

<i>Preface</i>	xiii
<i>Acknowledgments</i>	xvii
<i>Notation and Terminology</i>	xviii
Part I Introduction and Basics	1
1 Distributional Regression Models	3
1.1 The Data Analysis Circle	3
1.2 Statistical Models	6
1.3 Regression Models	7
1.4 From Linear Models to GAMLSS	8
1.4.1 The Linear Model	8
1.4.2 Generalized Linear Models	12
1.4.3 Generalized Additive Mixed Models	13
1.4.4 Mean and Dispersion Models	16
1.5 Generalized Additive Models for Location, Scale and Shape	17
1.5.1 GAMLSS as a Distributional Regression Model	17
1.5.2 Response Distributions	18
1.5.3 Link Functions	20
1.5.4 Structured Additive Predictors	21
1.5.5 Basis Function Representation	22
1.5.6 Compact Summary	23
1.6 Other Distributional Regression Approaches	24
1.6.1 Quantile Regression	24
1.6.2 Conditional Transformation Models	28
 2 Distributions	 30
2.1 Some Terminology and Notation	31
2.2 Continuous Distributions	31
2.2.1 Continuous Distributions on \mathbb{R}	32
2.2.2 Continuous Distributions on \mathbb{R}_+	35
2.2.3 Continuous Distributions on $\mathbb{R}_{(0,1)}$	38
2.2.4 Mixed Distributions	39
2.2.5 Inflated Distributions on $\mathbb{R}_{(0,1)}$	40
2.3 Discrete Distributions	41
2.3.1 Unbounded Count Distributions: $\mathcal{S} = \{0, 1, 2, \dots\}$	41

2.3.2	Bounded Count Distributions: $\mathcal{S} = \{0, 1, \dots, n\}$	45
2.4	Multivariate Continuous Distributions	46
2.4.1	Bivariate Normal Distribution	46
2.4.2	Dirichlet Distribution	47
2.4.3	Copula Models	48
2.5	Generated Distributions	49
2.6	Choice of Response Distribution	51
3	Additive Model Terms	53
3.1	Penalized Splines	53
3.2	Generic Representation	60
3.3	Tensor Product Penalized Splines	63
3.4	Radial Basis Functions and Geostatistical Smoothing	65
3.5	Markov Random Fields	67
3.6	Random Effects	69
3.7	Varying Coefficient Terms	71
3.8	Other Effect Types	71
3.8.1	Parametric Linear Effects	71
3.8.2	Regularized Linear Effects	72
3.8.3	Functional Effects	72
	Part II Statistical Inference in GAMLSS	75
4	Inferential Methods for Distributional Regression	77
4.1	Introduction	77
4.2	The Population, the Sample and the Model	78
4.3	Risk, Empirical Risk, and the Likelihood Function	83
4.4	Model Comparison with GAMLSS	85
4.4.1	Information Criteria for Evaluating and Comparing Model Fits	86
4.4.2	Predictive Accuracy of Models	88
4.4.3	Cross-validation	92
4.5	Model Selection Strategies for GAMLSS	92
4.5.1	Selection of a Distribution in GAMLSS	93
4.5.2	Selection of Covariates and Terms in GAMLSS	95
4.5.3	A Recipe for Model Selection	97
4.6	Parameter Orthogonality	98
4.7	Diagnostic Tools for GAMLSS	101
4.7.1	Normalized Quantile Residuals	102
4.7.2	Worm Plots	103
4.7.3	Bucket Plots	104
4.8	Visualizing and Interpreting GAMLSS	107
5	Penalized Maximum Likelihood Inference	109
5.1	Introduction	109
5.2	Fitting GAMLSS using Penalized Maximum Likelihood	110
5.2.1	Estimating the β and γ Coefficients	112
5.2.2	Estimating the Hyperparameters λ	115

