Hossain, A., Rigby, R.A., Stasinopoulos, D.M. and Enea, M. (March 2016)
Centile estimation fro a proportion response variable.
Statistics in Medicine, Vol 35, 6, 895-904

https://doi.org/10.1002/sim.6748

Statistics in Medicine

Centile Estimation for a Proportion Response Variable

Journal:	*Statistics in Medicine*
Manuscript ID:	Draft
Wiley - Manuscript type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Stasinopoulos, Dimitrios; London Metropolitan University, STORM Rigby, Robert; London Metropolitan University, STORM Hossain, Abu; London Metropolitan University, STORM Enea, Marco; University Of Palermo, Statistics
Keywords:	Beta inflated distribution, Centile Estimation, GAMLSS, generalised Tobit model, LMS

Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Centile Estimation for a Proportion Response Variable

Abu Hossain, [a] Robert Rigby, [a] [*] Mikis Stasinopoulos [a] and Marco Enea [b]

This paper introduces two general models for computing centiles when the response variable Y can take values between zero and one, inclusive of zero or one. The models developed are more flexible alternatives to the beta inflated distribution. The first proposed model employs a flexible four parameter logit skew student t (logitSST) distribution to model the response variable Y on the unit interval (0,1), excluding 0 and 1. This model is then extended to the inflated logitSST distribution for Y on the unit interval, including 0 or 1. The second model developed in this paper is a generalized Tobit model for Y on the unit interval, including 0 or 1. An application of the new models to real data shows that they can provide superior fits.

Keywords: Beta inflated distribution, Centile Estimation, GAMLSS, generalised Tobit model, LMS, logit skew student t distribution.

1. Introduction

The purpose of this paper 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 curve bracket indicates that the end point is excluded. In this paper we will focus on a response variable Y on $(0, 1]$. Extensions are available for Y on $[0, 1)$ and $[0, 1]$.

Specifically in this paper we develop two new innovations 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 generalised 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 generalised 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` package version $4.3 - 2$ in R, [1] based on the `GAMLSS` model, [2].

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 [3]. Warton and Hui [4] argued that the logit transformation worked better than the arcsine transformation

[a] *STORM, London Metropolitan University* [b] *University of Palermo*
[*] *Correspondence to: . E-mail: r.rigby@londonmet.ac.uk*

when analysing proportion data. Aitchinson [5] 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.

Given its relatively flexible nature, the beta distribution has been used widely in the statistical literature to model proportion data. For example Trenkler [6], Kieschnick and McCullough [7] and Ferrari and Cribari-Neto [8] 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 [7], Hoff [9], Cook et al. [10] and Ospina and Ferrari [11] presented empirical examples of 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. [12] use a Bayesian approach to augment probabilities of zeroes and ones with the beta density for modelling proportion data. Hoff [9] 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 [11] 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 [11] do not claim that the beta inflated model always provides a better fit than the Tobit model.

The rest of the paper proceeds as follows. Section 2 describes the statistical methodology implemented in this paper. In particular, subsection 2.1 includes a brief description of centile estimation using the LMS method and its extensions. Subsection 2.2 provides a general model for centile estimation, while subsections 2.3 and 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 2.5 describes the generalised Tobit model for Y on $(0, 1]$. Section 3 applies the methodologies proposed in section 2 to the lung function data. Conclusions are given in Section 4.

2. Statistical methodology

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 [13, 14]. 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 [15] and its extensions in [16, 17]. 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 [15] is that the response variable $Y > 0$ is defined by a transformation

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

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 [18] in `R` [19]) 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 [16, 17] extended the $BCCG$ distribution by introducing the four parameter $BCPE$ and BCT distributions. For the *BCPE* distribution the transformed random variable Z in (1) follows a truncated power exponential distribution, while for the *BCT* distribution Z follows a truncated *t* distribution.

