gamlss

for statistical modelling

Generalized additive models for location scale and shape

Bob Rigby

Mikis Stasinopoulos

Calliope Akantziliotou

LONDON metropolitan university

gamlss

Summary

1. Introduction
2. Model definition
3. Additive terms
4. Population distributions for Y
5. R packages
6. Examples
7. Conclusions

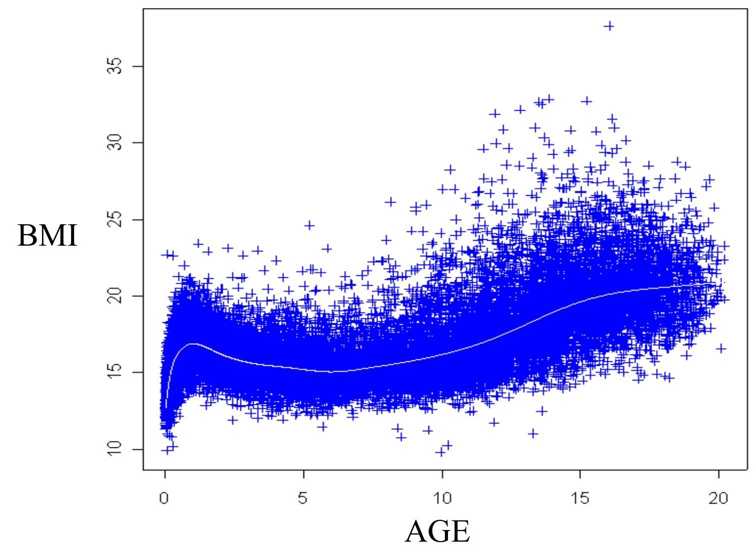
gamlss

Introduction: Why use gamlss?

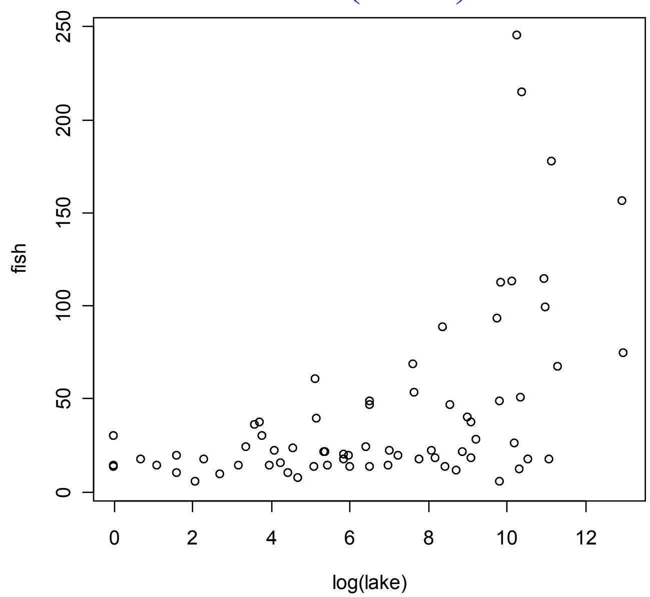
- Unified framework for univariate regression type of models
- The fitted algorithm is modular, where different components can be added easily
- Models can be fitted easily and fast
- Explanatory tool to find appropriate set of models (and then use your favourite mode of inference)
- It deals with
  1. Skewness
  2. Kurtosis
  3. Overdispersion

gamlss

Example 1: BMI against AGE for Dutch girls

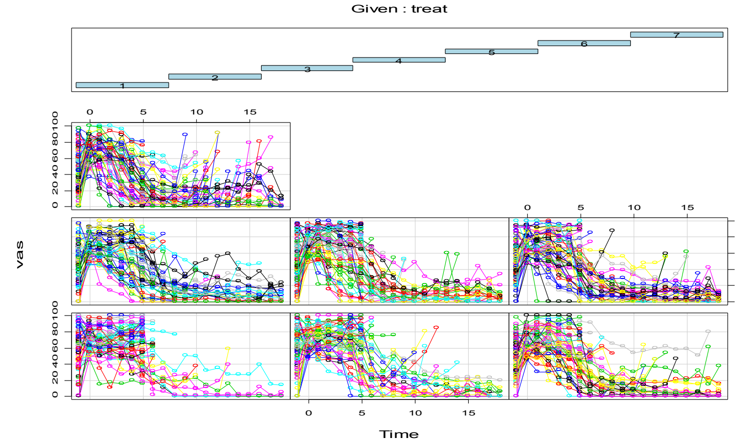


Example 2: The fish species data, Stein and Juritz (1988)



