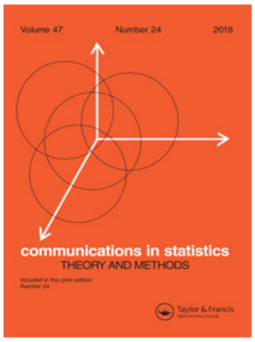

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

A survey of a hurdle model for heavy-tailed data based on the generalized lambda distribution

D. Marcondes, C. Peixoto & A. C. Maia

To cite this article: D. Marcondes, C. Peixoto & A. C. Maia (2018): A survey of a hurdle model for heavy-tailed data based on the generalized lambda distribution, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2018.1549251

To link to this article: https://doi.org/10.1080/03610926.2018.1549251

Published online: 31 Dec 2018.

Submit your article to this journal

Article views: 12

View Crossmark data

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2018.1549251

A survey of a hurdle model for heavy-tailed data based on the generalized lambda distribution

D. Marcondes[a], C. Peixoto[a], and A. C. Maia[b]

[a]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil; [b]Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo, São Paulo, Brazil

ABSTRACT

We present a literature review on the current state of Generalized Lambda Distribution (GλD) research and propose a highly flexible GλD hurdle model for heavy tailed data with excessive zeros. We apply the developed models to a dataset consisting of yearly health-care expenses, a typical example of heavy-tailed data with excessive zeros. The fitted GλDs are compared with models based on the Generalised Pareto Distribution and it is established that the GλD performs the best.

ARTICLE HISTORY
Received 19 April 2018
Accepted 11 November 2018

KEYWORDS
Generalized Lambda Distribution; Generalized Pareto Distribution; hurdle models; two-way models

1. Introduction

A motivation for the development of models for heavy-tailed data with excessive zeros arises from data on healthcare expenses, that is characterized by its heavy tails, great number of zeros and high skewness, which make fitting models to it a complex task (Jones, Lomas, and Rice 2014; Mihaylova et al. 2011). Indeed, a suitable choice of model for healthcare expenses are clumped-at-zero models, which are divided into two classes. The first class are the zero-inflated models, which are based on distributions that already have a probability mass at zero, that is then inflated. The zero-inflated Poisson is an element of this class (Lambert 1992). The second class are the two-part or hurdle models, that are those whose underlying distribution does not have a probability mass at zero, that is then added to it. They are called hurdle for the probability mass at zero may be seem as a *hurdle*, and they are also known as two-part models because the probability mass at zero and the non-zero values may be modeled independently, so the model has two parts. An example of hurdle model, for the demand of medical care, is presented in Duan et al. (1983). In the class of two-part models there are also those whose underlying distribution has a probability mass at zero, but are nonetheless two-part models, as the one in Mullahy (1986), since the inflation of the probability mass at zero is made by truncation, independently of the non-zero data.

The underlying distribution of the hurdle model treated in this paper is the Generalized Lambda Distribution (GλD), that is a highly flexible four-parameter continuous probability distribution. The GλD was first proposed by Ramberg and Schmeiser (1974), and then extended by Freimer et al. (1988), as a generalization of

CONTACT D. Marcondes dmarcondes@ime.usp.br Rua do Matão, 1010, São Paulo 05508-090, Brazil.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

Tukey's Lambda Distribution (Hastings et al. 1947; Tukey 1990). Even though the GλD is a *wild card* distribution, that well approximate others (Karian and Dudewicz 2000, Chapter 3), its use has been limited as there is no explicit expression for its probability density function, which makes the estimation of its parameters a complex task.

Nevertheless, there is a considerable amount of applications of it in the literature. As examples, we cite the evaluation of non-normal process capability indices (Pal 2004), option pricing (Corrado 2001), the fitting of solar radiation data (Öztürk and Dale 1982) and income data (Tarsitano 2004), and statistical process control (Fournier et al. 2006). Regarding the modeling of healthcare expenses, the GλD was studied by Balasooriya and Low (2008), where it was compared with the transformed kernel density and models of the exponential family, and it was established that the GλD fitted the data the best.

A limitation on the use of the GλD used to be the estimation of its parameters, which had been carried out by the methods of moments and a percentile method until Su (2007b) proposed a numerical maximum likelihood method. Another limitation used to be the lack of a regression model, that was just recently proposed by Su (2015), which extended the range of applications for the GλD. Therefore, due to recent advances in GλD theory, it is now possible to further apply this powerful distribution and compare it to other established models in order to assess its advantages.

In this paper, we develop hurdle GλD models and assess their goodness-of-fit on a yearly healthcare expenses dataset. The models developed seek to fit the data taking into account covariates (regression model) or not. The GλD models are compared with hurdle models based on the Generalized Pareto Distribution (GPD), which are special cases of the one in Couturier and Victoria-Feser (2010). The GPD is also a highly flexible continuous probability distribution, although we argue that it is not as flexible as, and do not fit the data as good as, the GλD. For an assessment of the goodness-of-fit of the GPD for healthcare expenses see Cebrián, Denuit, and Lambert (2003).

In Section 2 we present a survey on the current state of GλD research. In Section 3 we propose a hurdle GλD and present its main properties. In Section 4 we present a survey on GλD regression and develop a hurdle GλD regression model. In Section 5 we present a simulation study about the asymptotic properties of the hurdle GλD regression coefficients. In Section 6 we apply the developed methods to model healthcare expenses and compare GλD and GPD models.

2. The generalized lambda distribution

In this section we present the main properties of two distinct parametrizations of the GλD, known as RS and FKML GλD.

2.1. RS generalized lambda distribution

The RS GλD, proposed by Ramberg and Schmeiser (1974), is a generalization of Tukey's Lambda Distribution, obtained from an uniform random variable. Let U be an uniform random variable with range $[0,1]$ defined in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the random variable X_λ, also defined in $(\Omega, \mathcal{F}, \mathbb{P})$, and given by

$$X_{\boldsymbol{\lambda}} := Q_{\boldsymbol{\lambda}}(U) = \lambda_1 + \frac{U^{\lambda_3} - (1-U)^{\lambda_4}}{\lambda_2} \tag{1}$$

has an RS GλD with parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$, whose quantile functions is $Q_{\boldsymbol{\lambda}}(u), u \in [0, 1]$, as $Q_{\boldsymbol{\lambda}}(u) = F_{\boldsymbol{\lambda}}^{-1}(u)$, in which $F_{\boldsymbol{\lambda}}(x) = \mathbb{P}(X_{\boldsymbol{\lambda}} \leq x)$. The density of $X_{\boldsymbol{\lambda}}$ is given by

$$f_{\boldsymbol{\lambda}}(x) = \frac{1}{Q'_{\boldsymbol{\lambda}}(F_{\boldsymbol{\lambda}}(x))} = \frac{\lambda_2}{\lambda_3 F_{\boldsymbol{\lambda}}(x)^{\lambda_3 - 1} + \lambda_4 (1 - F_{\boldsymbol{\lambda}}(x))^{\lambda_4 - 1}} \tag{2}$$

in which $Q'_{\boldsymbol{\lambda}}(F_{\boldsymbol{\lambda}}(x))$ is the derivative of $Q_{\boldsymbol{\lambda}}$ at point $F_{\boldsymbol{\lambda}}(x)$. The parametric space $\Lambda = \{\boldsymbol{\lambda} \in \mathbb{R}^4 : F_{\boldsymbol{\lambda}} \text{ is a cumulative distribution function}\}$ of $\boldsymbol{\lambda}$ is a proper subset of $\mathbb{R}^4$ and is given implicitly by inequality

$$\frac{\lambda_3 u^{\lambda_3 - 1} + \lambda_4 (1-u)^{\lambda_4 - 1}}{\lambda_2} \geq 0 \tag{3}$$

for $0 \leq u \leq 1$, which is obtained by noting that $f_{\boldsymbol{\lambda}}(x) \geq 0$ if, and only if, $Q'_{\boldsymbol{\lambda}}(F_{\boldsymbol{\lambda}}(x)) \geq 0$.

The RS GλD is quite flexible, as it is possible to choose its parameters in order to obtain a distribution with given mean, variance, skewness and kurtosis. Indeed, the mean can be shifted to any value by choosing λ_1 properly, the skewness and kurtosis are determined by λ_3 and λ_4 and, given λ_3 and λ_4, the variance is determined by λ_2. The range of $X_{\boldsymbol{\lambda}}$ is $[Q_{\boldsymbol{\lambda}}(0), Q_{\boldsymbol{\lambda}}(1)]$ and depends on $\boldsymbol{\lambda}$ (see Karian and Dudewicz 2000, Theorem 1.4.23). The *kth* moment of the RS GλD exists if, and only if, $\min(\lambda_3, \lambda_4) > -k^{-1}$ and, when it exists and $\lambda_1 = 0$, it is given by

$$E\left(X_{\boldsymbol{\lambda}}^k\right) = \lambda_2^{-k} \sum_{i=0}^{k} \binom{k}{i} (-1)^i \beta\left(\lambda_3(k-i) + 1; \lambda_4 i + 1\right) \tag{4}$$

in which $\beta(a; b)$ is the beta function evaluated at (a, b). A proof for (4) is given in Ramberg and Schmeiser (1974). The central moments of $X_{\boldsymbol{\lambda}}$ when $\lambda_1 \neq 0$ may be obtained from (4) by applying the properties of the expectation operator. For instance, we have that

$$E(X_{\boldsymbol{\lambda}}) = \lambda_1 + \frac{(\lambda_3 + 1)^{-1} - (\lambda_4 + 1)^{-1}}{\lambda_2} \tag{5}$$

so that $E(X_{\boldsymbol{\lambda}}) = \lambda_1$ if, and only if, $\lambda_3 = \lambda_4$ and $X_{\boldsymbol{\lambda}}$ is symmetric.

The classical estimation technique for the RS GλD is the Method of Moments (MM), as introduced by Ramberg and Schmeiser (1974) and consolidated by Karian, Dudewicz, and Mcdonald (1996). Although easily implemented, the MM has some limitations. First of all, two different vectors $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda$ may yield the same first four moments of the RS GλD. As pointed out by Karian, Dudewicz, and Mcdonald (1996), this may be seen as a problem or an opportunity, for it enables a flexible fit for the data, as we may choose the parameters that best fulfill our objectives regarding the fit. Another limitation of the MM is the fact that the existence of the first four moments depends on $\boldsymbol{\lambda}$ and, therefore, it cannot be applied for a subset of Λ. Furthermore, simulation studies have showed that the MM performs worse than other methods, as the Numerical Maximum Likelihood Method (NMLM) and the percentile matching method, for example (Karian and Dudewicz 2003; Su 2007b).

Even though other estimation techniques, as the *least square* estimation method proposed by Öztürk and Dale (1985), the Starship Method developed by King and MacGillivray (1999), the flexible discretized approach proposed by Su (2005) and the percentile matching method, similar to the MM but with best results in simulation studies, as introduced by Karian and Dudewicz (1999) and further studied by Karian and Dudewicz (2000) and Karian and Dudewicz (2003), are available in the literature, this paper treats only estimation by the NMLM, as proposed by Su (2007b) and Su (2011). For a good account of other estimation techniques see Lakhany and Mausser (2000).

The log-likelihood of a sample $\{x_1, ..., x_n\}$ of an RS GλD random variable may be written in terms of the cumulative distribution function $F_{\boldsymbol{\lambda}}$, by denoting $u_i = F_{\boldsymbol{\lambda}}(x_i), i = 1, ..., n$, so that

$$l_{RS}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \log \left[\frac{\lambda_2}{\lambda_3 u_i^{\lambda_3 - 1} + \lambda_4 (1 - u_i)^{\lambda_4 - 1}} \right], \boldsymbol{\lambda} \in \Lambda. \tag{6}$$

In order to maximize (6) it is preferable to apply direct numerical methods than the usual method of differentiation, as they are much more reliable and efficient than solving the conventional linear equations on $\boldsymbol{\lambda}$, because, in many cases, the RS GλD may be undefined for certain $\boldsymbol{\lambda}$, as was pointed out by Su (2011). Therefore, we apply the algorithm proposed by Su (2007b) to maximize (6) numerically.

The main issue in maximizing (6) is in finding suitable initial values for the quantile sample $\{u_1, ..., u_n\}$. The most efficient way of obtaining them is through the estimation of $\boldsymbol{\lambda}$ by the percentile method, as this is the method that, apart from the NMLM, has had more efficient results estimating the RS GλD parameters (Karian and Dudewicz 2003).

