General Methods for Analyzing Bounded Proportion Data

Abu Hossain

Faculty of Life Sciences and Computing

London Metropolitan University

A thesis submitted in partial fulfilment of the requiremnets of London Metropoltan University for the degree of

Doctor of Philosophy

June 2017

I would like to dedicate this thesis to my loving parents ...

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

Abu Hossain

June 2017

Acknowledgements

This research would not have been possible without the support of many people. First and foremost the author wish to express his deepest gratitude to his supervisors Professor Mikis Stasinopoulos and Dr Robert Rigby of STORM research centre, who were abundantly helpful and offered invaluable assistant, support and guidance.

A special thanks to Dr Marco Enea for his continued support while he visited STORM research centre for a short period of time. The author also grateful to Dr Vlasios Voudouris for introducing him to the STORM research centre to start his PhD.

In addition the author is very grateful to Dr Robert Rigby and professor Mikis Stasinopoulos for all he has learned from them and for their continuous help and support in all stages of this thesis and research.

The author is also like to convey thanks to Professor Robert Gillchrist former Director of STORM research centre for his valuable comments and advice.

The author would like to take this opportunity to express love and gratitude to his beloved families, specially his parents and two sisters for their understanding and endless love and support, through the duration of his studies. A very special thanks to Omi, Pupa, Lupin and Laron.

Thanks to the supportive friends and colleagues, Dr Pargat Calay, Dr Sundus Tewfik, Dr Deepti Ratnayake, Dr Majid Djennad, Dr Fernanda De Bastiani, Dr Luiz Nakamura, Dr Khurram Majeed, Samson Habte and Roberta Freezor.

Abstract

This thesis introduces two general classes of models for analyzing proportion response variable when the response variable Y can take values between zero and one, inclusive of zero and/or one. The models are inflated GAMLSS model and generalized Tobit GAMLSS model. The inflated GAMLSS model extends the flexibility of beta inflated models by allowing the distribution on (0,1) of the continuous component of the dependent variable to come from any explicit or transformed (i.e. logit or truncated) distribution on (0,1) including highly skewed and/or kurtotic or bimodal distributions. The second proposed general class of model is the generalized Tobit GAMLSS model. The generalized Tobit GAMLSS model relaxes the underlying normal distribution assumption of the latent variable in the Tobit model to a very general class of distribution on the real line. The thesis also provides likelihood inference and diagnostic and model selection tools for these classes of models. Applications of both the models are conducted using different sets of data to check the robustness of the proposed models. The originality of the thesis starts from chapter 4 and in particular chapter 5, 6 and 7 with applications of models in chapter 8, 9 and 10.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background and introduction

Proportion data can be represented as percentage, fraction or ratio, Kieschnick and McCullough (2003). It can be observed on the open interval (0,1), semi-open interval including 0 or 1 (i.e [0,1), (0,1] respectively) or in closed form including 0 and 1 (i.e. [0,1]).[1]

The analysis of proportion data is a common phenomenon in many scientific fields. For example medical research (Hunger et al. (2011), Galvis et al. (2014)), finance (Cook et al. (2008)), econometrics (Papke and Wooldridge (1996), Ferrari and Cribari-Neto (2004)), operational research (Hoff (2007)) and ecology(Girão et al. (2007), Warton and Hui (2011), Nishii and Tanaka (2013), Baran and Nemoda (2016)).

In recent times modelling a proportion response variable has gained significant attention in the finance literature, especially for modelling loss given default values which are often bounded between 0 and 1 including 0 and 1. Researchers use various statistical methods for analysing LGD values in the finance literature. For example Hu and Perraudin (2002), Siddiqi and Zhang (2004), Gupton and Stein (2005), Bastos (2010), Qi and Zhao (2011), Yashkir and Yashkir (2013) and Li et al. (2014).

[1] [0,1], where square bracket indicates including the end point.

In the literature from various fields, different methods have been proposed to analyze proportion data. One of the well known strategies is to transform proportion response variable and then run ordinary least square regression using the transformed response variable; for example, the arcsine square root transformation Warton and Hui (2011), logit transformation ($\log(y/(1-y))$, inverse Gaussian transformation, Hu and Perraudin (2002), inverse Gaussian with beta transformation, Gupton and Stein (2005). Another popular approach beta regression has been established as a powerful technique for modelling proportion response variable on the open interval (0,1). [2]

Ospina and Ferrari (2012, 2010) proposed a distribution that is a mixture of a beta distribution and a Bernoulli distribution for modelling a response variable observed on [0,1) or (0,1] and termed it as a beta inflated model.

The censored normal distribution model (i.e. Tobit model) proposed by Tobin (1958) was originally used to analyse the data observed on [0,∞). The two sided version of the Tobit model by Rosett and Nelson (1975) is used to analyse the data observed on [0,1]. Sigrist and Stahel (2010) propose a more complex censored gamma and two tiered gamma regression model (i.e. a generalization of Tobit model) to analyze a response variable observed on [0,1].

Details of some of the recent parametric models for a proportion response variable along with transformed regression models are give in chapter 2.

This thesis focuses on modelling proportion response variable observed on the semi-closed intervals [0,1) and (0,1] and the closed interval [0,1], where the square bracket indicates that the end point is included. The proposed models generalize and extend two popular models (e.g. beta inflated and Tobit) to analyze a bounded proportion response variable.

1.2 Research motivation

Modelling proportion data can lead to misleading results, especially if the specific nature of the proportion data are not taken into account; Schmid et al. (2013). For example using

[2] Johnson et al. (1995), Kieschnick and McCullough (2003), Ferrari and Cribari-Neto (2004).

OLS regression for a response variable on (0,1) may result in a biased and inefficient model. [3] Similarly beta regression (Gupta and Nadarajah (2004)) is widely used in literature for modelling a response variable observed on the open interval, i.e. (0,1). However, the beta distribution may not be appropriate for modelling variable observed on the semi closed or closed interval i.e. [0,1), (0,1] or [0,1].

Moreover the survey of the published literature on modelling a proportion response variable by Kieschnick and McCullough (2003) and very recent observation by Li et al. (2014) suggest that there is no commonly accepted distributional models, nor any commonly accepted regression model for the bounded response variable proposed in the literature. For example the quasi likelihood method avoids the mean and variance specification error, however the model does not assume a proper distribution for the dependent variable, Hoff (2007). Galvis et al. (2014) pointed out that modelling a bounded proportion response variable using a logistic normal model Aitchison (1982) suffers from an interpretation problem given that the expected value of the response variable is not a simple logit function of the covariates. Li et al. (2014) criticize using a transform regression for modelling a proportion response variable on [0,1], since it may lead to biased estimates. The censored normal regression model (i.e. Tobit model) also has been criticized for its restrictive distributional (Normal) assumption. Similarly in the beta inflated model Ospina and Ferrari (2010), although the beta distribution can take different shapes, it only has two parameters so its flexibility is limited.

This research is motivated by the diversity of the practice of modelling a proportion response variable and questionable nature of some of these practices. The research focuses on modelling distributional categories [0,1), (0,1] and [0,1] and the proposed models in this research are motivated by the unique characteristics of those data types. The presence of skewness and heteroskedasticity are the common characteristics of continuous and bounded data type, Galvis et al. (2014). In order to address this issue, the proposed models can accommodate continuous distributions including highly skewed and kurtotic distributions. To address the excess zeroes

[3] Kieschnick and McCullough (2003) stated that, such an approach contravenes two conditions. First the conditional expectation must be nonlinear as it maps between 0 and 1. Second the variable must be heteroskedastic since the variance will approach zero as the mean approaches either boundary point.

and ones, in the proposed inflated GAMLSS model, the probability of the response variables equalling zero or one is modelled independently of the distribution on (0,1), and in the proposed generalized Tobit GAMLSS model, the probabilities that the response variable equals zero and one are directly related to the distribution on (0,1). Another distinct characteristic of the proposed models is that the models assume that all the parameters of the distribution of the response variable can be related to explanatory variables.

1.3 GAMLSS framework and proposed contributions

The proposed inflated GAMLSS model and generalized Tobit GAMLSS models adapt the generalized additive model for location, scale and shape (GAMLSS) model by Rigby and Stasinopoulos (2005) to focus on modelling the dependent variable. Applications include modelling distributional categories [0,1) (e.g. pass scheme data), (0,1] (e.g. lung function data) and [0,1] (e.g. loss given default data) using the inflated GAMLSS model and generalized Tobit GAMLSS models. This research applied to different data sets to check the robustness of the proposed models.

The proposed models are mixed continuous-discrete distributions, in particular a continuous distribution on (0,1)with point probabilities at 0 and/or 1. In this context the GAMLSS framework is used, where the response variable can have any distribution which may exhibit both positive or negative skewness and high or low kurtosis.

GAMLSS models are semi parametric regression type model. The GAMLSS model requires a parametric distribution assumption and is semi-parametric in the sense that, the parameters of the response variable distribution are each modelled as a function of explanatory variables, which may involve using non parametric smoothing functions.

In the proposed inflated GAMLSS model, the response variable is assumed to come from an explicit distribution on (0,1) or a transformed (e.g. logit or truncated) distribution from $(-\infty, \infty)$ to (0,1). In the generalized Tobit model, the response variable is assumed to come from a parametric censored distribution.

In this thesis the author follows the R implementation of GAMLSS and expands the distribution parameter vector θ to a maximum of six parameters (e.g. in the inflated GAMLSS model) and denotes the parameters as $\psi = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6) = (\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$, where μ and σ usually represent the location and scale parameters respectively and ν and τ usually denote shape parameters and ξ_0 and ξ_1 are parameters from the point probabilities at 0 and 1 respectively. In this research a four parameter bimodal distribution (bimodal skew symmetric normal distribution)[4] is also implemented within the package `gamlss.dist` in R. In the case of the bimodal skew symmetric normal distribution the parameter τ affects the bimodality of the distribution.

The inflated GAMLSS model comprises two components : the first component considers the discrete part of the response variable distribution and the second component is a continuous distribution on (0,1) which may consider the skewness and bimodality of the response variable. The inflated GAMLSS model can fit the discrete and continuous components separately, whereas in the generalized Tobit GAMLSS model the probabilities of the discrete end values depend on the distribution on (0,1). In both the methods each of the parameters of the response variable distribution can be modelled using explanatory variables.

1.4 Thesis outline

The thesis unfolds as follows. The thesis comprises eleven chapters and five appendices.

Chapter 2 provides a brief discussion of some of the commonly used approaches for modelling a proportion response variable.

Chapter 3 describes a number of explicit and transformed distributions on the open interval (0,1).

Chapter 4 presents a short introduction to the bimodal skew symmetric normal distribution and its implementation within the package `gamlss.dist` in R.

[4] Hassan and El-Bassiouni (2016)

Chapter 5 introduces the inflated GAMLSS model for the distributional categories [0,1), (0,1] and [0,1]. Chapter 5 includes an analysis of likelihood inference and residuals of the inflated GAMLSS model within GAMLSS framework.

Chapter 6 introduces the generalized Tobit GAMLSS model. Chapter 6 includes model definition, analysis of likelihood inference and residuals.

Chapter 7 introduces gamlssinf packages in R for fitting the inflated GAMLSS model. The package has been used for fitting real data in chapters 8, 9 and 10.

Chapter 8 presents an estimation of centile curve for the lung function data on (0,1] using the proposed models, along with other popular centile estimation techniques. Chapter 8 also highlights the comparison of model performance.

Chapter 9 shows an application of the proposed models to the PASS scheme data on [0,1).

Chapter 10 presents an application of proposed model to the loss given default data on [0,1]. Chapter 10 also shows the comparison of models performance using a cross validation technique.

Chapter 11 concludes the thesis suggesting future research opportunities in the area of modelling proportion response variable.

Chapters 2 and 3 essentially review previous work, while chapters 4 to 11 provide the original contributions of the thesis.

This thesis makes a number of original contributions to the area of modelling continuous and bounded data by developing a class of univariate inflated GAMLSS models which extend the flexibility of the beta inflated model (Hoff (2007), Ospina and Ferrari (2010), Cook et al. (2008)).

The inflated GAMLSS model is a mixed continuous-discrete distribution model which allows modelling of any or all the parameters of a distribution (up to four parameters for the continuous component and two extra parameters for the discrete components) using explanatory terms (e.g. linear and/or non-linear smoothing terms in explanatory variables).

For example the inflated GAMLSS model extends the two parameter beta inflated model by including two extra parameters in the continuous component for modelling the skewness and kurtosis or bimodality. Unlike the beta inflated model this thesis offers a comprehensive framework for the statistical analysis of the continuous data observed on the standard unit interval (0,1) with point masses at 0 and/or 1.

The inflated GAMLSS model is a general class of regression model for modelling a continuous proportion with discrete boundary values at zero and/or one. A method of estimating the parameters of the inflated GAMLSS model is explained in chapter 5. This research also explains the normalized randomized quantile residuals of the mixed continuous-discrete random variable for the inflated GAMLSS model in chapter 5.

In addition to the inflated GAMLSS model, this thesis also develops a new class of model, the generalized Tobit GAMLSS model, for the bounded proportion response variable. The generalized Tobit GAMLSS model extends the Tobit model in terms of the number of parameters and their flexibility. It includes two more parameters than the Tobit model to model the conditional skewness and/or kurtosis and bimodality of the response variable. The generalized Tobit GAMLSS model allows modelling all the parameters of the distribution of the latent variable V using linear and/or smoothing terms in explanatory variables.

This thesis also describes the normalized randomized quantile residuals of the generalized Tobit GAMLSS model to assess the overall adequacy of the model (see chapter 6). A method of estimating the parameters of the model is also described and explained in chapter 6.

The generalized Tobit GAMLSS for a proportion response variable comprises three special cases: censored below zero, censored above one and interval censored below 0 and above 1. Applications of the three sub-models of the generalized Tobit model together with popular models currently in the literature are shown in chapters 8, 9 and 10. In all the cases generalized Tobit GAMLSS model performed better than the other popular previous models.

Chapter 2

Review of models for proportion data on [0,1), (0,1] and [0,1]

2.1 Introduction

In this chapter the research surveys the literature and synthesizes the various practices used in prior literature for modelling distributional ranges [0,1), (0,1] and [0,1]. This chapter mainly focuses on more recent techniques as well as some of the old frequently used techniques. The chapter also composes a comprehensive review of the models investigated in previous literature. This research also investigates the performance of some of the methods described here.

2.2 PW fractional response regression

Based on the quasi likelihood function, Wedderburn (1974) , Papke and Wooldridge (1996) proposed a general method to model a continuous response variable on the interval 0 to 1 which may include 0 and/or 1. The basic assumption of the model is as follows

$$\mu = E(Y|\mathbf{x}) = G(\sum_{k=1}^{K} \beta_k x_k) \tag{2.1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_K)^T$ is a vector of explanatory variables of length K,

$$\beta = (\beta_1, \beta_2, \cdots, \beta_K)^T$$

is a vector of model parameters and $G(t)$ is the inverse of the logit function ($0 < G(t) < 1$ for all $t \varepsilon \mathbb{R}$) given by

$$G(t) = \frac{1}{1+e^{-t}} \tag{2.2}$$

A Bernoulli log likelihood function is used to estimate the β coefficients given by

$$l(\beta) = \sum y_i \log \mu_i + \sum (1-y_i) \log(1-\mu_i) \tag{2.3}$$

where $\mu_i = G(t_i)$ and $t_i = \sum_{k=1}^{K} \beta_k x_{ik}$

Estimators of the β coefficients are consistent and asymptotically normally distributed regardless of the distribution of the response variable Y.

Given the expectation value of Y estimated using fractional response regression, the effect of the explanatory variable x_m on the expected Y is given by

$$\frac{\partial \mu}{\partial x_m} = \frac{\beta_m e^{(-\Sigma \beta_k x_k)}}{(1+e^{-(\Sigma \beta_k x_k)})^2} \tag{2.4}$$

2.3 Tobit model (Tobin, 1958)

The Tobit model was first suggested by the pioneering work of Tobin (1958). Tobin (1958) proposed a hybrid model of probit analysis and multiple regression to model a lower bounded dependent variable. Many generalizations of the Tobit model, including estimations of those generalisations, have been proposed in the literature. The applications of the Tobit model and its various generalizations range over wide areas of economics, engineering and biometrics. The Tobit model and its generalizations are classified in different ways in the literature.

Tobin (1958) analysed the relationship between household expenditure on durable goods against household income. An important characteristics of the data were noted that there are several observations where the expenditure is zero (i.e. no house expenditure on luxury goods). A boundary constraint here can be expressed as $Y \geq L$ or $Y = 0$, where L is the lowest (non zero) value of Y. Suppose a latent variable V is introduced , where V is assumed to be normally distributed and L is the same for all the observations. Tobin solved the original problem in the following way

$$Y = \begin{cases} 0 & \text{if} \quad V \leq L \\ V & \text{if} \quad V > L \end{cases}$$

Assuming $L = 0$, then the Tobit model is defined in the following way.

$$Y = \begin{cases} 0 & \text{if} \quad V \leq 0 \\ V & \text{if} \quad V > 0 \end{cases}$$

where $V \sim N(\mu, \sigma^2)$, The cumulative distribution function of Y is given by

$$F_Y(y) = \begin{cases} P(V \leq 0) & \text{if} \quad y = 0 \\ F_V(y) & \text{if} \quad y \geq 0 \end{cases}$$

and mixed continuous-discrete probability (density) function given by

$$f_Y(y) = \begin{cases} P(V \leq 0) & \text{if} \quad y = 0 \\ f_V(y) & \text{if} \quad y \geq 0 \end{cases}$$

More generally for n observations from the Tobit model,

$$V_i \sim N(\mu_i, \sigma_i^2)$$

independently for $i = 1, 2, 3, \cdots, n$, where

$$Y_i = \begin{cases} 0 & \text{if} \quad V_i \leq 0 \\ V_i & \text{if} \quad V_i > 0 \end{cases}$$

and clearly V_i is latent when $V_i \leq 0$ Takeshi (1984). The cumulative distribution function of Y_i at 0 is given by

$$\begin{aligned} F_{Y_i}(0) &= p(V_i \leq 0) \\ &= p(Z_i \leq -\frac{\mu_i}{\sigma_i}) \qquad (2.5) \\ &= p(Z_i \geq \frac{\mu_i}{\sigma_i}) \qquad (2.6) \\ &= 1 - \Phi(\frac{\mu_i}{\sigma_i}) \qquad (2.7) \end{aligned}$$

where $Z_i \sim N(0,1)$ independently for $i = 1, 2, 3, \cdots, n$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable. The probability density function of the Tobit model for $y_i > 0$ is given by

$$\begin{aligned} f_{Y_i}(y_i) &= \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma_i}\right)^2} \\ &= \frac{1}{\sigma_i} \phi\left(\frac{y_i - \mu_i}{\sigma_i}\right) \end{aligned}$$

where $\phi(\cdot)$ is the probability density function of a standard normal variable.

The likelihood function for $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is given by

$$L = \prod_0 \left[1 - \Phi\left(\frac{\mu_i}{\sigma_i}\right)\right] \prod_1 \sigma_i^{-1} \phi\left[\frac{y_i - \mu_i}{\sigma_i}\right] \tag{2.8}$$

where the first product is over values of $y_i = 0$ and the second product is over values of $y_i > 0$. The resulting log likelihood function for $\mathbf{y}$ is given by

$$\log L = \sum_0 \log\left[1 - \Phi(\frac{\mu_i}{\sigma_i})\right] - \frac{n_1}{2}\log\sigma_i^2 - \frac{1}{2\sigma_i^2}\sum_1 (y_i - \mu_i)^2 \tag{2.9}$$

where n_1 is the number of non-zero values in $\mathbf{y}$

2.4 Two limit Tobit model (Rosett and Nelson, 1975)

Rosett and Nelson (1975) extended the limited (i.e. bounded) dependent variable technique developed by Tobin (1958) for both an upper limit and a lower limit. In the two limit Tobit regression model, the dependent variable is bounded below L_1 (e.g. $L_1 = 0$) and above L_2 (e.g. $L_2 = 1$) respectively. Let observed Y be a censored version of latent variable V. The observed variable Y is determined by

$$Y = \begin{cases} L_1 & \text{when} \quad V \leq L_1 \\ L_2 & \text{when} \quad V \geq L_2 \\ V & \text{when} \quad L_1 < V < L_2 \end{cases}$$

Assuming $L_1 = 0$ and $L_2 = 1$, then the mixed continuous-discrete probability (density) function of variable Y in the two sided Tobit model is defined by

$$f_Y(y|\mu,\sigma) = \begin{cases} P[Y=0] = \Phi(-\frac{\mu}{\sigma}) & \text{if} \quad y = 0 \\ \sigma^{-1}\phi(\frac{y-\mu}{\sigma}), & \text{if} \quad 0 < y < 1 \\ P[Y=1] = 1 - \Phi(\frac{(1-\mu)}{\sigma}) & \text{if} \quad y = 1 \end{cases}$$

where, $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of a standard normal distribution and $\mu = x^T\boldsymbol{\beta}$.

2.5 Censored gamma regression (Sigrist and Stahel, 2010)

Sigrist and Stahel (2010) introduce the censored gamma regression model as a generalization of the Tobit model. The model assumes that the underlying latent variable (V) of the model follows a gamma distribution shifted by $-\xi$. The probability density function of the shifted gamma distribution is given by

$$f_V(y|\mu,\sigma,\xi) = \frac{1}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}\Gamma(\frac{1}{\sigma^2})}(y+\xi)^{\frac{1}{\sigma^2}-1}e^{-\frac{(y+\xi)}{\sigma^2\mu}}$$

for $y > -\xi$, where, $\mu > 0$, $\sigma > 0$, $\xi > 0$. The mixed continuous-discrete probability (density) function of Y is then given by

$$f_Y(y|\mu,\sigma,\xi) = \begin{cases} P[Y=0] = F_V(0) \\ f_V(y|\mu,\sigma,\xi), & \text{if} \quad 0 < y < 1 \\ P[Y=1] = 1 - F_V(1) \end{cases}$$

$F_V(\cdot)$ and $f_V(\cdot)$ are the cdf and pdf of the latent variable V.

2.6 Inverse Gaussian regression (IGR)

Hu and Perraudin (2002) and Qi and Zhao (2011) use the inverse Gaussian regression (IGR) to model the response variable in the unit interval [0,1]. IGR includes transforming the response variable from unit interval (0,1) to $(-\infty,\infty)$ using the inverse Gaussian distribution function. The pdf of the Inverse Gaussian distribution function is given by

$$f_Y(y|\mu,\sigma) = \left[\frac{\sigma}{2\pi y^3}\right]^{\frac{1}{2}} e^{-\frac{\sigma(y-\mu)^2}{2\mu^2 y}}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$.

Using the transformed y variable on $(-\infty,\infty)$, they then run ordinary least square regression. The fitted value $(\hat{y})$ from ordinary least square regression is then transformed back from $(-\infty,\infty)$ to (0,1) using the Gaussian distribution function.

2.7 Inverse Gaussian regression with beta transformation (IGR-BT) (Gupton and Stein, 2005)

Gupton and Stein (2005) use the inverse-Gaussian regression with beta transformation to model the proportion response variable on [0,1]. The model assumes that the response variable Y follow a beta distribution.

$$Y = \Phi^{-1}[BE(y|a,b,min,max)]$$

where Φ^{-1} inverse of the normal cumulative distribution, a and b are the beta distribution parameters and min is set to 0 and max is set to 1. With the lower bound min fixed to 0, the beta distribution can be defined by

$$BE(y|a,b,min=0,max) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left(\frac{y}{max}\right)^{a-1}\left(1-\frac{y}{max}\right)^{b-1}\left(\frac{1}{max}\right)$$

The shape parameters (a,b) of the beta distribution can be estimated in terms of mean (μ) and standard deviation(σ) by the following way,

$$a = \frac{\mu}{max}\left[\frac{\mu(max-\mu)}{max\cdot\sigma} - 1\right] \qquad b = a\left[\frac{max}{\mu} - 1\right]$$

The cdf of Y is calculated using estimated parameters of the beta distribution and transform from (0,1) to $(-\infty,\infty)$ using the inverse Gaussian distribution function (Φ^{-1}). The next step of the model is to run ordinary least square regression. The fitted value ($\hat{y}$) from OLS regression is then transform back from $(-\infty,\infty)$ to (0,1) using the Gaussian distribution function. In the final step inverse beta regression is used to convert probabilities back from the unimodal Gaussian to bimodal beta distribution.

Both the transformation regression models (i.e. IGR & IGR-BT) are not defined when response variable y equals 0 or 1. Thus in transformation regression models, the Y variable on [0,1] needs to convert to (0,1) by adding (to 0) and substracting (from 1) a small adjustment factor. However transformation regression is very sensitive to the choice of adjustment factor, a very large or small adjustment factor may leads to a poor model fit.

2.8 Two-step approach (Gürtler and Hibbeln, 2013)

Modelling data on the closed interval, Gürtler and Hibbeln (2013) used a two-step approach. Step-1 includes an ordered logistic regression on the probability of Y falling into categories, where $Y = 0$, $Y = 1$ or $0 < Y < 1$. The steps are defined by

$$y_i = \begin{cases} p_0^i = lo(\pi_0 - x_i\beta), & \text{if} \quad y_i = 0 \\ p_{0,1}^i = lo(\pi_1 - x_i\beta) - lo(\pi_0 - x_i\beta), & \text{if} \quad 0 < y_i < 1 \\ p_1^i = 1 - lo(\pi_1 - x_i\beta), & \text{if} \quad y_i = 1 \end{cases}$$

where $lo(\cdot)$ denotes logistic function and π_0 and π_1 are the cut point parameters to be estimated.

In the second step the model considers ordinary least square (OLS) regression for response variable on $0 < Y < 1$. The predicted Y on $0 < Y < 1$ from OLS is defined by

$$\hat{\mu}_i = x_i\hat{\beta}$$

The the i^{th} value of Y (y_i) is predict as

$$\hat{E}(y_i) = \hat{\mu}_i \times (1 - \hat{p}_0^i - \hat{p}_1^i) + \hat{p}_1^i$$

Here $\hat{E}(y_i)$ is a weighted average of the model output from step 1 and step 2. The two-step normal model also suffers the interpretation problem given that the expected value of the response variable is not a simple logit function of the covariates, (Galvis et al., 2014).

2.9 Beta inflated model

Ospina and Ferrari (2012, 2010) developed the beta inflated model. The parameterizations in this section are used in the `gamlss` package.

2.9.1 Beta distribution inflated at 0 (BEINF0)

The beta inflated at 0 distribution for Y, denoted by $Y \sim BEINF0(\mu, \sigma, \nu)$, is a mixture of two components: a discrete value $Y = 0$ with probability p_0 and a continuous component on $(0 < Y < 1)$ with beta distribution, $BE(\mu, \sigma)$. The mixed continuous-discrete probability (density) function of $Y \sim BEINF0(\mu, \sigma, \nu)$ generated by the mixture is given by

$$f_Y(y|\mu, \sigma, \nu) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0)\frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0 < y < 1 \end{cases} \tag{2.10}$$

for $0 \leq y < 1$, where

$$\begin{pmatrix} \alpha \\ \beta \\ p_0 \end{pmatrix} = \begin{pmatrix} \frac{\mu(1-\sigma^2)}{\sigma^2} \\ \frac{(1-\mu)(1-\sigma^2)}{\sigma^2} \\ \frac{\nu}{(1+\nu)} \end{pmatrix}$$

Hence the parameters μ, σ, and ν can be defined by the following

$$\begin{pmatrix} \mu \\ \sigma \\ \nu \end{pmatrix} = \begin{pmatrix} \alpha(\alpha+\beta)^{-1} \\ (\alpha+\beta+1)^{-\frac{1}{2}} \\ p_0(1-p_0)^{-1} \end{pmatrix}$$

where $0 < \mu < 1$, $0 < \sigma < 1$ and $\nu > 0$.

Let $\eta = (\eta_1, \eta_2, \eta_3)$ be the predictors of parameters $\theta = (\mu, \sigma, \nu)$. The default link functions in the GAMLSS software for the parameters are given by

$$\begin{bmatrix} \log(\frac{\mu}{1-\mu}) \\ \log(\frac{\sigma}{1-\sigma}) \\ \log \nu \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}$$

where $\log \nu = \log[p_0/(1-p_0)]$.

The mean and the variance function Ospina and Ferrari (2010) of $Y \sim BEINF0(\mu, \sigma, \nu)$ are given by the following equations

$$\begin{aligned} E(Y) &= \frac{\mu}{1+\nu} \\ V(Y) &= \frac{\mu\sigma^2(1-\mu)}{1+\nu} + \frac{\mu^2\nu}{(1+\nu)^2} \end{aligned}$$

Let $\mathbf{y}^T = (y_1, y_2, ...y_n)$ be a random sample of independent response observations from a distribution with probability (density) function $f_Y(y|\mu, \sigma, \nu)$ given by (2.10), where $\theta^T = (\mu, \sigma, \nu)$ is a vector of parameters for the model BEINF0. Therefore the likelihood function for $\theta^T = (\mu, \sigma, \nu)$ given $\mathbf{y}$ is

$$\begin{aligned}
L(\theta) &= \prod_{0\leq y_i<1} f_Y(y_i|\mu,\sigma,\nu) \\
&= \prod_{y_i=0} f_Y(0|\nu) \prod_{0<y_i<1} f_Y(y_i|\mu,\sigma) \\
&= \prod_{y_i=0} p_0 \prod_{0<y_i<1} (1-p_0)\frac{1}{B(\alpha,\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1} \\
&= \prod_{y_i=0} \frac{\nu}{1+\nu} \prod_{0<y_i<1} \frac{1}{1+\nu}\left[\frac{1}{B(\alpha,\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}\right] \\
&= [(\tfrac{\nu}{1+\nu})^{n_0}(\tfrac{1}{1+\nu})^{n_1}] \prod_{0<y_i<1} [\frac{1}{B(\alpha,\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}]
\end{aligned}$$

where n_0 and n_1 are the number of zero and non-zero values of Y respectively, so $n_0+n_1=n$.

The likelihood function $L(\theta)$ of BEINF0 is factorises in two terms. The first term depends on only one parameter ν and the second term depends only on the parameters μ and σ (through α and β). The log likelihood function for the BEINF0 is given by

$$\begin{aligned}
l(\theta) &= \log(L(\theta)) \\
&= n_0\log(p_0) + n_1\log(1-p_0) + n_1\log[B(\alpha,\beta)] \\
&\quad +(\alpha-1)\sum_{0<y_i<1}\log(y_i) + (\beta-1)\sum_{0<y_i<1}\log(1-y_i)
\end{aligned}$$

2.9.2 Beta inflated distribution at 1 (BEINF1)

If we consider a probability mass at 1 instead of 0 in equation (2.10) then the model is called beta inflated at 1 (BEINF1). The BEINF1 model is a mixture of a discrete component $(Y=1)$ with probability p_1 and a continuous component $(0<Y<1)$ with a beta distribution $\text{BE}(\mu,\sigma)$. The probability (density) function of $Y\sim\texttt{BEINF1}(\mu,\sigma,\nu)$ on $(0,1]$ is given by

$$f_Y(y|\mu,\sigma,\nu) = \begin{cases} p_1 & \text{if } y=1 \\ (1-p_1)\frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0<y<1 \end{cases} \tag{2.11}$$

for $0 \leq y < 1$, where

$$\begin{pmatrix} \alpha \\ \beta \\ p_1 \end{pmatrix} = \begin{pmatrix} \frac{\mu(1-\sigma^2)}{\sigma^2} \\ \frac{(1-\mu)(1-\sigma^2)}{\sigma^2} \\ \frac{\nu}{(1+\nu)} \end{pmatrix}$$

Hence the parameters μ, σ, and ν can be defined by the following way

$$\begin{pmatrix} \mu \\ \sigma \\ \nu \end{pmatrix} = \begin{pmatrix} \alpha(\alpha+\beta)^{-1} \\ (\alpha+\beta+1)^{-\frac{1}{2}} \\ p_1(1-p_1)^{-1} \end{pmatrix}$$

where $0 < \mu < 1$, $0 < \sigma < 1$ and $\nu > 0$.

The default link functions of the parameters are given by

$$\begin{bmatrix} \log(\frac{\mu}{1-\mu}) \\ \log(\frac{\sigma}{1-\sigma}) \\ \log \nu \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}$$

where $\log \nu = \log(p_1/(1-p_1)$

The mean and the variance of the BEINF1 model Ospina and Ferrari (2010) are obtained by the following equations

$$\begin{aligned} E(y) &= \frac{\nu+\mu}{1+\nu} \\ V(y) &= \frac{\mu\sigma^2(1-\mu)^2}{1+\nu} + \frac{(1-\mu)(\mu+\nu)}{(1+\nu)} \end{aligned}$$

Let $\mathbf{y}^T = (y_1, y_2, ...y_n)$ be a random sample of observations from a distribution with probability (density) function $f_Y(y|\mu,\sigma,\nu)$ given by (2.11), where $\theta^T = (\mu,\sigma,\nu)$ is a vector of parameters for the model BEINF0. Therefore the likelihood function for $\theta^T = (\mu,\sigma,\nu)$ given $\mathbf{y}$ is

$$\begin{aligned}
L(\theta) &= \prod_{0<y_i\leq 1} f_Y(y_i|\mu,\sigma,\nu) \\
&= \prod_{y_i=1} f_Y(1|\nu) \prod_{0<y_i<1} f_Y(y_i|\mu,\sigma) \\
&= \prod_{y_i=1} p_1 \prod_{0<y_i<1} (1-p_1)\frac{1}{B(\alpha,\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1} \\
&= \prod_{y_i=1} \frac{\nu}{1+\nu} \prod_{0<y_i<1} \frac{1}{1+\nu}\left[\frac{1}{B(\alpha,\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}\right] \\
&= [(\tfrac{\nu}{1+\nu})^{n_1}(\tfrac{1}{1+\nu})^{n_2}] \prod_{0<y_i<1} [\frac{1}{B(\alpha,\beta)} y_i^{\alpha-1}(1-y_i)^{\beta-1}]
\end{aligned}$$

where n_1 and n_2 are the number of one and non-one values of Y respectively, so $n_1+n_2=n$

The likelihood function of BEINF1 is factorises in two terms. The first term depends on parameter ν and the second term depends on the parameters μ and σ. The log likelihood function is obtained by

$$\begin{aligned}
l(\theta) &= \log(L(\theta)) \\
&= n_1\log(p_1) + n_2\log(1-p_1) + n_2\log[B(\alpha,\beta)] \\
&\quad +(\alpha-1)\sum_{0<y_i<1}\log(y_i) + (\beta-1)\sum_{0<y_i<1}\log(1-y_i)
\end{aligned}$$

2.9.3 Beta inflated distribution at 0 and 1, BEINF(μ,σ,ν,τ)

The beta inflated distribution is suitable for a fractional response variable on $0\leq Y\leq 1$ that includes both zero and one.The beta inflated model is a mixture of a beta distribution and Bernoulli distribution. The model include three components: a discrete value 0 with probability p_0, a discrete value 1 with probability p_1 and a beta distribution BE (μ,σ) distribution on the unit interval $(0,1)$ with probability $(1-p_0-p_1)$.

Let Y be a random variable that assumes values in the closed interval $[0,1]$. The mixed continuous-discrete probability (density) function of the beta inflated distribution, denoted by BEINF (μ,σ,ν,τ), with respect to the measure generated by the mixture components is given by

$$f_Y(y) = \begin{cases} p_0 & \text{if } y = 0 \\ (1-p_0-p_1) f_W(y) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{2.12}$$

for $0 \leq y \leq 1$, where $W \sim BE(\mu, \sigma)$ has a beta distribution with $0 < \mu < 1$ and $0 < \sigma < 1$ and $p_0 = \nu/(1+\nu+\tau)$ and $p_1 = \tau/(1+\nu+\tau)$. Hence $\nu = p_0/p_2$ and $\tau = p_1/p_2$ where $p_2 = 1 - p_0 - p_1$. Since $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$, hence $\nu > 0$ and $\tau > 0$. Here $f_W(y)$ is given by

$$f_W(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

for $0 < y < 1$. where

$$\begin{pmatrix} \alpha \\ \beta \\ p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} \frac{\mu(1-\sigma^2)}{\sigma^2} \\ \frac{(1-\mu)(1-\sigma^2)}{\sigma^2} \\ \frac{\nu}{(1+\nu+\tau)} \\ \frac{\tau}{(1+\nu+\tau)} \end{pmatrix}$$

Hence

$$\begin{pmatrix} \mu \\ \sigma \\ \nu \\ \tau \end{pmatrix} = \begin{pmatrix} \alpha(1+\beta)^{-1} \\ (\alpha+\beta+1)^{-\frac{1}{2}} \\ \frac{p_0}{p_2} \\ \frac{p_1}{p_2} \end{pmatrix}$$

The default link functions relating the parameters (μ, σ, ν, τ) to the predictors $(\eta_1, \eta_2, \eta_3, \eta_4)$, which may depend on explanatory variables, are

$$\begin{bmatrix} \log(\frac{\mu}{1-\mu}) \\ \log(\frac{\sigma}{1-\sigma}) \\ \log \nu \\ \log \tau \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{bmatrix}$$

where $\log \nu = \log(p_0/(p_2)$ and $\log \tau = \log(p_1/(p_2)$

The mean and the variance of the BEINF Ospina and Ferrari (2010) distribution are defined by

$$\begin{aligned} E(y) &= \frac{\tau+\mu}{1+\nu+\tau} \\ V(y) &= \frac{\nu+\tau}{1+\nu+\tau}\left\{\frac{\tau}{(\nu+\tau)^2} + \frac{1}{(1+\nu+\tau)}\left(\frac{\tau}{\nu+\tau} - \mu\right)^2\right\} + \frac{\mu^2\sigma^2(1-\mu)}{1+\nu+\tau} \end{aligned}$$

The Beta inflated distribution BEINF (μ, σ, ν, τ) can be fitted explicitly in GAMLSS.

Model (2.12) is equivalent to a beta distribution $BE(\mu, \sigma)$ model for $0 < Y < 1$, together with a multinomial model with three levels, denoted $MULT3(\nu, \tau)$, for recoded variable Y_1 given by

$$Y_1 = \begin{cases} 0 & \text{if } Y = 0 \\ 1 & \text{if } Y = 1 \\ 2 & \text{if } 0 < Y < 1 \end{cases} \tag{2.13}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} p_0 & \text{if} \quad y_1 = 0 \\ p_1 & \text{if} \quad y_1 = 1 \\ (1 - p_0 - p_1) & \text{if} \quad y_1 = 2 \end{cases} \tag{2.14}$$

where $p_0 = \nu/(1+\nu+\tau)$ and $p_1 = \tau/(1+\nu+\tau)$.

The log likelihood function for the BEINF model (2.12) is equal to the sum of the log likelihood functions of the beta BE model and the multinomial MN3 model (2.14).

$$\begin{aligned} l(\theta) &= \log(L(\theta)) \\ &= n_0 \log(p_1) + n_1 \log(p_1) + n_2 \log(1 - p_0 - p_1) + n_2 \log[B(\alpha, \beta)] \\ &\quad + (\alpha - 1) \sum_{0<y_i<1} \log(y_i) + (\beta - 1) \sum_{0<y_i<1} \log(1 - y_i) \end{aligned}$$

Where n_0, n_1 and n_2 are the number of zero, one and non-zero-one values of Y, so $n_0 + n_1 + n_2 = n$.

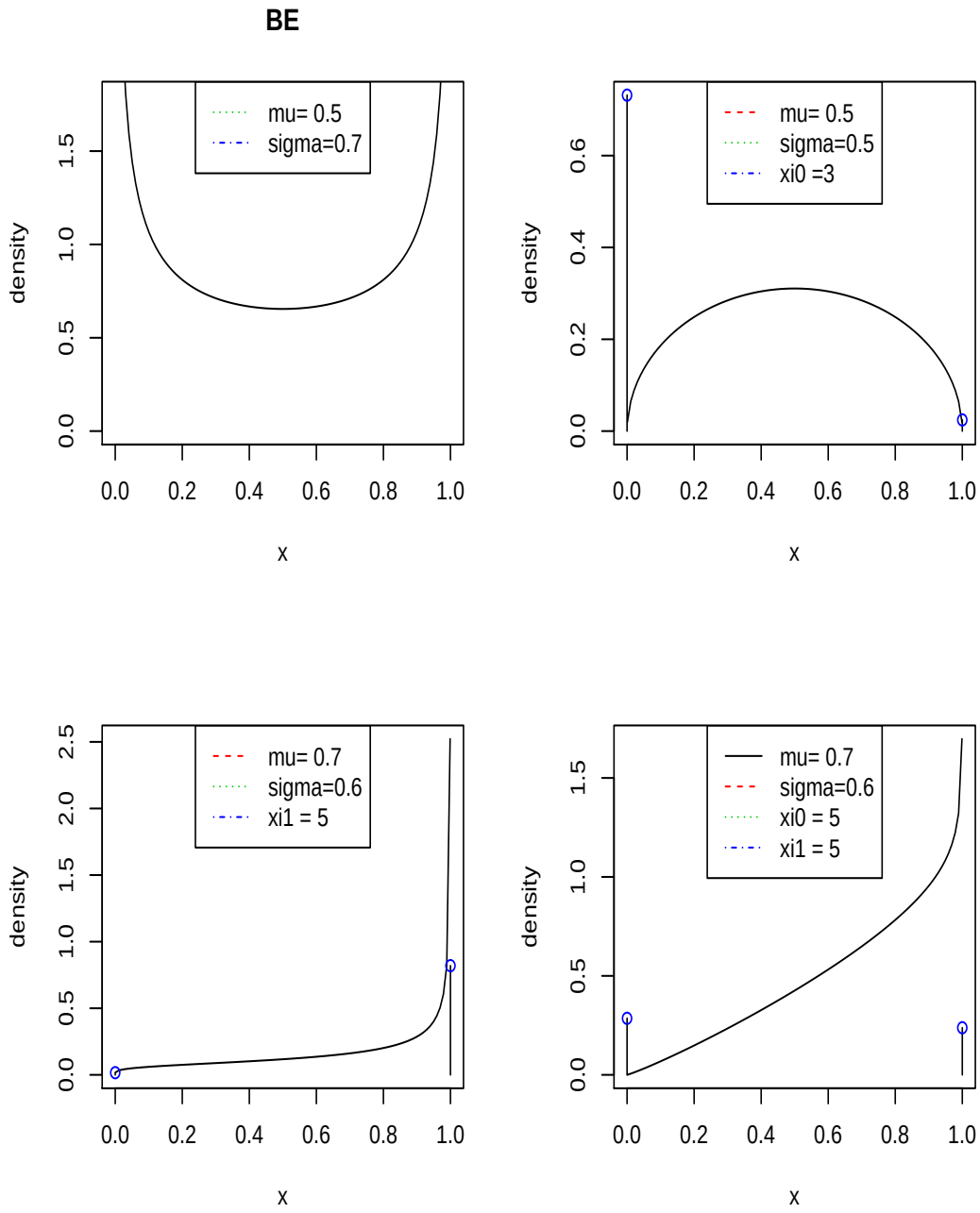


Fig. 2.1 Pdfs of BEINF0, BEINF1, BEINF and BE distribution.

Figure (2.1) presents BE, BEINF0, BEINF1 and BEINF distributions for different choices of μ, σ with fixed ν and τ. Note that for all μ and σ the BEINF0, BEINF1 and BEINF distributions are, in general asymmetrical because of the probability mass at 0 and/or 1. Densities of inflated beta distributions may be unimodal and uniantimodal and may have 'J', inverted 'J' and uniform shapes. In this graph a vertical bar with a circle above represents a probability mass at 0 and/or

1. BEINF0, BEINF1 and BEINF have same functional shape on the interval (0,1). However they differ at mass point, being at 0 for BEINF0, being at 1 for BEINF1 and being at 0 and 1 for BEINF.

Chapter 3

Distributions on (0,1)

3.1 Introduction

In this chapter we provide a review of some important explicit distributions on (0,1). The chapter also includes transformed logit and truncated distributions on (0,1).

3.1.1 Beta distribution

The beta distribution is a family of continuous distributions defined on the bounded support on the interval (0, 1). The beta distribution is a flexible two parameter distribution for a response variable taking values in the restricted range (0,1) not including 0 and 1. For example Trenkler (1996), Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004) have shown practical implementation of beta distribution in their work. The beta distribution was originally parameterised by two positive shape parameters and denoted here by $BE_o(a,b)$. The probability density function is given by

$$f_Y(y|a,b) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} \tag{3.1}$$

for $0 < y < 1$, $a > 0$ and $b > 0$ and B(a,b) is the beta function. Here the mean and variance functions are given by $E(Y) = \frac{a}{(a+b)}$ and $Var(Y) = ab(a+b)^{-2}(a+b+1)^{-1}$.

In the second parameterization of the beta distribution $BE(\mu, \sigma)$, the parameters μ and σ are location and scale parameters relate to the mean and standard deviation of a random variable Y. The beta distribution $BE(\mu, \sigma)$ with parameters μ and σ $(0 < \mu < 1, 0 < \sigma < 1)$ has probability density function (pdf), given by

$$f_Y(y|\mu, \sigma) = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1} \tag{3.2}$$

for $0 < y < 1$, where $a = \mu \left(\frac{1-\sigma^2}{\sigma^2}\right)$ and $b = (1-\mu)\left(\frac{1-\sigma^2}{\sigma^2}\right)$, $a > 0$ and $b > 0$. The relationship between two sets of parameters (μ, σ) and (a,b) is given by,

$$\begin{aligned} \mu &= \frac{a}{a+b} \\ \sigma &= (a+b+1)^{-\frac{1}{2}} \end{aligned}$$

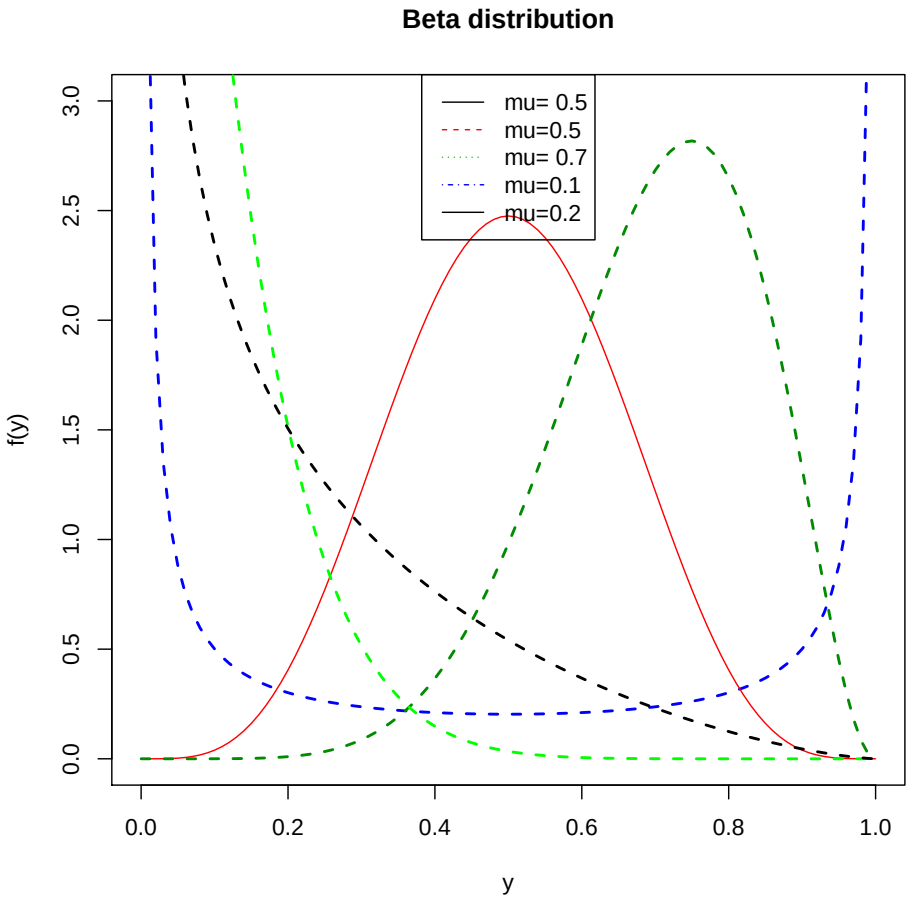


Fig. 3.1 Pdfs of beta distribution.

The mean and the variance function of Y are given by, $E(Y) = \mu$ and $V(Y) = \sigma^2\mu(1-\mu)$. In the `gamlss()` package in R this parameterization is denoted by $BE(\mu, \sigma)$.

The pdf of the beta distribution has different shapes: unimodal $(\mu > 1, \sigma > 1)$, uniantimodal $(\mu < 1, \sigma < 1)$, increasing $(\mu > 1, \sigma \leq 1)$, decreasing $(\mu \leq 1, \sigma > 1)$ or constant $(\mu = \sigma = 1)$ depending on the values of μ and σ relative to 1, see Figure 3.1.

3.1.2 Arcsine distribution

Arcsine distribution is a special case of beta distribution denoted by, $Y \sim BEo(\frac{1}{2}, \frac{1}{2})$. the probability density function of the standard arcsine distribution is given by

$$f_Y(y) = \frac{1}{\left\{\Gamma(\frac{1}{2})\right\}^2} y^{-\frac{1}{2}}(1-y)^{-\frac{1}{2}} \tag{3.3}$$

for $0 < y < 1$, since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, the probability density function is obtained by

$$f_Y(y) = \frac{1}{\pi\sqrt{y(1-y)}} \tag{3.4}$$

for $0 < y < 1$. The cumulative density function of the arcsine distribution is given by

$$F_Y(y) = \frac{2}{\pi} arcsine\sqrt{y} \tag{3.5}$$

for $0 \leq y \leq 1$, where arcsine function is the inverse of sine function, i.e. $arcsine(x) = sin^{-1}(x)$.

The distribution can be expanded to include any bounded support from $p \leq y \leq q$ by the simple transformation

$$F_Y(y) = \frac{2}{\pi} arcsine\left(\sqrt{\frac{y-p}{q-p}}\right) \tag{3.6}$$

for $p \leq y \leq q$ and therefore the probability density function of transformed random variable Y is given by

$$f_Y(y) = \frac{1}{\pi\left(\sqrt{\frac{y-p}{q-y}}\right)} \tag{3.7}$$

for $p \leq y \leq q$.

3.1.3 Kumarasawamy distribution

Jones (2009) introduced a beta like distribution which he called the Kumarasawamy distribution originally presented in the hydrological literature by Kumaraswamy (1980). The Kumarasawamy distribution has two positive shape parameters α and β and has many of the same properties that beta distribution has. The Kumarasawamy distribution is denoted by $Y \sim KU(\alpha, \beta)$ and the density function is given by

$$f_Y(y|\alpha,\beta) = \alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1} \tag{3.8}$$

for $0 < y < 1$, where $\alpha > 0$ and $\beta > 0$. Figure 3.2 shows plots of the Kumarasawamy distribution, $KU(\alpha, \beta)$, for different parameters α and β.

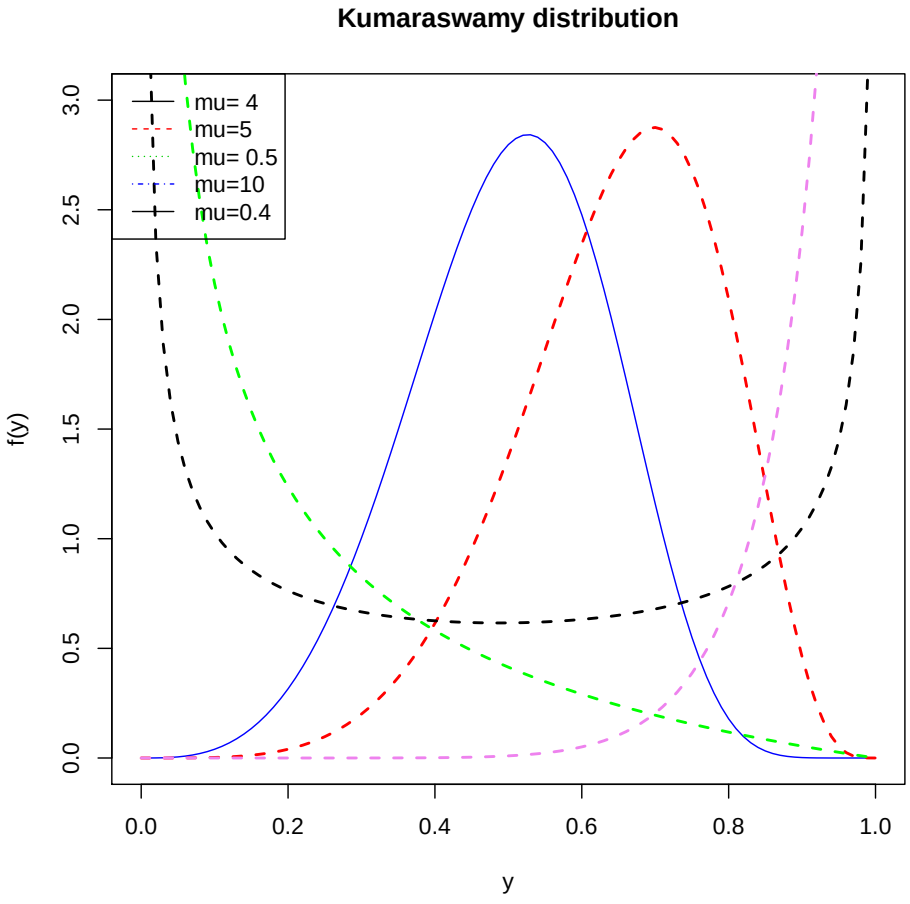


Fig. 3.2 Pdfs of Kumaraswamy distribution.

The cumulative distribution function of Kumarasawamy distribution for $0 < y < 1$ is given by

$$F_Y(y) = 1 - (1 - y^{\alpha})^{\beta} \tag{3.9}$$

3.1.4 Generalised beta distribution

McDonald and Xu (1995) introduced the five parameter generalised beta distribution of the first kind denoted by $Y \sim GB(a,b,c,p,q)$ the probability density function is defined by the probability density function

$$f_Y(y|a,b,c,p,q) = \frac{|a|y^{ap-1}\,(1-(1-c)(y/b)^a)^{q-1}}{b^{ap}B(p,q)\,(1+c(y/b)^a)^{p+q}} \tag{3.10}$$

for $0 < y^a < b^a/1-c$ and zero otherwise with $0 \leq c \leq 1$ and $b > 0$, $p > 0$ and $q > 0$.

