Modelling location, scale and shape parameters of the birnbaum-saunders generalized t distribution

Luiz R. Nakamura[1], Robert A. Rigby[2], Dimitrios M. Stasinopoulos[3], Roseli A. Leandro[4], Cristian Villegas[5], Rodrigo R. Pescim[6]

[1]*Departamento de Informatica e Estatstica, Universidade Federal de Santa Catarina*
[2]*STORM Research Centre, London Metropolitan University*
[3]*epartamento de Ci^encias Exatas, Escola Superior de Agricultura "Luiz deQueiroz", Universidade de São Paulo*
[4]*Departamento de Estatstica, Universidade Estadual de Londrina*

Abstract: The Birnbaum-Saunders generalized t (BSGT) distribution is a very flflexible family of distributions that admits different degrees of skewness and kurtosis and includes some important special or limiting cases available in the literature, such as the Birnbaum-Saunders and Birnbaum-Saunders t distributions. In this paper we provide a regression type model to the BSGT distribution based on the generalized additive models for location, scale and shape (GAMLSS) framework. The resulting model has high flflexibility and therefore a great potential to model the distribution parameters of response variables that present light or heavy tails, i.e. platykurtic or leptokurtic shapes, as functions of explanatory variables. For different parameter settings, some simulations are performed to investigate the behavior of the estimators. The potentiality of the new regression model is illustrated by means of a real motor vehicle insurance data set.

Key words*: Finance, GAMLSS, generalized additive models, penalized splines, positively skewed data.*

1. Introduction

The Birnbaum-Saunders (BS) distribution is the most popular model used to describe the lifetime process under fatigue. It was proposed by Birnbaum and Saunders (1969) due to the problems of vibration in commercial aircraft that caused fatigue in the materials. This distribution is also known as the fatigue life distribution and can be used to represent failure time in various scenarios. As reported by Pescim et al. (2014), the BS distribution has received wide ranging applications in past years that include: modelling of hourly SO2 concentrations at ten monitoring stations located in different zones in Santiago (Leiva et al., 2008); modelling of diameter at breast height distributions of near-natural complex structure silver fir-European beech forests (Podlaski, 2008); modelling of hourly dissolved oxygen (DO) concentrations observed at four monitoring stations located at different areas of Santiago (Leiva et al., 2009; Vilca et al., 2010); statistical analysis of redundant systems with one warm stand-by unit (Nikulin and Tahir, 2010), among others.

Because of the widespread study and applications of the BS distribution, there is a need for new generalizations of this distribution. Daz-Garca and Leiva (2005) proposed a family of generalized Birnbaum-Saunders (GBS) distributions based on countoured elliptical

distributions such as Pearson VII and Student's t distributions, Vilca and Leiva (2006) introduced a BS model based on skew normal distributions. Gomez et al. (2009) extended the BS distribution from the slash-elliptic model. Vilca et al. (2010) and Castillo et al. (2011) developed the epsilon-skew Birnbaum-Saunders distribution. More recently, Cordeiro and Lemonte (2011) and Pescim et al. (2014) dened the beta Birnbaum-Saunders and the Kummer beta Birnbaum-Saunders models, respectively.

Despite some of those BS generalized distributions induce asymmetry, symmetry and promote weight variation of the tail, they do not provide all these shapes in the same density function. A highly flexible model which admits light or heavy tails, shaper or flflatter peaked shape and it has some important special and/or limiting cases, is the Birnbaum-Saunders generalized t (BSGT) distribution pro-posed by Genc (2013). This generalization of the BS distribution contains some models previously studied in the literature and, therefore, the BSGT enables to study and t various types of data with different shapes by a unifed approach.

In many practical applications, the responses are affected by explanatory variables such as the socioeconomics and school levels, blood pressure, cholesterol level, soil quality, climate, among many others. BS regression models are widely used to estimate the reliability or predict the durability of non-repairable copies of materials. Among them, Rieck and Nedelman (1991) proposed a log-linear regression model based on the BS distribution. Diagnostic analysis for the BS regression model were developed by Galea et al. (2004), Leiva et al. (2007) and Xie and Wei (2007), while the Bayesian inference was introduced by Tisionas (2001). Barros et al. (2008) proposed a class of lifetime regression models that includes the log-Birnbaum-Saunders-t (BS-t) regression models as special case. Furthermore, Lemonte and Cordeiro (2009) and Villegas et al. (2011) studied the BS nonlinear and BS mixed models, respectively. However, those BS regression models follow the same idea of many previous regression type models in the literature such as generalized linear models (Nelder and Wedderburn, 1972), generalized additive models (Hastie and Tibshirani, 1990) and log location-scale models (Lawless, 2003; Cancho et al., 2009; Roman et al., 2012; Pascoa et al., 2013). These models use only the location parameter of the distribution of the response variable which is a major limitation since other parameters may need to be modelled.

In this context, Rigby and Stasinopoulos (2005) developed the generalized additive models for location, scale and shape (GAMLSS), a very general class of univariate regression models whose main advantage is that all parameters of a given distribution (that does not necessarily belong to the exponential family) can be modelled as parametric and/or additive nonparametric smooth functions of explanatory variables, which can lead to a simpler distribution for a given response variable Y, simplifying the interpretation of the problem in study. Within GAMLSS the shape of the conditional distribution of the response variable can vary according to the values of the explanatory variables, allowing great modelling flexibility.