The p*th* percentile of a sample $\{x_1, ..., x_n\}$ is defined as $\hat{\pi}_p = x_{(r)} + k(x_{(r+1)} - x_{(r)})$, in which $\{x_{(1)}, ..., x_{(n)}\}$ is the sample ordered in ascending order and r is the greatest integer lesser than $(n+1)p$, with $k = (n+1)p - r$. Rather than matching the sample moments to their theoretical values, in the percentile method we match the statistics

$$\hat{\rho}_1 = \hat{\pi}_{0.5} \quad \hat{\rho}_2 = \hat{\pi}_{1-v} - \hat{\pi}_v \quad \hat{\rho}_3 = \frac{\hat{\pi}_{0.5} - \hat{\pi}_v}{\hat{\pi}_{1-v} - \hat{\pi}_{0.5}} \quad \hat{\rho}_4 = \frac{\hat{\pi}_{0.75} - \hat{\pi}_{0.25}}{\hat{\rho}_2} \tag{7}$$

to their theoretical values, in which v is an arbitrary number between 0 and 0.25, that we choose to be 0.1, so that it is consistent with Karian and Dudewicz (2000) and Su (2007b).

Matching the theoretical values of ρ_1, ρ_2, ρ_3 and ρ_4 to the quantile function of an RS GλD we obtain the following identities:

$$\begin{aligned}
\rho_1(\boldsymbol{\lambda}) &= Q_{\boldsymbol{\lambda}}(0.5) = \lambda_1 + \frac{0.5^{\lambda_3} - 0.5^{\lambda_4}}{\lambda_2} \\
\rho_2(\boldsymbol{\lambda}) &= Q_{\boldsymbol{\lambda}}(1-v) - Q_{\boldsymbol{\lambda}}(v) = \frac{(1-v)^{\lambda_3} - v^{\lambda_3} + (1-v)^{\lambda_4} - v^{\lambda_4}}{\lambda_2} \\
\rho_3(\boldsymbol{\lambda}) &= \frac{Q_{\boldsymbol{\lambda}}(0.5) - Q_{\boldsymbol{\lambda}}(v)}{Q_{\boldsymbol{\lambda}}(1-v) - Q_{\boldsymbol{\lambda}}(0.5)} = \frac{(1-v)^{\lambda_4} - v^{\lambda_3} + 0.5^{\lambda_3} - 0.5^{\lambda_4}}{(1-v)^{\lambda_3} - v^{\lambda_4} + 0.5^{\lambda_4} - 0.5^{\lambda_3}} \\
\rho_4(\boldsymbol{\lambda}) &= \frac{Q_{\boldsymbol{\lambda}}(0.75) - Q_{\boldsymbol{\lambda}}(0.25)}{\rho_2} = \frac{0.75^{\lambda_3} - 0.25^{\lambda_4} + 0.75^{\lambda_4} - 0.25^{\lambda_3}}{\rho_2}.
\end{aligned} \tag{8}$$

The conditions $-\infty < \rho_1 < \infty, \rho_2 \geq 0$, $\rho_3 \geq 0$ and $\rho_4 \in [0,1]$ must be satisfied, as can be established from (7). In order to estimate $\boldsymbol{\lambda}$ we match the sample values (7) to their theoretical values (8) and solve numerically for $\boldsymbol{\lambda}$ by the Newton-Raphson method, for example, with the stopping rule given by the minimization of the Euclidean 2-norm $H(\boldsymbol{\lambda}) = ||(\rho_3(\boldsymbol{\lambda}), \rho_4(\boldsymbol{\lambda}))-(\hat{\rho}_3, \hat{\rho}_4)||_2$. Once λ_3 and λ_4 are obtained from the last two equations of (8), we substitute their values in the first two equations of (8) in order to obtain λ_1 and λ_2.

The percentile method is applied to get initial values in order to maximize (6). The maximization of (6) is performed by a 4-step algorithm proposed by Su (2007b),[1] which employs quasi random numbers and the percentile method. The algorithm is as follows:

1. Specify the range of initial values for λ_3 and λ_4 and the number of initial values to be generated. In this step, quasi random numbers are sampled as candidates for the initial values of λ_3 and λ_4. Su (2007b) proposes that 10, 000 quasi random values (scrambled so that the sampled values fill uniformly the considered space) be chosen from the square $[-1.5, 1.5]^2$.
2. Evaluate λ_1, λ_2 for each of the initial values of λ_3, λ_4 in the first two equations of (8). Remove all initial values that
   a. Do not result in a legal parametrization of the RS GλD by (3).
   b. Do not span the entire region of the dataset.
   Among the initial points not removed, find the initial set $\hat{\boldsymbol{\lambda}}_0$ that minimizes the norm $H(\boldsymbol{\lambda})$.
3. Calculate the quantiles $\{u_1, ..., u_n\}$ by solving numerically (1) with the initial values $\hat{\boldsymbol{\lambda}}_0$.
4. Once $\{u_1, ..., u_n\}$ is obtained, substitute them in (6) and solve it numerically for $\hat{\boldsymbol{\lambda}}$. It is convenient to repeat this process for different initials values, in order to check the consistency of the solution. The obtained estimator is called revised percentile estimator of the RS GλD under maximum likelihood estimation. The quality of the fitted distribution may be established by diagnostic techniques, as the data histogram superimposed by the estimated density, quantile plots and goodness-of-fit tests.

2.2. FKML generalized lambda distribution

The FKML GλD, proposed by Freimer, Kollia, Mudholkar, and Lin (1988), is also a four parameter generalization of Tukey's Lambda Distribution obtained from an uniform distribution. Indeed, let U be an uniform random variable with range $[0,1]$ defined in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then, the random variable $X_{\boldsymbol{\lambda}}$, also defined in $(\Omega, \mathcal{F}, \mathbb{P})$, and given by

$$X_{\boldsymbol{\lambda}} := Q_{\boldsymbol{\lambda}}(U) = \lambda_1 + \frac{1}{\lambda_2}\left[\frac{U^{\lambda_3}-1}{\lambda_3} - \frac{(1-U)^{\lambda_4}-1}{\lambda_4}\right] \tag{9}$$

has an FKML GλD with parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. The FKML G$\lambda$D is a probability distribution for all real-valued parameters $\boldsymbol{\lambda}$, with the restriction that $\lambda_2 > 0$ and the

[1] The algorithm in Su (2007b) has five steps, that we reduced to four, without loss of content.

conventions[2] that $X_{(\lambda_1,\lambda_2,0,\lambda_4)} = \lim_{\lambda_3 \to 0} X_{(\lambda_1,\lambda_2,\lambda_3,\lambda_4)}$ and $X_{(\lambda_1,\lambda_2,\lambda_3,0)} = \lim_{\lambda_4 \to 0} X_{(\lambda_1,\lambda_2,\lambda_3,\lambda_4)}$. The main motivation for generalizing Tukey's Lambda distribution to (9) is the weaker restrictions on its parametric space when comparing to the RS GλD, which facilitates the estimation of its parameters. Although both the RS and FKML GλD are generalizations of Tukey's Lambda Distribution, they are not equivalent, so the distribution fitted by one parametrization differs in general from the one fitted by the other.

The range of $X_{\boldsymbol{\lambda}}$ depends on $\boldsymbol{\lambda}$ and is given by $[Q_{\boldsymbol{\lambda}}(0), Q_{\boldsymbol{\lambda}}(1)]$. The density of the FKML GλD is obtained in a similar manner of (2) and is given by

$$f_{\boldsymbol{\lambda}}(x) = \frac{1}{Q'_{\boldsymbol{\lambda}}(F_{\boldsymbol{\lambda}}(x))} = \frac{\lambda_2}{F_{\boldsymbol{\lambda}}(x)^{\lambda_3 - 1} + (1 - F_{\boldsymbol{\lambda}}(x))^{\lambda_4 - 1}}. \tag{10}$$

The distribution of $X_{\boldsymbol{\lambda}}$ is symmetric if, and only if, $\lambda_3 = \lambda_4$, although its skewness measure may be zero for[3] $\lambda_3 \neq \lambda_4$. The parameters λ_3 and λ_4 determine single-handedly the nature and shape of the left and right tails of $X_{\boldsymbol{\lambda}}$, respectively, although the shape of the probability density function depends on both λ_3 and λ_4. Examples of FKML GλD may be found in Su (2015). Although the parameters of both the RS and FKML GλD are denoted by λ_1, λ_2, λ_3 and λ_4, and are related to the same properties of the distribution, they are not equivalent, nor comparable. Nevertheless, the FKML GλD is also highly flexible, as it is possible to choose $\boldsymbol{\lambda}$ so that $X_{\boldsymbol{\lambda}}$ has specific mean, variance, skewness and kurtosis. Furthermore, its tails are also flexible, so that the FKML GλD (and the RS GλD) provides a better fit for heavy tailed data than the usual Generalized Additive Models for Location, Scale and Shape (Rigby and Stasinopoulos, 2005), for example.

The *kth* moment of the FKML GλD also exists if, and only if, $\min(\lambda_3, \lambda_4) > -k^{-1}$. Denoting $a = 1/\lambda_2$ and $b = \lambda_1 - 1/\lambda_2\lambda_3 + 1/\lambda_2\lambda_4$, the *kth* moment of $X_{\boldsymbol{\lambda}}$ may be obtained from the moments of $(X_{\boldsymbol{\lambda}} - b)/a$ that, when exist, are given by

$$s_k := E\left(\left[\frac{X_{\boldsymbol{\lambda}} - b}{a}\right]^k\right) = \sum_{i=0}^{k} \binom{k}{i} (-1)^i \lambda_3^{-(k-i)} \lambda_4^{-i} \beta\big(\lambda_3(k-i) + 1; \lambda_4 i + 1\big) \tag{11}$$

as showed in Freimer, Kollia, Mudholkar, and Lin (1988) and Lakhany and Mausser (2000). The central moments of $X_{\boldsymbol{\lambda}}$ may be obtained from (11).

Although there is a vast literature about the estimation of the FKML GλD parameters, we treat only the NMLM as proposed by Su (2007b) and Su (2011). The log-likelihood of a sample $\{x_1, ..., x_n\}$ of an FKML GλD is given by

$$l_{FKML}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \log\left[\frac{\lambda_2}{u_i^{\lambda_3 - 1} + (1 - u_i)^{\lambda_4 - 1}}\right], \lambda_1, \lambda_3, \lambda_4 \in \mathbb{R}, \lambda_2 > 0 \tag{12}$$

in which $u_i = F_{\boldsymbol{\lambda}}(x_i), i = 1, ..., n$. The maximization of (12) is performed applying an algorithm slightly different from the one applied to maximize (6). The main issue in maximizing (12) is also in finding initial values for $\{u_1, ..., u_n\}$, and the estimation method, apart from the NMLM, that seems to perform the best for the FKML GλD is the method of moments, as outlined by the simulation studies of Lakhany and Mausser

[2] The same conventions apply to the RS GλD.

[3] This is also the case for the RS GλD.

(2000). Therefore, this is the method we use to find the initial values of $\{u_1, ..., u_n\}$ in a similar manner of what has been done for the RS GλD.

The method of moments for the FKML GλD, presented in Lakhany and Mausser (2000), consists on matching the first four sample moments of $\{x_1, ..., x_n\}$ given by

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i & \hat{\mu}_2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_1)^2 \\ \hat{\alpha}_3 &= \frac{1}{n(\hat{\mu}_2)^{1.5}}\sum_{i=1}^{n}(x_i - \hat{\mu}_1)^3 & \hat{\alpha}_4 &= \frac{1}{n(\hat{\mu}_2)^2}\sum_{i=1}^{n}(x_i - \hat{\mu}_1)^4 \end{aligned} \tag{13}$$

to their theoretical moments

$$\begin{aligned} \mu_1(\boldsymbol{\lambda}) &= \lambda_1 - \frac{1}{\lambda_2}\left(\frac{1}{\lambda_3 + 1} - \frac{1}{\lambda_4 + 1}\right) & \mu_2(\boldsymbol{\lambda}) &= \frac{1}{\lambda_2^2}\left(s_2 - s_1^2\right) \\ \alpha_3(\boldsymbol{\lambda}) &= \frac{s_3 - 3s_1s_2 + 2s_1^3}{\left(s_2 - s_1^2\right)^{3/2}} & \alpha_4(\boldsymbol{\lambda}) &= \frac{s_4 - 4s_1s_3 + 6s_1^2s_2 - 3s_1^4}{\left(s_2 - s_1^2\right)^2}. \end{aligned} \tag{14}$$

As proposed by Lakhany and Mausser (2000), we first solve numerically $(\alpha_3(\boldsymbol{\lambda}), \alpha_4(\boldsymbol{\lambda})) = (\hat{\alpha}_3, \hat{\alpha}_4)$ for λ_3 and λ_4 in the plane $(-1/4, \infty) \times (-1/4, \infty)$ by the minimization of the Euclidean 2-norm $H(\boldsymbol{\lambda}) = ||(\alpha_3(\boldsymbol{\lambda}), \alpha_4(\boldsymbol{\lambda})) - (\hat{\alpha}_3, \hat{\alpha}_4)||_2$, and then substitute their values in the first two equations of (14) to obtain λ_1 and λ_2. Using the estimates from the method of moments as initial values, we apply an algorithm analogous to the one applied to the RS GλD in order to obtain NMLM estimates. The algorithm was also proposed by Su (2007b), and is a slight modification of the algorithm of Section 2.1, in which the method of moments is used to find the initial values instead of the percentile method, and the FKML GλD likelihood is maximized, instead of the RS GλD one. More details about it may be found in Su (2007b).

