GAMLSS: A distributional regression approach

Mikis D Stasinopoulos[1], Robert A Rigby[1], and Fernanda De Bastiani[2]
[1]London Metropolitan University, London, United Kingdom.
[2]Universidade Federal de Pernambuco, Recife, PE, Brazil.

Abstract: A tutorial of the generalized additive models for location, scale and shape (GAMLSS) is given here using two examples. GAMLSS is a general framework for performing regression analysis where not only the location (e.g., the mean) of the distribution but also the scale and shape of the distribution can be modelled by explanatory variables.

Key words: beta negative binomial, Box–Cox *t*, centile estimation, head circumference, kurtosis, physician visits, skewness

1 Introduction

This article introduces the generalized additive models for location, scale and shape (GAMLSS) to a wider audience by concentrating on two intuitive examples, avoiding the technical jargon associated with statistical model definitions. We assume that the reader is familiar with simple regression analysis and generalized linear models (GLMs).

Since their introduction by Rigby and Stasinopoulos (2005), GAMLSS have been applied in a variety of different scientific fields such as: actuarial science (Heller et al., 2007), biology (Hawkins et al., 2013), economics (Voudouris et al., 2015), environment (Villarini et al., 2009), genomics (Khondoker et al., 2007), finance (International Monetary Fund, 2015; Giraud and Kockerols, 2015), fisheries, food consumption, management science (Budge et al., 2010), marine research, medicine (Rodrigues et al., 2009), meteorology, and vaccines.

GAMLSS have also become standard for centile estimation, for example, Visser et al. (2009), Villar et al. (2014) and Neuhauser et al. (2011). WHO (2007) and WHO (2009) uses GAMLSS for centile estimation to produce growth charts for children. Their charts are used by more than 140 countries as the standard charts monitoring the growth of children. The Global Lung Function Initiative (GLFI), [http://www.lungfunction.org, Quanjer et al. (2012)] uses GAMLSS to provide a unified worldwide approach to monitoring lung function, by obtaining centiles for lung function based on age and height.

GAMLSS is a general framework for performing a 'univariate' regression. 'Univariate' refers to a single response variable, while there can be many explanatory valuables. In a univariate regression we assume that the response (or target) variable depends on the explanatory variables. This dependance can be linear, non-linear or smooth non-parametric. For example, in the classical linear regression model (LM), the mean of the response variable is a linear function of the explanatory variables. In the GLMs, Nelder and Wedderburn (1972), a monotonic function of the mean, called the linear predictor, is a linear function of the explanatory variables. Since the late 1980s, non-linear relationships between the response variable and the explanatory variables, within both LM and GLM, are dealt with by using non-parametric smoothing functions, giving additive models (AM) and generalized additive models GAMs, respectively. The GAMs introduced by Hastie and Tibshirani (1990) and popularized by Wood (2017) have made the smoothing techniques within a regression framework available to a wide range of practitioners.

GAMLSS can be seen as an extension of the LM, GLM and GAM. GAMLSS have two main features. First, the GLMs and GAMs assume that the response variable has a distribution that belongs to the exponential family. However, in GAMLSS, the assumed distribution can be any parametric distribution. Second, within GAMLSS, all the parameters (not only the location, for example, the mean) of the distribution can be modelled as linear or smooth functions of the explanatory variables. As a result, the location, scale and shape of the distribution of the response variable is allowed to change according to explanatory variables. The GAMLSS models are an example of a 'Beyond Mean Regression' model, Kneib (2013). Because an explicit distributional assumption is made for the response variable, they also fall into the category of the 'distributional regression' modelling approach, Fahrmeir et al. (2013).

GAMLSS allows a variety of smooth functions of explanatory variables. Smoothers can be divided broadly into two categories: the ones that employ a quadratic penalty and the ones that do not. For more details see Stasinopoulos et al. (2017), Ch 10. The quadratic penalized smoothers include some very popular smoothers: P-splines, cubic splines, thin-plate splines, tensor-product splines and random effects. The second category of smoother includes: local regression, neural networks, decision trees, etc.

This article is organized as follows. Section 2 provides a general form of the GAMLSS model for any response variable distribution. Section 3 demonstrates GAMLSS analysis of a continuous response variable. Section 3.2 models height against a single explanatory variable age in Dutch boys, while Section 3.3 models head circumference against two explanatory variables height and age. Both sections focus on obtaining centiles for the response variable. Section 4 demonstrates GAMLSS analysis of a discrete response variable using different discrete distributions. Conclusions are provided in Section 5. The basic ideas of GAMLSS, see Stasinopoulos et al. (2017), have been implemented in **R** in a series of packages, see Stasinopoulos and Rigby (2007). All the material presented in this tutorial are reproducible using the **R** code provided as supplementary material.

2 The GAMLSS framework

GAMLSS are semi-parametric regression-type models. They are ‘semi’ in the sense that the modelling of the parameters of the distribution may involve using non-parametric smoothing functions of explanatory variables, and parametric in the sense that they require a parametric distribution assumption for the response variable. It provides a very general and flexible system for modelling a response variable.

The distribution of the response variable is selected from a very wide range of distributions available in the `gamlss.dist` package in `R`, Stasinopoulos and Rigby (2007), 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 distributions. A GAMLSS model assumes that, for $i = 1, 2, \ldots, n$, independent observations Y_i have probability (density) function $f_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i)$ conditional on up to four distribution parameters, each of which can be a function of the explanatory variables. The first two population distribution parameters μ_i and σ_i are usually characterized as location and scale parameters, and ν_i and τ_i are usually characterized as shape parameters, for example, skewness and kurtosis, respectively.

The model can be generalized to more than four distribution parameters; however, the **gamlss** package includes distributions with up to four parameters. All the parameters of the response variable distribution can be modelled using parametric and/or non-parametric smooth functions of explanatory variables, thus allowing modelling of the location, scale and shape parameters. Rigby and Stasinopoulos (2005) define an original formulation of a GAMLSS model as follows.

Response variable observations $Y_1, Y_2, \ldots, Y_n$ are independent with

$$Y_i \sim D(\mu_i, \sigma_i, \nu_i, \tau_i)$$

for $i = 1, \ldots, n$, where D is any distribution with (up to) four distribution parameters.

For $k = 1, 2, 3, 4$, let $g_k(.)$ be a known monotonic link function relating a distribution parameter to a predictor $\boldsymbol{\eta}_k$, where

$$\begin{aligned}
g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} s_{1j}(\mathbf{x}_{1j}) \\
g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} s_{2j}(\mathbf{x}_{2j}) \\
g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} s_{3j}(\mathbf{x}_{3j}) \\
g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} s_{4j}(\mathbf{x}_{4j}),
\end{aligned} \tag{2.1}$$

where $\mathbf{X}_k$ is a known design matrix, $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kJ'_k})^\top$ is a parameter vector of length J'_k, s_{kj} is a smooth non-parametric function of variable X_{kj} and the $\mathbf{x}_{kj}$'s are vectors of length n, for $k = 1, 2, 3, 4$ and $j = 1, \ldots, J_k$. That is, a GAMLSS model allows the modelling of the parameters of the distribution as linear, that is, $\mathbf{X}_k\boldsymbol{\beta}_k$ or smooth term functions $s_{kj}(x_{kj})$ for $k = 1, 2, 3, 4$.

3 Centile estimation for a continuous response variable

3.1 The Dutch boys data

The Fourth Dutch Growth Study, Fredriks et al. (2000a,b) is a cross-sectional study that measures growth and development of the Dutch population between the ages of 0 and 22 years. The study measured, among other variables, height, weight, head circumference and age for 7 482 males and 7 018 females.

Here we analysed 6 885 observations for head circumference, height and age of males (having removed missing observations of these three variables from the original dataset of 7 482 cases). The data are shown in Figure 1 where in panel (a) the head circumference is plotted against age, in panel (b) head circumference is plotted against height and in panel (c) height is plotted against age.

Section 3.2 shows the analysis of height (as response variable) against a single explanatory variable age. This is a typical 'centile' estimation problem, one of the most widely used applications of GAMLSS. In Section 3.3, we use the head circumference as the response variable, and the age and height as explanatory variables. This will serve us as an example of centile estimation for a response variable using two explanatory variables (by surface interaction fitting).

3.2 Centile estimation using a single explanatory variable

Centile estimation of a response variable is widely used in medicine, nutrition, sport science and other disciplines, where an individual is checked on whether they have an abnormally low or high value of the response variable (given their values of the explanatory variable(s)), and hence whether they are potentially at risk. For example, children are checked whether they have an abnormal height for their age. A typical centile estimation problem will involve two variables, the response variable Y of interest and an explanatory variable x usually age. In this section, we will analyse height (as the response variable) against age (as the explanatory variable).