Example 3: Visual analog scale (VAS) data

368 patients, measured at 18 times
with 7 treatment



2. Model definition

Univariate Regression type model

Y <output> ← X's <input>

response   Explanatory variables

GAMLSS philosophy: The response should have a distribution and all the parameters of the distribution could be modelled as functions of explanatory variables

The GAMLSS model

$Y \sim D(\mu, \sigma, \nu, \tau)$ where D is any distribution and

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

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

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

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

Known link

Predictor

Linear terms

Additive terms

Random effects

Here $\gamma_{jk} \sim N_{q_{jk}}\left(0,\ G_{jk}^{-1}\right)$ and $G_{jk} = G_{jk}(\lambda)$

MAP estimation of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ given $\boldsymbol{\lambda}$

Hence given $\boldsymbol{\lambda}$,

posterior mode (or MAP) estimation of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$

¿ maximising l_h, hierarchical log likelihood

¿ maximising l_p, penalised log likelihood

with respect to $(\boldsymbol{\beta}, \boldsymbol{\gamma})$

3. Parametric Additive terms

- Linear and interaction terms for variates and factors.
- Polynomials, inverse polynomials, piecewise polynomials (with fixed knots), fractional polynomials (Royston and Altman, 1994)
- Non-linear parametric terms

Additive terms

- Additive smoothing terms
  - ✓ cubic splines (Green and Silverman, 1994)
  - ✓ P-splines (Eilers and Marx, 1996)
  - ✓ varying coefficient models (Hastie and Tibshirani, 1993)
  - ✓ loess (Cleveland *et al*., 1993)
- Random effects (overdispersion, simple random effects, random coefficients)
- Parameter driven Time Series (random walks)

1. Population distributions for Y

4.1 General comments

2) A wide range of discrete, continuous and mixed distributions implemented, including highly skew and kurtotic distributions
3) Easy implementation of new distributions
4) Different parameterisations of a distribution can be implemented
5) Truncated distributions and censored data easily implemented
6) Finite mixture distributions easy to implement (new)

4.2 Discrete distributions for Y

Two parameter distributions

BB	Beta-Binomial
NBI	Negative Binomial type I
NBII	Negative Binomial type II
PIG	Poisson-Inverse Gaussian
ZIP	Zero inflated Poisson

Three parameter distributions

SICHEL	Sichel
DEL	Delaport

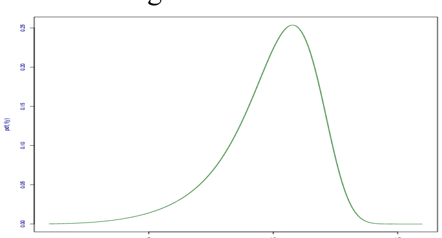
4.3 Continuous distributions for Y

Four parameters

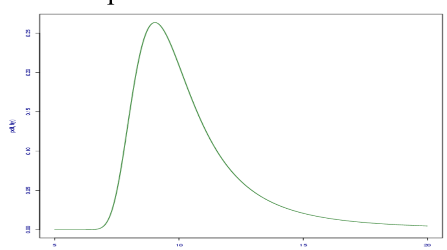
μ	location
σ	scale
ν	skewness
τ	kurtosis