3.1.4.1 Generalised beta type 1 (GB1)

The generalised beta type 1 distribution (GB1) in the `gamlss` package is a reparameterisation of the submodel with range $0 < y < 1$ of the five parameter generalised beta, GB (a,b,c,p,q), of McDonald and Xu (1995). GB1 has range $0 < y < 1$, whereas the range of the generalised beta of the first kind (McDonald and Xu, 1995) depends on the parameters. The generalised beta type 1 distribution is denoted by GB1(μ,σ,ν,τ). GB1 is defined by assuming that Z has a $BE(\mu,\sigma)$ distribution where

$$Z = \frac{Y^{\tau}}{\nu + (1-\nu)Y^{\tau}}$$

where $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$ and $\tau > 0$. Hence the probability density function of GB1, denoted by $Y \sim GB1(\mu,\sigma,\nu,\tau)$ is given by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{\tau\nu^{\beta}y^{\tau\alpha-1}(1-y^{\tau})^{\beta-1}}{B(\alpha,\beta)[\nu+(1+\nu)y^{\tau}]^{\alpha+\beta}} \tag{3.11}$$

for $0 < y \leq 1$, GB1 is a reparameterised submodel of generalised beta of the first kind (GB) of McDonald and Xu (1995) where

$$GB1(\mu, \sigma, \nu, \tau) = GB\left(\tau, \nu^{\frac{1}{\tau}}, (1-\nu), \mu(\sigma^{-2}-1), (1-\mu)(\sigma^{-2}-1)\right)$$

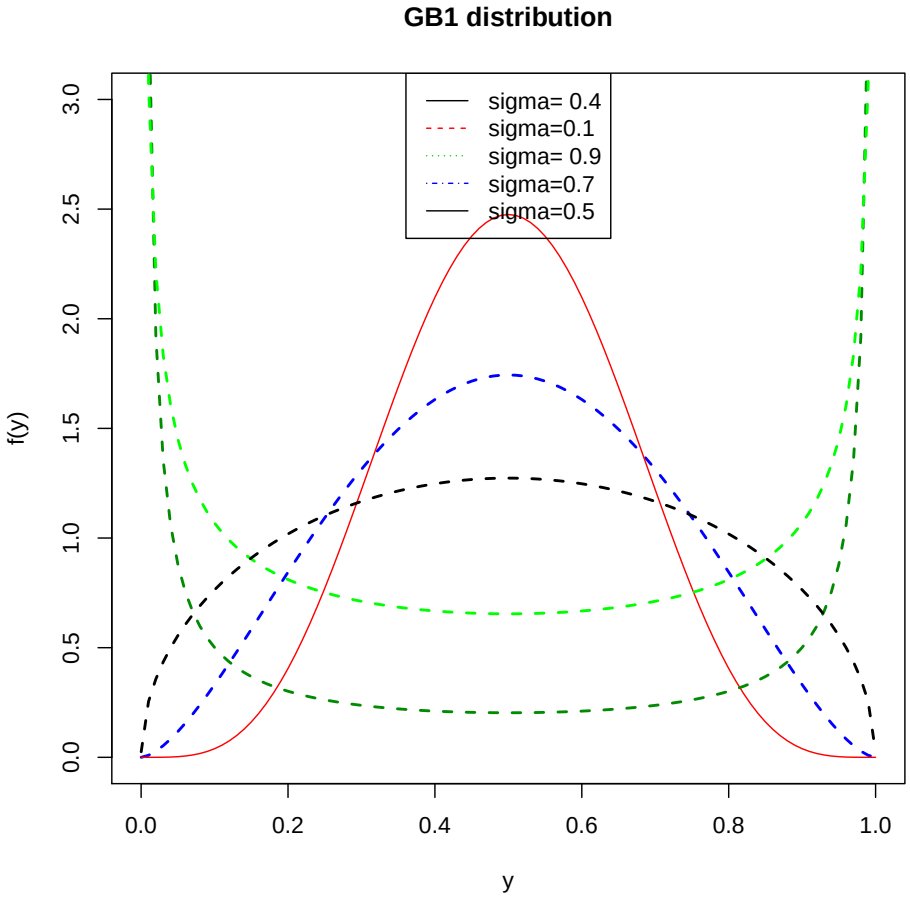


Fig. 3.3 Pdfs of GB1 distribution.

3.1.4.2 Generalised beta distribution type 3 (G3B)

The generalised beta distribution with three parameters, denoted by $Y \sim G3B(\alpha_1, \alpha_2, \gamma)$ was developed by Libby and Novick (1982). The probability density function of G3B is given by,

$$f_Y(y|\alpha_1, \alpha_2, \lambda) = \frac{\lambda^{\alpha} y^{\alpha-1}(1-y)^{\alpha_2-1}}{B(\alpha_1, \alpha_2)[1-(1-\lambda)y]^{\alpha_1+\alpha_2}} \tag{3.12}$$

. The generalised three parameter beta is a reparameterised submodel of GB1 given by

$$G3B(\alpha_1,\alpha_2,\lambda) = GB1\left(\alpha_1(\alpha_1+\alpha_2)^{-1},(\alpha_1+\alpha_2-1)^{\frac{-1}{2}},\lambda^{-1},1\right)$$

3.1.5 Triangular Distribution

Johnson (1997) showed the use of the triangular distribution as a proxy of the beta distribution. Like the beta distribution, the triangular distribution is a continuous distribution on the range [a,b]. In its most general form the pdf of the triangular distribution is given by

$$f_Y(y) = \begin{cases} \frac{2(y-a)}{(b-a)(c-a)} & \text{for } a \le y \le c \\ (\frac{2(b-y)}{(b-a)(b-c)} & \text{for } c < y \le b \end{cases} \tag{3.13}$$

. The cumulative distribution function of triangular distribution also obtained by

$$F_y(y) = \begin{cases} \frac{(y-a)^2}{(b-a)(c-a)} & \text{for } a \le y \le c \\ 1-\frac{(b-y)^2}{(b-a)(b-c)} & \text{for } c \le y \le b \end{cases} \tag{3.14}$$

where c is the mode.

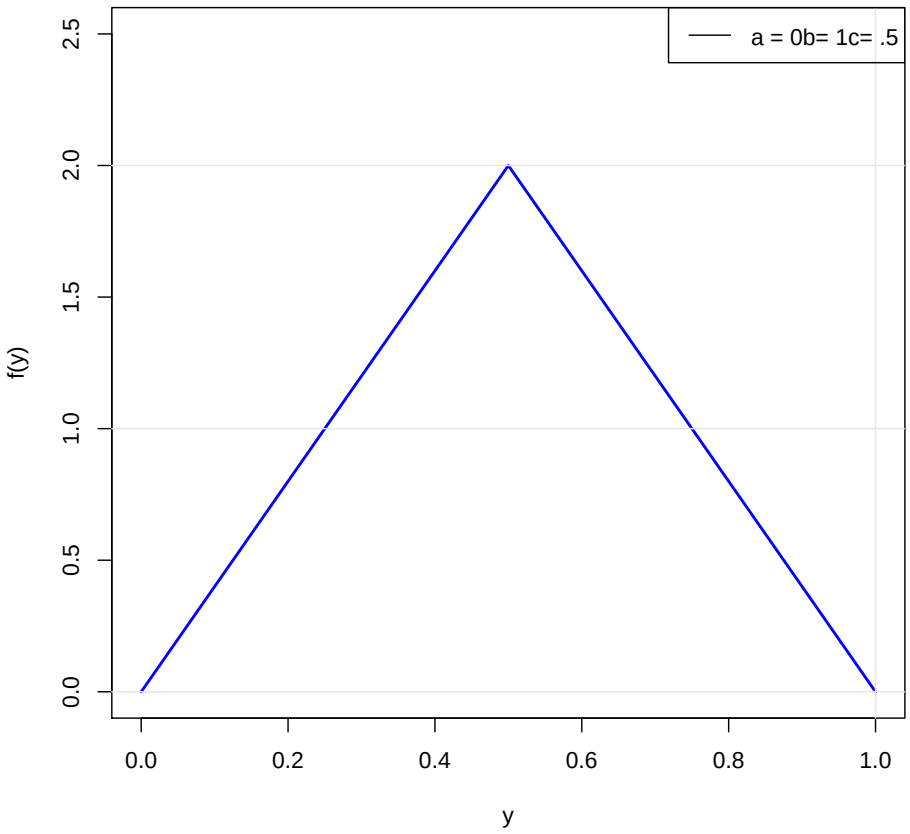


Fig. 3.4 Pdfs of triangular distribution.

Setting $a = 0$ and $b = 1$ in the triangular distribution (3.13) gives a distribution with range $0 \leq y \leq 1$, with probability density function

$$f_Y(y) = \begin{cases} \frac{2y}{c_o} & \text{if } 0 \leq y \leq c_o \\ \frac{2(1-y)}{(1-c_o)} & \text{if } c_o \leq y \leq 1 \end{cases} \tag{3.15}$$

for $0 \leq y \leq 1$, where $0 \leq c_o \leq 1$. The mode of the transformed distribution is $c_o = \frac{(c-a)}{(b-a)}$. If $c_o = \frac{1}{2}$, the distribution of Y is called the symmetric triangular distribution. Like the beta distribution, the triangular distribution can be symmetrical and positively or negatively skewed but can not be other than unimodal, see Figure 3.4.

It is worth mentioning some other distributions which can be used to address the bounded proportion data (e.g. Topp and Leone (1955), Vicari et al. (2008))

3.1.6 Simplex distribution

Song et al. (2004) use the simplex distribution of Barndorff-Nielsen and Jørgensen (1991) to model the marginal means of a longitudinal proportion response variable.

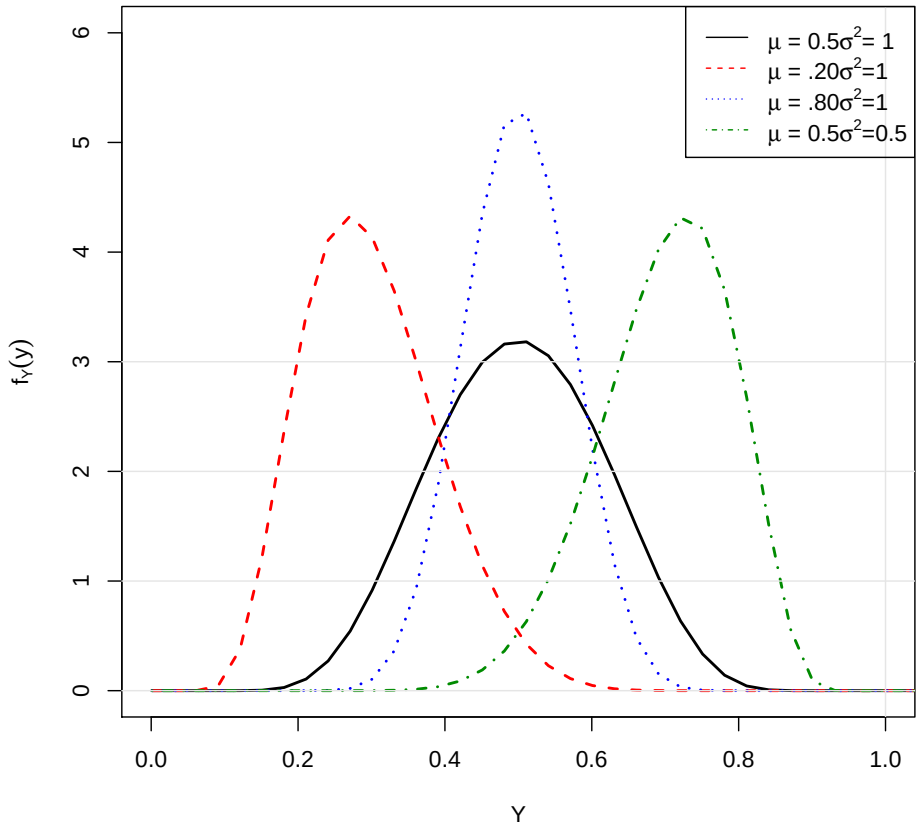


Fig. 3.5 Pdfs of simplex distribution.

The probability density function of the simplex distribution is given by

$$f_Y(y|\mu,\sigma^2) = \left[2\pi\sigma^2\{y(1-y)\}^3\right]^{-\frac{1}{2}} exp\left\{-\frac{1}{2\sigma^2}d(y;\mu)\right\} \tag{3.16}$$

for $0<y<1$, where, $d(y;\mu)=\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu^2)}$, with location parameter $0<\mu<1$ and dispersion parameter $\sigma^2>0$. The random variable Y follows a simplex distribution denoted by $Y\sim SIMPLEX(\mu,\sigma^2)$.

3.1.7 Logit distributions

Any distribution on range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ by using an inverse logit transformation $Y = 1/(1+e^{-Z})$. The distribution of Y is called a logit distribution. If Z has a four parameter distribution denoted D in general, i.e. $Z \sim D(\mu, \sigma, \nu, \tau)$, then the distribution of Y is called a logit D distribution denoted $Y \sim logitD(\mu, \sigma, \nu, \tau)$. If Z has the pdf $f_Z(z)$ and Y has pdf $f_Y(y)$, then

$$f_Y(y) = f_Z(z)\left|\frac{dz}{dy}\right| = \frac{1}{y(1-y)} f_Z(z) \tag{3.17}$$

for $0 < y < 1$, where $z = \log[y/(1-y)]$.

For example if Z has a skew exponential power distribution $Z \sim SEP(\mu, \sigma, \nu, \tau)$ on $(-\infty, \infty)$, Fernandez et al. (1995), then Y has a *logitSEP* distribution, $Y \sim logitSEP(\mu, \sigma, \nu, \tau)$ on $(0,1)$. The *logitSEP* distribution is created using the `gamlss` function `gen.Family`, which allows any `gamlss` distribution with range $(-\infty, \infty)$, (e.g. *SEP*), to be transformed to a new `gamlss` distribution, (e.g. *logitSEP*), with range $(0,1)$. Alternatives to the skew exponential power (*SEP*) distribution include the skew student t (*SST*) distribution, see Wurtz et al. (2006), reparameterized from Fernández and Steel (1998*a*) and the sinh-arcsinh (SHASHo) distribution, see Jones and Pewsey (2009).

3.1.7.1 Logit normal distribution

The logit normal distribution emerges by assuming Z has a normal $NO(\mu, \sigma)$ distribution and $Y = \frac{1}{(1+e^{-Z})}$. Then Y has a logit normal distribution denoted, $Y \sim logitNO(\mu, \sigma)$ on $(0,1)$. The probability density function of a logit normal distribution is given by

$$f_Y(y) = \frac{1}{y(1-y)} f_Z(z) \tag{3.18}$$

for $0 < y < 1$, where $Z \sim N(0,1)$ and $z = \log(\frac{y}{1-y})$, and $f_Z(z) = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}z^2\right]$ for $-\infty < z < \infty$. Therefore the pdf of the logit normal distribution logitNO(μ, σ), for Y is given by

$$f_Y(y) = \frac{1}{y(1-y)} \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}\left\{\log\left(\frac{y}{1-y}\right)\right\}^2\right] \tag{3.19}$$

.

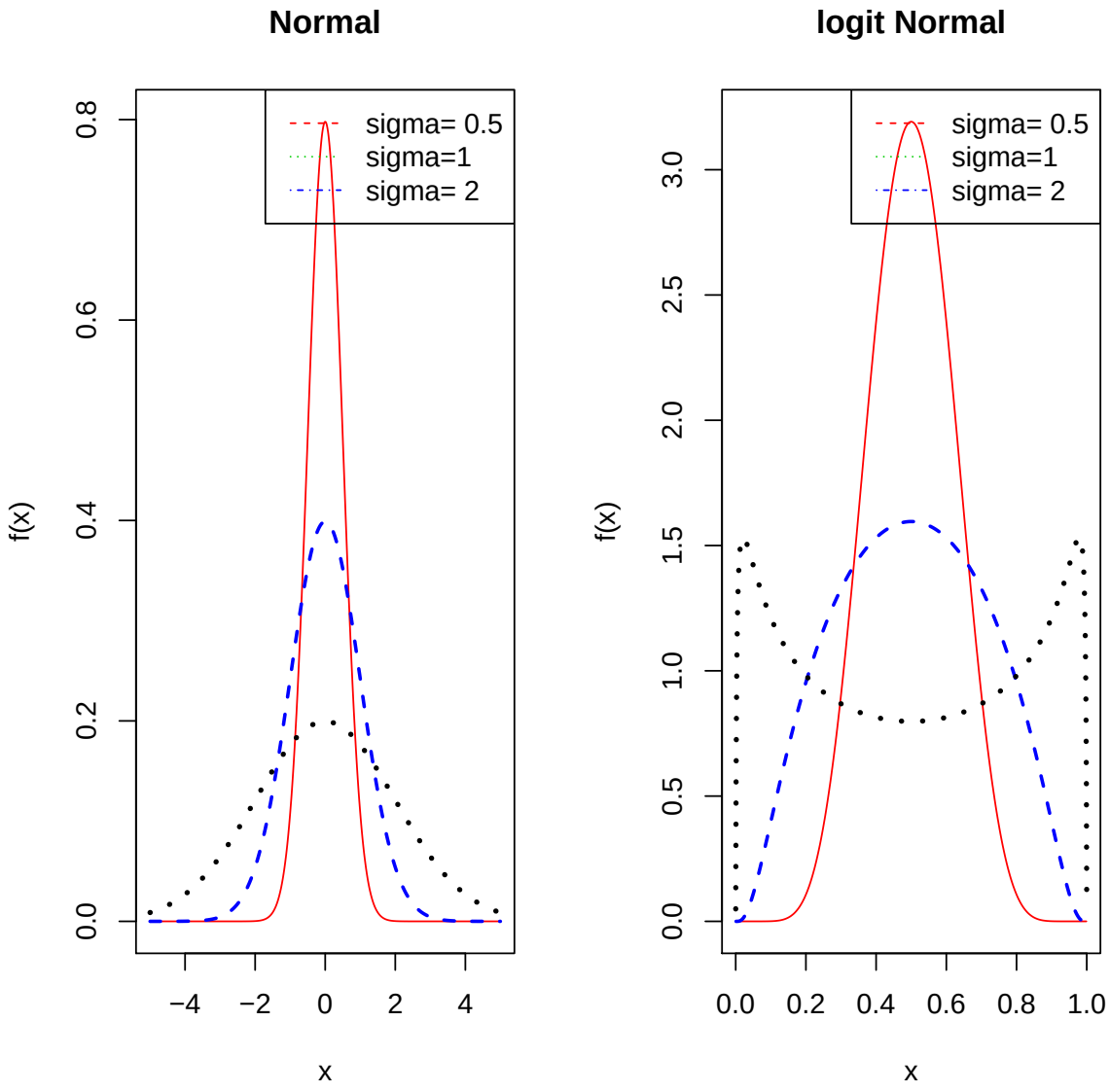


Fig. 3.6 Pdfs of NO and logit-normal distribution.

Figure 3.6 shows the logit normal distribution can take various shapes depending on the parameters μ and σ. Note that changing μ to $-\mu$ reflects the distribution about a vertical axis at $y = 0.5$. Aitchison and Begg (1976) indicate that the logit normal distribution is richer than and can approximate any beta density.

3.1.7.2 Logit skew t type 3 distribution

The skew t type 3 distribution is a spliced-scale distribution denoted by $Z \sim ST3(\mu, \sigma, \nu, \tau)$ for $-\infty < Z < \infty$ Fernández and Steel (1998*a*). The probability density function of distribution $ST3(\mu, \sigma, \nu, \tau)$ for Z is given by,

$$f_Z(z|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma}\left\{1 + \frac{(z-\mu)^2}{\sigma^2\tau}\left[\nu^2 I(z < \mu) + \frac{1}{\nu^2} I(z \geq \mu)\right]\right\} \tag{3.20}$$

for $-\infty < z < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, $\tau > 0$ and

$$c = 2\nu / \left[\sigma(1+\nu^2) B(\tfrac{1}{2}, \tfrac{\tau}{2}) \tau^{\frac{1}{2}}\right]$$

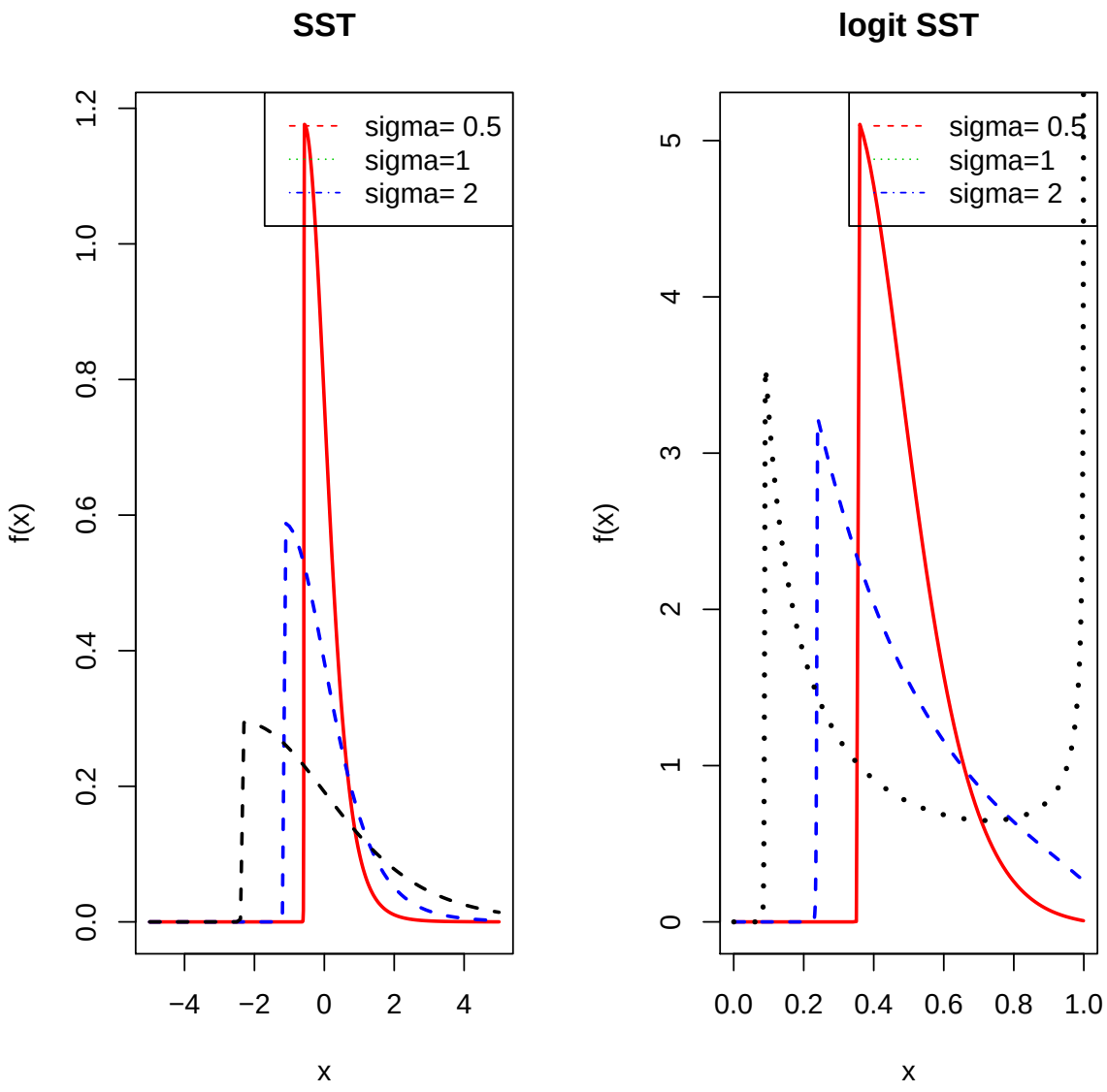


Fig. 3.7 Pdfs of SST and logitSST distribution.

The logit skew t type 3 distribution is an inverse logit transformation $Y = 1/(1+e^{-Z})$ of Z, where Z is assumed to have a ST3 distribution and is denoted by $Y \sim logitST3(\mu,\sigma,\nu,\tau)$. Hence the probability density function of distribution logitST3(μ,σ,ν,τ) for Y is given by,

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{1}{y(1-y)} f_Z(z|\mu,\sigma,\nu,\tau) \tag{3.21}$$

for $0 < y < 1$, where $z = \log[y/(1-y)]$ and $f_Z(z|\mu,\sigma,\nu,\tau)$ is given by equation (3.20).

Figure 3.7 shows on the left the pdf of the skew student t (SST) distribution on the real line and on the right the logitSST distribution on (0, 1). In the pdf of the logit skew student t distribution, it can be shown that the density at 0 is infinity, similarly density at 1 is infinity.

3.1.8 Truncated distributions

A truncated distribution can be obtained by restricting the range of the random variable Y. Truncation of a distribution can be below, above or both. If the truncation is from below, the mean of the truncated distribution is greater than the original distribution. In the case of the above truncation the mean of the truncated distribution is less than the original distribution .

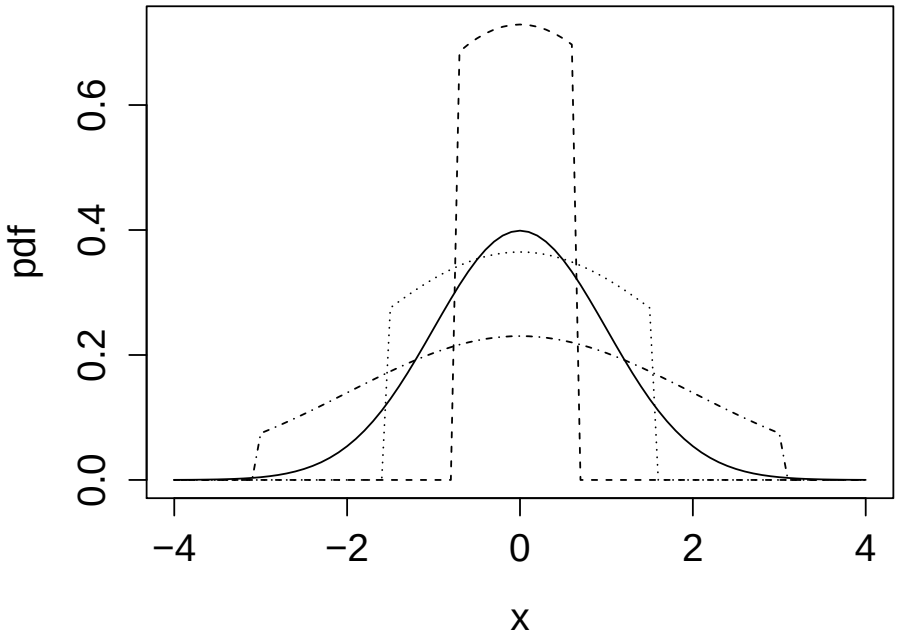


Fig. 3.8 Truncated pdfs of standard normal distribution.

3.1.8.1 Below truncation

If a continuous random variable Y has a density function $f_Y(y)$. Let T_l and T_r be constants in the range of Y, so that $T_r > T_l$. For left truncation $(i.e. Y \geq T_l)$, let Y_l be the resulting left truncated variable having pdf given by

$$f_{Y_l}(y) = \frac{f_Y(y)}{1 - F_Y(T_l)} \tag{3.22}$$

The cdf for the left truncated distribution is given by

$$F_{Y_l}(y) = \frac{F_Y(y) - F_Y(T_l)}{1 - F_Y(T_l)} \tag{3.23}$$

where $f_Y(y)$ and $F_Y(y)$ are the pdf and cdf of the original variable Y.

3.1.8.2 Above truncation

For above truncation $(i.e. Y < T_r)$, let Y_r be the resulting right truncated variable having probability density function given by

$$f_{Y_r}(y) = \frac{f_Y(y)}{F_Y(T_r)} \tag{3.24}$$

The cumulative distribution function for the right truncated distribution is given by

$$F_{Y_r}(y) = \frac{F_Y(y)}{F_Y(T_r)} \tag{3.25}$$

where $f_Y(y)$ and $F_Y(y)$ are the pdf and cdf of the original distribution.

3.1.8.3 Both truncation

For both truncation $(i.e. T_l \leq Y \leq T_r)$ let Y_{lr} be the resulting two sided truncated variable having the probability density function is given by

$$f_{Y_{lr}}(y) = \frac{f_Y(y)}{F_Y(T_r) - F_Y(T_l)} \tag{3.26}$$

with cumulative distribution function of the truncated distribution given by

$$F_{Y_{lr}}(y) = \frac{F_Y(y) - F_Y(T_l)}{F_Y(T_r) - F_Y(T_l)} \tag{3.27}$$

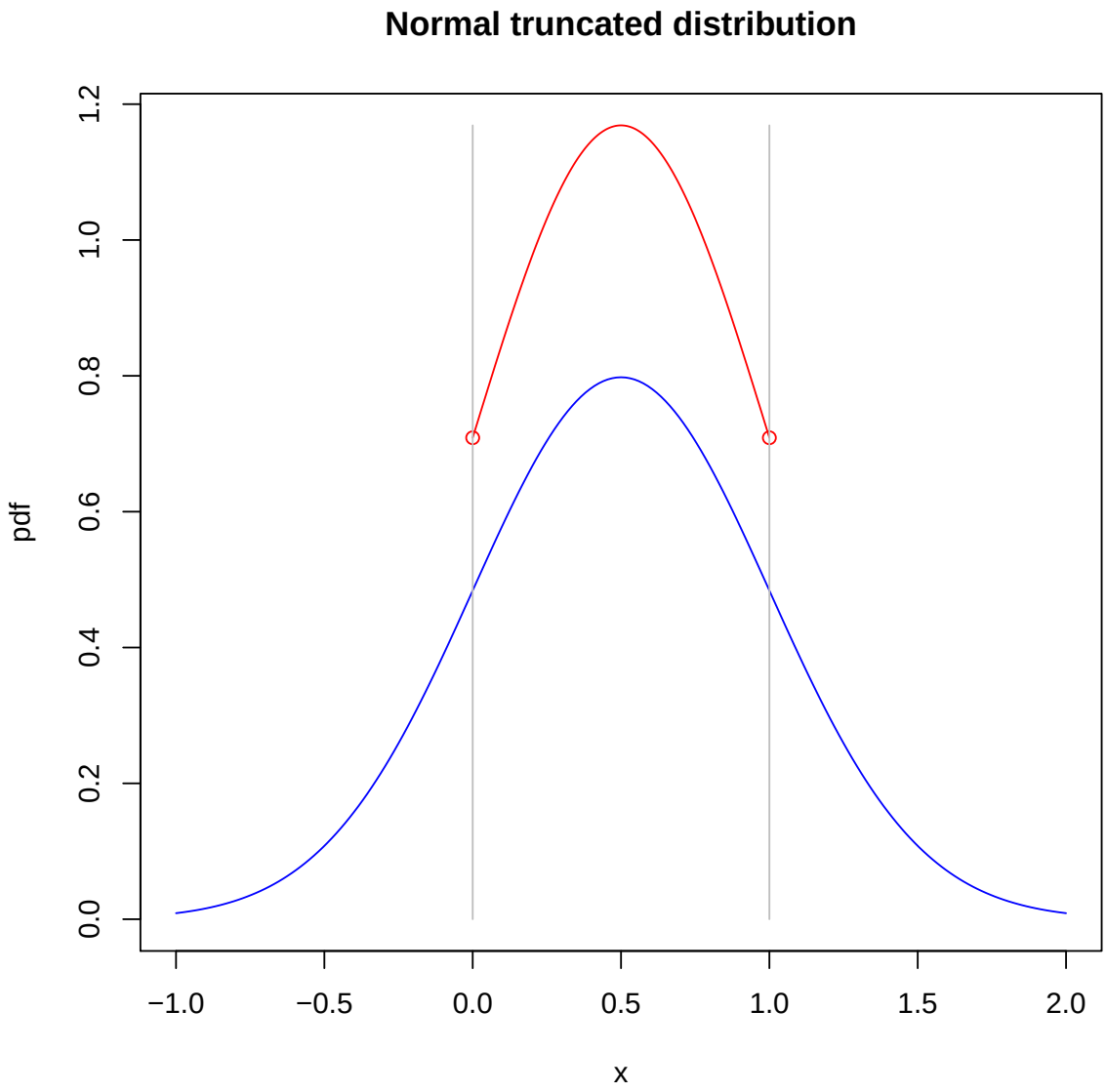


Fig. 3.9 Truncated normal distribution below 0 and above 1

For example any distribution for a variable Z on range $(-\infty,\infty)$ can provide a variable Y on a truncated range $(0,1)$ by truncating Z below 0 and above 1. The resulting probability density function for Y is given by

$$f_Y(y) = \frac{f_Z(y)}{F_Z(1) - F_Z(0)}$$

for $0 < y < 1$, where f_Z and F_Z are the probability density function and cumulative distribution function of Z respectively. If Z has a four parameter distribution denoted D in general, i.e. $Z \sim D(\mu,\sigma,\nu,\tau)$, then Y has a truncated D distribution, denoted $Y \sim Dtr(\mu,\sigma,\nu,\tau)$. For

example if $Z \sim SEP(\mu, \sigma, \nu, \tau)$ then $Y \sim SEPtr(\mu, \sigma, \nu, \tau)$. The *SEPtr* distribution is created using the GAMLSS function gen.trun() from the `R` package `gamlss.trun`, which allows any `gamlss` distribution (e.g. *SEP*) to be converted into a truncated distribution.

Figure 3.9 shows the normal distribution and the truncated normal distribution on 0 to 1. Unlike the beta or logit distributions, in the case of a truncated distribution the density of y at 0 and 1 is finite, neither 0 nor infinity.

Chapter 4

Bimodal skew symmetric normal (BSSN) distribution

4.1 Introduction

In this chapter a parsimonious bimodal distribution for a response variable with a range $(-\infty, \infty)$ will be described. The distribution is referred to as the bimodal skew symmetric normal distribution (BSSN) [1]. The BSSN distribution is used in research for its effectiveness in capturing bimodality, skewness and excess kurtosis. The mean, variance, skewness, excess kurtosis and likelihood estimation of the distribution are taken from Hassan and El-Bassiouni (2016). This chapter outlines the inclusion of BSSN distribution in the `gamlss.dist` package in `R` which allows fitting of the distribution (with any or all parameters modelled using explanatory variables). This work is original to this thesis.

The objective then is to apply the gamlss function `gen.Family()` (as described in section 3.1.7) in order to transform the BSSN distribution on $(-\infty, \infty)$ to the logitBSSN distribution on (0,1), creating a distribution on (0,1) which can be bimodal.

[1] Bimodal skew symmetric normal distribution by Hassan and El-Bassiouni (2016)

4.2 Bi-modal skew symmetric normal distribution and its logit transformation

This section, following Hassan and El-Bassiouni (2016), describes the bi-modal skew symmetric normal distribution (BSSN). The parameterizations used in the R code in Appendix E is defined as follows [2]. Let Φ and ϕ are the cdf and pdf of normal random variable with mean (μ) and standard deviation ($\frac{1}{\sqrt{(2\sigma)}}$). Let $\alpha(y)$ be a linear function of y. The cumulative distribution function of the bi-modal skew symmetric normal distribution is obtained by

$$F_Y(y|\mu,\sigma,\nu,\tau) = \Phi(y) - \alpha(y)\cdot\phi(y) \tag{4.1}$$

where

$$\alpha(y) = (y+\mu-2\nu)/(1+2\sigma[\tau+(\nu-\mu)^2])$$

Let Y be a random variable having a BSSN (μ,σ,ν,τ) distribution i.e.

$$Y \sim BSSN(\mu,\sigma,\nu,\tau)$$

Then the probability density function of the BSSN distribution is given by

$$f_Y(y|\mu,\sigma,\nu,\tau) = c[\tau+(y-\nu)^2]e^{-\sigma(y-\mu)^2} \tag{4.2}$$

for $-\infty < y < \infty$, where $c = 2\sigma^{\frac{3}{2}}/\gamma\sqrt{\pi}$, $\gamma = 1+2\sigma\theta$, $\theta = \tau+\delta^2$, $\delta = \nu-\mu$. $-\infty < \mu < \infty$ and $-\infty < \nu < \infty$ are location parameters and $\sigma > 0$ and $\tau \geq 0$ denote the scale and bi-modality parameters respectively.

[2] The distribution parameters (μ, σ, ν and τ) are equivalent to (μ, ψ, β and δ) respectively in Hassan and El-Bassiouni (2016)

The following results are obtained form Hassan and El-Bassiouni (2016). The mean and variance of the bi-modal skew symmetric normal distribution, $BSSN(\mu,\sigma,\nu,\tau)$, for Y are given by

$$\begin{aligned} E(Y) &= \mu - 2\frac{\delta}{\gamma} \\ Var(Y) &= \frac{(3+8\sigma\tau+4\sigma^2\theta^2)}{2\sigma\gamma^2} \end{aligned}$$

Skewness (ζ) of $Y \sim BSSN(\mu,\sigma,\nu,\tau)$ is measure by

$$\zeta = \frac{4\sqrt{2\sigma}[3-2\sigma(\delta^2-3\tau)]\delta}{(3+8\sigma\tau+4\sigma^2\theta^2)^{\frac{3}{2}}}$$

The BSSN distribution can be symmetric ($\zeta = 0$), skewed to the right ($\zeta > 0$) and skewed to the left ($\zeta < 0$) and the excess kurtosis ($\aleph$) is measured by

$$\aleph = \frac{3(5+64\tau\theta^2\sigma^3+16\theta^4\sigma^4+16\sigma(3\delta^2+2\tau)+8\sigma^2[\delta^4+9\tau(2\delta^2+\tau)]}{(3+8\sigma\tau+4\sigma^2\theta^2)^2} - 3$$

The BSSN distribution can be mesokurtic ($\aleph = 0$), leptokurtic ($\aleph > 0$) and platykurtic ($\aleph < 0$).

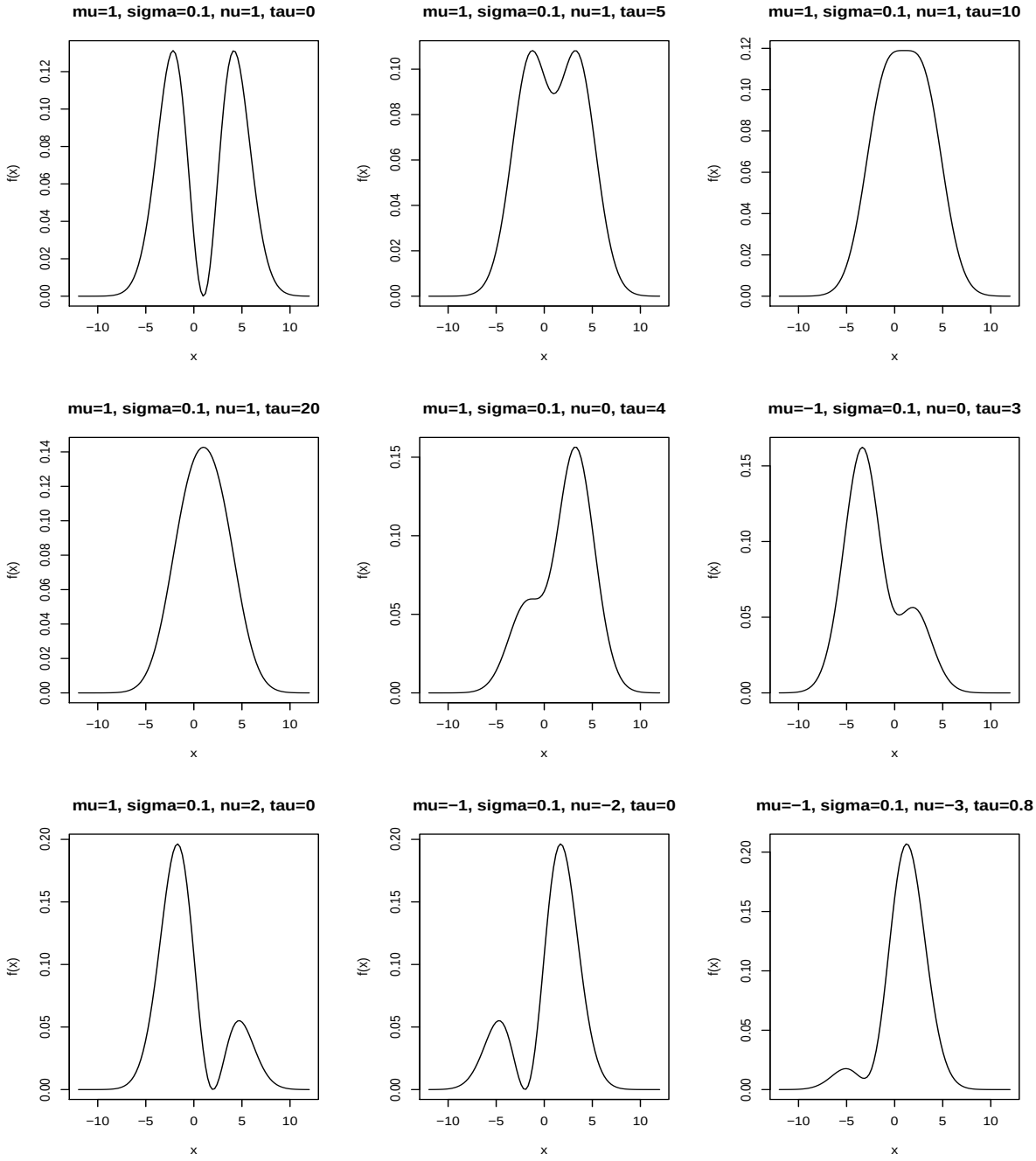


Fig. 4.1 Shapes of the pdfs of BSSN

4.2.1 Maximum likelihood estimation of BSSN

The loglikelihood function of the BSSN is given by

$$\log(l) = \log(c) + \log[\tau + (y-\nu)^2] - \sigma(y-\mu)^2$$

Following Hassan and El-Bassiouni (2016), the derivatives of the loglikelihood function of distribution BSSN with respect to μ, σ, ν and τ are given by

$$\begin{aligned}
\frac{\partial \log(l)}{\partial \mu} &= \frac{4\sigma\delta}{\gamma} + 2\sigma(y-\mu) \\
\frac{\partial \log(l)}{\partial \sigma} &= \frac{3}{2\sigma} - \frac{2\rho}{\gamma} - (y-\mu)^2 \\
\frac{\partial \log(l)}{\partial \nu} &= \frac{-4\sigma\delta}{\gamma} - 2\frac{y-\nu}{[\tau+(y-\nu)^2]} \\
\frac{\partial \log(l)}{\partial \tau} &= \frac{-2\sigma}{\gamma} + \frac{1}{[\tau+(y-\nu)^2]}
\end{aligned}$$

4.3 R implementation of BSSN

The R code for implementing the BSSN distribution in the gamlss.dist package in R is given in Appendix E and is original to this thesis.

4.3.0.1 Functions used in R implementation of BSSN

The following function given in Appendix E, were written to implement BSSN in gamlss.dist.

```
BSSN(mu.link = "identity", sigma.link = "log",
     nu.link = "identity", tau.link = "log")
```

```
dBSSN(x, mu=1, sigma=0.1, nu=1, tau=0, log = FALSE)
```

```
pBSSN(q, mu=1, sigma=0.1, nu=1, tau=0, lower.tail = TRUE,
      log.p = FALSE)
```

```
qBSSN(p, mu=1, sigma=0.1, nu=1, tau=0, lower.tail = TRUE,
      log.p = FALSE)
```

```
rBSSN(n, mu=1, sigma=0.1, nu=1, tau=0)
```

4.3.1 Arguments

mu.link Defines the mu.link, with identity link as the default for the mu parameter.

sigma.link Defines the sigma.link, with log link as the default for the sigma parameter.

nu.link Defines the nu.link, with identity link as the default for the nu parameter.

tau.link Defines the nu.link, with log link as the default for the nu parameter.

x, q vector of quantiles.

mu vector of mu parameter values.

sigma vector of scale parameter values.

nu vector of nu parameter values.

tau vector of tau parameter values.

log, log.p logical; if TRUE, probabilities p are given as log(p).

lower.tail logical; if TRUE (default), probabilities are $P[X \leq x]$,otherwise, $P[X > x]$.

p vector of probabilities.

n number of observations. If $length(n) > 1$, the length is taken to be the number required.

... for extra arguments.

4.3.2 Functions

Five functions of the distribution are :

dBSSN The pdf of the distribution, **d** function.

pBSSN The cdf of the distribution, **p** function.

qBSSN The inverse cdf (or quantile) of the distribution, **q** function.

rBSSN The random generating function of the distribution, **r** function.

BSSN The function for fitting the distribution.

4.3.3 Use of BSSN function

Figure 4.1 shows the various plots of the pdf of the BSSN distribution and demonstrates its ability to accommodate various shapes in terms of skewness, kurtosis and bimodality. The probability density plots of the BSSN distribution were achieved for the different parameter values as below.

```
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=0),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=5),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=10),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=20),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=0, tau=4),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=-1, sigma=0.1, nu=0, tau=3),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=2, tau=0),
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=-1, sigma=0.1, nu=-2, tau=0),
```

```
      -12, 12, ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=-1, sigma=0.1, nu=-3, tau=0.8),
      -12, 12, ylab="f(x)", main="BSSN").
```

Chapter 5

General inflated GAMLSS model on the unit interval

5.1 Introduction

This chapter introduces a mixed continuous-discrete distribution to model a proportion response variable observed on [0,1), (0,1] or [0,1], where a closed bracket indicates the inclusion of the end point in the interval range. The response variable distribution is modelled by combining any distribution on (0,1) together with point probabilities at 0, 1 or both, depending upon the range of the proportion response variable. Distributions on (0,1) were described in chapter 2 not only allowing the mean and variance but also skewness and fat tail features of the proportion response variable on (0,1) to be modelled.

5.2 General distribution on (0,1) inflated at 0 and/or 1

A distribution on the unit interval $(0, 1)$ inflated at 0 and 1 captures probability masses at 0 and 1. Distributions inflated at 0, or 1, or 0 and 1, are appropriate when the response variable Y takes values from 0 to 1 including 0, i.e. range [0,1) or including 1, i.e. range (0,1], or including

both 0 and 1, i.e. range [0,1], respectively. Hoff (2007) uses the unit inflated beta to model the response variable on the interval (0,1]. Ospina and Ferrari (2010) introduce the beta inflated distribution model for the response variable on the intervals [0,1), (0,1] and [0,1].

A general inflated model for a proportion response variable on [0,1] is a mixture of three components: a discrete value 0 with probability p_0, a discrete value 1 with probability p_1, and a continuous distribution on the unit interval $(0,1)$ with probability $(1-p_0-p_1)$. The mixed continuous-discrete probability (density) function (pdf) of Y is $f_Y(y|\theta,\xi_0,\xi_1)$ given by

$$f_Y(y|\theta,\xi_0,\xi_1) = \begin{cases} p_0 & \text{if } y = 0 \\ (1-p_0-p_1) f_W(y|\theta) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{5.1}$$

for $0 \leq y \leq 1$, where $f_W(y|\theta)$ is any probability density function defined on $(0,1)$, i.e. for $0 < y < 1$, with parameters $\theta^T = (\theta_1, \theta_2, \cdot\cdot, \theta_p)$ and $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$ and where $\xi_0 = \frac{p_0}{p_2}$, $\xi_1 = \frac{p_1}{p_2}$, where $p_2 = 1 - p_0 - p_1$, so $\xi_0 > 0$ and $\xi_1 > 0$. Hence

$$\begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} \frac{\xi_0}{(1+\xi_0+\xi_1)} \\ \frac{\xi_1}{(1+\xi_0+\xi_1)} \end{pmatrix}$$

Model (5.1) is equivalent to a distribution on (0,1) for $0 < Y < 1$, together with a multinomial model with three levels, $MULT3(\xi_0,\xi_1)$ for recoded variable Y_1 given by

$$Y_1 = \begin{cases} 0 & \text{if } Y = 0 \\ 1 & \text{if } Y = 1 \\ 2 & \text{if } 0 < Y < 1 \end{cases} \tag{5.2}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} p_0 & \text{if } y_1 = 0 \\ p_1 & \text{if } y_1 = 1 \\ (1 - p_0 - p_1) & \text{if } y_1 = 2 \end{cases} \tag{5.3}$$

The cumulative distribution function of Y (with pdf given by (5.1)) is given by

$$F_Y(y|\theta, \xi_0, \xi_1) = \begin{cases} p_0, & \text{if y=0} \\ p_0 + (1 - p_0 - p_1)F_W(y|\theta), & \text{if 0<y<1} \\ 1, & \text{if y=1} \end{cases} \tag{5.4}$$

for $0 \leq y \leq 1$.

Inflated distributions have the advantage of extra flexibility, in that the probilities of Y at 0 and 1 are modelled independently of the distribution on (0,1), but with the cost of introducing extra parameters (ξ_0, ξ_1) into the model.

Note p_0 and p_1 in equation (5.1) represent probability masses at 0 and 1 respectively. If the model (5.1) comprises only two components: a discrete value 0 and a continuous component on (0,1), the model is called a zero-inflated model and the probability (density) function of Y is $f_Y(y|\theta, p_0)$ given by

$$f_Y(y|\theta, p_0) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0) f_W(y|\theta) & \text{if } 0 < y < 1 \end{cases} \tag{5.5}$$

for $0 \leq y < 1$. If the model comprises only a discrete value 1 and a continuous component on (0,1), the model is called a one-inflated model and the probability density function is $f_Y(y|\theta)$ given by

$$f_Y(y|\theta, p_1) = \begin{cases} (1 - p_1) f_W(y|\theta) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{5.6}$$

5.3 Model Definition

Let $Y^T = (Y_1, Y_1, \cdots, Y_n)$ be a vector of independent response observations, where Y_i has probability (density) function

$$f_{Y_i}(y_i|\psi^i) = f_Y(y_i|\theta^i, \xi_{i1}, \xi_{i2})$$

given by (5.1), where $\theta^i = (\theta_{i1}, \theta_{i2}, \cdot\cdot, \theta_{ip})^T$, for $i = 1, 2, \cdot, \cdot, n$ and $\psi^i = (\theta^i, \xi_{i1}, \xi_{i2})$ is the full parameter vector for observation i.

Let $g_k(\cdot)$ be a known monotonic link function (for $k = 1, 2, 3, \cdots, K$) relating parameter ψ_k to explanatory variables through an additive model given by

$$g_k(\psi_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} s_{jk}(x_{jk}) \tag{5.7}$$

where ψ_k and η_k are vectors of length n, e.g.

$$\psi_k^T = (\psi_{1k}, \psi_{2k}, \cdots, \psi_{nk})$$
$$\eta_k^T = (\eta_{1k}, \eta_{2k}, \cdots, \eta_{nk})$$

β_k is a parameter vector of length $J_k{}'$,

$$\beta_k^T = (\beta_{1k}, \beta_{2k}, \cdots, \beta_{J_k{}'k})$$

X_k is a known design matrix of order $n \times J_k{}'$, x_{jk} is a vector of length n and the function s_{jk} is a non-parametric additive function of the explanatory variable x_{jk}, for $j = 1, 2, \cdots, J_k$ and $k = 1, 2, \cdots, p$. The amount of smoothing allowed in any of the smoothing function s_{jk} is determined by a set of smoothing parameters λ_{jk} for $j = 1, 2, \cdots, J_k$ and $k = 1, 2, 3, 4, 5, 6$. Assuming that $f_W(y|\theta)$ has four parameters $\theta = (\mu, \sigma, \nu, \tau)$ then equation (5.7) for the parameters $\psi^T = (\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ is given by

$$
\begin{aligned}
g_1(\mu) = \eta_1 &= X_1\beta_1 + \sum_{j=1}^{J_1} s_{j1}(x_{j1}) \\
g_2(\sigma) = \eta_2 &= X_2\beta_2 + \sum_{j=1}^{J_2} s_{j2}(x_{j2}) \\
g_3(\nu) = \eta_3 &= X_3\beta_3 + \sum_{j=1}^{J_3} s_{j3}(x_{j3}) \\
g_4(\tau) = \eta_4 &= X_4\beta_4 + \sum_{j=1}^{J_4} s_{j4}(x_{j4}) \\
g_5(\xi_0) = \eta_5 &= X_5\beta_5 + \sum_{j=1}^{J_5} s_{j5}(x_{j5}) \\
g_6(\xi_1) = \eta_6 &= X_6\beta_6 + \sum_{j=1}^{J_6} s_{j6}(x_{j6})
\end{aligned}
\tag{5.8}
$$

5.4 Model components

The inflated GAMLSS model (5.1) comprises a mixed continuous-discrete distribution with probability (density) function $f_Y(y|\psi) = f_Y(y|\theta,\xi_0,\xi_1)$, predictor η_k including parametric terms and smoothing terms, and link functions $g_k(\cdot)$, for $k = 1,2,\cdots,K$.

5.4.1 Population distribution $f_Y(y|\psi)$

The population probability (density) function $f_Y(y|\psi)$ in model (5.1) and (5.8) for a mixed continuous-discrete distribution is given by (5.1). This includes the probability density function $f_W(y|\theta)$ of any explicit continuous distribution (e.g.beta) on (0,1) or transformed distribution (e.g. logit or truncated) on (0,1), together with point probabilities at 0 and 1. The inflated GAMLSS model (5.1) and (5.8) with mixed continuous-discrete distribution denoted D_{mix}, can be presented by

$$Y \sim D_{mix}\{g_1(\mu) = t_1, g_2(\sigma) = t_2, g_3(\nu) = t_3, g_4(\tau) = t_4, g_5(\xi_0) = t_5, g_6(\xi_1) = t_6\}$$

where D_{mix} is the mixed continuous-discrete distribution of the response variable, $(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ are the parameters of D_{mix} and $(g_1, \cdot\cdot, g_6)$ are the link functions and $(t_1, \cdot\cdot, t_6)$ are the model formulae for the explanatory terms in the predictors $(\eta_1, \cdot\cdot, \eta_6)$ respectively.

5.4.2 Link function

The model (5.1) and (5.8) assume link function $g_k(\cdot)$ is strictly monotonic and twice differentiable for $k = 1, 2, \cdot\cdot, 6$. The default link functions relate the parameters $(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ to the predictors $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$, (which depend on explanatory variables) e.g.

$$\begin{bmatrix} \mu \\ \log \sigma \\ \log \nu \\ \log \tau \\ \log\left(\frac{p_0}{p_2}\right) = log(\xi_0) \\ \log\left(\frac{p_1}{p_2}\right) = log(\xi_1) \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \eta_6 \end{bmatrix}$$

Various different link functions may be used for a single distribution parameter ψ with range $0 < \psi < 1$, e.g. logit link $[\eta = \log(\frac{\psi}{1-\psi})]$, probit link $[\eta = \Phi^{-1}(\psi)]$, where $\Phi(\cdot)$ denotes the standard normal distribution function, complementary log-log link $[\eta = \log - \log(1-\psi)]$, log-log link $[\eta = -\log - \log(\psi)]$, the symmetric Aranda-Ordaz link (Aranda-Ordaz, 1981) $[\eta = \frac{2(\psi)^{\phi} - (1-\psi)^{\phi}}{\phi(\psi)^{\phi} + (1-\psi)^{\phi}}]$, the asymmetric Aranda-Ordaz transformation $[\eta = \log((1-\psi)^{-\phi} - 1)/\phi]$, where $\phi(\cdot)$ denotes the normal density function, the Pregibon two parameter link (Pregibon, 1980) $[\frac{(\psi)^{a-b}-1}{a-b} - \frac{(1-\psi)^{a+b}-1}{a+b}]$ and a generalization of logitlink (i.e. $\eta = \frac{\psi^{\alpha}-1}{\alpha} - \frac{(1-\psi)^{\beta}-1}{\beta}$) for $0 < \psi < 1$ proposed by Scallan et al. (1984). Also a genegalization of loglink (i.e. $\eta = \frac{\psi^{\alpha-1}}{\alpha}$) for $\psi > 0$ Scallan et al. (1984).

5.4.3 The predictor

5.4.3.1 Parametric terms

In equation (5.7) the predictor η_k for $k = 1,2,\cdot\cdot,K$ are usually consist of a parametric component $X_k\beta_k$ and additive smooth components, (e.g. penalized splines), $s_{jk}(x_{jk})$ for $j = 1,2,\cdot\cdot,J_k$. The parametric component can include linear, factor, interaction terms, polynomial, fractional polynomial (Royston and Altman (1994) and piecewise polynomial (Smith (1979), Stasinopoulos and Rigby (1992) terms for explanatory variables.