3. Hurdle generalized lambda distribution

In this section we propose a Hurdle Generalized Lambda Distribution (HGλD), which is obtained by adding a fifth parameter λ_0 to either the RS or FKML GλD to represent their probability mass at zero.

3.1. Hurdle RS generalized lambda distribution

Let U and V be independent random variables defined in $(\Omega, \mathcal{F}, \mathbb{P})$, such that U is uniformly distributed in $[0, 1]$ and $\mathbb{P}(V = 1) = 1 - \mathbb{P}(V = 0) = \lambda_0$. We say that the random variable $Y_{\boldsymbol{\lambda}^*}$ given by

$$Y_{\boldsymbol{\lambda}^*} := Q^*_{\boldsymbol{\lambda}^*}(U, V) = (1 - V)\left(\lambda_1 + \frac{U^{\lambda_3} - (1 - U)^{\lambda_4}}{\lambda_2}\right) \tag{15}$$

has a hurdle RS GλD (HRS GλD) with parameters $\boldsymbol{\lambda}^* = (\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4)$ in the parametric space $\Lambda^* = [0, 1] \times \Lambda$.

The random variable $Y_{\boldsymbol{\lambda}^*}$ follows a mixed probability distribution, that has a probability mass λ_0 at zero and a probability mass $1 - \lambda_0$ spread over $[Q^*_{\boldsymbol{\lambda}^*}(0, 0), Q^*_{\boldsymbol{\lambda}^*}(1, 0)]$

according to an RS GλD. As the flexibility of the GλD is maintained in our hurdle generalization, an advantage of fitting an HRS GλD is that it is suitable for modeling data with heavy tails and skewness that also has excessive zeros.

3.1.1. Estimation

The estimation of the HRS GλD parameters may be performed by the NMLM, by extending the method of Su (2007b). We represent a sample of $Y_{\boldsymbol{\lambda}^*}$ by $\{(y_1, v_1), ..., (y_n, v_n)\}$, in which y_i are the observed values and[4] $v_i = \mathbb{1}\{y_i = 0\}$, $i = 1, ..., n$, so that the log-likelihood of $\boldsymbol{\lambda}^* = (\lambda_0, \boldsymbol{\lambda})$ is given by

$$l^*_{RS}(\boldsymbol{\lambda}^*) = l^*_{RS(1)}(\lambda_0) + l^*_{RS(2)}(\boldsymbol{\lambda}) \tag{16}$$

in which

$$\begin{cases} l^*_{RS(1)}(\lambda_0) = \sum_{i=1}^{n} v_i \log \lambda_0 + (1-v_i)\log(1-\lambda_0) \\ l^*_{RS(2)}(\boldsymbol{\lambda}) = \sum_{i=1}^{n} (1-v_i)\log\left[\frac{\lambda_2}{\lambda_3 u_i^{\lambda_3 - 1} + \lambda_4 (1-u_i)^{\lambda_4 - 1}}\right] \end{cases}.$$

As he log-likelihood (16) may be factored into two functions, one depending on λ_0 and other depending on $\boldsymbol{\lambda}$, λ_0 and $\boldsymbol{\lambda}$ are orthogonal and, therefore, may be estimated independently.

On the one hand, the maximum likelihood estimator of λ_0 is $\hat{\lambda}_0 = \frac{1}{n}\sum_{i=1}^{n} v_n$. On the other hand, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ may be estimated by applying the algorithm of Section 2.1 to the non-zero data values, so that we obtain the revised percentile estimator $\hat{\boldsymbol{\lambda}}^*$ of the HRS GλD under maximum likelihood estimation. As the estimated distribution fits the zero data values perfectly, it is enough to apply diagnostic techniques to the non-zero data values, e.g., by comparing graphically their histogram with the density of an RS GλD with parameters $(\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$.

3.2. Hurdle FKML generalized lambda distribution

The hurdle FKML GλD (HFKML GλD) is constructed in the same manner as the HRS GλD, by letting U and V be independent random variables defined in $(\Omega, \mathcal{F}, \mathbb{P})$, such that U is uniformly distributed in $[0, 1]$ and $\mathbb{P}(V = 1) = 1 - \mathbb{P}(V = 0) = \lambda_0$, and defining the random variable $Y_{\boldsymbol{\lambda}^*}$ as

$$Y_{\boldsymbol{\lambda}^*} := Q^*_{\boldsymbol{\lambda}^*}(U, V) = (1-V)\left(\lambda_1 + \frac{1}{\lambda_2}\left[\frac{U^{\lambda_3} - 1}{\lambda_3} - \frac{(1-U)^{\lambda_4} - 1}{\lambda_4}\right]\right) \tag{17}$$

so that $Y_{\boldsymbol{\lambda}^*}$ has an HFKML GλD with parameters $\boldsymbol{\lambda}^* = (\lambda_0, \lambda_1, \lambda_2, \lambda_3, \lambda_4) \in [0, 1] \times \mathbb{R}^4$ with the restriction that $\lambda_2 > 0$ and the same conventions of (9). The random variable $Y_{\boldsymbol{\lambda}^*}$ also follows a mixed probability distribution with the same general characteristics of

[4] $\mathbb{1}$ is the indicator function.

the HRS GλD: it is highly flexible, has a probability mass λ_0 at zero and a probability mass $1-\lambda_0$ spread over $[Q^*_{\lambda^*}(0,0), Q^*_{\lambda^*}(1,0)]$ according to an FKML GλD.

3.2.1. Estimation

The estimation of the HFKML GλD is performed in a way analogous to that of the HRS GλD, as the log-likelihood of an HFKML GλD sample $\{(y_1,v_1),\dots,(y_n,v_n)\}$, $v_i = \mathbb{1}\{y_i=0\}, i=1\dots,n$, may be factored as

$$l^*_{FKML}(\boldsymbol{\lambda}^*) = l^*_{FKML(1)}(\lambda_0) + l^*_{FKML(2)}(\boldsymbol{\lambda}), \tag{18}$$

in which

$$\begin{cases} l^*_{FKML(1)}(\lambda_0) = \displaystyle\sum_{i=1}^{n} v_i \log\lambda_0 + (1-v_i)\log(1-\lambda_0) \\ l^*_{FKML(2)}(\boldsymbol{\lambda}) = \displaystyle\sum_{i=1}^{n}(1-v_i)\log\left[\frac{\lambda_2}{u_i^{\lambda_3-1}+(1-u_i)^{\lambda_4-1}}\right], \end{cases}$$

so that λ_0 and $(\lambda_1,\lambda_2,\lambda_3,\lambda_4)$ are orthogonal, and may be estimated independently.

In order to obtain the revised method of moments estimator $\hat{\boldsymbol{\lambda}}^*$ of the HFKML GλD under maximum likelihood estimation, we estimate λ_0 by the proportion of zero data values $\hat{\lambda}_0 = \frac{1}{n}\sum_{i=1}^{n} v_i$ and $(\lambda_1,\lambda_2,\lambda_3,\lambda_4)$ by applying the algorithm of Section 2.1 to the non-zero data values. Diagnostic methods may be applied to the non-zero data values in order to assess the quality of the fitted model.

4. Hurdle generalized lambda distribution regression

In this section we propose a regression model for the HGλD, in which we model its location and probability mass at zero as functions of covariates $\boldsymbol{W}$ and $\boldsymbol{Z}$, respectively, which are random vectors defined in $(\Omega,\mathcal{F},\mathbb{P})$, that may share some variables or be equal. Our method is an adaptation of the one presented in Su (2015). We first outline the method of Su (2015) and then extend it to the HGλD.

4.1. Flexible parametric quantile regression model

The algorithm of Su (2015) seeks to estimate the parameters $(\boldsymbol{\beta},\lambda_2,\lambda_3,\lambda_4)$ of model

$$X|\boldsymbol{W} = \boldsymbol{W}^T\boldsymbol{\beta} + \epsilon \tag{19}$$

in which $\epsilon \sim G\lambda D(\lambda_1^*,\lambda_2,\lambda_3,\lambda_4)$ and λ_1^* is such that $E(\epsilon)=0$, i.e.,

$$\lambda_1^* = \begin{cases} -\dfrac{(\lambda_3+1)^{-1}-(\lambda_4+1)^{-1}}{\lambda_2} & \text{for the RS G}\lambda\text{D} \\ \dfrac{(\lambda_3+1)^{-1}-(\lambda_4+1)^{-1}}{\lambda_2} & \text{for the FKML G}\lambda\text{D} \end{cases}. \tag{20}$$

In order to estimate the parameters of (19) we apply a 5-step algorithm that is analogous to the algorithm of Section 2.1: find initial values in order to evaluate and maximize the log-likelihood to get NMLM estimates.

Let $\{(x_1, \boldsymbol{w}_1), \ldots, (x_n, \boldsymbol{w}_n)\}$ be a sample of the response variable and covariates. The algorithm is as follows and more details about it are presented in Su (2015).

1. Obtain $\hat{\boldsymbol{\beta}}^{(0)}$ from the least square method by solving
$$\hat{\boldsymbol{\beta}}^{(0)} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(x_i - \boldsymbol{w_i}^T \boldsymbol{\beta}\right)^2$$
and calculate the initial residuals $\hat{e}_i^{(0)} = x_i - \boldsymbol{w_i}^T \hat{\boldsymbol{\beta}}^{(0)}$.
2. Obtain the initial estimates $(\hat{\lambda}_2^{(0)}, \hat{\lambda}_3^{(0)}, \hat{\lambda}_4^{(0)})$ by applying the algorithm of Section 2.1 to sample $\{e_1^{(0)}, \ldots, e_n^{(0)}\}$ of ϵ.
3. Calculate the log-likelihood of the model:
    (a) Evaluate $\hat{\lambda}_1^{*(0)}$ by (20) so that the initial estimated distribution of error ϵ has zero mean.
    (b) Force the residuals sample mean to be zero by making
$$e_i^* = \left(y_i - \boldsymbol{w}_i^T \boldsymbol{\beta}\right) - \frac{1}{n} \sum_{i=1}^{n} e_i^*.$$
    (c) Evaluate the log-likelihood of the zero mean residuals from equations (6) or (12):
        (i) For the RS GλD with $\boldsymbol{\lambda} \in \Lambda$
$$l_{e^*}(\boldsymbol{\beta}, \lambda_2, \lambda_3, \lambda_4) = \sum_{i=1}^{n} \log \left[\frac{\lambda_2}{\lambda_3 u_i^{\lambda_3 - 1} + \lambda_4 (1 - u_i)^{\lambda_4 - 1}}\right] \tag{21}$$
$$e_i^* = \lambda_1^* + \frac{u_i^{\lambda_3} - (1 - u_i)^{\lambda_4}}{\lambda_2} \tag{22}$$
        (ii) For the FKML GλD with $\lambda_3, \lambda_4 \in \mathbb{R}, \lambda_2 > 0$
$$l_{e^*}(\boldsymbol{\beta}, \lambda_2, \lambda_3, \lambda_4) = \sum_{i=1}^{n} \log \left[\frac{\lambda_2}{u_i^{\lambda_3 - 1} + (1 - u_i)^{\lambda_4 - 1}}\right] \tag{23}$$
$$e_i^* = \lambda_1^* + \frac{1}{\lambda_2}\left[\frac{u_i^{\lambda_3} - 1}{\lambda_3} - \frac{(1 - u_i)^{\lambda_4} - 1}{\lambda_4}\right] \tag{24}$$

    in which u_i is given implicitly by (22) and (24), depending on the parametrization, and is a function of $(\boldsymbol{\beta}, \lambda_2, \lambda_3, \lambda_4)$.
4. Maximize numerically, by the Nelder-Mead simplex algorithm (Nelder and Mead 1965), for example, the log-likelihood (21) or (23), depending on the parametrization, employing $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\lambda}_2^{(0)}, \hat{\lambda}_3^{(0)}$ and $\hat{\lambda}_4^{(0)}$ as initial values, in order to obtain $\hat{\boldsymbol{\beta}}, \hat{\lambda}_2, \hat{\lambda}_3$ and $\hat{\lambda}_4$.
5. Obtain $\hat{\lambda}_1^*$ substituting the estimated values $\hat{\lambda}_2, \hat{\lambda}_3$ and $\hat{\lambda}_4$ in (20).
6. Conduct simulations to obtain statistical properties of the estimated regression coefficients $\hat{\beta}$ as follows:

    a. Generate $\{\epsilon_1, \ldots, \epsilon_n\}$ from the GλD with parameters $(\hat{\lambda}_1^*, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$ and obtain a new sample $\{x_1^*, \ldots, x_n^*\}$ by adding $x_i^* = \boldsymbol{w_i}^T \hat{\boldsymbol{\beta}} + \epsilon_i, i = 1, \ldots, n$. Fit a regression model to $\{x_1^*, \ldots, x_n^*\}$ obtaining estimates for the regression coefficients.

b. Repeat step (a) 1000 times to obtain 1000 coefficients.[5]
c. Adjust each coefficient sample in (b) so that its mean is equal to the final estimated coefficients of step 5. The simulated coefficients histogram may be plotted and $(1-\alpha)\%$ confidence intervals may be found by evaluating the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulated samples. We use quantile type 8 from the quantile function in **R** (Hyndman and Fan 1996; R Core Team 2017) in order to be consistent to Su (2015).