Skewness and kurtosis

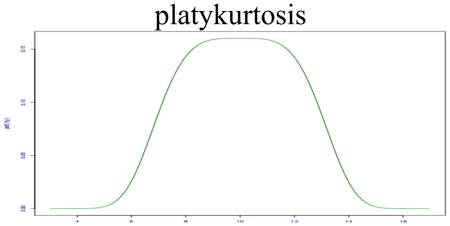
negative skewness



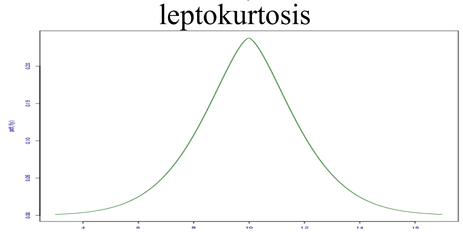
positive skewness



platykurtosis



leptokurtosis



Distributions in real line

Table 3.1: Continuous GAMLSS family distributions defined on $(-\infty, +\infty)$

Distributions	family	no parameters	skewness	kurtosis
Gumbel	GU()	2	(negative)	-
Johnson's SU	JSU()	4	both	lepto
Johnson's original SU	JSUo()	4	both	lepto
Logistic	LO()	2	(symmetric)	(lepto)
NET	NET()	2 + (2 fixed)	(symmetric)	lepto
Normal	NO()	2	(symmetric)	(meso)
Power Exponential	PE()	3	(symmetric)	both
Reverse Gumbel	RG()	2	positive	-
Sinh Arc Sinh	SHASH()	4	both	lepto
Skew Exponential Power	SEP()	4	both	both
Skew t	ST3()	4	both	lepto
t Family	TF()	3	(symmetric)	lepto

Distributions in positive real line

Table 3.2: Continuous GAMLSS family distributions defined on $(0, +\infty)$

Distributions	family	no parameters	skewness	kurtosis
Box-Cox Cole and Green	BCCG()	3	both	-
Box-Cox Power Exponential	BCPE()	4	both	both
Box-Cox-t	BCT()	4	both	lepto
Gamma	GA()	2	positive	-
Inverse Gaussian	IG()	2	positive	-
Log Normal	LNO()	2 +(1 fixed)	positive	
Weibull	WEI()	2	positive	-
Weibull (PH)	WEI2()	2	positive	-

5. The R packages

- gamlss
- gamlss.nl
- gamlss.tr
- gamlss.dist (to be released)
- gamlss.mx (new)
- gamlss.cen (future)

The gamlss package

- The gamlss() function creates a gamlss object
- Methods for a gamlss object

  AIC(), addterm(), coef(), deviance(), fitted(), formula(), lot(), print(), predict(), residuals(), update(),
- Others functions

  centiles(), fitted.plot(), GAIC(), gamlss.scope(), par.plot(), lperd(), pdf.plot(), prof.plot(), prof.term(), Q.stats(), refit(), rqres.plot(), stepGAIC(), term.plot()

The gamlss() function

```
gamlss(formula = formula(data), sigma.formula = ~1,
    nu.formula = ~1, tau.formula = ~1, family = NO(),
    data = sys.parent(), weights = NULL,
    contrasts = NULL, method = RS(), start.from =
 NULL,
    mu.start = NULL, sigma.start = NULL,
    nu.start = NULL, tau.start = NULL,
    mu.fix = FALSE, sigma.fix = FALSE, nu.fix =
 FALSE,
    tau.fix = FALSE, control = gamlss.control(...),
    i.control = glim.control(...), ...)
```

6. Examples: Modelling body mass index (BMI) against AGE

Variables

Body mass index (Y = BMI) against AGE,

for 20243 Dutch girls aged under 20

Study

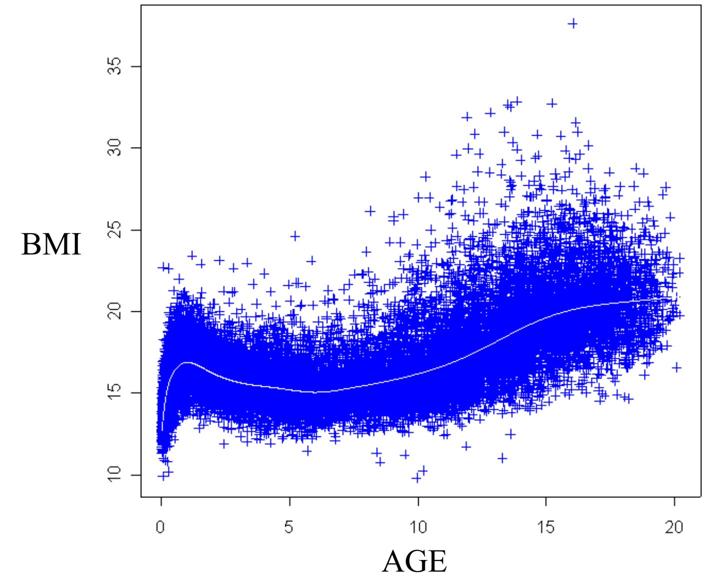
cross sectional data,

Dutch population study,

Cole and Roede (1999)