5.4.3.2 Penalized splines term

Penalized splines or P-splines are usually defined on equidistant knots with a B-spline basis, de Boor (2001), and difference penalty applied directly to the B-splines parameters to control function wiggliness, see Eilers and Marx (1996) and Wood (2001). As a rule of thumb for the number of knots to be used in P-splines Ruppert et al. (2003) and Lang and Brezger (2004) suggested a moderately large number of knots, i.e. $20-40$. The default in `gamlss` is 20, but it can be changed. For a P-spline representation in the class of `GAMLSS` models, each smooth function is modelled as a regression spline function denoted by

$$s(x) = Z\gamma$$

where s is a penalised smooth function, Z is an $n \times q$ design matrix defined using B-splines basis functions for the explanatory variable x and γ is a $q \times 1$ vector of B-splines parameters. The parameters γ are estimated locally subject to a penalty term stated in matrix notation as

$$\lambda\gamma^T G\gamma = \lambda\gamma^T D_r^T D_r\gamma$$

where λ is smoothing parameter, D_r is a $(q-r) \times$ q matrix giving the r^{th} order difference of the q dimension vector γ, e.g. D_2 is given by

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & 1 & -2 & 1 & & \\ & & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \\ & & & & \cdot & \cdot & \cdot \\ & & & & 1 & -2 & 1 \end{pmatrix}$$

and $D_2^T D_2$ is a $q \times q$ matrix

$$D_2^T D_2 = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & \cdot & & & & \\ & 1 & -4 & 6 & \cdot & \cdot & & & \\ & & 1 & -4 & \cdot & \cdot & 1 & & \\ & & & 1 & \cdot & \cdot & -4 & 1 & \\ & & & & \cdot & \cdot & 6 & -4 & 1 \\ & & & & & \cdot & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}$$

The penalty is equivalent to assuming that $D_k\gamma$ is normally distributed,

$$D_k\gamma \sim N_{q-k}(0, \lambda^{-1}\mathbf{I}_{q-k})$$

5.4.4 Model estimation

A parametric inflated GAMLSS model (i.e. (5.1) and (5.8) with no smoothing terms) is fitted by maximum likelihood estimation with respect to $\beta = (\beta_1, \beta_2, \cdot\cdot, \beta_6)$ based on a sample of n independent observations. The log likelihood function for model (5.1) and (5.8) is l_{inf} given by

$$\begin{aligned} l_{inf} &= \sum_{i=1}^{n} \log f_Y(y_i|\psi^i) \\ &= \sum_{i=1}^{n} \Big((\log p_{0i})(\text{if} \quad y_i = 0) + (\log p_{1i})(\text{if} \quad y_i = 1) \\ &\quad + (\log(1 - p_{0i} - p_{1i}))(\text{if} \quad 0 < y_i < 1) + (\log(f_W(y_i|\theta))) \\ &\quad (\text{if} \quad 0 < y_i < 1) \Big) \end{aligned}$$

Notice that the log likelihood function $l_{inf}(\cdot)$ factorizes in two terms. The first of which depends only on parameters (p_0, p_1) which depend on (ξ_0, ξ_1) and the second only on the parameters of the continuous distribution $\theta = (\mu, \sigma, \nu, \tau)$. Since the parameters are separable, Pace and Salvan (1997), the maximum likelihood inference for the two sets of parameters (ξ_0, ξ_1) and (μ, σ, ν, τ) can be performed separately.

The more general model with smooth functions is fitted by maximum penalized likelihood estimation Rigby and Stasinopoulos (2005) with respect to β and γ for fixed λ. The penalized log-likelihood function for model (5.1) and (5.8) is given by

$$l_{infp} = l - \frac{1}{2} \sum_{k=1}^{6} \sum_{j=1}^{J_k} \lambda_{jk} \gamma_{jk}^T G_{jk} \gamma_{jk} \tag{5.9}$$

where G_{jk} and γ_{jk} are matrices and vectors respectively. The first and second derivatives of equation (5.9) are obtained to give the Newton-Raphson step for maximizing equation (5.9) with respect to β_k and γ_{jk} for $j = 1, 2, \cdots, J_k$ and $k = 1, 2, 3, \cdots, 6$). Each step of Newton-Raphson algorithm is achieved by using a backfitting procedure which cycles through each of the linear and smoothing parameters respectively.

5.4.5 Local estimation of smoothing parameter λ

5.4.5.1 Local random effect model

Following Rigby and Stasinopoulos (2013), a smoothing parameter λ can be estimated using a local internal random effect model expressed as

$$\begin{aligned}\varepsilon &= Z\gamma + e \\ e &\sim N(0, \sigma_e^2 W^{-1}) \\ \gamma &\sim N(0, \sigma_b^2 G^{-1})\end{aligned}$$

where Z is a basis for smoothing, W is a diagonal matrix for iterative weights w, G is a known matrix for the smoothing method used and σ_e , σ_b and γ are parameters to be estimated. The smoothing parameter λ is obtained by

$$\lambda = \frac{\sigma_e^2}{\sigma_b^2}$$

The parameters σ_e^2, σ_b^2 and γ can be estimated using following algorithm (e.g. see Rigby and Stasinopoulos (2013)). Given $\hat{\lambda}$, parameter γ can be estimated by using the penalized least square procedure,

$$\hat{\gamma} = (Z^T W Z + \hat{\lambda} G)^{-1} Z^T W \varepsilon$$

hence $\hat{\varepsilon} = Z\hat{\gamma}$ and $\hat{\sigma_e^2}$ and $\hat{\sigma_b^2}$ can be obtained by

$$\hat{\sigma_e^2} = \frac{(\varepsilon - \hat{\varepsilon})^T (\varepsilon - \hat{\varepsilon})}{n - tr(\mathsf{S})}$$

and

$$\hat{\sigma_e^2} = \frac{\hat{\gamma}^T \hat{\gamma}}{tr(\mathsf{S})}$$

where

$$\mathsf{S} = Z(Z^T W Z + \lambda G)^{-1} Z^T W$$

so $\hat{\varepsilon} = \mathsf{S}\varepsilon$. Therefore the smoothing parameter λ can be estimated by

$$\hat{\lambda} = \frac{\hat{\sigma}_e^{\,2}}{\hat{\sigma}_b^{\,2}}$$

5.4.5.2 Local generalized Akaike information criteron

Following Stasinopoulos et al. (2015), an alternative method of estimating a smoothing parameter λ is by minimizing a local generalised Akaike information criterion defined by

$$\mathsf{GAIC} = \left\| \sqrt{w} \circ (\varepsilon - Z\hat{\gamma}) \right\|^2 + k \cdot \mathrm{tr}(\mathsf{S})$$

for given k, where $\circ$ is the Hadamard element by element product operator [e.g. let $y = (y_1, y_2, y_3)^T$ and $x = (x_1, x_2, x_3)^T$ then $y \circ x = (y_1 x_1, y_2 x_2, y_3 x_3)^T$], $\| \cdot \|$ is the Euclidean vector length [i.e. $\|x\| = (\sum x_i^2)^{1/2}$] and $\mathrm{tr}(\mathsf{S})$ is the trace of matrix S. GAIC is minimized with respect to λ (which affect S and $\hat{\gamma}$). Note that $k = 2$ gives a local AIC and $k = \log(n)$ gives a local SBC. The algorithm given in sections 5.4.4 and 5.4.5.2 are used in the `GAMLSSinf` function when the additive term pb is used with *method* = *GAIC* and penalty k for fitting the P-spline, e.g.

```
gamlssinf(y~pb(x1, k=6, method=GAIC)+ pb(x2, k=6, method = GAIC))
```

5.4.5.3 Local generalised cross validation criteron

Following Stasinopoulos et al. (2015), an alternative method of estimating a smoothing parameter λ is by minimizing a local generalised cross validation defined by

$$\mathsf{GCV} = \frac{n \| \sqrt{w} \circ (\varepsilon - Z\hat{\gamma}) \|^2}{[n - tr(\mathsf{S})]^2}$$

with respect to λ.

5.4.6 Model Diagnostics

5.4.6.1 Residuals

The normalized (randomized) quantile residuals are used to assess the overall adequacy of a zero and one inflated GAMLSS model. The true normalized (randomized) residuals of a model, if the model is correct, have a standard normal distribution, i.e. $\texttt{NO}(0,1)$, and therefore mean zero, variance one, skewness equal to zero and excess kurtosis equal to 0. This is true irrespective to the underlying model distribution. In the GAMLSS framework the normalized quantile residuals are used to check the adequacy of a GAMLSS fitted model. The fitted normalized (randomized) quantile residual, Dunn and Smyth (1996), is a (randomised) version of Cox and Snell (1968), and is given by

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i) \tag{5.10}$$

where $\Phi(\cdot)$ is the cdf of standard normal distribution function and u_i in equation (5.10) is defined differently for continuous and discrete cases. Let y_i be a continuous response variable then u_i is defined as

$$u_i = F(y_i|\theta_i)$$

and

$$\hat{u}_i = F(y_i|\hat{\theta}_i)$$

where $F(y_i|\theta_i)$ is a cumulative distribution function. If the model is correctly specified u_i has the uniform random distribution between 0 and 1. Hence the normalized quantile residual (i.e. z-score) for a continuous response variable is given by

$$\hat{r}_i = \Phi^{-1}(\hat{u}_i) = \Phi^{-1}[F(y_i|\hat{\theta}_i)].$$

If y_i is an observation from a discrete response variable, the uniform random variable $\hat{u}_i$ lies on the interval

$$[\hat{u}_1, \hat{u}_2] = [F(y-1|\hat{\theta}), F(y|\hat{\theta})]$$

Here $\hat{u}_i$ is selected randomly from the interval $(\hat{u}_1, \hat{u}_2)$ and then transformed into the residual (i.e.z-score), $\hat{r} = \Phi^{-1}(\hat{u})$.

In the zero-inflated `GAMLSS` model, $\hat{u}_i$ is a uniform random value on $(0, \hat{p}_{0i})$ if $y_i = 0$ and $\hat{u}_i = F_Y(y_i|\hat{p}_{0i}, \hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$ if $y_i \in (0,1)$. On the other hand in the one-inflated model, $\hat{u}_i$ is a uniform random value on $[\hat{p}_{1i}, 1)$, if $y_i = 1$ and $\hat{u}_i = F_Y(y_i|\hat{p}_{1i}, \hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i)$ if $y_i \in (0,1)$. For the zero and one inflated `GAMLSS` model, $\hat{u}_i$ is a uniform random value on,

$$\begin{cases} (0, \hat{p}_{0i}) & \text{if} \quad y_i = 0 \\ (1-\hat{p}_{i1}, 1) & \text{if} \quad y_i = 1 \\ u_i = F_Y(y_i|\hat{p}_{0i}, \hat{p}_{1i}, \hat{\mu}_i, \hat{\sigma}_i, \hat{\nu}_i, \hat{\tau}_i) & \text{if} \quad 0 < y_i < 1 \end{cases}$$

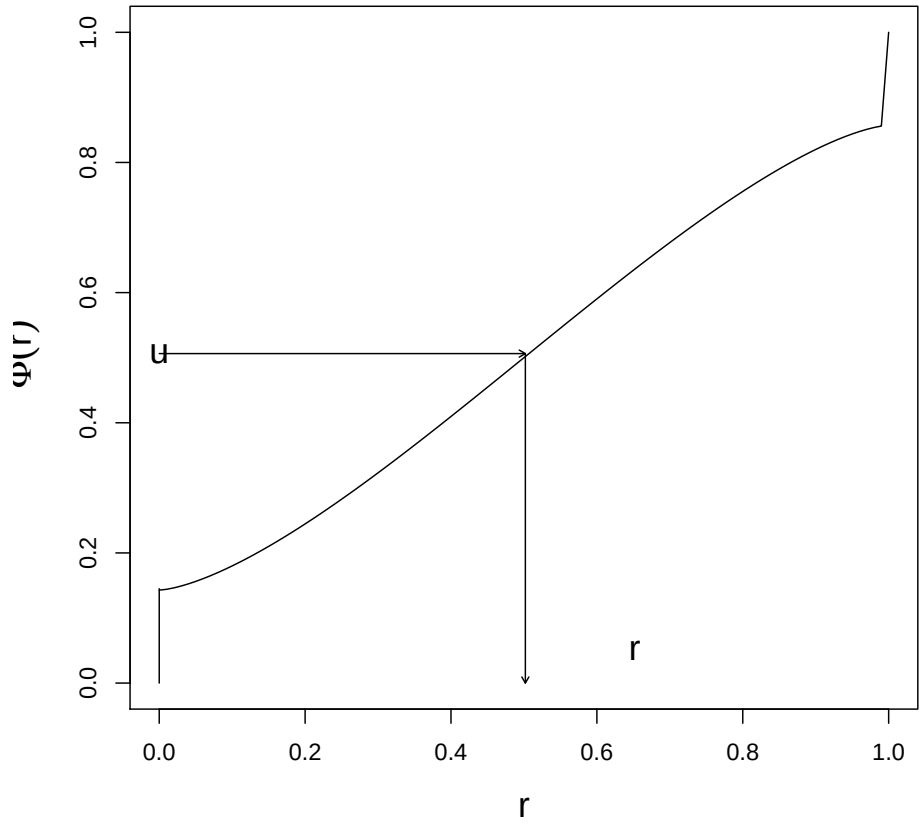


Fig. 5.1 Randomized quantile residual for inflated GAMLSS model.

A randomized procedure is conducted in order to produce a continuous residual. Note that randomized quantile residuals can vary from one realization to another. In practice several randomised set of residuals should be studied before a decision about the accuracy of the fitted model is taken. Figure 5.1 shows the randomized quantile residuals for zero and one inflated `GAMLSS` model.

A plot of the normalized (randomized) quantile residuals of a fitted model against the fitted values of the μ parameter, against the index, and against an explanatory variable should not show any detectable trends.

5.4.6.2 Global goodness-of-fit measure

Global goodness-of-fit measure pseudo R^2 can be used to check how model fits the data. The pseudo R^2 of Cox and Snell (1989)is defined by

$$R^2 = 1 - \left(\frac{\hat{L}_0}{\hat{L}_1}\right)^{\frac{2}{n}}$$

where n is the sample size, $\hat{L}_0$ and $\hat{L}_1$ are the maximum likelihood function of the null model and fitted model respectively.

5.4.7 Inflated logit skew t distribution: An example of an inflated GAMLSS model

Here a specific example of the general distribution on (0,1) inflated at 0 and 1 (given in section 5.2) is considered

The inflated logit skew t type 3 (InflogitST3) distribution is suitable for a proportion response variable on $0 \leq Y \leq 1$ that includes both 0 and 1. The inflated logitST3 distribution is a mixture of a logitST3 distribution for $0 < Y < 1$ and a Bernoulli distribution for Y at 0 or 1. The model includes three components: a discrete value 0 with probability p_0, a discrete value 1 with probability p_1 and a $logitST3(\mu, \sigma, \nu, \tau)$ distribution on the unit interval $(0,1)$ with probability $(1-p_0-p_1)$.

Let $Z \sim$ST3(μ, σ, ν, τ), let $W = \frac{1}{1+e^{-Z}} \sim logitST3(\mu, \sigma, \nu, \tau)$ therefore

$$f_W(y) = f_Z(z)\left|\frac{dz}{dy}\right|$$

where

$$\begin{aligned} z &= \log\left(\frac{y}{1-y}\right) \\ &= \log(y) - \log(1-y) \end{aligned}$$

and

$$\begin{aligned}\frac{dz}{dy} &= \frac{1}{y}+\frac{1}{1-y}\\ &= \frac{1}{y(1-y)}\end{aligned}$$

therefore

$$f_W(y) = \frac{1}{y(1-y)} f_Z(logit(y))$$

where $Z \sim ST3(\mu,\sigma,\nu,\tau)$. The mixed (continuous-discrete) probability (density) function of

$$Y \sim \text{logitST3Inf0to1}(\mu,\sigma,\nu,\tau,\xi_0,\xi_1)$$

is given by

$$f_Y(y|\mu,\sigma,\nu,\tau,\xi_0,\xi_1) = \begin{cases} p_0 & \text{if } y=0 \\ p_1 & \text{if } y=1 \\ (1-p_0-p_1) f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0<y<1 \end{cases} \tag{5.11}$$

for $0 \le y \le 1$, where $0<p_0<1$, $0<p_1<1$ and $0<p_0+p_1<1$ and $W \sim logitST3(\mu,\sigma,\nu,\tau)$ has a $logitST3$ distribution with $-\infty<\mu<\infty$ and $\sigma>0$, $\nu>0$, $\tau>0$ with probability density function given by

$$\begin{aligned} f_W(y|\mu,\sigma,\nu,\tau) &= \frac{1}{y_i(1-y_i)} \left(\frac{2\nu}{(\sigma^2(1+\nu^2)B(\frac{1}{2}\frac{\tau}{2})\tau^{\frac{1}{2}})} \right) \\ &\quad \left\{ 1+\frac{(z_i-\mu)^2}{\sigma^2\tau}[\nu^2 I(z_i<\mu)+\frac{1}{\nu^2}I(z_i\ge\mu)] \right\} \end{aligned}$$

where $z_i = logit(y_i) = log[y_i/(1-y_i)]$, and $B(\cdot,\cdot)$ is the beta function and $I(A)$ is an indicator function, where $I(A)=1$ if A is true and $I(A)=0$ if A is false.

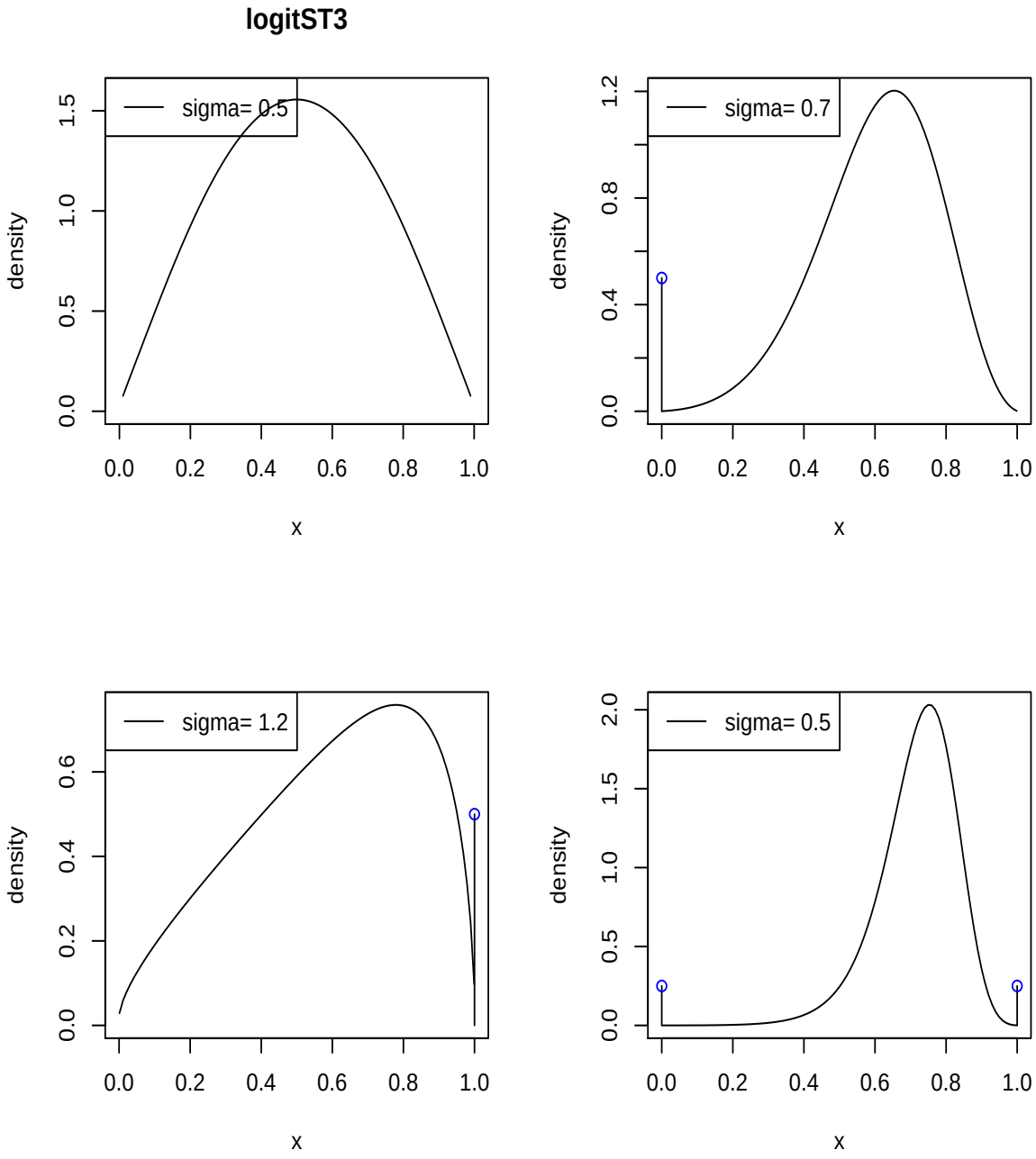


Fig. 5.2 Pdfs of logitST3, logitST3Inf0, logitST3Inf1 and logitST3Inf0to1 distribution.

The parameters ξ_0 and ξ_1, are related to p_0 and p_1 by $\xi_0 = p_0/p_2$, $\xi_1 = p_1/p_2$, where $p_2 = 1 - p_0 - p_1$, so $\xi_0 > 0$ and $\xi_1 > 0$. Hence $p_0 = \xi_0/(1+\xi_0+\xi_1)$ and $p_1 = \xi_1/(1+\xi_0+\xi_1)$.

Figure 5.2 presents logitST3, logitST3 inflated at 0, logitST3 inflated at 1 and logitST3 inflated at 0 and 1 distributions for different choices of μ, σ, ν, τ ξ_0 and ξ_1. Note that logitST3 inflated at

0, logitST3 inflated at 1 and logitST3 inflated at 0 and 1 distributions have same functional shape on the interval (0,1). However they differ at the mass point(s), being at 0 for logitST3Inf0 (i.e. inflated at 0), being at 1 for logitST3Inf1 (i.e. inflated at 1) and at 0 and 1 for logitST3Inf0to1 (i.e inflated at 0 and 1).

The default link functions (in the `GAMLSS` package) relate the parameters $(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$ to the predictors $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5, \eta_6)$, i.e.

$$\begin{bmatrix} \mu \\ \log \sigma \\ \log \nu \\ \log \tau \\ \log\left(\frac{p_0}{p_2}\right) = log(\xi_0) \\ \log\left(\frac{p_1}{p_2}\right) = log(\xi_1) \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \\ \eta_6 \end{bmatrix}$$

The dependence of the predictors of the parameters (i.e. η_1 to η_6) on explanatory variables may be linear, nonlinear, non-parametric smooth, regression trees or neural network models.

Model (5.11) can be fitted by fitting two models: a logitST3(μ, σ, ν, τ) distribution for $0 < Y < 1$, together with a multinomial distribution with three levels, denoted by $MN3(\xi_0, \xi_1)$ in the `GAMLSS` package, for a recoded response factor Y_1 given by (5.2) and (5.3), where $\xi_0 = p_0/p_2$ and $\xi_1 = p_1/p_2$ and $p_2 = 1 - p_0 - p_1$, giving $\xi_0 > 0$ and $\xi_1 > 1$. Alternatively model (5.11) can be fitted more easily using a new function `gamlssinf()`.

The log likelihood function for the `logitST3Inf0to1` model (5.11) is equal to the sum of the log likelihood functions of the `logitST3` model for $0 < y < 1$ and the multinomial $MN3$ model (5.2 and (5.3). Here the likehood function factorises in two different terms; the first term depends only on the ξ_0 and ξ_1 the second term depends on μ, σ, ν and τ. Observed likelihood quantities for inference about ξ_0 and ξ_0 only depend on the first term and inference about ξ_0 and ξ_1 can take place separately from that about μ, σ, ν, and τ as if the values of μ, σ, ν, and τ were known, Pace and Salvan (1997).

Hence the parameter sets (μ,σ,ν,τ) and (ξ_0,ξ_1) are 'information' orthogonal. Fisher's information matrix for the inflated logitST3 distribution can be written as

$$K(\theta)=\begin{pmatrix} k_{\xi_0\xi_0} & k_{\xi_0\xi_1} & 0 & 0 & 0 & 0 \\ k_{\xi_1\xi_0} & k_{\xi_1\xi_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & k_{\mu\mu} & k_{\mu\sigma} & k_{\mu\nu} & k_{\mu\tau} \\ 0 & 0 & k_{\sigma\mu} & k_{\sigma\sigma} & k_{\sigma\nu} & k_{\sigma\tau} \\ 0 & 0 & k_{\nu\mu} & k_{\nu\sigma} & k_{\nu\nu} & k_{\nu\tau} \\ 0 & 0 & k_{\tau\mu} & k_{\tau\sigma} & k_{\tau\nu} & k_{\tau\tau} \end{pmatrix}$$

5.4.8 Inflated truncated skew power exponential: An example of an inflated GAMLSS model

Here a second specific example of the general distribution on (0,1) inflated at 0 and 1 (given in section 5.2) is considered. Suppose a random variable Z has a distribution on $(-\infty,\infty)$ with pdf and cdf specified by $f_Z(\cdot)$ and $F_Z(\cdot)$ respectively. Let Y_{tr} be a random variable representing truncated version of the distribution over the interval [0,1]. The resulting probability density function for Y_{tr} is given by

$$f_{Ytr}(y) \quad = \quad \begin{cases} \frac{f_Z(y)}{F_Z(1)-F_Z(0)}, & \text{if } 0\le y\le 1 \\ 0, & \text{otherwise} \end{cases}$$

If Z has a four parameter distribution denoted D in general, i.e. $Z\sim D(\mu,\sigma,\nu,\tau)$, then Y_{tr} has a truncated D distribution, denoted $Y_{tr}\sim Dtr(\mu,\sigma,\nu,\tau)$.

For example if $Z\sim SEP(\mu,\sigma,\nu,\tau)$ then $Y_{tr}\sim$SEPtr(μ,σ,ν,τ). The $SEPtr$ distribution is created using the GAMLSS function gen.trun(), which allows any gamlss distribution (e.g. SEP) to be converted into a truncated distribution. The truncated distribution SEPtr(μ,σ,ν,τ) on (0,1) can be inflated with probabilities at 0 and 1. The resulting mixed continuous-discrete probability (density) function of

$$Y \sim \text{SEPtrInf0to1}(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$$

is given by

$$f_Y(y|\mu,\sigma,\nu,\tau,\xi_0,\xi_1) = \begin{cases} p_0, & \text{if } y = 0 \\ \frac{f_Z(y|\mu,\sigma,\nu,\tau)}{F_Z(1)-F_Z(0)}, & \text{if } 0 < y < 1 \\ p_1, & \text{if} y = 1 \end{cases}$$

for $0 \leq y \leq 1$ and $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, $\tau > 0$ and $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$, where $\xi_0 = p_0/p_2$, $\xi_1 = p_1/p_2$ and $p_2 = 1 - p_0 - p_1$.

Chapter 6

Generalized Tobit GAMLSS model

6.1 Introduction

This chapter outlines a generalized Tobit GAMLSS model for a proportion response variable on the interval [0,1), (0,1] or [0,1]. The Tobit model, Tobin (1958), and the two sided version of the Tobit model, Rosett and Nelson (1975), were considered to establish the ideas of the new model. The proposed model includes a flexible distribution assumption considering that the normal distribution is not adequate for all data. The model primarily focuses on censoring below 0 or above 1 or both. Each of the parameters of the assumed distribution can be modelled as linear, P-splines, neural network or decision tree functions of explanatory variables.

6.2 Tobit model

The basic assumption of the standard Tobit model is that the dependent variable has a normal distribution which is left censored to the semi closed interval $[a,\infty)$ (Tobin, 1958) or the closed interval $[a,b]$ for the two sided version (Rosett and Nelson, 1975) and is conditional on some covariates $x = (x_1, x_2, \cdot, \cdot, \cdot, x_p) \in \mathbb{R}^p$. Here we consider the situation where $a = 0$ and $b = 1$. Let observed variable Y be a censored version of latent variable V. The standard Tobit model

originally proposed by Tobin (1958) can be defined in the following way

$$V \sim NO(\mu, \sigma)$$

and

$$\begin{aligned} Y &= V \quad \text{if} \quad V > 0 \\ &= 0 \quad \text{if} \quad V \leq 0 \end{aligned} \tag{6.1}$$

Two sided version of Tobit model (Rosett and Nelson, 1975) can be defined by

$$\begin{aligned} Y &= 0 \quad \text{if} \quad V \leq 0 \\ &= V \quad \text{if} \quad 0 < V < 1 \\ &= 1 \quad \text{if} \quad V \geq 1 \end{aligned} \tag{6.2}$$

Furthermore the mean function (μ) of the latent variable V is modelled in the following way ,

$$\mu = x^T \beta, \quad \beta \in \mathbb{R}^p$$

See Maddala (1983) and Takeshi (1984, 1985) for an explanation of the mean function and associated details.

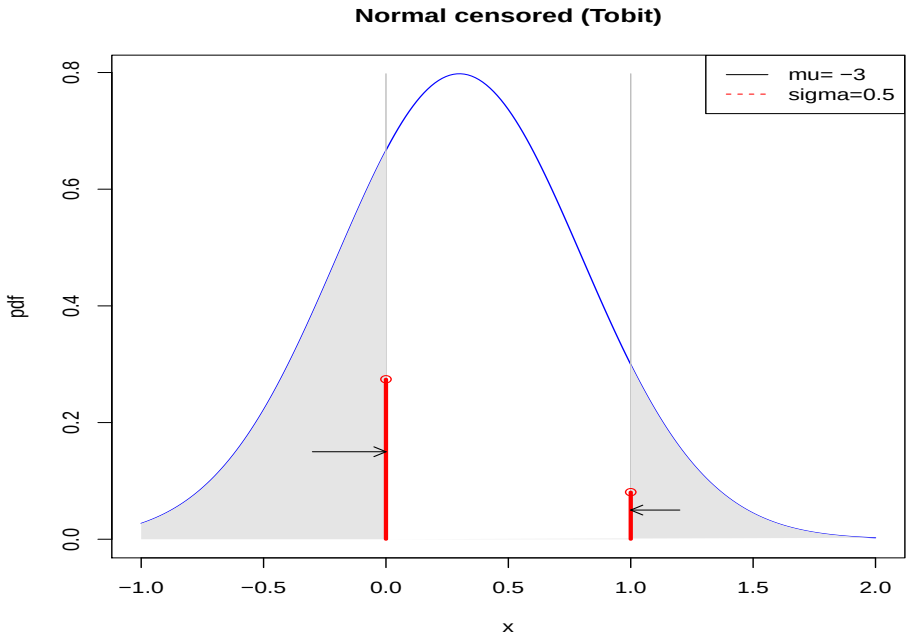


Fig. 6.1 Two sided version of Tobit model

Maddala and Nelson (1975) and Arabmazar and Schmidt (1982) have shown that the Tobit model is sensitive to distribution assumptions. It is clear that the underlying assumption of the normal distribution for the latent variable V may not be adequate for all data sets. Here an alternative approach is proposed to replace the normal distribution by a more flexible GAMLSS distribution on the real line ($\mathbb{R}$). Since the model is developed within the GAMLSS framework, the model can easily accommodate all the GAMLSS features. The generalised Tobit GAMLSS model is proposed for observed variable Y having range from 0 to 1 including 0 and (or) 1.

Figure 6.1 shows censored normal distribution on [0,1] which is obtained by censoring a normal distribution on $(-\infty, \infty)$ below 0 and above 1 to give point probabilities at 0 and 1.

6.3 Generalized Tobit model for $0 \leq y \leq 1$

The proposed model is a generalization of the Tobit model specified in equation 6.2. Let $V \sim D(\theta)$ be a latent variable whose parameters are conditional on covariates x. The model assumes that the latent variable V has a distribution on $(-\infty, \infty)$ with probability density function

$f_V(v|\theta)$ and cumulative distribution function $F_V(v|\theta)$. The observed variable Y then depends on the latent variable as in equation (6.2).

The distribution of the observed (interval censored) variable Y is denoted by

$$Y \sim D_{ic}(\theta)$$

where subscript ic indicates that the distribution D is interval censored to interval [0,1], i.e. below 0 and above 1. The resulting mixed continuous-discrete probability (density) function of Y is given by

$$f_Y(y|\theta) = \begin{cases} F_V(0), & \text{if y=0} \\ f_V(y|\theta), & \text{if 0<y<1} \\ 1 - F_V(1), & \text{if y=1} \end{cases} \tag{6.3}$$

for $0 \leq y \leq 1$, where θ is the parameter vector. The cumulative distribution function of Y [with pdf given by (6.3)] is given by (5.4), where $p_0 = F_V(0)$ and $p_1 = 1 - F_V(1)$.

In the generalised Tobit models the probabilities of Y at 0 and 1 are directly related to the distribution between 0 and 1 and so are less flexible, but the model is more concise (i.e. parsimonious) in that it has less parameters. Also the Tobit model is usually not sensitive to values of Y very close to 0 or 1.

6.3.1 Generalised Tobit model for $0 \leq y < 1$ and $0 < y \leq 1$

Two other general classes of models can be derived similar to model (6.3), when the response variable is recorded on the interval [0,1) or (0,1]. A generalized Tobit model on (0,1] can be obtained by censoring above 1 a flexible model response variable distribution on $(0, \infty)$ to give its positive probability at 1. Censoring refers to the transformation of observations outside the limiting interval to the border value, Hoff (2007). Here the values of Y in the model distribution above 1 are transformed to 1.

Let $V \sim D(\theta)$ be a flexible uncensored distribution on $(0,\infty)$. Let $Y \sim D_{rc}(\theta^T)$ be the corresponding right censored distribution on $(0,1]$, i,e censored above 1. Then

$$\begin{aligned} Y = V, \quad & \text{if} \quad 0 < V < 1 \\ = 1, \quad & \text{if} \quad V \geq 1 \end{aligned}$$

Hence the probability density function of Y is given by

$$f_Y(y|\theta) = \begin{cases} f_V(y|\theta), & \text{if 0<y<1} \\ 1 - F_V(1), & \text{if y=1} \end{cases} \tag{6.4}$$

for $0 < y \leq 1$. For the observed response variable on [0,1) a left censored distribution on $(0,\infty)$ can be used by transforming dependent variable Y in (6.4) into $(1-Y)$.

6.4 Generalized Tobit GAMLSS model

The generalized Tobit GAMLSS model is implemented using the generalized additive model for location, scale and shape (GAMLSS) framework by Rigby and Stasinopoulos (2005). For example assume that an observation is right censored at 1, then its contribution to the log likelihood is given by

$$\log[1 - F(1|\theta)]$$

where $F(1|\theta)$ is the cdf at 1. Hence incorporation of censoring require computation of $F(1|\theta)$.

In order to extend the model to the GAMLSS regression case we relate the parameters of the distribution of the latent variable V through link functions to the linear and smooth terms in explanatory variables by

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} s_{jk}(x_{jk}) \tag{6.5}$$

where $g_k(\cdot)$ is a monotonic link function, θ_k is a parameter vector of length n and β_k is a parameter vector of length J_k', x_{jk} is a explanatory variable vector of length n and s_{jk} is a unknown smooth function of the variable x_{jk}, for $J = 1,2,\cdot\cdot,J_k$ and $k = 1,2,\cdot\cdot,p$. Note that if no censoring occurs the generalised Tobit model would be a general form of GAMLSS regression model (Rigby and Stasinopoulos, 2005).

6.4.1 Likelihood inference

Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ be independent observations with range [0,1] from the model given by (6.3) and (6.5) and $\theta^T = (\theta_1, \theta_2, \cdot, \cdot, \cdot, \theta_p)$ be a parameter vector. The likelihood function for the observed $\mathbf{y}$ is given by

$$L_c = \prod_{i=1}^{n} \left\{ F_Y(y_i)^{\mathbf{I}(y_i=0)} f_Y(y_i)^{\mathbf{I}(0<y_i<1)} S_Y(y_i)^{\mathbf{I}(y_i=1)} \right\} \tag{6.6}$$

where $\mathbf{I}(A)$ denotes indicator function equating 1 if A is true, otherwise 0. $S_Y(y_i)$ is a survival function,which can be expressed as

$$S_Y(y|\theta) = 1 - F_Y(y|\theta)$$

Therefore the log likelihood function (l_c) is obtained by

$$\begin{aligned} l_c &= \sum_{i=1}^{n} \{\mathbf{I}(y_i = 0) \log F_Y(y_i|\theta_i) + \mathbf{I}(0 < y_i < 1) \log f_Y(y_i|\theta_i)\} \\ &\quad + \mathbf{I}(y_i = 1) \log S_Y(y_i|\theta_i)\} \\ &= \sum_{i=1}^{n} \{\mathbf{I}(y_i = 0) \log F_Y(y_i|\theta_i) + \mathbf{I}(0 < y_i < 1) \log f_Y(y_i|\theta_i)\} \\ &\quad + \mathbf{I}(y_i = 1) \log(1 - F_Y(y_i|\theta_i))\} \end{aligned} \tag{6.7}$$

where $\theta^i = (\theta_{i1}, \theta_{i2}, \cdot\cdot, \theta_{ip})^T$. Hence for fixed smoothing parameter λ_{jk}, the fixed and random effect parameters β and γ respectively are estimated by maximising a penalised likelihood

function (l_{cp})

$$l_{cp} = l_c + \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k}\lambda_{jk}\gamma_{jk}^T G_{jk}\gamma_{jk} \tag{6.8}$$

where l_c is the log likelihood from equation (6.7) and G_{jk} and γ_{jk} are matrices and vectors respectively defined in section 5.4.3.2. The maximum (penalised) likelihood estimation is achieved through a Newton-Raphson, Fisher scoring or quasi Newton-Raphson algorithm.

6.4.2 Residuals

Randomized quantile residuals discussed in Chapter 3 are also appropriate for a censored response variable. A fitted normalised (randomized) quantile residual (Dunn and Smyth, 1996) for model (6.3) is given by

$$\hat{r}_i = \Phi^{-1}[\hat{u}_i]$$

where $\hat{u}_i = F_V(y_i|\hat{\theta}^i)$ if $0 < y_i < 1$ and u_i is a random value from a uniform distribution given by

$$\begin{aligned} U(0,p_{0i}), \quad \text{where} p_{0i} &= F_V(0|\hat{\theta}^i), \quad \text{if} \quad y_i = 0 \\ U(1-p_{1i},1), \text{where} p_{1i} &= 1 - F_V(1|\hat{\theta}^i), \quad \text{if} \quad y_i = 1 \end{aligned}$$

and $\theta^i = (\theta_{i1}, \theta_{i2}, \cdot\cdot, \theta_{ip})^T$ for $i = 1, 2, \cdot\cdot, n$. The true residual r_i has a standard normal distribution if the model is correct.

6.5 Interval censored BSSN distribution: An example of the generalized Tobit GAMLSS model

Let latent variable V has a bi-modal skew symmetric normal distribution on $(-\infty, \infty)$ denoted by

$$V \sim \text{BSSN}(\mu, \sigma, \nu, \tau)$$

where the pdf of the BSSN distribution is given by

$$f_V(y|\mu,\sigma,\nu,\tau) = c\left[\tau+(y-\nu)^2\right]e^{-\sigma(y-\mu)^2} \tag{6.9}$$

where

$$c = \frac{2\sigma^{3/2}}{1+2\sigma(\tau+(\nu-\mu)^2)}$$

where μ and ν are location parameters and σ and τ are scale and bimodality parameters respectively with $\sigma > 0$ and $\tau > 0$. For large τ distribution is close to normal $N(\mu, 1/(2\sigma))$ distribution. Variable Y is obtained by censoring V below 0 and above 1 as defined by equation (6.2).

The resulting distribution of variable Y is denoted by

$$Y \sim \text{BSSNic}(\mu,\sigma,\nu,\tau)$$

with mixed continuous-discrete probability(density) function given by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \begin{cases} F_V(0|\mu,\sigma,\nu,\tau), & \text{if y=0} \\ f_V(y|\mu,\sigma,\nu,\tau), & \text{if 0<y<1} \\ 1-F_V(1|\mu,\sigma,\nu,\tau), & \text{if y=1} \end{cases}$$

for $0 \leq y \leq 1$.

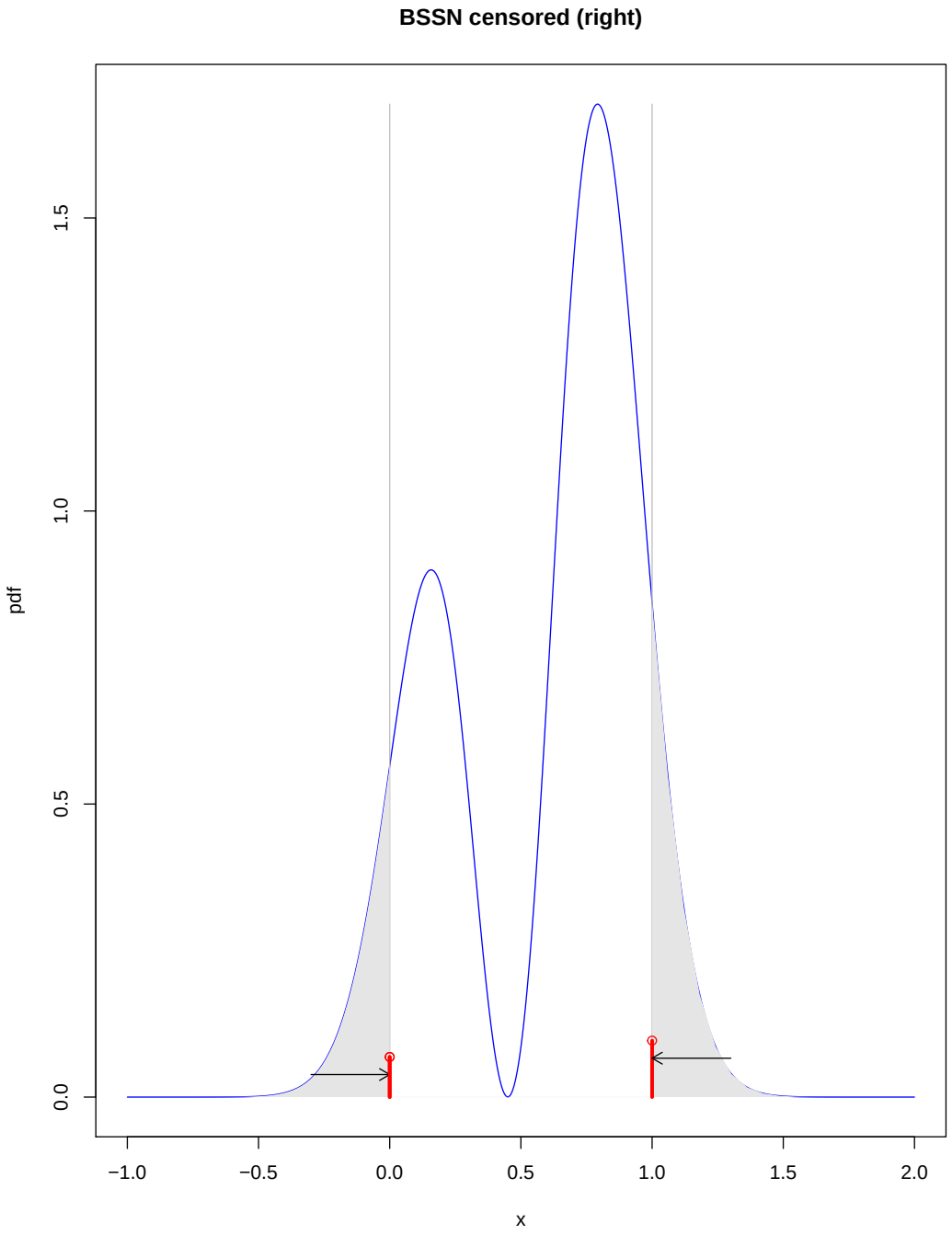


Fig. 6.2 Interval censored bimodal skew symmetric normal distribution

Figure 6.2 shows the pdf of the bimodal skew symmetric normal distribution censored below 0 and above 1.

All four parameters (μ, σ, ν, τ) of the interval censored BSSN can be modelled in terms of explanatory variables using suitable default link functions and can be given by

$$\begin{aligned}
\mu = \eta_1 &= X_1^T \beta_1 + \sum_{j=1}^{J_1} s_{j1}(x_{j1}) \\
\log(\sigma) = \eta_2 &= X_2^T \beta_2 + \sum_{j=1}^{J_2} s_{j2}(x_{j2}) \\
\nu = \eta_3 &= X_3^T \beta_3 + \sum_{j=1}^{J_3} s_{j3}(x_{j3}) \\
\log(\tau) = \eta_4 &= X_4^T \beta_4 + \sum_{j=1}^{J_4} s_{j4}(x_{j4})
\end{aligned}$$

The $s_{jk}(x_{jk})$ functions are modelled here using P-splines (Eilers and Marx, 1996). The advantage of inclusion of non-parametric P-spline terms is to identify non-linear relationships between the response and each explanatory variable. Penalised B-splines are able to select the degree of smoothing automatically using penalised maximum likelihood estimation. The selection of degree of smoothing can be achieved by selecting the corresponding smoothing parameters λ using either a local random effect model, a local generalized Akaike information criterion or a local generalized cross validation criterion, see section 5.4.5.

Chapter 7

The GAMLSSinf package in R

The new R package gamlss.inf is designed to fit inflated distributions on the interval $[0,1]$ which were described in chapter 5. The **gamlss** package already provides the inflated beta distribution, BEINF which allows the user to fit a beta distribution on $(0,1)$ with extra point probability at 0 and 1. The probability at the points 0 and 1 may depend on explanatory variables. Since the beta distribution has 2 parameters, the inflated beta (with the addition of the two mass points at 0 or/and 1) has a total of 4 parameters. In practice, and for complicated data sets, the part of the response which lies on $(0,1)$ may need more than 2 distribution parameters to be captured correctly. The **R** package **gamlss.dist** provides through the function `gen.Family(...,type="logit")` the facility of taking any distribution from $(-\infty,\infty)$ and mapping it into $(0,1)$ using an inverse logit transformation. The new **R** package gamlss.inf enhances this capability of the **gamlss.dist** package in that the distribution between $(0,1)$ (up to four parameters) can be inflated with probability at 0 and/or 1. The overall distribution can then have up to six parameters. Let μ,σ,ν,τ represent the four parameters of the the distribution defined on $(0,1)$ and ξ_0 and ξ_1 be parameters related to the probability at 0 and 1 respectively. Then the general inflated $[0,1]$ model that the new package gamlss.inf can fit can be written as:

$$
\begin{aligned}
Y &\overset{\text{ind}}{\sim} \mathscr{D}(\mu,\sigma,\nu,\tau,\xi_0,\xi_1) \\
\eta_1 &= g_1(\mu) = \mathbf{X}_1\beta_1 + s_{11}(\mathbf{x}_{11}) + \ldots + s_{1J_1}(\mathbf{x}_{1J_1}) \\
\eta_2 &= g_2(\sigma) = \mathbf{X}_2\beta_2 + s_{21}(\mathbf{x}_{21}) + \ldots + s_{2J_2}(\mathbf{x}_{2J_2}) \\
\eta_3 &= g_3(\nu) = \mathbf{X}_3\beta_3 + s_{31}(\mathbf{x}_{31}) + \ldots + s_{3J_3}(\mathbf{x}_{3J_3}) \\
\eta_4 &= g_4(\tau) = \mathbf{X}_4\beta_4 + s_{41}(\mathbf{x}_{41}) + \ldots + s_{4J_4}(\mathbf{x}_{4J_4}) \\
\eta_5 &= g_5(\xi_0) = \mathbf{X}_5\beta_5 + s_{51}(\mathbf{x}_{51}) + \ldots + s_{5J_5}(\mathbf{x}_{5J_5}) \\
\eta_6 &= g_6(\xi_1) = \mathbf{X}_6\beta_6 + s_{61}(\mathbf{x}_{61}) + \ldots + s_{6J_6}(\mathbf{x}_{6J_6})
\end{aligned}
\tag{7.1}
$$

where $\mathscr{D}(\mu,\sigma,\nu,\tau,\xi_0,\xi_1)$ is a distribution of the response variable Y defined on $[0,1]$ given by (5.1) where $\theta = (\mu,\sigma,\nu,\tau)$, $\mathbf{X}_k$ are the design matrices incorporating the linear additive terms in the model, β_k are the linear coefficient parameters and $s_{kj}(\mathbf{x}_{kj})$ represent smoothing functions for explanatory variables $\mathbf{x}_{kj}$, for $k = 1,2,3,4,5,6$ and $j = 1,\ldots,J_k$. Note that the quantitative explanatory variables in the $\mathbf{X}$'s can be the same or different for the ones defined in the smoothers. The vectors η_1, η_2, η_3, η_4 η_5 and η_6 are called the predictors of the distribution parameters μ, σ, ν, τ, ξ_0 and ξ_1 respectively.

7.1 Distributions on $(0,1)$

7.1.1 Explicit distributions on $(0,1)$

Within the `gamlss.dist` package there are currently three distributions defined on $(0,1)$,

1. the beta distribution, `BE`, with two parameters,
2. the logit normal distribution, `LOGITNO`, with two parameters and
3. the generalised beta type 1 distribution, `GB1`, with four parameters.

7.1.2 Logit distributions on $(0,1)$

In addition, as described in section 3.1.7, any continuous random variable say Z defined on $(-\infty,\infty)$ can be transformed by the inverse logit transformation $Y = 1/(1+\exp(-Z))$ to a random variable Y defined on $(0,1)$. For example if Z is a t-family distributed variable i.e. $Z \sim \texttt{TF}(\mu,\sigma,\nu)$, and the inverse logit transformation is applied, then $Y \sim \texttt{logitTF}(\mu,\sigma,\nu)$, i.e. a logit-$t$ family distribution on $(0,1)$.

The following is an example on how to take a `gamlss.family` distribution on $(-\infty,\infty)$ and create a corresponding logit distribution defined on $(0,1)$. The gamlss function `gen.Family()` of the **gamlss.dist** package generates the `d` (pdf), `p` (cdf), `q` (inverse cdf) and `r` (random generation) functions of the distribution together with the function which can be used for fitting within **gamlss**. Here first generate a logit-t distribution and and in Fig. 7.1 plot the distribution for different values of μ, σ and ν. Note that μ, σ and ν are defined on the original t-distribution ranges $(-\infty,\infty)$ for μ and $(0,\infty)$ for σ and ν. This implies that $1/(1+\exp(-\mu))$ is not the mean of the logit distribution but its median. Also σ and ν are related to the scale and shape of the distribution. Next use `gen.Family("TF", type="logit")` to generate a logit-t distribution and then plot the distribution (see Fig. 7.1) for different values of μ, σ and ν using the function `curve()`.

```
# generate the distribution
library(gamlss)
gen.Family("TF", type="logit")

## A  logit  family of distributions from TF has been generated
##  and saved under the names:
##  dlogitTF plogitTF qlogitTF rlogitTF logitTF

# different mu
curve(dlogitTF(x, mu=-5, sigma=1, nu=10), 0,1, ylim=c(0,3))
title("(a)")
curve(dlogitTF(x, mu=-1, sigma=1, nu=10), 0,1, add=TRUE, lty=2)
curve(dlogitTF(x, mu=0, sigma=1), 0,1, add=TRUE, lty=3)
curve(dlogitTF(x, mu=1, sigma=1), 0,1, add=T, lty=4)
curve(dlogitTF(x, mu=5, sigma=1), 0,1, add=T, lty=5)
# different sigma
curve(dlogitTF(x, mu=0, sigma=.5, nu=10), 0,1, ylim=c(0,3))
```

```
title(("(b)"))
curve(dlogitTF(x, mu=0, sigma=1, nu=10), 0,1, add=TRUE, lty=2)
curve(dlogitTF(x, mu=0, sigma=2, nu=10), 0,1, add=TRUE, lty=3)
curve(dlogitTF(x, mu=1, sigma=5, nu=10 ), 0,1, add=T, lty=4)
# different nu
curve(dlogitTF(x, mu=0, sigma=1, nu=1000), 0,1, ylim=c(0,3))
title("(c)")
curve(dlogitTF(x, mu=0, sigma=1, nu=10), 0,1, add=TRUE, lty=2)
curve(dlogitTF(x, mu=0, sigma=2, nu=5), 0,1, add=TRUE, lty=3)
curve(dlogitTF(x, mu=0, sigma=2, nu=1), 0,1, add=TRUE, lty=4)
```

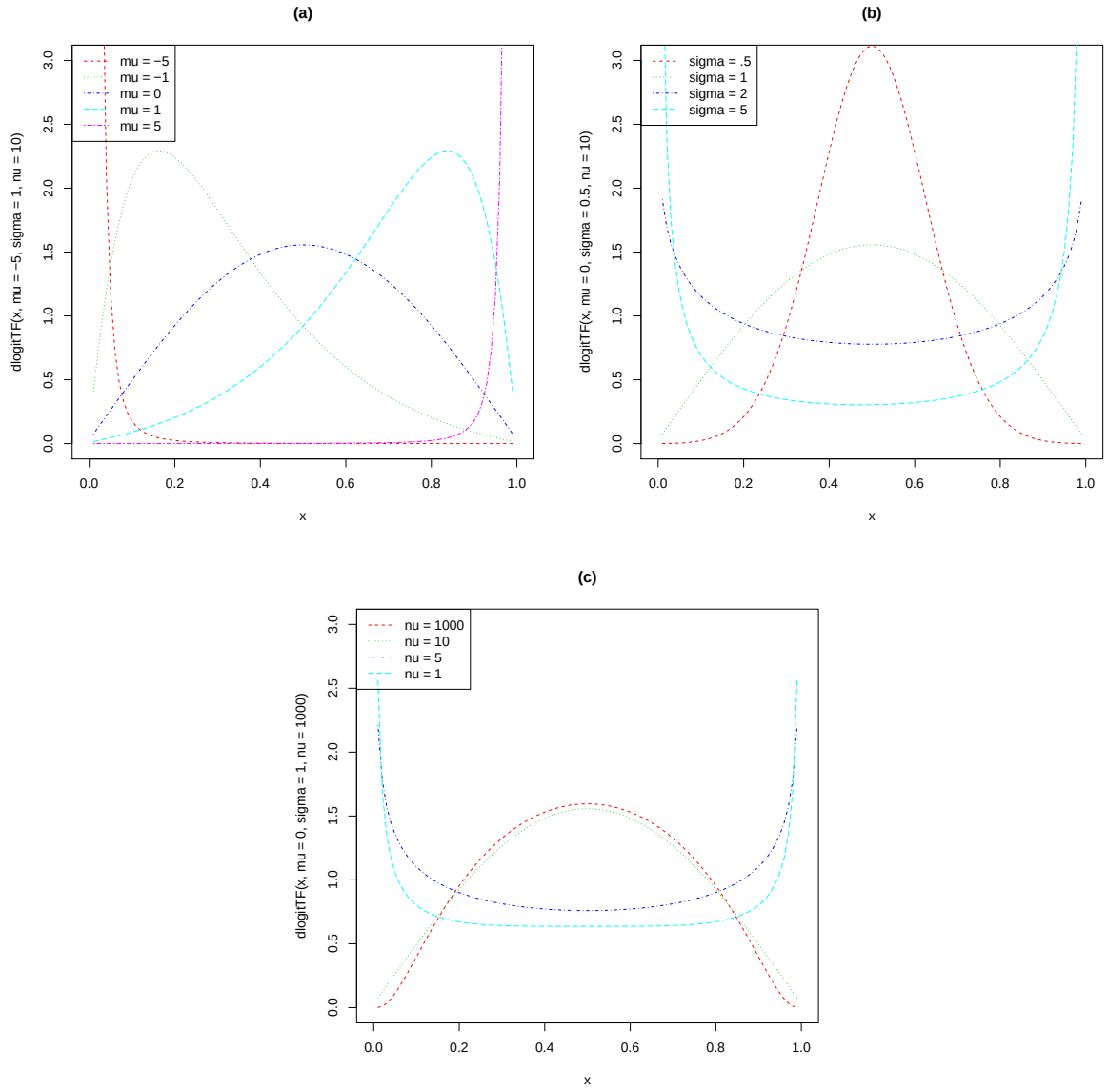


Fig. 7.1 A logit-t distribution: (a) with values $\mu = (-5,-1,0,1,5)$, $\sigma = 1$ and $\nu = 10$, (b) with values $\mu = 0$, $\sigma = (0.5,1,2,5)$ and $\nu = 10$ and (c) with values $\mu = 0$, $\sigma = 1$ and $\nu = (1000,10,5,1)$.

Figure 7.1 shows the different shapes the distribution can take. Panel (a) shows for fixed $\sigma = 1$ and $\nu = 10$ how the distribution changes for different values of $\mu = (-5, -1, 0, 1, 5)$. Panel (b) for fixed $\mu = 0$ and $\nu = 10$ varies $\sigma = (0.5, 1, 2, 5)$. Finally panel (c) fixes $\mu = 0$ and $\sigma = 1$ and varies $\nu = (1000, 10, 5, 1)$.