In this paper, we introduce the BSGT distribution into the GAMLSS frame-work in order to provide a very flexible regression model for this family, modelling all of its four parameters using explanatory variables. The new regression model may be fitted to a data set with light or heavy tails, i.e. a platykurtic or leptokurtic response variable as, for example, the total claim amount of an insurance company. The rest of this paper is outlined as follows. Section 2 provides a brief review of the BSGT family of distributions. The

BSGT is introduced into the GAMLSS framework in Section 3. Section 4 shows a simulation study with different values of the parameters. A real data set application regarding insurance is provided in Section 5 to show the BSGT flexibility, and comparing it with well-known models. Section 6 ends the paper with some concluding remarks.

2. The BSGT distribution - a brief review

Daz-Garca and Leiva (2005) proposed the GBS family of distributions defined by transformation from any random variable Z with symmetric distribution S (Gupta and Varga, 1993), with density given by

$$f_Z(z|\zeta,\phi,\delta) = |cK\left[\frac{(x-\delta)^2}{\phi^2}\right], -\infty < z < \infty,$$

as

$$\mathrm{Y} = \beta\left[\frac{\alpha Z}{2} + \sqrt{\left(\frac{\alpha Z}{2}\right)^2 + 1}\right]^2 \qquad (1)$$

where $\mathrm{Z} \sim \mathrm{S}(\zeta = 0, \phi = 1, \delta)$, δ correponds to the parameters inherited from the baseline distribution, c is the normalizing constant such that $f_Y(y)$ is a proper density, $K(\cdot)$ is the kernel of the density of Z, $\alpha > 0$ represents the shape parameter and $\beta > 0$ is the scale parameter and is also the median of the distribution. As $\alpha \to 0$, the GBS distribution becomes symmetrical around β, whereas when α grows the distribution becomes increasingly positively skewed. Its probability density function (pdf) can be expressed as

$$f_Y(\mathrm{y}|\alpha,\beta,\delta) = \frac{c}{2\alpha\beta^{\frac{1}{2}}} y^{-\frac{3}{2}}(\mathrm{y}+\beta)\mathrm{K}\left(\frac{1}{\alpha^2}\left[\frac{y}{\beta}+\frac{\beta}{y}-2\right]\right),$$

for $y > 0$.
If Z has a GT distribution, $\mathrm{Z} \sim \mathrm{GT}(0,1,\nu,\tau)$, with pdf given by

$$f_Z(z|\nu,\tau) = \frac{\tau}{2\nu^{\frac{1}{\tau}}B(\frac{1}{\tau},\nu)(1+\frac{|z|^\tau}{\nu})^{\nu+\frac{1}{\tau}}}$$

where $-\infty < \mathrm{z} < \infty$ then the random variable Y obtained from transformation (1) has a BSGT distribution with pdf given by

$$f_Y(y|\alpha,\beta,\nu,\tau) = \frac{\tau y^{-\frac{3}{2}}(y+\beta)}{4\alpha\beta^{\frac{1}{2}}\nu^{\frac{1}{2}}B\left(\frac{1}{\tau},\nu\right)}(1+\frac{1}{\nu\alpha^\tau}\left|\frac{y}{\beta}+\frac{\beta}{y}\right|^{\frac{\tau}{2}})^{-(\nu+\frac{1}{\tau})}$$

where $y > 0, \alpha > 1, \beta > 0, \upsilon > 0$ and $\tau > 0$. As in (1) if the shape parameter $\alpha \to 0$, the distribution becomes near symmetrical around β and when α grows the model becomes increasingly positively skewed; β is a scale parameter and is also the median of the distribution; and ν and τ are the parameters related to the peak and tails of the distribution.

Note that as $y \longrightarrow \infty$ then $f_Y(y|\alpha,\beta,\nu,\tau) = O(y^{-\frac{\nu\tau}{2}-1})$, the same order as a t distribution with $\nu\tau/2$ degrees of freedom.

Hence, small values of the product of and result in a heavier upper tail. Similarly, larger values of the product of ν and τ will result in a lighter upper tail. Parameter τ also affects the peak of the distribution: $\tau \leq 1$ results in a spiked peak in the distribution, with a sharp spike (i.e. infinite derivative) if $\tau < 1$, while a larger τ results in an increasingly flatter peak.

Despite its flexibility that can combine symmetrical/asymmetrical shapes with light or heavy tails (i.e. leptokurtic or platykurtic densities), the BSGT model is important since it has some special or limiting cases already proposed in the literature such as the BS, BS-t, BS-Laplace, BS-Cauchy and BS-power exponential (BSPE) distributions, as displayed in figure1.

