Journal of Statistical Software

June 2018, Volume 85, Book Review 2. *doi: 10.18637/jss.v085.b02*

Reviewer: Xavier Barber
Universitas Miguel Hernández

Flexible Regression and Smoothing: Using GAMLSS in R

Mikis D. Stasinopoulos, Robert A. Rigby, Gilllian Z. Heller, Vlasios Voudouris, Fernanda De Bastiani
Chapman & Hall/CRC, Boca Raton, 2017.
ISBN 978-1-138-19790-9. 549 pp. USD 99.95 (P).
https://www.crcpress.com/9781138197909

The GAMLSS book

On reading the first lines of the preface, the reader gets a clear picture of the book; it is about modeling and learning using GAMLSS: generalized additive models for location, scale, and shape. GAMLSS is a modern distribution-based approach to regression analysis that expands the traditional approach, which is focused on the mean, to accommodate distribution parameters that are modeled as additive functions of predictor variables.

The book is divided into six parts: Introduction to models and packages, Algorithms, Functions and inference, Distributions, Additive terms, Model selection and diagnostics, and Applications.

I like how the authors have presented and explained the methodology. However, I recognize that some readers who appreciate more mathematical details may be left waiting for more; this type of reader should look forward to the next two books that the authors are currently writing. The second book, titled "Distributions for Location Scale and Shape: Using GAMLSS in R", is being prepared. The third book, titled "Generalised Additive Models for Location Scale and Shape: A Distributional Regression Approach" is in the draft stage and will be available soon (the authors state on their webpage). Yet, this first book provides a broad overview of the GAMLSS methodology and will be useful for anyone who wants to use the **gamlss** package in R.

Introduction to models and packages

I appreciate the title of the first chapter: "*Why GAMLSS?*" When you finish reading this chapter, you already know whether you should continue, because the first and second chapters present the **gamlss** package along with only a few concepts regarding GAMLSS and their

wide range of applications. These chapters emphasize that GAMLSS is compatible with any parametric distribution for the response variable and allows all parameters (location, scale, and shape) of the distribution to be modeled as linear or smooth functions of the explanatory variables.

A demonstration of the linear regression, generalized linear regression, additive linear regression, and GAMLSS regression on the same dataset provides a good way to understand what a flexible regression could contribute to other analyses. In a similar way, the second chapter provides a comparison of a parametric regression to a non-parametric regression with the application of P-splines, cubic splines, locally weighted scatterplot smoother, and neural networks. Only details that can be easily understood are presented using `gamlss()` with real data examples, integrated code, and figures to illustrate the methods; there is no mathematical notation (this is going to be presented in subsequent chapters).

Algorithms, functions, and inference

Once I was certain that my dataset would benefit from a GAMLSS approach, I was interested in how it works. In this section the authors explain the difference between two algorithms in order to adjust the model. The algorithms are presented in flow diagrams (along with a few mathematical equations) to help readers understand, at the undergraduate level, how the location, shape, and scale parameters are obtained.

Chapter 3 talks about algorithms that the `gamlss` function uses. The authors present the difference between two inference algorithms: the `method = RS()` algorithm (default option) as a generalization of the algorithm presented in Rigby and Stasinopoulos (1996) or the `method = CG()` as a generalization of Cole and Green (1992). At the end of this chapter, some notes refer the reader to Rigby and Stasinopoulos 2005, Appendix A.1 where the authors explain, with mathematical expressions, the GAMLSS algorithms based on maximum (penalized) likelihood.

The description of the `gamlss()` function is presented in Chapter 4, with very informative comments. Subsections 1 to 3 focus on the function and Subsection 4 provides details about the `gamlss` object. Finally, Subsection 5 describes the *methods and functions associated with these objects*; this is very useful because in some cases the user may need to extract some information in a loop but there is no need to program these functions.

