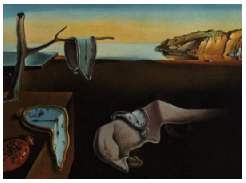

Journal of Statistical Software

December 2007, Volume 23, Issue 7. *http://www.jstatsoft.org/*

Generalized Additive Models for Location Scale and Shape (GAMLSS) in R

D. Mikis Stasinopoulos
London Metropolitan University

Robert A. Rigby
London Metropolitan University

Abstract

GAMLSS is a general framework for fitting regression type models where the distribution of the response variable does not have to belong to the exponential family and includes highly skew and kurtotic continuous and discrete distribution. GAMLSS allows all the parameters of the distribution of the response variable to be modelled as linear/non-linear or smooth functions of the explanatory variables. This paper starts by defining the statistical framework of GAMLSS, then describes the current implementation of GAMLSS in R and finally gives four different data examples to demonstrate how GAMLSS can be used for statistical modelling.

Keywords: Box-Cox transformation, centile estimation, cubic smoothing splines, LMS method, negative binomial, non-normal, non-parametric, overdispersion, penalized likelihood, skewness and kurtosis.

1. What is GAMLSS?

1.1. Introduction

Generalized additive models for location, scale and shape (GAMLSS) are semi-parametric regression type models. They are parametric, in that they require a parametric distribution assumption for the response variable, and "semi" in the sense that the modelling of the parameters of the distribution, as functions of explanatory variables, may involve using non-parametric smoothing functions. GAMLSS were introduced by Rigby and Stasinopoulos (2001, 2005) and Akantziliotou, Rigby, and Stasinopoulos (2002) as a way of overcoming some of the limitations associated with the popular generalized linear models, GLM, and generalized additive models, GAM (see Nelder and Wedderburn 1972; Hastie and Tibshirani 1990, respectively).

In GAMLSS the exponential family distribution assumption for the response variable (y) is relaxed and replaced by a general distribution family, including highly skew and/or kurtotic continuous and discrete distributions. The systematic part of the model is expanded to allow modelling not only of the mean (or location) but other parameters of the distribution of y as, linear and/or non-linear, parametric and/or additive non-parametric functions of explanatory variables and/or random effects. Hence GAMLSS is especially suited to modelling a response variable which does not follow an exponential family distribution, (e.g., leptokurtic or platykurtic and/or positive or negative skew response data, or overdispersed counts) or which exhibit heterogeneity, (e.g., where the scale or shape of the distribution of the response variable changes with explanatory variables(s)).

There are several R-packages that can be seen as related to the **gamlss** packages and to its R implementation. The original **gam** package (Hastie 2006), the recommenced R package **mgcv** (Wood 2001), the general smoothing splines package **gss** (Gu 2007) and the vector GAM package, **VGAM** (Yee 2007). The first three deal mainly with models for the mean from an exponential family distribution. The **VGAM** package allows the modelling from a variety of different distributions (usually up to three parameter ones) and also allows multivariate responses.

The remainder of Section 1 defines the GAMLSS model, available distributions, available additive terms and model fitting. Section 2 describes the R **gamlss** package for fitting the GAMLSS model. Section 3 gives four data examples to illustrate GAMLSS modelling.

1.2. The GAMLSS model

A GAMLSS model assumes independent observations y_i for $i = 1, 2, \ldots, n$ with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$ conditional on $\boldsymbol{\theta}^i = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$ a vector of four distribution parameters, each of which can be a function to the explanatory variables. We shall refer to $(\mu_i, \sigma_i, \nu_i, \tau_i)$ as the *distribution parameters*. The first two population distribution parameters μ_i and σ_i are usually characterized as location and scale parameters, while the remaining parameter(s), if any, are characterized as shape parameters, e.g., skewness and kurtosis parameters, although the model may be applied more generally to the parameters of any population distribution, and can be generalized to more than four distribution parameters.

Rigby and Stasinopoulos (2005) define the original formulation of a GAMLSS model as follows. Let $\mathbf{y}^\top = (y_1, y_2, \ldots, y_n)$ be the n length vector of the response variable. Also for $k = 1, 2, 3, 4$, let $g_k(.)$ be known monotonic link functions relating the distribution parameters to explanatory variables by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk}, \tag{1}$$

i.e.

