

The R implementation of Generalized Additive Models for Location, Scale and Shape

K. Akanztliotou, R. A. Rigby and D. M. Stasinopoulos ¹

¹ STORM: Statistics, OR and Probabilistic Methods Research Centre at University of North London, London, UK.

Abstract: In this paper we describe how the class of univariate statistical models called Generalised Additive Models for Location, Scale and Shape, **GAMLSS** is implemented into the R statistical package. Within GAMLSS models the distribution for the response variable y can be selected from a very general family of continuous and discrete distributions including highly skew or kurtotic distributions. The systematic part of the model is expanded to allow modelling not only the mean (or location), but other parameters of the distribution of y , as linear parametric or additive non-parametric functions of explanatory variables. In this paper the R functions to fit and display the GAMLSS models are described by way of a simple example.

Keywords: Additive models, Generalized t -distribution, Kurtosis, Logistic distribution, Penalized likelihood, Skewness, Smoothing Splines.

1 Introduction

Generalized Additive Models for Location, Scale and Shape (GAMLSS) were introduced by Rigby and Stasinopoulos (2001, 2002) as way of overcoming some of the limitations associated with Generalized Linear Models (GLM) and Generalized Additive Models (GAM) (Nelder and Wedderburn, 1972 and 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 distributions. The systematic part of the model is expanded to allow modelling not only the mean (or location) but other parameters of the distribution of y as linear parametric or additive non-parametric functions of explanatory variables. Maximum (penalised) likelihood estimation is used to fit the models. The algorithm used to fit the models is discussed in detail in Rigby and Stasinopoulos (2002). Section 2 defines a GAMLSS model. Section 3 shows the different distributions for the response variable included in our current R implementation of GAMLSS. The R functions are also discussed in more detail in section 3.

2 The GAMLSS Model

A GAMLSS model assumes independent observations y_i for $i = 1, 2, \dots, n$ with probability (density) function $f(y_i|\boldsymbol{\theta}^i)$ conditional on $\boldsymbol{\theta}^i = (\theta_{i1}, \theta_{i2}, \dots, \theta_{ip})$ is a vector of p parameters, each of which is related to the explanatory variables. In many practical situations at most $p = 4$ distribution parameters are required. The R implementation denotes these parameters as $(\mu_i, \sigma_i, \nu_i, \tau_i)$. The first two population 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, although the model may be applied more generally to the parameters of any population distribution. Let $\mathbf{y}^T = (y_1, y_2, \dots, y_n)$ be the n length vector of the response variable. Also for $k = 1, 2, 3, 4$, let $g_k(\cdot)$ be known monotonic link functions relating the k^{th} parameter $\boldsymbol{\theta}_k$ to J_k explanatory variables by semi-parametric Additive models given by

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}) \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} h_{j3}(\mathbf{x}_{j3}) \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} h_{j4}(\mathbf{x}_{j4}). \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$ and $\boldsymbol{\eta}_k$ and \mathbf{x}_{jk} for $j = 1, 2, \dots, J_k$ and $k = 1, 2, 3, 4$ are vectors of length n . The function h_{jk} is a non-parametric additive function of the explanatory variable X_{jk} evaluated at \mathbf{x}_{jk} . The explanatory vectors \mathbf{x}_{jk} are assumed fixed and known. Also X_k , for $k = 1, 2, 3, 4$, are fixed design matrices while $\boldsymbol{\beta}_k$ are the parameters vectors. Note that in typical applications a constant or other simple model is often adequate for each of the two shape parameters ($\boldsymbol{\nu}$ and $\boldsymbol{\tau}$).

The parametric vectors $\boldsymbol{\beta}_k$ and the additive terms h_{jk} for $j = 1, 2, \dots, J_k$ and $k = 1, 2, 3, 4$ are estimated within the GAMLSS framework (for fixed values of the regularization or smoothing parameters λ_{jk}) by maximising a penalized likelihood function ℓ_p given by $\ell_p = \ell - \frac{1}{2} \sum_{k=1}^4 \sum_{j=1}^{J_k} \lambda_{jk} \mathbf{Q}_{jk}(\mathbf{h}_{jk})$, where $\ell = \sum_{i=1}^n \log f(y_i|\boldsymbol{\theta}^i)$ is the log likelihood function, where for $j = 1, 2, \dots, J_k$ and $k = 1, 2, 3, 4$, $\mathbf{h}_{jk} = h_{ij}(\mathbf{x}_{jk})$ is the vector of evaluation of function h_{jk} at \mathbf{x}_{jk} and $\mathbf{Q}_{jk}(\mathbf{h}_{jk})$ are appropriate quadratic penalties introduced to penalize undesirable properties in the h_{jk} functions and to provide unique solutions for the functions h_{jk} to the otherwise ill-posed