gamlss

BMI against AGE for Dutch girls



gamlss

Model for BMI

$Y \sim BCT\,(\mu,\ \sigma,\ \nu,\ \tau)$ where BCT is the Box-Cox t distribution,

where $Y = BMI$ and $x = AGE^{\xi}$

$$\mu = cs\left(x,\ df_{\mu}\right)$$

$$\log\left(\sigma\right) = cs\left(x,\ df_{\sigma}\right)$$

$$\nu = cs\left(x,\ df_{\nu}\right)$$

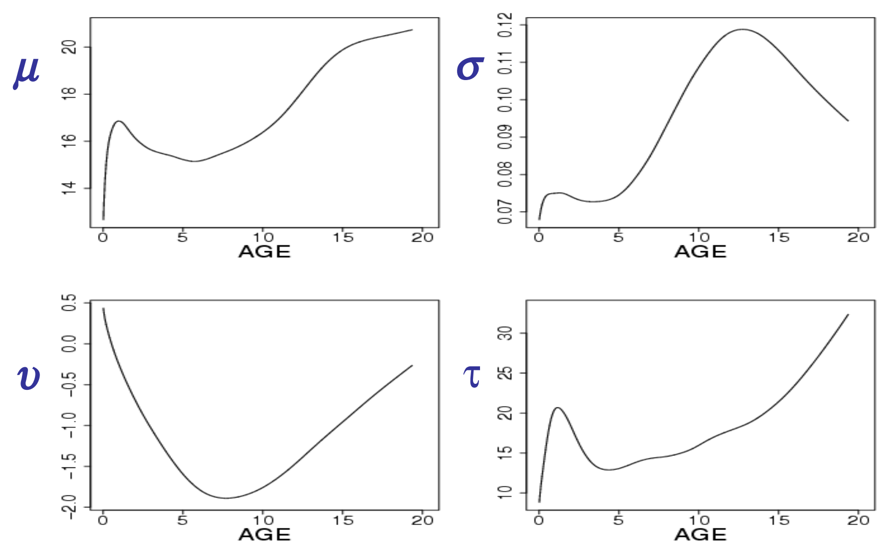
$$\log\left(\tau\right) = cs\left(x,\ df_{\tau}\right)$$

We need to select the five values $df_{\mu},\ df_{\sigma},\ df_{\nu},\ df_{\tau},\ \xi$

