Generalized Additive Models for Location, Scale and Shape

An emerging field in statistics, distributional regression facilitates the modeling of the complete conditional distribution, rather than just the mean. This book introduces generalized additive models for location, scale and shape (GAMLSS) – one of the most important classes of distributional regression. Taking a broad perspective, the authors consider penalized likelihood inference, Bayesian inference, and boosting as potential ways of estimating models, and illustrate their usage in complex applications.

Written by the international team who developed GAMLSS, the text's focus on practical questions and problems sets it apart. Case studies demonstrate how researchers in statistics and other data-rich disciplines can use the model in their work, exploring examples ranging from fetal ultrasounds to social media performance metrics. The R code and datasets for the case studies are available on the book's companion website, allowing for replication and further study.

MIKIS D. STASINOPOULOS is Professor of Statistics at the School of Computing and Mathematical Sciences, University of Greenwich. He is, together with Professor Bob Rigby, coauthor of the original Royal Statistical Society article on GAMLSS. He has also coauthored three books on distributional regression, and in particular the theoretical and computational aspects of the GAMLSS framework.

THOMAS KNEIB is a Professor of Statistics at the University of Göttingen, where he is the Spokesperson of the interdisciplinary Centre for Statistics and Vice-Spokesperson of the Campus Institute Data Science. His main research interests include semiparametric regression, spatial statistics, and distributional regression.

NADJA KLEIN is Emmy Noether Research Group Leader in Statistics and Data Science and Professor for Uncertainty Quantification and Statistical Learning at TU Dortmund University and the Research Center Trustworthy Data Science and Security. Nadja is member of the Junge Akademie and associate editor for *Biometrics*, *JABES*, and *Dependence Modeling*. Her research interests include Bayesian methods, statistical and machine learning, and spatial statistics.

ANDREAS MAYR is a Professor at the Institute for Medical Biometry, Informatics, and Epidemiology at the University of Bonn. He has authored more than 100 research articles both in statistics as well as medical research and is currently Editor of the *Statistical Modelling Journal*, Associate Editor of the *International Journal of Biostatistics*, and Editorial Board Member of the *International Journal of Eating Disorders*.

GILLIAN Z. HELLER is Professor of Biostatistics at the NHMRC Clinical Trials Centre, University of Sydney. She has coauthored four books in the regression modeling area, the first directed towards actuarial applications of the generalized linear model, and the remaining three focusing on distributional regression, in particular the GAMLSS framework.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

A complete list of books in the series can be found at www.cambridge.org/statistics.
Recent titles include the following:

41. *Confidence, Likelihood, Probability*, by Tore Schweder and Nils Lid Hjort
42. *Probability on Trees and Networks*, by Russell Lyons and Yuval Peres
43. *Random Graphs and Complex Networks* (*Volume 1*), by Remco van der Hofstad
44. *Fundamentals of Nonparametric Bayesian Inference*, by Subhashis Ghosal and Aad van der Vaart
45. *Long-Range Dependence and Self-Similarity*, by Vladas Pipiras and Murad S. Taqqu
46. *Predictive Statistics*, by Bertrand S. Clarke and Jennifer L. Clarke
47. *High-Dimensional Probability*, by Roman Vershynin
48. *High-Dimensional Statistics*, by Martin J. Wainwright
49. *Probability: Theory and Examples* (*Fifth Edition*), by Rick Durrett
50. *Model-based Clustering and Classification for Data Science*, by Charles Bouveyron *et al.*
51. *Spectral Analysis for Univariate Time Series*, by Donald B. Percival and Andrew T. Walden
52. *Statistical Hypothesis Testing in Context*, by Michael P. Fay and Erica Brittain
53. *The Fundamentals of Heavy Tails*, by Jayakrishnan Nair, Adam Wierman and Bert Zwart
54. *Random Graphs and Complex Networks* (*Volume 2*), by Remco van der Hofstad
55. *Modern Discrete Probability*, by Sébastien Roch
56. *Generalized Additive Models for Location, Scale and Shape*, by Mikis D. Stasinopoulos *et al.*