Any other method could be used to estimate the parameters of the error distribution in step 2. However, we prefer the NMLM for it provides better estimates, as has been established in the literature, although it may not converge in some cases. A limitation of this method is the lack of asymptotic theoretical results about the distribution of the estimators, so that we cannot construct asymptotic confidence intervals for the coefficients, nor test hypothesis. Nevertheless, computational methods for generating confidence intervals and establishing goodness-of-fit are implemented and can be applied (Su 2016).

4.2. HGλD regression model

In order to develop an HGλD regression model, we rely on the factorization of the log-likelihoods (16) and (18), which allows to model the probability mass at zero and the location of the distribution independently. Indeed, our regression model, whose response variable is Y and covariates are[6] $(\boldsymbol{W}, \boldsymbol{Z})$, may be written as

$$\begin{cases} Y|(\boldsymbol{W},\boldsymbol{Z}) = (1-(V|\boldsymbol{Z}))\left(\boldsymbol{W}^T\boldsymbol{\beta}+\epsilon\right) \\ \log\left(\dfrac{\mathbb{P}(V=1|\boldsymbol{Z})}{1-\mathbb{P}(V=1|\boldsymbol{Z})}\right) = \boldsymbol{Z}^T\boldsymbol{\gamma} \end{cases} \tag{25}$$

in which $\epsilon \sim G\lambda D(\lambda_1^*, \lambda_2, \lambda_3, \lambda_4)$ and λ_1^* is such that $E(\epsilon) = 0$.

Given a sample $\{(y_1, v_1, \boldsymbol{w}_1, \boldsymbol{z}_1), \ldots, (y_n, v_n, \boldsymbol{w}_n, \boldsymbol{z}_n)\}$ of model (25), in which $v_i = \mathbb{1}\{y_i = 0\}, i = 1, \ldots, n$, its log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda_2, \lambda_3, \lambda_4) &= \sum_{i=1}^{n} \log\left(\left[\frac{\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)}{1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)}\right]^{v_i}\left[\frac{f\left(y_i-\boldsymbol{w}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)}\right]^{(1-v_i)}\right) \\ &= \sum_{i=1}^{n} v_i\boldsymbol{z}_i^T\boldsymbol{\gamma} - \log\left(1+\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}\right)\right) + (1-v_i)\log\left(f\left(y_i-\boldsymbol{w}_i^T\boldsymbol{\beta}\right)\right) \\ &:= l_1^*(\boldsymbol{\gamma}) + l_2^*(\boldsymbol{\beta}, \lambda_2, \lambda_3, \lambda_4) \end{aligned} \tag{26}$$

in which $f(y_i - \boldsymbol{w}_i^T\boldsymbol{\beta})$ is either the density (2) or (10) with parameters $(\lambda_1^*, \lambda_2, \lambda_3, \lambda_4)$ evaluated at point $y_i - \boldsymbol{w}_i^T\boldsymbol{\beta}, i = 1, \ldots, n$.

[5] The number 1000 is arbitrary. It could be sampled more or less coefficients.

[6] $\boldsymbol{W}$ and $\boldsymbol{Z}$ are random vectors which may share some variables or be equal.

The estimation of model (25) may be performed by maximizing $l_1^*(\gamma)$ and $l_2^*(\boldsymbol{\beta}, \lambda_2, \lambda_3, \lambda_4)$ independently, so that we get the maximum likelihood estimator $\hat{\gamma}$ and the NMLM estimators $\hat{\boldsymbol{\beta}}, \hat{\lambda}_1^*, \hat{\lambda}_2, \hat{\lambda}_3$ and $\hat{\lambda}_4$. On the one hand, the maximization of $l_1^*(\gamma)$ is performed by fitting a logistic regression in the usual manner (see Hilbe (2009)) to sample $\{(v_1, \boldsymbol{z}_1), \ldots, (v_n, \boldsymbol{z}_n)\}$. On the other hand, the maximization of $l_2^*(\boldsymbol{\beta}, \lambda_2, \lambda_3, \lambda_4)$ is performed by applying the algorithm of Section 4.1 to the non-zero data values.

As γ and $(\boldsymbol{\beta}, \lambda_1^*, \lambda_2, \lambda_3, \lambda_4)$ are orthogonal, their maximum likelihood estimators are asymptotically independent (Cox and Reid 1987). Therefore, the usual methods of inference for logistic regression models may be applied to infer about γ, and logistic regression diagnostic techniques may be employed to asses the quality of the fit. However, as the estimators $\hat{\boldsymbol{\beta}}, \hat{\lambda}_1^*, \hat{\lambda}_2, \hat{\lambda}_3$ and $\hat{\lambda}_4$ are not of maximum likelihood, the usual inference techniques for maximum likelihood estimators cannot be applied to them. Nevertheless, we may construct numerical confidence intervals for $\boldsymbol{\beta}$ by applying the algorithm of Section 4.1 (step 6) to the non-zero data values.

The goodness-of-fit of HGλD regression models may be established by the study of two types of residuals: error residuals and normalized quantile residuals. The error residuals are given by $e = y - \boldsymbol{w}^T\hat{\beta}$ for $y \neq 0$ and their empirical distribution may be compared with the GλD$(\hat{\lambda}_1^*, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$, fitted to error ϵ, in order to establish goodness-of-fit. This comparison may be performed by the use of QQ-plots; a histogram of e superimposed by the estimated density; and a quantile plot that superimposes the estimated and empirical quantile functions of ϵ and e, respectively.

The normalized quantile residuals, presented in Dunn and Smyth (1996), are defined as $r = \Phi^{-1}(F_{\hat{\lambda}}^*(y - \boldsymbol{w}^T\hat{\beta}))$, in which Φ and $F_{\hat{\lambda}}^*$ are the cumulative distribution function of the standard normal distribution and the RS or FKML GλD$(\hat{\lambda}_1^*, \hat{\lambda}_2, \hat{\lambda}_3, \hat{\lambda}_4)$, respectively. The normalized quantile residuals are expected to be normally distributed if the model is properly fitted, so that we may regard the model as well fitted if the density estimate of r is close to the standard normal distribution density and the points of the normal QQ-plot of r are distributed around the line with intercept zero and slope one, for example. Normalised quantile residuals may also be employed to asses the goodness-of-fit of the logistic regression model (Rigby and Stasinopoulos 2005).

5. Simulation study

In this section we perform a simulation study in order to assess the asymptotic properties of the HGλD regression coefficients. We consider the model

$$\begin{cases} Y|(x_1, x_2) = (1 - (V|x_1, x_2))(6.13 - 0.021x_1 - 0.35x_2 + \epsilon) \\ \log\left(\dfrac{\mathbb{P}(V = 1|x_1, x_2)}{1 - \mathbb{P}(V = 1|x_1, x_2)}\right) = 1.6 - 0.13x_1 + 0.21x_2 \end{cases} \tag{27}$$

in which $x_1 \sim RS\ G\lambda D(3.87, 0.10, 0.024, 0.19)$ and $\mathbb{P}(x_2 = 1) = 1 - \mathbb{P}(x_2 = 0) = 0.6$. We simulate four different scenarios, in which the distribution of error ϵ is symmetric (RS GλD(0,2,0.13,0.13) and FKML GλD(0,2,0.13,0.13)), and right skewed (RS GλD($-$1.43,0.11,0.0023,0.19) and FKML GλD($-$0.147,$-$0.41,1.07,0.84,0.02)).

For each scenario and sample size $n = 100$, 200, 1000, we generate 1000 samples and, for each sample, we fit a HGλD regression model, estimating the coefficients of (27). We then study the mean, standard error, and 2.5th and 97.5th percentiles of the estimated coefficients over the 1000 samples. The results are presented in Table 1.

We observe, in all scenarios, that the mean of the estimated coefficients is close to the target value, especially for samples of size $n = 1000$, which is evidence that the estimators are unbiased. Furthermore, we see that as greater the sample size, the smaller is the standard error of the estimated coefficients, which is evidence that the estimators are consistent. Overall, the simulation study support the consistency of the estimators, so that it is not lost when we consider the hurdle model: the logistic regression consistency, theoretically established, and the consistency of the GλD regression, supported by the simulations of Su (2015), seems to be preserved when we consider the hurdle model.

6. Fitting an HGλD to healthcare expenses data

Healthcare expenses data has some peculiarities which make the HGλD a great option for modeling it: a great number of zeros, normally more than 50% of the data, as not every customer uses their health insurance in the period of a year; high skewness; and a heavy right tail that is hardly modeled by the usual distributions, as the Gamma, Weibull, Log-normal and Inverse-Gaussian.

In the following sections, we model a dataset that contains the yearly expenses of all insured customers of a Brazilian healthcare insurance company between 2006 and 2009. Our analysis focuses on modeling the yearly expenses in function of the covariates age, sex and previous year expenses. All expenses are in *Reais*[7] (R$) and were deflated to January 2006 value. The HGλD models are compared with GPD models in order to establish which better fits the data.

The GPD, introduced by Pickands (1975), is a three parameter positive probability distribution with density

$$f(y) = \begin{cases} \dfrac{1}{\tau}\left(1+\xi\dfrac{y-\alpha}{\tau}\right)^{-\frac{\xi+1}{\xi}} & \xi \neq 0 \\ \dfrac{1}{\tau}\exp\left(-\dfrac{y-\alpha}{\tau}\right) & \xi = 0 \end{cases}$$

for $y \geq \alpha$, in which $\alpha \geq 0$ is the location, $\tau > 0$ the scale and $\xi \in \mathbb{R}$ the shape parameter. The mean of the GPD is finite only for $\xi < 1$ and is given by

$$E(Y) := \mu = \alpha + \frac{\tau}{1-\xi}. \tag{28}$$

Note that the GPD may be re-parametrized so that μ is the scale parameter, instead of τ. Given a sample $\{y_1, \ldots, y_n\}$ and a known threshold α, (ξ, τ) (or (ξ, μ)) may be estimated by the Maximum Likelihood Method in the usual manner (see Hosking and Wallis (1987) and Grimshaw (1993)).

[7] Brazilian currency.

Table 1. Mean, standard error and 2.5th and 97.5th percentiles of samples of the coefficients of (27).

			Sample		Standard	Percentiles	
Distribution of ϵ	Coefficient	Target	size	Mean	error	2.5th	97.5th
HRS GλD(0,2,0.13,0.13)	Non-zero intercept	6.12	100	6.131	0.091	5.959	6.309
			200	6.130	0.059	6.009	6.244
			1000	6.130	0.022	6.085	6.174
	Non-zero x_1	−0.021	100	−0.021	0.016	−0.052	0.008
			200	−0.021	0.010	−0.040	0.001
			1000	−0.021	0.004	−0.028	−0.013
	Non-zero x_2	−0.35	100	−0.350	0.048	−0.446	−0.255
			200	−0.350	0.033	−0.416	−0.284
			1000	−0.350	0.012	−0.375	−0.326
	Zero intercept	1.6	100	1.659	0.944	−0.001	3.762
			200	1.627	0.650	0.401	2.984
			1000	1.611	0.272	1.097	2.136
	Zero x_1	−0.13	100	−0.136	0.164	−0.464	0.191
			200	−0.131	0.110	−0.362	0.083
			1000	−0.132	0.046	−0.223	−0.043
	Zero x_2	0.21	100	0.190	0.492	−0.768	1.139
			200	0.207	0.341	−0.481	0.829
			1000	0.216	0.149	−0.072	0.509
HFKML GλD(0,2,0.13,0.13)	Non-zero intercept	6.12	100	6.145	0.750	4.615	7.664
			200	6.114	0.489	5.215	7.215
			1000	6.123	0.176	5.780	6.470
	Non-zero x_1	−0.021	100	−0.022	0.129	−0.283	0.243
			200	−0.018	0.083	−0.201	0.131
			1000	−0.020	0.029	−0.077	0.034
	Non-zero x_2	−0.35	100	−0.359	0.409	−1.124	0.469
			200	−0.353	0.266	−0.876	0.165
			1000	−0.346	0.094	−0.532	−0.158
	Zero intercept	1.6	100	1.623	0.916	0.040	3.577
			200	1.620	0.649	0.388	2.975
			1000	1.611	0.272	1.097	2.136
	Zero x_1	−0.13	100	−0.136	0.159	−0.438	0.171
			200	−0.131	0.108	−0.362	0.081
			1000	−0.132	0.046	−0.223	−0.043
	Zero x_2	0.21	100	0.206	0.482	−0.735	1.149
			200	0.205	0.342	−0.484	0.834
			1000	0.216	0.149	−0.072	0.509
HRS GλD(−1.43,0.11, 0.0023,0.19)	Non-zero intercept	6.12	100	6.080	0.784	4.434	7.697
			200	6.102	0.448	5.200	7.003
			1000	6.115	0.179	5.742	6.451
	Non-zero x_1	−0.021	100	−0.014	0.134	−0.273	0.300
			200	−0.019	0.073	−0.158	0.134
			1000	−0.020	0.028	−0.080	0.041
	Non-zero x_2	−0.35	100	−0.337	0.403	−1.197	0.518
			200	−0.339	0.210	−0.738	0.078
			1000	−0.342	0.067	−0.462	−0.185
	Zero intercept	1.6	100	1.623	0.906	−0.045	3.534
			200	1.632	0.620	0.433	2.924
			1000	1.616	0.273	1.068	2.152
	Zero x_1	−0.13	100	−0.131	0.160	−0.442	0.188
			200	−0.134	0.107	−0.349	0.083
			1000	−0.132	0.047	−0.223	−0.038
	Zero x_2	0.21	100	0.211	0.493	−0.776	1.171
			200	0.209	0.328	−0.447	0.848
			1000	0.213	0.155	−0.073	0.507
	Non-zero intercept	6.12	100	6.139	0.777	4.584	7.906
			200	6.083	0.411	5.259	6.861
			1000	6.130	0.110	5.923	6.350

(*continued*)

Table 1. Continued.