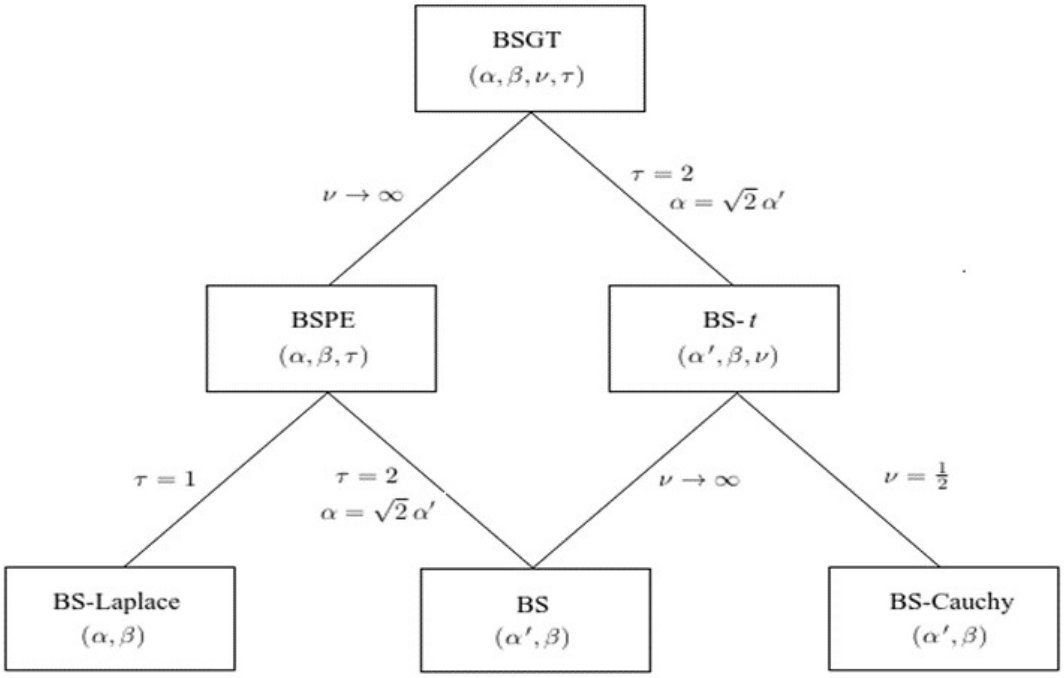

Figure 1: Relationships of the BSGT special models

3. GAMLSS model for the BSGT distribution

GAMLSS are semi-parametric regression models that involve a distribution for the response variable (parametric part) and may involve non-parametric smoothing terms when modelling parameters of the distribution as functions of explanatory variables. The GAMLSS R implementation, called the gamlss package, includes distributions with up to four parameters that are commonly represented by μ for location, σ for scale and ν and τ for shape (Rigby and Stasinopoulos, 2005). Hence, for the BSGT distribution, we consider $\mu = \beta, \sigma = \alpha, \nu = \nu$ and $\tau = \tau$ to obey the established notation in GAMLSS framework in R software (Stasinopoulos and Rigby, 2007). Moreover, from this point, we say that a random variable Y follows a BSGT distribution, denoted by $Y \sim BSGT(\mu, \sigma, \nu, \tau)$.

The GAMLSS model for the BSGT distribution assumes that conditional on its parameters $(\mu, \sigma, \nu \text{ and } \tau)$, observations Y_i are independent $BSGT(\mu, \sigma, \nu, \tau)$ variables with pdf given in (2), and can be expressed as

$$
\begin{aligned}
g_1(\mu) &= \eta_1 = X_1\beta_1 + \sum_{j=1}^{J_1} h_{j1}\,(x_{j1}) \\
g_2(\sigma) &= \eta_2 = X_2\beta_2 + \sum_{j=1}^{J_2} h_{j2}\,(x_{j2}) \\
g_3(\nu) &= \eta_3 = X_3\beta_3 + \sum_{j=1}^{J_3} h_{j3}\,(x_{j3}) \\
g_4(\tau) &= \eta_4 = X_4\beta_4 + \sum_{j=1}^{J_4} h_{j4}\,(x_{j4}) \qquad (3)
\end{aligned}
$$

where $g_k(\cdot)$, k = 1,2,3,4, are the link functions, $\beta_k^T = (\beta_{1k}, \ldots, \beta_{J_k'k})$ denotes the parameter vector associated to explanatory variables with design matrix X_k and each h_{jk} function is a smooth non-parametric function of an explanatory variable x_{jk}, being typically a smoothing spline (for more details, see e.g. Hastie and Tibshirani, 1990) or P-spline (Eilers and Marx, 1996).

Selecting the response variable distribution and diagnostics

Two different stages comprehend the strategy to t a GAMLSS model: fitting and diagnostics. In the fitting stage, we t different models using a generalized Akaike information criterion (GAIC, for more information, see Voudouris et al., 2012) to compare them (the model with the smallest GAIC is selected). The Akaike information criterion (AIC; Akaike, 1974) and Schwarz Bayesian criterion (SBC, Schwarz, 1978) are special cases of the GAIC(k) when k = 2 and k = log(n), respectively.

In the diagnostic stage, we use the normalized (randomized) quantile residuals (Dunn and Smyth, 1996) that are defined by

$$\hat{\tau}_i = \Phi^{-1}(\hat{u}_i)$$

Where Φ^{-1} is the inverse cumulative distribution function of a standard normal variable and $\hat{u} = F_Y(y|\hat{\theta})$ is the fitted cumulative distribution function. The main advantage of this type of residual is that its true values $r_i, i = 1, \ldots, n$ always have a standard normal distribution given the assumption that the model is correct, whatever the distribution of the response variable, i.e. if the model for the response variable is correct, the residuals have a standard normal distribution.