2.2. *General model for centile estimation*

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

$$\begin{aligned} Y &\sim D(\boldsymbol{\theta}) \\ g_k(\theta_k) &= h_k(x) \qquad k = 1, \ldots .p \\ x &= v^{\xi} \end{aligned} \tag{2}$$

where D is the assumed distribution, g_k and h_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 [15], [16] and [17] respectively. For example, the LMST method is given by

$$\begin{aligned} Y &\sim BCT(\mu, \sigma, \nu, \tau) \\ g_1(\mu) &= h_1(x) \\ g_2(\sigma) &= h_2(x) \\ g_3(\nu) &= h_3(x) \\ g_4(\tau) &= h_4(x) \\ x &= v^{\xi}. \end{aligned} \tag{3}$$

The default link functions for the BCT distribution in `gamlss` package are $g_1(\mu) = \mu$ (identity link), $g_2(\sigma) = \log \sigma$, $g_3(\nu) = \nu$ and $g_4(\tau) = \log \tau$. [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 variable on $[0, 1)$ and $[0, 1]$ are available.

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 range $(-\infty, \infty)$ transformed to range $(0, 1)$. 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 reason for the proposed model is that the beta distribution is often inadequate especially with large data for modelling a proportion response variable in real data sets. The inverse logit transformation of the skew student t (SST) distribution, called the $logitSST$ distribution, is introduced here 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 [20], reparameterized from [21] . 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 $(0, 1)$.

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 probability (density) function for $Y \sim Inf.logitSST(\mu, \sigma, \nu, \tau, p_1)$ is given by

$$f_Y(y|\mu,\sigma,\nu,\tau,p_1) = \begin{cases} p_1 & \text{if } y = 1 \\ (1-p_1) f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0 < y < 1 \end{cases} \tag{4}$$

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 $0 < p_1 < 1$, given by (4), 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.

$$\mu = h_1(x)$$

$$\log \sigma = h_2(x)$$
$$\log \nu = h_3(x)$$
$$\log \tau = h_4(x)$$
$$\log\left(\frac{p_1}{1-p_1}\right) = h_5(x).$$

The inflated logitSST distribution defined by (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 (4) can be fitted using the new package in R called `gamlss.inf`.

2.5. Generalised 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, [22].

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, [9]. Here the values of Y in the model distribution above 1 are transformed to 1.

Let $Y_1 \sim D(\mu, \sigma, \nu, \tau)$ be a flexible uncensored distribution on $(0, \infty)$. Let $Y \sim D_{rc}(\mu, \sigma, \nu, \tau)$ be the right censored distribution on $(0, 1]$, i,e censored above 1. Then

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

Hence the probability (density) function of Y is given by

$$f_Y(y) = \begin{cases} f_{Y_1}(y) & \text{if } 0 < y < 1 \\ P(Y_1 \geq 1) & \text{if } y = 1 \end{cases} \tag{5}$$

for $0 < y \leq 1$. In principle *D* can be any distribution on $(0, \infty)$. In the analysis in section 3 we use the three parameter *BCCGo* distribution for *D* with its default link functions.

3. Data Analysis

3.1. Fitting Models

In this paper we modelled 3164 male observations of lung function data previously analysed by Stanojevic et al. [23]. Data quality were assured and there were very few errors in the data since all data had been previously used in publications. The response variable is $Y = FEV_1/FVC$ and the explanatory variable is $x = log(height)$. Response variable Y is a ratio of forced expiratory variable in 1 second (FEV_1) to forced vital capacity (FVC). Spirometric lung function Y is an established index for diagnosing airway obstruction, [24].

Centile curves for Y against x are achieved by using five methods: LMS (*BCCGo*), *BEINF1* (beta inflated at 1), the original Tobit model and two new methods proposed in this paper, the $Inf.logitSST$ ($logitSST$ inflated at 1) model and the generalised Tobit model (*BCCGorc*, i.e, *BCCGo* right censored at 1). The methods were applied using the `gamlss` package version $4.3-2$ in R, [1]. For the inflated $logitSST$ model a new package was developed called `gamlss.inf`. The smoothing method P-splines, [25], was used for fitting each smooth function $h_k(x)$ 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, [26], 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 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 resulting in biased centile curves, leading to significant Z and Q residual test statistics. The `R` commands used in the analysis are available from the authors.

Table 1 summarises the values of the global Deviance, degrees of freedom (df), Akaike information criteron (AIC), Schwartz Bayesian criteron (SBC) and Generalised AIC (GAIC) with penalty $k = 6$, for four fitted models. The LMS (*BCCGo*) model could not be included in the comparison in Table 1 because it does not have a point probability at $Y = 1$. From Table 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 $BEINF1$ and standard Tobit models perform much worse.