TABLE 1. Implemented GAMLSS distributions

No of parameters	Distributions
Discrete One parameter	Poisson(PO), Positive.Poisson(PP), Geometric(GO), Logarithmic (LG), Yule (YU), Binomial (BI),
Continuous one parameter	Exponential (EX), Pareto (PA)
Discrete Two parameters	Negative.Binomial.type.I (NB), Negative.Binomial.type.II (BN), Poisson.Inverse.Gaussian (PI), Beta.Binomial (BB)
Continuous Two parameters	Normal (NO), Gamma (GA), Inverse.Gaussian (IG), Gumbel (GU), Reverse.Gumbel (RG) Logistic (LO), Log.Logistic (LL), Weibull (WE), Box.Cox (BC)
Discrete Three parameters	Sichel (SI)
Continuous Three parameters	Cole.Green (i.e. Box-Cox Normal) (CG) Generalized.Gamma.Family (GG) Exponential.Power.Family (EP) t .Family (TF), Generalized.Extreme.Family (GE)
Continuous Four parameters	Generalized. t .Family (i.e. Box-Cox t) (GT)

problem. The form of $\mathbf{Q}_{jk}(\mathbf{h}_{jk})$ depends on the different types of additive terms required. Quadratic penalties in the likelihood result from assuming a Normally distributed random effect exists in the linear predictor, see Rigby and Stasinopoulos (2001).

3 The R functions

Table 1 shows the response variable distributions implemented in the current R implementation of GAMLSS. The following are the R functions for the GAMLSS class.

3.1 The `gamlss()` function

The `gamlss()` is the R-function used to fit a GAMLSS model. Some essential arguments of this function are:

(i) **formula**: essential argument that specifies the model for the location parameter, μ , (ii) **sigma.formula**, (iii) **nu.formula** and (iv) **tau.formula** as optional arguments that specify the models for the appropriate parameters, σ, ν, τ . Also, an essential argument is (v) **family** that identifies the

distribution, (current options shown in table 1). Johnson *et al.* (1992, 1994, 1995) are the classical reference books for these distributions. The Cole and Green distribution in table 1 is the parameterization of the Box-Cox transformation model used by Cole and Green (1992). The Generalized t is obtained by modelling the Cole and Green transformation of y using a t rather than their normally distributed variable, hence incorporating an additional kurtosis parameter (the t distribution degrees of freedom $\nu > 0$ treated as a continuous parameter). Clearly table 1 provides a wide selection of distributions to choose from, but in addition the user can define their own distribution by reediting one of the existing family functions.

Starting values are also allowed by using the optional arguments: **mu.start**, **sigma.start**, etc. The algorithms used for fitting a GAMLSS model are given by Rigby and Stasinopoulos (2002).

To illustrate an example, the **abdominal** data set, will be used (data = abdom); the data are measurements of Abdominal circumference (response variable *abdomvar*) taken from 663 fetuses during ultrasound scans at Kings College Hospital, London, at gestational ages (variable *gest*) ranging between 12 and 42 weeks. The data were used to derive reference intervals by Chitty *et al.* (1994) and also for comparing different reference centile methods by Wright and Royston (1997), who were unable to find a satisfactory distribution for y (*abdomvar*) and commented that the distribution of residual Z-scores obtained from the different fitted models 'has somewhat longer tails than the normal distribution'. The R fit command is displayed below:

```
abdomfit <- gamlss(abdomvar ~ cs(gest, 3), sigma.formula =
~ cs(gest, 3), nu.formula = ~ 1, family = TF(mu.link = identity,
sigma.link = log, nu.link = log), data = abdom)
```

is a model where the response variable *abdomvar* has a t distribution with the location parameter μ modelled, using an identity link, as a smoothing cubic spline with 3 extra degrees of freedom on top of the linear term in *gest*, [i.e. $cs(x, 3)$], similarly for the scale parameter σ , and the t distribution degrees of freedom parameter ν (specified as the nu parameter in R) is constant, [i.e. model 1], but modelled in the log scale. The fit reached convergence at 5th iteration and the resulting output of this fit is displayed below:

```
GAMLSS iteration 1: Global Deviance = 4777.497
GAMLSS iteration 2: Global Deviance = 4776.573
GAMLSS iteration 3: Global Deviance = 4776.554
GAMLSS iteration 4: Global Deviance = 4776.553
GAMLSS iteration 5: Global Deviance = 4776.553
```