Selecting the explanatory variables

In order to select the explanatory variables for the BSGT model, we use a backward/forward algorithm implemented in gamlss package called StepGAICAll.A:

1) Step 1: select a model for μ using a forward GAIC selection procedure and fixing σ , ν and τ;

2) Step 2: select a model for σ using a forward GAIC selection procedure given the model for μ in Step 1 and fixing ν and τ;

3) Step 3: select a model for ν using a forward GAIC selection procedure given the models for μ and σ obtained in Steps 1 and 2, respectively, and fixing ;

4) Step 4: select a model for τ using a forward GAIC selection procedure given the models for μ, σ and ν obtained in Steps 1, 2 and 3, respectively;

5) Step 5: perform a backward GAIC selection procedure to select a model for ν given the models for μ, σ, and τ obtained from Steps 1, 2 and 4, respectively;

6) Step 6: perform a backward GAIC selection procedure to select a model for σ given the models for μ, ν, and *tau* obtained from Steps 1, 5 and 4, respectively;

7) Step 7: perform a backward GAIC selection procedure to select a model for μ given the models for σ, ν, and τ obtained from Steps 6, 5 and 4, respectively.

The resulting model may contain different terms for each of the parameters μ, σ, ν and τ.

Computational functions

In order to perform a simulation study with the BSGT distribution and a real data set application using the BSGT regression model, we implemented this family into the gamlss package in R (for more details about GAMLSS framework estimation processes, see Rigby and Stasinopoulos, 2005) and the following functions will be available in the gamlss.dist package:

1) dBSGT() gives the BSGT probability density function;
2) pBSGT() gives the BSGT cumulative distribution function (cdf);
3) qBSGT() gives the BSGT quantile function, i.e. inverse cdf; and
4) rBSGT() is the BSGT random number generator.

It is noteworthy that we can also t the special and/or limiting cases of the BSGT distribution in gamlss, e.g. in order to fit a BS-t distribution, we use the following arguments within the main function gamlss(): tau.fix = TRUE and tau.start = 2 in order to fit $\tau = 2$ in BSGT(μ, σ, ν, τ) model.

4. Simulation study

We performed a simulation study generating 12 different scenarios with two sample sizes (n = 100 and n = 500) using the rBSGT() function. The scenarios were chosen in such a way that all possible density shapes could be covered, using as true parameter values μ = 50 and

1) $\sigma = 0.5$ for near symmetrical (Table 1) and $\sigma = 1.5$ for very asymmetrical shapes (Table 2);

2) $\nu = 1.0$ for heavy-tailed and $\nu = 5.0$ for less heavy tailed;

3) $\tau = 1.5, \tau = 2.0$ and $\tau = 10.0$ since a low value of τ tends to a sharper peaked shape (leptokurtic), while a high value of tends to a flatter peaked shape (platykurtic).

The simulation study was performed in a HP Proliant M530e Gen8 Computer under a Debian Linux operating system. Tables 1 and 2 present the true simulated parameter values, average estimates (AE) and standard deviations (SD) for the estimated parameters for near symmetrical (σ = 0.5) and very asymmetrical (σ = 1.5) scenarios, respectively. The required numerical evaluations are implemented in R software through the gamlss function (Stasinopoulos and Rigby, 2007).

Table 1: Real parameter value, average estimates (AE) and standard deviations (SD) based on 1,000 simulations of the near symmetrical version of the BSGT distribution

Parameters	Real value	$n = 100$		$n = 500$	
		AE	SD	AE	SD
		Scenario 1			
μ	50	50.040	2.616	50.000	1.065
σ	0.5	0.461	0.116	0.490	0.050
ν	1.0	1.272	0.986	1.130	0.487
τ	1.5	1.775	1.176	1.526	0.349
		Scenario 2			
μ	50	50.160	2.373	49.940	1.091
σ	0.5	0.469	0.089	0.494	0.036
ν	1.0	1.266	1.028	1.145	0.539
τ	2.0	2.319	1.459	2.053	0.523
		Scenario 3			
μ	50	49.950	1.220	50.000	0.507
σ	0.5	0.482	0.037	0.499	0.018
ν	1.0	1.030	1.996	1.232	0.953
τ	10.0	12.02	6.040	10.900	4.111
		Scenario 4			
μ	50	50.060	2.305	50.040	0.993
σ	0.5	0.494	0.078	0.5012	0.042
ν	5.0	3.790	3.640	5.263	3.277
τ	1.5	1.857	0.722	1.578	0.301
		Scenario 5			
μ	50	50.020	1.943	50.020	0.865
σ	0.5	0.484	0.059	0.5002	0.031
ν	5.0	3.695	3.749	5.388	3.659
τ	2.0	2.606	1.259	2.115	0.438
		Scenario 6			
μ	50	50.020	0.981	50.010	0.379
σ	0.5	0.473	0.032	0.494	0.013
ν	5.0	3.047	4.068	4.406	4.012
τ	10.0	10.570	3.351	11.530	2.868