3.2. Centile estimation

Figure 1 shows centile curves constructed using four different fitted models: LMS ($BCCGo$), $BEINF1$, inflated $logitSST$ and generalised Tobit ($BCCGo$, right censored at 1). The fitted $(2, 10, 25, 50, 75, 90, 98)\%$ centile curves show that LMS model constructed less smooth curves than the other models. Table 2 shows the sample percentages below each centile curve for the fitted models against the nominal percentiles. 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.

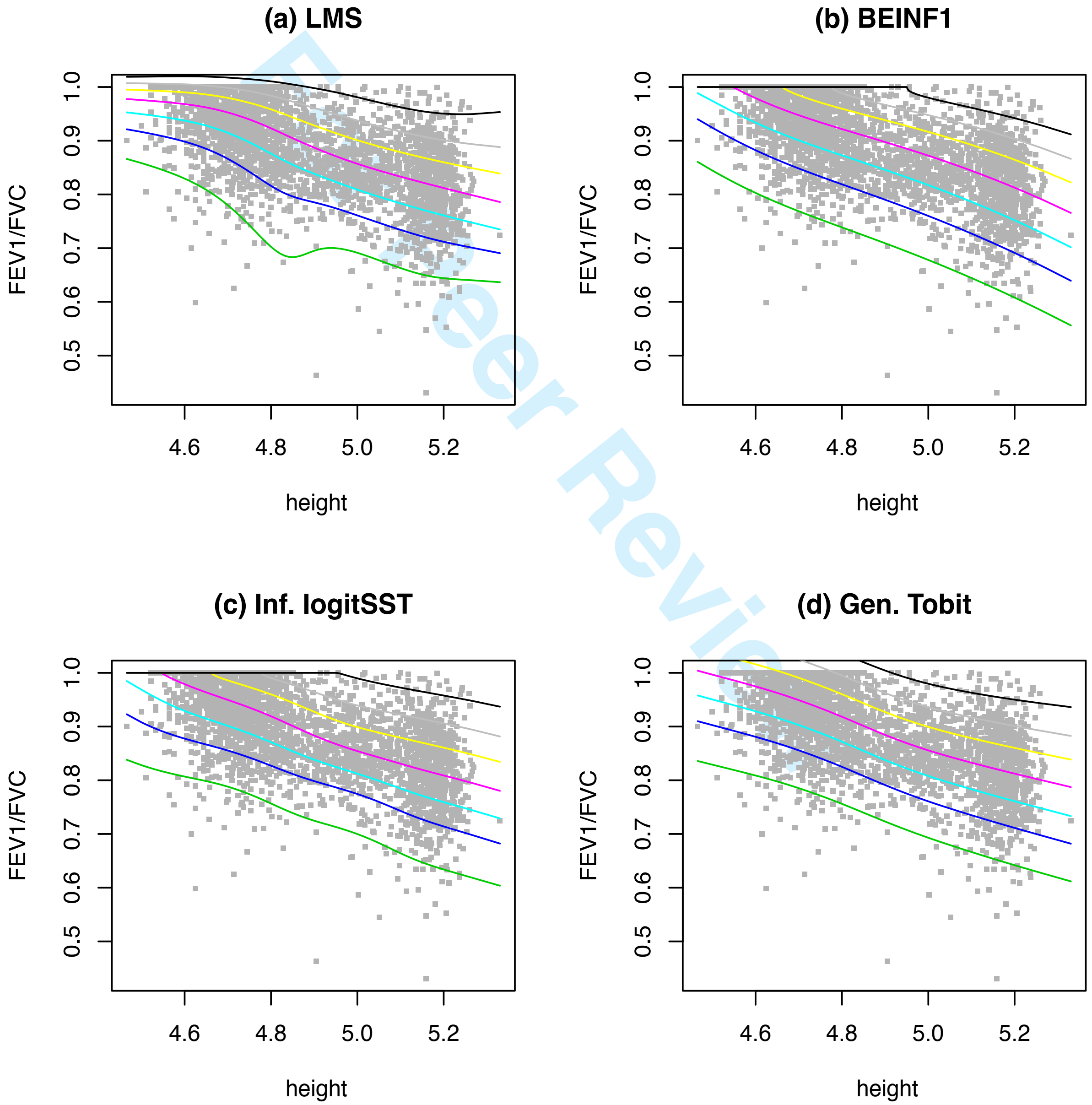

Figure 1. Centile curves for model a) LMS b) BEINF1 c) Inflated logitSST d) Generalised Tobit

3.3. *Model checking using residual based diagnostics*

The residuals used in GAMLSS are normalized (randomised) quantile residuals, [27], or 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.