3.2 The summary() function

The **summary()** is used to summarise the results produced by a GAMLSS model fitting. It displays the parameter estimates of each model (location,

scale, shape, ...) as well as their standard errors and p-values for significance tests and a brief summary of the degrees of freedom of the fit, the values of (i) the Global Deviance (**GD**), (ii) the Akaike Criterion (**AIC**), and (iii) the Schwarz Bayesian information Criterion (**SBC**), the number of iterations used on the fit, etc. Following our example, the summary is:

```
summary(object = abdomfit)
giving the following output
*****
Family: c("TF", "t.Family")
Call:  gamlss(formula = abdomvar ~ cs(gest,3), sigma.formula =
~ cs(gest,3), family = TF(), data = abdom)
-----
Mu link function: identity
Mu Coefficients:
              Estimate  Std. Error  t value  Pr(> |t|)
(Intercept)   -63.47    1.38017  -45.98   1.296e-199
cs(gest, 3)    10.67    0.05943   179.54   0.000e+00
-----
Sigma link function: log
Sigma Coefficients:
              Estimate  Std. Error  t value  Pr(> |t|)
(Intercept)    2.5750    0.216360   11.90   1.708e-29
cs(gest, 3)    0.0824    0.007579   10.87   2.841e-25
-----
Nu link function: log
Nu Coefficients:
              Estimate  Std. Error  t value  Pr(> |t|)
(Intercept)    2.485    0.2980    8.338   5.053e-16
-----
No. of observations in the fit: 610
Degrees of Freedom for the fit: 10.99969
Residual Deg. of Freedom: 599.0003
                        at cycle: 5

Global Deviance: 4776.553
                AIC: 4798.552
                SBC: 4847.099
*****
```

3.3 The plot() function

The `plot()` is used to produce a set of four graphs of the quantile residuals of a GAMLSS object. For each fitted GAMLSS model the (randomised) quantile residuals of Dunn and Smyth (1996) can be used to check the adequacy of the model and especially the distribution of the y variable.

The randomised quantile residuals are given by $r_i = \Phi^{-1}(u_i)$ where Φ^{-1} is the inverse cumulative distribution function of a standard normal variate and u_i is defined as $u_i = F(y_i|\hat{\theta}^i)$ if y_i is continuous or a random value from the uniform distribution in the interval $[F((y_i - 1)|\hat{\theta}^i), F(y_i|\hat{\theta}^i)]$ if y_i is discrete. A set of plots of the (randomised) quantile residuals are given by using the `plot` function. In our example:

`plot(object = abdomfit)`

gives the following output (including the graph):

Summary of the Randomised Quantile Residuals

bandwidth = 0.2942333
coef. of skewness = 0.1662309
coef. of kurtosis = 2.992121

Filliben correlation coefficient = 0.998519

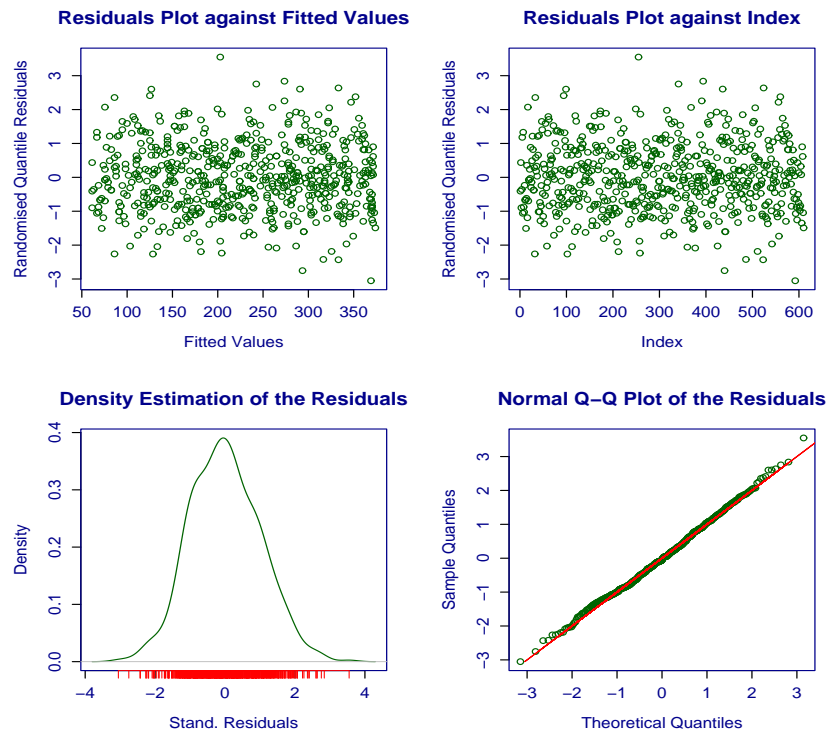


FIGURE 1. Randomised quantile residuals plots for the abdominal data.