7.2 Truncated distributions on (0,1)

As discussed in section 3.1.8 any distribution defined on the real line $(-\infty, \infty)$ can be left truncated at 0 and right truncated 1 to give a truncated distribution on (0,1) using the function `gen.trun()` from the `R` package **gamlss.trun**.

7.3 Generating inflated distributions on $[0, 1]$

Next it is shown how any `gamlss.family` distribution defined on $(0, 1)$ can be extended by inflation to $[0, 1]$.

The function `gen.Inf0to1()` takes as an argument a `gamlss.family` distribution on $(0, 1)$ and generates an inflated version of the distribution with point probabilities at 0 and/or 1. The function has two arguments, `family` and `type.of.Inflation`. The first specifies a distribution family on $(0, 1)$, while the second specifies the type of inflation. The options are i) `"Zero"`, `"One"` and `"Zero&One"`.

The resulting mixed continuous-discrete probability (density) function (pdf) for option "*Zero&One*" is given by

$$f_Y(y|\theta, \xi_0, \xi_1) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0 - p_1) f_W(y|\theta) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{7.2}$$

for $0 \leq y \leq 1$, where $f_W(y|\theta)$ is any probability density function defined on $(0,1)$, i.e. for $0 < y < 1$, with parameters $\theta^T = (\theta_1, \theta_2, \cdot\cdot, \theta_p)$ and $0 < p_0 < 1, 0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$ and where $\xi_0 = \frac{p_0}{p_2}$, $\xi_1 = \frac{p_1}{p_2}$, where $p_2 = 1 - p_0 - p_1$, so $\xi_0 > 0$ and $\xi_1 > 0$. Hence

$$\begin{pmatrix} p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} \frac{\xi_0}{(1+\xi_0+\xi_1)} \\ \frac{\xi_1}{(1+\xi_0+\xi_1)} \end{pmatrix}$$

However for option "*Zero*" the pdf is

$$f_Y(y|\theta,\xi_0) = \begin{cases} \xi_0 & \text{if } y = 0 \\ (1-\xi_0) f_W(y|\theta) & \text{if } 0 < y < 1 \end{cases} \tag{7.3}$$

so in this case $\xi_0 = P(Y = 0)$.

Also for option "*One*" the pdf is

$$f_Y(y|\theta,\xi_1) = \begin{cases} (1-\xi_1) f_W(y|\theta) & \text{if } 0 < y < 1 \\ \xi_1 & \text{if } y = 1 \end{cases} \tag{7.4}$$

so in this case $\xi_1 = P(Y = 1)$.

In the example below first take the skew t-family distribution, `SST`, and use the `gen.Family()` function in the **gamlss.dist** package to generate the distribution `logitSST` defined on $(0,1)$. By using the function `gen.Inf0to1()` on the new generated `logitSST` distribution, an inflated logitSST distribution, inflated at 0 and 1, is created.

```
library(gamlss.inf)
gen.Family(family="SST", type="logit")

## A  logit  family of distributions from SST has been generated
##  and saved under the names:
##  dlogitSST plogitSST qlogitSST rlogitSST logitSST

gen.Inf0to1(family="logitSST",  type.of.Inflation="Zero&One")
```

```
## A  0to1 inflated logitSST distribution has been generated
##  and saved under the names:
##  dlogitSSTInf0to1 plogitSSTInf0to1 qlogitSSTInf0to1 rlogitSSTInf0to1
##  plotlogitSSTInf0to1
```

There are five function generated here:

dlogitSSTInf0to1 The pdf of the distribution, d function.

plogitSSTInf0to1 The cdf of the distribution, p function.

qlogitSSTInf0to1 The inverse cdf of the distribution, q function.

rlogitSSTInf0to1 The random generating function of the distribution, r function.

logitSSTInf0to1 The function for fitting the distribution and

plotlogitSSTInf0to1 The function for plotting the pdf of the distribution.

7.4 Plotting inflated distributions on $[0, 1]$

The newly created `plotlogitSSTInf0to1()` function can be used to plot the pdf of the inflated distribution (which is a mixed continuous-discrete distribution). Figure 7.2 shows the use of the `plotlogitSSTInf0to1()` function. The function plots the inflated distribution function including point probabilities at zero and one. Unfortunately in its present form only one plot is allowed per figure. Figure 7.2 shows eight different realisations of the distribution for different values of the parameters.

```
plotlogitSSTInf0to1(mu= 1, sigma=1, nu=1, tau=10, xi0=.1, xi1=.2);
plotlogitSSTInf0to1(mu=-1, sigma=1, nu=1, tau=10, xi0=.1, xi1=.2);
plotlogitSSTInf0to1(mu=-1, sigma=2, nu=1, tau=10, xi0=.1, xi1=.2);
plotlogitSSTInf0to1(mu=0,  sigma=2, nu=1, tau=10, xi0=.1, xi1=.2);
plotlogitSSTInf0to1(mu=0,  sigma=1, nu=10,tau=10, xi0=.1, xi1=.2);
plotlogitSSTInf0to1(mu=0,  sigma=1, nu=1, tau=3,  xi0=.1, xi1=.2);
plotlogitSSTInf0to1(mu=0,  sigma=1, nu=2, tau=3,  xi0=.5, xi1=.1);
plotlogitSSTInf0to1(mu=0,  sigma=1, nu=.3,tau=100,xi0=.1, xi1=.5);
```

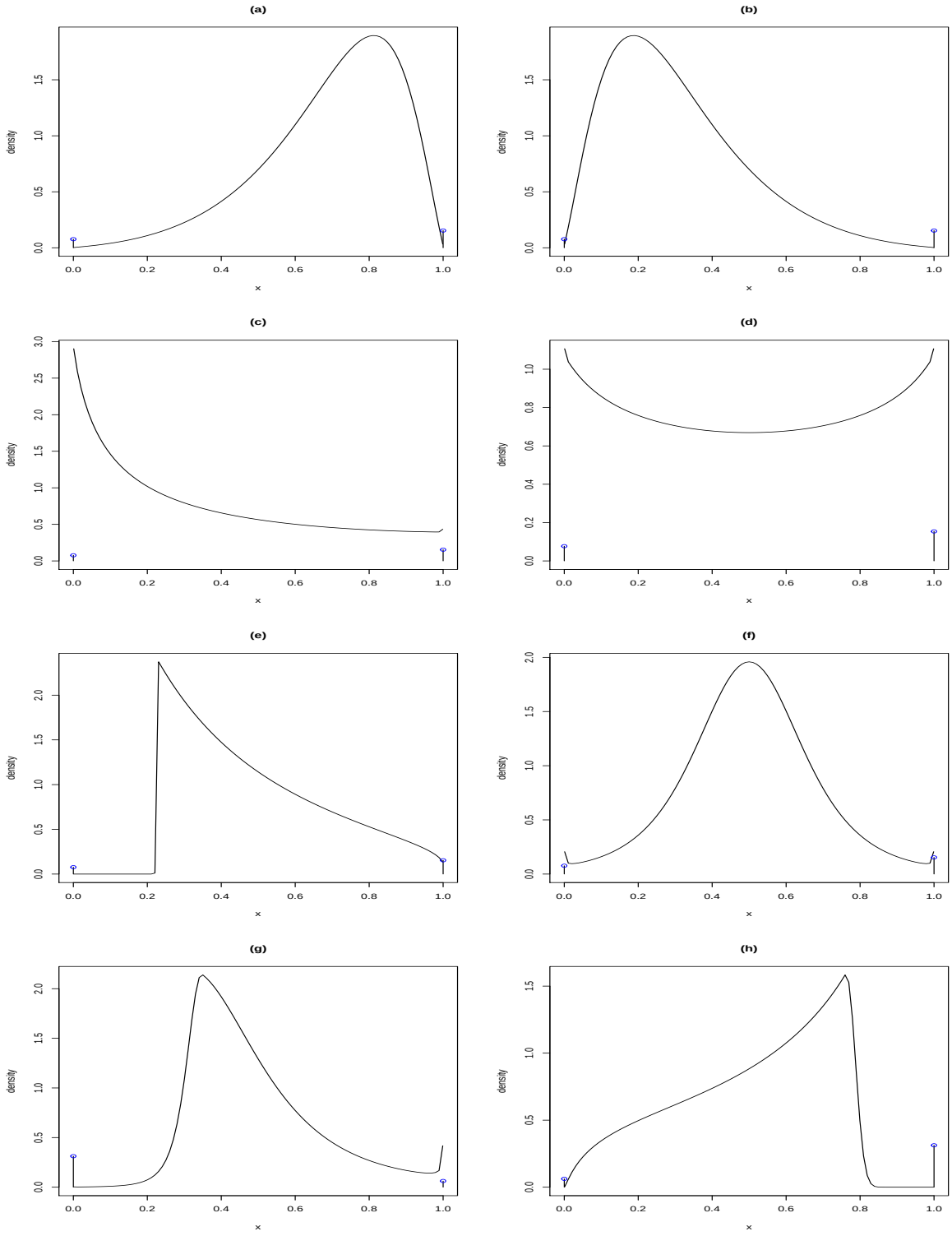


Fig. 7.2 A logit-SST distribution: (a) with values $\mu = 1$, $\sigma = 1$, $\nu = 1$, $\tau = 10$. $\xi_0 = .1$, and $\xi_1 = .2$ (b) with values $\mu = -1$ $\sigma = 2$, $\nu = 1$, $\tau = 10$. $\xi_0 = .1$, and $\xi_1 = .2$ (c) with values $\mu = -1$ $\sigma = 2$, $\nu = 1$, $\tau = 10$. $\xi_0 = .1$, and $\xi_1 = .2$ (d) with values $\mu = 0$ $\sigma = 2$, $\nu = 1$, $\tau = 10$. $\xi_0 = .1$, and $\xi_1 = .2$ (e) with values $\mu = 0$ $\sigma = 1$, $\nu = 2$, $\tau = 10$. $\xi_0 = .1$, and $\xi_1 = .2$ (f) with values $\mu = 0$ $\sigma = 1$, $\nu = 1$, $\tau = 3$. $\xi_0 = .1$, and $\xi_1 = .2$ (g) with values $\mu = 0$ $\sigma = 1$, $\nu = 2$, $\tau = 3$. $\xi_0 = .1$, and $\xi_1 = .2$ (h) with values $\mu = 0$ $\sigma = 1$, $\nu = 3$, $\tau = 3$. $\xi_0 = .1$, and $\xi_1 = .2$

The standard plotting functions of R can also be used to plot the created mixed distribution as is shown below. Figure 7.3 shows how the pdf, cdf, inverse cdf and randomisation functions can be displayed for different values of the distribution parameters.

```
# plotting the pdf ------------------------------
curve(dlogitSSTInf0to1(x, mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2),
      0.001,0.999, ylab="pdf",  main="(a)")
# getting the probabilities
p0 <- dlogitSSTInf0to1(x=0, mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2)
p1 <- dlogitSSTInf0to1(x=1, mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2)
points(c(0,1), c(p0,p1), col="blue")
lines(c(0,1), c(p0,p1), col="blue", type="h")
# plotting the cdf ------------------------------
curve(plogitSSTInf0to1(x, mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2),
      0.0001,0.999, ylim=c(0,1), ylab="cdf", main="(b)")
#points(c(0), c(p0), col="blue")
lines(c(0), c(p0), col="blue", type="h")
p1 <- plogitSSTInf0to1(q=.999, mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2)
lines(c(1,1),c(p1,1))
# plotting the inverse cdf ----------------------
curve(qlogitSSTInf0to1(x, mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2),
      0.0001,0.999, ylim=c(0,1), ylab="inverse cdf", main="(c)")
# plottind simulated data
truehist(rlogitSSTInf0to1(1000,mu=0, sigma=1, nu=.8, tau=10, xi0=.1, xi1=.2),
         main="(d)")
```

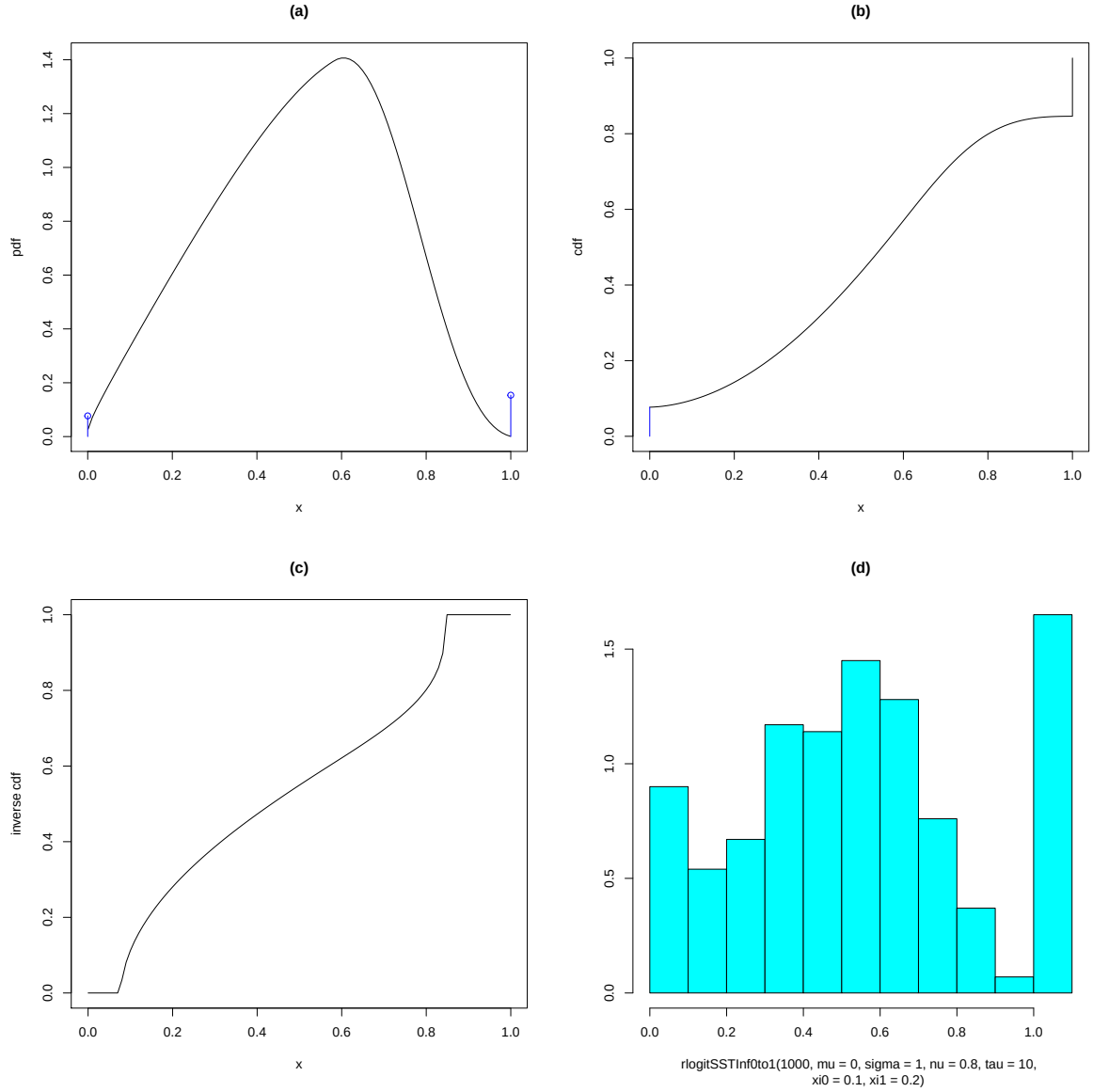


Fig. 7.3 The (a) pdf (b) cdf (c) inverse cdf and (d) simulated data from an inflated logitSST distribution with $\mu = 0$, $\sigma = 1$, $\nu = .8$, $\tau = 10$. $\xi_0 = .1$, and $\xi_1 = .2$

The next section demonstrates how to use the function `gamlssInf0to1()` to fit a model which has a response variable on the interval $[0, 1]$.

7.5 Fitting a distributions on $[0,1]$

7.5.1 The `gamlssInf0to1()` function

The main function for fitting a model with a response variable Y in the interval $[0,1]$ is `gamlssInf0to1()`. In an inflated distribution the parameters μ, σ, ν and τ are orthogonal to the parameters ξ_0 and ξ_1 in the sense that the log-likelihood function can be factorised in two components, one containing the μ, σ, ν and τ and another containing ξ_0 and ξ_1. This means that the two sets of parameters can be estimated separately. The function `gamlssInf0to1()` takes advantage of this separation and works as follows:

- It picks the argument `family` (see below) which defines a `gamlss.family` distribution defined on $(0,1)$
- Depending on whether the range of the response variable Y is $[0,1)$, $(0,1]$ or $[0,1]$, it creates an appropriate binary or multinomial response variable and it fits an appropriate GAMLSS model. For example
  - for $[0,1)$, $(0,1]$ it fits a binary logistic model (using the `gamlss.family` `BI`)
  - for $[0,1]$ it fits a multinomial model (using the `gamlss.family` `MN3`)
- Fits a GAMLSS model to the data cases with Y inside $(0,1)$ using the distribution defined by `family`, by weighting out the observations with zero and/or one.
- Creates the (randomized) quantile residuals for the whole model
- Saves the output as an `gamlssinf0to1` object which is a subclass of an `gamlss object`.

The idea is that the object `gamlssinf0to1` should behave similar to a `gamlss` object. For this purpose the following S3 methods are created.

1. `fitted.gamlssinf0to1()`,
2. `coef.gamlssinf0to1()`,

3. `print.gamlssinf0to1()`,
4. `deviance.gamlssinf0to1()`,
5. `vcov.gamlssinf0to1()`,
6. `summary.gamlssinf0to1()`,
7. `predict.gamlssinf0to1()`,
8. `formula.gamlssinf0to1`.

The above methods are demonstrated in the next sections.

The function `gamlssInf0to1()` has the following arguments:

y the proportion response variable (including values at zero and/or one)

mu.formula a model formula for the μ parameter

sigma.formula a model formula for the σ parameter

nu.formula a model formula for the ν parameter

tau.formula a model formula for the τ parameter

xi0.formula a model formula for the ξ_0 parameters which is related to the probability at zero

xi1.formula a model formula for the ξ_1 parameters which is related to the probability at one

data a data frame containing the variables occurring in the formula.

family any `gamlss()` distribution family defined on $(0,1)$

weights a vector of weights as in `gamlss()`

trace logical, if TRUE information on model estimation will be printed during the fitting

... for extra arguments which can be passed to `gamlss()`.

Since the individual models fitted within the algorithm used in `gamlssInf0to1()` are GAMLSS models, the parameter formulae above can take any linear or additive GAMLSS terms inclining smoothers and random effects.

To demonstrate the use of the `gamlssInf0to1()` function simulated examples are used below. In the examples there are no explanatory variables. That is, a response from different inflated distributions on $[0,1$, $(0,1]$ and $[0,1]$ is simulated and then a distribution is fitted to the response variable.

7.5.2 Simulating data

To compare the results obtained by the function `gamlssInf0to1()` to the ones obtained from standard `gamlss()`, simulate data from the inflated beta distributions `BEINF0`, `BEINF1`, `BEINF` which generate data on $[0,1)$, $(0,1]$ and $[0,1]$ respectively.

```
library(gamlss)     # loading gamlss package
library(gamlss.inf)
# creating  data
set.seed(324)
 y0 <- rBEINF0(1000, mu=.3, sigma=.3, nu=.15)# p0=0.13
 y1 <- rBEINF1(1000, mu=.3, sigma=.3, nu=.15)# p1=0.13
y01 <- rBEINF(1000, mu=.3, sigma=.3, nu=0.1, tau=0.2) # p0=0.769, p1=0.1538
```

The mixed continuous-discrete probability (density) function of $Y \sim BEINF(\mu,\sigma,\nu,\tau)$ is given by

$$f_Y(y) = \begin{cases} p_0 & \text{if } y = 0 \\ (1-p_0-p_1) f_W(y) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{7.5}$$

for $0 \leq y \leq 1$, where $W \sim BE(\mu,\sigma)$ has a beta distribution with $0 < \mu < 1$ and $0 < \sigma < 1$ and $p_0 = \nu/(1+\nu+\tau)$ and $p_1 = \tau/(1+\nu+\tau)$. Hence $\nu = p_0/p_2$ and $\tau = p_1/p_2$ where $p_2 = 1-p_0-p_1$. Since $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0+p_1 < 1$, hence $\nu > 0$ and $\tau > 0$.

Here $f_W(y)$ is given by

$$f_W(y) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

for $0 < y < 1$. where

$$\begin{pmatrix} \alpha \\ \beta \\ p_0 \\ p_1 \end{pmatrix} = \begin{pmatrix} \frac{\mu(1-\sigma^2)}{\sigma^2} \\ \frac{(1-\mu)(1-\sigma^2)}{\sigma^2} \\ \frac{\nu}{(1+\nu+\tau)} \\ \frac{\tau}{(1+\nu+\tau)} \end{pmatrix}$$

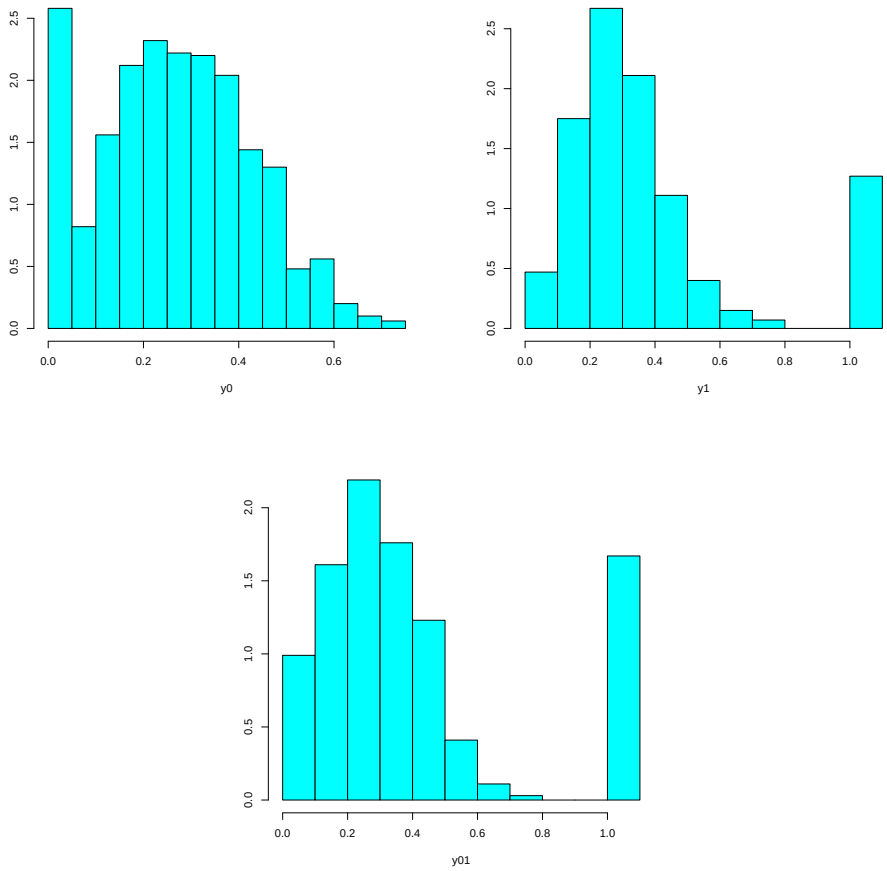
Hence

$$\begin{pmatrix} \mu \\ \sigma \\ \nu \\ \tau \end{pmatrix} = \begin{pmatrix} \alpha(1+\beta)^{-1} \\ (\alpha+\beta+1)^{-\frac{1}{2}} \\ \frac{p_0}{p_2} \\ \frac{p_1}{p_2} \end{pmatrix}$$

For $Y \sim BEINF0(\mu,\sigma,\nu)$ set $\tau = 0$ in the above probability (density) function (7.5). For $Y \sim BEINF1(\mu,\sigma,\nu)$ set $\nu = 0$ and the $\tau = \nu$ in the above probability (density) function (7.5).

All three simulated examples come from a beta distribution with $\mu = 0.3$ and $\sigma = 0.3$. For the distribution on $[0,1)$ the probability at zero is $p_0 = \frac{\nu}{1+\nu} = 0.15/(1+.15) = 0.1304348$. For the distribution on $(0,1]$ the probability at one is $p_1 = \frac{\nu}{1+\nu} = 0.15/(1+.15) = 0.1304348$. For the distribution on $[0,1]$ the probability at zero is $p_0 = \frac{\nu}{1+\nu+\tau} = 0.1/(1+0.1+0.2) = 0.0769231$ while the probability at one is $p_1 = \frac{\tau}{1+\nu+\tau} = 0.2/(1+0.1+0.2) = 0.1538462$. The proportions of zeros and ones in the sample are 0.123 for $[0,1)$, 0.127 for $(0,1]$ and $(0.07, 0.167)$ for $[0,1]$. Next plot the three data sets using `histdist()`.

```
library(MASS)
truehist(y0)
truehist(y1)
truehist(y01)
```



R code on page **??**

Fig. 7.4 Generated data using inflated beta distribution: with values $\mu = 0.3$, $\sigma = 0.3$, and $\nu = 0.15$ for the distribution on $[0,1)$, $\nu = 0.15$ for the distribution on $(0,1]$ and $\nu = 0.1$ and $\tau = 0.2$ for the distribution on $[0,1]$.

7.5.3 Fitting a distributions on $[0,1)$

Below an inflated distribution at 0 is fitted using both `gamlss()` and `gamlssInf0to1()` functions. Note that the `family` argument in `gamlssInf0to1()` takes a `gamlss.family` distribution defined on $(0,1)$. The `trace=TRUE` argument is used in `gamlssInf0to1()` to check the convergence of the two different models fitted, one using the `BI` family and the other using the `BE`.

```
g0 <- gamlss(y0~1, family=BEINF0)
```

```
## GAMLSS-RS iteration 1: Global Deviance = -239.6607
## GAMLSS-RS iteration 2: Global Deviance = -307.2383
```

```
## GAMLSS-RS iteration 3: Global Deviance = -307.6252
## GAMLSS-RS iteration 4: Global Deviance = -307.6258
```

```
library(gamlss.inf)
t0 <- gamlssInf0to1(y=y0, mu.formula=~1, family=BE, trace=TRUE)
```

```
## *****       The binomial model        *****
## GAMLSS-RS iteration 1: Global Deviance = 745.7199
## GAMLSS-RS iteration 2: Global Deviance = 745.7199
## ***** The continuous distribution model *****
## GAMLSS-RS iteration 1: Global Deviance = -985.3807
## GAMLSS-RS iteration 2: Global Deviance = -1052.958
## GAMLSS-RS iteration 3: Global Deviance = -1053.345
## GAMLSS-RS iteration 4: Global Deviance = -1053.346
##             The Final Global Deviance  = -307.6258
```

```
AIC(g0,t0, k=0)
```

```
##    df       AIC
## t0  3 -307.6258
## g0  3 -307.6258
```

Note that the global deviance of the fitted `t0` model, using `gamlssInf0to1()`, is obtained by adding the individual deviances from the binomial and the beta model. The third fitted parameter in both models, is related to the the probability at zero. The third parameter is called nu (i.e. ν) in `gamlss` but $xi\phi$ (i.e. ξ_0) in `gamlssInf0to1()`. The predictor η_3 for the third parameter is the same for both both model as shown below.

```
coef(g0, "nu")
```

```
## (Intercept)
##   -1.964323
```

```
coef(t0, "xi0")
```

```
## (Intercept)
##   -1.964323
```

The two fitted coefficients for predictor η_3 are identical, but the fitted values for ν and ξ_0 are not the same because the parameterizations used for the zero inflated distribution are different for gamlss using BEINF0 (see subsection 7.5.2) and for gamlssInf0to1 using BE (see section 7.3). Next only the first element of the fitted values vector is displayed (since all values are identical because we fit a constant model).

```
fitted(t0, "xi0")[1]
```

```
## [1] 0.123
```

```
fitted(g0, "nu")[1]
```

```
##         1
## 0.1402509
```

The differences in the fitted values is due to the way the two models are fitted. `gamlssInf0to1()` fits a binary distribution model with a `logit` link for the binomial distribution parameter. The vector `fitted(t0, "xi0")` contains the fitted probabilities at zero. For example let $\hat{\beta}_t = -1.964$ be the `coef(t0, "xi0")` then $1/(1+e^{-\hat{\beta}_t}) = \hat{\pi}_0 = 0.123$. In `gamlss`, ν is fitted using a log link. Let $\hat{\beta}_g = -1.964$ be the `coef(g0, "nu")` so $e^{\hat{\beta}_g} = \hat{\nu} = 0.1402509$ the fitted value for ν. In `BEINF0` ν is defined as the odds ratio for example $\hat{\nu} = \hat{\pi}_0/(1-\hat{\pi}_0)$ which implies that $\hat{\pi}_0 = \hat{\nu}/(1+\hat{\nu})$. This can be confirmed by:

```
fitted(g0, "nu")[1]/(1+fitted(g0, "nu"))[1]
```

```
##     1
## 0.123
```

which is the fitted probability of observing zero. The `summary()` function makes it clear that the two models use different link functions for the third parameters ν or ξ_0.

```
summary(t0)
```

```
## ******************************************************************
## Family:  "InfBE"
##
## Call:  gamlssInf0to1(y = y0, mu.formula = ~1, family = BE, trace = TRUE)
##
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  logit
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84294    0.02203  -38.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## -----------------------------------------------------------------
## Sigma link function:  logit
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84659    0.02989  -28.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----------------------------------------------------------------
## xi0 link function:  logit
## xi0 Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96432    0.09628   -20.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----------------------------------------------------------------
## No. of observations in the fit:  1000
## Degrees of Freedom for the fit:  3
##       Residual Deg. of Freedom:  997
##                       at cycle:
##
## Global Deviance:     -307.6258
##             AIC:     -301.6258
##             SBC:     -286.9026
## *****************************************************************

summary(g0)

## *****************************************************************
## Family:  c("BEINF0", "Beta Inflated zero")
##
## Call:  gamlss(formula = y0 ~ 1, family = BEINF0)
##
## Fitting method: RS()
##
## -----------------------------------------------------------------
## Mu link function:  logit
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84294    0.02203  -38.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## -----------------------------------------------------------------
## Sigma link function:  logit
```

```
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84659    0.02989  -28.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Nu link function:  log
## Nu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96432    0.09628   -20.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  1000
## Degrees of Freedom for the fit:  3
##       Residual Deg. of Freedom:  997
##                       at cycle:  4
##
## Global Deviance:     -307.6258
##             AIC:     -301.6258
##             SBC:     -286.9026
## ******************************************************************
```

The variance covariance matrix for the fitted `g0` and `t0` models can be obtained as follows.

```
vcov(t0)
```

```
##              (Intercept)  (Intercept) (Intercept)
## (Intercept) 0.0004854761 0.0001292291 0.000000000
## (Intercept) 0.0001292291 0.0008936886 0.000000000
## (Intercept) 0.0000000000 0.0000000000 0.009270331
```

```
vcov(g0)
```

```
##              (Intercept)  (Intercept) (Intercept)
## (Intercept) 0.0004854761 0.0001292291 0.000000000
## (Intercept) 0.0001292291 0.0008936886 0.000000000
## (Intercept) 0.0000000000 0.0000000000 0.009270331
```

Note that because of the partition of the likelihood function parameters μ and σ are orthogonal to ν or ξ_0.

The residuals for the two models should be identical for the not zeros response. Due to the randomization at discrete values (zero here) differences are expected when the response is zero. This is demonstrated in the lower part of Figure 7.5 where the residuals are plotted against the observation index.

```
plot(resid(t0), pch="+")
points(resid(g0), col="red")
```

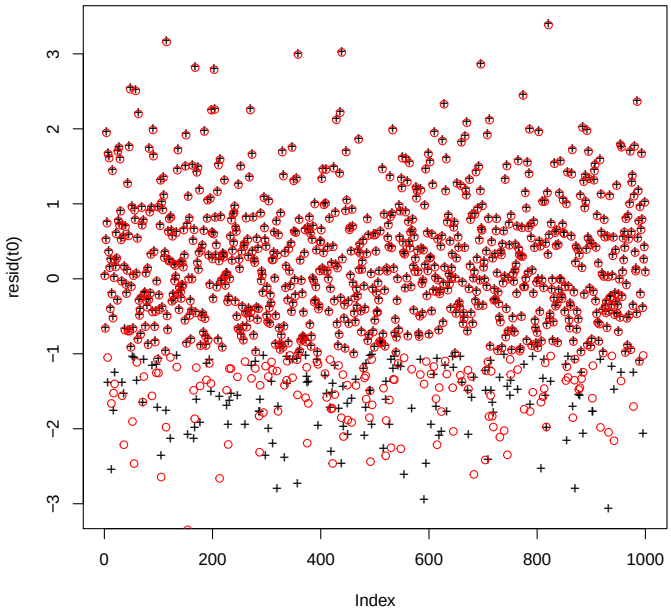


Fig. 7.5 Superimposed residuals from models `t0` and g0. Because of the randomization in the zero values of the response the lower part of the plot is not identical

Next we will plot the fitted distribution in Figure 7.6. The standard `BEINF0` distribution in **gamlss.dist** has its own plotting function called `plotBEINF0()` which can be used here. For the model fitted with `gamlssInf0to1()` such a function is created using the `gen.Inf0to1()` function.

```
# generate the
gen.Inf0to1("BE", type="Zero")

## A  0 inflated BE distribution has been generated
##  and saved under the names:
##  dBEInf0 pBEInf0 qBEInf0 rBEInf0
##  plotBEInf0
```

```
plotBEINF0(mu=fitted(g0, "mu")[1], sigma=fitted(g0, "sigma")[1],
           nu=fitted(g0, "nu")[1], main="(a)", ylab="density")
plotBEInf0(mu=fitted(t0, "mu")[1], sigma=fitted(t0, "sigma")[1],
           xi0=fitted(t0, "xi0")[1]) ; title("(b)")
```

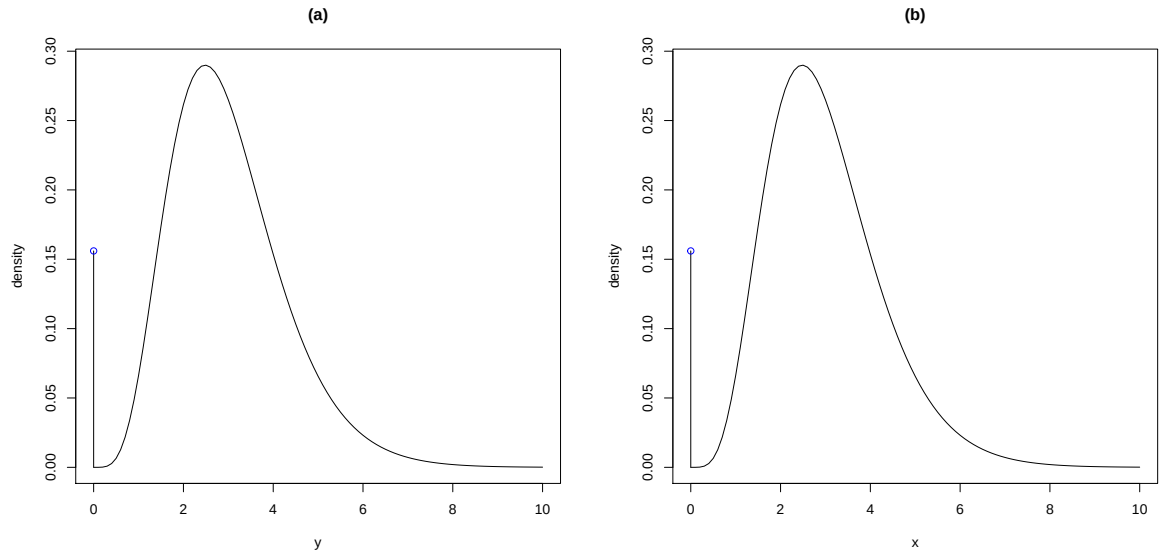


Fig. 7.6 The fitted distribution using (a) `gamlss()` and (b) `gamlssInf0to1()`

7.5.4 Fitting a distributions on $(0,1]$

Now the data inflated at 1 is analyzed.

```
 g1 <- gamlss(y1~1, family=BEINF1)

## GAMLSS-RS iteration 1: Global Deviance = -264.3046
## GAMLSS-RS iteration 2: Global Deviance = -343.1028
## GAMLSS-RS iteration 3: Global Deviance = -343.6308
## GAMLSS-RS iteration 4: Global Deviance = -343.6318

 t1 <- gamlssInf0to1(y=y1, mu.formula=~1, family=BE)
AIC(g1,t1, k=0)

##    df       AIC
## g1  3 -343.6318
## t1  3 -343.6318
```

The third fitted parameter in both models, is related to the probability at one. The third parameter is called nu (i.e. ν) in `gamlss` but xi1 (i.e. ξ_1) in `gamlssInf0to1()`.

```
coef(g1, "nu")
```

```
## (Intercept)
##   -1.927748
```

```
coef(t1, "xi1")
```

```
## (Intercept)
##   -1.927748
```

Again the two fitted coefficients are identical, but the fitted values for ν and ξ_1 are different.

```
fitted(t1, "xi1")[1]
```

```
## [1] 0.127
```

```
fitted(g1, "nu")[1]
```

```
##         1
## 0.1454754
```

The vector `fitted(t1, "xi1")` contains the fitted probabilities at one. For example let $\hat{\beta}_{t1}$ be the `coef(t1, "xi1")` then $1/(1+e^{-\hat{\beta}_{t1}}) = \hat{p}_1 = 0.127$. In `gamlss`, ν is fitted using the log link. Let $\hat{\beta}_{g1}$ be the `coef(g1, "nu")` so $e^{\hat{\beta}_{g1}} = \hat{\nu} = 0.1454754$ the fitted values for ν. In `BEINF1` ν is defined as the odds ratio for example $\hat{\nu} = \hat{p}_1/(1-\hat{p}_1)$ which implies that $\hat{p}_1 = \hat{\nu}/(1+\hat{\nu})$. This can be confirmed by:

```
fitted(g1, "nu")[1]/(1+fitted(g1, "nu"))[1]
```

```
##     1
## 0.127
```

which is the fitted probability of observing one. The `summary()` function makes it clear that the two models use different link functions for the third parameters ν or ξ_1.

```
summary(t1)
```

```
## ******************************************************************
## Family:  "InfBE"
##
## Call:  gamlssInf0to1(y = y1, mu.formula = ~1, family = BE)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
```

```
## Mu link function:  logit
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.88074    0.02172  -40.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  logit
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.88122    0.02987   -29.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## xi1 link function:  logit
## xi1 Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.92775    0.09497   -20.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  1000
## Degrees of Freedom for the fit:  3
##       Residual Deg. of Freedom:  997
##                       at cycle:
##
## Global Deviance:     -343.6318
##             AIC:     -337.6318
##             SBC:     -322.9085
## ******************************************************************

summary(g1)

## ******************************************************************
## Family:  c("BEINF1", "Beta Inflated one")
##
## Call:  gamlss(formula = y1 ~ 1, family = BEINF1)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  logit
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -0.88074    0.02172  -40.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  logit
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.88122    0.02987   -29.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Nu link function:  log
## Nu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.92775    0.09497   -20.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## No. of observations in the fit:  1000
## Degrees of Freedom for the fit:  3
##       Residual Deg. of Freedom:  997
##                       at cycle:  4
##
## Global Deviance:     -343.6318
##             AIC:     -337.6318
##             SBC:     -322.9085
## ******************************************************************
```

The variance covariance matrix for the fitted `g1` and `t1` models can be obtained as follows:

```
vcov(t1)
```

```
##              (Intercept)  (Intercept) (Intercept)
## (Intercept) 0.0004719565 0.0001297286  0.00000000
## (Intercept) 0.0001297286 0.0008923410  0.00000000
## (Intercept) 0.0000000000 0.0000000000  0.00901949
```

```
vcov(g1)
```

```
##              (Intercept)  (Intercept) (Intercept)
## (Intercept) 0.0004719565 0.0001297286  0.00000000
## (Intercept) 0.0001297286 0.0008923410  0.00000000
## (Intercept) 0.0000000000 0.0000000000  0.00901949
```

The fitted distributions can be plotted as follows:

```
# generate the
gen.Inf0to1("BE", type="One")

## A  1 inflated BE distribution has been generated
##  and saved under the names:
##  dBEInf1 pBEInf1 qBEInf1 rBEInf1
##  plotBEInf1

plotBEINF1(mu=fitted(g1, "mu")[1], sigma=fitted(g1, "sigma")[1],
           nu=fitted(g1, "nu")[1], main="(a)", ylab="density")
plotBEInf1(mu=fitted(t1, "mu")[1], sigma=fitted(t1, "sigma")[1],
           xi1=fitted(t1, "xi1")[1]) ; title("(b)")
```

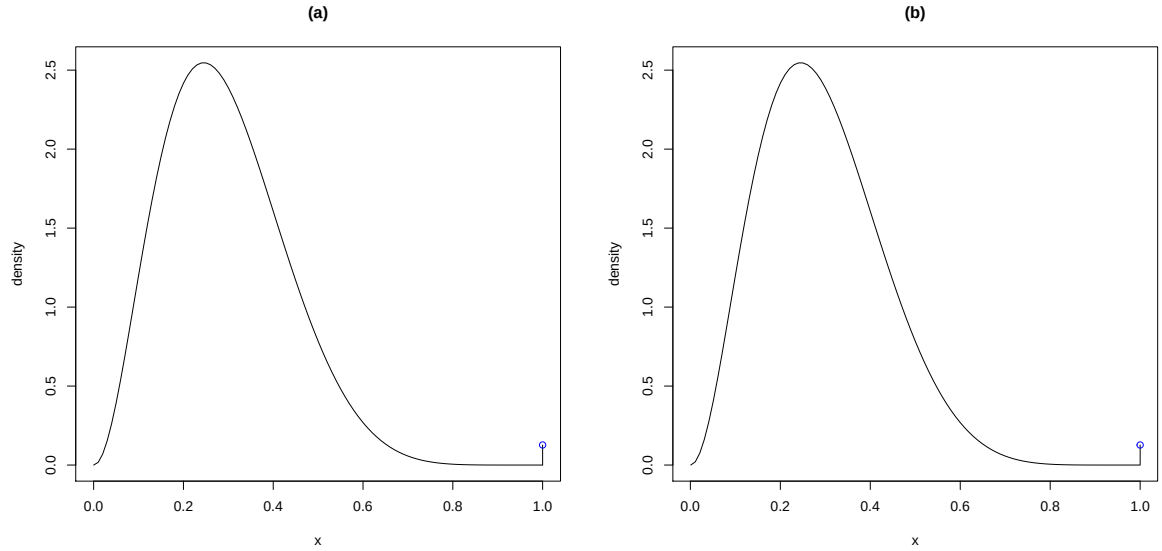


Fig. 7.7 The fitted distribution using (a) `gamlss()` and (b) `gamlssInf0to1()`

7.5.5 Fitting a distribution on $[0, 1]$

Now an inflated distribution on 0 and 1 is fitted using both `gamlss()` and `gamlssInf0to1()` functions.

```
 g01 <- gamlss(y01~1, family=BEINF)

## GAMLSS-RS iteration 1: Global Deviance = 471.2145
## GAMLSS-RS iteration 2: Global Deviance = 401.5136
## GAMLSS-RS iteration 3: Global Deviance = 401.1247
## GAMLSS-RS iteration 4: Global Deviance = 401.1241
```

```
 t01 <- gamlssInf0to1(y=y01, mu.formula=~1, family=BE)
AIC(g01,t01, k=0)
```

```
##     df      AIC
## g01  4 401.1241
## t01  4 401.1241
```

Note that in `gamlssInf0to()` it was not needed to specify that the distribution was on $[0,1]$ because the function detected whether there are zero and ones in the response and acts accordingly. The third and fourth fitted parameters in both models are related to the probability at zero and one. They are called ν and τ in `gamlss` but ξ_0 and ξ_1 in `gamlssInf0to1()`.

```
coef(g01, "nu")
```

```
## (Intercept)
##   -2.388763
```

```
coef(t01, "xi0")
```

```
## (Intercept)
##   -2.388747
```

```
coef(g01, "tau")
```

```
## (Intercept)
##   -1.519264
```

```
coef(t01, "xi1")
```

```
## (Intercept)
##   -1.519263
```

The fitted coefficients are (almost) identical for ν and ξ_0 and also for τ and ξ_1. Now look at the fitted values.

```
fitted(t01, "xi0")[1]
```

```
##          1
## 0.09174455
```

```
fitted(g01, "nu")[1]
```

```
##          1
## 0.09174314
```

```
fitted(t01, "xi1")[1]
```

```
##         1
```

```
## 0.2188732
```

```
fitted(g01, "tau")[1]
```

```
##         1
## 0.2188729
```

Note that contrary to the models with only zero or only one in the response variable, the models on $[0,1]$ use the same parametrization as `BEINF` so the fitted values are identical. In fact here the parameters are related to the probabilities at zero and ones as

$$\xi_0 = \nu = \frac{p_0}{(1-p_0-p_1)}$$

and

$$\xi_1 = \tau = \frac{p_1}{(1-p_0-p_1)}$$

so

$$p_0 = \frac{\xi_0}{(1+\xi_0+\xi_1)} = \frac{\nu}{(1+\nu+\tau)}$$

and

$$p_1 = \frac{\xi_1}{(1+\xi_0+\xi_1)} = \frac{\tau}{(1+\nu+\tau)}.$$

This can be verified by:

```
# probability for y=0
fitted(g01, "nu")[1]/(1+fitted(g01, "nu")+fitted(g01, "tau"))[1]
```

```
##          1
## 0.07000002
```

```
fitted(t01, "xi0")[1]/(1+fitted(t01, "xi0")+fitted(t01, "xi1"))[1]
```

```
##        1
## 0.070001
```

```
# probability for y=1
fitted(g01, "tau")[1]/(1+fitted(g01, "nu")+fitted(g01, "tau"))[1]
```

```
##     1
## 0.167
```

```
fitted(t01, "xi1")[1]/(1+fitted(t01, "xi0")+fitted(t01, "xi1"))[1]
```

```
##     1
## 0.167
```

The `summary()` produces the same results for both models, so only `t01` is represented here.

```
summary(t01)
```

```
## ******************************************************************
## Family:  "InfBE"
##
## Call:  gamlssInf0to1(y = y01, mu.formula = ~1, family = BE)
##
## Fitting method: RS()
##
## ------------------------------------------------------------------
## Mu link function:  logit
## Mu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.8484     0.0226  -37.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  logit
## Sigma Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.91148    0.03182  -28.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## xi0 link function:  log
## xi0 Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.3887     0.1249  -19.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## xi1 link function:  log
## xi1 Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.51926    0.08543  -17.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## ------------------------------------------------------------------
## No. of observations in the fit:  1000
## Degrees of Freedom for the fit:  4
##       Residual Deg. of Freedom:  996
##                       at cycle:
##
## Global Deviance:     401.1241
##             AIC:     409.1241
##             SBC:     428.7551
## ******************************************************************
```

The variance covariance matrix for the fitted `g01` and `t01` models is for all practical purposes identical.

```
vcov(t01)
```

```
##              (Intercept)  (Intercept) (Intercept) (Intercept)
## (Intercept) 0.0005107378 0.0001350223 0.000000000 0.000000000
## (Intercept) 0.0001350223 0.0010125764 0.000000000 0.000000000
## (Intercept) 0.0000000000 0.0000000000 0.015596125 0.001310618
## (Intercept) 0.0000000000 0.0000000000 0.001310618 0.007298641
```

```
vcov(g01)
```

```
##               (Intercept)   (Intercept)   (Intercept)   (Intercept)
## (Intercept)  5.107378e-04  1.350223e-04 -6.286964e-16 -3.501124e-15
## (Intercept)  1.350223e-04  1.012576e-03 -4.714802e-15 -2.625608e-14
## (Intercept) -6.286964e-16 -4.714802e-15  1.559632e-02  1.310616e-03
## (Intercept) -3.501124e-15 -2.625608e-14  1.310616e-03  7.298640e-03
```

Because of the randomization at zero and one, the residuals differ when the response variable is at those values, as is shown in Figure 7.8 where the residuals are plotted against the observation index.

```
plot(resid(t01), pch="+")
points(resid(g01), col="red")
```

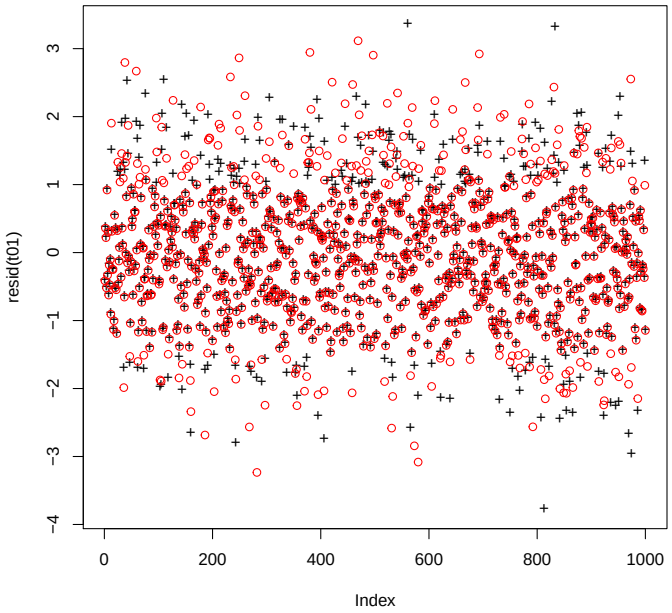


Fig. 7.8 Superimposed residuals from models `t01` and g01. Because of the randomization the values differ when the response variable is at zero and one

The fitted distributions are plotted next.

```
# generate
gen.Inf0to1("BE", type="Zero&One")

## A  0to1 inflated BE distribution has been generated
##  and saved under the names:
##  dBEInf0to1 pBEInf0to1 qBEInf0to1 rBEInf0to1
##  plotBEInf0to1

plotBEINF(mu=fitted(g01, "mu")[1], sigma=fitted(g01, "sigma")[1],
          nu=fitted(g01, "nu")[1], tau=fitted(g01, "tau")[1],
          main="(a)", ylab="density")
plotBEInf0to1(mu=fitted(t01, "mu")[1], sigma=fitted(t01, "sigma")[1],
           xi0=fitted(t01, "xi0")[1],  xi1=fitted(t01, "xi1")[1])
title("(b)")
```

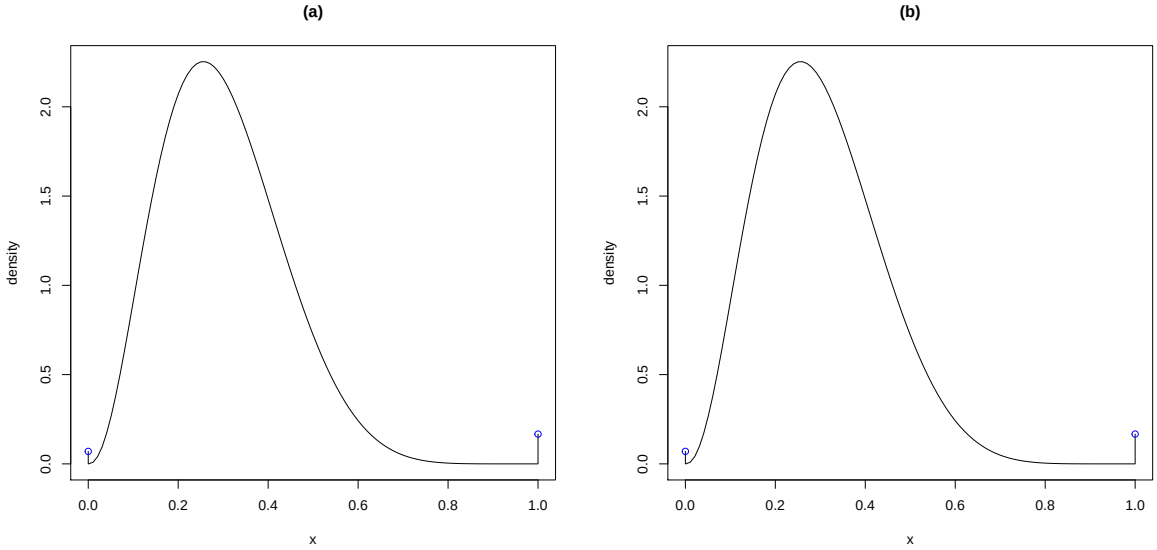


Fig. 7.9 The fitted distribution using (a) `gamlss()` and (b) `gamlssInf0to1()`

Chapter 8

Analysis of a proportion response variable on (0,1]

8.1 Introduction

This chapter follows the work of Hossain et al. (2016a). The purpose here is to provide flexible modelling approaches for centile curve estimation for a continuous proportion response variable measured on the interval from zero to one, i.e. intervals $(0,1)$, $[0,1)$, $(0,1]$ or $[0,1]$, where the square bracket indicates that the end point is included, while the curved bracket indicates that the end point is excluded. This chapter will focus on a response variable Y on $(0,1]$. Extensions are available for Y on $[0,1)$ and $[0,1]$ by following the approaches taken in chapters 9 and 10 respectively.

Specifically in this chapter two innovations are developed for modelling a proportion response variable Y on the interval $(0,1]$ including 1. The first is the development of a model employing a logit skew Student t (*logitSST*) distribution inflated at 1. The second innovation is the introduction of a generalized Tobit model [based on a flexible distribution on $(0,\infty)$, censored above 1], which allows modelling on the interval $(0,1]$. The inflated *logitSST* and generalized Tobit models can also be extended to model a proportion response variable on the intervals $[0,1)$

or $[0,1]$. The models are fitted to a lung function response variable using the `gamlss` packages (version 4.3.6) in *R*, Stasinopoulos and Rigby (2007) based on the `GAMLSS` model, (Rigby and Stasinopoulos, 2005).

Various alternative methods are used in the literature to model a proportion response variable, for example ordinary least squares (OLS) regression using a transformed response variable, e.g using the arcsine square root transformation, i.e. $sin^{-1}(\sqrt{Y})$, or the logit transformation, i.e. $\log[Y/(1-Y)]$. However OLS regression for modelling a proportion response variable has been questioned because of the potential mismatch of its underlying assumptions even after data transformation, Schmid et al. (2011). Warton and Hui (2011) argued that the logit transformation worked better than the arcsine transformation when analysing proportion data. Aitchison (1982) also proposed a logit transformation to a normal distribution to model compositional data in the form of proportions. The transformed normal distribution model also suffers from an interpretation problem, since the expected value of Y is not a simple transformation of the expected value of the logit transformed response. A logit transformation also used by Dawson et al. (2010) where they modified the LMS method to deal with the truncated response variable.

Given its relatively flexible nature, the beta distribution has been used widely in the statistical literature to model proportion data. For example Trenkler (1996), Kieschnick and McCullough (2003) and Ferrari and Cribari-Neto (2004) have shown the practical implementation of the beta distribution in their work. The beta distribution is a family of continuous distributions defined on the open interval $(0,1)$ not including 0 or 1. The probability density function (pdf) $f_Y(y)$ of the beta distribution was originally parameterised by two positive shape parameters α and β, i.e. $f_Y(y) = y^{\alpha-1}(1-y)^{\beta-1}/B(\alpha,\beta)$ for $0 < y < 1$, where $B(\alpha,\beta)$ is the beta function. The pdf of the beta distribution has different shapes: unimodal $(\alpha > 1, \beta > 1)$, uniantimodal or U shaped $(\alpha < 1, \beta < 1)$, increasing $(\alpha > 1, \beta \leq 1)$, decreasing $(\alpha \leq 1, \beta > 1)$ or constant $(\alpha = \beta = 1)$ depending on the values of α and β relative to 1.