3.2.1 Normal distribution model and extensions

In the classical linear regression model, a continuous response variable was modelled using a normal distribution with mean μ and standard deviation σ, that is, $Y \sim NO(\mu, \sigma)$, where μ is linear in explanatory variable x and σ is a constant value β_{02}, that is,

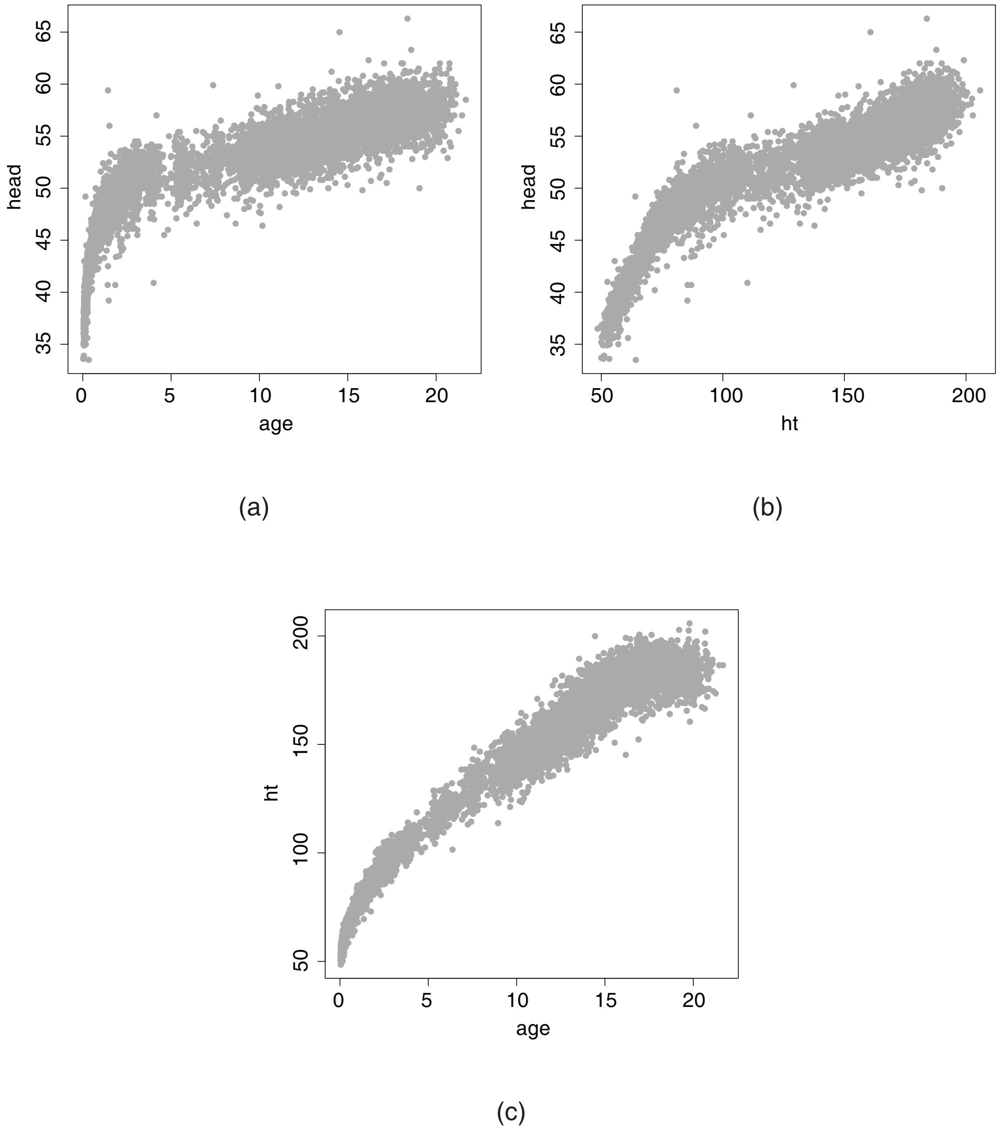

Figure 1 Plots of (a) head circumference (`head`) against age (`age`), (b) head circumference (`head`) against height (`ht`) and (c) height (`ht`) against age (`age`)

$$
\begin{aligned}
Y &\sim NO(\mu, \sigma)\\
\mu &= \beta_{01} + \beta_{11}x\\
\sigma &= \beta_{02}.
\end{aligned}
$$

However, it is often found that modelling μ as linear in x is inadequate, so μ is modelled using a smooth function of x, $s_1(x)$, giving a simple additive model. Also, Y is often found to have heterogeneous variance, so σ is not constant but potentially depends on a smooth function of x giving the following model,

$$
\begin{aligned}
Y &\sim NO(\mu, \sigma)\\
\mu &= s_1(x)\\
\log(\sigma) &= s_2(x).
\end{aligned}
\tag{3.1}
$$

The function $\log(\sigma)$ ensures that σ is always positive.

However, it is often found that the response variable Y (given x) has a skew distribution, so a three-parameter distribution, for example, the Box–Cox Cole and Green, $BCCG(\mu, \sigma, \nu)$, distribution, where ν is a skewness parameter, may be needed to model Y, for example,

$$
\begin{aligned}
Y &\sim BCCG(\mu, \sigma, \nu)\\
\mu &= s_1(x)\\
\log(\sigma) &= s_2(x)\\
\nu &= s_3(x).
\end{aligned}
\tag{3.2}
$$

Finally, the response variable Y may exhibit varying kurtosis, so a four-parameter distribution, for example, the Box-Cox t, $BCT(\mu, \sigma, \nu, \tau)$, where τ is a kurtosis parameter, may be needed to model Y, for example,

$$
\begin{aligned}
Y &\sim BCT(\mu, \sigma, \nu, \tau)\\
\mu &= s_1(x)\\
\log(\sigma) &= s_2(x)\\
\nu &= s_3(x)\\
\log(\tau) &= s_4(x).
\end{aligned}
\tag{3.3}
$$

This model can be generalized to model (3.4), given in the next section. Model (3.1), (3.2), (3.3)and (3.4) are examples of GAMLSS models that can be fitted with the **gamlss** R software.

3.2.2 LMS model and extensions

There are currently two major methodologies for creating centile curves, (a) the lambda, mu and sigma (LMS) method and its extensions (Cole and Green, 1992); (Rigby and Stasinopoulos, 2004; Rigby and Stasinopoulos, 2006) and (b) the quantile regression method (Koenker et al., 1994; He and Ng, 1999; Ng and Maechler, 2007).

More details about both methodologies can be found in Stasinopoulos et al. (2017), Ch 13. Here, we concentrate on the LMS methodology and its extensions, which are a subclass of GAMLSS and were also adopted by the World Health Organization for the construction of worldwide standard growth (centile) curves for children (see WHO, 2006, 2007, 2009).

The model for the extended LMS methodology can be written as:

$$\begin{aligned} Y &\sim \mathcal{D}(\mu, \sigma, \nu, \tau) \\ g_1(\mu) &= s_1(u) \\ g_2(\sigma) &= s_2(u) \\ g_3(\nu) &= s_3(u) \\ g_4(\tau) &= s_4(u) \\ u &= x^{\xi}, \end{aligned} \tag{3.4}$$

where $\mathcal{D}$ represents the distribution of the response variable Y, and μ, σ, ν and τ are parameters of this distribution. The $g(\cdot)$ functions represent appropriate link functions (i.e., known monotonic functions of the distribution parameters, which can also guarantee that the distribution parameter will be in the appropriate range). The $s(\cdot)$ are non-parametric smoothing functions of u, where u is a power transform function of the explanatory variable x and ξ is the power transform parameter. The reason why a power transformation for x may be needed in the model is to facilitate the estimation of the smoothing functions when spells of sharp growth in Y occur for low values of x. The model in Equation (3.4) needs five parameters to be estimated, that is, four smoothing parameters for the functions $s_1(.)$, $s_2(.)$, $s_3(.)$ and $s_4(.)$, and the power transform parameter ξ.

The original LMS method of Cole and Green (1992) uses only three distribution parameters, and it is equivalent to using the Box-Cox Cole and Green distribution, $BCCG(\mu, \sigma, \nu)$, as $\mathcal{D}$. The parameters μ, σ and ν, in this case, are the approximate median, approximate coefficient of variation and skewness parameters. That is, μ controls the location, σ controls the scale and ν controls the skewness of the distribution $\mathcal{D}$. The introduction of a fourth parameter τ for modelling the kurtosis of the distribution leads to the creation of the Box–Cox power exponential, $BCPE(\mu, \sigma, \nu, \tau)$, and the Box-Cox-$t$, $BCT(\mu, \sigma, \nu, \tau)$, distributions, see Rigby and Stasinopoulos (2004) and Rigby and Stasinopoulos (2006), respectively. The resulting centile estimation methods were called LMSP and LMST respectively.

3.2.3 Analysis of height against age

The current `R` implementation of GAMLSS has two similar sets of functions to fit the *LMS*, *LMSP* and *LMST* distributions: the functions *BCCGo*, *BCPEo* and *BCTo*, and the functions *BCCG*, *BCPE* and *BCT*. Their only difference is in the default link function $g_1(\mu)$ for μ. The functions *BCCGo*, *BCPEo* and *BCTo* have a log link for μ as a default, that is, $g_1(\mu) = \log(\mu)$, (which ensures that μ is always positive),

Table 1 Summary of the fitted models for the Dutch boys data, showing the effective degrees of freedom (*df*) used in the model, the global deviance (GD), the AIC ($k = 2$) and BIC

Distribution	*df*	*GD*	*AIC*	*BIC*
NO	25.7	41 842.5	41 893.8	42 069.4
BCCGo	27.7	41 828.7	41 884.1	42 073.3
BCPEo	29.9	41 807.6	41 867.5	42 072.0
BCTo	30.4	41 806.4	41 867.3	42 075.5

while *BCCG*, *BCPE* and *BCT* have the identity, that is, $g_1(\mu) = \mu$ (as was originally used by Cole and Green (1992)). Note that alternative link functions can be tried and the preferred link function is the one which gives the smallest value of the generalized Akaike information criterion, $GAIC(k)$, Akaike (1983), for chosen value of k.