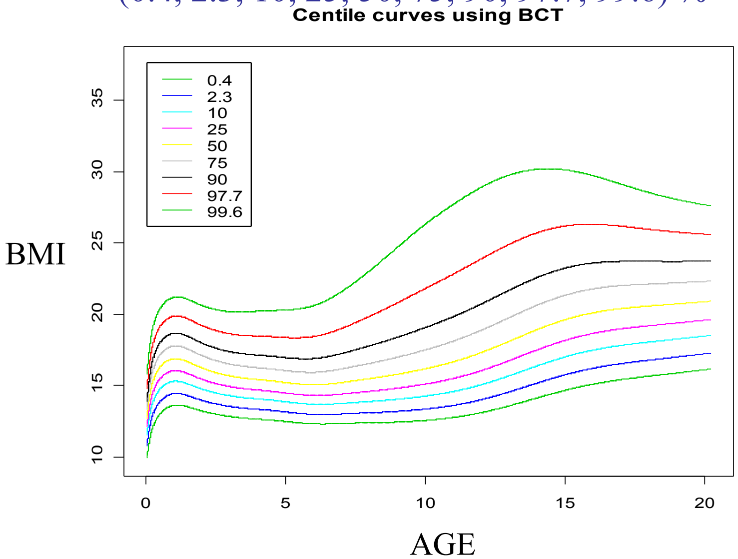
gamlss

Fitted parameters $\mu,\ \sigma,\ \upsilon,\ \tau$ against AGE

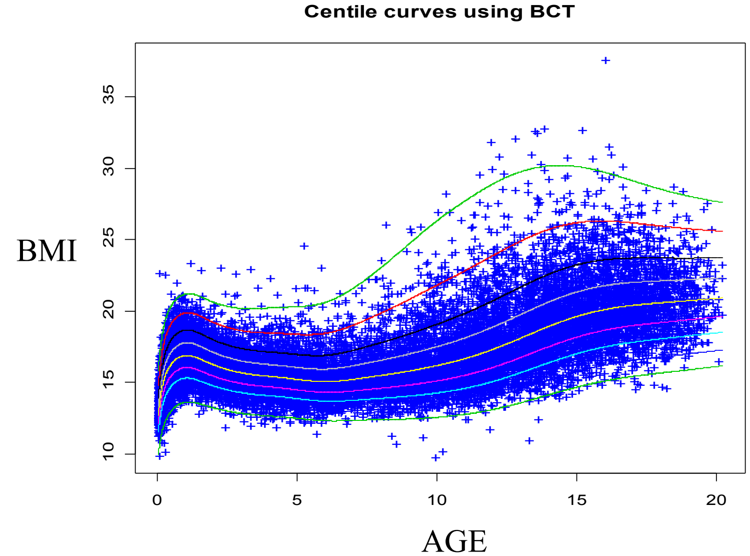
(for BCT model for BMI chosen with penalty # = 2.4)



Centile curves of BMI against AGE (# = 2.4)
(0.4, 2.3, 10, 25, 50, 75, 90, 97.7, 99.6) %



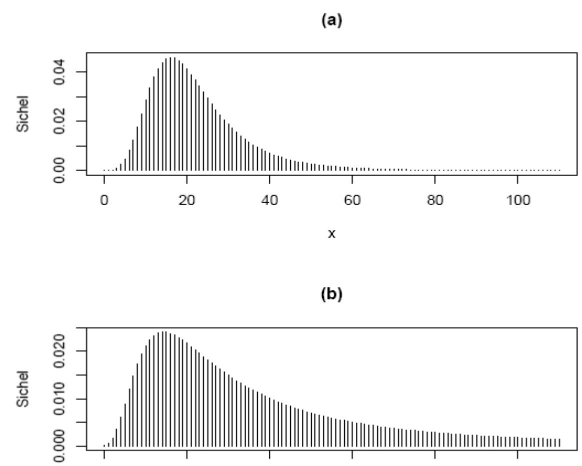
Centile curves of BMI against AGE (# = 2.4)
(0.4, 2.3, 10, 25, 50, 75, 90, 97.7, 99.6) %



Fish species data: Stein and Juritz (1988)

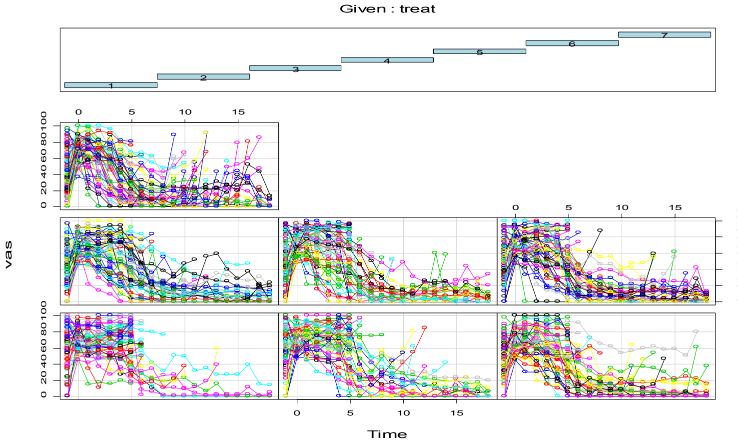
$$
\begin{aligned}
\log(\boldsymbol{\mu}) &= h_1(\mathbf{x}) & & \\
\log(\boldsymbol{\sigma}) &= h_2(\mathbf{x}) & & \\
\boldsymbol{\nu} &= h_3(\mathbf{x}) & \text{for} \quad & SI(\mu,\sigma,\nu) \\
\text{logit}(\boldsymbol{\nu}) &= h_3(\mathbf{x}) & \text{for} \quad & DEL(\mu,\sigma,\nu)
\end{aligned}
$$