However the beta distribution is limited to modelling data on the open interval $(0,1)$, not including 0 or 1. To model data with a significant number of zeros and/or ones a mixed

continuous-discrete distribution could be used. Kieschnick and McCullough (2003), Hoff (2007), Cook et al. (2008) and Ospina and Ferrari (2010) present empirical examples of the implementation of the beta inflated model as a mixed continuous-discrete distribution to model proportion data on the intervals $[0,1)$, $(0,1]$ or $[0,1]$. The beta inflated distribution comprises a beta distribution on $(0,1)$ together with the point probabilities at 0 and/or 1. Galvis et al. (2014) use a Bayesian approach to augment probabilities of zeroes and ones with the beta density for modelling proportion data. Hoff (2007) compares four different approaches for modelling $(0,1]$ data, i.e Tobit regression, OLS regression, the Papke-Wooldbridge (PW) model and the unit inflated beta model. Hoff's results suggest that the beta model performed worse than the other models. Ospina and Ferrari (2010) conclude that the complexity of interpretation of parameters and the assumed normality of the latent variable do not allow the Tobit model to be as flexible as the beta inflated distribution to model a response variable on the unit interval $[0,1]$. However Ospina and Ferrari (2010) do not claim that the beta inflated model always provides a better fit than the Tobit model. More recently Li et al. (2014) compare different models for a proportion response variable Y on [0,1] although their focus is on using the fitted mean of Y for prediction purposes rather than estimating centiles of Y.

The rest of the chapter proceeds as follows. Section 8.2 describes the statistical methodology implemented in this chapter. In particular, subsection 8.2.1 includes a brief description of centile estimation using the LMS method and its extensions. Subsection 8.2.2 provides a general model for centile estimation, while subsections 8.2.3 and 8.2.4 provide the logit skew Student t $(logitSST)$ and inflated $logitSST$ distributions, appropriate for centile estimation for a response variable Y on the intervals $(0,1)$ and $(0,1]$, respectively. Subsection 8.2.5 describes the generalized Tobit model for Y on $(0,1]$. Section 8.3 applies the methodologies proposed in section 8.2 to the lung function data. Conclusions are given in section 8.4.

8.2 Statistical methodology

8.2.1 LMS centile estimation method and extensions

The estimation of different centiles of a response variable at each level of one (or more) explanatory variables, is a major statistical problem in many applied human sciences, for example the WHO growth curves, WHO (2006, 2007). A statistical approach widely used for creating growth centile references for individuals from a population is the λ, μ and σ (LMS) method of Cole and Green (1992) and its extensions in Rigby and Stasinopoulos (2004, 2006). Note that subsequently we will use the notation ν rather than λ to refer to the third parameter of the model.

The main assumption of the LMS method, Cole and Green (1992), is that the response variable $Y > 0$ is defined by a transformation

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right] & \text{if } \nu \neq 0 \\ \frac{1}{\sigma}\log\left(\frac{Y}{\mu}\right) & \text{if } \nu = 0 \end{cases} \tag{8.1}$$

where Z is assumed to have truncated standard normal distribution. The truncated part comes from the fact that since $Y > 0$, Z has to satisfy the condition $-1/\sigma\nu < Z < \infty$ if $\nu > 0$ and $-\infty < Z < -1/\sigma\nu$ if $\nu < 0$. The LMS method uses the power transformation $(Y/\mu)^{\nu}$ to correct for skewness. The resulting distribution for Y (called the Box-Cox, Cole and Green, *BCCG*, distribution within the `gamlss` package, Stasinopoulos and Rigby (2007), in `R`, R Core Team (2014)) has three parameters, approximate median μ, approximate coefficient of variation σ and skewness parameter ν (where $\nu < 1$ and $\nu > 1$ correspond to positive and negative skewness, respectively).

However the *BCCG* distribution does not handle kurtosis. For modelling kurtosis Rigby and Stasinopoulos (2004, 2006) extended the *BCCG* distribution by introducing the four parameter *BCPE* and *BCT* distributions. For the *BCPE* distribution the transformed random variable Z

in (8.1) follows a truncated power exponential distribution, while for the *BCT* distribution Z follows a truncated t distribution.

8.2.2 General model for centile estimation

Let us assume that Y is the response variable, V is a single explanatory variable and $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ is a vector of p distribution parameters, then a general model for creating centiles for Y conditional on the value v of V is:

$$\begin{aligned} Y &\sim D(\theta) \\ g_k(\theta_k) &= s_k(x) \quad k = 1, \ldots, p \\ x &= v^{\xi} \end{aligned} \tag{8.2}$$

where D is the assumed distribution, g_k and s_k are link and smooth functions respectively for $k = 1, 2, ..., p$ and ξ is a power parameter applied to v to accommodate rapid growth of Y for low or high values of v. Here D represents any distribution. Letting D represent the *BCCG*, *BCPE* and *BCT* distributions gives respectively the LMS, LMSP and LMST methods of centile estimation, see Cole and Green (1992), Rigby and Stasinopoulos (2004) and Rigby and Stasinopoulos (2006) respectively. For example, the LMST method is given by

$$\begin{aligned} Y &\sim BCT(\mu, \sigma, \nu, \tau) \\ g_1(\mu) &= s_1(x) \\ g_2(\sigma) &= s_2(x) \\ g_3(\nu) &= s_3(x) \\ g_4(\tau) &= s_4(x) \\ x &= v^{\xi}. \end{aligned} \tag{8.3}$$

The default link functions for the *BCT* distribution in the `gamlss` package are $g_1(\mu) = \mu$ (identity link), $g_2(\sigma) = \log\sigma$, $g_3(\nu) = \nu$ and $g_4(\tau) = \log\tau$. Note that the BCCG distribution only has three parameters μ, σ and ν. [In the `gamlss` package, the notation *BCCGo*, *BCTo* and *BCPEo* refers to the *BCCG*, *BCT* and *BCPE* distributions respectively, except that the default link function for μ is $g_1(\mu) = \log\mu$.] The *BCCG*, *BCPE* and *BCT* distributions are suitable for modelling a response variable $Y > 0$. However they may not provide adequate models for Y on the unit interval $(0,1)$. Also they do not allow the value $Y = 0$. Next we investigate different distributions D appropriate for a response variable on $(0,1)$ and $(0,1]$. Extensions to response variables on $[0,1)$ and $[0,1]$ are available by following the approaches taken in chapter 9 and 10 respectively.

8.2.3 Logit skew student t distribution (logitSST)

The idea of the proposed model is to replace the beta distribution on (0,1) with any distribution on the range $(-\infty,\infty)$ transformed to the range $(0,1)$. Any distribution on the range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ by using an inverse logit transformation $Y = 1/(1+e^{-Z})$.

The reason for the proposed model is that the beta distribution often provides a poor fit to a proportion response variable on (0,1) in real data sets. The inverse logit transformation of the skew student t (SST) distribution, called the *logitSST* distribution, is introduced to provide an improved model on the interval $(0,1)$. Note that if $Z \sim SST(\mu,\sigma,\nu,\tau)$ for $-\infty < Z < \infty$, then $Y = 1/(1+e^{-Z}) \sim logitSST(\mu,\sigma,\nu,\tau)$ for $0 < Y < 1$. Details of the skew student t (SST) distribution are given in Wurtz et al. (2006), reparameterized from Fernández and Steel (1998*a*) . The *logitSST* distribution is created using the `gamlss` function `gen.Family()`, which allows any `gamlss` distribution with range $(-\infty,\infty)$, (e.g. SST), to be transformed to a new `gamlss` distribution, (e.g. logitSST), with range $(0,1)$.

8.2.4 LogitSST distribution inflated at 1

The *logitSST* distribution inflated at 1 is a mixture of two components: a discrete value 1 with probability p_1 and a $logitSST(\mu,\sigma,\nu,\tau)$ distribution on the unit interval $(0,1)$ with probability $(1-p_1)$. The resulting mixed continuous-discrete probability (density) function for $Y \sim logitSSTInf1(\mu,\sigma,\nu,\tau,p_1)$ is given by

$$f_Y(y|\mu,\sigma,\nu,\tau,p_1) = \begin{cases} (1-p_1)f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0<y<1 \\ p_1 & \text{if } y=1 \end{cases} \tag{8.4}$$

for $0 < y \leq 1$, where $W \sim logitSST(\mu,\sigma,\nu,\tau)$ has a *logitSST* distribution, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, $\tau > 0$ and $0 < p_1 < 1$, subsequently called the inflated logitSST distribution. The default link functions relate the parameters $(\mu,\sigma,\nu,\tau,p_1)$ to smooth functions of x, i.e.

$$\mu = s_1(x)$$

$$\log\sigma = s_2(x)$$

$$\log\nu = s_3(x)$$

$$\log\tau = s_4(x)$$

$$\log\left(\frac{p_1}{1-p_1}\right) = s_5(x).$$

The inflated logitSST distribution defined by (8.4) can be fitted by fitting two models: a *logitSST* (μ,σ,ν,τ) distribution model for $0 < Y < 1$, together with a binary model for recoded variable Y_1 given by

$$Y_1 = \begin{cases} 0 & \text{if } 0<Y<1 \\ 1 & \text{if } Y=1 \end{cases}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} (1-p_1) & \text{if } y_1 = 0 \\ p_1 & \text{if } y_1 = 1 \end{cases}$$

Alternatively the inflated logitSST distribution (8.4) can be fitted using the new package in R called `gamlss.inf` described in chapter 7. The inflated logit distributions (e.g. $logitSSTInf1$) have the advantage of extra flexibility, in that the probability of Y equal 1 is modelled independently of the distribution on (0,1), (e.g. $logitSST$), but at the cost of introducing an extra parameter (p_1) into the model. Note that the logit transformation is sensitive to values of Y very close to 0 and 1.

8.2.5 Generalized Tobit model

The original Tobit model of a response variable Y on $[0,1]$ assumes that the response follows a normal distribution censored below 0 and above 1, Tobin (1958).

Here we assume the response variable Y is recorded on $(0,1]$. The generalised Tobit model on (0,1] requires data censoring above 1 of a flexible model response variable distribution on $(0,\infty)$ for its positive probability at 1. Censoring refers to the transformation of observations outside the limiting interval to the border value, Hoff (2007). Here the values of Y in the model distribution above 1 are transformed to 1.

Let $V \sim D(\mu,\sigma,\nu,\tau)$ be a flexible uncensored distribution on $(0,\infty)$. Let $Y \sim Drc(\mu,\sigma,\nu,\tau)$ be the corresponding right censored distribution on $(0,1]$, i,e censored above 1. Then

$$Y = \begin{cases} V & \text{if} \quad 0 < V < 1 \\ 1 & \text{if} \quad V \geq 1 \end{cases}$$

Hence the mixed continuous-discrete probability (density) function of Y is given by

$$f_Y(y) = \begin{cases} f_V(y) & \text{if} \quad 0 < y < 1 \\ P(V \geq 1) & \text{if} \quad y = 1 \end{cases} \tag{8.5}$$

for $0 < y \leq 1$. Note that Y is treated as having a proper mixed continuous- discrete (generalised Tobit) distribution D_{rc} on (0,1] with a point probability at 1. In principle D can be any distribution on $(0, \infty)$. In the analysis in section 8.3 we use the three parameter *BCCGo* distribution for D with its default link functions. In the generalised Tobit model the probability of Y equal 1 is directly related to the distribution between 0 and 1 and so is less flexible, but the model is more concise (i.e. parsimonious) in that it has one less distribution parameter than if $P(Y = 1)$ were modelled independently. Also the generalised Tobit model is not sensitive to values of Y very close to 1.

8.3 Data Analysis

8.3.1 Data and fitted models

Here 3164 male observations of lung function, previously analysed by Stanojevic et al. (2009), are modelled. Lung data set has been privately collected from Stanojevic et al. (2009).The response variable is $Y = FEV_1/FVC$ and the explanatory variable is $x = log(height)$. The response variable Y is a ratio of forced expiratory volume in 1 second (FEV_1) to forced vital capacity (FVC). Spirometric lung function Y is an established index for diagnosing airway obstruction, e.g. Quanjer et al. (2010). Figure (8.1) shows the histogram and box plot of response variable Y.

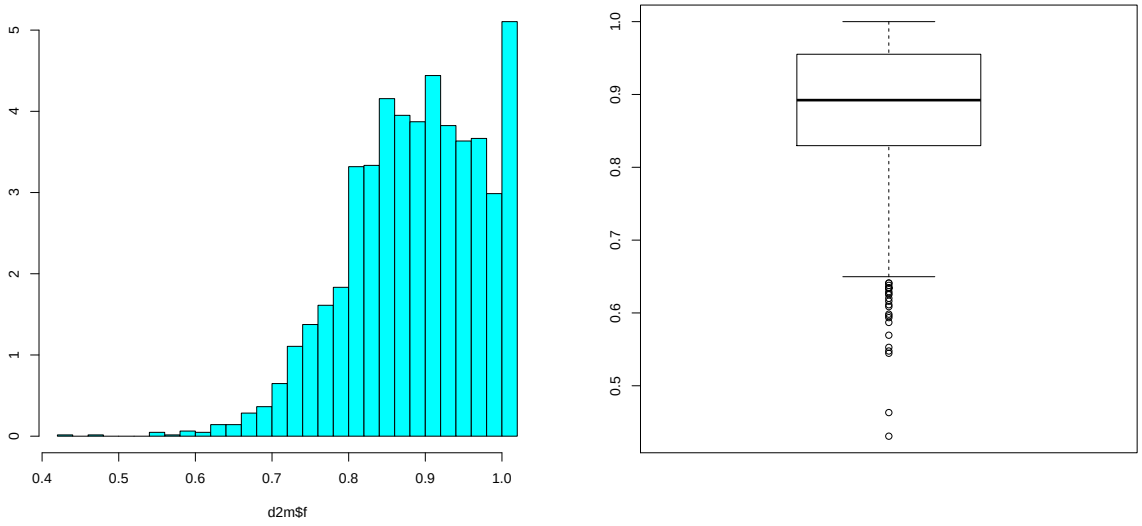


Fig. 8.1 Frequency histogram and boxplot of observed variable Y ($Y = FEV_1/FVC$)

Figure 8.2 shows a scatter plot of the response variable against height with marginal histogram. The marginal histogram of the ratio depicts that the response lies on the interval (0,1) including 1 and shows an excess number of ones.

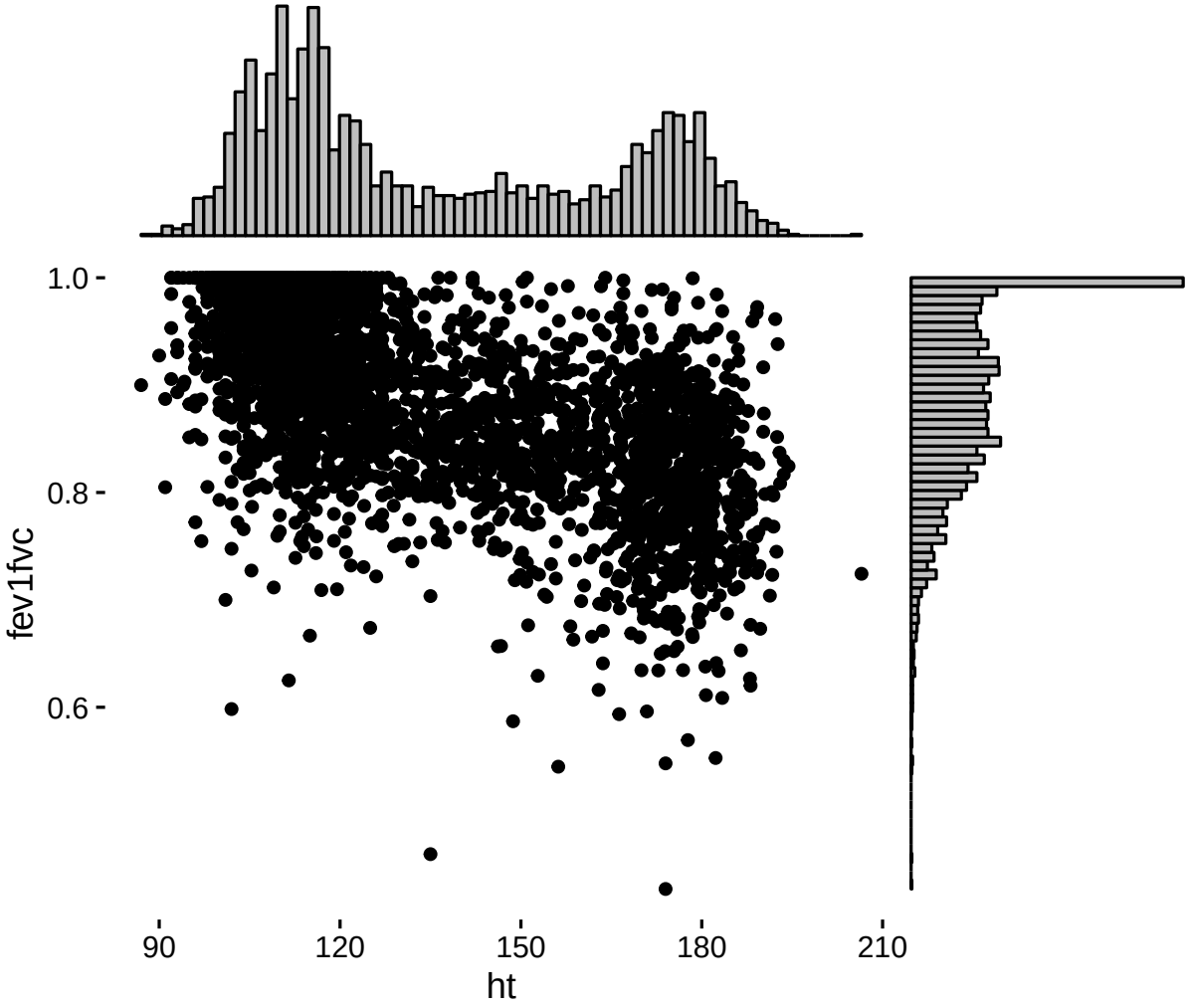


Fig. 8.2 Scatter plot with marginal histogram of observed variable Y ($Y = FEV_1/FVC$) against height

A log transformation of height is used following Cole et al. (2009) and Quanjer et al. (2012). The resulting centiles transform back height variable automatically by centiles plot function `centiles()` to create centiles in the Figure 8.3. Centile curves for Y against $x = \log(height)$ are achieved by using five methods: LMS (*BCCGo* with $x = \log(height)$) so no power parameter ξ was estimated in equation (8.2), *BEINF1* (beta inflated at 1), the original Tobit model and two new methods proposed in this chapter, the generalised Tobit model (*BCCGorc*, i.e, *BCCGo* right censored at 1) and the *logitSSTinf*1 (*logitSST* inflated at 1) models. The methods were applied using the `gamlss` package version $4.3-2$ in R, Stasinopoulos and Rigby (2007). For the inflated *logitSST* model a new package was developed called `gamlss.inf` described in chapter 7. The smoothing method P-splines, Eilers and Marx (1996), was used for fitting each smooth function

$s_k(x)$ in each of the 5 models. The P-splines method is a combination of B-splines regression and quadratic penalties imposed on the estimated coefficients. The degrees of freedom used for smoothing was estimated locally using a local generalised Akaike Information Criterion, Akaike (1983), with penalty $k = 6$ for each degree of freedom in the smooth function.

The penalty $k = 6$ was chosen as a compromise between a low value of k (eg. $k = 2$ for the Akaike information criterion, AIC) which can lead to overfitting (i.e. undersmoothing) resulting in erratic fitted centile curves, and a high value of k (eg. $k = \log(n) = 8.06$ for the Schwartz Bayesian criterion, SBC) which can lead to underfitting (i.e. oversmoothing) resulting in biased centile curves, leading to significant Z and Q residual test statistics. The R commands used in the analysis are given in Appendix A.

Table 8.1 Comparison of fitted models

Method	Parameters	df	Deviance	AIC	GAIC (k=6)	SBC
BEINF1	3	6.0	210	222	246	258
Tobit	2	7.0	118	132	160	175
GenTobit	3	9.0	29	47	82	101
logitSSTInf1	5	14.3	0	29	86	115

Table 8.1 summarises the number of distribution parameters and the total degrees of freedom (df) used for four of the models. Also given are the values of the global deviance (Deviance), Akaike information criteron (AIC), Generalised AIC (GAIC) with penalty $k = 6$, and Schwarz Bayesian criteron (SBC) for the four fitted models (where 6390.7 was added to all the values to make the comparison of values clearer). The LMS (*BCCGo*) model could not be included in the comparison in Table 8.1 because it does not have a point probability at $Y = 1$. From Table 8.1, the inflated *logitSST* model is best as judged by AIC (as it has the lowest value), while the generalised Tobit model is best as judged by SBC. Using criterion GAIC with $k = 6$, the inflated *logitSST* and generalised Tobit models are almost equally good. The *BEINF*1 and standard Tobit models perform much worse.

8.3.2 Centile estimation

Figure 8.3 shows centile curves constructed using four different fitted models: LMS (*BCCGo*), *BEINF*1, inflated *logitSST* and generalised Tobit (*BCCGo*, right censored at 1). The fitted $(2, 10, 25, 50, 75, 90, 98)\%$ centile curves show that the BEINF1 and generalised tobit models constructed the most smooth curves, while the LMS model constructed the least smooth curves of the four models. Note that, unlike the other three models, the upper centile curves for the LMS model can reach above 1, since for the LMS model the response variable is not bounded by 1.

Table 8.2 Comparison of fitted centile percentages

Nominal Centile %	LMS	BEINF1	logitSSTInf1	GenTobit
2	1.74	1.33	1.93	2.02
10	9.96	8.31	10.02	9.26
25	27.12	25.16	24.59	24.62

Table 8.2 shows the sample percentages at or below the $(2, 10, 25)\%$ centile curve for the fitted models against the nominal percentiles. [Percentages above 25% were not given since for at least one model these centile curves reach 1 and hence the sample percentage *at* or below the centile curve provides a distorted overestimate of the correspondent model centile curve percentage.] Among the four models given, the inflated *logitSST* model generally performs best. The beta inflated model performs much worse than the other models because its sample percentages below each centile curve are far from the nominal centile percentages.

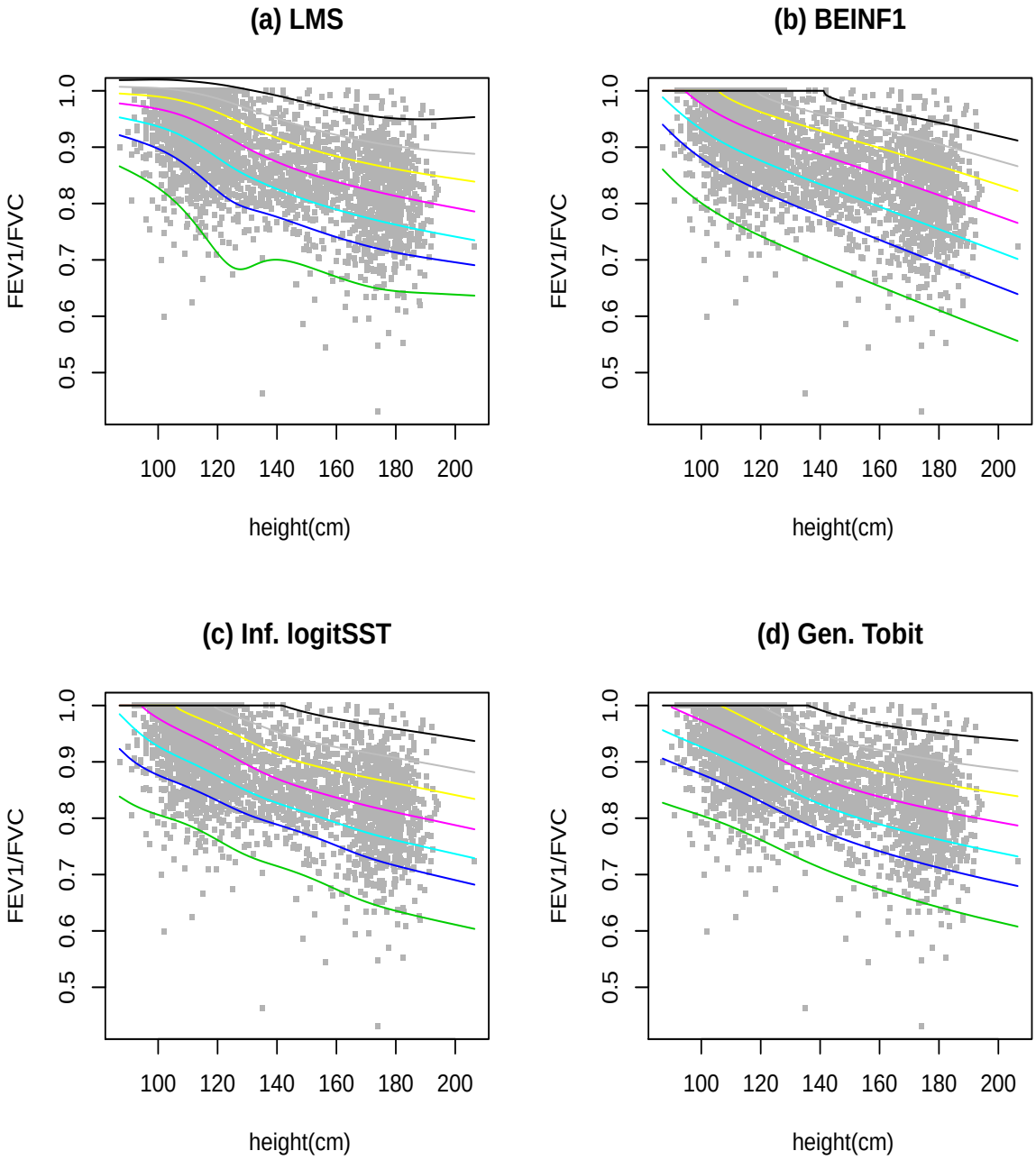


Fig. 8.3 Centile curves for model a) LMS b) BEINF1 c) logitSSTInf1 d) Generalized Tobit

8.3.3 Data analysis using two explanatory variables

The total number of subjects included in the study is 3164. This analysis finds centiles of a response variable dependent on two explanatory variables. The response variable is forced

expiratory volume in 1 second to forced vital capacity and the explanatory variables are height and age. Table 8.3 of descriptive statistics includes the number of observations (N), mean, standard deviation and minimum and maximum values of variables fev1/fvc, age and height.

Table 8.3 Lung data

Statistic	N	Mean	St.Dev.	Min	Max
fev1fvc	3,164	0.885	0.085	0.431	1.000
age	3,164	13.557	15.239	2.500	80.000
ht	3,164	135.952	28.506	87.000	206.500

Figure 8.4 shows the scatter plot matrix of lung data with histogram, kernel density and absolute correlation of variables fev1/fvc, age and hight.

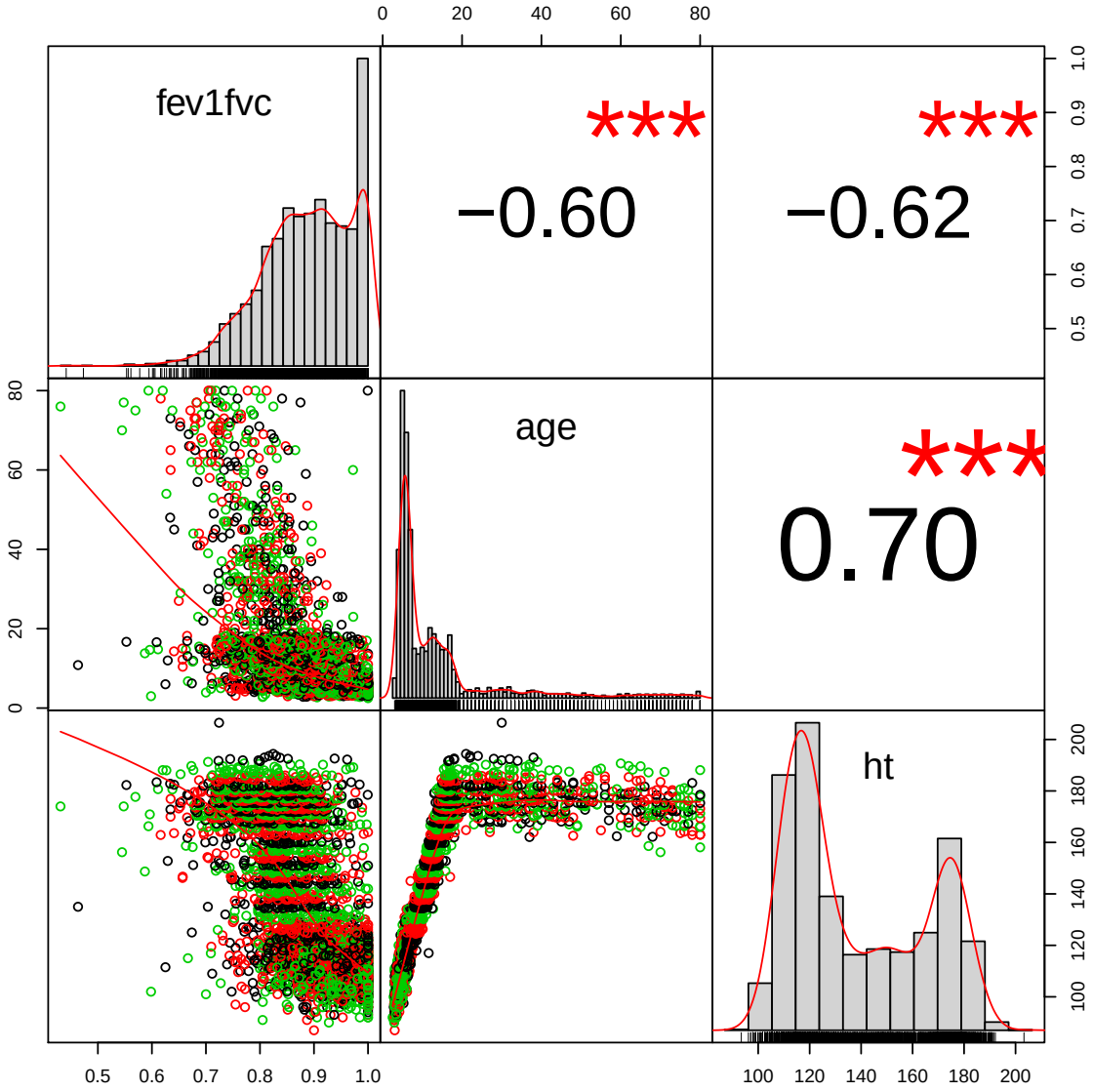


Fig. 8.4 Summary of Lung data.

The subjects age range is between 2.5 to 80 years; 59.73 % were aged 2.5 to 10 years and 25.22 % were aged between 11 to 20 and only 15.4% were aged above 20 years. Box plot in Figure 8.5 shows the ratio of FEV1/FVC against different age ranges.

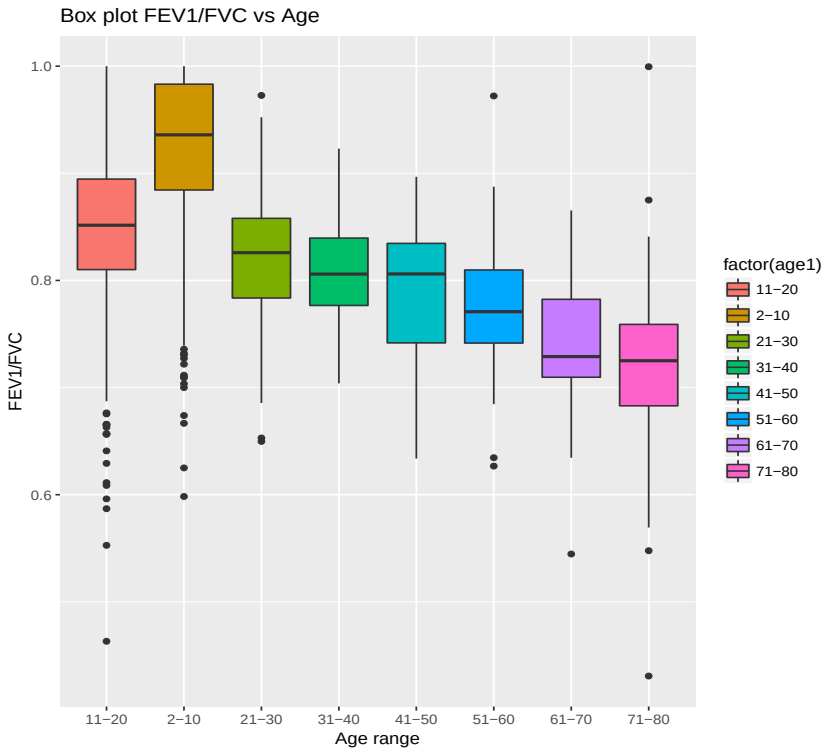


Fig. 8.5 Box plot of FEV1/FVC against age range

Another explanatory variable height of subjects lies between 80 cm to 206.5 cm. 41.18% of the subjects are 101cm to 120cm tall, 55.27% subjects have height between 121cm to 195cm and only one subject is 206.5cm tall. Box plot in Figure 8.6 shows the ratio of FEV1/FVC against different age ranges.

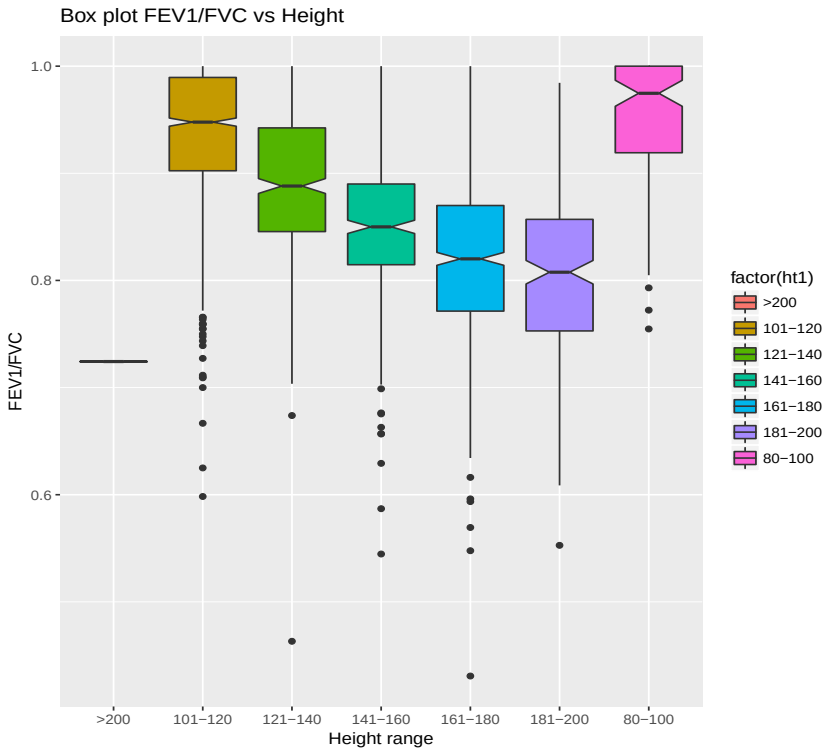


Fig. 8.6 Box plot of FEV1/FVC against height range

Scatter plot in Figure 8.7 shows the distribution of spirometric index (FEV1/FVC) against two quantitative explanatory variables age and height.

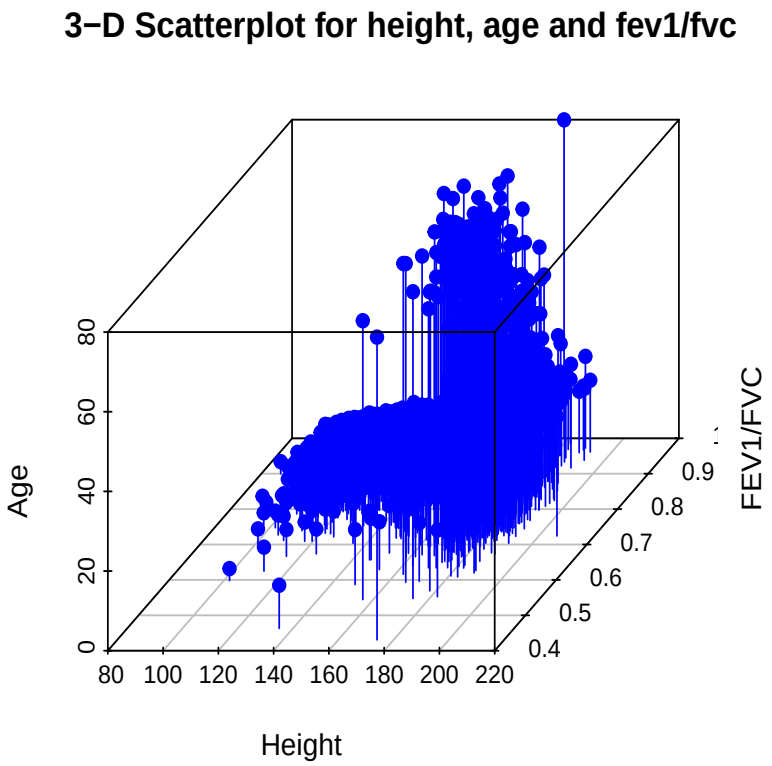


Fig. 8.7 Scatter plot of FEV1/FVC against height and age

It is generally accepted that the spirometric index (FEV1/FVC) decreases from childhood to old age (Quanjer et al., 1989). Table 8.4 shows the numbers of subjects by age group with data on the ratio of forced expiratory volume in 1 second(FEV1) to forced vital capacity (FEV1/FVC).

Table 8.4 Number of subjects with age and height group with FEV1/FVC

Age (years)	FEV1/FVC	Height(cm)	FEV1/FVC
$2-10$	0.93	$80-100$	0.95
$11-20$	0.85	$101-120$	0.94
$21-30$	0.82	$121-140$	0.89
$31-40$	0.81	$141-160$	0.85
$41-50$	0.79	$161-180$	0.82
$51-60$	0.78	$181-200$	0.81
$61-70$	0.74	>200	0.72
$71-80$	0.72		

`stepGAICAll.A()` function is then used to search for a suitable model for FEV1/FVC using the inflated logit skew student t distribution at 1. A local GAIC with $k=6$ is used to choose the effective degrees of freedom for smoothing in the smoothing function *pb*. A global penalty $k=6$ also used to select terms in the `stepGAICAll.A()` procedure. The reason for using $k=6$ is to compromise between higher value (i.e. $k=\log(n)$) and a lower value ($k=2$) value of k. A lower value of k (i.e.$k=2$) would result less smooth centiles but a better fit to the data, while a higher value of k would result in even smoother centiles but a worse fit to the data. Table 8.5 shows the chosen models for the parameters.

Table 8.5 Chosen model for the parameters

Parameters	chosen models
μ	$ga(\sim s(\log(height), \log(age)))$
σ	$pb(\log(age), method = "GAIC", k = 6)$
ν	$pb(\log(age), method = "GAIC", k = 6)$
τ	$pb(\log(age), method = "GAIC", k = 6)$
ξ_1	$pb(\log(age), method = "GAIC", k = 6)$
$+$	$pb(\log(height), method = "GAIC", k = 6)$

Figure 8.8 shows the residuals of the fitted model against fitted value and index along with Q-Q plot and density plot of the quantile residuals.

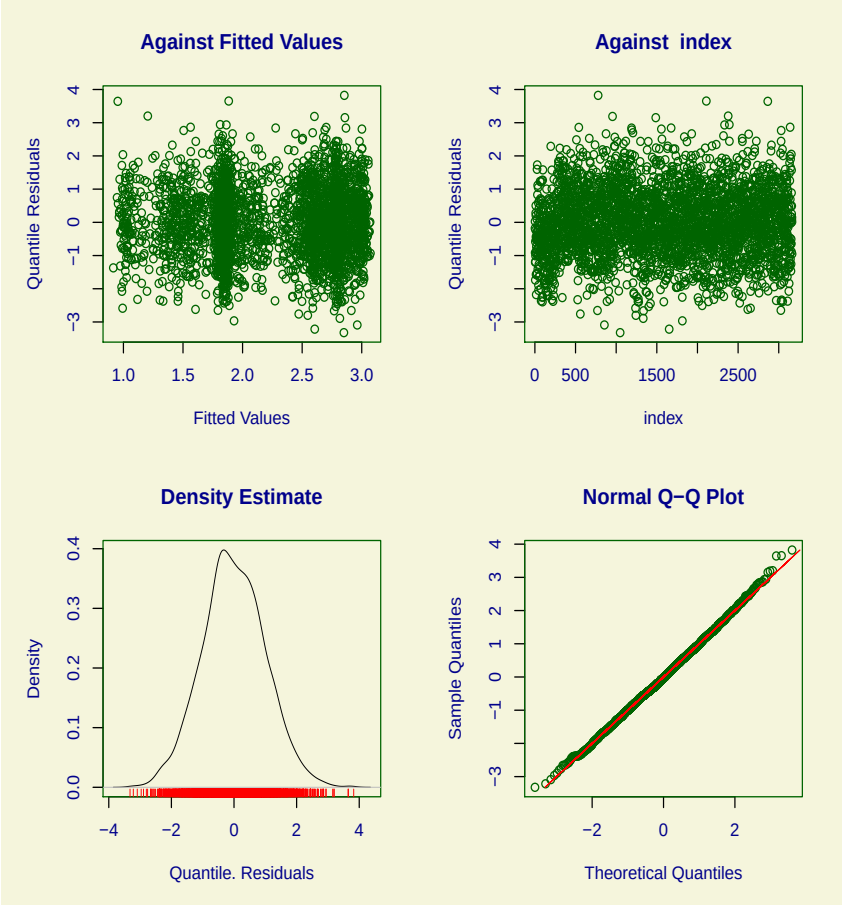


Fig. 8.8 Residual plot for fitted model

Now a model is fit for the height against age. The purpose of this is to find lower and upper centile limit (i.e. 1% and 99.9%) of height against age. Figure 8.9 shows the centiles (0.1, 0.4, 2, 10, 25, 50, 75, 90, 98, 99.6, 99.9)% for height against age for a fitted model (height against age).

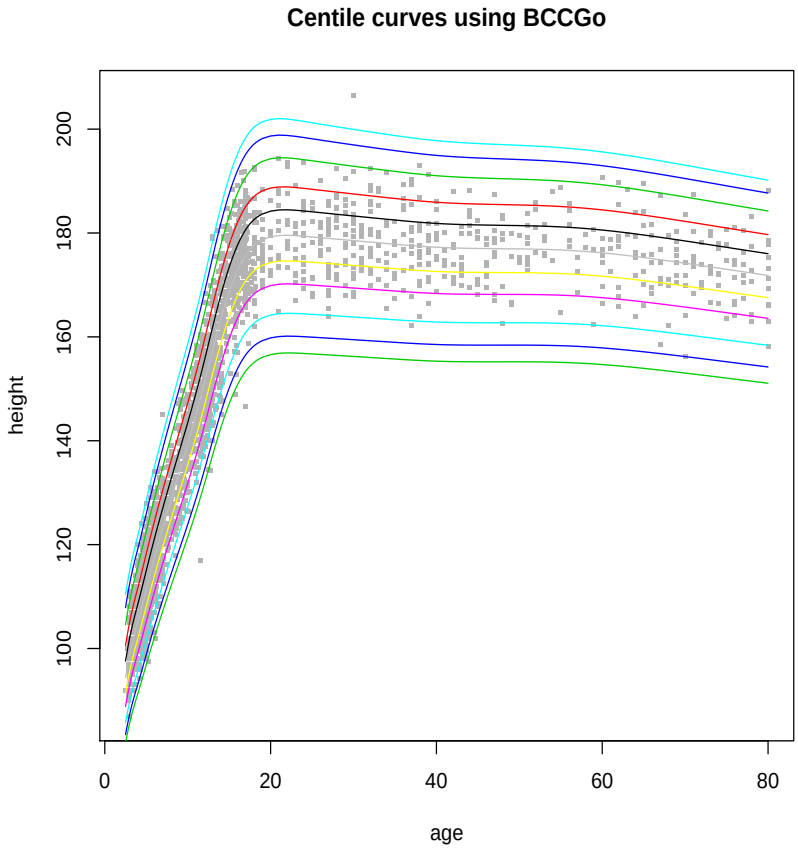


Fig. 8.9 Centiles for height against age

Contour plot of the 5th centile of FEV1/FVC against height and age constructed in figure 8.10. Figure 8.10 shows that 5% centile value of FEV1/FVC is higher for younger and shorter people and lower for older and taller people.

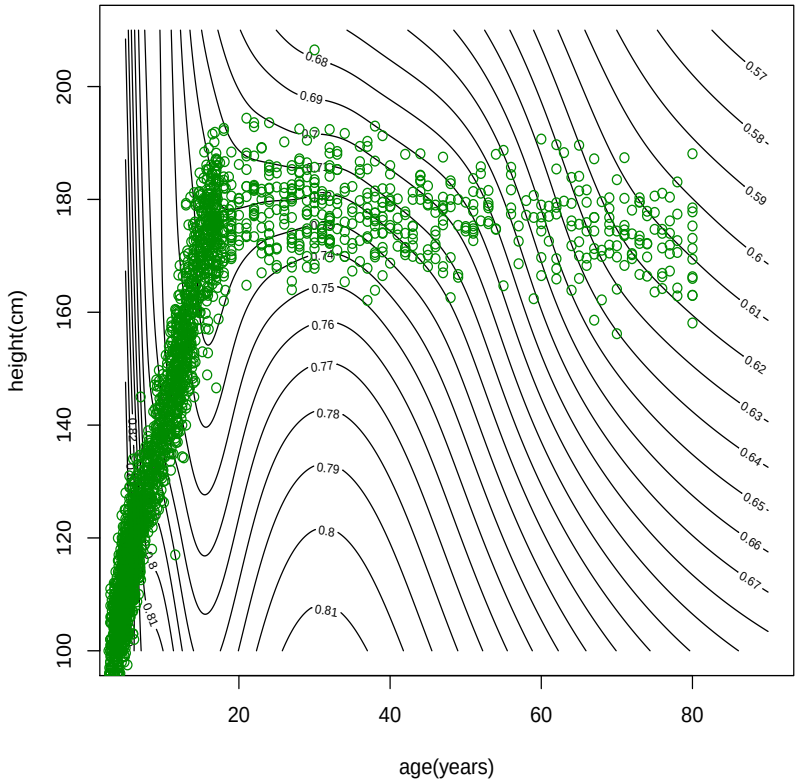


Fig. 8.10 Contour plot of the 5th centile of FEV1/FVC

8.3.4 Model checking using residual based diagnostics

The residuals used in GAMLSS are normalized (randomized) quantile residuals, Dunn and Smyth (1996) also called z-scores. In this paper two residual based diagnostic tools, the worm plot and Z and Q statistics, are used to check the adequacy of each model.

8.3.4.1 Worm Plots

van Buuren and Fredriks (2001) introduced the worm plot, which consists of de-trended Q-Q residual plots. The explanatory variable is split into (non-overlapping) intervals (with equal numbers of observations) and a detrended Q-Q plot of the residuals is obtained for residuals in each interval. The shape of the worm plot indicates how the observed response variable

distribution differs from assumed underlying model distribution within each interval of the explanatory variable. In this spirometry data example we need to check whether the model fits well within different intervals of height.

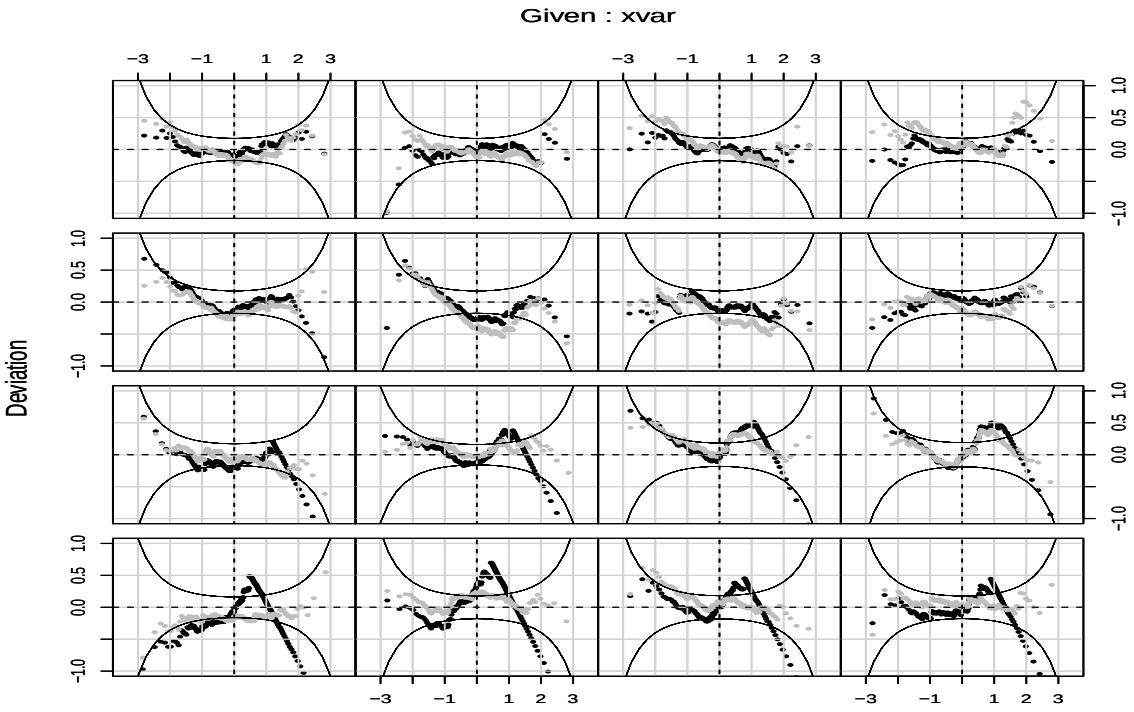


Fig. 8.11 Twin worm plot for LMS (dark points) and BEINF1 (light points) models.

In Figure 8.11 worm plots of the LMS and *BEINF*1 models are shown in 16 intervals of height with equal number of observations. The 16 height intervals are given in Figure 8.12 and correspond to the 16 worm plots from the bottom left plot to the top right plot in rows in Figures for generalized Tobit model (BCCGrc) and logitSSTInf1. For an adequate model, 95% of the points in each plot should lie between the elliptical 95% pointwise confidence band curves. For interpretation of the worm plot see van Buuren and Fredriks (2001) and Stasinopoulos and Rigby (2007). The shapes of the worm plots can indicate the type of model failure, see Rigby and Stasinopoulos (2004). Different shapes of the worm plot, i.e. a vertical shift, a slope, a parabola or a S shape, indicate differences (i.e. misfits) in the mean, variance, skewness and excess kurtosis of the residuals respectively from their assumed values 0,1,0 and 0 for a standard normal distribution.

A worm plot with vertical shift above (or below) the horizontal origin line indicates that the fitted model location is too low (or too high) respectively (within the corresponding interval of height). A worm plot with a positive (or negative) slope indicates that the fitted model scale is too low (or too high) respectively. A worm plot with a parabola U-shape (inverted U-shape) indicates that the fitted model distribution skewness is too low (or too high) i.e. too left skewed (or too right skewed) respectively. A worm plot with an S-shape with left bend down (or left bend up) indicates that the fitted model distribution kurtosis is too low (or too high) so the tails are too light (or too heavy) respectively.

Inadequacies in the fitted model described above may be reduced (or eliminated) by increasing the flexibility of the fitted model e.g., by reducing the local penalty k (for each smoothing curve degree of freedom used in the model). The worm plots in Figure 8.11 show that the LMS and beta inflated ($BEINF1$) models fit badly in most height intervals.

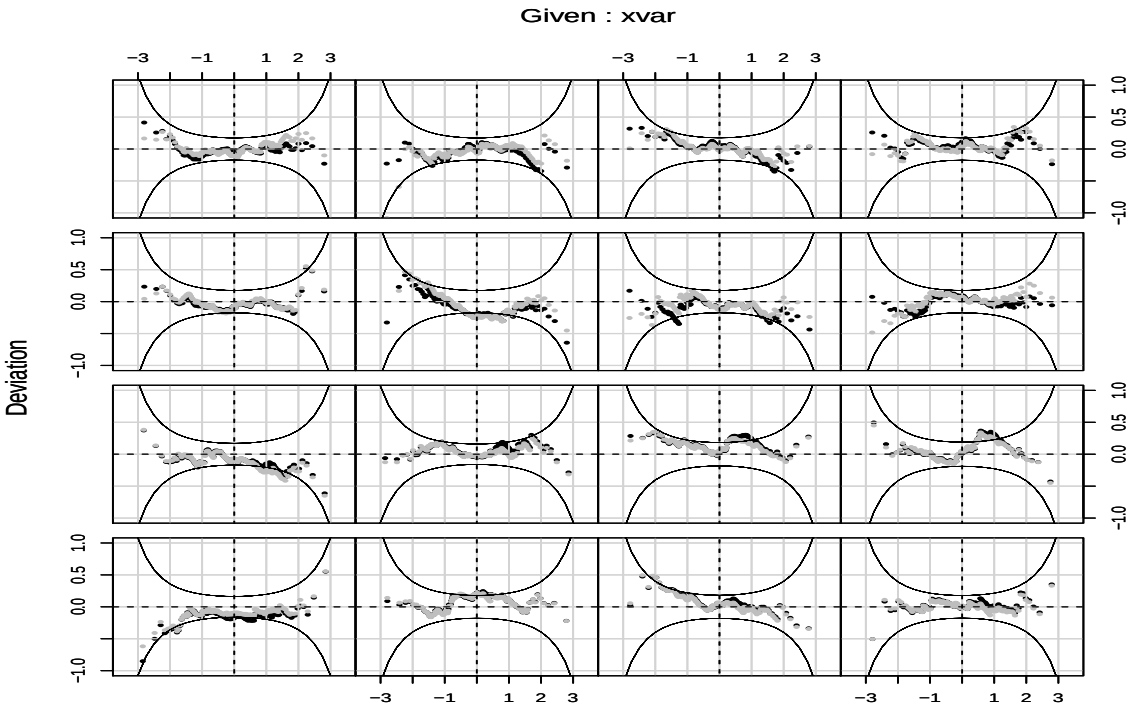


Fig. 8.12 Twin worm plot for logitSSTInf1 (dark points) and Gen.Tobit (light points) models.

The worm plots of the two proposed models, the inflated logitSST and generalised Tobit models are shown in Figure 8.12. Based on the worm plots the proposed inflated lositSST and

generalised Tobit models fit well to the data, since approximately 95% of the points of the worm plots lie between the two elliptic 95% pointwise confidence band curves.

8.3.4.2 Z and Q statistics

Another residual based diagnostic tool, Z and Q statistics, is used in this section. Z and Q statistics are useful to test the standard normality of the residuals. If the model is correct the true residuals have a standard normal distribution. The Z_{gj} statistics for $j = 1,2,3,4$ test whether the mean, variance, skewness and excess kurtosis of the residuals are the standard normal values of $0,1,0$ and 0 respectively within each explanatory variable interval group $g = 1,2,3.....G$, Royston and Wright (2000). The test statistics for skewness and kurtosis are given by D'agostino et al. (1990).

Royston and Wright (2000) also computed Q-statistics by

$$Q_j = \sum_{g=1}^{G} Z_{gj}^2 \tag{8.6}$$

The test statistics Q_1, Q_2, Q_3 and Q_4 provide global test statistics, combining all G groups, that the mean, variance, skewness and excess kurtosis of the residuals are correct (i.e, $0,1,0$ and 0 respectively), see Royston and Wright (2000) and Rigby and Stasinopoulos (2004). Royston and Wright (2000) suggest an approximate Chi square distribution with adjusted degrees of freedom $G - df_\mu$, $G - [df_\sigma + 1]/2$ and $G - df_\nu$ for Q_1, Q_2 and Q_3 respectively, if the model is correct. Rigby and Stasinopoulos (2004) suggest adjusted degrees of freedom $G - df_\tau$ for the Q_4 statistic. If any of the Q-statistics is significant this provides an indication that the model may be inadequate.