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1}$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2}$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3}$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4}.$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ $\boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$ are vectors of length n, $\boldsymbol{\beta}_k^\top = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J'_k k})$ is a parameter vector of length J'_k, $\mathbf{X}_k$ is a fixed known design matrix of order $n \times J'_k$, $\mathbf{Z}_{jk}$ is a fixed known $n \times q_{jk}$ design matrix and $\boldsymbol{\gamma}_{jk}$ is a q_{jk} dimensional random variable which is assumed to be distributed as $\boldsymbol{\gamma}_{jk} \sim N_{q_{jk}}(\mathbf{0}, \mathbf{G}_{jk}^{-1})$, where $\mathbf{G}_{jk}^{-1}$ is the (generalized) inverse of a $q_{jk} \times q_{jk}$ symmetric matrix $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$ which may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$, and where if $\mathbf{G}_{jk}$ is singular then $\boldsymbol{\gamma}_{jk}$ is understood to have an improper prior density function proportional to $\exp\left(-\frac{1}{2}\boldsymbol{\gamma}_{jk}^\top \mathbf{G}_{jk}\boldsymbol{\gamma}_{jk}\right)$.

The model in (1) allows the user to model each distribution parameter as a linear function of explanatory variables and/or as linear functions of stochastic variables (random effects). Note that seldom will all distribution parameters need to be modelled using explanatory variables.

There are several important sub-models of GAMLSS. For example for readers familiar with smoothing, the following GAMLSS sub-model formulation may be more familiar. Let $\mathbf{Z}_{jk} = \mathbf{I}_n$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix, and $\boldsymbol{\gamma}_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ for all combinations of j and k in (1), then we have the *semi-parametric additive* formulation of GAMLSS given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{2}$$

where to abbreviate the notation use $\boldsymbol{\theta}_k$ for $k = 1, 2, 3, 4$ to represent the distribution parameter vectors $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$, and where $\mathbf{x}_{jk}$ for $j = 1, 2, \ldots, J_k$ are also vectors of length n. The function h_{jk} is an unknown function of the explanatory variable X_{jk} and $\mathbf{h}_{jk} = h_{jk}(\mathbf{x}_{jk})$ is the vector which evaluates the function h_{jk} at $\mathbf{x}_{jk}$. If there are no additive terms in any of the distribution parameters we have the simple *parametric linear* GAMLSS model,

$$g_1(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k \tag{3}$$

Model (2) can be extended to allow non-linear parametric terms to be included in the model for μ, σ, ν and τ, as follows (see Rigby and Stasinopoulos 2006)

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{4}$$

where h_k for $k = 1, 2, 3, 4$ are non-linear functions and $\mathbf{X_k}$ is a known design matrix of order $n \times J''_k$. We shall refer to the model in (4) as the *non-linear semi-parametric additive* GAMLSS model. If, for $k = 1, 2, 3, 4$, $J_k = 0$, that is, if for all distribution parameters we do not have additive terms, then model (4) is reduced to a *non-linear parametric* GAMLSS model.

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k). \tag{5}$$

If, in addition, $h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) = \mathbf{X}_k^\top\boldsymbol{\beta}_k$ for $i = 1, 2, \ldots, n$ and $k = 1, 2, 3, 4$ then (5) reduces to the linear parametric model (3). Note that some of the terms in each $h_k(\mathbf{X_k}, \boldsymbol{\beta_k})$ may be linear,

in which case the GAMLSS model is a combination of linear and non-linear parametric terms. We shall refer to any combination of models (3) or (5) as a *parametric* GAMLSS model.

The parametric vectors $\boldsymbol{\beta}_k$ and the random effects parameters $\boldsymbol{\gamma}_{jk}$, for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, 3, 4$ are estimated within the GAMLSS framework (for fixed values of the smoothing hyper-parameters λ_{jk}'s) by maximising a penalized likelihood function ℓ_p given by

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\boldsymbol{\gamma}_{jk}'\mathbf{G}_{jk}\boldsymbol{\gamma}_{jk} \tag{6}$$

where $\ell = \sum_{i=1}^{n}\log f(y_i|\boldsymbol{\theta}^i)$ is the log likelihood function. More details on how the penalized log likelihood ℓ_p is maximized are given in Section 1.5. For parametric GAMLSS model (3) or (5), ℓ_p reduces to ℓ, and the β_k for $k = 1, 2, 3, 4$ are estimated by maximizing the likelihood function ℓ. The available distributions and the different additive terms in the current GAMLSS implementation in R are given in Sections 1.3 and 1.4 respectively. The R function to fit a GAMLSS model is `gamlss()` in the package **gamlss** which will be described in more detail in Section 2.

1.3. Available distributions in GAMLSS

The form of the distribution assumed for the response variable y, $f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$, can be very general. The only restriction that the R implementation of GAMLSS has is that the function $\log f(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ and its first (and optionally expected second and cross) derivatives with respect to each of the parameters of $\boldsymbol{\theta}$ must be computable. Explicit derivatives are preferable but numerical derivatives can be used.