First, to avoid estimating the power parameter ξ in model (3.4), we use an empirical method, which consists of plotting the response variable height (its log since we use a log link $g_1(\mu) = \log(\mu)$ function for μ) against the age, log age and square root of age (corresponding to ξ effectively equals to 1, 0 or 0.5, respectively), and choose the one which looks more linear and is therefore easier to smooth. The square root of age was the best choice in our case. In the rest of the analysis, we use $u = \sqrt{age}$. Note that estimation of the smoothing parameters is done automatically in GAMLSS, using the methodology described in Rigby and Stasinopoulos (2013) and in Stasinopoulos et al. (2017), Ch 3. [Note that there are also functions in GAMLSS for automatically choosing the power parameter ξ see Stasinopoulos et al. (2017), Ch 3].

To choose between the *NO*, *BCCGo*, *BCPEo* and *BCTo* distributions, we use the generalized Akaike information criterion $GAIC(k)$, where $GAIC(2)$ is the standard AIC, while $GAIC(\log n)$ is the bayesian information criterion, BIC, see Table 1. According to the AIC, the *BCTo* distribution is best, while for BIC, all the values were similar with the *NO* distribution best. Figure 2(a) presents the centiles curves for the *BCTo* distribution.

Diagnostics plots based on the residuals are a good way of checking the adequacy of a model. The GAMLSS methodology uses the normalized quantile residuals (or z-scores; Dunn and Smyth, 1996) which apply to all distributions. More on the residual diagnostic checks, including normalized quantile residuals and the worm plots shown in Figure 2(b), can be found in Stasinopoulos et al. (2017), Ch 12. Figure 2(b) presents the worm plot for the *BCTo* distribution. A worm plot is a detrended Q-Q plot of the normalized quantile residuals, with elliptical curves indicating approximate 95% point-wise confidence bands. Ideally the points in the worm plot should be close to the horizontal line in the middle with no systematic shape and 95% or more of the points inside the elliptical curve. The worm plot in Figure 2(b) is acceptable as there are no points outside the confidence bands.

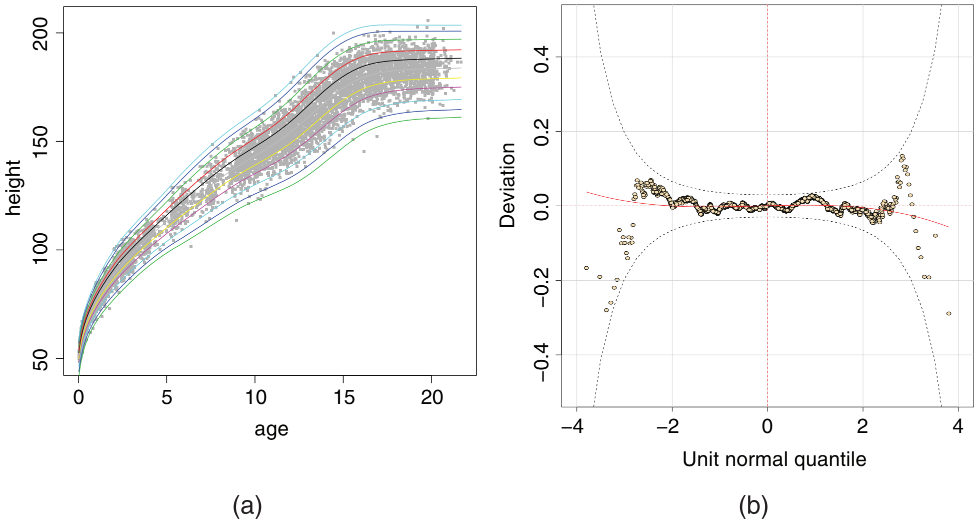

Figure 2 (a) Plot of the centile curves for the *BCTo* model and (b) worm plot of its residuals

The resulting *BCTo* model is given by

$$\begin{aligned} \text{height} &\sim BCTo(\mu, \sigma, \nu, \tau) \\ \log(\mu) &= s_1(u) \\ \log(\sigma) &= s_2(u) \\ \nu &= s_3(u) \\ \log(\tau) &= s_4(u) \\ u &= \sqrt{age}. \end{aligned} \tag{3.5}$$

[We also used the `lms()` function in the **gamlss** package, which uses an automated procedure for the LMS (*BCCGo*), LMSP (*BCPEo*) and LST (*BCTo*) methods of centile estimation, including an automated estimation of both the power parameter ξ and the smoothing parameters for all the distribution parameters. It also chooses between *BCCGo*, *BCPEo* and *BCTo* distribution models using criterion $GAIC(k)$, with k chosen by the user. Here, we used $k = 4$, a compromise between $AIC(k = 2)$ and BIC (with $k = log(n)$). The `lms()` function chose *BCPEo* with estimated power parameter approximately 0.68, quite close to 0.5, that is, to the square root that we chose earlier. The resulting fitted centile curves are very similar to the fitted centile curves presented in Figure 2(a).]

3.3 Centile estimation using two explanatory variables

This section shows how centile estimation can be constructed when we have two explanatory variables. The response variable is head circumference (`head`), and the explanatory variables are height (`ht`) and age (`age`). The objective here is to model the distribution of head circumference using height and age. The data are shown in two dimensions in Figure 1(a) and 1(b), and in three dimensions in Figure 3, where we can see that the y variable, head circumference, is defined only in a limited joint range of the age and height space. This has consequences in fitting a model to the head circumference because prediction outside the data space of the explanatory variables will rely on extrapolation and therefore will be unreliable.

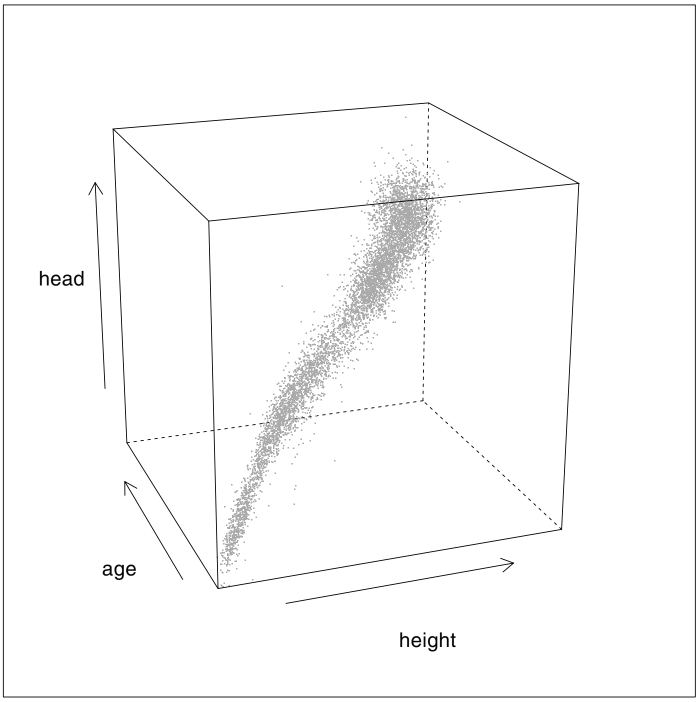

Figure 3 A three-dimensional plot of head circumference against age and height

In order to fit a GAMLSS model to the data, we need:

1. to consider whether transformed versions of height and age will help the analysis;
2. a suitable distribution for head circumference; and

3. to determine how height and age affect the parameters of the distribution of the response variable.

In order to choose a suitable transformation for age and height, we plotted head circumference (its log since we use a log link $g_1(\mu) = \log(\mu)$ function for μ) against each of (a) age and height, (b) the log age and height and (c) the square root of age and height (the six plots are not shown here). The transformation that makes the relationship between log (head) and each transformed explanatory variable closest to linear was the log transformation for both of height and age.

To choose a suitable distribution, we used three distributions (*BCCGo*, *BCPEo* and *BCTo*) defined for a positive response variable (since head circumference is always positive) and fit initially a smooth function in log height for μ only. The distribution that was 'best' using the *AIC* or *BIC* was the $BCTo(\mu, \sigma, \nu, \tau)$ distribution.

In order to find out how the explanatory variables, age and height, affect the different distribution parameters of the $BCTo(\mu, \sigma, \nu, \tau)$ distribution, we use a selection technique, which chooses between the following four possible models for each distribution parameter:

- $s(u_h)$: main smooth effect for height;
- $s(u_a)$: main smooth effect for age;
- $s(u_a) + s(u_h)$: additive smooth effect for age and height; and
- $s(u_a, u_h)$: smooth interaction of age and height,

where $u_h = \log\text{(height)}$ and et $u_a = \log\text{(age)}$.

The chosen model was

$$\begin{aligned} \text{head} &\sim BCTo(\mu, \sigma, \nu, \tau) \\ \log(\mu) &= s_1(u_a, u_h) \\ \log(\sigma) &= s_2(u_h) \\ \nu &= s_3(u_a) \\ \log(\tau) &= s_4(u_h), \end{aligned} \tag{3.6}$$

where $s_1(.)$ is a smooth surface and $s_2(.)$, $s_3(.)$ and $s_4(.)$ are smooth functions. The worm plot for the final chosen model (3.6), given in Figure 4(a), showed some extreme outliers in the tails (6 in the upper tail and 5 in the lower tail, out of 6 885 observations). These extreme outliers distort the fitted centiles of head circumference. When the 11 extreme outliers were removed, the worm plot improved greatly. Figure 4(b) shows the worm plot of the residuals for the *BCTo* model removing the 11 observation with extreme residuals.