5.3	Inference within Penalized Maximum Likelihood Estimation	116
5.3.1	Inference for the Coefficients β	116
5.3.2	Profile Likelihood	118
5.3.3	Inference Based on Data Reuse Techniques	119
5.4	Covariate Selection via Shrinkage and Regularization	122
6	Bayesian Inference	125
6.1	A Brief Introduction to Bayesian Inference	125
6.2	Priors for the Regression Coefficients in GAMLSS	129
6.3	Posterior Mode vs. Penalized Likelihood	130
6.4	The General Idea of Markov Chain Monte Carlo Simulations	132
6.4.1	The Basic Setup	132
6.4.2	A Generic MCMC Algorithm	135
6.4.3	Monitoring Mixing and Convergence	140
6.5	MCMC for GAMLSS	143
6.5.1	Iteratively Weighted Least Squares Proposals	143
6.5.2	Sampling Subject to Linear Constraints	147
6.5.3	Auxiliary Variable Approaches	150
6.6	Hierarchical Predictor Specifications	151
6.7	Hyperpriors for the Smoothing Variances	152
6.7.1	Inverse-Gamma-Type Priors	152
6.7.2	Scale-Dependent Priors	154
6.7.3	Further Prior Choices for the Smoothing Variances	155
6.7.4	Effect Selection Priors	156
6.8	Posterior Inference in GAMLSS	157
6.8.1	Pointwise Credible Intervals and Simultaneous Credible Bands	157
6.8.2	MCMC-Based Inference for Parameter Transformations	159
6.8.3	Model Choice Based on Information Criteria	160
6.8.4	Posterior Predictive Checks	162
6.9	Other Bayesian Approaches	163
7	Statistical Boosting for GAMLSS	165
7.1	Boosting in Machine Learning	166
7.2	Statistical Boosting	170
7.2.1	Base-learners for Statistical Boosting	170
7.2.2	Loss Functions for Statistical Boosting	171
7.2.3	The Algorithm	172
7.2.4	Tuning of the Algorithm	174
7.2.5	Properties of Statistical Boosting in Practice and Current Developments	175
7.2.6	Implementation in R	179
7.3	Boosting for Distributional Regression	183
7.3.1	The Algorithm for Boosting Distributional Regression	184
7.3.2	Implementation in R	186
7.3.3	Model Tuning	187
7.3.4	Post Selection Inference and Stability Selection	188
7.3.5	Diagnostics and Selecting the Most Suitable Distribution	191

Part III	Applications and Case Studies	195
8	Fetal Ultrasound	197
8.1	Data and Research Question	197
8.2	Model Building	197
8.3	Hypothesis Testing	202
8.4	Variable Selection	205
8.5	Selection of Distribution	208
8.6	The Partial Effects	211
8.6.1	Partial Effects on the Predictors η_θ	212
8.6.2	Partial Effects on the Distribution Parameters θ_k	213
8.6.3	Partial Effects on the Moments	216
8.6.4	Partial Effects on the Quantiles y_p	218
8.6.5	Partial Effects on the Distribution $\mathcal{D}()$	219
8.6.6	Prediction	221
9	Speech Intelligibility Testing	222
9.1	Data and Research Question	222
9.2	Model Building	223
9.2.1	Response Distribution	223
9.2.2	Variable Selection	225
9.3	Model Interpretation	226
10	Social Media Post Performance	230
10.1	Data and Research Question	230
10.2	Model Building	230
10.3	Prediction	233
10.4	Model Interpretation	234
10.4.1	Partial Effects on the Distribution Parameters	235
10.4.2	Partial Effects on the Distribution $\mathcal{D}()$	235
11	Childhood Undernutrition in India	239
11.1	Data and Research Question	239
11.2	Univariate Analysis	240
11.2.1	Model Specification	240
11.2.2	MCMC Diagnostics	241
11.2.3	Model Comparison	242
11.2.4	Regression Effects and Hyperprior Sensitivity	243
11.3	Bivariate Analysis	245
11.3.1	Model Specification	246
11.3.2	Estimated Contour Lines and Model Comparison	247
11.3.3	Effects on the Dependence	250
11.4	Summary	252
12	Socioeconomic Determinants of Federal Election Outcomes in Germany	253
12.1	Data and Research Question	253
12.2	Estimation Results	254
12.3	Scenario Analysis	256

13	Variable Selection for Gene Expression Data	260
13.1	Data and Research Question	260
13.2	Model Comparison and Variable Selection via Boosting	260
13.2.1	Comparison with Cyclical Boosting Approach	262
13.2.2	Additional Approaches towards Enhanced Variable Selection	263
Appendix A	Continuous Distributions	267
Appendix B	Discrete Distributions	269
<i>References</i>		271
<i>Index</i>		283

What this Book is About

The 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 include 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 by

1. relaxing the assumption that the response distribution belongs to the exponential family such that any parametric distribution can be assumed for the response, and

2. attaching separate regression predictors for each parameter of the distribution in particular parameters determining scale and shape features.

The assumption of a parametric response distribution under the generalized linear model 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 estimation, Bayesian inference, and functional gradient descent estimation as specific inferential approaches for implementing inference in GAMLSS.

The main contribution of the GAMLSS framework is that it challenges the common notion of modeling only the mean of the distribution of the response variable. Instead, all parameters (including location, scale, and shape) of that distribution can be modeled. This allows the analyst to approach a much wider range of phenomena, including heterogeneity in variance, positive or negative skewness, plateaus or asymptotic distributions, heavy tails and extremes, overdispersion and underdispersion, missing data, excess or shortage of certain parts of the support (e.g., accumulated counts), and multivariate responses. Importantly, the aspects beyond the mean