The value of squared Z_{gj} helps to identify which height group is causing the Q_j statistic to be significant. If the model is correct Z_{gj} should be approximately normally distributed. Hence a rough guide for the Z_{gj} value to be significant at the 5% significance level is $Z_{gj} > 1.96$

or $Z_{gj} < -1.96$ indicating that the model may be inadequate within the corresponding height interval.

Figure 8.13 shows the visual display of the Z_{gj} statistics for $g = 1,2,3....16$ and $j = 1,2,3,4$ for each of the four models. The corresponding intervals of height for the 16 groups are given on the left of each plot. The larger the circles, the larger the value of $|Z_{gj}|$. The radius of the circle is proportional to $|Z_{gj}|$. A square within the circle indicates that $|Z_{gj}| > 1.96$ indicating a misfit of the model to the response variable within the corresponding interval of height. In colour, red indicates $Z_{gj} > 0$ and blue indicates $Z_{gj} < 0$.

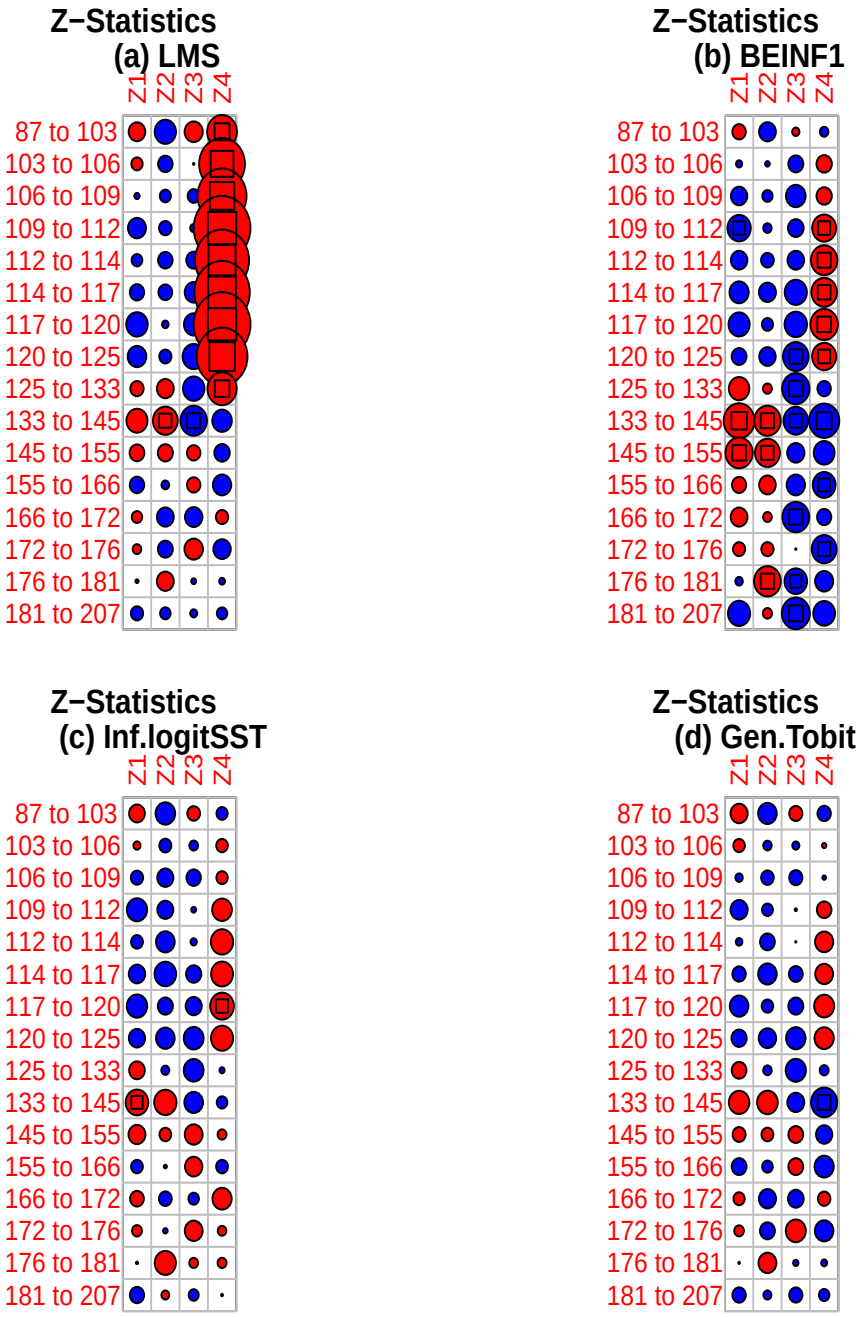


Fig. 8.13 Z statistics for a) LMS b) BEINF1 c) logitSSTInf1 d) Generalised Tobit

Hence Figure 8.13 shows the presence of many misfits in the LMS and beta inflated models. Four misfits were found in the inflated *logitSST* model and one in the generalised Tobit model. Analysis of the residuals using Z statistics is also consistent with the diagnostic tools used earlier in this paper. The corresponding Q-statistics, i.e. Q_j for $j = 1,2,3,4$ given by (8.6), for each of the four models are given in Table 3. This indicates that the generalised Tobit model provides an

adequate fit (i.e. no significant Q statistics), while the other three models provide inadequate fits. Note that the local penalty $k = 6$ was used for each of the smoothing degrees of freedom. When the penalty was increased to $k = \log n = 8.06$, for the SBC criterion, then the generalized Tobit model became inadequate, while if the penalty was decreased to $k = 2$, for the AIC criterion, then its centile curves became erratic.

Table 8.6 Q statistics

Method	Q_1 (p-val)	Q_2 (p-val)	Q_3 (p-val)	Q_4 (p-val)
LMS	14.01(0.258)	16.8 (0.300)	24.41 (0.013)	839.6 (0.000)
BEINF1	42.5 (0.000)	26.6(0.026)	55.5 (0.000)	63.8(0.000)
GenTobit	11.3 (0.498)	12.9 (0.550)	13.9 (0.457)	20.8 (0.185)
logitSSTInf1	15.9 (0.151)	19.6 (0.163)	14.2 (0.426)	21.2 (0.060)

It can be concluded that the generalized Tobit model (with $k = 6$) performed well compared to other models. It has smooth centile curves and provides a good fit to the data.

Next the fitted (or predicted) mixed continuous-discrete probability (density) function of $Y = FEV1/FVC$ is plotted for for six new values of height (80, 100, 120, 140, 160, 180) cms, first for the logitSSTinf1 model in Figure 8.14 with corresponding fitted parameter values given in Table 8.7, and second for the generalized Tobit (BCCGorc) model in Figures 8.15 to 8.17 with corresponding fitted parameters values given in Table 8.8. Note that from Table 8.7 $P(Y = 1) = \xi_1$, while from Table 8.8 $P(Y = 1) = P(V \geq 1)$ where $V \sim BCCGo(\mu, \sigma, \nu)$.

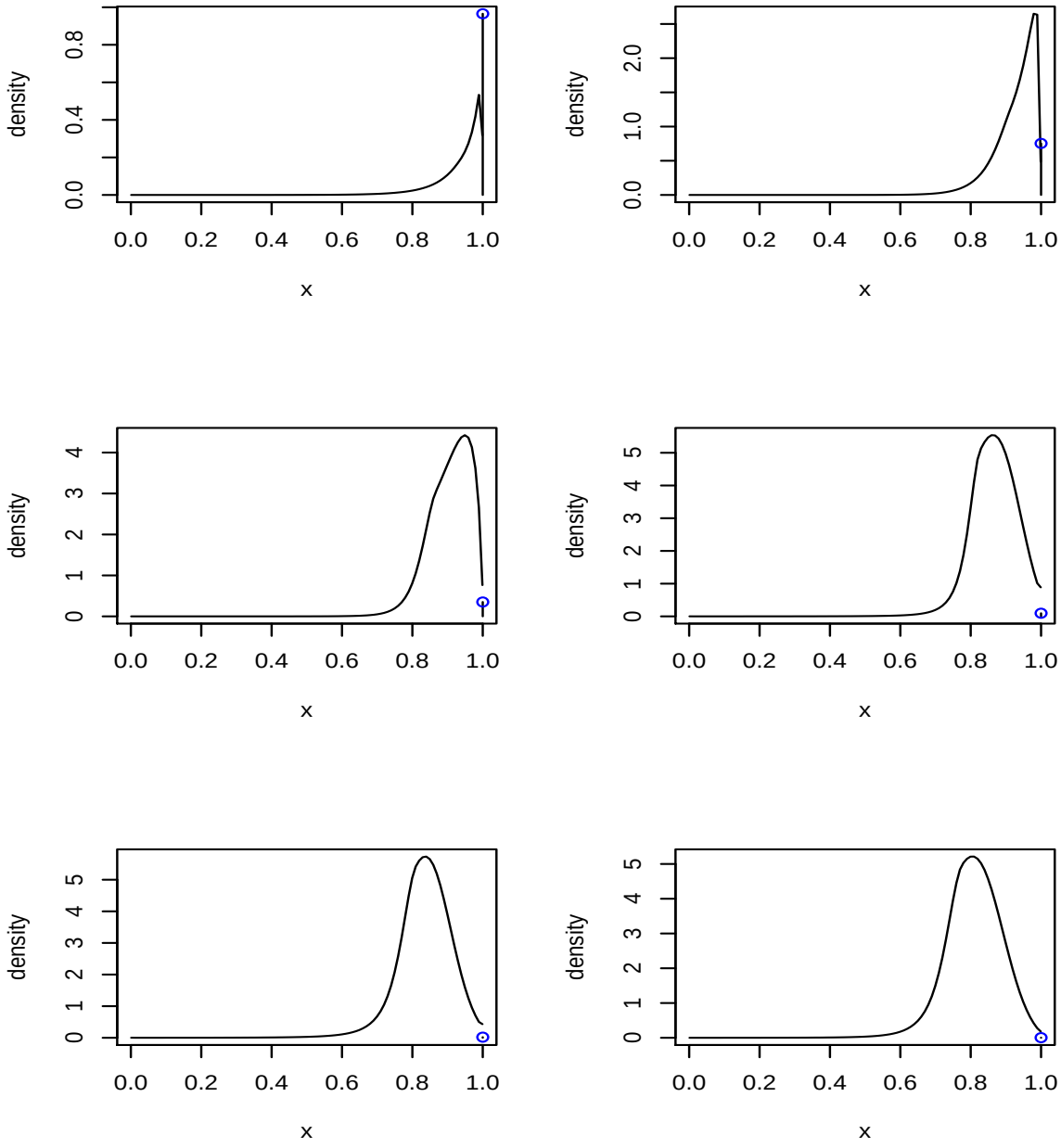


Fig. 8.14 Plot of the predicted pdf of Y for the logitSSTInf1 model for height, from top left in rows 80, 100, 120, 140, 160, 180 (cm)

Table 8.7 Predicted parameter values using logitSSTInf1

Height	μ	σ	ν	τ	ξ_1
80	3.59	1.61	1.91	82894.18	0.99
100	3.23	1.25	1.78	1701.88	0.83
120	2.96	1.03	1.69	85.34	0.37
140	2.52	0.88	1.62	9.31	0.09
160	2.01	0.76	1.56	3.67	0.02
180	1.74	0.68	1.51	3.71	0.006

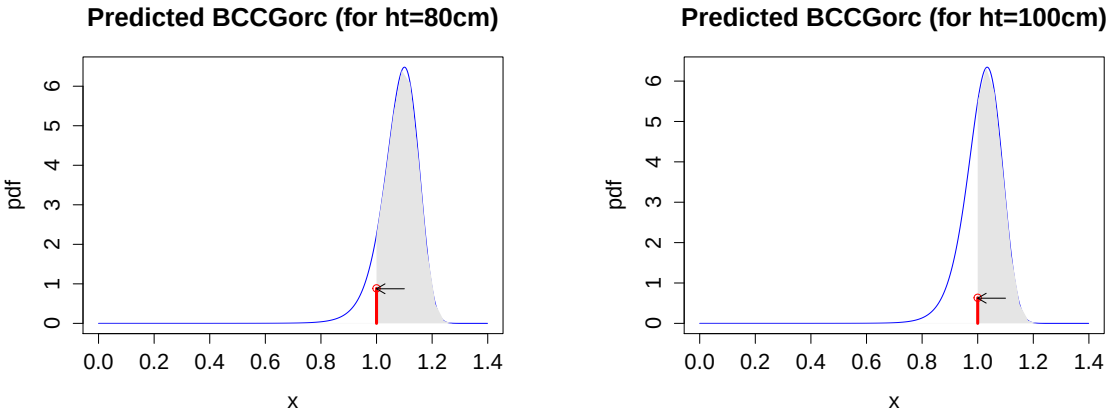


Fig. 8.15 Plot of the predicted pdf of Y for the BCCGorc model at height (80cm and 100cm)

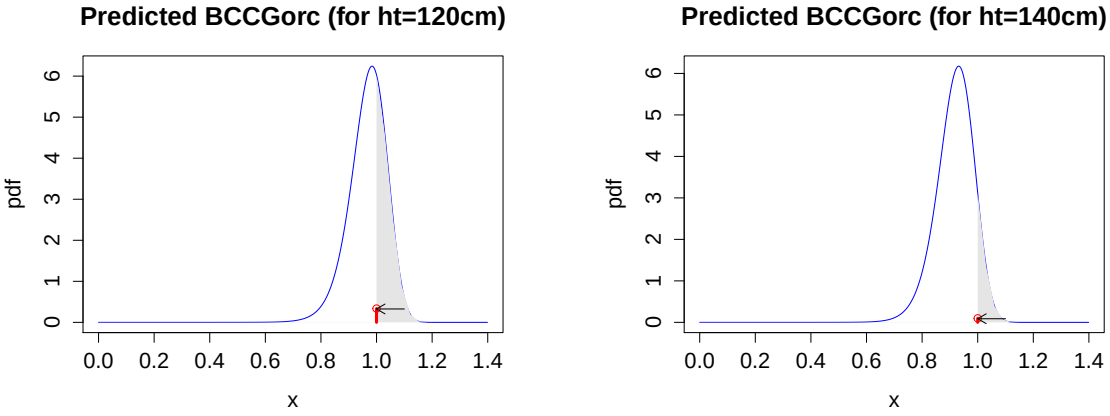


Fig. 8.16 Plot of the predicted pdf of Y for the BCCGorc model at height (120cm and 140cm)

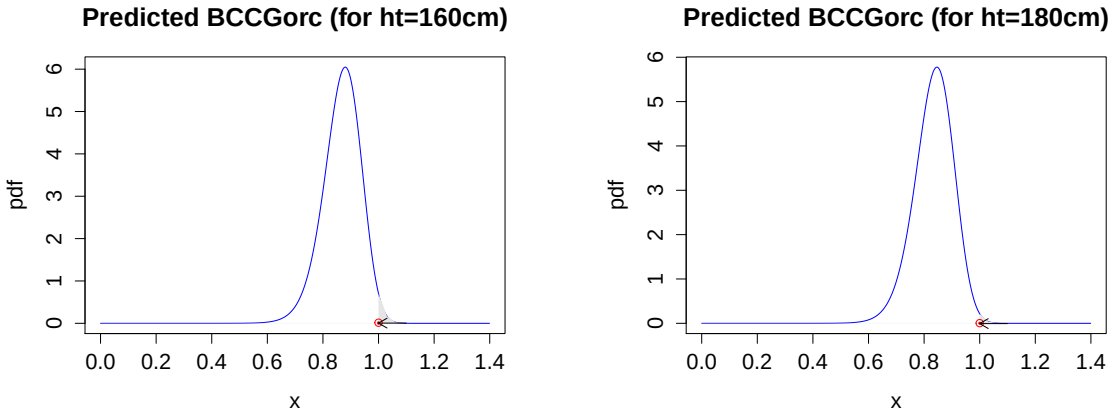


Fig. 8.17 Plot of the predicted pdf of Y for the BCCGorc model at height (160cm and 180cm)

Table 8.8 Predicted parameter values using BCCGorc

Height (cm)	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$
80	1.09	0.06	4.74
100	1.02	0.06	4.06
120	0.97	0.07	3.53
140	0.92	0.07	3.09
160	0.87	0.08	2.73
180	0.84	0.08	2.42

8.4 Conclusion

This chapter proposed the inflated logit skew student t (i.e. inflated $logitSST$) distribution and a generalized Tobit model as mixed continuous-discrete distributions to model a response variable recorded on the unit interval (0,1], including 1. The main purpose of this chapter is to offer two models for centile estimation as viable alternatives to the LMS and beta inflated models. The chapter focuses on a response variable recorded on the interval $(0,1]$. Extension to the interval [0,1) can be achieved simply by analysing a new response variable $(1-Y)$, which is on the interval (0,1] instead of Y. Extension to the interval $[0,1]$ can be achieved by inflating a flexible

distribution on $(0,1)$, e.g. the *logitSST* distribution, at both 0 and 1, or by censoring a flexible distribution on $(-\infty,\infty)$, e.g. the *SST* distribution, below 0 and above 1 for the generalized Tobit model.

An empirical application to real data has been presented modelling a lung function response variable on the interval $(0,1]$. This chapter uses the Akaike information Criterion (AIC), Akaike (1974), and Schwarz Bayesian Criterion (SBC), Schwarz (1978), to compare the relative performance of the models. The model with lowest AIC or SBC is ranked as best model. Worm plots and Z statistics were used to the check the adequacy of each of the models. The LMS, beta inflated and Tobit models were clearly inadequate at fitting the response variable, while the generalized Tobit (i.e, *BCCGo* right censored at 1) and inflated *logitSST* models provided better fits. From the empirical example it can be concluded that generalized Tobit and the inflated *logitSST* models can provide better fits than the LMS, beta inflated and Tobit models and can provide powerful tools for modelling proportion data.

Chapter 9

Application of proposed models to a response variable on [0,1)

9.1 Introduction

In this chapter an application of the proposed models to response variable observations on [0,1) is given. For the sake of comparison two proposed models: inflated GAMLSS and generalized Tobit GAMLSS models, together with the beta inflated at 0 model and standard Tobit model are fitted. In addition to the application, model adequacy is checked by investigating the residuals of each fitted model.

9.2 Data

A peer assisted student success data set consisting of a total of 1679 observations is analysed. Here the proportion of attendance (PA) is the explanatory variable and the average module mark (AMM) is the response variable. Note that the observed variable AMM is converted to a proportion, $AMM1 = AMM/100$, which has a number of zeroes but no values at one, i.e.

$0 \leq AMM1 < 1$. Table 9.1 of descriptive statistics includes the number of observations (N), mean, standard deviation and minimum and maximum values of variables AMM and PA.

Table 9.1 Pass scheme data

Statistic	N	Mean	St. Dev.	Min	Max
AMM	1,669	44.445	24.368	0.000	90.500
PA	1,669	0.253	0.290	0.000	1.000

Using the variables, the proposed models are fitted together with the beta inflated at 0 model and standard Tobit model. Figure 9.1 shows the scattter plot with marginal histogram of the variable average module mark as a proportion (AMM1) against proportion attendance (PA). Pass scheme data set used in chapter 9 has been privately collected from the PASS Scheme Project, Center for Professional and Educational Development (CPED),LMU.

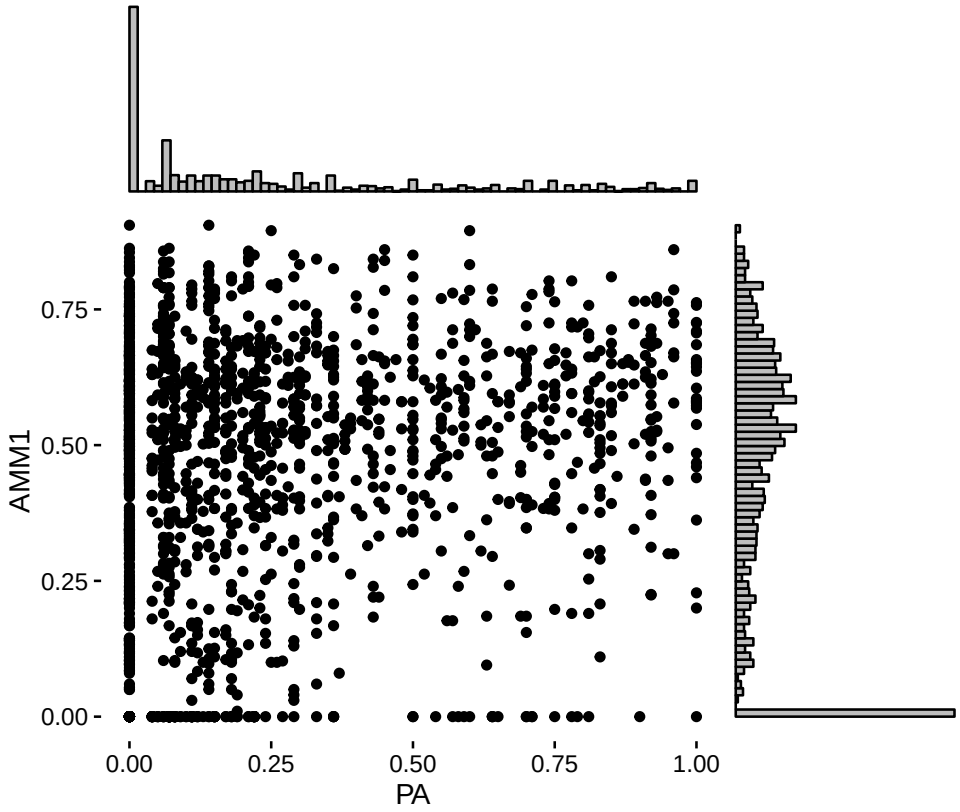


Fig. 9.1 Scatter plot of average module mark as a proportion against proportion attendance

9.3 Inflated at 0 GAMLSS model

The GAMLSS model inflated at zero is a special case of the general inflated GAMLSS model described in chapter 5. The model use here is the inflated logit skew t (type 3) distribution which is a mixture of two components: a discrete component zero ($Y = 0$) with probability p_0 and a continuous component $(0 < Y < 1)$ with a logitST3(μ, σ, ν, τ) distribution with probability $(1 - p_0)$. The observed variable Y then follows an inflated logit skew t (type 3) distribution

$$Y \sim \text{InflogitST3}(\mu, \sigma, \nu, \tau, p_0)$$

with mixed continuous-discrete probability (density) function is given by

$$f_Y(y|\mu,\sigma,\nu,\tau,p_0) = \begin{cases} p_0, & \text{if } y = 0 \\ (1-p_0) f_W(y|\mu,\sigma,\nu,\tau), & \text{if } 0 < y < 1 \end{cases}$$

for $0 \leq y < 1$ and $0 < p_0 < 1$, where $f_W(y|\mu,\sigma,\nu,\tau)$ is a logit skew t (type 3) pdf with $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, $\tau > 0$. For ease of interpretation parameter p_0 used.

9.4 Generalised Tobit model

Since $0 \leq AMM1 < 1$, AMM1 is transformed to $Y = 1 - AMM1$, so $0 < Y \leq 1$, then a generalized Tobit model for Y can be obtained by right censoring at 1 a variable V with any distribution with range $0 < V < \infty$.

For example let V have a Box-Cox t distribution, Rigby and Stasinopoulos (2006), denoted by

$$V \sim \text{BCT}(\mu,\sigma,\nu,\tau)$$

for $0 < V < \infty$. Let

$$Y = \begin{cases} V, & \text{if } 0 < V < 1 \\ 1, & \text{if } V \geq 1 \end{cases}$$

Then Y has a right censored Box-Cox t distribution denoted

$$Y \sim \text{BCTrc}(\mu,\sigma,\nu,\tau)$$

with pdf is given by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \begin{cases} f_V(y|\mu,\sigma,\nu,\tau), & \text{if 0<y<1} \\ 1 - F_V(1), & \text{if y=1} \end{cases}$$

for $0 < y \leq 1$.

9.5 Model selection

In this section the generalized Akaike information criterion GAIC (Akaike, 1983) is used to compare different non nested inflated and generalized Tobit GAMLSS models. The main advantage of using GAIC is that it allows different penalties to be tried to penalize over fitting models. For the sake of comparison GAIC is assessed by adding to the fitted global deviance (GD) a fixed penalty k for each effective degree of freedom used in the model, i.e, $GAIC = GD + k * df$. The fitted global deviance is obtained by $GD = 2l(\hat{\theta})$ where $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4) = (\hat{\mu}, \hat{\sigma}, \hat{\nu}, \hat{\tau})$ where the maximised log likelihood function is given by $l(\hat{\theta}) = \sum_{i=1}^{n} l(y_i|\hat{\theta}^i)$ where $\theta^i = (\mu_i, \sigma_i, \nu_i, \tau_i)$ and $l(y_i|\hat{\theta}^i)$ is the fitted log likelihood function for observation y_i.

Two special cases of GAIC are the Akaike information Criteron (AIC) Akaike (1974), Schwartz Baysian Criteron (SBC) Schwarz (1978). The special cases of GAIC are given by, GAIC =AIC, if k = 2, while $GAIC = SBC$, if $k = \log(n)$. AIC is considered a more generous approach (potentially leading to overfitting), while SBC is a more restrictive approach (potentially leading to underfitting). In this example both approaches have been used for a more robust decision.

Finding an optimal value of k is another important aspect of model selection. A selection of different value of k, eg $k = 2, 2.5, 3.5, 4$ could be used for robustness of the model selection, (Rigby and Stasinopoulos, 2006). Kim and Gu (2004) suggested a penalty value 2.8, whereas Rigby and Stasinopoulos (2006), suggested a range $2.5 < k <= 3$ worked well for their data set. In this chapter in addition to AIC (i.e. $k = 2$) and SBC (i.e.$k = \log(n) = \log(1679) = 7.4$), a penalty value $k = 6$ is used as a compromise between AIC and SBC. Table 9.2 summarises the

Table 9.2 Relative quality of fitted models

Models	AIC	SBC	GAIC(k=6)
logitST3Inf1	0	56.93	42.01
BCTrc	1.3	45.01	33.56
Tobit (Norc)	442.10	485.23	473.92
BEINF1	100.86	135.73	126.59

values of AIC, SBC and GAIC(k=6) of the fitted models, each value having 292.27 substracted

for clarity of presentation. The models fitted to $Y = 1 - AMM1$ are the inflated logit skew t (type 3) (i.e. logitST3Inf1 distribution), the generalized Tobit right censored BCT (BCTrc), the Tobit right censored normal (NOrc) and the beta inflated at 1 (BEINF1). Among the four models (i.e. BEINF1, Tobit (NOrc), gen.Tobit (BCTrc) and InflogitST3), the generalized Tobit (BCTrc) model and inflated logit skew student t (type 3) distribution models have lower AIC, SBC and GAIC values. From Table 9.2, the logitST3Inf1 model is best as judged by AIC (as it has the lowest value), while the generalised Tobit model (BCTrc) is best as judged by SBC. Using criterion GAIC with $k = 6$, the inflated $logitSST$ and generalised Tobit models are almost equally good. The $BEINF1$ and standard Tobit models perform much worse.

9.6 Residual based diagnostics

A model may suffer misspecification. Moreover the presence of outliers in the data set may impair the model accuracy (Ospina and Ferrari, 2012). With this in view a residual analysis of the fitted model is used to check the model adequacy. In this chapter residual based diagnostic tools (e.g.worm plot) will be used to check the adequacy of each fitted model.

9.6.1 Worm plot

van Buuren and Fredriks (2001) introduced the worm plot which consists of de-trended Q-Q plots. The shape of the worm plot indicates how the observed response variable distribution differ from the assumed distribution.

A further diagnostic of residuals is obtained by analysing the coefficients of parameters in a cubic function fitted to each individual worm plot, van Buuren and Fredriks (2001). The four columns of coefficients are the fitted constant, linear, quadratic and cubic coefficients denoted by $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$. These shape coefficients can be used for quantitative assessment of the model fit. $\hat{\beta}_0$ provides a measure of the difference between the theoretical and empirical mean

of the residuals. $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ provide a measure of the difference between the theoretical and empirical variation, skewness and kurtosis of the residuals respectively. van Buuren and Fredriks (2001) provide threshold values $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ of 0.10, 0.10, 0.05 and 0.03 respectively. If the absolute value of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ or $\hat{\beta}_3$ exceeds the corresponding threshold value, this will be termed as misfit.

Table 9.3 Fitted coefficient values (logitST3Inf1)

PA-group	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
[1]	-0.12	0.04	0.01	-0.02
[2]	-0.40	0.29*	-0.16	-0.09
[3]	0.07	0.27*	-0.06	-0.04
[4]	-0.13	0.17*	0.06	-0.03
[5]	0.09	0.07	0.01	-0.02
[6]	0.07	-0.19	0.01	0.02
[7]	0.13*	-0.16	0.05	0.01
[8]	0.00	-0.14	0.02	0.02
[9]	0.21*	-0.14	-0.03	0.015

$*$ misfit

Table 9.4 Fitted coefficient values (BCTrc)

PA-group	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
[1]	0.06	-0.05	-0.01	-0.00
[2]	0.34*	0.03	-0.01	-0.10
[3]	-0.11	0.18*	0.02	-0.03
[4]	0.087	0.11	-0.05	-0.02
[5]	-0.10	0.08	-0.03	-0.03
[6]	-0.05	-0.19	-0.02	0.03
[7]	-0.10	-0.09	-0.05	-0.00
[8]	0.09	-0.02	0.04	0.03
[9]	-0.00	0.06	0.03	-0.02

$*$ misfit

Table 9.5 Fitted coefficient values (BEINF1)

PA-group	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
[1]	0.06	-0.01	-0.06	-0.02
[2]	-0.22	0.15*	-0.09	-0.10
[3]	0.17*	0.16*	-0.08	-0.05
[4]	-0.05	0.15*	0.01	-0.05
[5]	0.13*	0.07	-0.01	-0.05
[6]	0.13*	-0.15	-0.08	0.02
[7]	0.16*	-0.14	-0.05	0.01
[8]	-0.06	0.00	-0.05	-0.01
[9]	0.10	0.00	-0.07	-0.02

$*$ misfit

Table 9.6 Fitted coefficient values (NOrc)

PA-group	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
[1]	-0.12	0.01	0.14*	-0.02
[2]	0.27*	0.48*	0.19*	-0.15
[3]	-0.20	0.14*	0.16*	-0.05
[4]	-0.04	0.13*	0.13*	-0.04
[5]	-0.21	0.06	0.15*	-0.03
[6]	-0.24	-0.11	0.18*	0.02
[7]	-0.29	-0.15	0.17*	0.04
[8]	-0.05	-0.03	0.20*	0.01
[9]	-0.19	-0.14	0.19*	0.02

$*$ misfit

Tables 9.3, 9.4, 9.5 and 9.6 present the β shape coefficient values of the four fitted models inflated logitST3, BCTrc, BEINF0 and NOrc respectively. The inflated logit skew student t (type 3) distribution and generalized Tobit (BCTrc) models show less misfits than the beta inflated and Tobit model. However BCTrc shows the least misfits with one misfit in $\hat{\beta}_0$ (PA group 2) and one misfit in $\hat{\beta}_1$ (PA group 3). The BEINF1 and Tobit models show many misfits as their fitted model's (absolute) coefficient values exceed the threshold values.

The twin worm plot of the proposed logitST3 inflated at 1 and generalized Tobit models is given in Figure 9.2. In the figure the points generally lie close to the horizontal line and are mostly between the 95% confidence interval curves given by the elliptic curves, providing evidence of the adequacy of the proposed models.

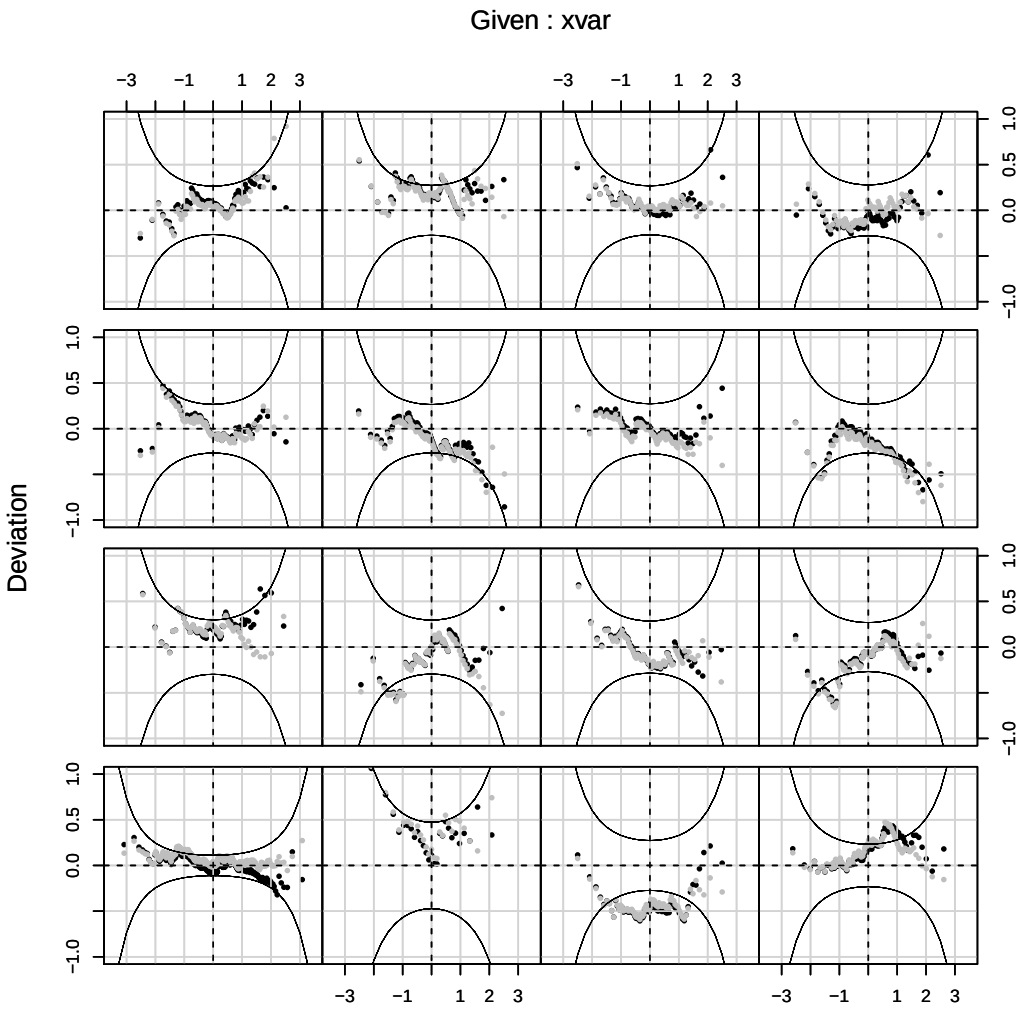


Fig. 9.2 Twin worm plot of logitST3Inf1 (light points) and BCTrc (dark points)

Figure 9.3 shows the twin worm plot of the fitted beta inflated at 0 and standard Tobit model. The residuals are shown in Figure 9.3 in nine non-overlapping intervals. In many intervals a higher percentage of the points than 5% lie outside the region between two elliptic curves, indicating that the fitted distribution of the model is inadequate to explain the response variable.

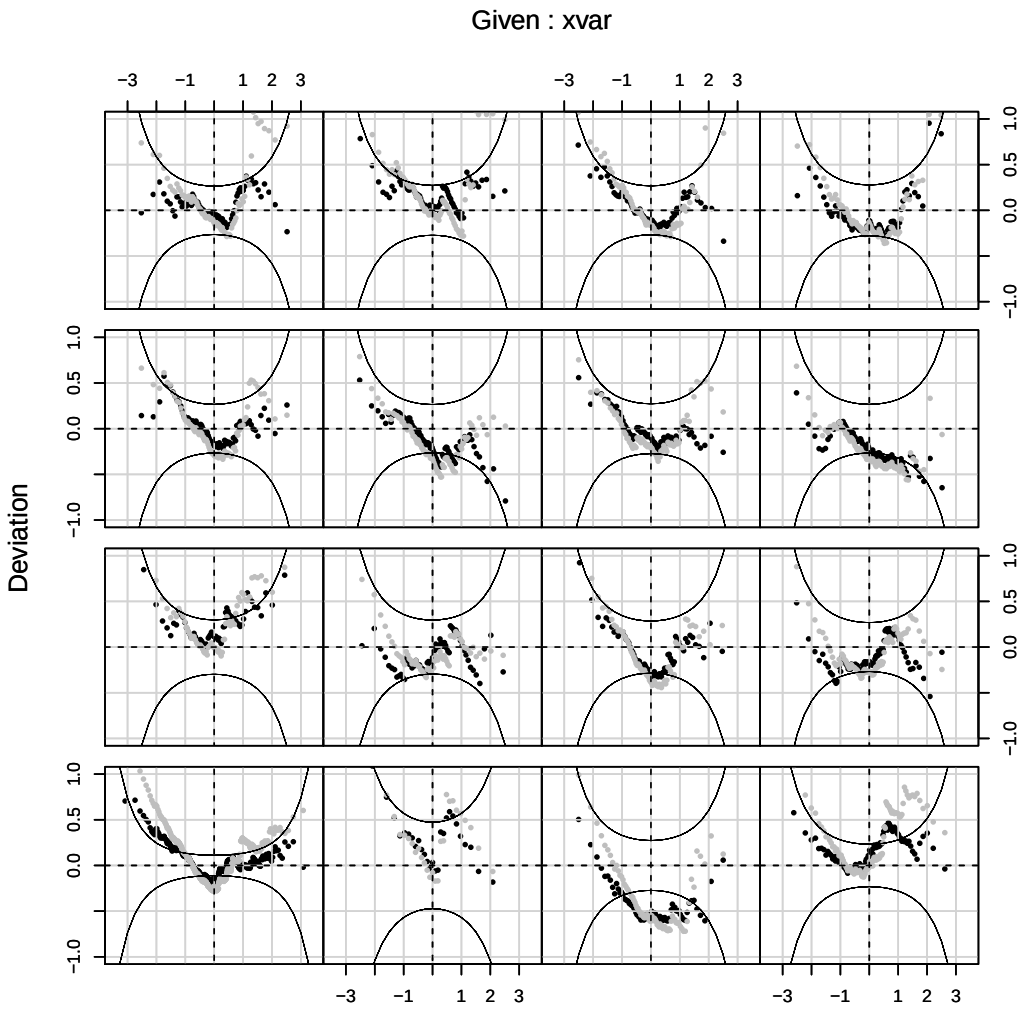


Fig. 9.3 Twin worm plot of Tobit model (light points) and BEINF1 (dark points

9.7 Fitted centile curves

Figure 9.4 shows the centile curves constructed using Tobit model, beta inflated at 1, Inflated logit skew student t type 3 and generalized Tobit (BCTrc) models. The fitted $(2, 10, 25, 50, 75, 90, 98)\%$ centile curves show that the BEINF1, logitST3Inf1 and generalized Tobit (BCTrc) models constructed the most smooth curves, while Tobit model constructed the least smooth curves.

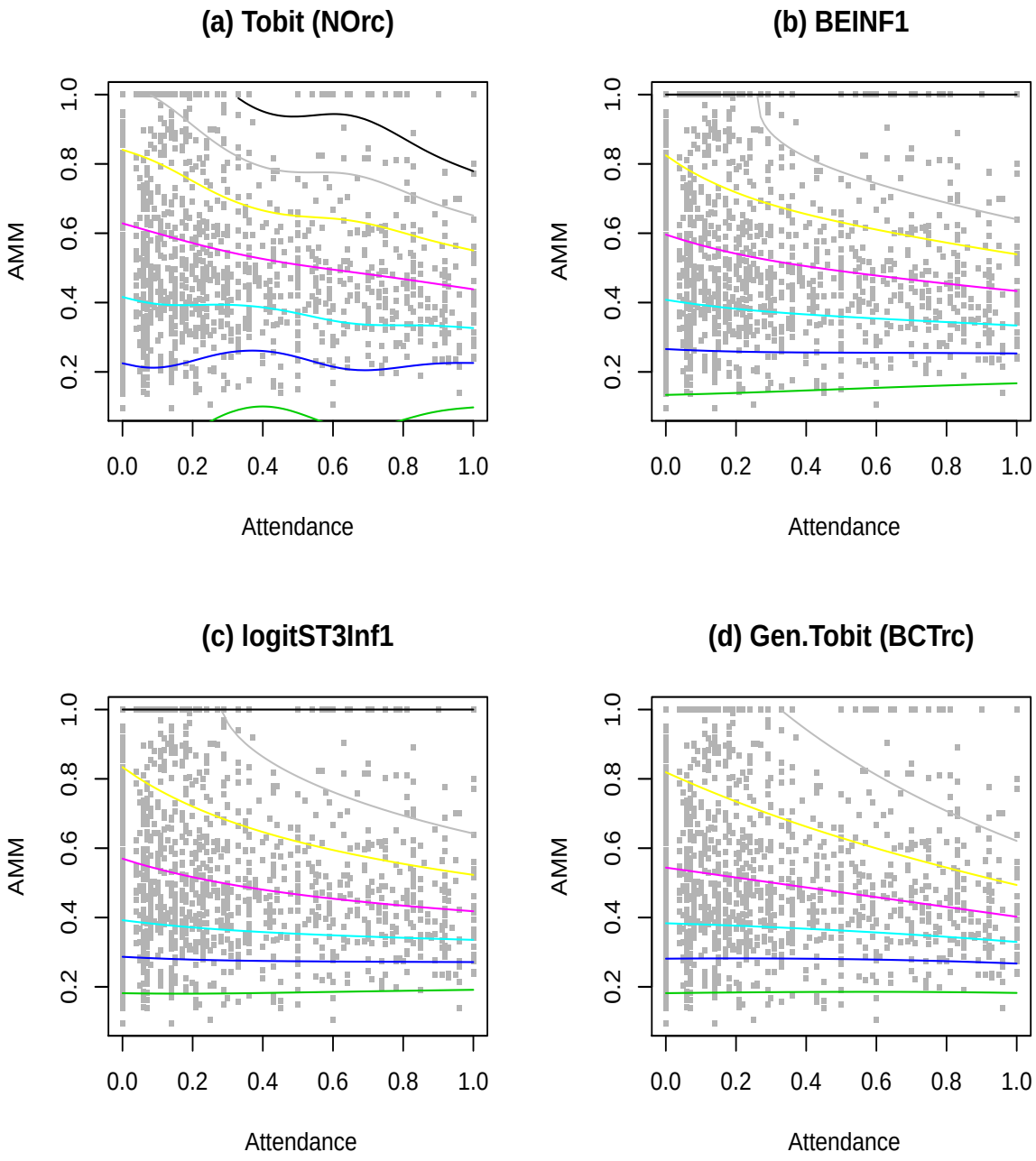


Fig. 9.4 Centile curves for model a) NOrc b) BEINF1 c) logitST3Inf1 d) Generalised Tobit (BCTrc)

Table (9.7) shows the comparison between fitted centile percentages at or below the $(2, 10, 25)\%$ and nominal percentages. [1] Among the four models the genralized Tobit (BCTrc) and inflated logit skew student t distribution type 3 at 1 (logitST3Inf1) models perform best.

Table 9.7 Comparison of fitted centile percentages for $Y = 1 - AMM1$

Nominal Centile %	NOrc	BEINF0	GenTobit (BCTrc)	logitST3Inf1
2	0	0.42	2.27	2.21
10	5.27	8.39	9.82	9.58
25	27.92	26.54	24.20	24.20

9.8 Fitted distributions of $Y = (1 - AMM)$ for different values of PA

The fitted (i.e. predicted) distribution of $Y = 1 - AMM1$ was plotted in Figure 9.5 for five values of the explanatory variable PA (10%, 30%, 50%, 70%, 80% and 90%) using the fitted inflated logitST3 model. Predicted values of the parameters are also given in the Table 9.8.

[1] Percentages above 25% were not given since for at least one model these centile curves reach 1 and hence the sample percentage *at* or below the centile curve provides a distorted overestimate of the correspondent model centile curve percentage.

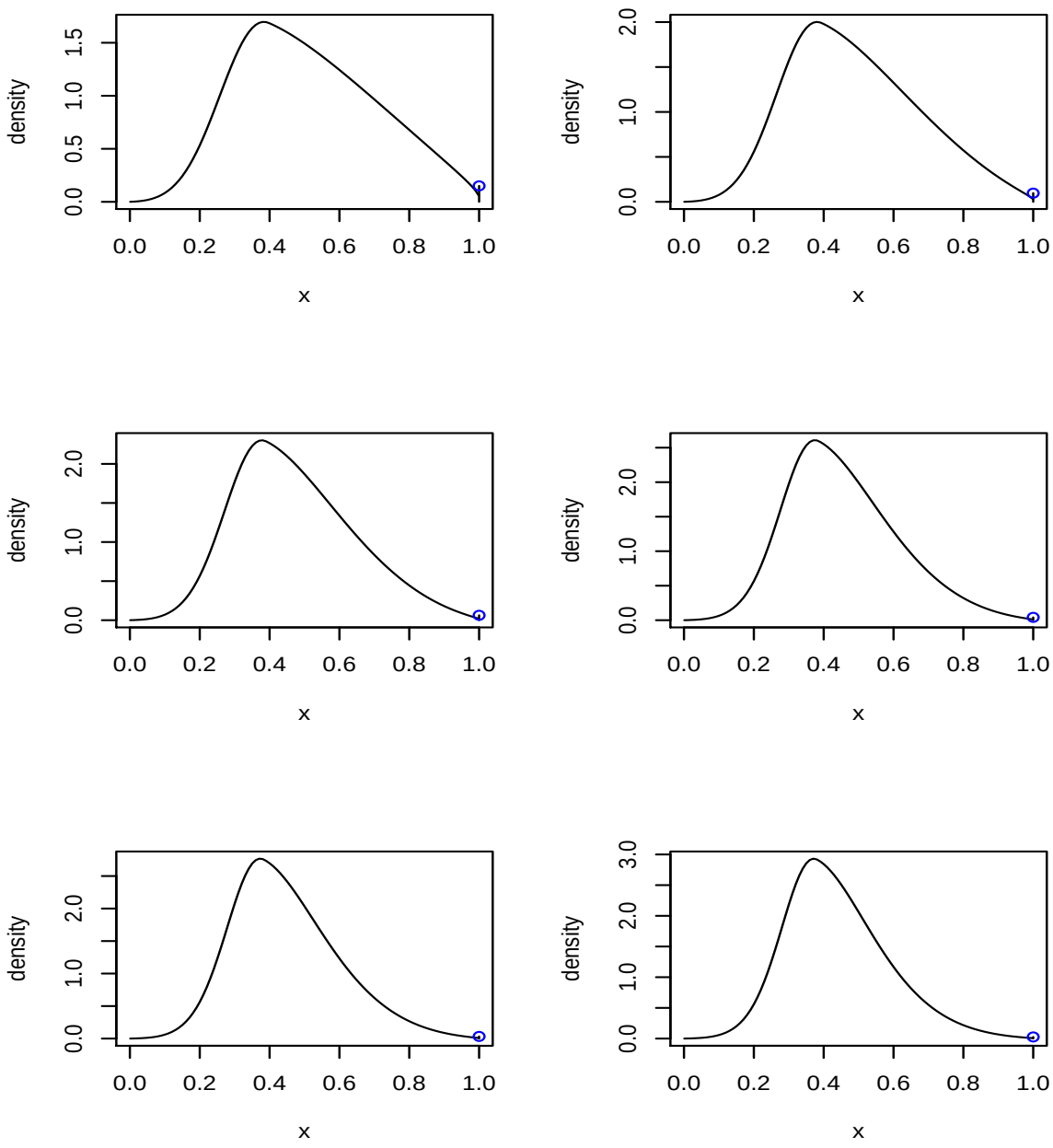


Fig. 9.5 Predicted distribution for Y for the inflated logitST3 distribution for different values of PA from top left in rows

Table 9.8 Fitted parameter values for the logitST3Inf1 model for different values of attendance(%)

Attendance	μ	σ	ν	τ	ξ_1
10%	-0.41	0.77	1.43	8.77	0.15
30%	-0.43	0.70	1.39	8.26	0.10
50%	-0.45	0.64	1.35	7.79	0.063
70%	-0.47	0.58	1.31	7.34	0.041
80%	-0.48	0.55	1.29	7.13	0.03
90%	-0.49	0.53	1.27	6.92	0.027

Fitted (predicted) distribution of $Y = 1 - AMM1$ for generalized Tobit model are given in Figures 9.6, 9.7 and 9.8 and corresponding values are given in Table 9.9. Note that from Table 9.8 $P(Y = 1) = \xi_1$, while from Table 9.9 $P(Y = 1) = P(V \geq 1)$ where $V \sim BCT(\mu, \sigma, \nu, \tau)$.

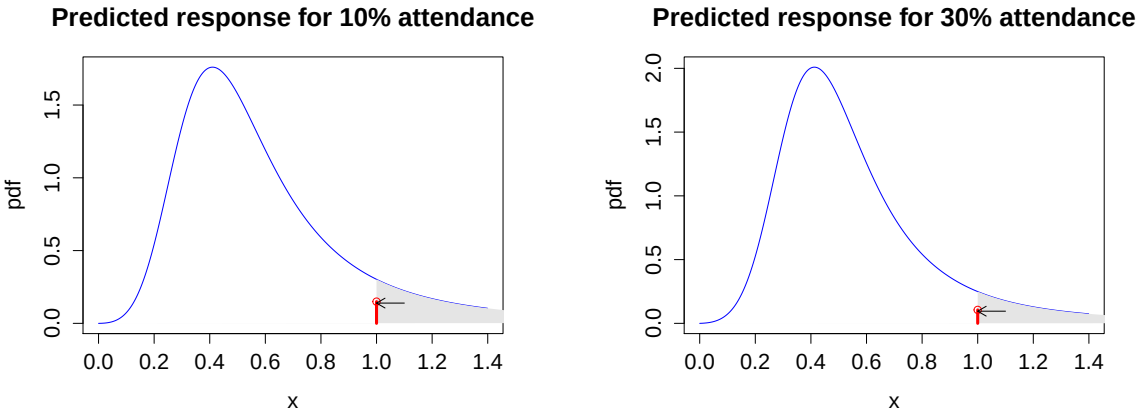


Fig. 9.6 Predicted value for attendance 10 % and 30 %

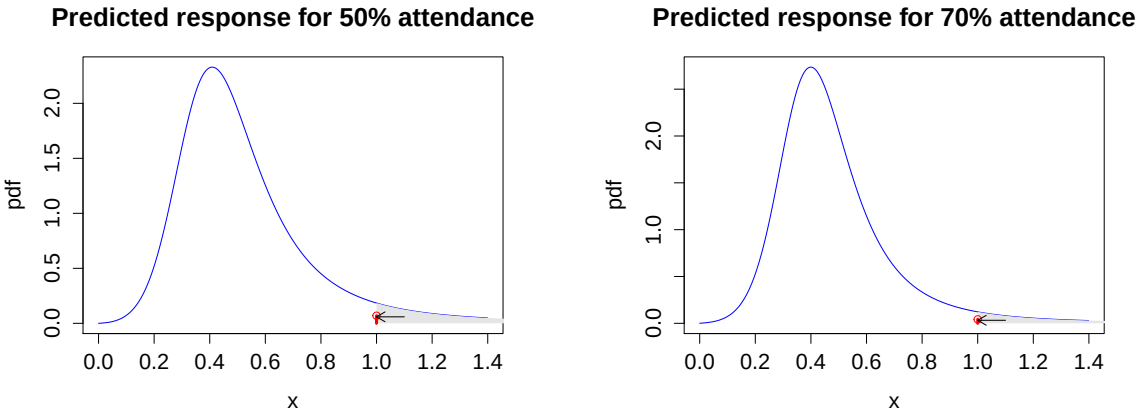


Fig. 9.7 Predicted value for attendance 50 % and 70 %

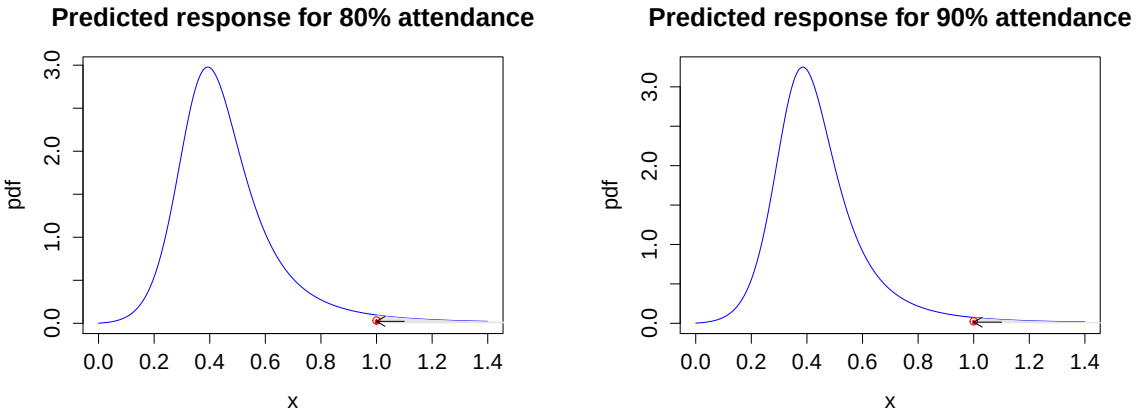


Fig. 9.8 Predicted value for attendance 80 % and 90 %

Table 9.9 Fitted parameter values for the *BCTrc* model for different values of attendance

Attendance	μ	σ	ν	τ
10%	0.53	0.49	-0.39	5.40
30%	0.50	0.43	-0.34	4.98
50%	0.47	0.38	-0.28	4.59
70%	0.44	0.33	-0.23	4.23
80%	0.43	0.31	-0.20	4.07
90%	0.42	0.29	-0.17	3.91

9.9 Conclusion

Methods for handling a proportion response variable on $[0, 1)$ were investigated. The inflated GAMLSS model and generalized Tobit model were developed. The class of models widen the flexibility of the beta inflated and Tobit models. The models were developed within the GAMLSS framework which allows them to adopt all the GAMLSS features i.e. a flexible distribution and models for all the distribution parameters using explanatory variables. Application to the real data set also supports with evidence that the proposed models fit better than the beta inflated at 0 and Tobit models.

Chapter 10

Application on loss given default, a proportion response on [0,1]

10.1 Introduction

Loss given default (LGD) is a proportion of a credit exposure that is lost if the obligor defaults on a loan. Response variable LGD contains values between 0 and 1 including both 0 and 1, where 0 means that the balance is fully recovered while 1 means total loss of exposure at default. This chapter addresses two alternative semi parametric approaches for modelling loss given default, which is measured on the interval [0,1]. The class of models are very flexible and can accommodate skewness and bimodal characteristics of LGD data. The dependence of the predictors of each of the parameters (of the proposed model distribution for LGD) on explanatory variables can be additive P- splines, regression trees or neural network models. The proposed models are applied to a loss given default data set and compared with current popular models.

Loss given default is the key variable for a bank's minimum regulatory capital requirement based on the Basel II framework. Therefore modelling LGD is pivotal for financial regulators and retailers. However modelling LGD poses substantial challenges due to the bounded nature

of LGD data and its unusual distribution, (see Bellotti and Crook (2012)). LGD values often lie on the interval [0,1] and the distribution tends to be bimodal with modes close to the end values.

Previous approaches for modelling (the distribution of) LGD on [0,1] include ordinary least squares, e.g. Qi and Yang (2009), fractional response regression (FRR), Papke and Wooldridge (1996), transformation models, e.g. Qi and Zhao (2011) and Li *et al.* (2014), the inflated beta model, Ospina and Ferrari (2010), a two step approach combining an ordinal logistic regression model and normal error model, Li *et al.* (2014), and Tobit models obtained by censoring a normal distribution or one or two shifted gamma distributions Li *et al.* (2014). In a very recent paper Hossain et al. (2016a,b) proposed inflated `logitSST` and generalized Tobit models for the proportion response variable on the intervals (0,1] and [0,1] respectively.