Distribution of ϵ	Coefficient	Target	Sample size	Mean	Standard error	Percentiles	
						2.5th	97.5th
HFKML GλD(−0.147,−0.41, 1.07,0.84,0.02)	Non-zero x_1	−0.021	100	−0.022	0.133	−0.309	0.236
			200	−0.011	0.066	−0.149	0.128
			1000	−0.021	0.015	−0.050	0.011
	Non-zero x_2	−0.35	100	−0.352	0.408	−1.232	0.513
			200	−0.366	0.187	−0.742	0.019
			1000	−0.351	0.042	−0.446	−0.268
	Zero intercept	1.6	100	1.635	0.916	−0.098	3.433
			200	1.615	0.652	0.402	2.962
			1000	1.613	0.279	1.054	2.154
	Zero x_1	−0.13	100	−0.132	0.157	−0.434	0.187
			200	−0.132	0.111	−0.353	0.086
			1000	−0.132	0.048	−0.225	−0.035
	Zero x_2	0.21	100	0.194	0.509	−0.820	1.246
			200	0.200	0.338	−0.459	0.847
			1000	0.212	0.154	−0.073	0.507

In order to fit a GPD when there are covariates, we employ a generalized linear model (GLM) framework, as introduced by Nelder and Baker (1972). In this framework, we suppose that the location parameter α is known and independent of the covariates, and that the shape parameter ξ is unknown, but is lesser than one and independent of the covariates. Then, we model the mean μ as $E(Y|\boldsymbol{x}_i) := \mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$, in which $\boldsymbol{x}_i$ are the covariates of the $i - th$ observation and $\boldsymbol{\beta}$ are the coefficients of the model. The coefficients $(\xi, \boldsymbol{\beta})$ are estimated by Maximum Likelihood numerically and their asymptotic distributions are obtained by the asymptotic properties of Maximum Likelihood Estimators.

A Hurdle Generalized Pareto Distribution (HGPD) may be developed in a similar manner as the HGλD model, as it is enough to add a parameter λ_0 to the GPD to represent its probability mass at zero, and then estimate the parameters accordingly: the estimate of λ_0 is the proportion of zeros in the sample and the estimate of (ξ, τ) is the Maximum Likelihood estimate of the GPD fitted to the non-zero data values.

A HGPD GLM is obtained by replacing $(\boldsymbol{W}^T\boldsymbol{\beta} + \epsilon)$ in expression (25) by a random variable $U|\boldsymbol{W}^T$ that has a GPD with parameters $(\alpha, \xi, \mu = \exp(\boldsymbol{W}^T\boldsymbol{\beta}))$. The parameters related to the probability mass at zero and to the GPD GLM for the non-zero values are orthogonal, so that their estimation may be performed independently. This hurdle model is a special case of the Zero-inflated Truncated Generalized Pareto Distribution introduced by Couturier and Victoria-Feser (2010).

In order to establish the goodness-of-fit of the HGPD GLM we may consider the zero and non-zero data values separately. For the zero values we consider logistic regression diagnostic techniques and for the non-zero values we propose the study of two types of residuals: normalized quantile residuals and error residuals, that are given respectively by

$$r = \Phi^{-1}\left(F_{(\alpha,\hat{\xi},\hat{\mu})}(y)\right) \quad e = \frac{y - \alpha}{\hat{\mu}}$$

in which $F_{(\alpha,\hat{\xi},\hat{\mu})}$ is the cumulative probability function of a GPD with parameters $(\alpha, \hat{\xi}, \hat{\mu})$, for $y \neq 0$. If the model is well-fitted then r is normally distributed and e has a

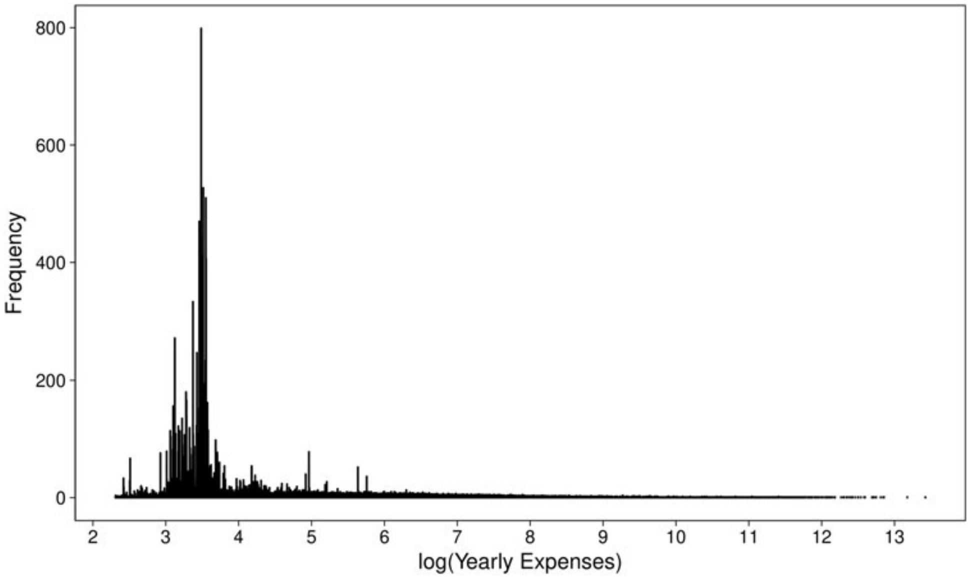

Figure 1. Frequency of each yearly expense value greater than zero in the logarithm scale.

GPD with parameters $\alpha_e = 0, \xi_e = \hat{\xi}$ and $\mu_e = 1$, so that graphical tools, as QQ-plots, may be used to establish goodness-of-fit. The error residuals were proposed by Couturier and Victoria-Feser (2010) where more details are presented.

6.1. The dataset

In order to fit a model to the data at hand, we first observe some systematic behavior of the data and transform it to obtain a better fit. First of all, there are some yearly expenses which are observed in the dataset hundreds of times, as can be seen in Figure 1, for there are some simple medical procedures that have standardized costs. These repeated values make it hard to fit a continuous model, as some expenses have a probability mass greater than zero. Therefore, we truncate the yearly expenses at R$100, and consider all yearly expenses lesser than R$100 to be zero. This truncation is justified by the practical application of the fitted model, as the main interest in modeling healthcare expenses is in properly fitting the tail of the distribution, i.e., the yearly expenses that are dozens of times the expected one, so that low expenses, as those less than R$100, may be regarded as zero without any loss for the practical application of the model. Even though 69% of the dataset has an expense less than R$100, their expenses sum to R$2,552,800, which is less than 2% of the total expenses of the dataset, that is R$137,382,575.

Truncating the dataset at R$100, we have, for each year and for the whole dataset, the proportion of zeros, selected percentiles, mean, standard deviation and maximum expense displayed in Table 2. The percentiles, mean and standard deviation refer to the truncated data, i.e., are calculated considering only data values greater than R$100. From Table 2 it can be seen that the 99th percentile is approximately twice the 98th percentile, the same occurring with the 99.5th and 99th percentiles. Furthermore, the

Table 2. Descriptive statistics of the yearly expenses. The percentiles, mean and standard deviation refer to the truncated data values, i.e., consider only the yearly expenses which are greater than R$100.

Year	2006	2007	2008	2009	All data
Size	70,186	71,814	73,038	74,418	289,456
Percentage of < R$ 100	60	81	82	51	69
Percentiles					
25	195	151	154	249	190
50	367	237	259	514	366
75	807	453	510	1,147	830
90	1,901	1,033	1,180	2,782	2,040
95	3,610	2,309	2,737	5,662	4,145
96	4,412	3,035	3,632	7,123	5,186
97	5,756	4,227	4,950	9,593	6,878
98	8,337	6,633	7,906	14,754	10,629
99	15,971	13,168	15,583	27,675	20,023
99.5	28,700	22,990	29,614	50,549	35,821
99.9	106,744	61,111	68,552	151,114	116,643
Maximum	377,862	295,736	279,450	675,440	675,440
Mean	1,313	870	991	2,028	1,485
Standard deviation	7,082	4,689	4,898	10,966	8,387

99.9th percentile is around three times the 99.5th percentile and the maximum is up to almost five times the 99.9th percentile, which shows that the dataset has heavy tails, as can be also seen in the box-plots of the logarithm of the yearly expenses in Figure 2.

Figure 3 shows the dispersion of the logarithm of the yearly expenses by each one of the covariates that are considered on the regression model, i.e., age, sex and the logarithm of the previous year expenses. The data considered for the regression model contemplate the yearly expenses of 2007, 2008 and 2009, and regards only customers that were enrolled in the insurance program in the considered year and in the previous year, which amounts to 214,925 observations. Figure 3 does not yield any clear relation between the logarithm of the yearly expenses and age or previous year expenses, although it seems that women tend to have greater yearly expenses than men.

6.2. HGλD model fit

We first fit HGλDs and HGPDs to the yearly expenses of each year (2006, 2007, 2008 and 2009) without considering any covariate. All models are fitted to the logarithm of the yearly expenses in order to obtain better fitted models and for computation optimization, as the non-transformed data has some extreme outliers, which makes it hard to fit a model properly. The goodness-of-fit is established graphically by the use of QQ-plots and the histogram of the data superimposed by the estimated curves. The fitted distributions are also compared with the kernel density estimate in order to establish which is the model that best fit the data objectively. See Bickel and Rosenblatt (1973) and Fan (1994) for examples of how the kernel density estimation is used for assessing goodness-of-fit. We apply the method proposed by Sheather and Jones (1991) in order to choose the bandwidth of the kernel estimate, and select the probability density function of a standard normal distribution as the kernel. For more details on kernel estimation see Silverman (1986).

In order to compare the fitted distributions to the kernel density estimate we use three different distance measures: the global distance, the L^2 norm and the L^∞ norm

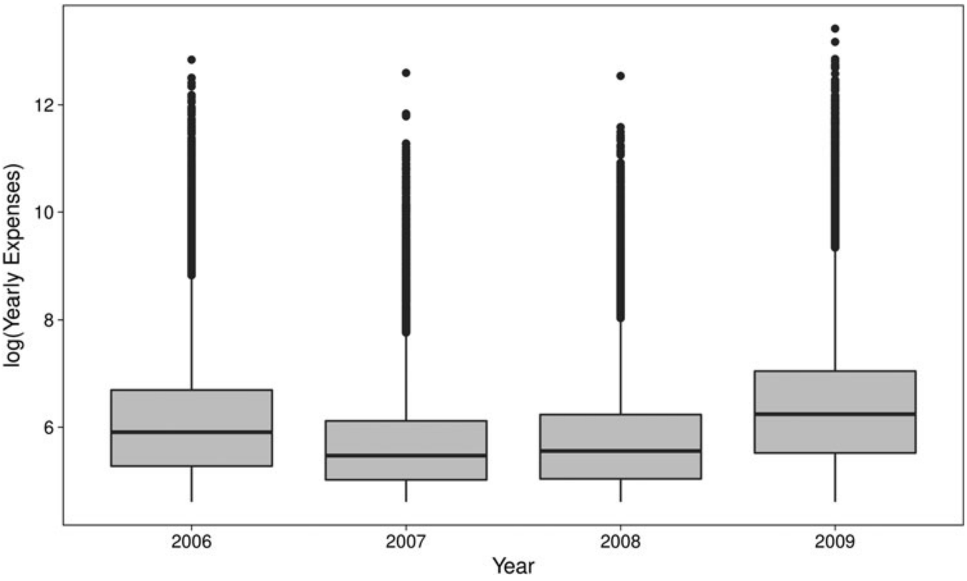

Figure 2. Box-plot of the logarithm of the yearly expenses by year. The expenses lesser than R$100 were omitted for a better visualization.