Generalized Additive Models for Location, Scale and Shape

A Distributional Regression Approach, with Applications

Mikis D. Stasinopoulos
University of Greenwich

Thomas Kneib
University of Göttingen

Nadja Klein
TU Dortmund University

Andreas Mayr
University of Bonn

Gillian Z. Heller
University of Sydney

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009410069

DOI: 10.1017/9781009410076

First published 2024
First hardback edition 2024

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Names: Stasinopoulos, Mikis D., author. | Kneib, Thomas, author. | Klein, Nadja, 1987- author. | Mayr, Andreas, 1983- author. | Heller, Gillian Z., author.
Title: Generalized additive models for location, scale and shape : a distributional regression approach, with applications / Mikis D. Stasinopoulos, Thomas Kneib, Nadja Klein, Andreas Mayr, Gillian Z. Heller.
Description: First edition. | Cambridge ; New York, NY : Cambridge University Press, 2024. | Series: Cambridge series in statistical and probabilistic mathematics 56 | Includes bibliographical references and index.
Identifiers: LCCN 2023035047 (print) | LCCN 2023035048 (ebook) | ISBN 9781009410069 (hardback) | ISBN 9781009410076 (epub)
Subjects: LCSH: Regression analysis–Mathematical models. | Theory of distributions (Functional analysis) Classification: LCC QA278.2 .S674 2024 (print) | LCC QA278.2 (ebook) | DDC 519.5/36–dc23/eng/20231016
LC record available at https://lccn.loc.gov/2023035047
LC ebook record available at https://lccn.loc.gov/2023035048

ISBN 978-1-009-41006-9 Hardback

During the course of writing this book, we lost two people close to us.

Our dear friend and colleague
Professor Brian D. Marx (1960–2021)

and

María Belén Avila (1983–2020)

We dedicate this book to their memories.

Contents

Preface

What this Book is About

This book is devoted to a special class of regression models that takes a *distributional* perspective on regression modeling. Instead of focusing on the expectation of the response variable, as most classical regression approaches such as linear models, generalized linear models, and generalized additive models do, we deal with regression models that relate more general features of the response distribution to covariates. This yields characterizations of, for example, location, scale, and shape of the response distribution conditional on covariate information. Since the models also involve flexible forms of regression modeling based on a variety of covariate types, including nonlinear effects of continuous covariates, spatial effects, and random effects, the model class is called *generalized additive models for location, scale and shape (GAMLSS)*.

More precisely, GAMLSS builds upon the classical framework of generalized linear models but

- relaxes the assumption that the response distribution belongs to the exponential family such that any parametric distribution can be assumed for the response, and
- specifies separate regression predictors for any parameter of this distribution, in particular parameters determining scale and shape features.

The assumption of a parametric response distribution implies the immediate availability of the likelihood such that inferential procedures based on the likelihood can be used for inference. In this book, we will consider (penalized) maximum likelihood inference, Bayesian inference, and functional gradient descent boosting as specific algorithmic approaches for implementing inference in GAMLSS.

The main contribution of the GAMLSS framework is that it challenges the classical notion of modeling only the mean of the distribution of the response variable. Rather, all parameters (including location, scale, and shape) of that distribution may be modeled. This allows the analyst to approach a much wider range of problems, including heterogeneity in variance, positive or negative skewness, platykurtic or leptokurtic response distributions, heavy tails and extremes, overdispersion and underdispersion in count data, excess or shortage of certain parts of the support (e.g. zero-inflated count data), and multivariate responses. Importantly, the aspects *beyond the mean* are not treated as a nuisance but rather can be central to the analysis of interest. There are other approaches to distributional regression such as quantile regression or conditional transformation models that will not be considered here, but we provide a brief comparison in Section 1.6.

Types of Applications and Questions that Can be Approached with GAMLSS

GAMLSS can be useful in any area that relies on regression modeling for the generation of insights from observed data. This includes *exploratory* as well as *confirmatory* analyses (depending on the research question, data collection strategy, and exact model specification). Moreover, the goals of the analysis may range from *interpreting* the resulting model coefficients to pure *prediction-oriented* applications. For the latter, GAMLSS have the specific advantage of providing covariate-dependent predictive distributions rather than only a point prediction.