The purpose of this chapter is to provide two flexible modelling approaches for a proportion response variable measured on the interval from 0 to 1, including both 0 and 1, i.e. range [0,1], following Hossain et al. (2016a,b). In the first approach a flexible distribution for Z with range $(-\infty,\infty)$ is transformed to Y with range (0,1), using an inverse logit transformation, $Y = 1/(1+e^{-Z})$, which is then inflated by including point probabilities for Y at 0 and 1. The second approach is a generalized Tobit model, in which a flexible distribution for Z on $(-\infty,\infty)$ is censored below 0 and above 1 to provide range $0 \leq Y \leq 1$ with probabilities at 0 and 1.

In practice, for each of the two modelling approaches, any available distribution on $(-\infty,\infty)$ within the `gamlss` package, Stasinopoulos and Rigby (2007), can be used for Z, for example the flexible four parameter skew exponential power (SEP), skew student t (SST), sinh arc-sinh $(SHASHo)$ or bi-modal skew symmetric normal $(BSSN)$ distribution, Hasan and El-Bassiouni (2016). In the `gamlss` package the dependence of the predictors of each of the parameters of the proposed model distributions for Y on explanatory variables can be linear, non-linear, non-parametric smooth functions, regression trees or neural network models. Note that Qi and Zhao (2011) and Li *et al.* (2014) found that regression tree and neural network models outperformed linear parametric models.

10.2 Data

Loss Given Default (LGD) is the proportion of the exposure lost following a default. It is also called the severity of loss.

$$LGD = Severity = 1 - RecoveryRate$$

Table 10.1 of descriptive statistics includes the number of observations (N), mean, standard deviation and minimum and maximum values of variables SEVERITY, ORIGIN_YR, DEFAULT_YR, MOB and hrate.

Table 10.1 Loss Given Default

Statistic	N	Mean	St. Dev.	Min	Max
SEVERITY	7,713	0.256	0.347	0.000	1.000
ORIGIN_YR	7,713	2,000.098	2.703	1,994	2,006
DEFAULT_YR	7,713	2,003.449	2.291	2,000	2,007
MOB	7,713	44.068	31.228	0	269
hrate	7,713	0.407	0.079	0.290	0.554

Figure 10.1 shows the scatter plot matrix of LGD data with histogram, kernel density and absolute correlation of variables SEVERITY, ORIGIN_YR, DEFAULT_YR, MOB and hrate.

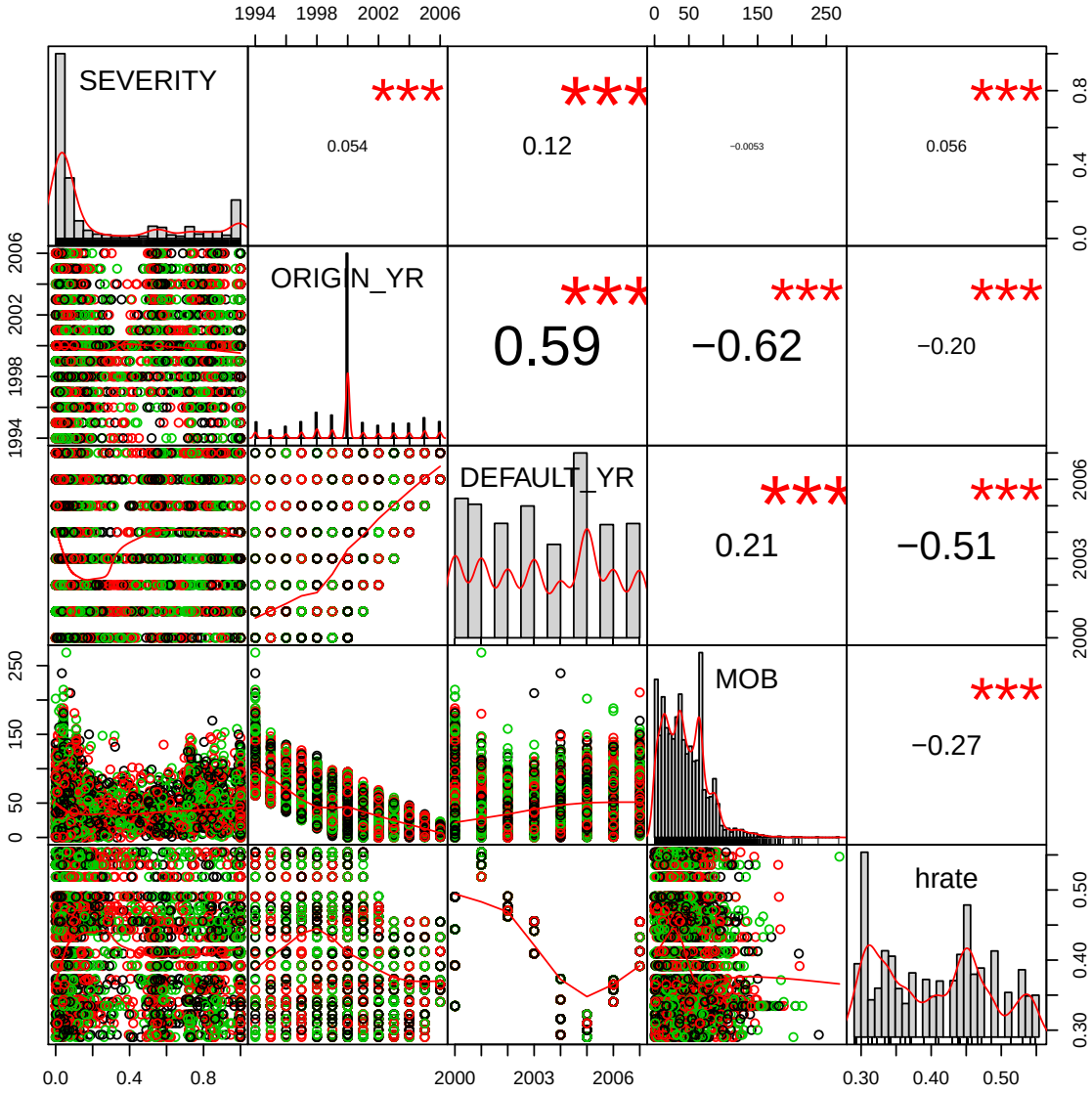


Fig. 10.1 Summary of LGD data.

The range of LGD is bounded on $[0,1]$. The LGD value also tends to follow a bi-modal distribution.

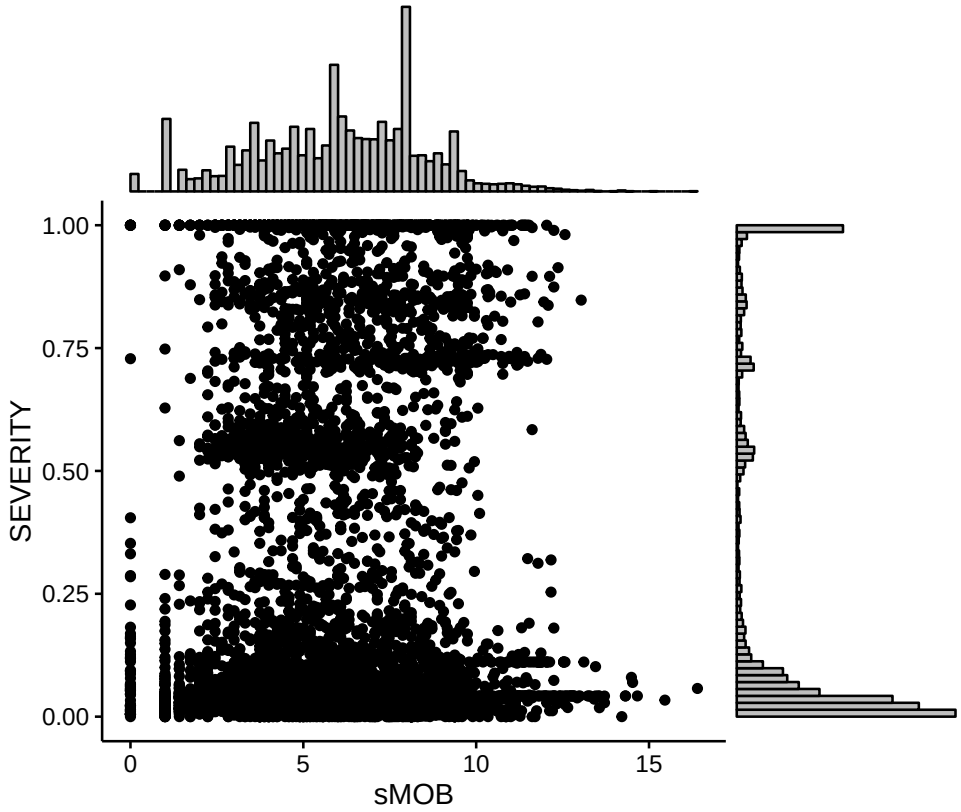


Fig. 10.2 Distribution of observed SEVERITY

The motivating data example is the LGD values collected from one of the leading banks in the USA. The data frame comprises 7713 small business loan defaults between 2000 and 2007. In this analysis the response variable SEVERITY (LGD) is modelled using four covariates: Month-on-Books (MOB), hazard rate (hrate), year of origin (ORIGIN_YR) and year of default (DEFAULT_YR). The four explanatory variables are treated as quantitative variables. Variable MOB was transformed to

$$sMOB = \sqrt{(MOB)}$$

Square root transformation is used to reduce the skewness of the covariate Month on Book value (MOB). Figure 10.2 shows a scatter plot of SEVERITY against explanatory variable sMOB, the square root of month on book, with a marginal histogram of the response variable SEVERITY (i.e., loss given default) .

10.3 Models

10.3.1 Logit distribution

Any distribution on range $-\infty < Z < \infty$ can be transformed to a restrictive range $0 < Y < 1$ by using an inverse logit transformation $Y = 1/(1+e^{-Z})$. The distribution of Y is called a logit distribution. If Z has a four parameter distribution denoted D is general, i.e. $Z \sim D(\mu, \sigma, \nu, \tau)$, then the distribution of Y is called a logit D distribution denoted $Y \sim logitD(\mu, \sigma, \nu, \tau)$. For example if Z has a bi-modal skew symmetric normal distribution $Z \sim BSSN(\mu, \sigma, \nu, \tau)$ on $(-\infty, \infty)$, then Y has a $logitBSSN$ distribution, $Y \sim logitBSSN(\mu, \sigma, \nu, \tau)$ on $(0,1)$. The probability density function $f_Y(y)$ of Y is given by

$$f_Y(y) = f_Z(z)\left|\frac{dz}{dy}\right| = \frac{1}{y(1-y)} f_Z(z) \tag{10.1}$$

where $z = \log[y/(1-y)]$

The $logitBSSN$ distribution is created using the function `gen.Family()` in `gamlss` which allows any `gamlss` distribution with range $(-\infty, \infty)$, (e.g. $BSSN$), to be transformed to a new `gamlss` distribution, (e.g. $logitBSSN$), with range $(0,1)$.

10.3.2 Logit distribution, inflated at 0 and 1

An inflated logit distribution is suitable for a proportion response variable on $0 \leq Y \leq 1$, that includes both 0 and 1. An inflated logit distribution is a mixture of a logit distribution for $0 < Y < 1$ and a Bernoulli distribution for Y at 0 or 1. The model includes three components: a discrete value 0 with probability p_0, a discrete value 1 with probability p_1 and a logit distribution on the unit interval $(0,1)$ with probability $(1-p_0-p_1)$. For a general four parameter logit distribution, $logitD(\mu, \sigma, \nu, \tau)$, then the inflated logit distribution is denoted $Y \sim InflogitD(\mu, \sigma, \nu, \tau, \xi_0, \xi_1)$

with mixed (continuous-discrete) probability (density) function given by

$$f_Y(y|\mu,\sigma,\nu,\tau,\xi_0,\xi_1) = \begin{cases} p_0 & \text{if } y = 0 \\ p_1 & \text{if } y = 1 \\ (1-p_0-p_1) f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0 < y < 1 \end{cases} \tag{10.2}$$

for $0 \leq y \leq 1$, where $W \sim logitD(\mu,\sigma,\nu,\tau)$ has a $logitD$ distribution, where $0 < p_0 < 1$, $0 < p_1 < 1$ and $0 < p_0 + p_1 < 1$. The parameters ξ_0 and ξ_1, are related to p_0 and p_1 by $\xi_0 = p_0/p_2$, $\xi_1 = p_1/p_2$, where $p_2 = 1 - p_0 - p_1$, so $\xi_0 > 0$ and $\xi_1 > 0$. Hence $p_0 = \xi_0/(1+\xi_0+\xi_1)$ and $p_1 = \xi_1/(1+\xi_0+\xi_1)$. For example if $W \sim logitBSSN(\mu,\sigma,\nu,\tau)$ then Y has a inflated $logitBSSN$ distribution $Y \sim InflogitBSSN(\mu,\sigma,\nu,\tau,\xi_0,\xi_1)$ with $-\infty < \mu < \infty$ and $\sigma > 0$, $\nu > 0$, $\tau > 0$, $\xi_0 > 0$, and $\xi_1 > 0$.

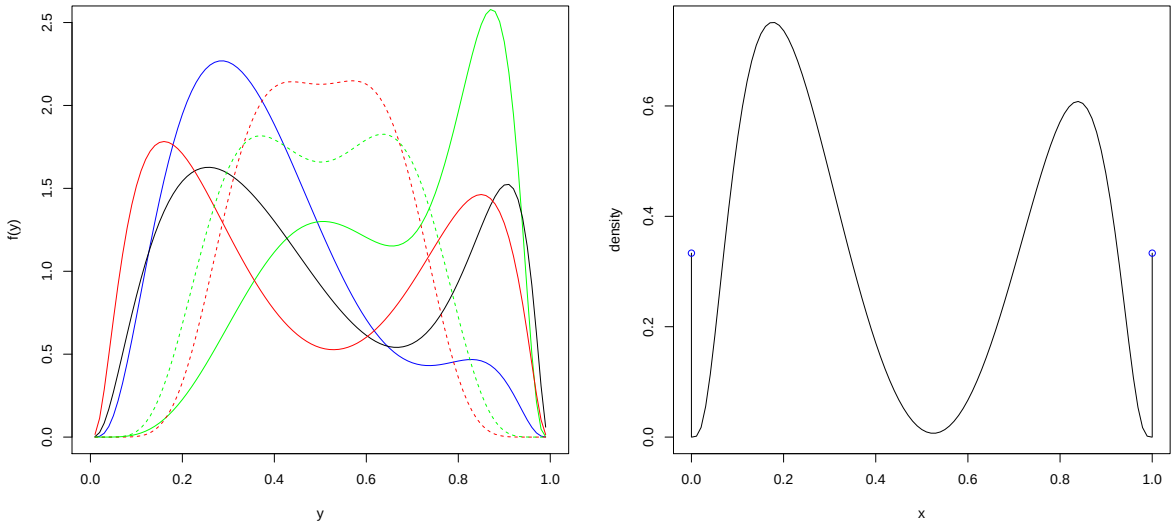


Fig. 10.3 PDF of lositBSSN and InflogitBSSN

For $Y \sim InflogitBSSN(\mu,\sigma,\nu,\tau,\xi_0,\xi_1)$ default link functions relate the parameters $(\mu,\sigma,\nu,\tau,\xi_0,\xi_1)$ to the predictors $(\eta_1,\eta_2,\eta_3,\eta_4,\eta_5,\eta_6)$, i.e.

$$\begin{aligned}
\mu &= \eta_1 \\
\log\sigma &= \eta_2 \\
\log\nu &= \eta_3 \\
\log\tau &= \eta_4 \\
\log(p_0/p_2) = \log(\xi_0) &= \eta_5 \\
\log(p_1/p_2) = \log(\xi_1) &= \eta_6.
\end{aligned}$$

The dependence of the predictors of the parameters (i.e. η_1 to η_6) on explanatory variables may be linear, nonlinear, non-parametric smooth, regression trees or neural network models.

Model (10.2) can be fitted by fitting two models: a $logitD(\mu,\sigma,\nu,\tau)$ distribution for $0<Y<1$, together with a multinomial distribution with three levels, denoted by $MN3(\xi_0,\xi_1)$ in the gamlss.dist package, for a response factor Y_1 given by:

$$Y_1 = \begin{cases} 0 & \text{if } Y=0 \\ 1 & \text{if } Y=1 \\ 2 & \text{if } 0<Y<1 \end{cases} \tag{10.3}$$

i.e.

$$p(Y_1=y_1) = \begin{cases} p_0 & \text{if } y_1=0 \\ p_1 & \text{if } y_1=1 \\ 1-p_0-p_1 & \text{if } y_1=2 \end{cases} \tag{10.4}$$

where $\xi_0 = p_0/p_2$ and $\xi_1 = p_1/p_2$ and $p_2 = 1-p_0-p_1$, giving $\xi_0>0$ and $\xi_1>1$ Alternatively model (10.2) can be fitted more easily using a new function `gamlssInf0to1()` in the new package `gamlss.inf` described in chapter 7. The log likelihood function for the $InflogitD$ model (10.2) is equal to the sum of the log likelihood functions of the $logitD$ model and the multinomial ($MN3$) model (10.4). Hence the parameter sets (μ,σ,ν,τ) and (ξ_0,ξ_1) are 'information' orthogonal.

The inflated logit distributions (i.e. $InflogitD$) have the advantage of extra flexibility, in that the probabilities of Y at 0 and 1 are modelled independently of the distribution on (0,1), (i.e. $logitD$), but with the cost of introducing extra parameters (ξ_0, ξ_1) into the model.

10.3.3 Inflated logit distribution with global adjustment

Note that the logit transformation is sensitive to response variable Y values very close to 0 or 1. To avoid this problem it may be necessary for values of Y close to 0 or 1 to be adjusted. The typical local adjustment approach is applied only to the boundary Y values of 0 or 1, which are adjusted to c and $1-c$ respectively, for a small value c, prior to fitting the transformed regression (e.g. Qi and Zhao (2011) and Altman and Kalotay (2014). Qi and Zhao (2011) has pointed out that the result of transformation regression are very sensitive to the value of the adjustment factor c.

Li et al. (2014) propose an alternative global adjustment approach specifically to adjust all Y values from [0,1] to variable Y' with values in range $(c, 1-c)$ prior to fitting the transformation regression. The global adjustment is achieved through the following equation $Y' = c + (1-2c)Y$, where c is a predetermined adjustment factor. Li et al. (2014) conducted an investigation of c values ranging from 10^{-11} to 0.45 and found optimal value of $c = 0.1$ for their Y variable.

However our investigation shows that the global adjustment proposed by Li et al. (2014) failed to take account of the mode close to 0 and 1 of the Y variable and consequently led to poor model performance. To resolve this issue an alternative adjustment is made to the values close to 1 prior to fitting the inflated logit model given by $Y' = Y(\text{ if } Y \leq 1-c) + 1(\text{ if } Y > 1-c)$, where c is a predetermined adjustment factor. An ad hoc approach is taken to selecting c by investigating the QQ plot of the residuals from the fitted inflated model with different values of the adjustment factor c. The residual plot was found to improve dramatically as the c value increases and a value of $c = 0.1$ was selected, which is consistent with the results found by Li et al. (2014).

10.4 Generalized Tobit model

The original Tobit model for a response variable Y on $[0,1]$ assumes that the response follows a normal distribution censored below 0 and above 1, Tobin (1958).

The generalised Tobit model on [0,1] requires censoring below 0 and above 1 of a flexible model distribution on $(-\infty,\infty)$ for its positive probabilities at 0 and 1. Censoring refers to the transformation of observations outside the limiting interval to the border values, Hoff (2007). Here the values in the model distribution below 0 and above 1 are transformed to 0 and 1 respectively.

Let $Z \sim D(\mu,\sigma,\nu,\tau)$ be a flexible uncensored distribution on $(-\infty,\infty)$. Let $Y \sim Dic(\mu,\sigma,\nu,\tau)$ be the corresponding distribution left censored below 0 and right censored above 1 (called interval censoring, ic) with resulting range $[0,1]$. Then

$$Y = \begin{cases} 0 & \text{if } Z \leq 0 \\ Z & \text{if } 0 \leq Z \leq 1 \\ 1 & \text{if } Z \geq 1. \end{cases}$$

Hence the (mixed continuous-discrete) probability (density) function of Y is given by

$$f_Y(y) = \begin{cases} P(Z \leq 0) & \text{if } y = 0 \\ f_Z(y) & \text{if } 0 < y < 1 \\ P(Z \geq 1) & \text{if } y = 1 \end{cases} \tag{10.5}$$

for $0 \leq y \leq 1$. In principle D can be any distribution on $(-\infty,\infty)$, for example the four parameter *SEP*, *SST* or *SHASHo* distribution. Interval censoring is achieved using gamlss function gen.cens() in the gamlss package `gamlss.cens`.

In the generalised Tobit models the probabilities of Y at 0 and 1 are directly related to the distribution between 0 and 1 and so are less flexible, but the model is more concise (i.e. parsimonious) in that it has two less parameters. Also the Tobit model is not so sensitive to values of Y very close to 0 or 1.

10.4.1 Inflated truncated censored model

The model assumes that Y has a distribution (truncated below 0 and right censored above 1) inflated at 0, defined by $Y \sim InfDtrrc(\mu, \sigma, \nu, \tau)$ with probability density function defined by

$$f_Y(y|\mu,\sigma,\nu,\tau,p_0) = \begin{cases} p_0 & \text{if } y = 0 \\ (1-p_0) f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0 < y < 1 \\ (1-p_0) P(W > 1) & \text{if } y = 1 \end{cases} \tag{10.6}$$

where $W \sim \text{Dtr}(\mu, \sigma, \nu, \tau)$ has a distribution $D(\mu, \sigma, \nu, \tau)$ which is left truncated below 0, and $Dtrrc(\mu, \sigma, \nu, \tau)$ is a left truncated right censored distribution.

10.4.2 Model selection

For each model fitted each distribution parameter was modelled additively using P-splines for sMOB and hrate and factors for $ORIGIN_{YR}$ and $DEFAULT_{YR}$. Table 10.2 shows the models together with the degrees of freedom (df) used in the model and their fitted global deviance (GD), Akaike information criteron (AIC), Akaike (1974), and Schwarz Bayesian Criteron (SBC), Schwarz (1978), and generalized AIC (GAIC) with penalty $k = 4$ for each parameter in the model, where -12556 was substracted from each value of global deviance, AIC, SBC and GAIC in the table to make the comparison clearer. The logitBSSNInf0to1 model is identified as the best model, based on having the lowest values of AIC, SBC and $GAIC(k = 4)$.

Table 10.2 In-sample model section criterion

Models	df	GD	AIC	SBC	GAIC (k=4)
logitBSSNInf0to1	146	0	292	1306	579
BEINF	104	5754	5962	6684	6170
Tobit (NOic)	87	18277	18452	19059	18626
GenTobit (BSSNic)	83	18337	18504	19082	18670

As an alternative method of model selection, 10-fold cross validation results for each of the models were obtained. The typical choice of k is 5 or 10 in k-fold cross validation Friedman et al. (2001). 10-fold cross validation primarily relies on randomly partitioning of the data (i.e. n cases) into 10 sub samples of approximately equal size $(n/10)$. Each subsample successively plays the role of validation sample. Table 10.3 shows each of the 10 folds using $(10-1)=9$ folds for training and the remaining one for validation.

Table 10.3 k-fold cross validation

k	Y	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\nu}$	$\hat{\tau}$
$k=1$	y_{11}	$\hat{\mu_{11}}$	$\hat{\sigma_{11}}$	$\hat{\nu_{11}}$	$\hat{\tau_{11}}$
	y_{12}	$\hat{\mu_{12}}$	$\hat{\sigma_{12}}$	$\hat{\nu_{12}}$	$\hat{\tau_{12}}$
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	y_{1n}	$\hat{\mu_{1n}}$	$\hat{\sigma_{1n}}$	$\hat{\nu_{1n}}$	$\hat{\tau_{1n}}$
$k=2$	y_{21}	$\hat{\mu_{21}}$	$\hat{\sigma_{21}}$	$\hat{\nu_{21}}$	$\hat{\tau_{21}}$
	.	.	.	.	.
	y_{2n}	$\hat{\mu_{2n}}$	$\hat{\sigma_{2n}}$	$\hat{\nu_{2n}}$	$\hat{\tau_{2n}}$
.	.	.	.	.	.
.	.	.	.	.	.
$k=10$	y_{1k}	$\hat{\mu_{1k}}$	$\hat{\sigma_{1k}}$	$\hat{\nu_{1k}}$	$\hat{\tau_{1k}}$
	.	.	.	.	.
	.	.	.	.	.
	y_{nk}	$\hat{\mu_{nk}}$	$\hat{\sigma_{nk}}$	$\hat{\nu_{nk}}$	$\hat{\tau_{nk}}$

Therefore the cross validation (CV) global deviance is defined by the following equation,

$$CV = -2\sum_{k=1}^{10}\sum_{i=1}^{n_k}\log f_Y(y_{ki}|\hat{\mu}_{ki},\hat{\sigma}_{ki},\hat{\nu}_{ki},\hat{\tau}_{ki})$$

where n_k is the sample size of the k^{th} subsample for $k = 1,2,\cdot\cdot,10)$

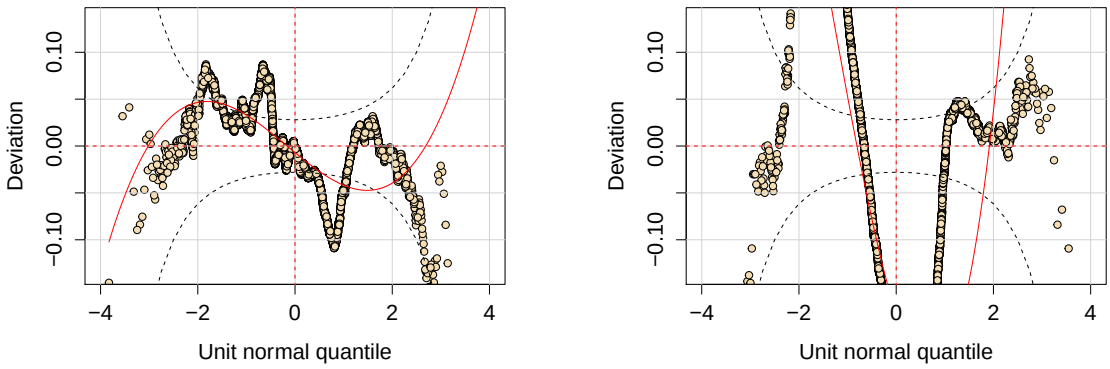
Table 10.4 Cross validation

Models	Terms	CV(Deviance)
logitBSSN	pb()	-18446.07
logitBSSN	linear	-17831.23
logitSST	linear	-16175.82
BE	linear	-12374.06

Table 10.4 shows the cross validated (CV) global deviance for the different models fitted to SEVERITY values between 0 and 1. According to Table 10.4, the inflated logitBSSN model (logitBSSNInf0to1) is selected, as it has the smallest cross validated global deviance.

10.4.3 Residuals of the fitted model

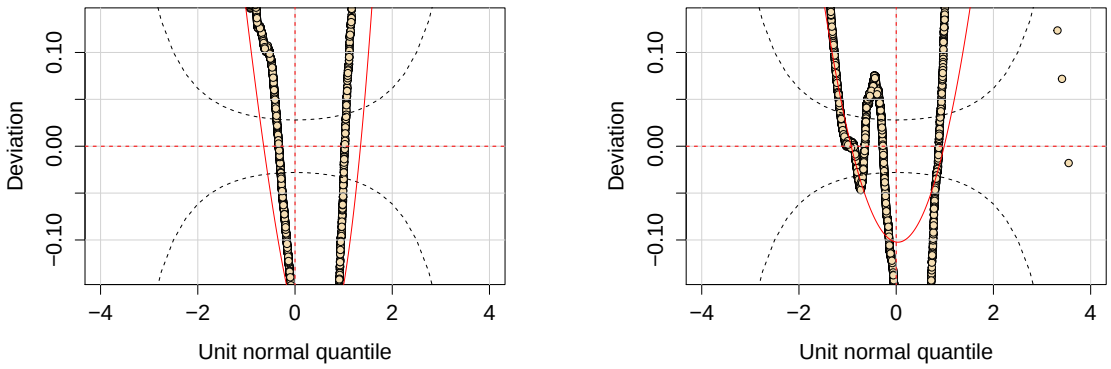
Figure 10.4 and 10.5 show the worm plots of the fitted logitBSSNInf0to1, BEINF, Tobit and genralized Tobit (BSSNic) models respectively. Based on the worm plots the proposed inflated logitBSSN model fits much better than all the other models for the loss given default data set.



(a) worm plot(logitBSSNInf0to1)

(b) Worm plot (BEINF)

Fig. 10.4 Worm plot of logitBSSNInf0to1 and BEINF



(a) Worm plot (Tobit)

(b) Worm plot (TobitBSSN)

Fig. 10.5 Worm plot of Tobit and GenTobit model

10.4.4 Fitted distribution

Figure 10.6 shows the fitted distribution of Y for the fitted logitBSSNInf0to1 model using six data cases (i.e. 100, 500, 1000, 2000, 3000, 5000). The corresponding values of explanatory variables are given in Table 10.5

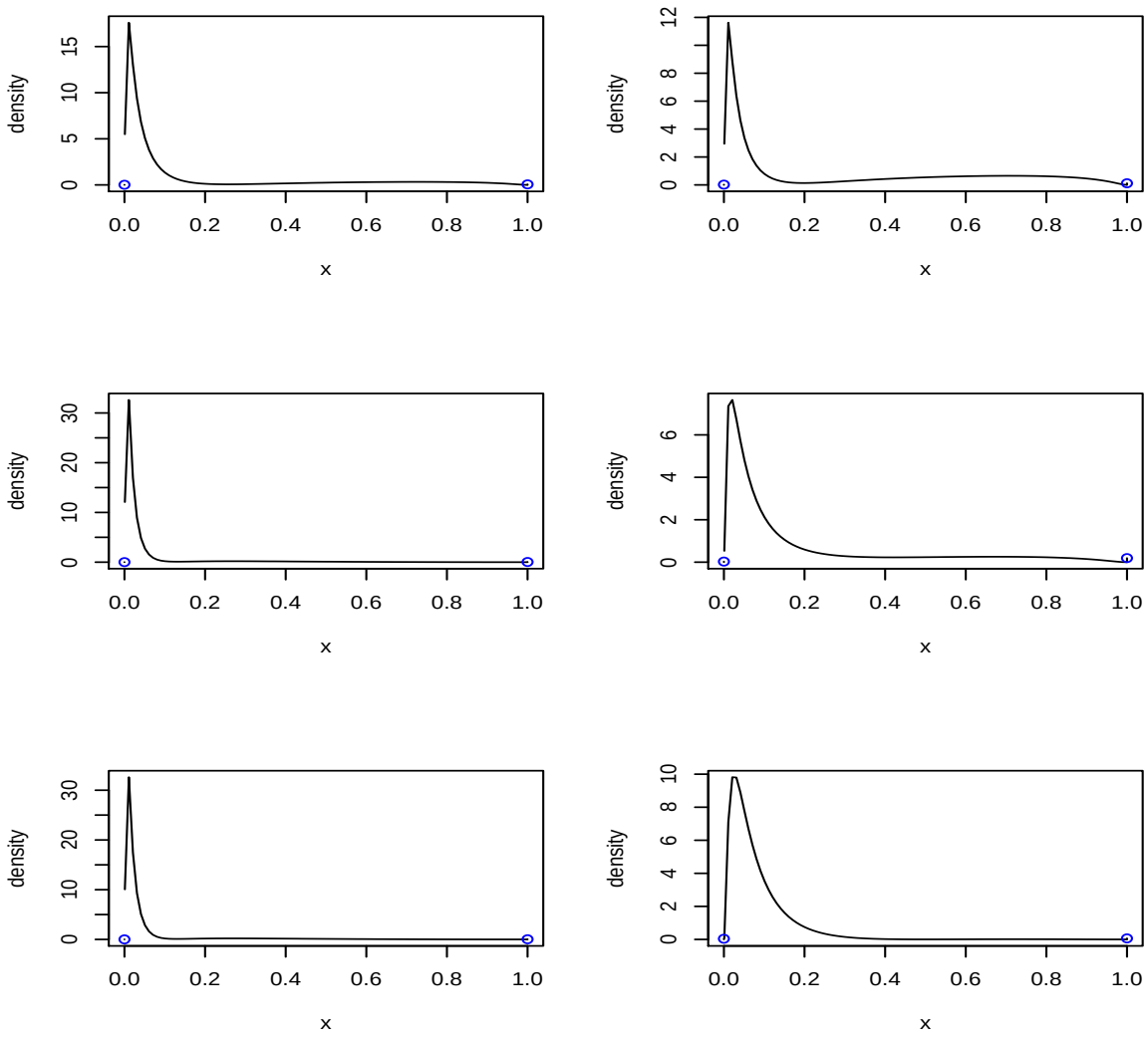


Fig. 10.6 Fitted distribution of the InflogitBSSN for six data cases

Table 10.5 Corresponding values of the explanatory variables for fitted distributions

	sMOB	hrate	$ORIGIN_{YR}$	$DEFAULT_{YR}$
1	9.746794	0.454601	2000	2007
2	7.81025	0.357487	2001	2006
3	8.124038	0.309862	2000	2005
4	5	0.443965	1998	2000
5	8.062258	0.300251	2000	2005
6	10.86278	0.334559	1994	2000

10.5 Conclusion

This chapter proposes an inflated logit distribution and a generalized type I Tobit model for loss given default (LGD). Both models use the four parameter bi-modal skew symmetric normal (BSSN) distribution (used in order to model the bimodality of the distribution of LGD). Flexible nonparametric P-splines were used to model the parameters of the distribution of the response variable using covariates. The dependence of each of the parameters of the two proposed models on explanatory variable can be replaced with linear, regression trees or neural network models. The proposed inflated logit `GAMLSS` model was compared with the beta inflated model. Based on the AIC and SBC criterion, the study concluded that the inflated logitBSSN provided the best fit to the loss given default data.

Chapter 11

Conclusion and Future developments

The thesis investigated methods for handling a proportion response variable and presents new classes of models for a bounded proportion. The proposed models primarily address the bounded nature, skewness and bimodality of the proportion dependent variable. The proposed inflated GAMLSS and generalized Tobit GAMLSS model are discussed in 5 and 6 respectively. The proposed inflated GAMLSS model can be fitted using new package `gamlssinf` in R. The generalized Tobit GAMLSS model can be fitted using the package `gamlss.cens`. The proposed models can be applied to data from a variety of disciplines. This chapter outlines the main contributions of the thesis in modelling a proportion response variable and proposes some directions for future developments.

11.1 Originality of inflated GAMLSS model

This thesis makes a number of original contributions to the area of modelling continuous and bounded data by developing a class of univariate inflated GAMLSS models which extend the flexibility of the beta inflated model by (Hoff (2007), Ospina and Ferrari (2010), Cook et al. (2008)).

The inflated GAMLSS model is a mixed continuous-discrete distribution model which allows modelling of any or all the parameters of a distribution (up to four parameters for the continuous component and two extra parameters for the discrete components) using explanatory terms (e.g. linear and/or non-linear smoothing terms in explanatory variables).

For example the inflated GAMLSS model extends the beta inflated model by including two extra parameters in the continuous component for modelling the skewness and kurtosis or bimodality. Unlike the beta inflated model this thesis offers a comprehensive framework for the statistical analysis of the continuous data observed on the standard unit interval (0,1) with point masses at 0 and/or 1.

The inflated GAMLSS model is a general class of regression model for modelling a continuous proportion with discrete boundary values at zero and/or one. A method of estimating the parameters of the inflated GAMLSS model is explained in chapter 5. This research also explained the randomized quantile residuals of the mixed continuous-discrete random variable for the inflated GAMLSS model in chapter 5.

11.2 Important applications of inflated GAMLSS model

The inflated GAMLSS model allows the use of a flexible explicit or transformed (e.g. logit or truncated) distribution on (0,1) for the response variable including highly skew and/or kurtotic distributions, for example the explicit distribution on (0,1) GB1(μ, σ, ν, τ) of McDonald and Xu (1995), logit or truncated version of the power exponential distribution of Nelson (1991), Johnson's SU of Johnson et al. (1994), the sinh-arcsinh of Jones and Pewsey (2009), the skewed t family Fernandez and Steel (1998*b*) and bimodal skew symmetric normal of Hassan and El-Bassiouni (2016) distributions.

The use of flexible explicit or transformed (i.e. logit or truncated) distributions on (0,1) together with a binomial or multinomial model allows fitting of the inflated GAMLSS model using penalised

likelihood estimation algorithm discussed in chapter 5 and uses of variety of diagnostic tools for model checking and selection (see for example chapters 8, 9 and 10).

The inflated GAMLSS model also expands the centile estimation techniques by allowing a more flexible model for the data containing ones and zeroes.

In chapter 8 an inflated GAMLSS at 1 model was used to model a response variable (spirometric lung function on (0,1]) using an explanatory variable . Two other cases of inflated GAMLSS model (inflated GAMLSS at 0 and inflated GAMLSS at 0 and 1) are used to analyse the response variable in two different data sets, i.e. PASS scheme data with a response variable on [0,1) and loss given default data set with a response variable on [0,1] respectively. Using three different data sets helps justify the usefulness of the model. In all the three cases, the inflated GAMLSS model outperformed other previous models.

11.3 Originality of the generalized Tobit GAMLSS model

In addition to the inflated GAMLSS model, this thesis also developed a new class of model, the generalized Tobit GAMLSS model, for the bounded proportion response variable.The generalized Tobit GAMLSS model allows modelling all the parameters of the distribution of the latent variable V using linear or non-linear terms and/or smoothing terms in explanatory variables.

The genralized Tobit GAMLSS model extends the Tobit model in terms of the number of parameters and their flexibility. The generalized Tobit GAMLSS model includes two more parameters than the Tobit model to model the conditional skewness and/or kurtosis and bimodality of the response variable.

This thesis also describes the randomized quantile residuals of the generalized Tobit GAMLSS model to assess the overall adequacy of the model (see chapter 6). A method of estimating the parameters of the model is also described and explained in chapter 6.

The generalized Tobit GAMLSS for a proportion response variable comprises three special cases: censored below zero, censored above one and interval censored below 0 and above 1. Applications of the three sub-models of generalized Tobit model together with popular models currently in the literature are shown in chapters 8, 9 and 10 respectively. In all the cases generalized Tobit GAMLSS model performed better than the other popular previous models.

11.4 Limitations and future developments

The inflated logit distributions have the advantage of extra flexibility, in that the probability of Y equals 0 or 1 is modelled independently of the distribution (0,1) but at the cost of introducing extra parameters. Note that logit transformation is very sensitive to values close to 0 and 1.

In the generalized Tobit GAMLSS model, the probability of Y equals 0 or 1 is directly related to the distribution between 0 and 1 and so is less flexible but the model is more concise, because it has two less distribution parameters than the inflated GAMLSS model. The generalized Tobit GAMLSS model was not adequate in modelling a response variable containing a large number of values close to 0 and 1 (in the loss given default data example in chapter 10).

11.4.1 Future developments

Further work includes a model consisting of censoring above one (as introduced in the generalized Tobit GAMLSS model) and additional zeroes (as addressed in inflated GAMLSS model).

The distribution of the response variable can also be extended to a general distribution on $(0, \infty)$, which is left shifted and interval censored to [0,1]. The inflated GAMLSS model and generalized Tobit GAMLSS models can be extended to include spatial terms in the model for a bounded proportion response variable.

The comparison of the inflated GAMLSS model and generalized Tobit GAMLSS model with different additive terms (e.g. neural network, decision tree) is also unexplored.

The above is important further work that has been considered by the author, but due to the time constraint, all those important directions have been proposed for future development of the inflated GAMLSS and the generalized Tobit GAMLSS models.

References

Aitchison, J. (1982), 'The statistical analysis of compositional data', *Journal of the Royal Statistical Society. Series B (Methodological)* **44**(2), 139–177.

Aitchison, J. and Begg, C. B. (1976), 'Statistical diagnosis when basic cases are not classified with certainty', *Biometrika* **63**(1), 1–12.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**(6), 716–723.

Akaike, H. (1983), 'Information measures and model selection', *Bulletin of the International Statistical Institute* **50**, 277–290.

Altman, E. I. and Kalotay, E. A. (2014), 'Ultimate recovery mixtures', *Journal of Banking & Finance* **40**, 116–129.

Arabmazar, A. and Schmidt, P. (1982), 'An investigation of the robustness of the tobit estimator to non-normality', *Econometrica: Journal of the Econometric Society* pp. 1055–1063.

Aranda-Ordaz, F. J. (1981), 'On two families of transformations to additivity for binary response data', *Biometrika* pp. 357–363.

Baran, S. and Nemoda, D. (2016), 'Censored and shifted gamma distribution based emos model for probabilistic quantitative precipitation forecasting', *Environmetrics* .

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991), 'Some parametric models on the simplex', *Journal of Multivariate Analysis* **39**(1), 106–116.

Bastos, J. A. (2010), 'Forecasting bank loans loss-given-default', *Journal of Banking & Finance* **34**(10), 2510–2517.

Bellotti, T. and Crook, J. (2012), 'Loss given default models incorporating macroeconomic variables for credit cards', *International Journal of Forecasting* **28**(1), 171–182.

Cole, T. J. and Green, P. J. (1992), 'Smoothing reference centile curves: the lms method and penalized likelihood', *Statistics in Medicine.* **11**, 1305–1319.

Cole, T., Stanojevic, S., Stocks, J., Coates, A., Hankinson, J. and Wade, A. (2009), 'Age-and size-related reference ranges: A case study of spirometry through childhood and adulthood', *Statistics in Medicine* **28**(5), 880–898.

Cook, D. O., Kieschnick, R. and McCullough, B. (2008), 'Regression analysis of proportions in finance with self selection', *Journal of Empirical Finance* **15**(5), 860 – 867.

Cox, D. R. and Snell, E. . J. (1968), 'A general definition of residuals', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 248–275.

Cox, D. R. and Snell, E. J. (1989), *Analysis of binary data*, Vol. 32, CRC Press.

D'agostino, R. B., Belanger, A. and D'Agostino Jr, R. B. (1990), 'A suggestion for using powerful and informative tests of normality', *The American Statistician* **44**(4), 316–321.

Dawson, J. A., Kamlin, C. O. F., Vento, M., Wong, C., Cole, T. J., Donath, S. M., Davis, P. G. and Morley, C. J. (2010), 'Defining the reference range for oxygen saturation for infants after birth', *Pediatrics* **125**(6), e1340–e1347.

de Boor, C. (2001), *A Practical Guide to Splines. Revised edition*, Springer-Verlag, New-York.

Dunn, P. K. and Smyth, G. K. (1996), 'Randomized quantile residuals', *Journal of Computational and Graphical Statistics* **5**, 236–244.

Eilers, P. H. C. and Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties (with comments and rejoinder)', *Statist. Sci* **11**, 89–121.

Fernandez, C., Osiewalski, J. and Steel, M. F. (1995), 'Modeling and inference with υ-spherical distributions', *Journal of the American Statistical Association* **90**(432), 1331–1340.

Fernández, C. and Steel, M. F. (1998*a*), 'On bayesian modeling of fat tails and skewness', *Journal of the American Statistical Association* **93**(441), 359–371.

Fernandez, C. and Steel, M. F. J. (1998*b*), 'On bayesian modelling of fat tails and skewness', *J. Am. Statist. Ass.* **93**, 359–371.

Ferrari, S. and Cribari-Neto, F. (2004), 'Beta regression for modelling rates and proportions', *Journal of Applied Statistics* **31**(7), 799–815.

Friedman, J., Hastie, T. and Tibshirani, R. (2001), *The elements of statistical learning*, Vol. 1, Springer series in statistics Springer, Berlin.

Galvis, D. M., Bandyopadhyay, D. and Lachos, V. H. (2014), 'Augmented mixed beta regression models for periodontal proportion data', *Statistics in medicine* **33**(21), 3759–3771.

Girão, L. C., Lopes, A. V., Tabarelli, M. and Bruna, E. M. (2007), 'Changes in tree reproductive traits reduce functional diversity in a fragmented atlantic forest landscape', *PLoS One* **2**(9), e908.

Gupta, A. K. and Nadarajah, S. (2004), *Handbook of beta distribution and its applications*, CRC press.

Gupton, G. M. and Stein, R. M. (2005), 'Dynamic prediction of lgd modeling methodology'.

Gürtler, M. and Hibbeln, M. (2013), 'Improvements in loss given default forecasts for bank loans', *Journal of Banking & Finance* **37**(7), 2354–2366.

Hassan, M. and El-Bassiouni, M. (2016), 'Bimodal skew-symmetric normal distribution', *Communications in Statistics-Theory and Methods* **45**(5), 1527–1541.

Hoff, A. (2007), 'Second stage dea: Comparison of approaches for modelling the {DEA} score', *European Journal of Operational Research* **181**(1), 425 – 435.

Hossain, A., Rigby, R., Stasinopoulos, M. and Enea, M. (2016a), 'Centile estimation for a proportion response variable', *Statistics in medicine* **35**(6), 895–904.

Hossain, A., Rigby, R., Stasinopoulos, M. and Enea, M. (2016b), A flexible approach for modelling a proportion response variable, *in* J.-F. Dupuy and J. Josse, eds, 'Proceeding of the 31st international workshop on statistical modelling', Vol. 1, INSA, pp. 127–132.

Hu, Y.-T. and Perraudin, W. (2002), 'The dependence of recovery rates and defaults', *Birkbeck College* .

Hunger, M., Baumert, J. and Holle, R. (2011), 'Analysis of sf-6d index data: is beta regression appropriate?', *Value in Health* **14**(5), 759–767.

Johnson, D. (1997), 'The triangular distribution as a proxy for the beta distribution in risk analysis', *Journal of the Royal Statistical Society. Series D (The Statistician)* **46**(3), pp. 387–398.
URL: *http://www.jstor.org/stable/2988573*

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994), *Continuous Univariate Distributions, Volume I, 2nd edn.*, Wiley, New York.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1995), *Continuous Univariate Distributions, Volume II, 2nd edn.*, Wiley, New York.

Jones, M. (2009), 'Kumaraswamy's distribution: A beta-type distribution with some tractability advantages', *Statistical Methodology* **6**(1), 70–81.

Jones, M. C. and Pewsey, A. (2009), 'Sinh-arcsinh distributions', *Biometrika* **96**, 761–780.

Kieschnick, R. and McCullough, B. D. (2003), 'Regression analysis of variates observed on (0, 1): percentages, proportions and fractions', *Statistical modelling* **3**(3), 193–213.

Kim, Y.-J. and Gu, C. (2004), 'Smoothing spline gaussian regression: more scalable computation via efficient approximation', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(2), 337–356.

Kumaraswamy, P. (1980), 'A generalized probability density function for double-bounded random processes', *Journal of Hydrology* **46**(1), 79–88.

Lang, S. and Brezger, A. (2004), 'Bayesian p-splines', *Journal of computational and graphical statistics* **13**(1), 183–212.

Li, P., Qi, M., Zhang, X. and Zhao, X. (2014), Further investigation of parametric loss given default modeling, Technical report, Citeseer.

Libby, D. L. and Novick, M. R. (1982), 'Multivariate generalized beta distributions with applications to utility assessment', *Journal of Educational and Behavioral Statistics* **7**(4), 271–294.

Maddala, G. (1983), 'Qualitative and limited dependent variable models in econometrics'.

Maddala, G. S. and Nelson, F. D. (1975), 'Specification errors in limited dependent variable models'.

McDonald, J. B. and Xu, Y. J. (1995), 'A generalisation of the beta distribution with applications', *Journal of Econometrics* **66**, 133–152.

Nelson, D. B. (1991), 'Conditional heteroskedasticity in asset returns: a new approach', *Econometrica* **59**, 347–370.

Nishii, R. and Tanaka, S. (2013), 'Modeling and inference of forest coverage ratio using zero-one inflated distributions with spatial dependence', *Environmental and ecological statistics* **20**(2), 315–336.

Ospina, R. and Ferrari, S. L. (2012), 'A general class of zero-or-one inflated beta regression models', *Computational Statistics & Data Analysis* **56**(6), 1609–1623.

Ospina, R. and Ferrari, S. L. P. (2010), 'Inflated beta distributions', *Statistical Papers* **23**, 111–126.

Pace, L. and Salvan, A. (1997), *Principles of statistical inference: from a Neo-Fisherian perspective*, Vol. 4, World scientific.

Papke, L. E. and Wooldridge, J. (1996), 'Econometric methods for fractional response variables with an application to 401 (k) plan participation rates', *Journal of Applied Econometrics* **11**(6), 619–632.

Pregibon, D. (1980), 'Goodness of link tests for generalized linear models', *Applied statistics* pp. 15–14.

Qi, M. and Zhao, X. (2011), 'Comparison of modeling methods for loss given default', *Journal of Banking & Finance* **35**(11), 2842–2855.

Quanjer, P. H., Stanojevic, S., Cole, T. J., Baur, X., Hall, G. L., Culver, B. H., Enright, P. L., Hankinson, J. L., Ip, M. S., 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**(6), 1324–1343.

Quanjer, P., Stanojevic, S., Stocks, J., Hall, G., Prasad, K., Cole, T., Rosenthal, M., Perez-Padilla, R., Hankinson, J., Falaschetti, E. et al. (2010), 'Changes in the fev1/fvc ratio during childhood and adolescence: an intercontinental study', *European Respiratory Journal* **36**(6), 1391–1399.

Quanjer, P., Stocks, J., Polgar, G., Wise, M., Karlberg, J. and Borsboom, G. (1989), 'Compilation of reference values for lung function measurements in children.', *The European respiratory journal. Supplement* **4**, 184S–261S.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rigby, R. A. and Stasinopoulos, D. M. (2004), 'Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution', *Statistics in Medicine* **23**, 3053–3076.

Rigby, R. A. and Stasinopoulos, D. M. (2005), 'Generalized additive models for location, scale and shape, (with discussion)', *Appl. Statist.* **54**, 507–554.

Rigby, R. A. and Stasinopoulos, D. M. (2006), 'Using the Box-Cox *t* distribution in gamlss to model skewness and kurtosis', *Statistical Modelling* **6**, 209–229.

Rigby, R. A. and Stasinopoulos, D. M. (2013), 'Automatic smoothing parameter selection in gamlss with an application to centile estimation', *Statistical methods in medical research* **23**, 318–332.

Rosett, R. N. and Nelson, F. D. (1975), 'Estimation of the two-limit probit regression model', *Econometrica: Journal of the Econometric Society* pp. 141–146.

Royston, P. and Altman, D. G. (1994), 'Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion)', *Appl. Statist.* **43**, 429–467.

Royston, P. and Wright, E. M. (2000), 'Goodness-of-fit statistics for age-specific reference intervals', *Statistics in Medicine* **19**, 2943–2962.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003), *Semiparametric regression*, Cambridge university press.

Scallan, A., Gilchrist, R. and Green, M. (1984), 'Fitting parametric link functions in generalised linear models', *Computational Statistics & Data Analysis* **2**(1), 37–49.

Schmid, M., Hothorn, T., Maloney, K. O., Weller, D. E. and Potapov, S. (2011), 'Geoadditive regression modeling of stream biological condition', *Environmental and Ecological Statistics* **18**(4), 709–733.

Schmid, M., Wickler, F., Maloney, K. O., Mitchell, R., Fenske, N. and Mayr, A. (2013), 'Boosted beta regression', *PloS one* **8**(4), e61623.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Ann. Statist.* **6**, 461–464.

Siddiqi, N. A. and Zhang, M. (2004), 'A general methodology for modeling loss given default', *RMA JOURNAL* **86**(8), 92–95.

Sigrist, F. and Stahel, W. A. (2010), 'Using the censored gamma distribution for modeling fractional response variables with an application to loss given default', *arXiv preprint arXiv:1011.1796* .

Smith, P. L. (1979), 'Splines as a useful and convenient statistical tool', *Amer. Statist.* **33**, 57–62.

Song, P. X.-K., Qiu, Z. and Tan, M. (2004), 'Modelling heterogeneous dispersion in marginal models for longitudinal proportional data', *Biometrical journal* **46**(5), 540–553.

Stanojevic, S., Wade, A., Cole, T. J., Lum, S., Custovic, A., Silverman, M., Hall, G. L., Welsh, L., Kirkby, J., Nystad, W. et al. (2009), 'Spirometry centile charts for young caucasian children: The asthma uk collaborative initiative.', *American journal of respiratory and critical care medicine* **180**(6), 547–552.

Stasinopoulos, D. M. and Rigby, R. A. (1992), 'Detecting break points in generalised linear models.', *Comp. Stat. Data Anal.* **13**, 461–471.

Stasinopoulos, D. M., Rigby, R., Voudouris, V., Heller, G. and Bastiani, F. (2015), 'Flexible regression and smoothing, the gamlss packages in r'.
URL: *http://www.gamlss.org/wp-content/uploads/2015/07/FlexibleRegressionAndSmoothingDraft-1.pdf*

Stasinopoulos, D. and Rigby, R. (2007), 'Generalized additive models for location scale and shape (GAMLSS) in R', *Journal of Statistical Software* **23**(7), 1–46.

Takeshi, A. (1984), 'Tobit models: A survey', *Journal of econometrics* **24**(1-2), 3–61.

Takeshi, A. (1985), *Advanced econometrics*, Harvard university press.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica* **26**(1), 24–36.

Topp, C. W. and Leone, F. C. (1955), 'A family of j-shaped frequency functions', *Journal of the American Statistical Association* **50**(269), 209–219.

Trenkler, G. (1996), 'Continuous univariate distributions : N.L. Johnson, S. Kotz and N. Balakrishnan Vol. 1, 2nd Edition. John Wiley, New York, 1994, pp. xix + 756', *Computational Statistics & Data Analysis* **21**(1), 119–119.

van Buuren, S. and Fredriks, M. (2001), 'Worm plot: a simple diagnostic device for modelling growth reference curves', *Statistics in Medicine* **20**, 1259–1277.

Vicari, D., Van Dorp, J. R. and Kotz, S. (2008), 'Two-sided generalized topp and leone (ts-gtl) distributions', *Journal of Applied Statistics* **35**(10), 1115–1129.

Warton, D. I. and Hui, F. K. (2011), 'The arcsine is asinine: the analysis of proportions in ecology', *Ecology* **92**(1), 3–10.

Wedderburn, R. W. M. (1974), 'Quasi-likelihood functions, generalised linear models and the gauss-newton method', *Biometrika* **61**, 439–447.

WHO, M. G. R. S. (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, M. G. R. S. (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.

Wood, S. N. (2001), 'mgcv: Gams and generalised ridge regression for r', *R News* **1**, 20–25.

Wurtz, D., Chalabi, Y. and Luksan, L. (2006), 'Parameter estimation of arma models with garch/aparch errors an r and splus software implementation', *Journal of Statistical Software* pp. 28–33.

Yashkir, O. and Yashkir, Y. (2013), 'Loss given default modeling: a comparative analysis', *The Journal of Risk Model Validation* **7**(1), 25.