3.4 The profile() function

The **profile()** is used to produce a profile global deviance graph with respect to the one of the parameters of a GAMLSS model, to give the maximum likelihood estimate (i.e. optimal value) of the parameter (where the GD is minimized), together with a 95% confidence interval for the parameter. The **which** argument specifies the desired parameter ("mu", "sigma", "nu", "tau") that the profile Global Deviance graph will be produced with respect to this parameter. In our example

```
profile(object = abdomfit, which = "nu", min = 4, max = 50, step = 1)
gives the following output (including the graph):
*****
```

Profile Global Deviance

```
Best estimate of the fixed parameter nu is 12
with a Global Deviance equal to 4776.553 at position 9
A 95% Confidence interval for nu is: ( 6.332382 , 42.89995 )
*****
```

Figure 2 gives the profile global deviance plot.

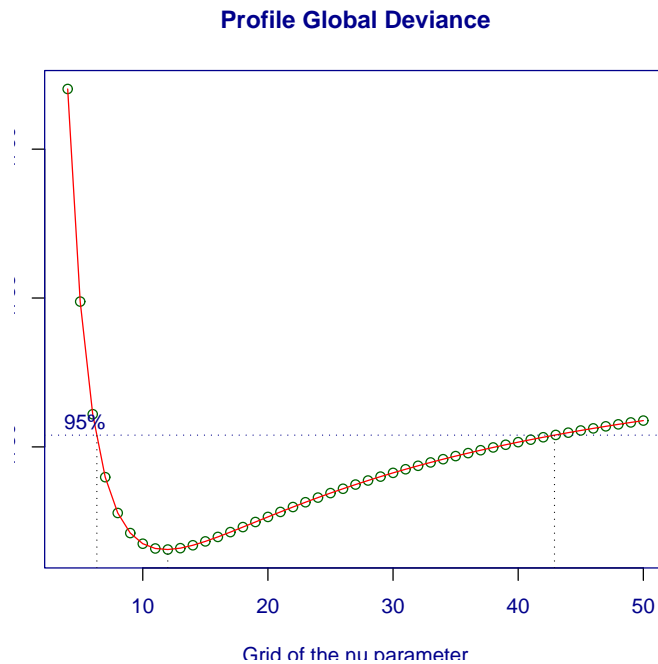


FIGURE 2. The profile global deviance plot for parameter ν (=nu) for the abdominal data.

3.5 The `refit()` function

The `refit()` is used to refit the model, if the maximum number of iterations has been reached, but the Global deviance has not yet converged. The default maximum number of (outer) iterations is 20. In our example the command to use is

```
refit(object = abdomfit)
```

3.6 The `control()` function

The `control()` is used to control the iterations for GAMLSS model fitting, e.g. to change the maximum number of (outer) iterations needed for the fit or the constant of convergence, etc.

References

- Chitty, L.S., Altman, D.G., Henderson, A., and Campbell, S. (1994) Charts of fetal size: 3, abdominal measurements. *Br. J. Obstetr.*, **101**, 125-131.
- Cole, T. and Green, P. (1992) Smoothing reference centile curves: The LMS method and penalized likelihood. *Statist. in Med*, **11**, 1305-1319.
- Dunn, P.K. and Smyth, G.K. (1996) Randomized Quantile Residuals. *Journal of Computational Graph. Statist.*, **5**, 236-244.
- Hastie, T.J., and Tibshirani, R.J. (1990) *Generalized Additive Models*. London: Chapman & Hall.
- Johnson, N.L., Kotz, S. and Kemp, A.W. (1992) *Univariate discrete distributions*. New York: Wiley.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate distributions, Volume I*. New York: Wiley.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate distributions, Volume II*. New York: Wiley.
- Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *J. R. Statist. Soc. A*, **135**, 370-384.
- Rigby, R.A. and Stasinopoulos, D.M (2001) The GAMLSS project: a flexible approach to statistical modelling. In *New trends in Statistical Modelling*, proceedings of the 16th International Workshop on Statistical Modelling, editors: B. Klein and L. Korsholm. Odense, Denmark.
- Rigby, R.A. and Stasinopoulos, D.M (2002) Generalized Additive Models for Location, Scale and Shape. Submitted for publication.

Wright, E. M. Royston, P. (1997) A comparison of statistical methods for age-related reference intervals. *J. R. Statist. Soc. A*, Vol: **2**, 47-69.