Table 1 shows a variety of one, two, three and four parameter families of continuous distributions implemented in our current software version. Table 2 shows the discrete distributions. We shall refer to the distributions in Tables 1 and 2 as the `gamlss.family` distributions, a name to coincide with the R object created by the package **gamlss**. Johnson, Kotz, and Balakrishnan (1994, 1995); Johnson, Kotz, and Kemp (2005) are the classical reference books for most of the distributions in Tables 1 and 2. The BCCG distribution in Table 1 is the Box-Cox transformation model used by Cole and Green (1992) (also known as the LMS method of centile estimation). The BCPE and BCT distributions, described in Rigby and Stasinopoulos (2004, 2006) respectively, generalize the BCCG distribution to allow modelling of both skewness and kurtosis. For some of the distributions shown in Tables 1 and 2 more that one parameterization has been implemented. For example, the two parameter Weibull distribution can be parameterized as $f(y|\mu, \sigma) = (\sigma y^{\sigma-1}/\mu^{\sigma})\exp\{-(y/\mu)^{\sigma}\}$, denoted as WEI, or as $f(y|\mu, \sigma) = \sigma\mu y^{\sigma-1}e^{-\mu y^{\sigma}}$, denoted as WEI2, or as $f(y|\mu, \sigma) = (\sigma/\beta)(y/\beta)^{\sigma-1}\exp\{-(y/\beta)^{\sigma}\}$ denoted as WEI3, for $\beta = \mu/[\Gamma(1/\sigma) + 1]$. Note that the second parameterization WEI2 is suited to proportional hazard (PH) models. In the WEI3 parameterization, parameter μ is equal to the mean of y. The choice of parameterization depends upon the particular problem, but some parameterizations are computationally preferable to others in the sense that maximization of the likelihood function is easier. This usually happens when the parameters μ, σ, ν and τ are orthogonal or almost orthogonal. For interpretation purposes we favour parameterizations where the parameter μ is a location parameter (mean, median or mode). The specific parameterizations used in the `gamlss.family` distributions are given in the appendix of Stasinopoulos, Rigby, and Akantziliotou (2006).

Distributions	R Name	μ	σ	ν	τ
beta	BE()	logit	logit	-	-
beta inflated (at 0)	BEOI()	logit	log	logit	-
beta inflated (at 1)	BEZI()	logit	log	logit	-
beta inflated (at 0 and 1)	BEINF()	logit	logit	log	log
Box-Cox Cole and Green	BCCG()	identity	log	identity	-
Box-Cox power exponential	BCPE()	identity	log	identity	log
Box-Cox-t	BCT()	identity	log	identity	log
exponential	EXP()	log	-	-	-
exponential Gaussian	exGAUS()	identity	log	log	-
exponential gen. beta type 2	EGB2()	identity	identity	log	log
gamma	GA()	log	log	-	-
generalized beta type 1	GB1()	logit	logit	log	log
generalized beta type 2	GB2()	log	identity	log	log
generalized gamma	GG()	log	log	identity	-
generalized inverse Gaussian	GIG()	log	log	identity	-
generalized y	GT()	identity	log	log	log
Gumbel	GU()	identity	log	-	-
inverse Gaussian	IG()	log	log	-	-
Johnson's SU (μ the mean)	JSU()	identity	log	identity	log
Johnson's original SU	JSUo()	identity	log	identity	log
logistic	LO()	identity	log	-	-
log normal	LOGNO()	log	log	-	-
log normal (Box-Cox)	LNO()	log	log	fixed	-
NET	NET()	identity	log	fixed	fixed
normal	NO()	identity	log	-	-
normal family	NOF()	identity	log	identity	-
power exponential	PE()	identity	log	log	-
reverse Gumbel	RG()	identity	log	-	-
skew power exponential type 1	SEP1()	identity	log	identity	log
skew power exponential type 2	SEP2()	identity	log	identity	log
skew power exponential type 3	SEP3()	identity	log	log	log
skew power exponential type 4	SEP4()	identity	log	log	log
shash	SHASH()	identity	log	log	log
skew t type 1	ST1()	identity	log	identity	log
skew t type 2	ST2()	identity	log	identity	log
skew t type 3	ST3()	identity	log	log	log
skew t type 4	ST4()	identity	log	log	log
skew t type 5	ST5()	identity	log	identity	log
t Family	TF()	identity	log	log	-
Weibull	WEI()	log	log	-	-
Weibull (PH)	WEI2()	log	log	-	-
Weibull (μ the mean)	WEI3()	log	log	-	-
zero adjusted IG	ZAIG()	log	log	logit	-

Table 1: Continuous distributions implemented within the **gamlss** packages (with default link functions).

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.