Fitted model for observation 40 (a) and 67 (b)



Visual analog scale (VAS) data

368 patients, measured at 18 times
with 7 treatment



Model for all data

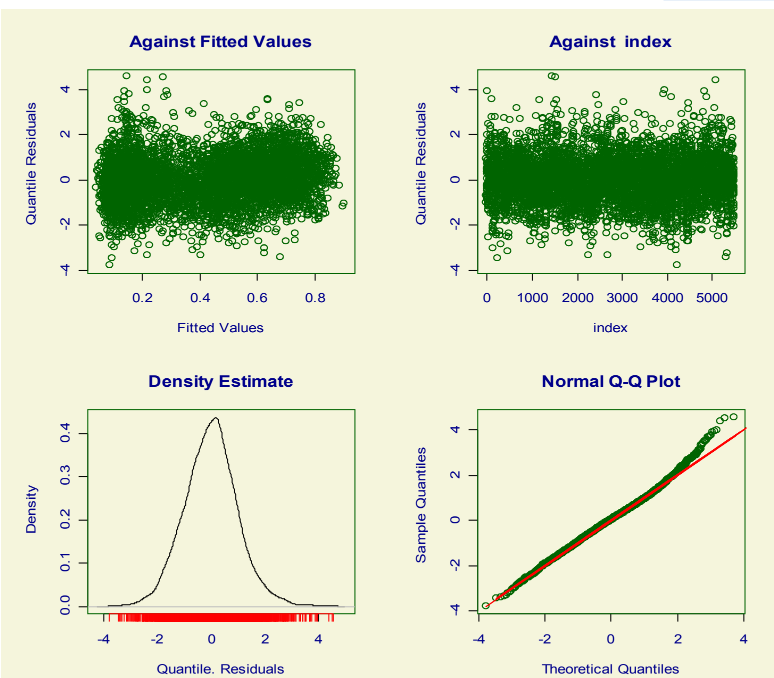
$Y \sim BEINF(\mu, \sigma, \nu, \tau)$ where *BEINF* is the Beta inflated distribution

$$\mu = cs(time, df{=}10) + treat + random(patient, 250)$$

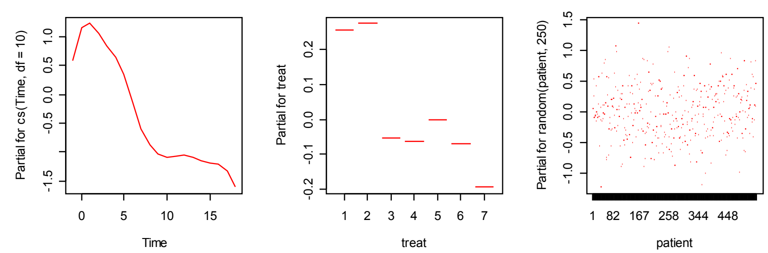
$$\log(\sigma) = cs(time, df{=}10) + treat$$