Table 2: Real parameter value, average estimates (AE) and standard deviations (SD) based on 1,000 simulations of the very asymmetrical version of the BSGT distribution

Parameters	Real value	$n = 100$		$n = 500$	
		AE	SD	AE	SD
Scenario 7					
μ	50	51.010	8.381	50.100	3.198
σ	1.5	1.389	0.345	1.478	0.153
ν	1.0	1.245	0.970	1.151	0.502
τ	1.5	1.718	1.081	1.511	0.326
Scenario 8					
μ	50	50.240	7.263	50.120	3.061
σ	1.5	1.415	0.273	1.487	0.106
ν	1.0	1.177	1.004	1.153	0.603
τ	2.0	2.418	1.540	2.051	0.492
Scenario 9					
μ	50	50.100	3.068	50.040	1.222
σ	1.5	1.444	0.113	1.494	0.052
ν	1.0	1.056	1.872	1.322	1.288
τ	10.0	13.360	7.911	10.790	3.944
Scenario 10					
μ	50	50.42	6.671	50.09	2.753
σ	1.5	1.476	0.230	1.498	0.124
ν	5.0	4.339	4.229	5.090	3.061
τ	1.5	1.919	0.931	1.577	0.303
Scenario 11					
μ	50	50.14	5.430	49.95	2.258
σ	1.5	1.455	0.173	1.496	0.089
ν	5.0	3.815	4.008	4.965	3.120
τ	2.0	2.570	1.099	2.118	0.416
Scenario 12					
μ	50	50.05	2.678	50.020	1.011
σ	1.5	1.419	0.097	1.482	0.040
ν	5.0	3.041	8.159	4.732	5.874
τ	10.0	13.850	8.780	12.720	5.166

As expected, we observe (from Tables 1 and 2) that when n = 500 we obtain closer estimates compared to the true generating value and the SD values decrease. Moreover, we can note that parameters ν and τ are slightly more imprecise than μ and σ which could be happening since they are often high correlated. Also, distribution of the parameter estimators of ν and τ are highly positively skewed.

5. Application: motor vehicle insurance data

In this Section, we illustrate the usefulness of the BSGT regression model, using the GAMLSS framework, to the total claim amount (response variable, Y) from motor vehicle insurance policies over a twelve-month period in 2004−2005 (De Jong and Heller, 2008, p. 15). The original data set was composed of approximately 68,000 policies, but here, we used only those with at least one claim (totalling 3,911 policies). Using this reduced data set, Y ranges from 1.09 to 55,720.00, with mean=2,145.00, median=844.70, standard deviation=3,765.86, skewness=4.74 and kurtosis=38.58.

Since Y is a very positively skewed variable we used four different distributions besides the BSGT distribution which are possible suitable candidates for the response variable: the Box-Cox t (BCTo), generalized gamma (GG), inverse Gaussian (IG) distributions that are already available in gamlss.dist package and the BS distributions which is a special case of the BSGT distribution. The co-variates used to build the models in order to explain Y are displayed in Table 3.

Table 3: Covariates of the motor vehicle insurance data

Variable	Type	Range
Vehicle value (X_1)	Quantitative	\$0–\$139,000
Number of claims (X_2)	Factor	1, 2, 3, 4
Automobile manufacturing company (X_3)	Factor	A, B, C, D
Vehicle age (X_4)	Factor	1, 2, 3, 4 (1 is recent)
Driver gender (X_5)	Factor	male, female
Drivers area of residence (X_6)	Factor	A, B, C, D, E, F
Age band of policy holder (X_7)	Factor	1, 2, 3, 4, 5, 6 (1 is the youngest)
Amount of exposure during the year (X_8)	Quantitative	0–1

Here, we replaced X_1 by $X_1^* = \log(X_1 + 1)$ to modify the high skewness exhibited by this variable. After some previous analysis, we excluded six observations that presented $X_1 = 0$, i.e. the vehicles with value equals zero and the only two observations with $X_2 = 4$, i.e. when there were four claims, since they were considered as outliers. Finally, we fitted several regression models using the backward/forward algorithm available in Section 3. Moreover, we considered a P-spline (pb; for more details, see Eilers and Marx, 1996) in both quantitative covariates (X_1^* and X_8). Appropriate link functions for each of the parameters were chosen in all five distributions: when a distribution parameter θ has range $-\infty < \theta < \infty$, we used the identity link function, whereas, when $\theta > 0$ the logarithm link function was adopted.

A backward/forward selection of explanatory terms as described in Section 3 was performed for all parameters through stepGAICAll.A function in gamlss package (Stasinopoulos and Rigby, 2007) and values of global deviance (GD), Akaike information criterion (AIC) and Schwarz Bayesian criterion (SBC) were computed in order to compare all fitted models. Table 4 displays those statistics from the best tted models for each used distribution, and so, the BSGT regression model could be chosen as the more suitable model since it returned the smallest GD, AIC and SBC values (65,534.1, 65,607.1 and 65,836.0,respectively).