We illustrate the applicability of GAMLSS in various kinds of research questions in case studies in Part III. More precisely, we consider the following applications:

- fetal ultrasound (Chapter 8) where we are considering the prediction of birthweight from a number of covariates derived from ultrasound measurements to illustrate basic components and steps of GAMLSS analyses;
- speech intelligibility testing (Chapter 9) where a mixed discrete–bounded continuous response is considered in combination with random effects to adjust for a repeated measurements design;
- social media post performance (Chapter 10) to illustrate the consideration of overdispersed count data regression involving cyclic P-splines and a comparison of different inferential approaches for prediction;
- childhood undernutrition in India (Chapter 11) as a case study on the Bayesian approach to GAMLSS and its advantages including the extension to bivariate responses and the consideration of spatial effects;
- federal election outcomes in Germany (Chapter 12) as another case of Bayesian inference, dealing with multivariate fractional responses as well as nonlinear and spatial effects;
- riboflavin production (Chapter 13) as a case study on high-dimensional regression with variable selection via boosting.

Goal of this Book

With this book, we provide a comprehensive introduction to the concepts of GAMLSS that serves applied scientists interested in conducting data analyses with GAMLSS and statisticians interested in the statistical foundations and background of GAMLSS. Rather than focusing on one specific way of conducting statistical inference, we aim at a unified treatment of different inferential approaches, including applications highlighting their specific advantages and disadvantages. For this, we draw on the ample experience of the authors in developing and applying penalized likelihood approaches, Bayesian inference, and functional gradient descent boosting. While the book includes a solid introduction to the foundations of the statistical inference approaches, our emphasis is always on the practical applications rather than a theoretical dogma. This is supported by devoting a considerable part of the book to hands-on examples in terms of complex applications that illustrate the dos and don'ts of GAMLSS modeling.

Readership of the Book and Assumed Background Knowledge

This book is written for:

- practitioners and applied researchers interested in understanding modern regression approaches and applying them to their own datasets,
- applied statisticians aiming at a better understanding of the relevance and application of GAMLSS as well as the different inferential approaches for GAMLSS,
- data analysts who are interested in models emphasizing interpretability over pure prediction power, and
- students in statistics and data science who wish to go beyond basic forms of regression modeling.

We assume that readers are familiar with the basic concepts of standard regression analysis (including the corresponding matrix algebra) as presented in, for example, Fahrmeir et al. (2021). All computations underlying the examples presented in the book have been conducted in the **R** environment for statistical computing. **R** commands are not provided in this text, but are available in the online supplementary material of the book published on `https://gamlssbook.bitbucket.io`.

Structure of the Book and Additional Resources

The book consists of three distinct parts.

- In Part I (Chapters 1 to 3), we introduce the basic concepts and history of GAMLSS and review some of the relevant ingredients for setting up a GAMLSS model, that is, different types of response distributions, and the various regression effects that can appear in the predictor specifications.
- Part II (Chapters 4 to 7) explains the theoretical background of the different inferential procedures that can be used to estimate GAMLSS models. This part comprises dedicated chapters on penalized likelihood inference, Bayesian inference, and functional gradient descent boosting.
- In Part III (Chapters 8 to 13), we present several practical examples and case

studies to illustrate the different inferential approaches as well as various aspects relevant in GAMLSS modeling.

Depending on the reader's background and interest, there are different ways of reading the book. While Chapter 1 serves as a good starting point for any reader, it is not necessary to work through the rest of the book sequentially. Rather, it is possible to pick the inferential approach that appears most relevant, first, or to start with reading some of the case studies.