3.3.1. Worm Plots van Buuren and Fredriks [28] 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. In Figure 2 worm plots of the LMS and $BEINF1$ models are shown in 9 intervals of height with equal number of observations. The 9 height intervals are given in Figure 4 and corespond to the 9 worm plots from the bottom left plot to the top right plot in rows in Figures 2 and 3. 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 [28] and [1]. The shapes of the worm plots can indicate the type of model failure, see [16]. The worm plots in Figure 2 show that the LMS and beta inflated ($BEINF1$) models fit badly in most height intervals.

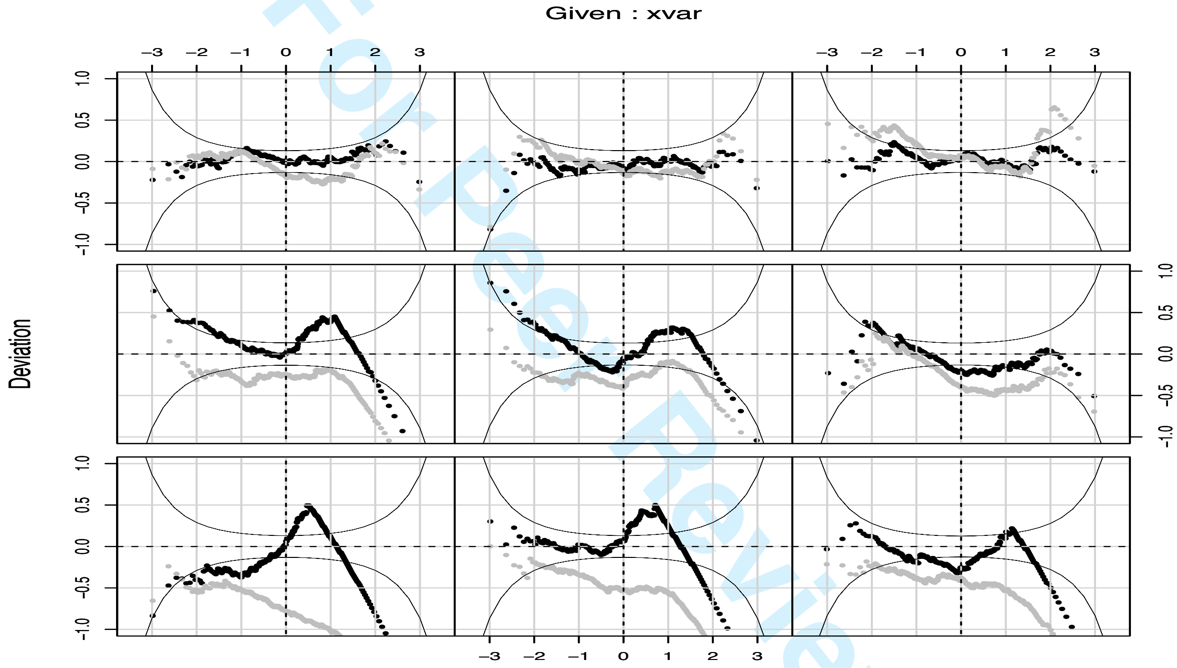

Figure 2. Twin worm plot for LMS (dark points) and BEINF1 (light points) models.

The worm plots of the two proposed models, the inflated logitSST and generalised Tobit models are shown in Figure 3. 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.