Model (3.6) can now be used to obtain the 5% and 95% centiles of head, each of which is plotted as a contour plot against height and age in Figure 5. Figure 6 reduces the age range to 0–2 years to see more clearly the head contour values for that age group. From a practical viewpoint, given the height and age of a Dutch boy, an observed head circumference less than the 5% centile value indicates

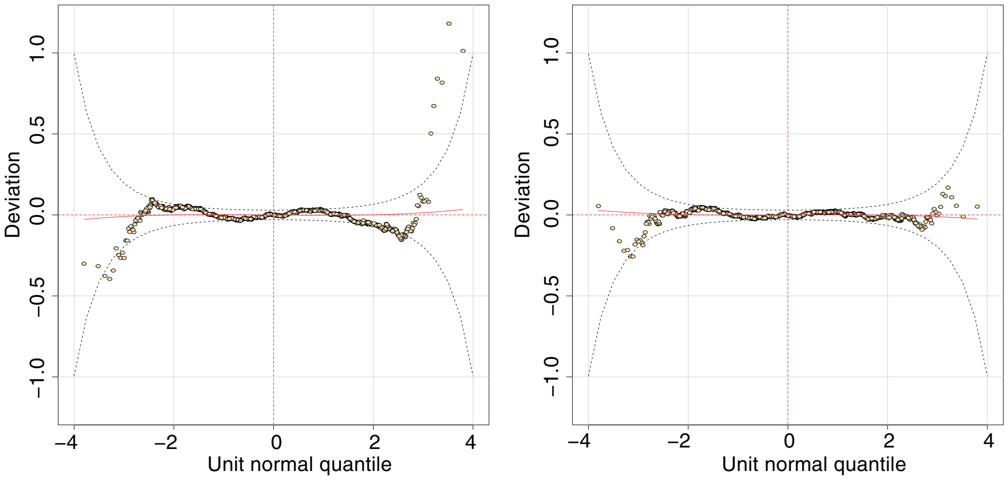

Figure 4 Worm plot of the residuals for the *BCTo* model (a) using all 6 885 observations and (b) removing 11 outlier observations with extreme residuals (right)

an unusually small head circumference (given the height and age), while a value greater than the 95% centile value indicates an unusually large head circumference (given the height and age). This can be used as a medical diagnostic tool. Note that for the very young boys, (less then 1 year), the contours are slanted (i.e., not horizontal or vertical), so both the height and age are important for determining the 5% and 95% centile values of head. However, for older boys, the contours are closer to horizontal, indicating that mainly the height determines the centile values of head.

4 Modelling a discrete count response variable within GAMLSS

4.1 Demand for medical care data

The data analysed in this section originates from the United States National Medical Expenditure Survey (NMES) conducted in 1987 and 1988, and is available from the **AER** package in `R` and is called NMES1988. [The data is cross-sectional, i.e., not repeated measurements.]

The response variable is the `visits` (i.e., number of physician office visits). The data frame `NMES1988` has 4 406 observations on 19 variables, but we consider only the following:

- `visits`: number of physician office visits,
- `hospital`: number of hospital stays,

- `health`: health status—a factor indicating whether self-perceived health is poor, average (reference category) or excellent,
- `chronic`: number of chronic conditions,
- `gender`: a factor indicating gender,
- `school`: number of years of education and

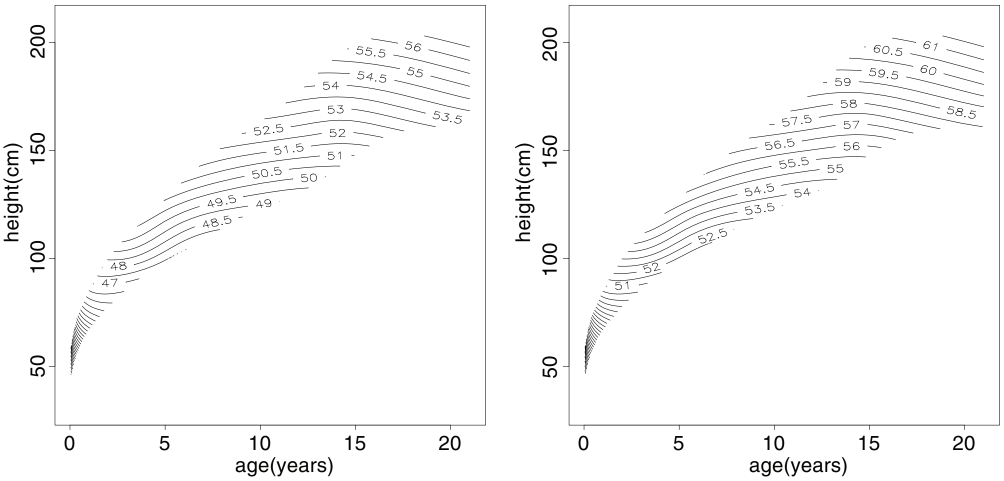

Figure 5 Contour plots for the (a) 5% and (b) 95% centiles of head against height and age (for all ages)

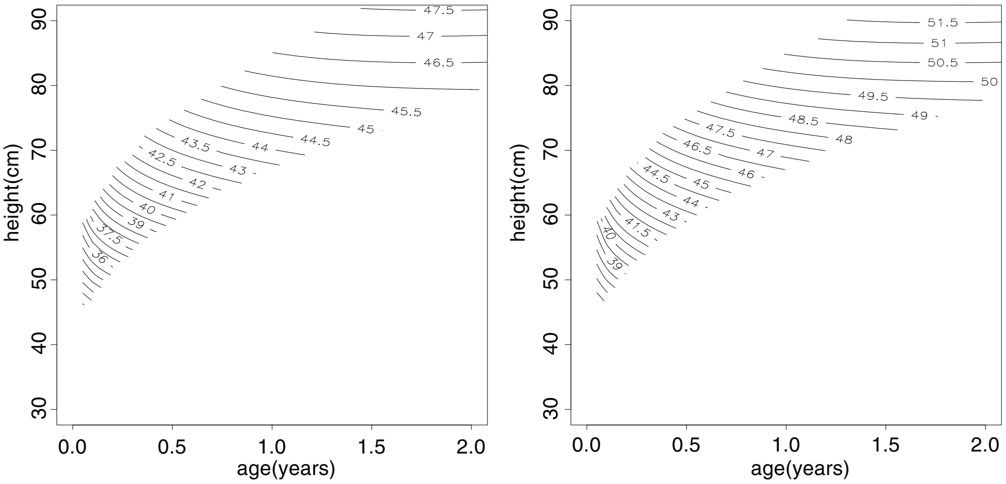

Figure 6 Contour plots for (a) 5% and (b) 95% centiles of head against height and age (for age from zero to two)

- `insurance`: a factor indicating whether the individual is covered by private insurance.

Figure 7 shows plots of the `visits` against each of the earlier explanatory variables and shows the complexity of this data. The response variable `visits` is a count type of variable with range $0, 1, 2, \ldots$. The plots of `visits` suggest possible relationships between mean `visits` and each of `hospital`, `chronic` and `school`. The remaining box and whisker plots display how the number of visits varies according to the categorical explanatory factors. The median of the number of visits (the horizontal bar in the middle of each box) decreases as the health status improves from poor to average and then to excellent. The median of the number of visits is similar for both male and female with a few higher values for male. The median of the number of visits is slightly higher for a person covered by a private insurance than for a person not covered. The variation (as measured by the interquartile range, the vertical length of the box) varies with the health status. The problem of skewness is prominent with longer upper than lower tails.

Any statistical model used for the analysis of the earlier data should be able to deal with overdispersion, high positive skewness and also an excess of zeros. The mean of the response variable `visits` depends on explanatory variables. Also the variance of the response variable `visits` may depend on its mean and/or explanatory variables. There is a clear indication of skewness in the distribution, which may also depend on explanatory variables.

4.2 Fitting different discrete distributions

One approach to deal with the complexity in this data is to fit different distributions and model each of the parameters of the distribution as linear or smooth functions of the explanatory variables.

The first natural attempt do deal with overdispersion is to fit the negative binomial distribution, $NBI(\mu, \sigma)$, modelling the mean parameter μ and the dispersion parameter σ with linear or smoothing terms in the explanatory variables. However, since it is only a two parameter distribution, it cannot also model the skewness or kurtosis in the response variable. A three– or four–parameter discrete distribution is needed for that. The Sichel, $SICHEL(\mu, \sigma, \nu)$, and the beta negative binomial, $BNB(\mu, \sigma, \nu)$, are three–parameter distributions, but did not provide good fits to the response variable in our case. This could be due to the excess of zeros in the response variable. A solution to the excess or deficiency of zero values in a specific discrete distribution is provided by the zero–inflated (ZI) or the zero–altered versions of the discrete distribution, respectively. In the analysis of this data, we have tried the zero inflated, *ZI*, and zero adjusted, *ZA*, distribution for the negative binomial distribution (*ZINBI*, *ZANBI*), the Sichel (*ZISICHEL*, *ZASICHEL*), the negative binomial family (*ZINBF*, *ZANBF*) and the beta negative binomial (*ZIBNB*, *ZABNB*). See Rigby et al. (2017) for the probability function and properties of each of these distributions.