Appendix A

R code for application on lung function data

```
library(gamlss)     # loading gamlss package
#-------------------------------------------------------------------
# files needed
source('Inf0to1-11-5-15.R') # the inflated distribution functions
                            # gamlss.Inf0to1
source('qstats-TEST.R')     # a modified verion of Q.stats

source('centTobit.R')       # a modified version of centiles
source("wptwimT.R")         # a modified version of the wptwim()
source("rqres1.R")          # a modified version of rqres()

#Data
alldata<-read.table("fev1fvc_all_best.txt", header=T, na.strings="NA")
dim(alldata)
# select out just the males
                    d1m<-subset(alldata,sex==1)
                    dim(d1m)
                    d2m<-na.omit(d1m)
                    dim(d2m)
                    #> dim(d2m)
                    #[1] 3164
                    d2m$f <- d2m$fev1fvc
                    d2m$lht<- log(d2m$ht)
                    k1<- 6
```

```
# The response variable is d2m$f

# FITTING THE MODELS
-----------------------------------------------------
                  LMS
-----------------------------------------------------
m1 <- gamlss(f~ pb(lht,method = "GAIC",k=k1),
             sigma.formula=~pb(lht,method="GAIC", k=k1),
             nu.formula=~pb(lht,method="GAIC", k=k1),
             family = BCCGo,data = d2m,n.cyc=100)
#m1<- rqres1(m1, setseed=351, save.resid=TRUE)

#-------------------------------------------------------------------------------
                   Beta Inflated
#-------------------------------------------------------------------------------
m2 <- gamlss( f ~ pb(lht,method= "GAIC", k=k1),
             sigma.formula=~pb(lht,method= "GAIC", k=k1),
             nu.formula=~pb(lht,method= "GAIC", k=k1),
             data=d2m, family=BEINF1)
# this will recalculate the residuals with set seed for comparisons
m2<- rqres1(m2, setseed=351, save.resid=TRUE)
#-----------------------------------------------------------------
#          logitSST
#-------------------------------------------------------------------------------
gen.Family("SST", "logit")
# choosing the smoothing parameters with GAIC and penalty 6
m3 <- gamlssInf0to1( y=f,mu.formula=~ pb(lht, method="GAIC", k=k1),
                     sigma.formula=~pb(lht,method="GAIC", k=k1),
                     nu.formula=~pb(lht,method="GAIC", k=k1),
                     tau.formula=~pb(lht,method="GAIC", k=k1),
                     xi1.formula=~pb(lht,method="GAIC", k=k1),
                     data=d2m, family=logitSST,
                     trace = T, setseed=351,
                     gd.tol=100)
# choosing the smoothing parameters with GAIC and penalty 2
m31 <- gamlssInf0to1( y=f,mu.formula=~ pb(lht, method="GAIC", k=2),
                    sigma.formula=~pb(lht,method="GAIC", k=2),
                    nu.formula=~pb(lht,method="GAIC", k=2),
                    tau.formula=~pb(lht,method="GAIC", k=2),
                    xi1.formula=~pb(lht,method="GAIC", k=2),
                    data=d2m, family=logitSST,
                    trace = T, setseed=351,
                    gd.tol=100)
# choosing the smoothing parameters with GAIC and penalty 8.059592
m32 <- gamlssInf0to1( y=f,mu.formula=~ pb(lht, method="GAIC", k=log(3164)),
                    sigma.formula=~pb(lht,method="GAIC", k=log(3164)),
                    nu.formula=~pb(lht,method="GAIC", k=log(3164)),
```

```
                tau.formula=~pb(lht,method="GAIC", k=log(3164)),
                xi1.formula=~pb(lht,method="GAIC", k=log(3164)),
                data=d2m, family=logitSST,
                trace = T, setseed=351,
                gd.tol=100)

# TOBIT type model
#-----------------------------------------------

library(survival)
# creating the Y variable
d2m$fs<- Surv(d2m$f, d2m$f!=1, type="right")
# creating the distribution
library(gamlss.cens)
gen.cens("BCCGo", type="right")
gen.cens("NO", type="right")
# fitting the  model
# choosing the smoothing parameters with GAIC and penalty 6
m4 <- gamlss( fs ~ pb(lht,method="GAIC", k=k1),
              sigma.formula=~pb(lht,method="GAIC", k=k1),
              data=d2m, family=NOrc)
m4<- rqres1(m4, setseed=351, save.resid=TRUE)

m5 <- gamlss( fs ~ pb(lht,method="GAIC", k=k1),
      sigma.formula=~pb(lht,method="GAIC", k=k1),
      nu.formula=~pb(lht,method="GAIC", k=k1),
      data=d2m, family=BCCGorc)
m5<- rqres1(m5, setseed=351, save.resid=TRUE)
# choosing the smoothing parameters with GAIC and penalty 2
m51 <- gamlss( fs ~ pb(lht,method="GAIC", k=2),
              sigma.formula=~pb(lht,method="GAIC", k=2),
              nu.formula=~pb(lht,method="GAIC", k=2),
              data=d2m, family=BCCGorc)
# choosing the smoothing parameters with GAIC and penalty  8.059592
m52 <- gamlss( fs ~ pb(lht,method="GAIC", k=log(3164)),
              sigma.formula=~pb(lht,method="GAIC", k=log(3164)),
              nu.formula=~pb(lht,method="GAIC", k=log(3164)),
              data=d2m, family=BCCGorc)
m52<- rqres1(m52, setseed=351, save.resid=TRUE)


#-------------------------------------------------
               Centile curves
-------------------------------------------------
centiles(m1,xvar= d2m$ht, cent=c(2,10,25,50,75,90,98),
ylab="FEV1/FVC", xlab=" height",ylim = range(d2m$f),
legend=F, main="(a) LMS")
```

```
centiles(m2,xvar= d2m$ht, cent=c(2,10,25,50,75,90,98),
ylab="FEV1/FVC", xlab=" height",ylim = range(d2m$f),
legend=F,main="(b) BEINF1")

centiles.Inf0to1(m3,xvar= d2m$ht, cent=c(2,10,25,50,75,90,98),
ylab="FEV1/FVC", xlab=" height",ylim = range(d2m$f),
legend=F,main="(c) Inf. logitSST")

centiles.T(m5,xvar= d2m$ht, cent=c(2,10,25,50,75,90,98),
         ylab="FEV1/FVC", xlab=" height", ylim = range(d2m$f),
         legend=F,main="(d) Gen. Tobit")
centiles.Inf0to1(m3,xvar= d2m$ht, cent=c(2,10,25,50,75,90,98),
                 ylab="FEV1/FVC", xlab=" height",ylim = range(d2m$f),
                 legend=F, main=" Inf. logitSST")
centiles.T(m5,xvar= d2m$ht, cent=c(2,10,25,50,75,90,98),
           ylab="FEV1/FVC", xlab=" height", ylim = range(d2m$f),
           legend=F,main=" Gen. Tobit")

#-------------------------------------------------
#                 GAIC (comparing models)
#-------------------------------------------------
GAIC(m1,m2,m3,m4,k=6)
#-------------------------------------------------
#          Q-stat
#-------------------------------------------------
qstat.m1<- Q.stats(m1,xcut.points= NULL,n.inter=16,
     xvar=d2m$ht, digits.xvar=3), title (" (a) LMS")
qstat.m2<- Q.stats(m2,xcut.points= NULL,n.inter=16,
   xvar=d2m$ht,digits.xvar=3),title (" (b) BEINF1")
qstat.m3<- Q.stats(m3,xcut.points= NULL,n.inter=16,
  xvar=d2m$ht,digits.xvar=3),title("(c) Inf.logitSST")
qstat.m5<- Q.stats(m5,xcut.points= NULL,n.inter=16,
      xvar=d2m$ht,digits.xvar=3)
title("(d) Gen.Tobit")

-----------------------------------------------------
Twin worm plot
-----------------------------------------------------
wp.twinT(m1, m2, xvar= d2m$lht,  xvar.column=2,
       n.inter=16, show.given= FALSE ,ylim.worm = 1,
       cex = .5,  col1= "black", col2 = "grey",
       warning= FALSE,  pch=21)

wp.twinT(m3, m5, xvar= d2m$lht,  xvar.column=2,
    n.inter=16, show.given= FALSE ,ylim.worm = 1,
    cex = .5,  col1= "black", col2 = "grey",
```

```
    warning= FALSE,  pch=21)
-----------------------------------------------------------
Fitted (prediced)distribution of d2m$f
-----------------------------------------------------------
########## logitSSTInf1###############

# inflogitSST
#logit model
d3m<- subset(d2m, d2m$f!=1)
# fitting continuous part of the model for prediction
m3logit <- gamlss(f~ pb(lht, method="GAIC", k=k1),
            sigma.formula=~pb(lht,method="GAIC", k=k1),
            nu.formula=~pb(lht,method="GAIC", k=k1),
            tau.formula=~pb(lht,method="GAIC", k=k1),
            data=d3m, family=logitSST,trace = T,
            setseed=351, gd.tol=100)

newd2m3<-data.frame(lht=c(4.094345,4.382027,4.60517,4.787492,
                          4.941642,5.075174))

pm3logit<-predictAll(m3logit, newdata=newd2m3, use.weights=TRUE)

# BI model & Prediction (fitting binary model for discrete part)
d2m$f0<- ifelse(d2m$f==1, 1 && d2m$f!=1, 0)
m3BI <- gamlss(f0~ pb(lht, method="GAIC", k=k1),sigma.formula=~pb(lht,
        method="GAIC", k=k1),data=d2m, family=BI)
pm3BI<-predictAll(m3BI, newdata=newd2m3, use.weights=TRUE)


# Generating inflated logit distribution
gen.Inf0to1(family = "logitSST", type.of.Inflation = "One")
x11(width = 6, height = 8, pointsize = 12)

op<-par(mfrow=c(3,2))
plotlogitSSTInf1(mu=3.248380,  sigma=1.2483134, nu= 1.352479,
                 tau=1491.913581, xi1=0.965443806)
plotlogitSSTInf1(mu=2.958775,  sigma=1.0265913, nu= 1.508604,
                 tau=80.938310, xi1=0.753542382)
plotlogitSSTInf1(mu=2.520048,  sigma=0.8749899, nu= 1.802432,
                 tau=9.352586, xi1= 0.351104068)
plotlogitSSTInf1(mu=2.017535,  sigma=0.7643921, nu= 1.693014,
                 tau=3.712501, xi1=0.095667080)
plotlogitSSTInf1(mu=1.739814 , sigma=0.6799430, nu= 1.402725,
                 tau=3.717171, xi1=0.019854368)
plotlogitSSTInf1(mu=1.541542 , sigma=0.6132293, nu= 1.457597,
                 tau=4.542058, xi1=0.004406673)
par(op)
```

```
#############gen.Tobit model#############################
m5all2<-predictAll(m5,newdata=newd2m3,use.weights=TRUE)
```

Appendix B

R code for the application on PASS scheme data

```
library(gamlss) # loading gamlss package
library(gamlssinf) # loading gamlssinf package
x<-read.table("C:/Users/user/Desktop/Pass scheme.csv", header=TRUE, sep=",")
x$AMM1<-x$AMM/100
x$AMM2<-(1-x$AMM1)# transform Y variable
#------------------------------------------------------------------------
                     Inflated logit ST3 at 0
#------------------------------------------------------------------------

mST3 <- gamlssInf0to1( y=AMM2,mu.formula=~ pb(PA, method="GAIC", k=k1),
       sigma.formula=~pb(PA,method="GAIC", k=k1),
       nu.formula=~pb(PA,method="GAIC", k=k1),
       tau.formula=~pb(PA,method="GAIC", k=k1),
       xi1.formula=~pb(PA,method="GAIC", k=k1),
       data=x, family=logitST3,trace = T,gd.tol=100)
############### BEINF1 ###################
mBEINF <- gamlss(AMM2~ pb(PA, method="GAIC", k=k1),
          sigma.formula=~pb(PA,method="GAIC", k=k1),
          nu.formula=~pb(PA,method="GAIC", k=k1),
          family=BEINF1,data=pass,trace = T,gd.tol=100)

#-------------------------------------------------
    Generalized Tobit GAMLSS model (censored at 0)
#-------------------------------------------------
library(survival)
library(gamlss.cens)
```

```
x$AMM2<-(1-x$AMM1)
#creating Surv object
x$AMM3<- Surv(x$AMM2, x$AMM2!=1, type="right")

# Generating right censored distribution

gen.cens("BCT", type="right")
#------------------------ Tobit model------------
mNorc<-gamlss( pass$AMM3 ~ pb(pass$PA,method="GAIC", k=6),
             sigma.formula=~pb(pass$PA,method="GAIC", k=6),
             data=pass, family=NOrc)
#-----------------------------------------------------------
#BCTrc
m3 <- gamlss( x$AMM3 ~ pb(x$PA,method="GAIC", k=6),
            sigma.formula=~pb(x$PA,method="GAIC", k=6),
            nu.formula=~pb(x$PA,method="GAIC", k=6),
            tau.formula=~pb(x$PA,method="GAIC", k=6),
            data=x, family=BCTrc)
#-------------------------------------------------
                          GAIC(k=6)
#-------------------------------------------------
GAIC(m3,mST3,mBEINF, mNOrc)

#-----------------------------------------------------------
#-----------Fitted centile curves---------
op <- par(mfrow=c(2,2))
centiles(mNorc,xvar= pass$PA, cent=c(2,10,25,50,75,90,98),
         ylab="AMM", xlab="Attendance", ylim = range(pass$AMM3),
         legend=F, main="(a) Tobit (NOrc)")

centiles(mBEINF,xvar= pass$PA, cent=c(2,10,25,50,75,90,98),
         ylab="AMM", xlab=" Attendance",ylim = range(pass$AMM2),
         legend=F,main="(b) BEINF1")

centiles.Inf0to1(mST3,xvar= pass$PA, cent=c(2,10,25,50,75,90,98),
                 ylab="AMM", xlab=" Attendance",ylim = range(pass$AMM2),
                 legend=F,main="(c) logitST3Inf1")

centiles.T(mBCTrc,xvar= pass$PA, cent=c(2,10,25,50,75,90,98),
           ylab="AMM", xlab=" Attendance", ylim = range(pass$AMM3),
           legend=F,main="(d) Gen.Tobit (BCTrc)")
par(op)

#-----------------------------------------------------------
worm plot and twin worm plot
#-----------------------------------------------------------
wp.twin(mST3, m3, xvar = x$PA, xvar.column = 2, n.inter = 2,
```

```
      show.given = FALSE, ylim.worm = 0.5, line = FALSE, cex = 1,
      col1 = "black", col2 = "orange", warnings = FALSE)

wp.twin(mBEINF, mNOrc, xvar = x$PA, xvar.column = 2, n.inter = 10,
  show.given = FALSE, ylim.worm = 0.5, line = FALSE, cex = 1,
  col1 = "black", col2 = "orange", warnings = FALSE)

##############Prediction###################
#-------------using logitST3Inf1-----------------
# LOGIT sT3 # fitting model for continuous part of the data
p2m<- subset(pass, AMM2!=1)
m9logitST3 <- gamlss(AMM2~ pb(PA, method="GAIC", k=k1),
            sigma.formula=~pb(PA,method="GAIC", k=k1),
            nu.formula=~pb(PA,method="GAIC", k=k1),
            tau.formula=~pb(PA,method="GAIC", k=k1),data=p2m,
            family=logitST3,trace = T,gd.tol=100)
#### Predict using 7 data cases
newP2m<-data.frame(PA=c(0.1,0.3,0.5,0.7,0.8,0.9))# ne2w data

pm9logit<-predictAll(m9logitST3, newdata=newP2m, use.weights=TRUE)

########## BI model# fitting binary model
pass$AMM0<- ifelse(pass$AMM2==1, 1 && pass$AMM2!=1, 0)

mbi9<- gamlss(AMM0~ pb(PA, method="GAIC", k=k1),
                sigma.formula=~pb(PA, method="GAIC", k=k1),
                data=pass, family=BI)

pm9BI<-predictAll(mbi9, newdata=newP2m, use.weights= TRUE)

library(gamlss.inf)
gen.Family(family="ST3", type="logit")
# generating inflated distribution function
gen.Inf0to1(family="logitST3",  type.of.Inflation="One")
x11(width = 6, height = 8, pointsize = 12)
op<-par(mfrow=c(3,2))
plotlogitST3Inf1 (mu= -0.4140361, sigma=0.7656729, nu= 1.428967,
                  tau= 8.767638, xi1= 0.15066964, xlab= "y", ylab="fy")
plotlogitST3Inf1 (mu= -0.4320217, sigma=0.6981526, nu= 1.387453,
                  tau= 8.264562, xi1= 0.09686596)
plotlogitST3Inf1 (mu= -0.4500436, sigma=0.6365360, nu= 1.347151,
                  tau= 7.790544, xi1= 0.06290156)
plotlogitST3Inf1 (mu= -0.4681511, sigma=0.5803502, nu= 1.308046,
                  tau= 7.343815, xi1= 0.04123668)
plotlogitST3Inf1 (mu= -0.4772450, sigma=0.5541458, nu=1.288914,
                 tau= 7.130169, xi1= 0.03333156)
plotlogitST3Inf1 (mu= -0.4863611, sigma=0.5291258, nu= 1.270052,
```

```
                tau= 6.922741, xi1= 0.02689532)
par(op)
###    Prediction using BCTrc model

newpass<-data.frame(PA=c(0.1,0.3,0.5,0.7,0.8,0.9))# ne2w data
predict(mBCTrc1, what="mu", newdata=newpass)
predictAll(mBCTrc, newdata= newpass)
```

Appendix C

R code for the application on LGD data

[0,1]

```
library(lattice)    # loading lattice plotting package
library(gamlss)     # loading gamlss package
library(gamlss.inf) # loading gamlssinf package
source('/Users/user/Desktop/Bi-modal-dist/GAMLSS_BSSN.txt', chdir = TRUE)

#  DATA
da<-read.table('/Users/user/Dropbox/inflateddistributions/paper/
               LGD.Data.csv',header = TRUE, sep=",")

#-----------------------------------------------------------
           Global adjustment
#-------------------------------------------------
da1<- subset(da,da$SEVERITY!=0&da$SEVERITY!= 1)
da2<- subset(da,da$SEVERITY>=0&da$SEVERITY<= 0.5)
da3<- subset(da,da$SEVERITY>0.5&da$SEVERITY<= 1)
da4<-(da1[da1$SEVERITY<=0.98&da1$SEVERITY>=0.02,])
#-------factor----
da$FORIGIN_YR <- factor(da$ORIGIN_YR)
da$FDEFAULT_YR <- factor(da$DEFAULT_YR)

da$sMOB <- sqrt(da9$MOB)
#---------Data processing for generalised tobit
library(survival)
library(gamlss.cens)
#creating Survival object
v1<- ifelse(da$SEVERITY==0,NA,da$SEVERITY)
```

```
v2<- ifelse(da$SEVERITY==1,NA,da$SEVERITY)
Y1 <- Surv(v1,v2, type="interval2")
Y2<- Surv(da$L, type="interval2")
da$Y1 <- Surv(v1,v2, type="interval2")

# generating interval censored BSSN distribution
gen.cens("BSSN",type="interval")

#-----BETA--Inflated----------------------
mBEINF<- gamlss(SEVERITY~pb(sMOB,df=7)+pb(hrate,df=7)
    + FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~pb(hrate,df=1)+pb(sMOB,df=1) ,
    nu.formula=~pb(hrate,df=9) + FORIGIN_YR
    + FDEFAULT_YR+pb(sMOB,df=9),
    tau.formula=~pb(hrate,df=1)
    + pb(sMOB,df=1),data = da,family=BEINF,
    n.cyc=200, gd.tol=Inf,c.crit=0.1)

mlogitSST<- gamlss(SEVERITY~pb(sMOB,df=7)+pb(hrate,df=7)
    + FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~pb(hrate,df=1)+pb(sMOB,df=1) ,
    nu.formula=~pb(hrate,df=9)
    + FORIGIN_YR + FDEFAULT_YR+pb(sMOB,df=9),
    tau.formula=~pb(hrate,df=1)+ pb(sMOB,df=1),
    data = da4,family=logitSST,
    n.cyc=200, gd.tol=Inf,c.crit=0.1)

mlogitBSSN<- gamlss(SEVERITY~pb(sMOB,df=7)+pb(hrate,df=7)
    + FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~pb(hrate,df=1)+pb(sMOB,df=1) ,
    nu.formula=~pb(hrate,df=9)
    + FORIGIN_YR + FDEFAULT_YR+pb(sMOB,df=9),
    tau.formula=~pb(hrate,df=1)+ pb(sMOB,df=1),
    data = da4,family=logitBSSN,
    n.cyc=200, gd.tol=Inf,c.crit=0.1)

-----------------------------------------------------
# logitBSSN model with gamlssinfto1
-----------------------------------------------------
m2logitBSSN <- gamlssInf0to1(y=SEVERITY,
    mu.formula= ~ pb(MOB)+ pb(hrate)
    + FORIGIN_YR + FDEFAULT_YR,
    sigma.formula = ~pb(sMOB) + pb(hrate)
    + FORIGIN_YR,
    nu.formula = ~ pb(MOB)+ pb(hrate)
    + FORIGIN_YR+ FDEFAULT_YR,
    tau.formula = ~~pb(sMOB) + pb(hrate),
```

```
    xi1.formula= ~scci+hrate + FORIGIN_YR,
    xi0.formula= ~hrate + FORIGIN_YR,
    data=da, family=logitBSSN,
    trace = T, n.cyc=200,
    gd.tol=Inf,c.crit=0.1)
#--------Tobit and genTobit-----
Tobit<- gamlss(Y1~pb(sMOB,df=7)+pb(hrate,df=7)
    + FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~pb(hrate,df=1)+pb(sMOB,df=1) ,
    data = da,family=NOic,n.cyc=200,
    gd.tol=Inf,c.crit=0.1)

TobitBSSN<- gamlss(Y1~pb(sMOB,df=7)+pb(hrate,df=7)
    +FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~pb(hrate,df=1)+pb(sMOB,df=1) ,
    nu.formula=~pb(hrate,df=9)+pb(sMOB,df=9)
    + FORIGIN_YR + FDEFAULT_YR,
    tau.formula=~pb(hrate)+ pb(sMOB),data = da,
    family=BSSNic,n.cyc=200,
    gd.tol=Inf,c.crit=0.1)

#-----------------------------------------------
comparing model using gen cross validation(GCV)
#-----------------------------------------------
set.seed(123)
rand <- sample(2, 6738, replace=TRUE, prob=c(0.6,0.4))

rand1<- sample(10, 6738, replace=TRUE)
rand2<- sample(6, 6738, replace=TRUE)
table(rand)/6738
olddata<-da4[rand==1,] # training data
newdata<-da4[rand==2,] # validation data

g1bssn <- gamlssCV(SEVERITY~sMOB+hrate+FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~hrate+sMOB, nu.formula=~1,
    tau.formula=~1, data=da4,
    family=logitBSSN,rand=rand1)

g1SST <- gamlssCV(SEVERITY~sMOB+hrate+FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~hrate+sMOB, nu.formula=~1,
    tau.formula=~1, data = da4,family=logitSST,
    rand=rand1, parallel="no", ncpus=nC)

g1BE <- gamlssCV(SEVERITY~sMOB+hrate+FORIGIN_YR+FDEFAULT_YR,
    sigma.formula=~hrate+sMOB, data = da4,
    family=BE,rand=rand1,parallel="no", ncpus=nC)
```

```
############## Fitted Distribution of SEVERITY for 6 new cases#####
# generating inflated logit BSSN distribution

gen.Inf0to1(family="logitBSSN",  type.of.Inflation="Zero&One")

x11(width = 7, height = 8, pointsize = 12)
op<-par(mfrow=c(3,2))
plotlogitBSSNInf0to1(mu=fitted(m2logitBSSN)[100],
sigma=fitted(m2logitBSSN, "sigma")[100],
nu= fitted(m2logitBSSN, "nu")[100], tau=fitted(m2logitBSSN, "tau")[100],
fitted(m2logitBSSN, "xi0")[100],fitted(m2logitBSSN, "xi1")[100])
plotlogitBSSNInf0to1(mu=fitted(m2logitBSSN)[500],
sigma=fitted(m2logitBSSN, "sigma")[500],
nu= fitted(m2logitBSSN, "nu")[500], tau=fitted(m2logitBSSN, "tau")[500],
fitted(m2logitBSSN, "xi0")[500], fitted(m2logitBSSN, "xi1")[500])
plotlogitBSSNInf0to1(mu=fitted(m2logitBSSN)[1000],
sigma=fitted(m2logitBSSN, "sigma")[1000],
nu= fitted(m2logitBSSN, "nu")[1000], tau=fitted(m2logitBSSN, "tau")[1000],
fitted(m2logitBSSN, "xi0")[1000], fitted(m2logitBSSN, "xi1")[1000])
plotlogitBSSNInf0to1(mu=fitted(m2logitBSSN)[2000],
sigma=fitted(m2logitBSSN, "sigma")[2000],
nu= fitted(m2logitBSSN, "nu")[2000], tau=fitted(m2logitBSSN, "tau")[2000],
fitted(m2logitBSSN, "xi0")[2000],fitted(m2logitBSSN, "xi1")[2000])
plotlogitBSSNInf0to1(mu=fitted(m2logitBSSN)[3000],
sigma=fitted(m2logitBSSN, "sigma")[3000],
nu= fitted(m2logitBSSN, "nu")[3000], tau=fitted(m2logitBSSN, "tau")[3000],
fitted(m2logitBSSN, "xi0")[3000],fitted(m2logitBSSN, "xi1")[3000])
plotlogitBSSNInf0to1(mu=fitted(m2logitBSSN)[5000],
sigma=fitted(m2logitBSSN, "sigma")[5000],
nu= fitted(m2logitBSSN, "nu")[5000], tau=fitted(m2logitBSSN, "tau")[5000],
fitted(m2logitBSSN, "xi0")[5000],fitted(m2logitBSSN, "xi1")[5000])
```

Appendix D

Box-Cox t distribution

Box-Cox t distribution of Rigby and Stasinopoulos (2006) is denoted as $Y \sim BCT(\mu, \sigma, \nu, \tau)$ and probability density function is given by

$$\begin{aligned} f_Y(y|\mu,\sigma,\nu,\tau) &= f_Z(z)|\tfrac{dz}{dy}| \\ &= \tfrac{y^{\nu-1}}{\mu^{\nu}\sigma} f_Z(z) \end{aligned}$$

where $f_Z(z)$ is the truncated t probability density function of Z given by

$$Z = \begin{cases} \frac{1}{\sigma\nu}[(\frac{Y}{\mu})^{\nu} - 1], & \text{if } \nu \leq 0 \\ \frac{1}{\sigma}\log(\frac{Y}{\mu}), & \text{if } \nu = 0 \end{cases}$$

for $0 < Y < \infty$, $\mu > 0$, $\sigma > 0$ and $\tau > 0$ (treated as continuous parameter).

Appendix E

R code for bi modal skew symmetric normal distribution

```
BSSN <- function (mu.link="identity", sigma.link="log",
        nu.link ="identity", tau.link="log")
{
        mstats <- checklink(   "mu.link",
                "Bimodal skew-symmetric normal",
                substitute(mu.link),
                c("inverse", "log", "identity", "own"))
        dstats <- checklink("sigma.link",
                "Bimodal skew-symmetric normal",
                substitute(sigma.link),
                c("inverse", "log", "identity", "own"))
        vstats <- checklink(   "nu.link",
                "Bimodal skew-symmetric normal",
                substitute(nu.link),
                c("inverse", "log", "identity", "own"))
        tstats <- checklink(  "tau.link",
                "Bimodal skew-symmetric normal",
                substitute(tau.link),
                c("inverse", "log", "identity", "own"))
        structure(
                list(family = c("BSSN", "Bimodal skew-symmetric normal"),
                     parameters = list(mu=TRUE, sigma=TRUE,
                     nu=TRUE, tau=TRUE),
                     nopar = 4,
                     type = "Continuous",
                     mu.link = as.character(substitute(mu.link)),
```

```
      sigma.link = as.character(substitute(sigma.link)),
      nu.link = as.character(substitute(nu.link)),
      tau.link = as.character(substitute(tau.link)),
      mu.linkfun = mstats$linkfun,
      sigma.linkfun = dstats$linkfun,
      nu.linkfun = vstats$linkfun,
      tau.linkfun = tstats$linkfun,
      mu.linkinv = mstats$linkinv,
      sigma.linkinv = dstats$linkinv,
      nu.linkinv = vstats$linkinv,
      tau.linkinv = tstats$linkinv,
      mu.dr = mstats$mu.eta,
      sigma.dr = dstats$mu.eta,
      nu.dr = vstats$mu.eta,
      tau.dr = tstats$mu.eta,
# The second derivatives of this log likelihood
  function with respect to mu
      dldm = function(y, mu, sigma, nu, tau)
      {     lambda <- nu-mu
            theta <- tau + (lambda)^2
            gamma <- 1 + 2*sigma*theta
            c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
            dldm<- 4*sigma*lambda/gamma
            + 2*sigma*(y-mu)

            dldm
      },
#The second derivatives of this log likelihood
 function with respect to mu
      d2ldm2 = function(y, mu, sigma, nu, tau)
      {   lambda <- nu-mu
          theta <- tau + (lambda)^2
          gamma <- 1 + 2*sigma*theta
          c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
          dldm<- 4*sigma*lambda/gamma
                 + 2*sigma*(y-mu)
              d2ldm2<- - (dldm) * (dldm)

              d2ldm2
      },

#The first derivatives of this log likelihood
 function with respect to sigma
      dldd= function(y,mu, sigma, nu, tau)
      {       lambda <- nu-mu
              theta <- tau + (lambda)^2
              gamma <- 1 + 2*sigma*theta
```

```
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldd <- 3/(2*sigma) - 2*theta/gamma
                - (y-mu)^2
                dldd
        },
    #The second derivatives of this log likelihood
     function with respect to sigma
        d2ldd2 = function(y, mu, sigma, nu, tau)
        {       lambda <- nu-mu
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldd <- 3/(2*sigma) - 2*theta/gamma
                - (y-mu)^2
                d2ldd2<- -(dldd) * (dldd)

                d2ldd2
        },
 #The first derivatives of this log likelihood
  function with respect to nu
        dldv= function(y,mu, sigma, nu, tau)
        {       lambda <- nu-mu
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldv <- -4*sigma*lambda/gamma
                - 2*(y-nu)/(tau
                + (y-nu)^2)
                dldv
        },
#The second derivatives of this log likelihood
  function with respect to nu
        d2ldv2 = function(y, mu, sigma, nu, tau)
        {       lambda <- nu-mu
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldv <- -4*sigma*lambda/gamma - 2*(y-nu)/(tau
                + (y-nu)^2)
                d2ldv2<- -(dldv) * (dldv)

                d2ldv2
        },
 #The first derivatives of this log likelihood
   function with respect to tau
        dldt= function(y,mu, sigma, nu, tau)
        {       lambda <- nu-mu
```

```
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldt <- -2*sigma/gamma + 1/(tau
                + (y-nu)^2)
                dldt
        },
#The second derivatives of this log likelihood
 function with respect to tau
        d2ldt2 = function(y, mu, sigma, nu, tau)
        {       lambda <- nu-mu
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldt <- -2*sigma/gamma + 1/(tau
                + (y-nu)^2)
                d2ldt2<- -(dldt) * (dldt)

                d2ldt2
        },

        d2ldmdd = function(y, mu, sigma, nu, tau)
        {       lambda <- nu-mu
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldm<- 4*sigma*lambda/gamma + 2*sigma*(y-mu)
                dldd <- 3/(2*sigma) - 2*theta/gamma
                - (y-mu)^2
                d2ldmdd<- -(dldm) * (dldd)

                d2ldmdd
        },

        d2ldmdv = function(y, mu, sigma, nu, tau)
        {       lambda <- nu-mu
                theta <- tau + (lambda)^2
                gamma <- 1 + 2*sigma*theta
                c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
                dldm<- 4*sigma*lambda/gamma + 2*sigma*(y-mu)
                dldv <- -4*sigma*lambda/gamma - 2*(y-nu)/(tau
                + (y-nu)^2)

                d2ldmdv<- -(dldm) * (dldv)

                d2ldmdv
        },
```

```
d2ldmdt = function(y, mu, sigma, nu, tau)
{       lambda <- nu-mu
        theta <- tau + (lambda)^2
        gamma <- 1 + 2*sigma*theta
        c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
        dldm<- 4*sigma*lambda/gamma + 2*sigma*(y-mu)
        dldt <- -2*sigma/gamma + 1/(tau + (y-nu)^2)

        d2ldmdt<- -(dldm) * (dldt)

        d2ldmdt
},

d2ldddv = function(y, mu, sigma, nu, tau)
{       lambda <- nu-mu
        theta <- tau + (lambda)^2
        gamma <- 1 + 2*sigma*theta
        c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
        dldd <- 3/(2*sigma) - 2*theta/gamma - (y-mu)^2
        dldv <- -4*sigma*lambda/gamma - 2*(y-nu)/(tau
        + (y-nu)^2)
        d2ldddv<- -(dldd) * (dldv)

        d2ldddv
},

d2lddlt = function(y, mu, sigma, nu, tau)
{       lambda <- nu-mu
        theta <- tau + (lambda)^2
        gamma <- 1 + 2*sigma*theta
        c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
        dldd <- 3/(2*sigma) - 2*theta/gamma
        - (y-mu)^2
        dldt <- -2*sigma/gamma + 1/(tau + (y-nu)^2)
        d2lddlt<- -(dldd) * (dldt)

        d2lddlt
},

d2ldvdt = function(y, mu, sigma, nu, tau)
{       lambda <- nu-mu
        theta <- tau + (lambda)^2
        gamma <- 1 + 2*sigma*theta
        c <- 2 *(sigma)^(3/2)/(gamma*(pi)^(1/2))
        dldv <- -4*sigma*lambda/gamma - 2*(y-nu)/(tau
        + (y-nu)^2)
```

```
                        dldt <- -2*sigma/gamma + 1/(tau + (y-nu)^2)
                        d2ldvdt<- -(dldv) * (dldt)

                        d2ldvdt
                    },

                    G.dev.incr  = function(y,mu,sigma,nu,tau,...)
                    {
                        -2*dBSSN(y,mu,sigma,nu,tau,log=TRUE)
                    } ,
                    rqres = expression(
                        rqres(pfun="pBSSN", type="Continuous", y=y,
                        mu=mu, sigma=sigma, nu=nu, tau=tau)) ,
                    mu.initial = expression(mu <- (y+mean(y))/2),
                   sigma.initial = expression(sigma <- rep(0.1, length(y))),
                    nu.initial = expression(nu <- rep(1, length(y))),
                    tau.initial = expression(tau <-rep(1, length(y))),
                    mu.valid = function(mu) TRUE,
                    sigma.valid = function(sigma) all(sigma > 0),
                    nu.valid = function(nu) TRUE,
                    tau.valid = function(tau) all(tau > 0),
                    y.valid = function(y)  TRUE
                ),
                class = c("gamlss.family","family"))
}
#-------------------Probability density function of BSSN--------------------
dBSSN <- function(x, mu = 0, sigma = 1, nu = 1, tau = .5, log = FALSE)
{
        if (any(sigma < 0))  stop(paste("sigma must be positive",
            "\n", ""))
        if (any(tau < 0))  stop(paste("tau must be positive",
             "\n", ""))
        lambda <- nu-mu
        theta <- tau + (lambda)^2
        gamma1 <- 1 + 2*sigma*theta
        c <- 2 *(sigma)^(3/2)/(gamma1*(pi)^(1/2))
        d <- c*(tau+(x-nu)^2)*exp(-sigma*(x-mu)^2)
        loglik <- log(c)+log(tau+(x-nu)^2)- sigma*(x-mu)^2
        if(log==FALSE) ft  <- exp(loglik) else ft <- loglik
        ft
}
#---------------------CDF of BSSN--------------------------
# pfun needs to check
pBSSN <- function(q, mu = 0, sigma = 1, nu = 1, tau = .5,
        lower.tail = TRUE, log.p = FALSE,log=T)
{
        if (any(sigma < 0))  stop(paste("sigma must be positive",
```

```
        "\n", ""))
          if (any(tau < 0))  stop(paste("tau must be positive",
          \n", ""))
       # lambda <- nu-mu
      #  theta <- tau + (lambda)^2
      #  gamma <- 1 + 2*sigma*theta
      #      c <- (2 *(sigma)^(3/2))/(gamma*((pi)^(1/2)))
          ax <- (q + mu-(2*nu))/(1 + (2*sigma)*(tau+(nu-mu)^2))
        #  dNx <- dNO(q,0,1)
        #  pNx <- pNO(q,0,1)
           p <- pNO(q, mu=mu, sigma=sqrt(1/(2*sigma)))
           -ax * dNO(q, mu=mu, sigma=sqrt(1/(2*sigma)))
        if(lower.tail==TRUE) p  <- p else  p <- 1-p
        if(log.p==FALSE) p  <- p else  p <- log(p)
        p
}
#---------------Inverse CDF---------------------------------
# needs to get q function
qBSSN <-  function(p, mu = 0, sigma = 1, nu = 1, tau = .5,
        lower.tail = TRUE, log.p = FALSE)
{
        #---functions-------------------------------------------
        h1 <- function(q)
        {
                pBSSN(q , mu = mu[i], sigma = sigma[i], nu = nu[i],
                    tau = tau[i]) - p[i]
        }
        h <- function(q)
        {
                pBSSN(q , mu = mu[i], sigma = sigma[i], nu = nu[i],
                    tau = tau[i])
        }
        #-------------------------------------------------------------
        #if (any(mu <= 0))  stop(paste("mu must be positive",
        "\n", ""))
        if (any(sigma <= 0))  stop(paste("sigma must be positive",
        "\n", ""))
        if (log.p==TRUE) p <- exp(p) else p <- p
        if (lower.tail==TRUE) p <- p else p <- 1-p
        if (any(p < 0)|any(p > 1))  stop(paste("p must be between 0 and 1",
        "\n", ""))
        lp <-  max(length(p),length(mu),length(sigma),length(nu),
        length(tau))
        p <- rep(p, length = lp)
        sigma <- rep(sigma, length = lp)
        mu <- rep(mu, length = lp)
        nu <- rep(nu, length = lp)
```

```
        tau <- rep(tau, length = lp)
        q <- rep(0,lp)
        for (i in  seq(along=p))
        {
                if (h(mu[i])<p[i])
                {
                        interval <- c(mu[i], mu[i]+sigma[i])
                        j <-2
                        while (h(interval[2]) < p[i])
                        {interval[2]<- mu[i]+j*sigma[i]
                         j<-j+1
                        }
                }
                else
                {
                        interval <-  c(mu[i]-sigma[i], mu[i])
                        j <-2
                        while (h(interval[1]) > p[i])
                        {interval[1]<- mu[i]-j*sigma[i]
                         j<-j+1
                        }
                }
                q[i] <- uniroot(h1, interval)$root
                #interval <- c(.Machine$double.xmin, 20)
        }
        q
}
#---------------------Random function---------------------------------------
rBSSN <- function(n, mu=0, sigma=1, nu=1, tau=.5)
{
        if (any(sigma < 0))  stop(paste("sigma must be positive",
        "\n", ""))
        if (any(tau < 0))  stop(paste("tau must be positive",
        "\n", ""))
        n <- ceiling(n)
        p <- runif(n)
        r <- qBSSN(p,mu=mu,sigma=sigma,nu=nu,tau=tau)
        r
}
```

Appendix F

R code for Spirometric data analysis using two explanatory variables

```
#-----------------------------------------------------------------
The following are the R code for the analysis of lung data using
two explanatory  variables (age and height)

load(file="C:/Users/hossaina/Dropbox/InflatedDistributions
           /new_lung_analysis/d2m.Rdata")

#-----------------------------------------------------------------
# selection of variables for the binary  model
library(gamlss.add)
b0 <- gamlss(y~1, dat=d2m, family=BI)
b1 <- gamlss(y~pb(lht), dat=d2m, family=BI)
b2 <- gamlss(y~pb(lage), dat=d2m, family=BI)
b3 <- gamlss(y~pb(lht)+pb(lage), dat=d2m, family=BI)
b4 <- gamlss(y~ga(~s(lht,lage)), dat=d2m, family=BI)
AIC(b0,b1,b2,b3,b4)
# df      AIC
# b3 4.571516 1584.023
# b2 2.044672 1584.439
# b4 3.000000 1586.253
# b1 4.269015 1635.187
# b0 1.000000 2087.976
#-----------------------------------------------------------
# selection of variables for the the logitSST model
# reduced data
d2mR <- subset(d2m, y==0)
```

```
head(d2mR)
dim(d2mR)
#- generate logitSST
gen.Family("SST", "logit")
# fit null model
# the followin is the selection model for logitSST
# m0 <- gamlss(fev1fvc~1, family=logitSST, data=d2mR)
# m1<- stepGAICAll.A(m0, scope=list(lower=~1,
#            upper=~ lht+lage+pb(lht)+pb(lage)+ga(~s(lht,lage)) ))
 # the
#m1<- stepGAICAll.A(m0, scope=list(lower=~1,
#upper=~ lht+lage+pb(lht)+pb(lage)+ga(~s(lht,lage)) ))
 # ------------------------------------------------
 #   Distribution parameter:  mu
 # Start:  AIC= -6811.69
 # fev1fvc ~ 1
 #
 # Df     AIC
 # + ga(~s(lht, lage)) 17.2128 -7958.7
 # + pb(lage)           6.4977 -7950.1
 # + lage               1.0000 -7856.6
 # + pb(lht)            5.0526 -7770.7
 # + lht                1.0000 -7750.2
 # <none>                      -6811.7
 #
 # Step:  AIC= -7958.73
 # fev1fvc ~ ga(~s(lht, lage))
 #
 # Df     AIC
 # <none>            -7958.7
 # + lht      1.0000 -7956.7
 # + lage     1.0000 -7956.7
 # + pb(lht)  1.0045 -7956.7
 # + pb(lage) 1.0085 -7956.7
 # ------------------------------------------------
 #   Distribution parameter:  sigma
 # Start:  AIC= -7958.73
 # ~1
 #
 # Df     AIC
 # + pb(lage)           8.25483 -8310.6
 # + ga(~s(lht, lage)) 10.10103 -8307.4
 # + lage               0.86574 -8273.8
 # + pb(lht)            3.06881 -8225.4
 # + lht               -2.73408 -8223.1
 # <none>                       -7958.7
 #
```

```
# Step:  AIC= -8310.57
# ~pb(lage)
#
# Df     AIC
# <none>                       -8310.6
# + lht               0.62600 -8309.7
# + pb(lht)           0.62843 -8309.7
# + ga(~s(lht, lage)) 1.62548 -8307.7
# ------------------------------------------------
#   Distribution parameter:  nu
# Start:  AIC= -8310.57
# ~1
#
# Df     AIC
# + pb(lage)          3.25310 -8328.4
# + ga(~s(lht, lage)) 5.12998 -8326.8
# + pb(lht)           4.26102 -8326.4
# + lage              0.24767 -8321.1
# + lht               0.59362 -8317.7
# <none>                       -8310.6
#
# Step:  AIC= -8328.35
# ~pb(lage)
#
# Df     AIC
# <none>                      -8328.4
# + lht               1.0260 -8326.4
# + pb(lht)           1.0270 -8326.4
# + ga(~s(lht, lage)) 2.0235 -8324.4
# ------------------------------------------------
#   Distribution parameter:  tau
# Start:  AIC= -8328.35
# ~1
#
# Df     AIC
# + pb(lage)          3.61785 -8363.2
# + ga(~s(lht, lage)) 5.56324 -8358.7
# + pb(lht)           3.10702 -8353.0
# + lht               1.09025 -8341.0
# + lage              0.91452 -8341.0
# <none>                      -8328.4
#
# Step:  AIC= -8363.17
# ~pb(lage)
#
# Df     AIC
# <none>                       -8363.2
```

```
# + lht                 1.66079 -8361.9
# + ga(~s(lht, lage))  2.65262 -8359.9
# + pb(lht)           -0.50741 -8352.2
# - pb(lage)           3.61785 -8328.4
# ----------------------------------------------------
#   Distribution parameter:  nu
# Start:  AIC= -8363.17
# ~pb(lage)
#
# Df     AIC
# <none>            -8363.2
# - pb(lage) 2.4887 -8343.0
# ----------------------------------------------------
#   Distribution parameter:  sigma
# Start:  AIC= -8363.17
# ~pb(lage)
#
# Df     AIC
# <none>            -8363.2
# - pb(lage) 3.4514 -8336.0
# ----------------------------------------------------
#   Distribution parameter:  mu
# Start:  AIC= -8363.17
# fev1fvc ~ ga(~s(lht, lage))
#
# Df     AIC
# <none>                     -8363.2
# - ga(~s(lht, lage)) 14.821 -7388.5
# ----------------------------------------------------
# > m1
#
# Family:  c("logitSST", "logit SST")
# Fitting method: RS()
#
# Call:  gamlss(formula = fev1fvc ~ ga(~s(lht, lage)), sigma.formula = ~pb(lage),
#               nu.formula = ~pb(lage), tau.formula = ~pb(lage),
#               family = logitSST, data = d2mR, trace = FALSE)
#
# Mu Coefficients:
#   (Intercept)  ga(~s(lht, lage))
# 2.104                 NA
# Sigma Coefficients:
#   (Intercept)     pb(lage)
# 0.5289      -0.3524
# Nu Coefficients:
#   (Intercept)     pb(lage)
# 0.6812      -0.1168
```

```
 # Tau Coefficients:
 #   (Intercept)     pb(lage)
 # 2.3553      -0.5709
 #
 # Degrees of Freedom for the fit: 36.34 Residual Deg. of Freedom   2805
 # Global Deviance:     -8435.84
 # AIC:     -8363.17
 # SBC:     -8146.88

#-----------------------------------------------------------
# The final model
#-----------------------------------------------------------
gen.Family("SST", "logit")
library(gamlss.add)
library(gamlss.inf)
mf <- gamlssInf0to1(y=fev1fvc,
                        mu.formula = ~ ga(~s(lht, lage)),
                     sigma.formula = ~pb(lage),
                        nu.formula = ~pb(lage),
                       tau.formula = ~pb(lage),
                       xi1.formula = ~pb(lage)+pb(lht),
                            family = logitSST,
                              data = d2m,
                             n.cyc = 100,
                             trace = TRUE)
# *****       The binomial model        *****
#   GAMLSS-RS iteration 1: Global Deviance = 1580.418
# GAMLSS-RS iteration 2: Global Deviance = 1580.35
# GAMLSS-RS iteration 3: Global Deviance = 1580.35
# ***** The continuous distribution model *****
#   GAMLSS-RS iteration 1: Global Deviance = -8334.247
# GAMLSS-RS iteration 2: Global Deviance = -8409.004
# GAMLSS-RS iteration 3: Global Deviance = -8424.92
# GAMLSS-RS iteration 4: Global Deviance = -8435.827
# GAMLSS-RS iteration 5: Global Deviance = -8441.401
# GAMLSS-RS iteration 6: Global Deviance = -8442.847
# GAMLSS-RS iteration 7: Global Deviance = -8442.8
# GAMLSS-RS iteration 8: Global Deviance = -8442.51
# GAMLSS-RS iteration 9: Global Deviance = -8442.3
# GAMLSS-RS iteration 10: Global Deviance = -8442.179
# GAMLSS-RS iteration 11: Global Deviance = -8442.106
# GAMLSS-RS iteration 12: Global Deviance = -8442.053
# GAMLSS-RS iteration 13: Global Deviance = -8442.012
# GAMLSS-RS iteration 14: Global Deviance = -8441.975
# GAMLSS-RS iteration 15: Global Deviance = -8441.945
# GAMLSS-RS iteration 16: Global Deviance = -8441.92
# GAMLSS-RS iteration 17: Global Deviance = -8441.892
```

```
# GAMLSS-RS iteration 18: Global Deviance = -8441.879
# GAMLSS-RS iteration 19: Global Deviance = -8441.86
# GAMLSS-RS iteration 20: Global Deviance = -8441.851
# GAMLSS-RS iteration 21: Global Deviance = -8441.841
# GAMLSS-RS iteration 22: Global Deviance = -8441.833
# GAMLSS-RS iteration 23: Global Deviance = -8441.826
# GAMLSS-RS iteration 24: Global Deviance = -8441.82
# GAMLSS-RS iteration 25: Global Deviance = -8441.815
# GAMLSS-RS iteration 26: Global Deviance = -8441.811
# GAMLSS-RS iteration 27: Global Deviance = -8441.807
# GAMLSS-RS iteration 28: Global Deviance = -8441.804
# GAMLSS-RS iteration 29: Global Deviance = -8441.802
# GAMLSS-RS iteration 30: Global Deviance = -8441.8
# GAMLSS-RS iteration 31: Global Deviance = -8441.797
# GAMLSS-RS iteration 32: Global Deviance = -8441.795
# GAMLSS-RS iteration 33: Global Deviance = -8441.794
# The Final Global Deviance  = -6861.444
#### fitted model##########



plot(mf)

wp(mf, xvar=age, n.inter = 9)
wp(mf, xvar=ht, n.inter = 9)
wp(mf, xvar=~age+ht, ylim.worm=1)

Q.stats(mf, xvar=d2m$age)
Q.stats(mf, xvar=d2m$ht)
#-------------------------------------------------------------------
# refit the final with age and height rather than log's

mf1 <- gamlssInf0to1(y=fev1fvc,
                     mu.formula = ~ ga(~s(log(ht), log(age))),
                     sigma.formula = ~pb(log(age)),
                     nu.formula = ~pb(log(age)),
                     tau.formula = ~pb(log(age)),
                     xi1.formula = ~pb(log(age))+ pb(log(ht)),
                     family = logitSST,
                     data = d2m,
                     n.cyc = 100,
                     trace = TRUE)
x11()
plot(mf1)
## ---------------------------------------------------------------
newdata<-expand.grid(age=seq(5,90,0.1), ht=seq(100,210,1))
## ---------------------------------------------------------------
```

```
# get prediction values
# for xi1
xi1<-predict(mf1$multinom, newdata=newdata, type="response")
# for mu sigma nu and tau
all<-predictAll(mf1$dist, newdata=newdata, type="response")

xi10 <- predict(mf1, newdata=newdata, type="response", parameter="xi1")
 mu0 <- predict(mf1, newdata=newdata, type="response", parameter="mu")
 sigma0 <- predict(mf1, newdata=newdata, type="response", parameter="sigma")
 nu0 <- predict(mf1, newdata=newdata, type="response", parameter="nu")
 tau0 <- predict(mf1, newdata=newdata, type="response", parameter="tau")

plot(xi1~xi10)
plot(all$mu~mu0)

plot(all$sigma~sigma0)
plot(all$nu~nu0)
plot(all$tau~tau0)

names(all)
all$xi1 <- xi1
names(all)
# or using predict
pall <- list()
pall$mu<-predict(mf1,  newdata=newdata, type="response")
#plot(pall$mu~all$mu)
pall$sigma<-predict(mf1,newdata=newdata, parameter="sigma",type="response")
pall$nu<-predict(mf1,newdata=newdata, parameter="nu", type="response")
pall$tau<-predict(mf1,newdata=newdata, parameter="tau", type="response")
pall$xi1<-predict(mf1,newdata=newdata, parameter="xi1", type="response")
## ----------------------------------------------------------------

# we need to generate the distrbution
gen.Inf0to1("logitSST", "One")

fev5<-qlogitSSTInf1(0.05, mu=all$mu,sigma=all$sigma,
                    nu=all$nu,tau=all$tau, xi1=all$xi1)
# this can be used to plot fitted observations
pl <- function(i=1)
plotlogitSSTInf1(all$mu[i],all$sigma[i],all$nu[i],all$tau[i], all$xi1[i])

pl(1)
pl(1000)
pl(100)
pl(3045)

## ----eval=FALSE----------------------------------------------------------
```

```
# lower<-rep(maty[,2],111)
# upper<-rep(maty[,4],111)
# fev5a<-ifelse(((newdata$height<lower) | (newdata$height>upper)),
#                NaN,fev5)

## ----eval=FALSE-------Contour plot-----------------------------------------------
newheight<-seq(100,210,1)
newage<-seq(5,90,0.1)
mfev5<-matrix(data=fev5,nrow=851,ncol=111)
x11()
contour(newage,newheight,mfev5, nlevels=40, xlab="age(years)",
        ylab="height(cm)")
points(d2m$age, d2m$ht, col="green4")
```

Appendix G

Help file for BSSN distribution in R

The following are the source Rd text to generate help file for the bimodal skew symmetric normal distribution (BSSN).

```
\name{BSSN}
\alias{BSSN}
\alias{dBSSN}
\alias{pBSSN}
\alias{qBSSN}
\alias{rBSSN}

%- Also NEED an '\alias' for EACH other topic documented here.
\title{Bimodal Skew Symmetric Normal Distribution}
\description{
These functions define the Bimodal Skew Symmetric Normal Distribution.
This is a four parameter distribution and can be used to fit a GAMLSS model.
The functions \code{dBSSN}, \code{pBSSN}, \code{qBSSN} and \code{rBSSN} define
the probability distribution function, the cumulative distribution function,
the inverse cumulative distribution functions and  the random generation
for the Bimodal Skew Symmetric Normal Distribution; respectively.
}
\usage{
BSSN(mu.link = "identity", sigma.link = "log", nu.link = "identity",
     tau.link = "log")
dBSSN(x, mu = 0, sigma = 1, nu = 1, tau = 0.5, log = FALSE)
pBSSN(q, mu = 0, sigma = 1, nu = 1, tau = 0.5, lower.tail = TRUE,
     log.p = FALSE, log = TRUE)
qBSSN(p, mu = 0, sigma = 1, nu = 1, tau = 0.5, lower.tail = TRUE,
      log.p = FALSE)
```

```
}
%- maybe also 'usage' for other objects documented here.
\arguments{
  \item{mu.link}{Defines the \code{mu.link}, with "identity"
   link as the default for the mu parameter}
  \item{sigma.link}{Defines the \code{sigma.link}, with "log"
   link as the default for the sigma parameter}
  \item{nu.link}{Defines the \code{nu.link}, with "identity"
   link as the default for the nu parameter}
  \item{tau.link}{Defines the \code{tau.link}, with "identity"
   link as the default for the tau parameter}
  \item{x,q}{vector of quantiles}
  \item{mu}{vector of location parameter values}
  \item{sigma}{vector of scale parameter values}
  \item{nu}{vector of nu parameter values}
  \item{tau}{vector of tau parameter values}
  \item{log, log.p}{logical; if TRUE, probabilities p are given as log(p).}
  \item{lower.tail}{logical; if TRUE (default), probabilities
   are P[X <= x], otherwise, P[X > x]}
  \item{p}{vector of probabilities}
    \item{n}{number of observations. If \code{length(n) > 1},
    the length is taken to be the number required}
}
\details{
Then the probability density function of the BSSN distribution is given by

  \deqn{f_Y(y|\mu, \sigma, \nu, \tau)= c[\tau + (y-\nu)^(2)]e^{-\sigma(y-\mu)^(2)}}

for \eqn{-\infty < y < \infty}, where \eqn{c = 2\sigma^(3/2) / \gamma \sqrt\pi},
\eqn{\gamma= 1 + 2 \sigma \theta},  eqn{\theta= \tau + \delta^{2}},
\eqn{\delta= \nu - \mu}. \eqn{-\infty <\mu <\infty} and
\eqn{-\infty < \nu < \infty} are location parameters and
\eqn{\sigma >0} and \eqn{\tau \geq 0 }
denote the scale and bi-modality parameters respectively.
}
\value{
%%  ~Describe the value returned
%%  If it is a LIST, use
%%  \item{comp1 }{Description of 'comp1'}
%%  \item{comp2 }{Description of 'comp2'}
%% ...
}
\references{
Hassan, M. Y. and  El-Bassiouni M. Y. (2015).
Bimodal skew-symmetric normal distribution,
\emph{Communications in Statistics-Theory and Methods},
\bold{45}, part 5, pp 1527--1541.
```

```
}
\author{Abu Hossain, Bob Rigby and Mikis Stasinopoulos}
\note{
%%  ~~further notes~~
}

%% ~Make other sections like Warning with \section{Warning }{....} ~

\seealso{
%% ~~objects to See Also as \code{\link{help}}, ~~~
}
\examples{
op<-par(mfrow=c(3,3))
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=1),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=5),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=10),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=1, tau=20),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=0, tau=4),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=-1, sigma=0.1, nu=0, tau=3),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=1, sigma=0.1, nu=2, tau=0),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=-1, sigma=0.1, nu=-2, tau=0),-12, 12,
            ylab="f(x)", main="BSSN")
curve(dBSSN(x,  mu=-1, sigma=0.1, nu=-3, tau=0.8),-12, 12,
            ylab="f(x)", main="BSSN")
par(op)
  }
}
% Add one or more standard keywords, see file 'KEYWORDS' in the
% R documentation directory.
\keyword{ ~kwd1 }% use one of  RShowDoc("KEYWORDS")
\keyword{ ~kwd2 }% __ONLY ONE__ keyword per line
```