Journal of Statistical Software

*July 2006, Volume 16, Book Review 3.* *http://www.jstatsoft.org/*

Reviewer: John Maindonald
Australian National University

## Generalized Additive Models: An Introduction with R

Simon N. Wood
Chapman & Hall/CRC, Boca Raton, Florida, 2006.
ISBN 1-58488-474-6. 391+xviii pp. USD 79.95.
http://www.maths.bath.ac.uk/~sw283/

This attractively written advanced level text shows its style by starting with the question "How old is the universe?", plotting data from a 2003 paper from the astronomical literature in the hope of finding an answer. It serves also as a manual for the author's **mgcv** package, which is one of R's recommended packages. Data sets that are used in the exposition are included in the R package **gamair** that is available from the Comprehensive R Archive Network (CRAN).

It is a premise that "most people are interested in statistical models in order to use them, rather than to gaze upon the mathematical beauty of their structure ...". The handling of the substantial number of practical examples reflects the reality that statistical modelling problems "rarely require only straightforward brain-free application of some standard model". There are many insightful comments, on theoretical considerations, on competing numerical methods, on computational cost, on theoretical approximations, and on practical issues in regression modeling. A first appendix is devoted to matrix algebra, while a second appendix gives solutions to exercises. The style and emphasis, and the attention to practical data analysis issues, make this a highly appealing volume.

While Generalized Additive Models (GAMs) are perhaps the main course of a very ample meal, this text offers much else besides. Chapters 1 and 2 (120 pages in all), perhaps forming the appetizer, are taken up with an elegant and remarkably complete account of linear models and generalized linear models. The first 39 pages of the final chapter (Chapter 6) are devoted to linear and generalized linear mixed models. It is perhaps a pity that the title does not make it clear while GAMs may be the main course, this text offers a great deal more than GAMs *per se*. There are suggestions for different paths through the book, a necessary accommodation to the different demands of different readers and to the reality that "Life is too short to spend much of it reading statistics texts".

A GAM is described as a generalized linear model in which part of the linear predictor is specified as a sum of smooth functions of predictor variables. The challenge is to find suitable parametric representations for the smooth functions, and to control and choose the degree of smoothness appropriately. Chapter 3 begins with a brief treatment of regressions splines.

These have the big advantage that the classical theory of linear models is available. Objections to their use are arbitrariness in the choice of knots, and that there are complications in trying to nest models as required for comparisons based on the likelihood ratio text.

Regression splines lead naturally into full-blown GAMs. The methods presented and bases used in the initial discussion in Chapter 3 are chosen to be conducive to simple exposition of the basic framework, rather than for their suitability for practical use. Penalized likelihood maximization is used to control the degree of smoothness. The choice of smoothness is made using some form of cross-validation. In subsequent chapters, AIC or Mallows $C_p$ may be used. Readers are encouraged to work with a computer alongside, alternating between study of the statistical ideas and the use of R to work through examples. Chapter 4 continues the discussion of Chapter 3, giving greater theoretical detail, discussing the smoothers that are proposed for practical use, and discussing the calculation of confidence intervals and $p$-values.

Chapter 5 is devoted to the modeling functions in the **mgcv** package. Additionally, there are brief introductory accounts of two competing R packages – **gss** by Chong Gu, and **gam** by Trevor Hastie. Two helpful tables note advantages and disadvantages of different approaches. The first compares the different bases that are built into **mgcv** – thin plate regression splines, cubic regression splines (both these first two kinds without and with shrinkage), cyclic cubic regression splines, and P-splines. The second compares tensor product with thin plate regression splines, for smoothing with respect to multiple covariates.

The distributional theory for hypothesis tests and confidence intervals is in general approximate, involving theoretical and numerical approximations that leave loose ends untied. Simulations of frequentist intervals for model components were, in the cases investigated, unreliable, though the overall coverage for the whole model usually proved acceptable. Bayesian credible intervals, based on sampling from the posterior, did better. Presumably this is in part because the chosen prior for the joint distribution of the parameters (independent exponentials) is to an extent informative.

The complications are such that there seem to this reviewer clear advantages, when regression splines seem to meet the case, in working within this framework. It would be interesting to try to identify the examples, from among those presented in this text, where regression splines seem inadequate.

I strongly recommend this book.

### Reviewer:

John Maindonald
Australian National University
Centre for Mathematics and Its Applications
Canberra, ACT 0200, Australia
E-mail: john.maindonald@anu.edu.au
URL: http://www.maths.anu.edu.au/~johnm/

*Journal of Statistical Software* http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 16, Book Review 3 *Published:* 2006-07-12
July 2006