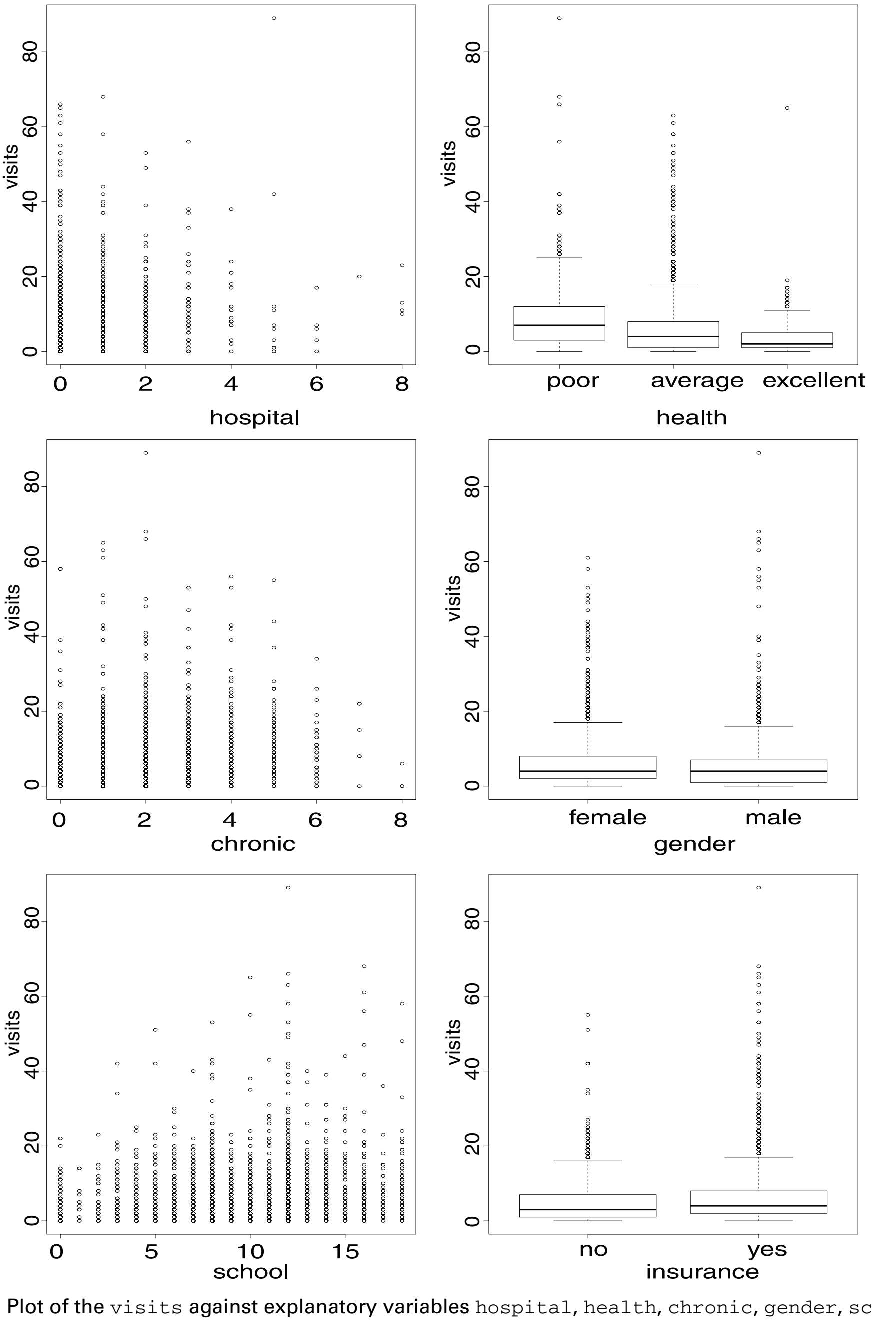

Figure 7 Plot of the `visits` against explanatory variables `hospital`, `health`, `chronic`, `gender`, `school` and `insurance`

The selection of the appropriate distribution is done in two stages, the fitting stage and the diagnostic stage. The fitting stage involves the comparison of different fitted models using a generalized Akaike information criterion (GAIC). The diagnostic stage involves the normalized randomized quantile residuals, or 'z-scores' (Dunn and Smyth, 1996), which provide information about the adequacy of the model and can be used with diagnostic plots such as worm plots, van Buuren and Fredriks (2001), or other test statistics, for example, Z-statistcs and Q-statistcs, Royston and Wright (2000). See also Stasinopoulos et al. (2017), Ch 12 for an explanation of these diagnostics. The initial selection of the link function is usually determined by the range of parameters, but can then be selected using the generalized Akaike information criterion, GAIC.

For each distribution parameter of each of the distributions considered, a step-wise selection of explanatory variable terms, $\sqrt{\text{hospital}}$, health, $\sqrt{\text{chronic}}$, gender, school and insurance, was applied using $GAIC(4)$ with penalty equal to 4 for each parameter, [since most terms have a single parameter and a 5% significance level, generalized likelihood ratio test for a single parameter being different from zero is based on an (asymptotic) Chi-squared distribution with critical value $\chi^2_{1,0.05} = 3.84 \approx 4$]. The transformations $\sqrt{\text{hospital}}$ and $\sqrt{\text{chronic}}$ were used because they were found to substantially improve the fit as judged by $GAIC(4)$. For more details about model selection, see Stasinopoulos et al. (2017), Ch 11.

First consider the negative binomial, *NBI*, distribution where the chosen fitted model using step-wise selection is given by

$$
\begin{aligned}
Y &\sim NBI(\hat{\mu}, \hat{\sigma}) \\
\log(\hat{\mu}) &= 0.80655 + 0.35579\sqrt{\text{hospital}} + 0.37545\sqrt{\text{chronic}} \\
&\quad + 0.02645\text{school} + 0.23068(\text{if health} = \text{poor}) \\
&\quad - 0.29802(\text{if health} = \text{excellent}) - 0.10838(\text{if gender} = \text{male}) \\
&\quad + 0.19523(\text{if insurance} = \text{yes}) \\
\log(\hat{\sigma}) &= 0.61390 - 0.45973\sqrt{\text{chronic}} + 0.26774(\text{if health} = \text{poor}) \\
&\quad - 0.08061(\text{if health} = \text{excellent}) + 0.20590(\text{if gender} = \text{male}) \\
&\quad - 0.16696\sqrt{\text{hospital}} - 0.49579(\text{if insurance} = \text{yes}).
\end{aligned}
\tag{4.1}
$$

The $NBI(\mu, \sigma)$ distribution for Y has mean μ and variance $\mu + \sigma\mu^2$.

However, the final chosen distribution was the zero inflated beta negative binomial, *ZIBNB* with chosen fitted model using step-wise selection, given by

$$
\begin{aligned}
Y &\sim \mathrm{ZIBNB}(\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau}),\\
\log(\hat{\mu}) &= 0.980 + 0.382\sqrt{\text{hospital}} + 0.332\sqrt{\text{chronic}}\\
&\quad + 0.025\text{school} + 0.255(\text{if health} = \text{poor})\\
&\quad - 0.313(\text{if health} = \text{excellent}) - 0.112(\text{if gender} = \text{male})\\
&\quad + 0.123(\text{if insurance} = \text{yes})\\
\log(\hat{\sigma}) &= -\,1.7026 - 0.208\sqrt{\text{chronic}} + 0.394(\text{if health} = \text{poor})\\
&\quad - 0.345(\text{if health} = \text{excellent}) + 0.197(\text{if gender} = \text{male})\\
\log(\hat{\nu}) &= -\,2.679 + 0.966\sqrt{\text{hospital}}\\
\log[\hat{\tau}/(1-\hat{\tau})] &= -\,1.077 - 0.744\sqrt{\text{chronic}} - 1.546(\text{if insurance} = \text{yes}),
\end{aligned}
\tag{4.2}
$$

where Y = visits. The *ZIBNB* model (4.2) has a $GAIC(4)$ of 24 030.55 which was clearly better than that of the *NBI* model (4.1) with $GAIC(4) = 24\,150.62$.

The four-parameter zero inflated beta negative binomial distribution, denoted $Y \sim ZIBNB(\mu, \sigma, \nu, \tau)$, is a mixture of $Y = 0$ with probability τ and $Y = Y_1$ with probability $(1-\tau)$, where $Y_1 \sim BNB(\mu, \sigma, \nu)$, (Rigby et al., 2017). Hence, τ is the probability of excess zeros.

The beta negative binomial, $BNB(\mu, \sigma, \nu)$ distribution is a reparameterization given by Rigby et al. (2017) of the distribution given in Wimmer and Altmann (1999), p.19. [However, the parameterization $BNB(\mu, \sigma, \nu)$ only includes distributions with a finite mean μ.] It is also called the beta Pascal distribution or the generalized Waring distribution.

The $BNB(\mu, \sigma, \nu)$ distribution has mean μ and is an overdispersed negative binomial distribution. The Waring, $WARING(\mu, \sigma)$, distribution (which is an overdispersed geometric distribution) is a special case of $BNB(\mu, \sigma, \nu)$, where $\nu = 1$. The negative binomial distribution is a limiting case of the beta negative binomial, since $BNB(\mu, \sigma, \nu) \to NBI(\mu, \nu)$ as $\sigma \to 0$ (for fixed μ and ν). If $Y \sim BNB(\mu, \sigma, \nu)$, then for large y, $P(Y = y) \sim ay^{-(\sigma^{-1}+2)}$, where a does not depend on y. Hence, the $BNB(\mu, \sigma, \nu)$ distribution has a heavy right tail, specially for large σ. Clearly, parameter σ is a right-tail heaviness parameter.

In order to interpret the parameters of $Y \sim ZIBNB(\mu, \sigma, \nu, \tau)$, μ is the mean of the $BNB(\mu, \sigma, \nu)$ component, σ is a right-tail heaviness parameter for the $BNB(\mu, \sigma, \nu)$ component, ν increases the variance (for $\nu^2 > \sigma/\mu$ and $\sigma < 1$, while the variance is infinite for $\sigma \geq 1$), and τ is the probability of excess of zeros. The mean of Y is $E(Y) = (1-\tau)\mu$.