3.3.2. Z and Q statistics Another residual based diagnostic tool, Z and Q statistics, is used in this paper. Z and Q statistics are useful to test the 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 $0, 1, 0$

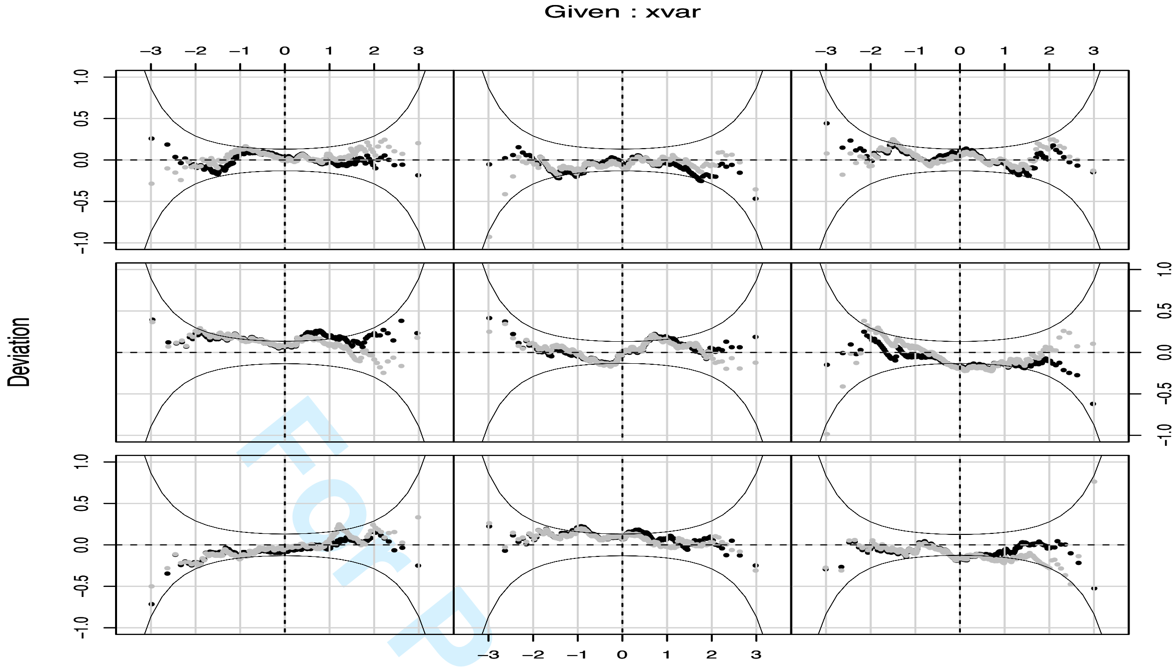

Figure 3. Twin worm plot for Inf logitSST (dark points) and Gen.Tobit (light points) models.

and 0 respectively within each explanatory variable interval group $g = 1, 2, 3.....G$, [29]. The test statistics for skewness and kurtosis are given by D'Agostino et al. [30].

Royston and Wright [29] also computed Q-statistics by

$$Q_j = \sum_{g=1}^{G} Z_{gj}^2. \tag{6}$$

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 [29] and [16]. 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 is $Z_{gj} > 2$ or $Z_{gj} < -2$ indicating that the model may be inadequate within the corresponding height interval.

Figure 4 shows the visual display of the Z_{gj} statistics for $g = 1, 2, 3....9$ and $j = 1, 2, 3, 4$ for each of four models. The corresponding intervals of height for the 9 groups are given on the left of each plot. The larger the circles, the larger the value of $|Z_{gj}|$. A square within the circle indicates that $|Z_{gj}| > 2$ indicating a misfit of the model to the response variable within the coresponding interval of height.

Hence Figure 4 shows the presence of many misfits in the LMS and beta inflated models. Five misfits were found in the inflated $logitSST$ model and four in the generalised Tobit model. Analysis of the residuals using Z statistics is also consistent with the diagnostic tools used earlier in this paper.

It can be concluded that the inflated $logitSST$ and generalised Tobit models perform well compared to other models and provide powerful tools to model proportion data.