Table 4: Statistics from the best fitted models for each used distribution

Model	GD	AIC	SBC
BSGT	65,534.1	65,607.1	65,836.0
BCTo	65,903.2	65,972.9	66,191.6
GG	65,953.8	66,019.4	66,225.2
IG	66,638.5	66,684.9	66,830.6
BS	66,205.5	66,271.5	66,478.6

The final and best model from the BSGT distribution under the GAMLSS framework (3) is given by

$$\begin{aligned}\log(\mu) &= 7.796 - 0.094X_1^* + 0.759(if\ \mathrm{X}_2 = 2) + 1.231(if\ \mathrm{X}_2 = 3)\\ &\quad +0.208(if\ \mathrm{X}_6 = B) + 0.215(if\ \mathrm{X}_6 = C) + 0148(if\ \mathrm{X}_6 = D)\\ &\quad +0.390(if\ \mathrm{X}_6 = E) + 0.473(if\ \mathrm{X}_6 = F) - 0.563\mathrm{X}_8\end{aligned}$$

$$\begin{aligned}\log(\sigma) &= 1.344 - pbX_1^* - 0.306(if\ \mathrm{X}_2 = 2) - 0.610(if\ \mathrm{X}_2 = 3)\\ &\quad +0.160(if\ \mathrm{X}_6 = B) + 0.157(if\ \mathrm{X}_6 = C) + 0.098(if\ \mathrm{X}_6 = D)\\ &\quad +0.129(if\ \mathrm{X}_6 = E) + 0.129(if\ \mathrm{X}_6 = F) - 0.397\mathrm{X}_8\end{aligned}$$

$$\begin{aligned}\log(\nu) &= -3.948 + 0.003(if\ \mathrm{X}_4 = 2) + 0.073(if\ \mathrm{X}_4 = 3) + 0.166(if\ \mathrm{X}_4 = 4)\\ &\quad -0.086(if\ \mathrm{X}_6 = B) - 0.262(if\ \mathrm{X}_6 = C) - 0.442(if\ \mathrm{X}_6 = D)\\ &\quad -0.0635(if\ \mathrm{X}_6 = E) - 0.856(if\ \mathrm{X}_6 = F)\end{aligned}$$

and

$$\begin{aligned}\log(\tau) &= 4.747 + 0.449(if\ \mathrm{X}_6 = \mathrm{B}) + 0.475(if\ \mathrm{X}_6 = \mathrm{C}) + 1.009(if\ \mathrm{X}_6 = \mathrm{D})\\ &\quad +1.185(if\ \mathrm{X}_6 = E) + 1.099(if\ \mathrm{X}_6 = F)\end{aligned} \tag{4}$$

We can note that the four covariates were considered on the location parameter μ in the final BSGT model and both of the quantitative ones did not require any smoothing function. From the model for the median μ in (4), we observe that the higher is the vehicle value the lower is the median total claim amount which is somewhat unexpected. The same occurs with the exposure during the year. From the other two covariates considered in μ, we can say analyzing Figures 2(a) and (b) that the greater is the number of claims, greater will be the median total claim amount and that people living in areas E and F tend to have higher median claim amounts, respectively.