Figure 8 displays the fitted parametric terms in $\log(\hat{\mu})$ in the final chosen model (4.2). Their effects are additive for $\log(\hat{\mu})$ and hence multiplicative for the fitted mean visits $(1-\hat{\tau})\hat{\mu}$. Assuming all other explanatory variables are fixed, then, due to $\hat{\mu}$, the fitted mean `visits` increases with the number of hospital stays (`hospital`), the number of chronic conditions (`chronic`) and the number of years schooling (school). A poor self-perceived health status results in a 29% [calculated from the

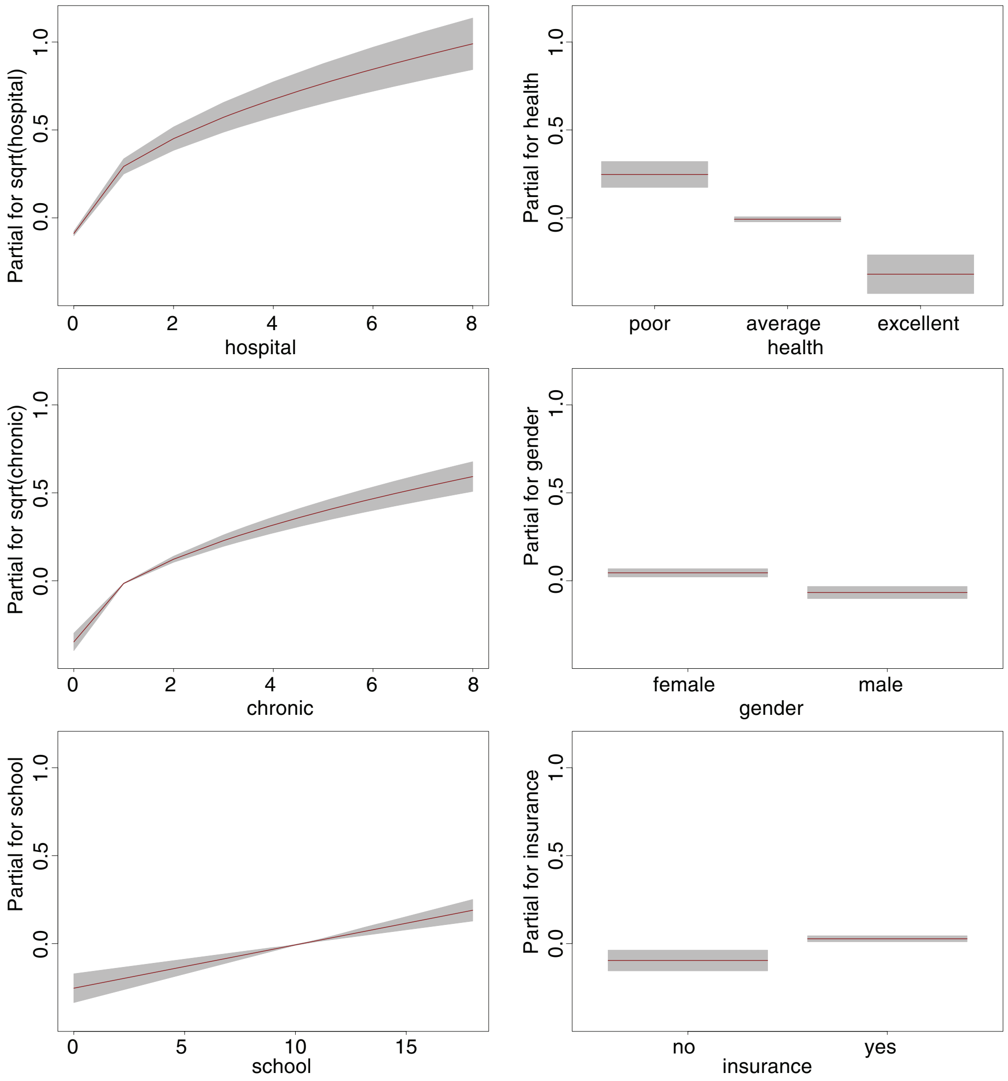

Figure 8 Term plots for μ for the fitted ZIBNB distribution

parameter estimate by $(e^{0.255} - 1) \times 100]$ 'increase' in fitted mean visits (relative to an average health) and an excellent health results in a 26.8% 'decrease' in fitted mean visits (relative to an average health). A male results in a 10.6% decrease in fitted mean `visits`. Being covered by a private insurance increases the fitted mean visits by 13.1%, due to $\hat{\mu}$, but also results in an additional increase due to $(1 - \hat{\tau})$.

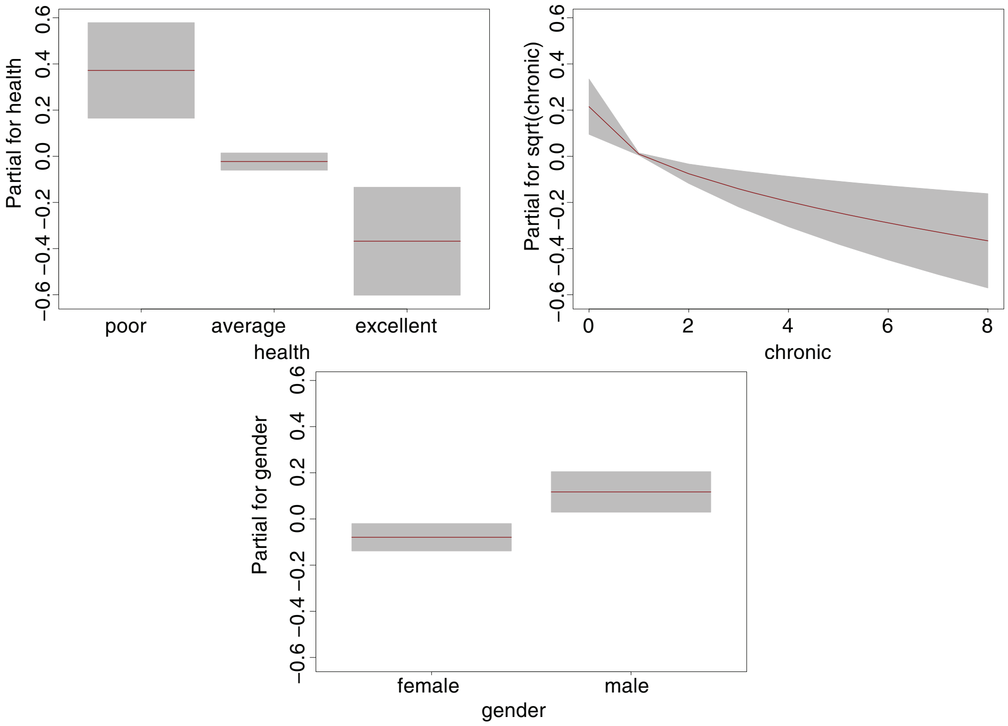

Figure 9 Term plots for σ for the fitted ZIBNB distribution

Also increasing `chronic` results in an additional increase in the fitted mean visits due to $(1 - \hat{\tau})$.

Figure 9 displays the fitted parametric terms in $\log(\hat{\sigma})$ in (4.2). Since $\hat{\sigma}$ controls the heaviness of the right tail of the distribution of visits, this heaviness increases with gender male and with poor health (relative to average health), but decreases with the number of chronic conditions and if the health is excellent (relative to average health), assuming all other explanatory variables are fixed.

Figure 10 displays the fitted parametric terms in $\log(\hat{\nu})$ in (4.2), showing that $\hat{\nu}$ increases with the number of hospital stays.

Figure 11 displays the fitted parametric terms in $\log[\hat{\tau}/(1 - \hat{\tau})]$ in (4.2). Since $\hat{\tau}$ is the fitted probability of excess zeros, this decreases with the number of chronic conditions and if the person is covered by private insurance. Since $E(Y) = (1 - \tau)\mu$, the fitted mean number of visits increases with the number of chronic conditions and with private insurance due to $(1 - \hat{\tau})$.

Figure 12 compares the worm plots of the $NBI(\mu, \sigma, \nu)$ and the $ZIBNB(\mu, \sigma, \nu, \tau)$ models given in (4.1) and (4.2), respectively, each with their distribution parameter terms chosen by a step-wise selection.