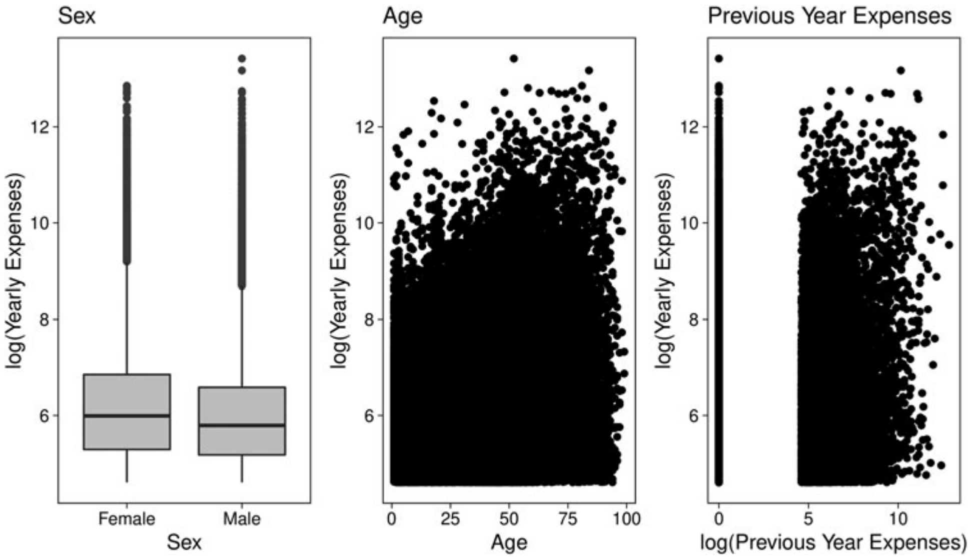

Figure 3. Dispersion of the logarithm of the yearly expenses by each one of the covariates. The yearly expenses lesser than R$100 were omitted for a better visualization.

that are given respectively by $D(\hat{f},\hat{f}_k) = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}(y_i)-\hat{f}_k(y_i)]^2$, $||\hat{f}-\hat{f}_k||_2 = [\int_0^{+\infty}|\hat{f}(y)-\hat{f}_k(y)|^2dy]^{1/2}$ and $||\hat{f}-\hat{f}_k||_\infty = \max_{y\in\mathbb{R}^+}|\hat{f}(y)-\hat{f}_k(y)|$, in which $\hat{f}$ is the parametric probability density function (GλD or GPD) fitted to the non-zero yearly expenses and $\hat{f}_k$ is the kernel density estimate. Note that the probability mass at zero is the same for all fitted distributions, so there is no need to compare them regarding the zero valued yearly expenses.

Table 3. Estimated parameters for the HGλD and HGPD models fitted to the yearly expenses, for each year.

			GλD				GPD		
Year	λ_0	Par	λ_1	λ_2	λ_3	λ_4	Scale	Shape	Location
2006	0.60	RS	4.74	0.12	0.0032	0.20	1.80	−0.22	4.61
		FKML	5.74	1.13	0.78	0.03			
2007	0.81	RS	4.62	0.07	0.0002	0.08	1.22	−0.09	4.61
		FKML	5.30	1.37	1.05	−0.07			
2008	0.82	RS	4.61	0.10	0	0.14	1.33	−0.12	4.61
		FKML	5.39	1.43	0.89	−0.10			
2009	0.51	RS	5.20	0.11	0.02	0.18	2.17	−0.25	4.61
		FKML	6.06	1.07	0.64	0.04			

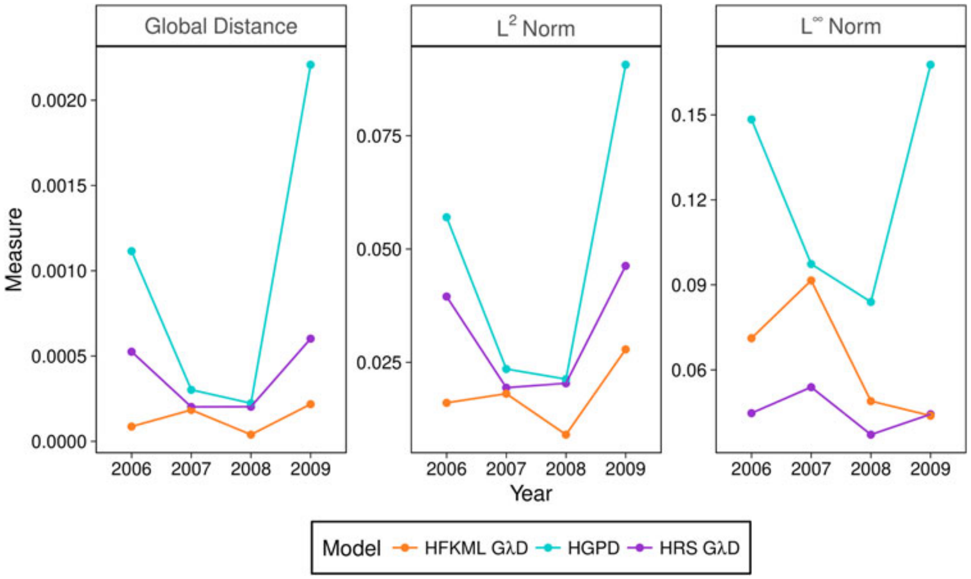

Figure 4. The distance between the fitted curve and the kernel density estimate for each model and year.

The estimated parameters for each year and model are displayed in Table 3. The estimated parameters differ significantly from one year to another, for all fitted models, although we observe in every year that the fitted GλDs are highly skewed, as the values of λ_3 and λ_4 are quite different. In Figure 4 we see that the densities estimated by the HRS and HFKML GλD are closer to the kernel estimate density for all years, by all distance measures. Furthermore, Figure 5 displays the histogram of the logarithm of the yearly expenses superimposed by the fitted GλDs and HGPD, and the QQ-plots between the empirical and fitted distributions, for all years, from which it can be seen that the HGλDs fit the data better for low values (near the threshold 4.61), and that the HRS GλD and HGPD fit as good the tail, while the HFKML GλD seems to fit it poorer.

From the diagnostic plots in Figure 5 we see that the major advantage of the HGλDs over the HGPD is that they are not necessarily threshold modal and monotonically decreasing so that they fit better the bulk of the distribution, i.e., the values near the threshold, when the distribution mode is greater than the threshold. Nevertheless, the

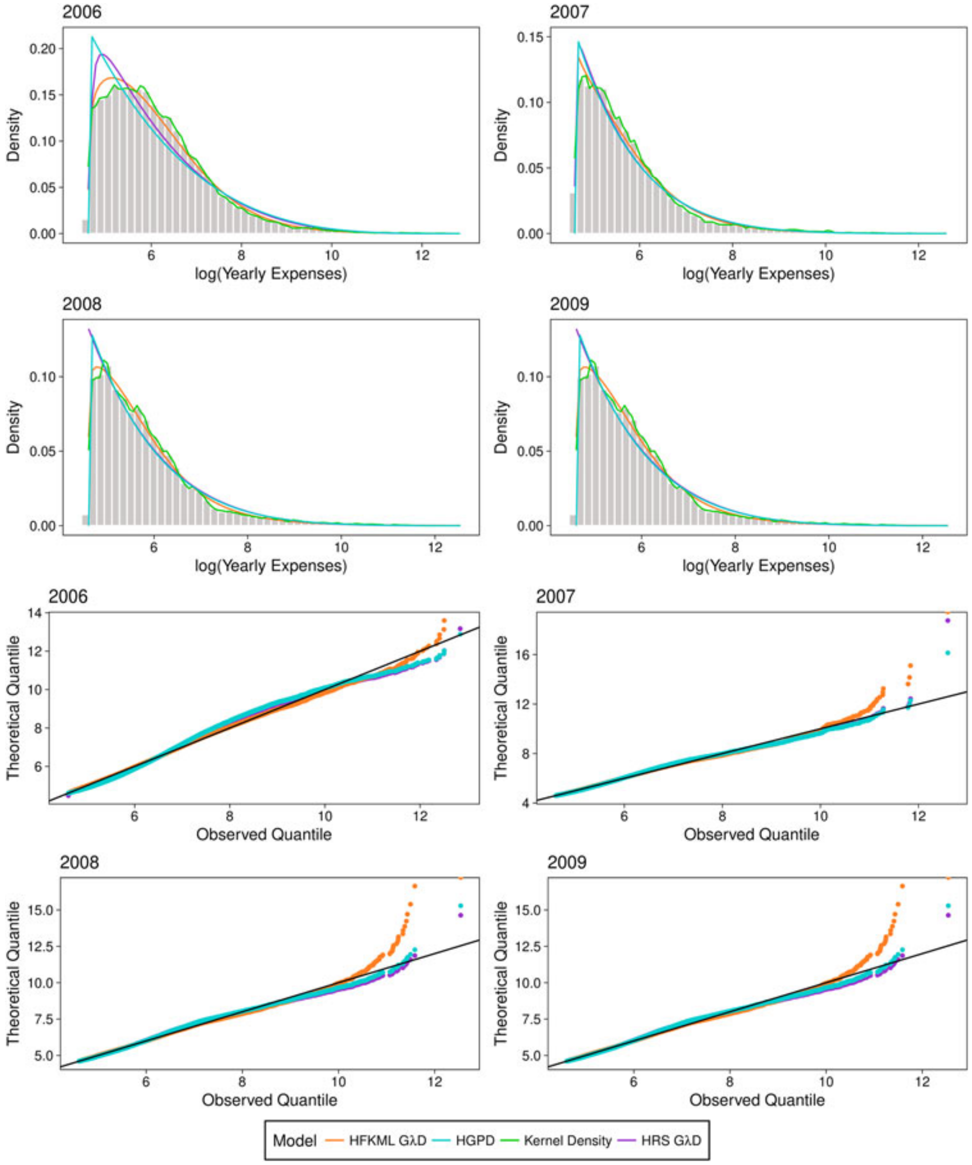

Figure 5. The top four plots display the histogram of the data superimposed by the estimated HGλDs and HGPD, for each year. The bottom four plots display the QQ-plot between the sample quantiles and the theoretical quantiles of the HGλDs and HGPD, for each year.

HRS GλD and the HGPD fit the tail of the distribution better, while the HFKML GλD fits its bulk better, for it is the distribution with best overall fit according to the distance measures. Therefore, the HGλDs fit the data better, especially the HRS GλD, although the HGPD fits the right tail of the distribution as good as them.

In order to choose between the proposed hurdle models, one should observe the nature of the data he seeks to fit. Indeed, although the GPD has a highly flexible right tail, which makes it useful for fitting heavy tailed data, its left tail is not quite flexible, which makes it a poor choice for modeling data that demands flexibility in both tails.

Table 4. Estimated parameters of the logistic regression that models the logit of the yearly expenses being less than R$100.

Parameter	Estimate	SE	*t* value	*p*-value
Intercept	1.6266	0.0121	134.8826	$< 2e{-}16$
LE	−0.1253	0.0018	−71.4760	$< 2e{-}16$
Male	0.2093	0.0098	21.2574	$< 2e{-}16$
Age	−0.0159	0.0002	−63.9644	$< 2e{-}16$

SE, standard deviation; LE, logarithm of the previous year expenses.

Table 5. Estimated parameters and numerical confidence intervals of the HRS GλD and HFKML GλD regression models.

			Confidence interval	
Parametrization	Parameter	Estimate	LB (0.5%)	UB (99.5%)
HFKML GλD	Intercept	6.13	6.11	6.24
	LE	−0.0000215	−0.0050842	0.0016275
	Male	−0.0003554	−0.2338218	0.0392448
	Age	0.0000259	−0.0002445	0.0003319
	λ_1	−0.41	−0.59	−0.29
	λ_2	1.07	0.94	1.56
	λ_3	0.84	0.50	1.02
	λ_4	0.02	−0.26	0.09
HRS GλD	Intercept	6.10	6.08	6.11
	LE	0.0013937	0.0005463	0.0023140
	Male	−0.0126310	−0.0182945	−0.0074669
	Age	0.0009363	0.0007947	0.0010634
	λ_1	−1.41	−1.43	−1.40
	λ_2	0.1102	0.1061	0.1142
	λ_3	0.0023749	0.0021813	0.0025770
	λ_4	0.19	0.18	0.20

LE, logarithm of the previous year expenses; LB, lower bound; UB, upper bound.

On the other hand, both tails of the GλD are flexible, so that it is a more robust choice when comparing to the GPD. As the parametrizations of the RS and FKML GλD are not equivalent, one must also choose between them, what may be done by observing the quality of each fit by applying tools as the distance to the kernel estimate or diagnostic plots.