This book is the third in a series of texts on GAMLSS. The first, *Flexible Regression and Smoothing: Using GAMLSS in R* (Stasinopoulos et al., 2017), concerns the implementation of GAMLSS in the suite of **R** packages **gamlss**, **gamlss.dist**, **gamlss.add**, etc. The second, *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R* (Rigby et al., 2019), describes the more than 100 continuous, discrete, and mixed distributions available in the package **gamlss.dist**, and also shows how more distributions can be generated by transformation, truncation, censoring, and zero-inflation. In contrast, this book is less associated with a specific group of **R** packages but serves as a general introduction to GAMLSS. The first two books are therefore not assumed knowledge, but they may nonetheless be useful as resources for distributions and details on the usage of the penalized likelihood **R** packages from the **gamlss** suite.

Additional resources for this book have been collected on the website `https://gamlssbook.bitbucket.io`. These comprise code and datasets for the case studies presented in the book, further supporting material and a list of errata.

Acknowledgments

We are indebted to Professor Bob Rigby, who was unable to participate in the authorship of this book, for his substantial intellectual contribution and support. We thank Stephan Klasen (1966–2020) for joint research on childhood malnutrition and distributional modeling thereof that provided the foundation for the application in Chapter 11; Stefan Lang and Nikolaus Umlauf, with whom we have collaborated on Bayesian inference for GAMLSS and corresponding statistical software; Marah-Lisanne Thormann for assistance with graphics in Chapters 3 and 6; Benjamin Hofner, Nora Fenske, and Matthias Schmid for joint research on boosting GAMLSS; Janek Thomas and Tobias Hepp for further enhancements; Jost von Petersdorff-Campen and Maarten Jung for assistance with code and graphics in Chapters 11 and 12; Annika Strömer for her help with the genetic application in Chapter 13; Tobias Wistuba for his help with graphics in Chapter 7; Stanislaus Stadlmann for his help with graphics in Chapter 10; Wenli Hu and Brett Swanson for joint research on models for speech intelligibility that provided the basis for Chapter 9; Fernanda de Bastiani for her contribution to the bucket plots; and the production team at Cambridge University Press for their great support throughout the production process of this book. Gillian Heller and Nadja Klein thank their previous employers, Macquarie University and Humboldt-Universität zu Berlin, respectively, for their support during the book-writing period. Finally, many of our productive collaborations were initiated and fostered at the annual meetings of the Statistical Modelling Society, the International Workshop on Statistical Modelling. To the IWSM community: Thank you for the very supportive research environment, and many useful discussions.

Notation and Terminology

Packages and code	
GAMLSS	the statistical model
gamlss	the **R** package
`gamlss()`	the **R** function
General mathematical notation	
$X^\top$	vector or matrix transpose
$f'(x)$	derivative of a function f
$\mathbb{1}(y = x)$	indicator function
$\mathbf{1}_n$	$(1, \ldots, 1)^\top$
$\boldsymbol{I}_n$	$n \times n$ identity matrix
Distributions	
y	a univariate response variable, or a single realization of the response variable
$\boldsymbol{y}$	the vector of observed values of the response variable y, namely $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$
$\boldsymbol{x}_i$	covariates (features)
n	sample size
$\mathbb{P}(\cdot)$	probability
$f(y)$	probability function; probability mass function for discrete random variables and probability density function for continuous random variables; occasionally, for clarity, $f_y(y)$
$F(y)$	cumulative distribution function (cdf)
y_p	$F^{-1}(p)$, inverse cdf (quantile function)
$\mathcal{D}(\cdot)$	generic distribution
$\mathcal{E}(\cdot)$	exponential family distribution
$\mathcal{S}$	support of y
$\mathcal{N}(\mu, \sigma^2)$	(univariate) normal distribution with mean μ and variance σ^2
$\phi(\cdot)$	probability density function of $\mathcal{N}(0, 1)$ (the standard normal)
$\Phi(\cdot)$	cumulative distribution function of the standard normal
$\mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$	k-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix Σ
$\mathcal{U}(a, b)$	uniform distribution on the interval (a, b)