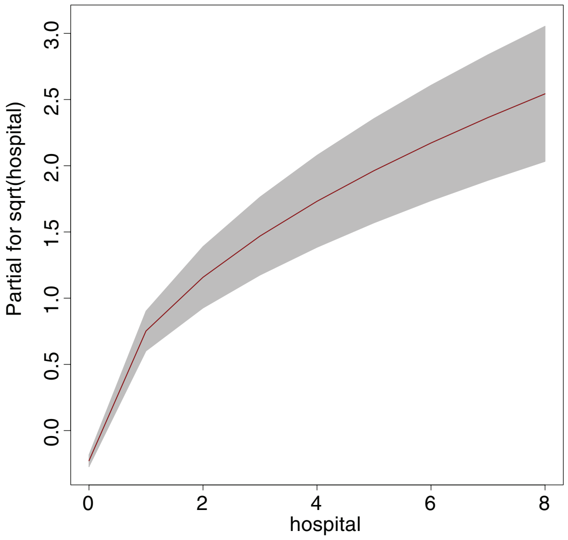

Figure 10 Term plot for ν for the fitted ZIBNB distribution

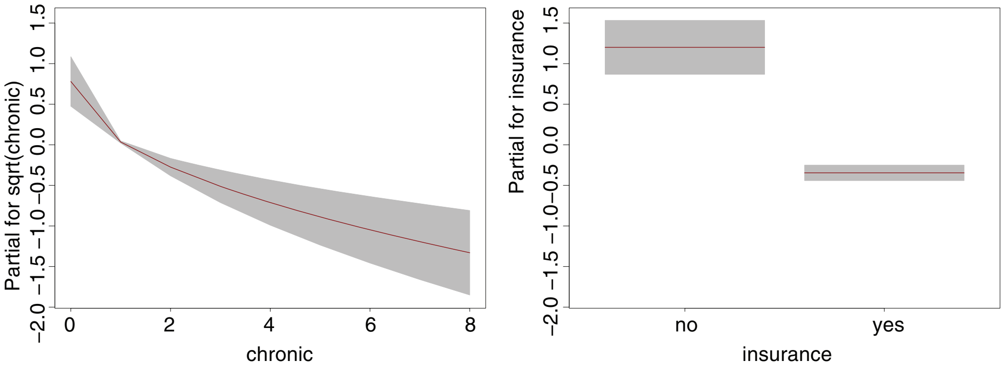

Figure 11 Term plots for τ for the fitted ZIBNB distribution

The worm plot in Figure 12(a) indicates that *NBI* does not provide a good fit to the data since many points lie well outside the elliptical (dashed) 95% point-wise interval bands, in the right tail. The worm plot in Figure 12(b) shows that *ZIBNB* provides a reasonable fit to the data. Clearly, the *ZIBNB* model provides a better fit as judged by the worm plot, but is still inadequate in the right tail.

There is an outlier not show in the right tail of Figure 12(a) because of its (vertical axis) deviation values is greater than 2. The outlier observation is case 1 522 and has `visits` = 65 and `health` = 'excellent'. It can be seen in the top-right plot of Figure 7, where it is very unusual. Omitting case 1 522 change the variable selected

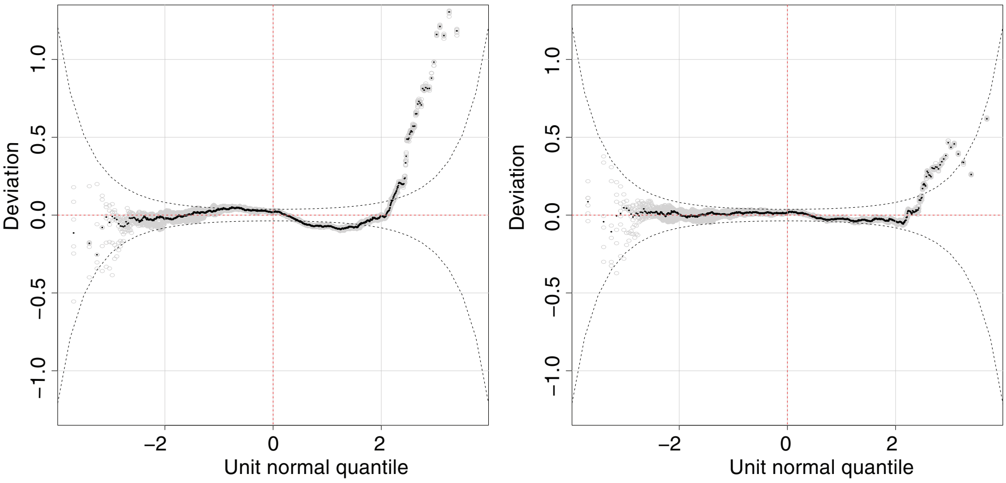

Figure 12 Worm plot of the randomized quantile residuals (a) for the *NBI* and (b) for the *ZIBNB* models. Grey points show 10 different realizations of the randomized quantile residuals, while black points show their median

in the NBI and ZIBNB models, but makes a considerable difference to the fitted parameter estimates for `health` =‘excellent’ in μ and σ for both the NBI and ZIBNB models, that is, (4.1), and (4.2) respectively. So maybe observation 1 522 should be omitted.

The worm plot provides a good diagnostic check of the fit of the discrete (count) distribution (especially of the right tail). An alternative diagnostic check (especially of the fit of the left tail of the discrete distribution) is given by the hanging rootogram, Kleiber and Zeileis (2016). This plot compares the (square root) observed (O_v) and expected (E_v) frequencies for $v = 0, 1, 2, \ldots, V$, where v is the number of physician visits. Hence, $O_v = \sum_{i=1}^{n} \mathrm{I}(y_i = v)$ is the observed number of patients having v visits, while $E_v = \sum_{i=1}^{n} P(Y_i = v|\hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$ is the expected number of patients having v visits obtained from the fitted model. Note I(.) is an indicator variable.

Figures 13(a) and 13(b) show the hanging rootogram for the chosen *NBI* and *ZIBNB* models, respectively. The curve shows the values of $\sqrt{E_v}$, while the vertical-shaded bars are drawn from $\sqrt{E_v}$, down to $\sqrt{E_v} - \sqrt{O_v}$, and hence the heights of the shaded bars are $\sqrt{O_v}$, for $v = 0, 1, 2, \ldots, 40$. For a ‘perfect’ fitted model, the bottom of the shaded bars would be aligned along the horizontal axis at 0. The plots also show the ‘warning limits’ of (Tukey, 1972, p. 314), set at ±1 [which are very rough 95% limits using the very rough approximate normal distribution with mean

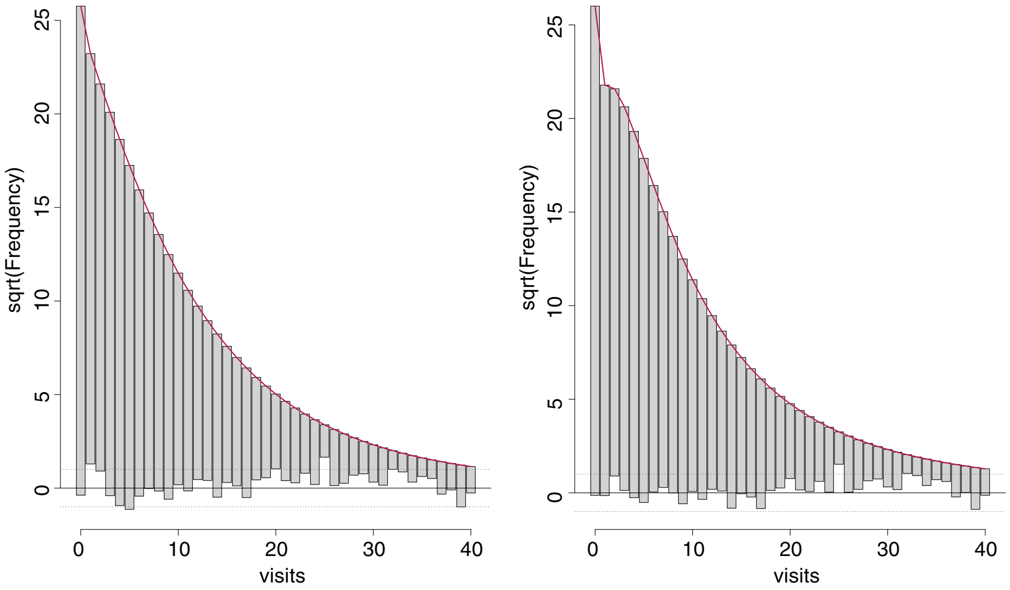

Figure 13 Rootogram (a) for the chosen *NBI* model and (b) for the chosen *ZIBNB* model

$\sqrt{E_v}$ and standard deviation 1/2, $NO(\sqrt{E_v}, 1/2)$, for $\sqrt{O_v}$]. He also suggests 'control' limits set at ±1.5 [which are very rough 99.8% limits].

Figure 13(a) shows six violations of the warning limits, including a potentially important one at visits = 1, suggesting that the *NBI* model may be inadequate. In addition, there is an important violation of the control limits for visits > 40 in the NBI model [where $E_{>40} = 10.20$ and $O_{>40} = 25$ and $\sqrt{E_{>40}} - \sqrt{O_{>40}} = -1.81$], indicating that the right tail of the NBI model is inadequate. In contrast, Figure 13(b) shows only two violations of the warning limits (which is to be expected for 41 values of v), suggesting that the *ZIBNB* model, is an improved and potentially adequate model according to this diagnostic. In the ZIBNB model, there is no violation of the control or warning limits for visits > 40, [since $E_{>40} = 17.88$ and $O_{>40} = 25$ and $\sqrt{E_{>40}} - \sqrt{O_{>40}} = -0.77$].

The adequacy of the fitted models can be further investigated by multiple worm plots (i.e., for different ranges of an explanatory variable), see van Buuren and Fredriks (2001) or Stasinopoulos et al. (2017), pp. 428–433 and analogously multiple rootograms.

5 Conclusions

This article illustrates the GAMLSS model using two real data examples, one with a continuous response and one with a discrete response variable. The examples show the flexibility of the GAMLSS model, in that different distributions can be fitted to the response variable and all distribution parameters can be modelled using linear functions or smooth non-parametric functions or surfaces of explanatory variables. The first example also shows how to obtain centile estimation curves of a continuous response variable using either one or two continuous explanatory variables.