6.3. HGλD regression model

In this section, HGλD regression models are fitted to the logarithm of the yearly expenses and compared with the HGPD GLM by the use of error and normalized quantile residuals. The estimated coefficients of the logistic regression, i.e., the parameters of the model for the probability mass at zero, are the same for all fitted models, as they are orthogonal to the parameters of the models for the non-zero values. Also, the logit modeled in the logistic regression is the logit of the expense being less than R$100, as the yearly expenses were truncated at R$100. We assume that, given the logarithm of the previous year expenses, the age and the sex, the logarithm of the yearly expenses are independent, even the expenses that refer to the same customer in different years, so that we have a sample of the model variables.

The estimated parameters of the logistic regression for the zero-valued data are presented in Table 4, in which the contrast used for the sex is “treatment” in which the female sex is

Table 6. Estimated parameters and *p*-values of the HGPD model.

Parameter	Estimate	SE	*p*-value
Shape	0.9924	2.58e-13	$< 2e{-}16$
Intercept	4.6576	0.0091	$< 2e{-}16$
LE	0.0152	0.0021	$8.26e{-}13$
Male	−0.1198	0.0125	$< 2e{-}16$
Age	0.0082	0.0003	$< 2e{-}16$

SE, standard deviation; LE, logarithm of the previous year expenses.

the base. The minus sign of the estimated coefficients of the logarithm of the previous year expenses and age shows that as greater the previous year expense or the age of a customer, the lesser is the probability of him having less than R$100 in yearly expenses, while the plus sign of the estimated coefficient for the male sex shows that men are more likely to have yearly healthcare expenses lesser than R$100 than women.

The estimated parameters of both parametrizations of the GλD regression and of the GPD GLM for the non-zero data values are presented in Tables 5 and 6, in which the female sex is again taken as the base for the “treatment” contrast of sex. On the one hand, as the zero is in the 99% confidence interval of all coefficients of the HFKML GλD model, there is no evidence that the location of the distribution depends on any of the covariates at a significance of 1% and we may regard these coefficients as zero. On the other hand, all the coefficients of the HRS GλD and HGPD model are different of zero at a significance of 1%, so that we regard only the estimated coefficients of these models.

The signs of the estimated coefficients of the HRS GλD and HGPD models are exchanged when comparing with the signs of the ones in Table 4, which is consistent. Indeed, we see that as greater the previous year expenses or the age, the greater is the location parameter of the HGλD and the mean of the HGPD, and that the location (and mean) of the male sex is lesser than the female’s, confirming what were was observed in the box-plot in Figure 3. Therefore, we obtain the same kind of interpretation for the yearly expenses from the logistic regression, HRS GλD model and HGPD GLM: as greater the previous year expenses or the age, the greater the expense; and women have greater expense than men.

The diagnostic plots for the HGλD models and the HGPD GLM are presented in Figures 6 and 7. Figure 6 displays plots of the normalized quantile residuals, while Figure 7 displays plots of the error residuals. Figure 6 yields that the HRS GλD and HFKML GλD regression models are fairly fitted, as the distributions of their normalized quantile residuals do not greatly deviate from the normal distribution. Furthermore, from Figure 7 it may be established that the HRS GλD and HFKML GλD models are well-fitted, as the points of their error residuals QQ-plot are distributed around the line with intercept zero and slope one. In fact, when comparing with the HGPD GLM, the HRS GλD and HFKML GλD regression models seem to better fit the data.

On the other hand, the fit of the HGPD GLM is not good, as its error residuals do not seem to be distributed as a GPD and its normalized quantile residuals are highly skewed. The HGPD GLM does not properly fit the residuals because the data is not threshold modal and the fitted distribution is supposed to have infinity mean, as can be seem from the estimate of the shape parameter that is close to one. The lack of flexibility of its left tail makes the GPD improper to fit data that presents a behavior on it that is not threshold modal and monotonically decreasing. Furthermore, the GLM

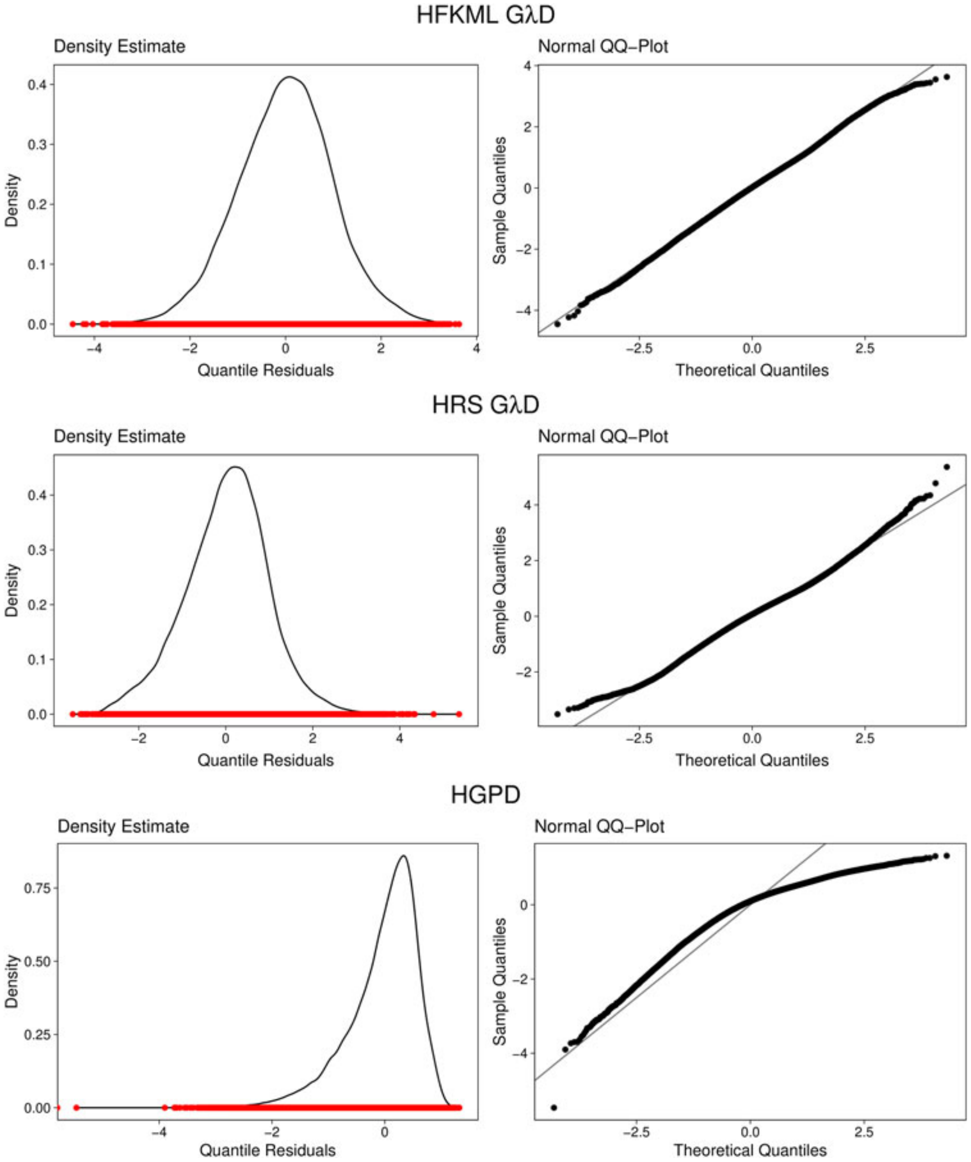

Figure 6. Estimated density and Normal QQ-plot of the normalized quantile residuals of the HRS GλD and HFKML GλD regression models, and the HGD GLM, for the non-zero yearly expenses.

framework is restricted to GPDs that have finite mean, i.e., such that $\xi < 1$. On the other hand, the GλD is exactly the opposite of the GPD in the matter of tail flexibility, as its tails may have different shapes. Moreover, the HGλD models the location of the distribution, so that it may fit distributions with infinite mean.

In general, when choosing between the proposed hurdle regression models, one must take into account the statistical significance of their coefficients, and carefully analyze the behavior of the normalized quantile and error residuals. The GλD regression models are more robust, as are also adequate when the conditional distribution of the response variable given the covariates has infinite mean or is not monotonically decreasing with the threshold as the mode.

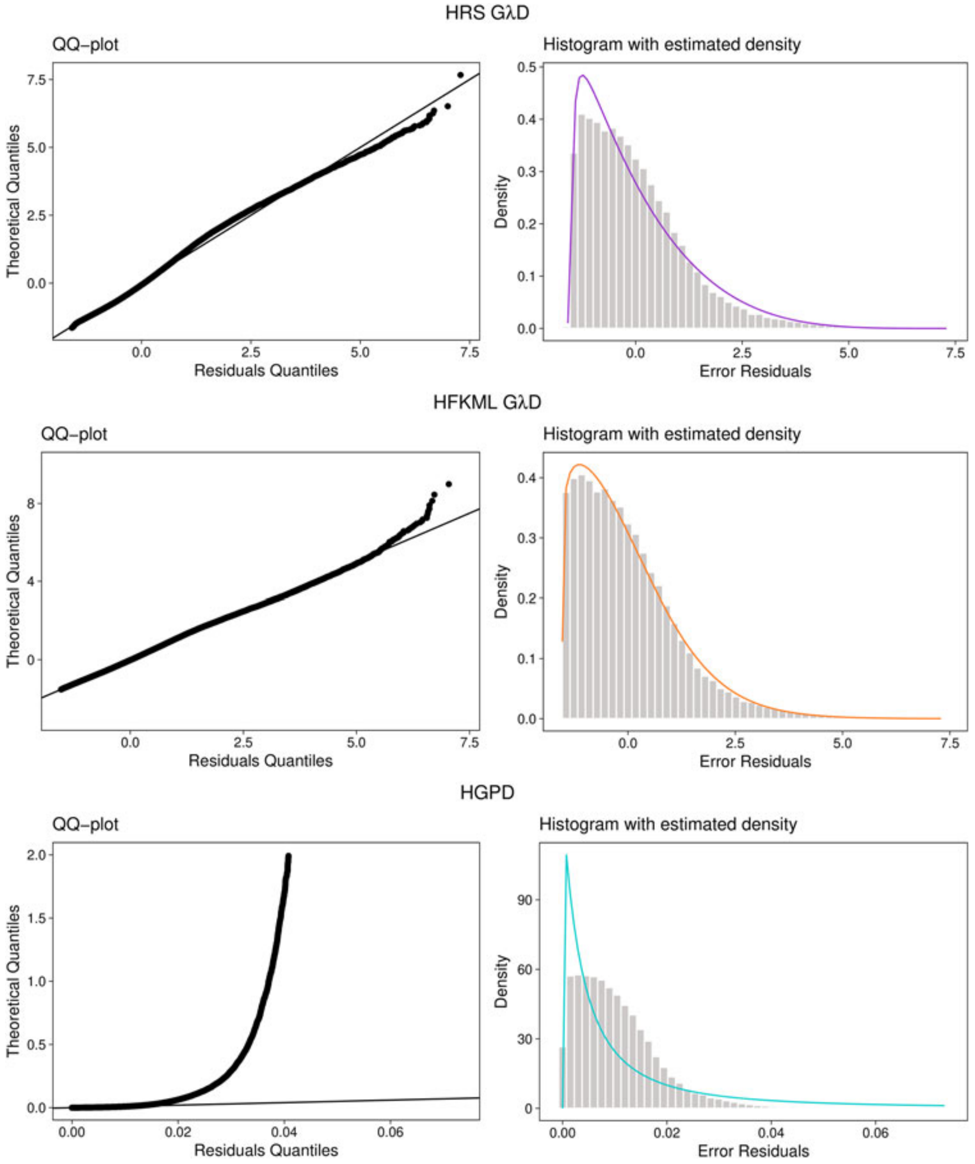

Figure 7. Diagnostics for the HRS GλD and HFKML GλD regression models, and the HGPD GLM, for the non-zero yearly expenses. The histograms are that of the respective error residuals and are superimposed by their theoretical distribution. The QQ-plots compare the empirical quantiles of the error residuals with their theoretical quantiles.

Nevertheless, one has also to choose between the RS and FKML GλD, which are not equivalent and, in order to do so, must carefully analyze both models, and choose the one that best fulfills the objective of the regression, e.g., best predicts an outcome or best fit the dataset.

An interesting feature of the HGλD regression models is that the fitted curve takes into account the probability mass at zero, so that we may readily see which are the profiles, i.e., combinations of the covariate's levels, that tend to have great and low expenses. As an example, we consider 12 profiles, that are presented in Table 7 and

Table 7. The covariates of each profile, their location, λ_0 and selected estimated percentiles for the yearly expenses from the HRS GλD regression model.