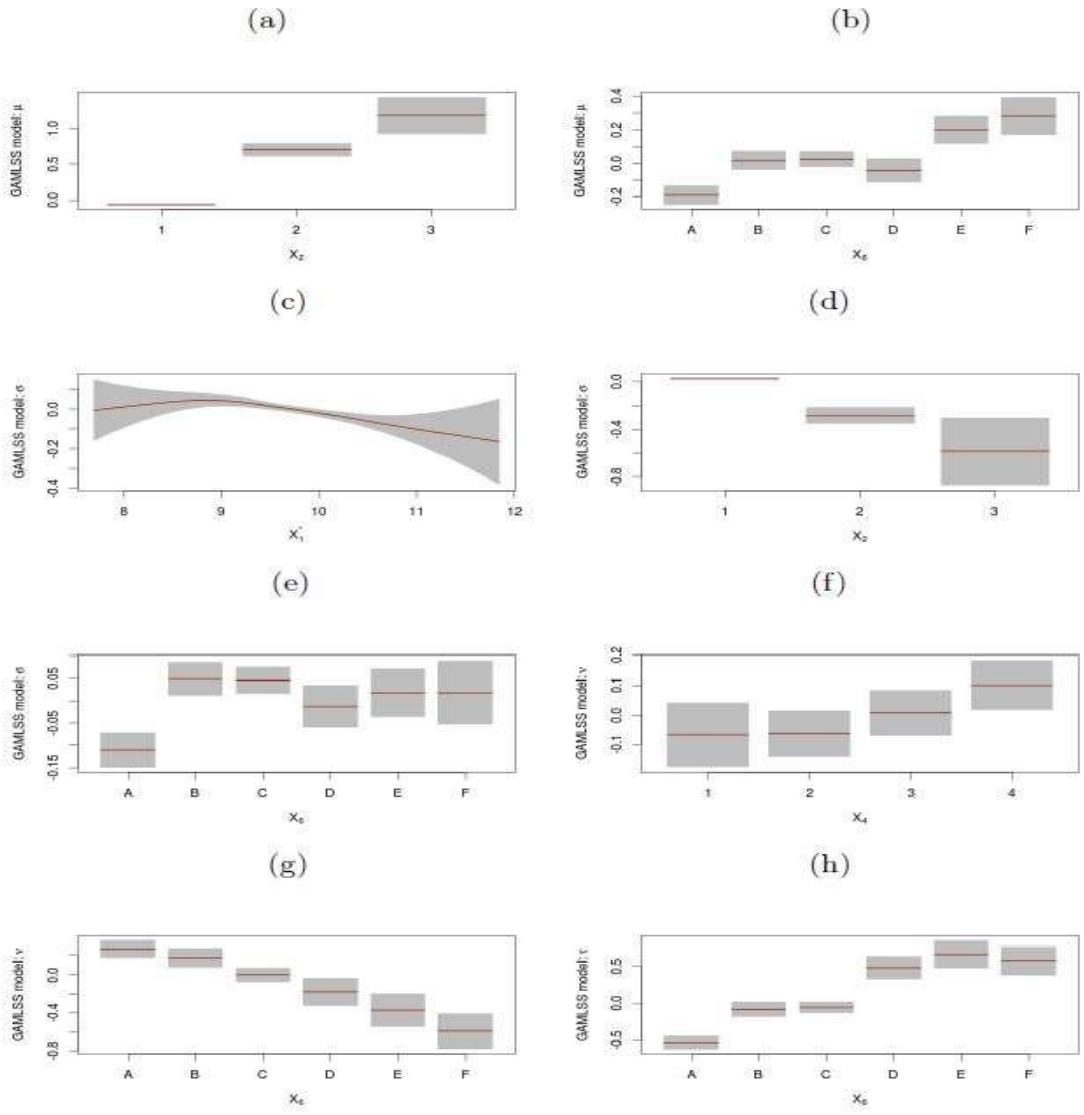

Figure 2: Regression terms for parameter (a) and (b) μ; (c), (d) and (e) σ; (f) and (g) ν; and (h) τ. Note that linear relationships were omitted

As it was observed in model (4), we jusy need a smoothing function to model the covariate X_1^* in σ. This relationship is showed in Figure 2(c) and we can note that for lower vehicle values there is a positive effect on the dispersion and after a certain point this relation becomes negative. Figures (d) and (e) present the relationship between the number of claims and driver's area of residence, respectively, with the dispersion. Further, the exposure during the year has a negative linear effect on dispersion. Figures 2(f)−(h) represent the relationships between selected covariates with the tails of the distribution.

Finally, the histogram and Q-Q plot of the normalized quantile residuals (Dunn and Smyth, 1996) of model (4) are displayed in Figure 3. Figure 3(a), apart from one outlier, show us that the residuals adequately follow a normal distribution. Figure 3(b) confirms

this outlier and also shows that there are a few points off the line in the high end of the range, but in general, the BSGT regression model provides a good fit to these data.

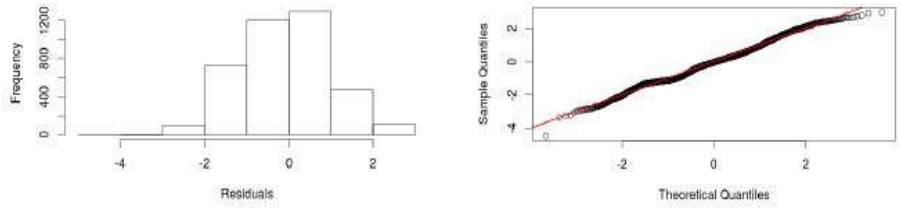

Figure 3: (a) Histogram and (b) Q-Q plot of the normalized quantile residuals from the BSGT fitted regression model

6. Concluding remarks

In this paper, we used the Birnbaum-Saunders generalized t (BSGT) distribution proposed by Genc (2013) which admits light or heavy tails, shaper or flatter peaked shape and it has some important special cases studied in the literature. Based on this distribution, we proposed a BSGT regression model using the flexibility of the GAMLSS framework (Rigby and Stasinopoulos, 2005). The new regression model can be used as an alternative to model light and heavy-tailed response variables as parametric and/or additive nonparametric smooth functions of explanatory variables. Hence, this extended regression model is very flexible in many practical situations. Moreover, we conducted a simulation study using 12 different scenarios in order to cover all possible BSGT density shapes: near symmetrical and very asymmetrical, light and heavy-tailed (i.e. platykurtic and leptokurtic). We also discussed model checking analysis using the normalized quantile residuals in the new regression model fitted to a real data. An application to insurance data set demonstrated that it can be used quite effectively to provide better fits than others flexible regression models.

Acknowledgements

The first author gratefully acknowledge grant from CAPES (Brazil) under the process number 99999.009857/2014-01.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.

[2] Barros, M., Paula, G.A. and Leiva, V. (2008). A new class of survival regression models with heavy-tailed errors: robustness and diagnostics. *Lifetime data analysis* 14, 316-332.

[3] Birnbaum, Z.W., and Saunders, S.C. (1969). A new family of life distributions. *Journal of Applied Probability* 6, 319-327.

[4] Cancho, V.G, Ortega, E.M.M. and Bolfarine, H. (2009). The log-exponentiated-Weibull regression models with cure rate: local influence and residual analysis. *Journal of Data Science* 7, 433-458.

[5] Castillo, N.O., Gomes, H.W. and Bolfarine, H. (2011). Epsilon Birnbaum-Saunders distribution family: properties and inference. *Statistical Papers* 52, 871-883.