The GAMLSS model is especially useful for data where modelling a response variable *Y* using the GLM or GAM is inadequate. In particular, GLM and GAM assume an exponential family distribution for *Y*. The exponential family is quite restrictive in the shape of the distribution. For example, it is unsuitable if a continuous response variable is negatively skew, or platykurtic, or leptokurtic unless positively skew. The GLM and GAM also model only the mean parameter using explanatory variables, and assume that the dispersion parameter (if there is one) is constant. Therefore, GLM and GAM cannot model the scale or shape of the distribution independently of the mean. The GAMLSS model in principle allows any distribution for the response variable. The current implementation in the **gamlss** package in `R` allows the user to choose between around 100 distributions with up to four parameters, allowing changes in location, scale and shape (e.g., skewness and kurtosis) to be modelled. The GAMLSS model includes the GLM and GAM as submodels so they can be fitted in the **gamlss** package, although an alternative `R` package is **mgcv** (Wood, 2017).

In a regression situation for a continuous response variable, especially for quantile (or centile) estimation, two alternative approaches to the GAMLSS model and packages are quantile regression Koenker (2017a), [using the **quantreg** `R` package Koenker (2017b)] and conditional transformation models (e.g., Hothorn (2018a)) [using the **mlt** `R` package Hothorn (2018b)]. An introduction to quantile regression is given by Waldmann (2018).

An alternative approach to GAMLSS for mean (and variance) estimation is generalized estimation equation (GEE) Hardin and Hilbe (2003).

For regression models for count data, an alternative to the gamlss packages is the **countreg** package on R-Forge, which also includes hurdle and ZI models to incorporate excess (or depleted) zeros, and functions for zero-truncated regression and finite mixture models. A Bayesian version of GAMLSS has been developed called BAMLSS and implemented in the **bamlss** `R` package.

The GAMLSS model and **gamlss** package have become standard for centile estimation (in particular using the *BCCGo*, *BCPEo* and *BCTo* distributions giving the *LMS*, *LMSP* and *LMST* methods of centile estimation, respectively; see, e.g., WHO, 2006, 2007, 2009). The alternative quantile regression approach is implemented in the **quantreg** (Koenker, 2017b) and **COBS** (Ng and Maechler, 2017) `R` packages.

The GAMLSS models are implemented in several packages existing in CRAN. The `R` code used in the two data analysis is available from www. gamlss.org. Further information about GAMLSS and using the **gamlss** packages is given in Stasinopoulos et al. (2017). Further information about the distributions used in the **gamlss** packages and their properties is given in Rigby et al. (2017).

References

Akaike H (1983) Information measures and model selection. *Bulletin of the International Statistical Institute*, **50**, 277–90.

Budge S, Ingolfsson A and Zerom D (2010) Empirical analysis of ambulance travel times: The case of calgary emergency medical services. *Management Science*, **56**, 716–23.

Cole TJ and Green PJ (1992) Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–19.

Dunn PK and Smyth GK (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–44.

Fahrmeir L, Kneib T, Lang S and Marx B (2013) *Regression: Models, methods and applications*. New York, NY: Springer-Verlag.

Fredriks A, van Buuren S, Burgmeijer R, Meulmeester J, Beuker R, Brugman E, Roede M, Verloove-Vanhorick S and Wit JM (2000a) Continuing positive secular change in The Netherlands, 1955–1997. *Pediatric Research*, **47**, 316–23.

Fredriks A, van Buuren S, Wit J and Verloove-Vanhorick SP (2000b) Body index measurements in 1996–7 compared with 1980. *Archives of Childhood Diseases*, **82**, 107–12.

Giraud G and Kockerols T (2015) *Making the European banking union macro-economically resilient: Cost of non-Europe report* (Report to the European Parliament). URL www.europart.europa.eu/thinktank/en document.html?reference=EPRS-STU(2015) 558771

Hardin JW and Hilbe JM (2003) *Generalized Estimating Equations*. Boca Raton, FL: Chapman & Hall/CRC.

Hastie TJ and Tibshirani RJ (1990) *Generalized Additive Models*. London: Chapman & Hall.

Hawkins E, Fricker TE, Challinor AJ, Ferro CA, Ho CK and Osborne TM (2013) Increasing influence of heat stress on French maize yields from the 1960s to the 2030s. *Global Change Biology*, **19**, 937–47.

He X and Ng P (1999) Cobs: Qualitative constrained smoothing via linear programming. *Computational Statistics*, **14**, 315–37.

Heller G, Stasinopoulos D, Rigby R and De Jong P (2007) Mean and dispersion modelling for policy claims costs. *Scandinavian Actuarial Journal*, **2007**, 281–92.

Hothorn T (2018a) Top-down transformation choice. *Statistical Modelling* (to be published).

——— (2018b) mlt: Most likely transformations. *R package version 0.2-2*. URL https:// CRAN. R-project.org/package=mlt (last accessed 5 March 2018).

International Monetary Fund (2015) United States. Financial Sector Assessment Program. Stress Testing: Technical note (IMF Country Report No. 15/173). URL https:// www.imf.org/en/Publications/CR/issues/ 2016/12/31/United-States-Financial-Assessment-Program-Stress-Testing-Technical Notes-43058

Khondoker MR, Glasbey C and Worton B (2007) A comparison of parametric and non-parametric methods for normalising cDNA microarray data. *Biometrical Journal*, **49**, 815–23.

Kleiber C and Zeileis A (2016) Visualizing count data regressions using rootograms. *The American Statistician*, **70**, 296–303.

Kneib T (2013) Beyond mean regression. *Statistical Modelling*, **13**, 275–303.

Koenker R (2017a) Quantile regression: 40 years on. *Annual Review of Economics*, **9**, 155–76.

——— (2017b) *quantreg: Quantile Regression*. URL R package version 5.3.5. https:// CRAN.R-project.org/package=quantreg. (last accessed 5 March 2018).

Koenker R, Ng P and Portnoy S (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–80. doi: 10.1093/biomet/81.4.673

Nelder JA and Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–84.

Neuhauser HK, Thamm M, Ellert U, Hense HW and Rosario AS (2011) Blood pressure percentiles by age and height from nonoverweight children and adolescents in Germany. *Pediatrics*, doi: 10.1542/peds.2010-1290

Ng P and Maechler M (2007) A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, **7**, 315–28.

——— (2017) cobs: COBS, constrained B-splines (Sparse matrix based). *R package version* 1.3-1. URL http://CRAN.R-project.org/package=cobs (last accessed 5 March 2018).

Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, Enright PL, Hankinson JL, Ip MS, Zheng J et al. (2012) Multi-ethnic reference values for spirometry for the 3–95-yr age range: The global lung function 2012 equations. *European Respiratory Journal*, **40**, 1324–43.

Rigby RA and Stasinopoulos DM (2004) Smooth centile curves for skew and kurtotic data modelled using the Box–Cox power exponential distribution. *Statistics in Medicine*, **23**, 3053–76.

——— (2005) Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, **54**, 507–54.

——— (2006) Using the Box–Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**, 209–29.

——— (2013) Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statistical Methods in Medical Research*, **23**, 318–32. doi: 10.1177/0962280212473302

Rigby RA, Stasinopoulos DM, Heller GZ and De Bastiani F (2017) Distributions for modelling location, scale, and shape: Using GAMLSS in R. URL www.gamlss.org. (last accessed 5 March 2018).

Rodrigues J, de Castro M, Cancho VG and Balakrishnan N (2009) Com–Poisson cure rate survival models and an application to a cutaneous melanoma data. *Journal of Statistical Planning and Inference*, **139**, 3605–11.

Royston P and Wright EM (2000) Goodness-of-t statistics for age-specic reference intervals. *Statistics in Medicine*, **19**, 2943–62.

Stasinopoulos DM and Rigby RA (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, **23**, 1–46.

Stasinopoulos DM, Rigby RA, Heller GZ, Voudouris V and De Bastiani F (2017) *Flexible Regression and Smoothing: Using GAMLSS in R*. Boca Raton, FL: Chapman & Hall.

Tukey J (1972) Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor,* edited by TA Bancroft. Vol. V, pages, 293–316.

van Buuren S and Fredriks M (2001) Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259–77.

Villar J, Ismail LC, Victora CG, Ohuma EO, Bertino E, Altman DG, Lambert A, Papageorghiou AT, Carvalho M, Jaer YA, et al. (2014) International standards for newborn weight, length, and head circumference by gestational age and sex: the newborn cross-sectional study of the intergrowth-21st project. *The Lancet*, **384**, 857–68.

Villarini G, Smith J, Serinaldi F, Bales J, Bates P and Krajewski W (2009) Flood frequency analysis for nonstationary annual peak records in an urban drainage basin. *Advances in Water Resources*, **32**, 1255–66.

Visser GH, Eilers PH, Elferink-Stinkens PM, Merkus HM and Wit JM (2009) New Dutch reference curves for birthweight by gestational age. *Early Human Development*, **85**, 737–44.

Voudouris V, Ayres R, Serrenho AC and Kiose D (2015) The economic growth enigma revisited: The EU-15 since the 1970s. *Energy Policy*, **86**, 812–32.

Waldmann E (2018) Quantile regression: A short story on the how and why. *Statistical Modelling* (to be published).

WHO MGRSG (2006) *WHO Child Growth Standards: Length/Height-for-Age, Weight-*

for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development. Geneva: World Health Organization.

WHO MGRSG (2007) *WHO Child Growth Standards: Head Circumference-for-Age, Arm Circumference-for-Age, Triceps Circumference-for-Age and Subscapular Skinford-for-Age: Methods and Development*. Geneva: World Health Organization.

——— (2009) *WHO Child Growth Standards: Growth Velocity Based on Weight, Length and Head Circumference: Methods and Development*. Geneva: World Health Organization.

Wimmer G and Altmann G (1999) *Thesaurus of Univariate Discrete Probability Distributions*. Essen: Stamm Verlag.

Wood SN (2017) *Generalized Additive Models. An Introduction with R*. Second edition. Boca Raton, FL: Chapman & Hall.