						Selected Percentiles					
Profile	Age	Sex	LE	λ_0	Location	75th	90th	95th	99th	99.5th	99.9th
1	20	F	0	0.79	452.53	0	360.78	948.72	5768.93	10763.16	34689.11
2	20	F	7	0.61	456.97	228.98	865.29	2040.61	10176.05	17736.15	50318.16
3	40	F	0	0.73	461.08	122.14	522.95	1314.88	7374.60	13383.05	40952.71
4	40	F	7	0.53	465.60	309.58	1115.03	2554.16	12069.66	20650.44	56586.22
5	60	F	0	0.66	469.80	183.74	725.48	1754.89	9169.98	16241.66	47487.74
6	60	F	7	0.45	474.40	398.82	1381.68	3089.37	13958.85	23516.69	62600.83
7	20	M	0	0.82	446.85	0	275.79	749.75	4834.56	9202.52	30793.22
8	20	M	7	0.65	451.23	182.43	716.00	1726.10	8963.52	15842.01	46133.72
9	40	M	0	0.77	455.30	0	411.71	1065.33	6293.60	11626.53	36784.42
10	40	M	7	0.58	459.76	255.32	947.96	2212.09	10817.96	18728.74	52469.37
11	60	M	0	0.71	463.90	141.94	587.78	1457.46	7969.18	14336.51	43159.53
12	60	M	7	0.50	468.45	339.34	1204.89	2735.73	12717.94	21637.32	58667.27

The location and percentiles are exponentiated. LE, logarithm of the previous year expenses.

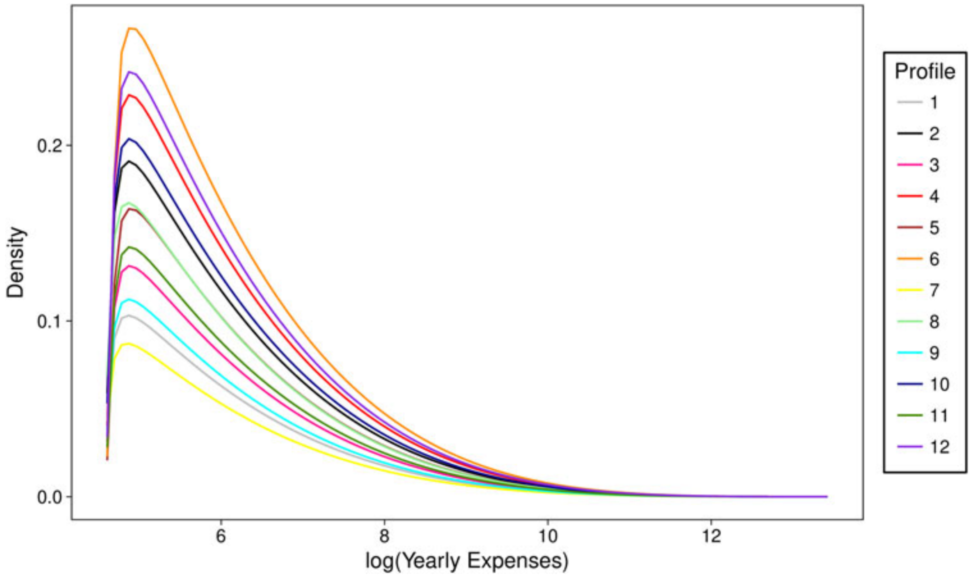

Figure 8. Estimated curves for the profiles in Table 7 given by the HRS GλD regression model.

whose HRS GλD fitted curves are displayed in Figure 8. On the one hand, the location of the curves is almost the same for all profiles, even though there are profiles that differ reasonably on all the covariates. On the other hand, the probability mass at zero differs significantly from one profile to another, as can be seen from the area under each curve. The exponential of selected percentiles for the 12 profiles are presented in Table 7, in which we observe that the percentiles differ significantly from one profile to another and their values are a reflex of the estimated coefficients of Tables 4 and 5.

7. Final remarks

The HGλD models proposed in this paper have a great potential for applications, not only to healthcare expenses data, but also to any highly skewed data, with excessive zeros and heavy tails. According to the results obtained in Section 6, we may argue that the HGPD is in general as good as the HGλD when fitting unimodal monotonically decreasing distributions, while the HGλD seems to better fit data that demands a higher flexibility on the left tail. Therefore, the methods developed in this paper bring contributions to the state-of-the-art in modeling heavy tailed clumped-at-zero data.

Although the HGλD fits best some kinds of data, it is still necessary to improve its estimation techniques, especially what concerns the asymptotic properties of the estimators and the computation of the estimates, which may take days, depending on the size of the dataset and the number of parameters. Therefore, a more theoretical research about the HGλD and the optimization of the algorithms used to estimate its parameters are interesting topics for future researches.

Supplementary material

The data analysis has been performed in the 3.4.2 version of **R** (R Core Team, 2017) by the adaptation of functions of the **GAMLSS** (Rigby and Stasinopoulos, 2005), GLDEX (Su, 2007a) and **GLDReg** (Su, 2016) packages. In the on-line supplementary material we provide an **R** package with functions to all the models of this paper and an **R** script that reproduce all tables and figures of this paper.

Acknowledgements

We would like to thank *Sabesprev* who kindly provided the dataset used in this paper.

ORCID

D. Marcondes http://orcid.org/0000-0002-6087-4821

References

Balasooriya, U., and C.-K. Low. 2008. Modeling insurance claims with extreme observations: transformed kernel density and generalized lambda distribution. *North American Actuarial Journal* 12 (2):129–42. doi:10.1080/10920277.2008.10597507.

Bickel, P. J., and M. Rosenblatt. 1973. On some global measures of the deviations of density function estimates. *The Annals of Statistics* 1 (6):1071–95. http://www.jstor.org/stable/2958266. doi:10.1214/aos/1176342558.

Cebrián, A. C., M. Denuit, and P. Lambert. 2003. Generalized Pareto fit to the society of actuaries large claims database. *North American Actuarial Journal* 7 (3):18–36. doi:10.1080/10920277.2003.10596098.

Corrado, C. J. 2001. Option pricing based on the generalized lambda distribution. *Journal of Futures Markets* 21 (3):213–36. doi:10.1002/1096-9934(200103)21:3<213::AID-FUT2>3.0.CO;2-H.

Couturier, D.-L., and M.-P. Victoria-Feser. 2010. Zero-inflated truncated generalized Pareto distribution for the analysis of radio audience data. *The Annals of Applied Statistics* 4 (4): 1824–46. http://www.jstor.org/stable/23362450. doi:10.1214/10-AOAS358.

Cox, D. R., and N. Reid. 1987. Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)* 49 (1):1–39. http://www.jstor.org/stable/2345476.

Duan, N., W. G. Manning, C. N. Morris, and J. P. Newhouse. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* 1 (2): 115–26. doi:10.2307/1391852.

Dunn, P. K., and G. K. Smyth. 1996. Randomized quantile residuals. *Journal of Computational and Graphical Statistics* 5 (3):236–44. doi:10.2307/1390802.

Fan, Y. 1994. Testing the goodness of fit of a parametric density function by Kernel method. *Econometric Theory* 10 (2):316–56. doi:10.1017/S0266466600008434.

Fournier, B., N. Rupin, M. Bigerelle, D. Najjar, and A. Iost. 2006. Application of the generalized lambda distributions in a statistical process control methodology. *Journal of Process Control* 16 (10):1087–98. doi:10.1016/j.jprocont.2006.06.009.

Freimer, M., G. Kollia, G. S. Mudholkar, and C. T. Lin. 1988. A study of the generalized Tukey lambda family. *Communications in Statistics-Theory and Methods* 17 (10):3547–67. doi:10.1080/03610928808829820.

Grimshaw, S. D. 1993. Computing maximum likelihood estimates for the generalized Pareto distribution. *Technometrics* 35 (2):185–91. doi:10.1080/00401706.1993.10485040.

Hastings, C., F. Mosteller, J. W. Tukey, and C. P. Winsor. 1947. Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics* 18 (3):413–26. http://www.jstor.org/stable/2235737. doi:10.1214/aoms/1177730388.

Hilbe, J. 2009. *Logistic regression models*. Abingdon, UK: Taylor & Francis. https://books.google.com.br/books?id=eJcMIAAACAAJ.

Hosking, J. R., and J. R. Wallis. 1987. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics* 29 (3):339–49. doi:10.2307/1269343.

Hyndman, R. J., and Y. Fan. 1996. Sample quantiles in statistical packages. *The American Statistician* 50 (4):361–5. doi:10.2307/2684934.

Jones, A. M., J. Lomas, and N. Rice. 2014. Going beyond the mean in healthcare cost regressions: A comparison of methods for estimating the full conditional distribution (Technical Report). HEDG, c/o Department of Economics, University of York.

Karian, Z. A., and E. J. Dudewicz. 1999. Fitting the generalized lambda distribution to data: a method based on percentiles. *Communications in Statistics-Simulation and Computation* 28 (3): 793–819. doi:10.1080/03610919908813579.

Karian, Z. A., and E. J. Dudewicz. 2000. *Fitting statistical distributions: The generalized lambda distribution and generalized bootstrap methods*. Boca Raton, FL: CRC Press.

Karian, Z. A., and E. J. Dudewicz. 2003. Comparison of GLD fitting methods: Superiority of percentile fits to moments in l2 norm. *Journal of the Iranian Statistical Society* 2 (2):171–87.

Karian, Z. A., E. J. Dudewicz, and P. Mcdonald. 1996. The extended generalized lambda distribution system for fitting distributions to data: History, completion of theory, tables, applications, the" final word" on moment fits. *Communications in Statistics-Simulation and Computation* 25 (3):611–42. doi:10.1080/03610919608813333.

King, R. A., and H. MacGillivray. 1999. A starship estimation method for the generalized lambda distributions. *Australian & New Zealand Journal of Statistics* 41 (3):353–74. doi:10.1111/1467-842X.00089.

Lakhany, A., and H. Mausser. 2000. Estimating the parameters of the generalized lambda distribution. *ALGO Research Quarterly* 3 (3):47–58.

Lambert, D. 1992. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34 (1):1–14. doi:10.2307/1269547.

Mihaylova, B., A. Briggs, A. O'Hagan, and S. G. Thompson. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Economics* 20 (8):897–916. doi:10.1002/hec.1653.

Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33 (3):341–65. doi:10.1016/0304-4076(86)90002-3.

Nelder, J. A., and R. J. Baker. 1972. *Generalized linear models*. Hoboken, NJ: Wiley Online Library.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal* 7 (4):308–13. doi:10.1093/comjnl/7.4.308.

Öztürk, A., and R. Dale. 1982. A study of fitting the generalized lambda distribution to solar radiation data. *Journal of Applied Meteorology* 21 (7):995–1004. doi:10.1175/1520-0450(1982)021<0995:ASOFTG >2.0.CO;2.

Öztürk, A., and R. F. Dale. 1985. Least squares estimation of the parameters of the generalized lambda distribution. *Technometrics* 27 (1):81–4. doi:10.2307/1270473.

Pal, S. 2004. Evaluation of nonnormal process capability indices using generalized lambda distribution. *Quality Engineering* 17 (1):77–85. doi:10.1081/QEN-200028709.

Pickands, J. 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* 3 (1):119–31. http://www.jstor.org/stable/2958083.

R Core Team. 2017. *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria: R: The R Foundation. https://www.R-project.org/.

Ramberg, J. S., and B. W. Schmeiser. 1974. An approximate method for generating asymmetric random variables. *Communications of the ACM* 17 (2):78–82. doi:10.1145/360827.360840.

Rigby, R. A., and D. M. Stasinopoulos. 2005. Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54: 507–54. doi:10.1111/j.1467-9876.2005.00510.x.

Sheather, S. J., and M. C. Jones. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 53 (3): 683–90.

Silverman, B. W. 1986. *Density estimation for statistics and data analysis*. Vol. 26. Boca Raton, FL: CRC Press.

Su, S. 2005. A discretized approach to flexibly fit generalized lambda distributions to data. *Journal of Modern Applied Statistical Methods* 4 (2):7.

Su, S. 2007a. Fitting single and mixture of generalized lambda distributions to data via discretized and maximum likelihood methods: GLDEX in R. *Journal of Statistical Software* 21 (9):1–17.

Su, S. 2007b. Numerical maximum log likelihood estimation for generalized lambda distributions. *Computational Statistics & Data Analysis* 51 (8):3983–98. doi:10.1016/j.csda.2006.06.008.

Su, S. 2011. Maximum log likelihood estimation using em algorithm and partition maximum log likelihood estimation for mixtures of generalized lambda distributions. *Journal of Modern Applied Statistical Methods* 10 (2):17.

Su, S. 2015. Flexible parametric quantile regression model. *Statistics and Computing* 25 (3): 635–50. doi:10.1007/s11222-014-9457-1.

Su, S. 2016. Fitting flexible parametric regression models with GLDreg in R. *Journal of Modern Applied Statistical Methods* 15 (2):46.

Tarsitano, A. 2004. Fitting the generalized lambda distribution to income data. COMPSTAT 2004 symposium, 1861–7.

Tukey, J. W. 1990. Practical relationship between the common transformations of percentages or fractions and of amounts. The Collected Works of John W. Tukey, Volume VI: More Mathematical, 211–9.