The next chapter, "*Inference and Prediction*", continues in line with the package explanation. Given the title, one might expect a more mathematical discussion but as mentioned above, this is not the main goal of this book. This chapter provides information about tools for the inferential process (via likelihood-based and bootstrapping methods) and how the confidence intervals and predictions work within the `gamlss` package, with references to some theoretical properties of GLM and GAMLSS.

Distributions

Chapter 6 shows the different types of distribution available in GAMLSS and a subsection for deciding whether complex distributions are needed. The **gamlss.dist** package admits more than one hundred continuous and discrete distributions. Moreover, there is an option to extend the GAMLSS family distributions (`gamlss.family`) in order to obtain a classical model:

log and *logit*, *truncating*, *censored*, and *finite mixtures* which are described in detail in Chapter 7. I appreciate that the authors explain the code that was used for obtaining parameters estimates and for the creation of an *own* link function and the underlying mathematics.

The next chapter explains finite-mixture distributions, and their syntax with `gamlssMX()` function is described. Finally, as this is a book about using R, the chapter finishes with a good example that the reader can follow, including a code that can be reproduced. At the end of this chapter, the authors present finite mixtures with parameters in common, i.e. parameter sets are not disjoint, and how to use these with the `gamlssNP()` function.

The authors provide additional information, in the http://www.gamlss.com/ webpage, regarding how to use mixed distributions in two vignettes: inflated distributions in the interval of [0, 1] and zero-adjusted distributions on the positive real number line.

Model terms

The largest section of the book covers model terms and includes three chapters describing linear parametric additive terms, additive smoothing terms, and random effects. Figure 8.1 on page 226 provides a very good summary of the classification of various additive terms within the GAMLSS models. A wide range of model terms are provided, which can be proved by the GAMLSS regression:

- **Linear**: Polynomials, fractional polynomials, piecewise polynomials, and B-splines.
- **Penalized**: P-splines, cubic splines, tensor products, thin plate splines, and Gaussian Markov random fields.
- **Others**: Loess, neural networks, decision trees, and multivariate adaptive regression splines.

This tree, encompassing Chapters 8, 9, and 10, presents a real example for each of the model terms, in some cases using the same data from different approaches.

There is a section dedicated to penalized smoothers with a univariate approach and another for the multivariate approach in Chapter 9. At the end of this chapter, a few ideas appear about machine learning methods such as neural networks and decision trees. In the era of big data, I think this content is very important.

Chapter 10 explains how random effects models can be used within GAMLSS. For example, *random intercept*, *random intercept and slopes*, *multilevel modeling*, and *repeated measurements* can be modeled with the join o the marginal likelihood.

After these ten chapters, the reader should be able to adjust a wide range of models by defining the distribution of the response variable, the inference algorithm, and the model terms. If a particular dataset needs a GAMLSS, the reader should be equipped with the capacity to adjust the model by this point.

Model selection and diagnostics

In this section the authors show two chapters related to the *goodness of fit* of a model, one for *model selection* and one for *diagnostics*.

The authors present four components for model selection: $\mathcal{D}, \mathcal{G}, \mathcal{T}$, and $\mathcal{L}$ for distributions, *link functions*, *additive terms*, and *smoothing parameters*, respectively. Alternatively, the authors propose model selection using a *validation data* method. The `addterm()` and `dropterm()` have parallel arguments that can be used for parallel computation (this is very interesting for high-dimensional datasets). Finally, the `stepGAIC()` function can be used to build a model for any of the distribution parameters based on a *forward*, *backward*, or *stepwise* procedure using the generalized Akaike information criterion. The authors recommend that if a dataset contains 5000 or more observations that it be split into (i) a training dataset, (ii) a validation dataset, and (iii) a testing data set.

The residuals of the fitted models for the GAMLSS models are the tools used using the normalized quantile residuals for continuous response variables and randomized normalized quantile residuals for discrete response variables.