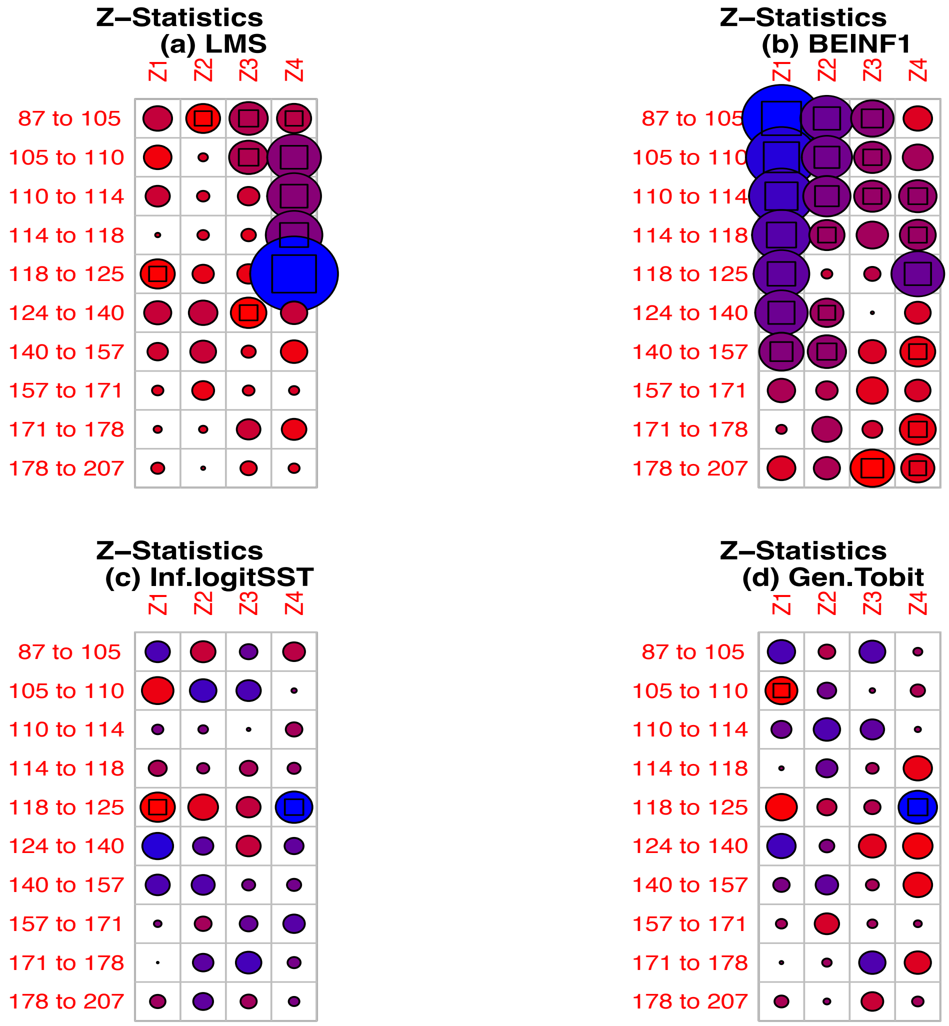

Figure 4. Z statistics for a) LMS b) BEINF1 c) Inflated logitSST d) Generalised Tobit

4. Conclusion

This paper proposed the inflated logit skew student t (i.e. inflated $logitSST$) distribution and a generalised 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 paper is to offer two models for centile estimation as viable alternatives to the LMS and beta inflated models. The paper focuses on a response variable recorded on the interval $(0, 1]$. Extension to the interval [0,1) can be achieved simply by analysing $(1 - Y)$ instead of Y. Extension to the interval $[0, 1]$ may be obtained 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 generalised Tobit model.

An empirical application to real data has been presented modelling a lung function response variable on the interval $(0, 1]$. This paper uses the Akaike information Criterion (AIC), [31] and Schwartz Bayesian Criterion (SBC), [32] 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 inflated $logitSST$ and generalised Tobit (i.e, $BCCGo$ right censored at 1) models provided adequate fits. From the empirical example it can be concluded that the inflated $logitSST$ and generalised Tobit models can provide better fits than the LMS, beta inflated and Tobit models and can be more flexible for modelling proportion data.

