bib/R-publications.html` gives a current list of publications about and using R. Most cover either introductory statistics or particular topics (regression, mixed-effects models, etc.). A few (Maindonald and Braun [6] and Heiberger and Holland [4]) are vaguely similar to this book, but more focused on classical approaches and probably intended more for reference or advanced use.

Table of contents

Chapter 1. Introduction. Custom-built models; close correspondence between ecological questions/hypotheses and statistical models. Philosophy: method of multiple working hypotheses. De-emphasis of null models, limitations of p-values: estimation. Parsimony. Common sense. (sketchy: I've been depending on ED) (Chunks 1, 2).

Chapter 2. Math review: calculus reminders. How to approach a new function: critical points and limits, Taylor expansion. Bestiary of mathematical functions, with examples in R. Combining functions. (Mostly written.) (Chunk 3)

Chapter 3. Probability/stats review: reminders of basic probability theory. Bestiary of probability distributions, with examples in R. Compounded and generalized distributions. (Mostly written.) (Chunks 4,5,6)

Chapter 4. Exploratory data analysis and graphics in R. (Unwritten.)

Chapter 5. Likelihood and all that: likelihood curves and surfaces. Confidence limits and confidence regions. Model selection. The likelihood ratio test. Parsimony/bias vs. variance tradeoff. Information criteria. (Partly written.) (Chunks 8,11,12)

Chapter 6. Practical optimization techniques: Direct search, Nelder-Mead, derivative-based methods, simulated annealing. Optimization tips and trouble-shooting. (Partly written.) (Chunks 9 and 10)

These are the main points, and they are enough to fit models to answer a wide range of ecological questions. There are other topics that I'd like to cover, although they may be covered as optional chapters or in some other way outside of the main flow, including:

- Bayesian methods (it *may* make sense to try to introduce these earlier in the book. One of the drawbacks of ED is that readers have often run out of steam by the time they get to the Bayesian chapters) (Chunk 15)
- Mixed/hierarchical models, Markov Chain Monte Carlo approaches (Chunk 16)
- Connections and interrelationships of classical models (ANOVA, linear regression, nonlinear regression, GLMs, time-series analysis, survival analysis, etc.) (Chunk 13)
- Dynamic models (Chunk 14)

Running data examples?
See attached document for note "chunks".

References

[1] Michael J. Crawley. GLIM *for ecologists*. Blackwell Scientific, Boston, 1993.

[2] Michael J. Crawley. *Statistical Computing: An Introduction to Data Analysis using S-PLUS*. John Wiley & Sons, 2002.

[3] J. W. Haefner. *Modeling Biological Systems: Principles and Applications*. Kluwer, 1996.

[4] Richard M. Heiberger and Burt Holland. *Statistical Analysis and Data Display: An Intermediate Course with examples in S-PLUS, R, and SAS*. Springer Texts in Statistics. Springer, 2004.

[5] R. Hilborn and M. Mangel. *The ecological detective : confronting models with data*. Princeton University Press, Princeton, New Jersey, USA, 1997.

[6] John Maindonald and John Braun. *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, 2003.

[7] J. Roughgarden. *Primer of Ecological Theory*. Prentice Hall, 1997.

[8] Will Wilson. *Simulating Ecological and Evolutionary Systems in C*. Cambridge University Press, Cambridge, UK, 2000.