The authors explain the main advantage of normalized quantile residuals, i.e., whatever the distribution of the response variable, the true residuals always have a standard normal distribution given that the assumptions of the model are correct; this is well established in the literature. The normalized (randomized) quantile residuals provide an easy way to check the adequacy of a GAMLSS fitted model (the reader can find several references at the end of this chapter). Several functions in the package were implemented, but I particularly like the name of one of them: the *worm* plot, which was introduced by Van Buuren and Fredriks (2001) to identify the regions of an explanatory variable within which the model does not adequately fit the data. Other functions described in this chapter are the *detrended Own's* plot function, the *Q-statistics* function, and the randomized residual plot function.

This section is one of the few that is quite short, and I would have liked a more thorough discussion. However, this is a personal preference. The procedures for the diagnosis and validation of models should be more popular among non-statisticians. In fact, the process described in applied papers are very rarely validated.

Applications

At the end of the first section of the preface, the authors note "*This book is written for ...*" followed by, firstly, "*the practitioners*" and secondly, "*students*". Both practitioners and students require an "Applications" section to demonstrate an example from the beginning to the end. The authors present in these two chapters specific applications where the GAMLSS methodology works fine: overdispersed count and binomial data, where the authors have shown the flexibility of the methodology, and especially, the advantage of GAMLSS to estimate the centile curves.

For the Centile estimation example (Chapter 13), the authors state on their webpage:

> *"The LMS (lambda, mu, sigma) method and its extensions are a subclass of GAMLSS. The LMS method within the* `gamlss()` *function is equivalent to assuming the Box-Cox Cole and Green distribution (BCCG) for the response variable. The BCCG distribution is suitable for both positively or negatively skew data but does cope with kurtotic data".*

Rigby and Stasinopoulos (2004) extended the LMS method by introducing the Box-Cox power exponential distribution in order to solve this kurtotic data problem within GAMLSS. This

approach implies that this would not be the expected *application* chapter but rather another procedure and its package functions.

Chapter 14 provides three real applications that would be useful to a student or practitioner. Unfortunately, some of these people may only want a *template* to publish; I would hope that such readers would read the entire book and not only this last chapter.

Conclusion

After reading the book, I am already waiting for Volumes 2 and 3 to be published. I hope they will delve deeper into some aspects and describe new advances that could be developed; and I'd like a more mathematical discussion in the "*Inference and prediction*" future chapters.

I think that the examples used in this book show how flexible GAMLSS is, considering that this book is about "using R". A great point of the book are the **Bibliographic notes** that appear in all the chapters, because they justify the importance of the cited references in four to five lines. I also like the last chapter and its epilogue.

References

Cole TJ, Green PJ (1992). "Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood." *Statistics in Medicine*, **11**(10), 1305–1319. doi:10.1002/sim.4780111005.

Rigby RA, Stasinopoulos DM (1996). "A Semi-Parametric Additive Model for Variance Heterogeneity." *Statistics and Computing*, **6**(1), 57–65. doi:10.1007/bf00161574.

Rigby RA, Stasinopoulos DM (2004). "Smooth Centile Curves for Skew and Kurtotic Data Modelled Using the Box-Cox Power Exponential Distribution." *Statistics in Medicine*, **23**(19), 3053–3076. doi:10.1002/sim.1861.

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society C*, **54**(3), 507–554. doi:10.1111/j.1467-9876.2005.00510.x.

Van Buuren S, Fredriks M (2001). "Worm Plot: a Simple Diagnostic Device for Modelling Growth Reference Curves." *Statistics in Medicine*, **20**(8), 1259–1277. doi:10.1002/sim.746.

Reviewer:

Xavier Barber
Universitas Miguel Hernández

Center for Operations Research
Elche, Spain 03202
E-mail: xbarber@umh.es
URL: http://cio.umh.es/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
June 2018, Volume 85, Book Review 2 *Published:* 2018-06-10
doi:10.18637/jss.v085.b02