References

1. Stasinopoulos D, Rigby R. Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software* 2007; **23**(7):1–46.
2. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape, (with discussion). *Appl. Statist.* 2005; **54**:507–554.
3. Schmid M, Hothorn T, Maloney KO, Weller DE, Potapov S. Geoadditive regression modeling of stream biological condition. *Environmental and Ecological Statistics* 2011; **18**(4):709–733.
4. Warton DI, Hui FK. The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 2011; **92**(1):3–10.
5. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)* 1982; **44**(2):139–177.
6. Trenkler G. 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* January 1996; **21**(1):119–119.
7. Kieschnick R, McCullough BD. Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical modelling* 2003; **3**(3):193–213.
8. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 2004; **31**(7):799–815.
9. Hoff A. Second stage dea: Comparison of approaches for modelling the {DEA} score. *European Journal of Operational Research* 2007; **181**(1):425 – 435, doi:http://dx.doi.org/10.1016/j.ejor.2006.05.019.
10. Cook DO, Kieschnick R, McCullough B. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance* 2008; **15**(5):860 – 867.
11. Ospina R, Ferrari SLP. Inflated beta distributions. *Statistical Papers* 2010; **23**:111–126.
12. Galvis DM, Bandyopadhyay D, Lachos VH. Augmented mixed beta regression models for periodontal proportion data. *Statistics in medicine* 2014; **33**(21):3759–3771.
13. WHO MGRSG. *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, 2006.
14. WHO MGRSG. *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, 2007.
15. Cole TJ, Green PJ. Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine.* 1992; **11**:1305–1319.
16. Rigby RA, Stasinopoulos DM. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine* 2004; **23**:3053–3076.

17. Rigby RA, Stasinopoulos DM. Using the Box-Cox *t* distribution in gamlss to model skewness and kurtosis. *Statistical Modelling* 2006; **6**:209–229.

18. Stasinopoulos D, Rigby R. Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* 2007; **23**(7):1–46.

19. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2014. URL http://www.R-project.org/.

20. Wurtz D, Chalabi Y, Luksan L. Parameter estimation of arma models with garch/aparch errors an r and splus software implementation. http://www-stat.wharton.upenn.edu/~steele/Courses/956/RResources/GarchAndR/WurtzEtAlGarch.pdf. Accessed: 2014-12-19.

21. Fernández C, Steel MF. On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 1998; **93**(441):359–371.

22. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica* 1958; **26**(1):24–36.

23. Stanojevic S, Wade A, Cole TJ, Lum S, Custovic A, Silverman M, Hall GL, Welsh L, Kirkby J, Nystad W, *et al.*. Spirometry centile charts for young caucasian children: The asthma uk collaborative initiative. *American journal of respiratory and critical care medicine* 2009; **180**(6):547–552.

24. Quanjer P, Stanojevic S, Stocks J, Hall G, Prasad K, Cole T, Rosenthal M, Perez-Padilla R, Hankinson J, Falaschetti E, *et al.*. Changes in the fev1/fvc ratio during childhood and adolescence: an intercontinental study. *European Respiratory Journal* 2010; **36**(6):1391–1399.

25. Eilers PHC, Marx BD. Flexible smoothing with b-splines and penalties (with comments and rejoinder). *Statist. Sci* 1996; **11**:89–121.

26. Akaike H. Information measures and model selection. *Bulletin of the International Statistical Institute* 1983; **50**:277–290.

27. Dunn PK, Smyth GK. Randomised quantile residuals. *J. Comput. Graph. Statist.* 1996; **5**:236–244.

28. van Buuren S, Fredriks M. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine* 2001; **20**:1259–1277.

29. Royston P, Wright EM. Goodness-of-fit statistics for age-specific reference intervals. *Statistics in Medicine* 2000; **19**:2943–2962.

30. D'Agostino, R B, Balanger, A and D'Agostino Jr, R B. A suggestion for using powerful and informative tests of normality. *American Statistician* 1990; **44**:316–321.

31. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**(6):716–723.

32. Schwarz G. Estimating the dimension of a model. *Ann. Statist.* 1978; **6**:461–464.

Table 1. Comparison of Fitted Models

Method	Deviance	df	AIC	SBC	GAIC (k=6)
BEINF1	-6180.9	6.0	-6168.9	-6132.5	-6144.9
Inf. logitSST	-6390.7	14.3	-6362.1	-6275.5	-6304.9
GenTobit	-6362.2	9.0	-6344.2	-6289.6	-6308.8
Tobit	-6272.6	7.0	-6258.6	-6215.9	-6230.4

Table 2. Comparison of fitted centile percentages

Nominal Centile %	LMS	BEINF1	Inf. logitSST	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