[6] Cordeiro, G.M. and Lemonte, A.J. (2011). The beta Birnbaum-Saunders distribution: An improved distribution for fatigue life modeling. *Computational Statistics and Data Analysis* 55, 1445{1461.

[7] de Jong, P. and Heller, G.Z. (2008). *Generalized Linear Models for Insurance Data. Cambridge: Cambridge University Press*. 196 p.

[8] Daz-Garca, J.A. and Leiva, V. (2005). A new family of life distributions based on the elliptically contoured distributions. *Journal of Statistical Planning and Inference* 128, 445-457.

[9] Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5, 236-245.

[10] Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11, 89-102.

[11] Galea, M., Leiva, V. and Paula, G.A. (2004). Influence diagnostics in log-BirnbaumSaunders regression models. *Journal of Applied Statistics* 31, 1049-1064.

[12] Genc, A.I. (2013). The generalized T Birnbaum-Saunders family. *Statistics: A Journal of Theoretical and Applied Statistics* 47, 613-625.

[13] Gomes, H.W., Olivares-Pacheco, J.F., Bolfarine, H. (2009). An extension of the generalized Birnbaum-Saunders distribution. *Statistics and Probability Letters* 79, 331-338.

[14] Gupta, A.K. and Varga, T. (1993). *Elliptically Contoured Models in Statistics*. Boston: Springer. 335 p.

[15] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models. London*: Chapman and Hall/CRC. 352 p.

[16] Lawless, J.F. (2003). *Statistical models and methods for lifetime data.* 2.ed. New York: Wiley. 630p.

[17] Leiva, V., Barros, M., Paula, G.A. and Galea, M. (2007). Influence diagnostics in log-BirnbaumSaunders regression models with censored data. *Computational Statistics and Data Analysis* 79, 5694-5707.

[18] Leiva, V., Barros, M., Paula, G.A. and Sanhueza, A. (2008). Generalized Birnbaum-Saunders distributions applied to air pollutant concentration. *Environmetrics* 19, 235-249.

[19] Leiva, V., Sanhueza, A. and Angulo, J. M. A length-biased version of the Birnbaum-Saunders distribution with application in water quality. *Stochastic Environmental Research and Risk Assessment* 23, 299-307, 2009.

[20] Lemonte, A.J. and Cordeiro, G.M. (2009). Birnbaum-Saunders nonlinear regression models. *Computational Statistics and Data Analysis* 53, 4441-4452.

[21] McDonald, J.B. and Newey, W.K. (1988). Partially adaptive estimation of regression models via the generalized t distribution. *Econometric Theory* 4, 428-457.

[22] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* 135, 370-384.

[23] Nikulin, M. S. and Tahir, R. (2010). Application of Sedyakin's model and Birnbaum-Saunders family for statistical analysis of redundant systems with one warm stand-by unit. *Veroyatnost' i Statistika* 17, 155-171.

[24] Pascoa, M.A.R., de Paiva, C.M.M., Cordeiro, G.M. and Ortega, E.M.M. (2013). The log-Kumaraswamy generalized gamma regression model with application to chemical dependency data. *Journal of Data* Science 11, 781-818.

[25] Pescim, R.R., Cordeiro, G.M., Nadarajah, S., Demetrio, C.G.B. and Ortega, E.M.M. (2014). The Kummer beta Birnbaum-Saunders: An alternative fatigue life distribution. *Hacettepe Journal of Mathematics and Statistics* 43, 473-510.

[26] Podlaski, R. (2008). Characterization of diameter distribution data in near-natural forests using the Birnbaum-Saunders distribution. *Canadian Journal of Forest Research* 38, 518-527.

[27] Rieck, J.R. and Nedelman, J.R. (1991). A log-linear model for the Birnbaum-Saunders distribution. *Technometrics* 33, 51-60.

[28] Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics* 54, 507-554.

[29] Roman, M., Louzada, F., Cancho, V.G. and Leite, J.G. (2012). A new long-term survival distribution for cancer data. *Journal of Data Science* 10, 241-258.

[30] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics 6, 461{464.

[31] Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software* 23, 1-10.

[32] Tisionas, E.G. (2001). Bayesian inference in BirnbaumSaunders regression *Communications in Statistics - Theory and Methods* 30, 179-193.

[33] Vilca, F. and Leiva, V. (2006). A new fatigue life model based on the family of skew-elliptical distributions. *Communications in Statistics|Theory and Methods* 35, 229-244.

[34] Vilca, F., Sanhueza, A., Leiva, V. and Christakos, G. (2010). An extended Birnbaum-Saunders model and its application in the study of environmental quality in Santiago, Chile. *Stochastic Environmental Research and Risk Assessment* 24, 771-782.

[35] Villegas, C. Paula, G.A. and Leiva, V. (2011). Birnbaum-Saunders Mixed Models for Censored Reliability Data Analysis. *IEEE Transactions on Reliability* 748, 748-758.

[36] Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J. and Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics* 39, 1279-1293.

[37] Xie, F.C. and Wei, B.C. (2007). Diagnostics analysis for log-BirnbaumSaunders regression models *Computational Statistics and Data Analysis* 51, 4692- 4706.