$$\log(\nu) = cs(time, df{=}5) + treat$$

$$\log(t) = cs(time, df{=}5) + treat$$

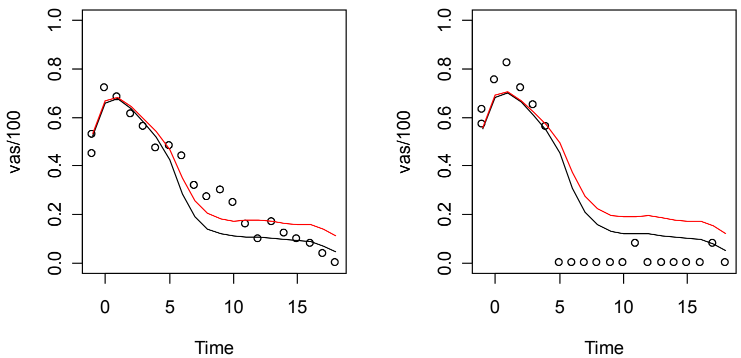


Model for mu



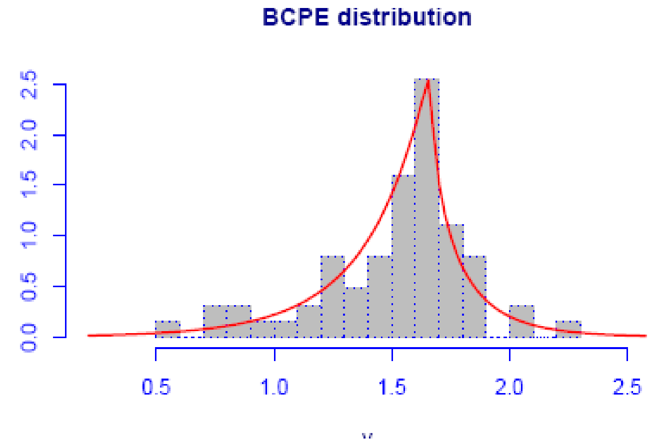
gamlss

Model for 50 and 550 participant



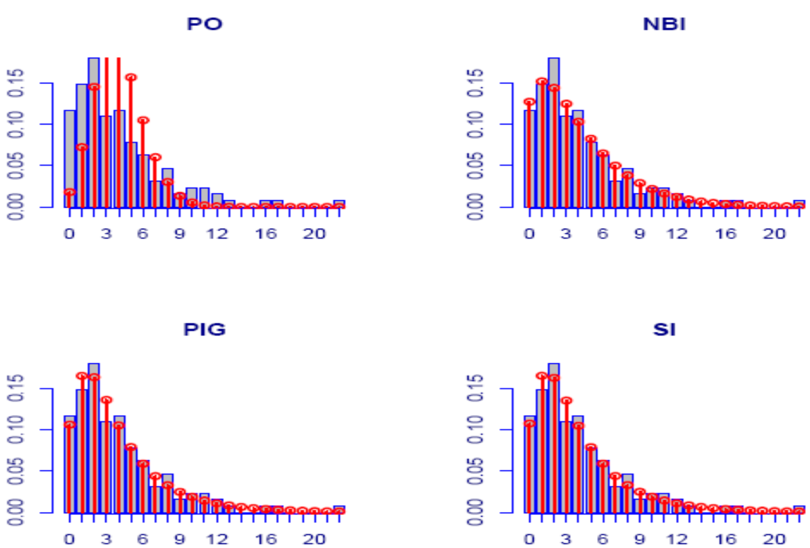
gamlss

Fitting distributions to data



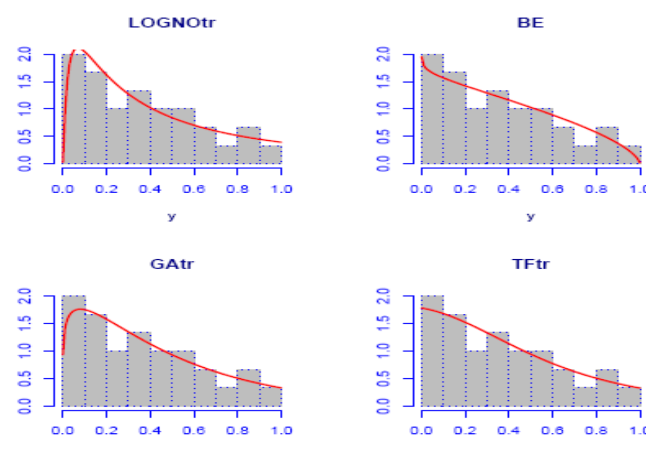
gamlss

Fitting discrete distributions to computer failure data



gamlss

tensile strength data: plots



6. Conclusions

GAMLSS allows flexible modelling of both:

i) the distribution of Y, including models for skewness and kurtosis

ii) the dependence of the distribution parameters, e.g. μ, σ, ν, τ, on explanatory variables and random effect additive terms.

GAMLSS papers, manual and related publications, available from website

http://www.londonmet.ac.uk/gamlss/

END