Intervals	
$\mathbb{R}$	real line
$\mathbb{R}_+$	positive real line
$\mathbb{R}_{(0,1)}$	unit interval $(0, 1)$
$\mathbb{N}$	nonnegative integers $(0, 1, 2, \ldots)$
$\mathbb{N}_+$	positive integers $(1, 2, \ldots)$
Distribution parameters	
K	number of distribution parameters
θ_k	kth distribution parameter, $k = 1, \ldots, K$
$\boldsymbol{\theta}$	vector of distribution parameters $(\theta_1, \ldots, \theta_K)^\top$ in GAMLSS
θ_1	first distribution parameter, sometimes denoted as μ
θ_2	second distribution parameter, sometimes denoted as σ
θ_3	third distribution parameter, sometimes denoted as ν
θ_4	fourth distribution parameter, sometimes denoted as τ
λ	a hyperparameter
$\boldsymbol{\lambda}$	vector of all hyperparameters in the model
$\boldsymbol{\vartheta}$	vector of all parameters in the model
Systematic part of the GAMLSS model	
$\boldsymbol{X}_k$	$n \times p_k$ fixed effects design matrix for θ_k
p_k	number of columns in the design matrix $\boldsymbol{X}_k$ for the kth parameter
J_k	total number of smoothers for θ_k
q_{kj}	dimension of the random effect vector $\boldsymbol{\gamma}_{kj}$
$\boldsymbol{x}_{kj}$	jth explanatory variable vector for θ_k
$\boldsymbol{\beta}_k$	vector of fixed effect coefficients of length J_K
$\boldsymbol{z}_{kj}$	jth random effects explanatory variable vector
$\boldsymbol{Z}_{kj}$	$n \times q_{kj}$ random effect design matrix
$\boldsymbol{\alpha}_{kj}$	jth random effect coefficient vector of length q_{kj}
$\boldsymbol{B}_{jk}$	$n \times L_{jk}$ generic design matrix for θ_k
$\boldsymbol{\gamma}_k$	vector of basis coefficients for this effect
$\boldsymbol{\eta}_k$	predictor for θ_k
$g_k(\cdot)$	link function for θ_k
$s_{kj}(\boldsymbol{x}_{kj})$	jth nonparametric or nonlinear function in $\boldsymbol{\eta}_k$
$\boldsymbol{W}$	$n \times n$ diagonal matrix of weights
$\boldsymbol{w}$	n-dimensional vector of weights (the diagonal elements of $\boldsymbol{W}$)
$\boldsymbol{K}_{kj}$	jth smoothing or penalty matrix for θ_k
Likelihood and information criteria	
L	likelihood function
ℓ	log-likelihood function
$\mathcal{I}$	Fisher's expected information matrix
$\mathcal{H}$	observed information matrix
df	total effective degrees of freedom used in the model
κ	penalty for each degree of freedom used in the model
GDEV	global deviance = minus twice the fitted log-likelihood

AIC	Akaike information criterion = GDEV + 2 × df
BIC	Bayesian information criterion = GDEV + $\log(n)$ × df
GAIC	generalized AIC = GDEV + κ × df
Residuals	
$\boldsymbol{u}$	vector of (randomized) quantile residuals
$\boldsymbol{r}$	vector of normalized (randomized) quantile residuals
$\boldsymbol{\varepsilon}$	vector of (partial) residuals
Moment measures and functions	
$\mathbb{E}(y)$	expected value (or mean) of random variable y
$\mathbb{V}(y)$	variance of random variable y
$\mathrm{Cov}(\boldsymbol{y})$	variance–covariance matrix of the random vector $\boldsymbol{y}$
μ_y	location parameter (usually mean)
$\tilde{\mu}_k$	$\mathbb{E}(y^k)$, kth population moment about zero
μ_k	$\mathbb{E}\{[y - \mathbb{E}(y)]^k\}$, kth population central moment
γ_1	moment skewness $= \mu_3/\mu_2^{1.5}$
β_2	moment kurtosis $= \mu_4/\mu_2^2$
γ_2	moment excess kurtosis $= \beta_2 - 3$
Boosting notation	
$\rho(y_i, \eta(x_i))$	loss function evaluated at observation i
$h(\cdot)$	base-learner
$h_{jk}^{[m]}(\cdot)$	base-learner at boosting iteration m for explanatory variable j in parameter θ_k